

$$
\begin{aligned}
& \text { Social Statistics: } \\
& \text { The Interplay among } \\
& \text { Censuses, Surveys and } \\
& \text { Administrative Data }
\end{aligned}
$$

Statistics Canada's XXVI International Symposium on Methodological Issues

## Catber 26-29,2010

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to $4: 30$ p.m., at the following numbers:

## Statistics Canada's National Contact Centre

Toll-free telephone (Canada and the United States):
Inquiries line 1-800-263-1136
National telecommunications device for the hearing impaired 1-800-363-7629
Fax line
1-877-287-4369
Local or international calls:
Inquiries line 1-613-951-8116
Fax line 1-613-951-0581
Depository Services Program
Inquiries line
1-800-635-7943
Fax line 1-800-565-7757

Mail:
Statistics Canada
100, Tunney's Pasture Driveway
Ottawa, Ontario
K1A 0T6

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" >"The agency" > "Providing services to Canadians."

## Symposium 2010 - Catalogue no. 11-522-XCB

## End-use licence agreement

Published by authority of the Minister responsible for Statistics Canada
© Minister of Industry, 2011
All rights reserved. The content of this electronic publication may be reproduced, in whole or in part, and by any means, without further permission from Statistics Canada, subject to the following conditions: that it be done solely for the purposes of private study, research, criticism, review or newspaper summary, and/or for non-commercial purposes; and that Statistics Canada be fully acknowledged as follows: Source (or «Adapted from", if appropriate): Statistics Canada, year of publication, name of product, catalogue number, volume and issue numbers, reference period and page(s). Otherwise, no part of this publication may be reproduced, stored in a retrieval system or transmitted in any form, by any means-electronic, mechanical or photocopy-or for any purposes without prior written permission of Licensing Services, Client Services Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

## For further information please contact:

Licensing Services
Client Services Division, Statistics Canada
R.H. Coats Building, 9th floor, section A

Ottawa, Ontario K1A 0T6, Canada
E-mail: licensing@statcan.gc.ca
Telephone: (613) 951-1122
Fax: (613) 951-1134

## SOCIAL STATISTICS: THE INTERPLAY AMONG CENSUSES, SURVEYS AND ADMINISTRATIVE DATA

## TABLE OF CONTENTS

PREFACE ..... 8
OPENING REMARKS
François Maranda, Statistics Canada. ..... 10
KEYNOTE ADDRESS
01A-1 Statistics without surveys? About the past, present and future of data collection in The Netherlands ..... 14
J. Bethlehem, Statistics Netherlands
SESSION 2A: SAMPLING FRAMES
02A-1 The OCTOPUSSE System of the New Master Sample at the Institut National de la Statistique et des Études Économiques (INSEE). ..... 23
M. Christine and S. Faivre, Institut National de la Statistique et des Études Économiques (France)
02A-2 The National Statistical Registry on Persons in Italy ..... 24
R. Vivio and C. Runci, Italian National Statistical Institute (Italy)
02A-3 Canadian Health Measures Survey, Cycle 2: Combining census and administrative data to improve survey frame efficiency ..... 33
F. Labrecque, S. Giroux and A. Quigley, Statistics Canada
SESSION 02B: POTENTIAL USES
02B-1 Accounting for temporal effects, sampling variability, incomplete data and reporting error in the integration of consumer expenditure data from survey and administrative-record sources ..... 42
J.L. Eltinge, Bureau of Labor Statistics (U.S.A.)
02B-2 The potential of administrative education data used to measure subnational migration in England. ..... 43
S. Jivraj, University of Manchester (U.K.)
02B-3 From neighbourhood statistics to beyond 2011 - Making better use of administrative data. ..... 50
L. Morris and M. Phillips, Office for National Statistics (U.K.)
02B-4 Validation of financial data using tax records ..... 56
P. Wanner, University of Geneva (Switzerland)
SESSION 02C: MULTI-MODEL DATA COLLECTION
02C-2 Using administrative data to find the best medium: Examples of mixed sources and mixed modes ..... 63
J. Hartkamp and H. Rutjes, DESAN Research Solutions, (The Netherlands)

02C-3 Accounting for the effects of data collection modes in population surveys surveys.

## SESSION 03A: MICROSIMULATIONS


03A-2 Using multiple data sources to project U.S. Social Security Finances and the Federal Budget in the long term. ..... 76
J.A. Schwabish, Congressional Budget Office (U.S.A.)03A-3 Integrating infectious and chronic disease information to model cervical cancer.81
M. Wolfson, Universityof Ottawa (Canada)

## SESSION 03B: SAMPLING FRAMES

$$
\begin{aligned}
& \text { 03B-1 Building the Dwelling Frame in Gauteng Province using a combination of statistical } \\
& \text { methods namely: Census, administrative data and survey data...................................................... } 83 \\
& \text { M. Mohale and H. Moletsane, Statistics South Africa (South Africa) }
\end{aligned}
$$

03B-2 Improving and using the Statistical Administrative Records System ..... 89L.B. Marshall and A.B. O'Hara, U.S. Census Bureau (U.S.A.)
03B-3 Use of the Canada Child Tax Benefit as a frame for the Survey of Young Canadians. ..... 95
M. Pantel, Statistics Canada03B-4 Targeting populations of interest in the sample design using the Census of Population:An example from the Ontario Survey on the Prevalence of Hypertension.96L. Stratychuk and J. Dumais, Statistics Canada

## SESSION 03C: POVERTY DATA

03C-2 Using a 'community survey' and previous census estimates to produce composite estimators for the South African Index of Multiple Deprivation 2007.103
C. Dibben, University of St. Andrews (U.K.), G. Wright and M. Noble, University of Oxford (U.K.)

03C-3 Using complete administration data for nonresponse analysis: The PASS Survey of LowIncome Households in Germany.
R. Schnell and T. Gramlich, University of Duisburg-Essen (Germany), A. Mosthaf and S. Bender, Institute for Employment Research (Germany)

## SESSION 04A: JUSTICE DATA

$$
\begin{aligned}
& \text { 04A-1 Operational needs versus data requirements: Challenges in providing policing data to } \\
& \text { Statistics Canada.................................................................................................................... } \\
& \text { T. McDonald, Royal Canadian Mounted Police (Canada) }
\end{aligned}
$$

04A-2 The Impact of changing concepts on estimates from administrative data: The case of the Integrated Criminal Courts Survey.
A. Matarazzo, Statistics Canada
04A-3 Opportunities and challenges in using administrative justice data for longitudinal and network research ..... 119
P.J. Carrington, University of Waterloo (Canada)
SESSION 04B: SAMPLING FRAMES 2
04B-2 Use of address databases as a basis for household frames: Confirmation bias in housing unit listing ..... 128
S. Eckman and F. Kreuter, Institute for Employment Research (Germany)
04B-3 Experiences using the census as a sampling frame: The case of the Access and Support to Education and Training Survey. ..... 129
E. Neusy and Y. Bélanger, Statistics Canada
04B-4 Using administrative data sources to target sampling in the Wealth and Assets Survey in the UK ..... 135
P. Smith, K. Ashworth and C. Lound, Office for National Statistics (U.K.)
SESSION 04C: LINKAGE OF HEALTH DATA
04C-1 Impact of identifying information requested on consent rates ..... 142
J. Fulton, University of Maryland (U.S.A.)
04C-2 When the Census and administrative data combine to illustrate social inequalities in health in Canada ..... 143
D. Hamel, R. Pampalon and P. Gamache, Institut national de santé publique du Québec (Canada)
04C-3 Indigenous life expectancy using multiple Australian data sources149R. Madden, University of Sydney (Australia), L. Tickle, Macquarie University (Australia), L.Jackson Pulver, University of New South Wales (Australia), I. Ring, University of Wollongong(Australia) and L. Taylor, New South Wales Department of Health (Australia)
04C-4 1991-2011 Canadian Census Mortality and Cancer Follow-up Study. ..... 150
P.A. Peters and M. Tjepkema, Statistics Canada
SESSION 05A: WAKSBERG AWARD WINNER ADDRESS
05A-1 The organization of statistical methodology and methodological research in national statistical offices ..... 158
I.Fellegi, Statistics Canada
SESSION 06A: NEW DEVELOPMENTS IN RECORD LINKAGE METHODS
06A-1 Glink - A Probabilistic Record Linkage System ..... 160
A. Chevrette and M. Wenzowski, Statistics Canada
06A-2 Record linkage methods and techniques as proposed in RELAIS ..... 161
N. Cibella, M. Fortini, D. Ichim and T. Tuoto, Italian National Statistical Institute (Italy)
06A-3 Obtaining estimates of parameters for probabilistic record linkage using the EM algorithm ..... 170
W.E. Yancey, U.S. Census Bureau (U.S.A.)
SESSION 06B: MULTIPLE FRAMES
06B-1 The optimal estimator in dual frame surveys using indirect sampling. ..... 172
M. Maia, Catholic University of Portugal and P. Vicente, Lisbon University Institute (Portugal)
06B-2 Demographic analysis in 2010: Methodological challenges ..... 173
K. West, J. G. Robinson, J. Devine and R. Bhaskar, U.S. Census Bureau (U.S.A.)
06B-3 Dual-frame weighting and estimation challenges for the Canadian Community Health Survey ..... 179
K. Wilder and S. Thomas, Statistics Canada
SESSION 06C: DATABASES I
06C-1 The Research Data Centre-in-Research Data Centre approach. ..... 187
S. Bender and J. Heining, Institute for Employment Research (Germany)
06C-2 The Canadian Forces portrait: A Case study on analyzing administrative and survey data to develop a personnel management resource ..... 188
K. Daley, Royal Canadian Mounted Police (Canada)
SESSION 07A: TREATMENT OF ITEM NONRESPONE IN SOCIAL SURVEYS
07A-1 Imputing underreported treatments using multiple sources of treatment information in a Cancer Services Study ..... 195
A. Zaslavsky and Y. He, Harvard Medical School (U.S.A.)
07A-2 Proxy pattern-mixture analysis for survey nonresponse ..... 196
R.R. Andridge, Ohio State University (U.S.A.) and R.J.A. Little, University of Michigan (U.S.A.)
07A-3 Calibrated robust imputation in surveys. ..... 205D. Haziza and P. Duchesne, University of Montreal (Canada)
SESSION 07B: ADMINISTRATIVE CENSUSES
07B-1 Methodology in the Swedish register-based census. ..... 207
M. Axelson, D. Hedlin, A. Holmberg and I. Jansson, Statistics Sweden
07B-2 Making better use of administrative data: How far can New Zealand go?. ..... 211
C.Bycroft, Statistics New Zealand
07B-3 First register-based census in Slovenia: How to convert administrative sources toStatistics.217
D. Dolenc, Statistical Office of the Republic of Slovenia
SESSION 07C: DATABASES II
07C-1 Historical data linkage of tax records on labour and income: The case of the Living in Canada Survey pilot ..... 225
A. Heisz, M Langevin, J. Randle, Statistics Canada
07C-2 Creation of a child-centric inter-agency data warehouse: The Longitudinal Study ofEarly Development232M.R. Pfeiffer, M. Mavinkurve, M.E. Slopen, S. Sedlar, J. Ho and K.H. McVeigh, New YorkCity Department of Health and Mental Hygiene (U.S.A.), A.E. Curry, The Children's Hospitalof Philadelphia (U.S.A.)
07C-3 The Interplay between administrative data and survey data in the Rotterdam Early School Leavers Monitor ..... 239
J. Hartkamp, DESAN Research Solutions (The Netherlands)
SESSION 08A: USE OF AUXILIARY SOURCES IN WEIGHTING
08A-1 Combining information from two independent surveys through the Pseudo Empirical Likelihood Method. ..... 245
C.Wu, University of Waterloo (Canada) and J.K. Kim, Iowa State University (U.S.A.)
08A-2 Estimation of correlations between cross-sectional estimates from repeated surveys - An Application to the variance of change ..... 246
Y.G. Berger and R. Priam, University of Southampton (U.K.)
08A-3 Use of auxiliary sources in weighting: The case of integration of social survey data ..... 256
T. Merkouris, Athens University of Economics and Business (Greece)
SESSION 08B: THE CANADIAN CENSUS
08B-1 Integrating Census and Household Survey Listing Operations ..... 265
T. Haymes and J.L. Bernier, Statistics Canada
08B-2 Overview of the Canadian Postcensal Survey Program ..... 270
E. Langlet, Statistics Canada
08B-3 An Integrated approach to collection of Census data in Canada ..... 275
P. Tallon and L. Ramsay, Statistics Canada
08B-4 The 2009 Census test supplementary study of the new Census collection methodology ..... 282
J. Taylor, Statistics Canada
SESSION 08C: LINKAGE METHODS
08C-1 An Empirical comparison of approaches to approximate string matching in private record linkage ..... 290
T. Bachteler, R. Schnell and J. Reiher, University of Duisburg-Essen (Germany)
08C-2 GLink - Constructing an avatar ..... 296
A. Chevrette, Statistics Canada
08C-3 Privacy-preserving record linkage using Bloom filters ..... 304
R. Schnell, T. Bachteler and J. Reiher, University of Duisburg-Essen (Germany)
SESSION 09A: SMALL AREA ESTIMATION
09A-1 Estimation of average design-based mean squared error of synthetic small area estimators ..... 311
P. Lahiri, University of Maryland (U.S.A.) and S. Pramanik, University of Chicago (U.S.A.)
09A-2 Estimation of poverty measures for small areas ..... 321
J.N.K. Rao, Carleton University (Canada)
09A-3 Alternative forms of small area models for using survey, census and administrative records data ..... 322
W.R. Bell, and J.T. Huang, U.S. Census Bureau (U.S.A.)

## SESSION 09B: EMPLOYMENT DATA

09B-1 Combining administrative and survey data in a German research data centre: Linkage, quality and future developments ..... 324
S. Bender, J. Heining, T. Hethey and P. Schioch, Institute for Employment Research (Germany)
09B-2 Employee flows to study firm and employment dynamics ..... 325
K. Guerts, HIVA - Katholieke Universiteit Leuven (Belgium), M. Ramioul and P. Vets, NationalSocial Security Office (Belgium)
09B-3 Dealing with inconsistencies in linking administrative data. ..... 331
D. Hochfellner and A. Voigt, Institute for Employment Research (Germany)
09B-4 You can match my data! Biasing effect in the use of linked administrative and survey Data. ..... 337
M. Huber, A. Schmucker and S. Bender, Institute for Employment Research (Germany)
SESSION 09C: NON-RESPONSE
09C-1 Toward a more robust nonresponse propensity model: The interplay among NAEP restricted use data, contact history information and school administrative data ..... 339
A. Young Chun and F. Scheuren, University of Chicago (U.S.A.)
09C-2 Multivariate fractional imputation in survey sampling ..... 340
J.K. Kim and W. Fuller, Iowa State University (U.S.A.)
09C-3 Low response rates and sample representativeness in an online survey of college students ..... 341
L. Lawson, E. O’Connor, C. Rouse, D. Street and S. Kulneva, Kennesaw State University (U.S.A.)
SESSION 10A: REALTIME ACCESS TO DATA FROM MULTIPLE SOURCES
10A-1 A service platform providing remote access to data for research purposes ..... 348
A. Bélanger and C. Gilbert, Institut de la statistique du Québec (Canada)
10A-2 The Real Time Remote Access at Statistics Canada ..... 356
D. Price and M. Lessard Statistics Canada
10A-3 Methods for permitting differential access to sensitive data while maintaining confidentiality ..... 357
P. Meyer, National Center for Health Statistics (U.S.A.)

## SESSION 10B: SMALL AREA ESTIMATION

10B-1 Feasibility study report for measuring the shadow population in northern Alberta ..... 359
B. Brooks, Statistics Canada, M. Haan, University of New Brunswick (Canada), and J. Lee,Alberta Office of Statistics and Information (Canada)
10B-2 More for less: Using statistical modelling tocombine survey data and administrativesources.363
A. Noble, S. Haslett, G. Jones, Massey University (New Zealand) and D. Ballas, University of Sheffield (New Zealand)10B-3 Estimation of monthly unemployment figures in a rotating panel: The use of auxiliaryseries in a structural time series model370
J. van den Brakel and S. Krieg, Statistics Netherlands

10B-4 Comparison of intercensal updating techniques for local level poverty statistics
M. Isidro, S. Haslett and G. Jones, Massey University (New Zealand)

## SESSION 10C: RECONCILIATION

10C-1 Use of mental health services in Québec based on data from a population survey data
and administrative data: Study of underreporting and overreporting of services ..................... 385
A. Drapeau, R. Boyer, F. B. Diallo, University of Montreal (Canada)
10C-2 Understanding victimization statistics from two national crime data sources: The
Uniform Crime Reporting Survey and the General Social Survey on Victimization
A. Halladay, M. Sinha, K. AuCoin and S. Franklin, Statistics Canada

10C-3 School education data in India: Emerging issues in increasing multiplication of sources of data.
A. N. Reddy, National University of Educational Planning and Administration (India)

10C-4 Validating diabetes disease status in claims and survey data400
J.W. Sakshaug, D.Weir and L.H. Nicholas, University of Michigan (U.S.A.)

## SESSION 11A: COMPLEMENTARY SURVEYS

11A-1 Use of the Census as a supplementary frame for the Canadian Survey of Household Spending
D. Malo, Statistics Canada

11A-2 Surveys and administrative files in France, the impact of hashing............................. 415
B. Riandey, Institut National d'Études Démographiques (France)

11A-3 American Community Survey design and statistical methodology............................. 422
A. Navarro, U.S. Census Bureau (U.S.A.)

## SESSION 11C: VALIDATION

11C-2 Linking the National Health and Nutrition Examination Survey to Supplemental Nutrition Assistance Program administrative data: a one state pilot study424
L.B. Mirel, V. Burt, D. Miller, M. Wiseman, J.Kirlin and C. Cox, Centre for Disease Control (U.S.A.)

11C-3 Quality and quantity: Using administrative data for scientific purposes in labour market research.
P. Schioch, Institute for Employment Research (Germany)

## SESSION 12A: CLOSING SPEECH

12A-1 A Strategic vision for the use of administrative data at Statistics Canada ..... 432M. Cloutier, Statistics Canada

## Preface

Symposium 2010 was the twenty-sixth in Statistics Canada's series of international symposia on methodological issues. Each year the symposium focuses on a particular theme. In 2010, the theme was: "Social Statistics: The Interplay among Censuses, Surveys and Administrative Data".

The 2010 symposium was held from October 26-29, 2010 at the Crowne Plaza Hotel in Ottawa, Ontario, and it attracted more than 500 people from 25 countries. Three workshops and eighty-four papers were presented. Aside from translation and formatting, the papers, as submitted by the authors, have been reproduced in these proceedings.

The organizers of Symposium 2010 would like to acknowledge the contribution of the many people, too numerous to mention individually, who helped make it a success. The organizers would also like to thank the presenters and authors for their presentations and for putting them in written form. Finally, the organizers would like to recognize the work of Annette Everett and Denis Lemire and thank them for their contributions.

## The Symposium 2010 Organizing Committee

Normand Laniel (Chair)<br>Logistic Committee: Sylvie Gauthier (Chair)<br>Marie-Claude Duval<br>Patrick Gallifa<br>Paul Matthews<br>Paula Versolato<br>Scientific Committee: Pierre Lavallée (Chair)<br>Edward Chen<br>André Cyr<br>Sarah Franklin

OPENING REMARKS

# OPENING REMARKS 

François Maranda ${ }^{1}$

Good morning,
On behalf of Statistics Canada, I would like to welcome you all, friends and colleagues, to Symposium 2010. This is the $26^{\text {th }}$ International Symposium organized by Statistics Canada on survey methodology.

Some of you are familiar with our Symposium, and we are glad to have you back. Others are joining us for the first time, and I would like to extend our most cordial welcome to you.

One of the goals of the annual Symposium is to bring together people from a variety of backgrounds to share their views, experience and expertise on current and emerging methodological issues. This year, the theme of the Symposium is "Social Statistics: The Interplay among Censuses, Surveys and Administrative Data".

This is not the first time that we have addressed this topic. In 1999, the theme of the Symposium was Combining Data from Different Sources; in 1997 it was New Directions in Surveys and Censuses, and in 1987 we discussed the Statistical Uses of Administrative Data. So, why have a Symposium now on the Interplay among Censuses, Surveys and Administrative Data in the field of Social Statistics?

There are several reasons - the increasing availability of administrative files in a number of areas such as justice and health; the growing capacity to merge information from different sources which provides opportunities for methodological developments; and the need to improve efficiencies and data quality, and to reduce respondent burden.

On the business statistics' side, the change in direction to a more integrated approach to surveys and administrative data started over ten years ago, and lead to the creation of the Unified Enterprise Survey and a to vast increase in the use of taxation data from the Canada Revenue Agency.

On the social statistics side, it is only more recently that a more holistic approach to surveys, censuses and administrative data has been adopted. Even though the experience is more recent, we nonetheless have many examples.

Several health surveys use multiple sources of data. For example, the Longitudinal Health and Administrative Database combine health information from sample surveys and administrative sources. And the Canadian Cancer Registry, which is a central database located at Statistics Canada, is constructed from administrative data on cancer from provincial and territorial registers.

We also have the Address Register which is constructed and maintained by combining information from the Census, the Labour Force Survey and from a variety of administrative sources such as telephone billing files, tax files, new building files, and files from Canada Post. The Address Register started as a tool to improve coverage in the Census but now is used as a direct source for some of the frame for the Census and the Labour Force Survey. It is also used as a listing tool for these two surveys.

[^0]Another example of combining multiple sources of data is our population estimates' program which uses census data, estimates derived from the census coverage studies and information from various administrative data sources, including tax records.

Administrative data are often used in combination with survey or census data. For example, this year the Survey of Household Spending asked respondents for permission to use their income tax data in lieu of filling out the survey's income module, as did the 2006 Census. The Survey of Labour and Income Dynamics has been doing this successfully since the 1990s. The Survey of Household Spending and the Survey of Labour and Income Dynamics also use tax data to calibrate their survey estimates.

Census information is used in a variety of ways by social surveys at Statistics Canada. It is used to build the frame for the Labour Force Survey, to benchmark survey estimates through calibration or post-stratification, and to draw samples for post-censal surveys.

Like the Census, the Labour Force Survey is used to design other sample surveys. For example, some surveys are supplements to the Labour Force Survey; others use respondents who have rotated out of the Labour Force Survey; and still others, such as the Canadian Community Health Survey, use the Labour Force Survey's area frame.

We have also recently started to conduct methodology research on combining multiple data sources for the development a new household survey frame. As it becomes more difficult to reach respondents for household surveys, the need for alternate sampling frames grows. A challenge at Statistics Canada in the next few years will be to integrate current frame activities into a household survey frame service that can be used by all household surveys. This frame will not only improve our ability to select samples, but it will also provide better contact information, give surveys more flexibility in targeting different types of individuals and provide useful information for data collection planning and for nonresponse adjustments.

There are, however, many challenges to building datasets from multiple sources. For example, there may be issues with the quality and stability of a data source over time. In the case of administrative data, it may require considerable time and effort to reach an agreement between Statistics Canada and the data supplier. And there are the challenges related to record linkage and privacy.

Over the next three days, there will be presentations on many of these aspects, including sampling frames, data collection issues, the use of auxiliary data in weighting, microsimulations, and record linkage. I would like to thank all of the presenters for sharing the fruits of their labour.

For our keynote speech today we are extremely privileged to have Jelke Bethlehem discuss the increasingly important role that population registers play at Statistics Netherlands. And on Friday morning, for the closing speech, Michel Cloutier of Statistics Canada will discuss A strategic vision for the use of Administrative data at Statistics Canada.

It is a tradition at the Symposium to invite the recipient of the Waksberg award to give a presentation on the topic of his or her choice. Tomorrow morning, this year's recipient, Dr. Ivan Fellegi, will give a presentation on The organization of statistical methodology and methodological research in national statistical offices. Developing and maintaining strong research programs are key to ensuring the relevancy of any statistical office.

Yesterday, almost 185 of you attended one of the three full-day workshops. One was on record linkage methods; another was about the evolution from traditional demographic calculations to projections by microsimulations; and a third was on using administrative and operating systems to strengthen statistical survey and census systems. I would particularly like to thank the leaders of yesterday's workshops for their important contributions: Young Chun and Fritz Scheuren from the National Opinion Research Centre; and Julien Bérard-Chagnon, Éric Caron-Malenfant, André Cyr, Patrice Dion, Karla Fox, Dominic Grenier and Lori Stratychuk from Statistics Canada.

From the point of view of Statistics Canada, the Symposium is an invaluable source of information. The relevance and value of exchanges that occur at the Symposium enable us to share views, experiences and diverse expertise that benefit all participants. By giving us the opportunity to learn how other organizations tackle challenges similar to ours, the Symposium helps to resolve our issues, critically examine our work and develop our staff.

So, before I declare this $26^{\text {th }}$ Symposium officially open, I would like to thank the participants: in addition to the 372 from Statistics Canada, there are 157 more from 25 countries around the world. Together, you represent 82 organizations from the academic communities as well as the public and private sectors. Thank you all for coming. I would also like to thank the organizing committee, session organizers, chairpersons and all the volunteers who have made this event possible. Welcome everybody.

Now, to get underway with the program, I would like to invite Pierre Lavallée, Chairman of the Scientific Committee of this $26^{\text {th }}$ Symposium, to introduce the keynote speaker.

## KEYNOTE ADDRESS

# Statistics without surveys? About the past, present and future of data collection in The Netherlands? 

Jelke Bethlehem ${ }^{1}$


#### Abstract

Registers play an increasingly important role in the production of statistics by Statistics Netherlands. Already from the early beginnings, the population register forms the backbone of a system of social statistics. It is used as a data source for population statistics, as a sampling frame and as a source of auxiliary information for weighting adjustment. After the last traditional census in 1971 Statistics Netherlands has developed the Social Statistical Database (SSD), in which the population register, other registers and surveys are combined. The SSD is the central data source for the virtual censes and many of the social statistics published by Statistics Netherlands. The SSD is also used for other applications like the treatment of nonresponse. Political pressure to reduce the response burden, the high costs of surveys and nonresponse problems have forced Statistics Netherlands to rethink its data collection strategy. The focus changed from surveys to registers. This approach has advantages, but there are also potential risks. The future of data collection is not clear.


Key Words: Census, survey, population register, social statistical database, mixed-mode survey.

## 1. The emergence of the population register

The first official population census in The Netherlands took place in 1795. From 1829 to 1971, censuses were carried out at intervals of approximately ten years. The government made several attempts to use the census data to set up a population register. This succeeded in 1859 . This register consisted of books that were maintained by municipalities. There was a separate page for each household. In case of changes like birth, marriage or deaths, a new entry was written on the page or an old entry was removed by scratching it out. The head of the household was made responsible for reporting births and deaths in the household. If a child married and started a new household, a new household page was created, and the child was scratched out in the original household. If a household moved to a new municipality, all entries in the old municipality were scratched out. One of the consequences of this new system was that moving from one town to the other was only possible with a certificate issued by the authorities of the current municipality of residence.

In 1909 it was realized that a register based on households was not convenient for recording major events in the live of persons, like a change of address or of marital status. It was recommended to introduce a system of personal cards. The discussion about this proposal lasted until 1938. Then the personal cards were introduced. This population register was used by Statistics Netherlands in the early days of survey sampling as a sampling frame. Usually twostage samples were selected, where the first stage was a sample of municipalities (drawn proportional to size) and the second stage was a systematic sample drawn from the registers in the selected municipalities.

The electronic version of the population register emerged in 1994. This was the start of the computerized population administration as it now still exits. It is called GBA (Gemeentelijke Basisadministratie, or Municipal Base Administration). There is an electronic register for each municipality. Each municipality is obliged (by law) to keep its register up-to-date, using only software that is approved by the national government. All municipalities are linked by a closed network. This makes it possible to exchange information between municipalities. Statistics Netherlands has access to GBA. It uses the GBA for three different purposes: as a data source for population statistics, as a sampling frame, and as a source of auxiliary variables for weighting adjustments.

[^1]The GBA is the obvious data source for population statistics. The register data are used to compile monthly and yearly population statistics. These statistics are obtained simply by counting. There are no surveys involved. The data need some edited before they can be used. The counting process is also complicated by administrative delays in the register data.

The second use of the GBA is as a sampling frame. The GBA is in fact a list of all persons in The Netherlands and therefore corresponds (according to the de jure concept) exactly to the population. So it can be used to draw a random sample of persons. Statistics Netherlands usually selects two-stage samples for social surveys. This sampling design made it possible to reduce travel costs of interviewers in face-to-face surveys.

Statistics Netherlands not only produces statistics on persons, but also statistics on households. The GBA can be used as a sampling frame of households, but this is not straightforward. The problem is that this register contains persons, and also the address at which they live, but there is no information as to which persons at an address form a household. There can be several households living at the same address. In order to use the GBA as a sampling frame for household surveys, households are constructed by Statistics Netherlands using the available data. In a few cases this requires some imputation.

The third use of the GBA is as a source of auxiliary variables for weighting adjustment to correct for nonresponse. Surveys in The Netherlands suffer from substantial amounts of nonresponse. Effective correction for nonresponse requires auxiliary variables. These variables must have been measured in the survey, and moreover, their population (or complete sample) distribution must be available. The population register contains auxiliary variables like gender, age, marital status, municipality of residence and ethnic background. These variables can be sued for weighting. Weighting adjustment will be more effective as relationships between survey variables and auxiliary variables are stronger. This may not always be the case for the auxiliary variables in the population register. So it is not unlikely that more auxiliary information is required for effective nonresponse correction.

## 2. The Social Statistical Database

The last traditional census in The Netherlands took place in 1971. Results of a pilot for the 1981 census predicted a nonresponse rate of $26 \%$. Therefore the reliability of the results was at stake. The parliament decided that participation had to be voluntary and not mandatory, and that the collected data had to be absolutely anonymous. The government came to the conclusion that the social-psychological climate in the country would make successful completion of a new census very difficult, and therefore postponed the census. The parliament decided in 1981 to cancel the traditional census completely. The Census Act of the Netherlands was officially repealed in 1991.

Although the 1981 was cancelled by Dutch parliament, The Netherlands still had to meet its European Union obligation to conduct a census. The European Commission gave permission to replace the census by a set of tables based on combining data from registers and surveys. This was repeated for the 1991 census. Demographic data for these censuses were obtained from the population register. Socio-economic data came from the Labor Force Survey. This included data on labor and education. These data sources, however, were used separately, without paying special attention to consistency. No attempt was made to link data from different sources at the micro-level. Consequently, tables based on data from different sources could be inconsistent. To overcome these inconsistencies, a weighting adjustment technique was applied in which table totals were made to fit population register totals.

The table sets compiled for the 1981 and the 1991 census were less detailed than those of subsequent censuses. For example, there were no tables at the level of the municipalities. More details for 1981 census are described in Vliegen \& Van de Stadt (1988) and for the 1991 census in Corbey (1994).

For the census of 2001, the European Union and other international organizations required more detailed information than for the census rounds of 1981 and 1991. To be able to cope with this request, Statistics Netherlands initiated a new approach. Core component of this approach was the creation of the Social Statistical Database (SSD). This database contains detailed and comprehensive information on persons and households. This information is collected from various sources and linked at the micro-level. The backbone of the SSD is formed by basic demographic variables that are retrieved from the population register. Socio-economic data are added from other administrative sources, such as employment, benefits and income. Not all data included in the SSD are retrieved from registers.

Some variables are only measured in sample surveys. Examples are occupation and level of education, which are recorded in the Labor Force Survey. The variable job size was obtained from the (large) Survey on Employment and Earnings. Figure 3-1 gives a schematic and simplified overview of the Social Statistical Database as it was used for the virtual census of 2001.

Records from various administrative sources could be linked to records in the SSD, because all these records contain the unique Social Security and Fiscal Number (the so-called SoFi number). The SoFi number is not recorded in the sample surveys. The SoFi number was replaced in 2007 by a Citizen Service Number. Linkage for survey data records was established using a key based on gender, data of birth and address. More on these record linkage issues can be found in Linder (2004) and Arts, Bakker \& Van Lith (2000).

Figure 3-1
Schematic view of the Social Statistical Database


The way tables are compiled depends on the source of the variables. There is no problem for tables based on variables coming from the registers and other administrative sources. Compiling a table is just a matter of counting. The situation is different for tables containing survey variables. Then, population counts have to be estimated. Moreover, if a variable is measured in more than one survey, there will be different estimates for its population distribution. Although difference may be within margins of sampling error, Statistics Netherlands prefers, and also Eurostat demands, to have consistent figures. To achieve this consistency, a new methodology was developed. This was called the method of repeated weighting.

The repeated weighting technique applied by Statistics Netherlands comes down to repeated application of the generalized regression estimator. To produce tables that are as accurate as possible, they should be computed using the largest number of possible records. This means that tables just based on register variables use all records in the register. Tables that contain at least one survey variable are estimated using the largest possible set of records that contain all variables concerned. To obtain a consistent set of tables, a sequential procedure is followed. For each new table, the margins are determined it has in common with tables that already have been estimated. The next step is to estimate the table while calibrating on common margins. The estimates, apart from being consistent, will also be more accurate, particularly if the margins can be estimated from larger data sets or counted from register data, and as such serve as auxiliary information. In case of traditional weighting there will usually be one set of weights for the survey. Repeated weighting, however results in a set of weights for each table. This is necessary in order to obtain consistency between tables. More detail about repeated weighting can be found in Houbiers et al. (2003) and Houbiers (2004).

The work on the census of 2001 started at a later moment in time than would have been the case for a traditional census. One of the reasons was that first all data sources (including registers and administrative sources) had to be
available. Notwithstanding the late start, The Netherlands was one of the first to send its tables to Eurostat, the statistical office of the European Union.

The costs of virtual census of 2001 were much lower than that of a traditional census. At the time, it was estimated that a traditional census would cost approximately 300 million euro. The costs of the virtual census were 3 million euro. More information about the Dutch Virtual Census of 2001 can be found in Gouweleeuw and Hartgers (2004), Van Maarseveen (2004), Schulte Nordholt (2004) and Schulte Nordholt (2007).

For the 2011 census the new European Census legislation of 2008 makes the production of European Census tables mandatory. This regulation leaves the countries free to decide on the data sources to use. The census topics for the 2011 census have been defined in the regulation, and decisions have been made about the tables to produce and the level of detail of these tables. Due to the required level of detail, it will be a challenge for register-based countries to produce all European census tables in a consistent way.

The Social Statistical Database (SSD) is not only an important instrument for the decennial census. It becomes a more and more important data source for Statistics Netherlands. By 2013 all statistics on households will be published from this database. The information in the SSD can also be used for other purposes than just publishing statistics. Two examples are given. A first example is nonresponse analysis and correction. Surveys suffer from nonresponse. Estimates of population characteristics will be biased if, due to non-response, some groups in the population are over- or under-represented, and these groups behave differently with respect to the survey variables. To reduce such a bias, usually some kind of weighting adjustment is carried out. Weighting adjustment requires auxiliary variables. The SSD is a rich source of such auxiliary information. Survey respondents and nonrespondents can be linked to the SSD records, and therefore their values for all kinds of auxiliary variables can be compared. Bethlehem (2009), and Bethlehem et al. (2011) contain examples of a response analysis using SSD-data.

A second example is the R-indicator for survey quality. Survey agencies often use the survey response rate as an indicator of survey quality. However, a low response rate does not necessarily imply that the accuracy of survey estimates is poor. If there is no correlation between response behavior and the survey variables, estimates will be unbiased, notwithstanding a low response rate. There is a need for additional survey quality indicators that provide more insight in the possible risk of biased estimators. Schouten et al. (2009) propose the $R$-indicator. The R stands for 'representativity'. The R-indicators measures how representative the survey response is. It is assumed that every element in the population has a certain probability $\rho$ to respond when selected in the sample. The R-indicator is defined as $R(\rho)=1-2 S(\rho)$, where $S(\rho)$ is the standard deviation of the response probabilities. The R-indicator assumes the value 1 if all response probabilities are equal. Then, the response is representative. The more the response probabilities vary, the less representative the response will be, and the closer the value of $R(\rho)$ will be to 0 .

The values of the individual response probabilities are unknown in practice. This problem is solved by estimating the response probabilities. This can be accomplished if proper auxiliary information is available, i.e. variables that have been measured for both respondents and non-respondents. Usually, a logit model is used for this. The Social Statistical Database of Statistics Netherlands contains the data to compute R-indicators.

## 3. From surveys to administrative data

A number of trends in society have forced Statistics Netherlands to rethink its data collection strategy. In the first place, there were increasing complaints about the response burden imposed by Statistics Netherlands on companies. Many companies received questionnaire forms for several different surveys each year. They complained that they sometimes had to provide the same information again and again, not only to Statistics Netherlands, but also to other government agencies (like the tax authorities). Statistics Netherlands measures the response burden of companies since 1994. It succeeded in reducing this response burden (in number of hours) by $70 \%$ in the period from 1994 to 2008. The total administrative burden of government was measured in 2007. It turned out that the share of Statistics Netherlands in this burden was only $0.25 \%$. The cost of the response burden for companies was estimated to be 23.1 million euro in 2007. The response burden of Statistics Netherlands in 2008 was 21.3 million euro, implying that the response burden was reduced by $8 \%$ from 2007 to 2008. Although Statistics Netherlands realized a substantial reduction of the response burden over the years, and it could show the response burden was only a very small part of the total (perceived) response burden, it remains a serious political issue. It should be noted that currently about $50 \%$
of the response burden is not caused by Dutch government but by European Union regulations. Furthermore, 30\% of the response burden is caused by international trade statistics, production statistics and turnover statistics.

A second trend is that Statistics Netherlands experiences an increasing demand for more and for more detailed information about all aspects of society. Particularly, there are regular requests for more detailed regional information. The sample sizes of most current surveys of Statistics Netherlands are insufficient to provide this level of detail. In some situations it may be possible to apply small area estimation techniques to produce model-based estimates for small areas. This raises the question whether official statistics can be based on models that cannot always be checked for validity. Are such statistics still reliable and accurate?
A third trend is that Statistics Netherlands has faced a number of budget cuts. This forced the institute to consider various of cost reductions of its operations. Conducting a survey is an expensive activity. Reducing the sample size may help to reduce costs, but smaller samples also reduce the accuracy of the results, and it may therefore make publication of statistics about small areas impossible.

A fourth trend is that surveys in The Netherlands are always affected by relatively high nonresponse rates. Nonresponse causes estimates to be biased. Reducing nonresponse in the field is costly and usually very difficult, if not impossible. Weighting adjustment can be applied to reduce the effect of nonresponse as much as possible, but this is also not always completely successful.

The way out of these data collection problems was to avoid as much as possible conducting surveys (primary data collection) and to focus more on retrieving data from other existing sources (secondary data collection). A new law was passed in November 2003 giving Statistics Netherlands access to these sources. With respect to data collection, the new law states that

1) Statistics Netherlands is authorized to use, for statistical purposes, data from registers that are maintained by government and related organizations. This use of data is free of charge;
2) If these registers fail to provide the required data, or the data is of insufficient quality, Statistics Netherlands can conduct surveys.

From the point of view costs, use of registers has clear advantages. Since data is already collected by some other organization, there are no data collection costs for Statistics Netherlands. Bakker (2009) warns, however, that register data often still need extensive data editing in order to get data of acceptable quality. Data editing may require extensive resources.

From the point of view of the amount and type of data that become available, use of register data has a number of advantages, particularly if data from various sources can be linked at the micro-level. Some of them are described by Bakker (2009):

- It becomes possible to investigate small sub-populations.
- It becomes possible to investigate relationships between variables that are difficult to measure in a survey. Costs may be too high or questions may be too difficult to ask (sensitive questions). An example is investigation of criminal behavior of young people.
- It opens new possibilities for longitudinal research;
- It is possible to carry out studies that combine data from different generations. An example is investigating the relation between criminal behavior of parents and their children.

To some extent the collection of register data is similar to the collection of survey data, particularly if questionnaires forms are used. An example is the yearly collection of income information by the tax authorities. This implies that register data are at least partly affected by the same problems as survey data are. Hence, register data are not perfect. Administrative registers are usually owned and maintained by other organizations than a statistical institute. When using register data for statistics, it should therefore be realized that these data have been collected by different agencies, using different definitions, for different purposes at different times. This gives raise to questions about the usefulness and quality of register data.

Register data may suffer from non-sampling errors. These errors are caused by phenomena like lack of population coverage and measurement errors. For example, population registers and business registers usually suffer from both over- and under-coverage due to the dynamics of these populations. Registers containing information about jobs and
allowances have measurement errors due to time delays in their registration of employment and unemployment. The register of real estate values maintained by the land register office does not contain information on new buildings.

Data from many registers and other administrative sources need extensive editing and imputation before they can be used for statistics. The degree to which these registers contain missing data and measurement errors differs for each source and is related to the objectives of the register holders and the extent to which register holders monitor the quality of their data. If some variables are of less importance to the register holder than others and the costs of monitoring the quality are high, the quality of the data can be very poor.

Probably one of the most important differences between registers and surveys is the complete absence of unit nonresponse. Unit non-response plays an important role in surveys, especially when it is caused by the topic of the survey. This seems to be far less the case for registers, because participation in the data collection is often mandatory or beneficiary for the data providers.

Given the fact that register data are not of perfect quality, the next question is to what extent a statistical institute should monitor and improve the quality of register data. See also Wallgren \& Wallgren (2007). And if some quality control system is needed, then the question is whether to use the quality control system of the register holder (if it exists at all), or to develop a new system that satisfies the requirements of the statistical institute. If the register holder has some quality control system, the statistical institute may seek co-operation with the register holder in an attempt to adapt this system in such a way that it is useful for making statistics. The question is whether the register holder is prepared to incorporate extensive data editing procedures in its quality system. It is a current practice at Statistics Netherlands to monitor the quality of the register data it receives. The administrative register data are processed and transformed, and so a statistical register is formed.

Close contact and intense communication with the register holders is vital. Only if the quality control system implemented by the register holder is known, a statistical institute can perform the necessary steps to evaluate quality. In the ideal situation the quality control of the register holder and the statistical institute are fully integrated. The situation in The Netherlands is that Statistics Netherlands is often a small player with relatively little influence on the process of data collection and processing performed by the register holder.

## 4. Beyond registers?

Using data from registers and other administrative sources for statistics involves some risk. This is at least partly caused by the fact that Statistics Netherlands is not the owner of the register, and therefore is not in control of the data being collected. It may be confronted with variables being excluded, with variable definitions being changed, with reduced data quality and even with a register being discontinued.

An example shows what can happen. Statistics Netherlands publishes monthly statistics on the construction sector. Turnover statistics for small companies (less than 10 employees) are not based on surveys but on register data that are collected by the tax administration. Monthly VAT declaration forms are used to construct these statistics.

The economic crisis led to a sudden change in the system of VAT declaration. The government decided in April 2009 that not all companies have to submit their VAT declarations every month any more. Those companies could change to a quarterly system, thereby postponing payments, and as a result improve their financial position somewhat. The consequences for the monthly statistics seemed serious. Many companies changed from monthly declaration to quarterly declaration. Statistics Netherlands was confronted with the question whether it was possible to continue producing reliable monthly statistics. There were still companies sending in monthly VAT declaration forms, but analysis showed they were not representative for all companies. So they could not be seen as a random sample.

The obvious solution of this problem was to re-introduce primary data collection. A new survey was set up to obtain the required information. However, this was not in line with the principle that the response burden of companies should be reduced as much as possible. There were questions in parliament and stories in the newspapers. Data collection was stopped after two months. An extensive study was carried out to compare the survey data with the data of the left-over monthly VAT declaration companies. The results of the analysis were not clear-cut. In the end it
was decided to continue producing preliminary monthly figures based on the VAT-data of the left-over companies. After three months all VAT-data become available. These data are used to revise the monthly statistics. This is a solution for the time being, but as more companies may decide to use the quarterly declaration system, the situation may become worse. More about this case can be found in Van Delden et al. (2010).

It is because of these circumstances that are beyond control of the statistical institute, that it may be wise to think about what to do if at some point in time register data may turn out not be useable any more. Statistics Netherlands has organized some brainstorm meetings to generate ideas about a data collection future that is less dependent on registers. Some of the, sometimes wild, ideas were:

- Statistics Netherlands could become the central coordinating body of all government registers in the Netherlands. That would give Statistics Netherlands more control about the contents of the registers. Moreover, this could help to achieve some form of standardization among registers (data and metadata).
- New innovative ways could be found to collect data straightaway from the Internet. For example, data on persons can be retrieved from social networks like Facebook, LinkedIn and Hyves.
- Collecting prices of products from company websites. For example, there is a website in The Netherlands containing all houses that are for sale (www.funda.nl). The house prices mentioned here could be used to construct indicators for the housing. And websites with fuel prices and airline ticket process could be used for the CPI. More on experiments collecting data straightaway form the Internet can be found in Hoekstra et al. (2010).
- Directly retrieving data (instead of asking) from the administrative and financial systems of companies. The Dutch Government promotes use of the XBRL (eXtensible Business Reporting Language) standard for the exchanges of financial data between government organizations and companies. This standard offers new possibilities here. See also Roos (2009).
- Use of RFID tags. RFID stands for Radio Frequency Identification. It is a small microchip that can be attached to an object. A RFID tag can pick up signals from a RFID reader or scanner, and returns a signal with additional information, usually a unique serial number or other customized information. If RFID tags are attached to products or trucks, they can be used to collect data for transport statistics.
- Use of data from GPS systems. GPS stands for Global Positioning System. Many trucks in The Netherlands have such a system. This makes it possible to determine the geographical position of the truck at any moment. Statistics Netherlands could use this type of information for transport statistics.
- Use of data from GSM systems. GSM stands for Global System for Mobile Communication. Mobile phones use GSM system. Many people in The Netherlands have mobile phones. There was a mobile phone in $90 \%$ of the households in 2008. The GSM system not only may provide information about calling behavior, but also about the location and movement of the owners of these phones. Such data may be used in mobility surveys.
- Use of Google search phrases. Google allows for accessing and analyzing search phrases that have been entered by users. Economic indicators can be defined using these data. Time series can be constructed broken down by, for example, country and region.

It will be clear that several ideas are interesting, but not yet ready for implementation. It is not clear whether they will ever contribute to reliable and accurate statistics. Nevertheless, this type of 'out of the box' thinking may help to develop creative, novel ideas.

## Acknowledgements

The author thanks Bart Bakker and Eric Schulte Nordholt for their valuable contributions to this paper.

## References

Arts, K., Bakker, B.F.M. and Van Lith, E. (2000). Linking Administrative Registers and Household Survey in P. Al et al. (eds.) Re-engineering Social Statistics by Micro-integration of Different Sources, Netherlands Official Statistics, 15, pp. 16-22.

Bakker, B.F.M. (2009). Trek alle registers open!, Inaugural address, 29 November 2009, Free University, Amsterdam, The Netherlands.

Bethlehem, J.G. (2009). Applied Survey Methods, A Statistical Approach. Hoboken: John Wiley \& Sons.
Bethlehem, J.G., Cobben, F. and Schouten, B. (2011). Handbook of Nonresponse in Household Surveys, Hoboken: John Wiley \& Sons.

Corbey, P. (1994). Exit the Population Census. Netherlands Official Statistics, 9, pp. 41-44.
Gouweleeuw, J. and Hartgers, M. (2004). The Method of Repeated Weighting in the 2001 Census in E. Schulte Nordholt et al. (eds.) The Dutch Virtual Census of 2001, Voorburg: Statistics Netherlands, pp. 261-276.

Hoekstra, R., Ten Bosch, O. and Harteveld, F. (2010). Automated Data Collection from Web Sources for Official Statistics: First Experiences, Technical Report, The Hague: Statistics Netherlands.

Houbiers, M. (2004). Towards a Social Statistical Database and Unified Estimates at Statistics Netherlands. Journal of Official Statistics, 20, pp. 55-75.

Houbiers, M., Knottnerus, P., Kroese, A.H., Renssen, R.H. and Snijders, V. (2003). Estimating Consistent Table Sets: Position Paper on Repeated Weighting, Discussion Paper 03005, Voorburg/Heerlen: Statistics Netherlands.

Linder, F. (2004). The Use of Administrative Registers and Sample Surveys in the Dutch Census of 2001 in E. Schulte Nordholt et al. (eds.), The Dutch Virtual Census of 2001, Voorburg: Statistics Netherlands, pp. 243260.

Roos, M. (2009). The Dutch Taxonomy Project in a Statistical Context. Proceedings of Statistics Symposium 2008 on Data Collection: Challenges, Achievements and New Directions, Gatineau: Statistics Canada.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. Survey Methodology, 35, pp. 101-113.

Schulte Nordholt, E. (2004). Introduction to the Dutch Virtual Census of 2001 in E. Schulte Nordholt et al. (eds.), The Dutch Virtual Census of 2001, Voorburg: Statistics Netherlands, pp. 9-22.

Schulte Nordholt, E. (2007). Virtual Censuses, International Statistical Seminar 47, Instituto Vasco de Estadistica, Vitoria-Gasteiz, Basque Country, Spain.

Van Delden, A., Scholtus, S. and Wolf, P.P. de (2010). The Case of the Missing Tax Data. Report, Report, The Hague: Statistics Netherlands.

Van Maarseveen, J. (2004). The Dutch Virtual Census of 2001 Compared to Previous Censuses. E. Schulte Nordholt et al. (eds.), The Dutch Virtual Census of 2001, Voorburg: Statistics Netherlands, pp. 119-142.

Vliegen, J.M. and Van de Stadt, H. (1988). Is a Census Still Necessary? Experiences and Alternatives. Monthly Bulletin of Population Statistics, Vol 2000/02, pp. 9-15.

Wallgren, A. and Wallgren, B. (2007). Register-based Statistics. Chichester: John Wiley \& Sons.

## SESSION 2A

SAMPLING FRAMES

# The OCTOPUSSE System of the New Master Sample of the Institut National de la Statistique et des Études Économiques (INSEE) 

Marc Christine and Sébastien Faivre ${ }^{1}$


#### Abstract

Since the 1960s, the two-stage sample design used in INSEE household surveys traditionally used exhaustive population censuses as a survey frame.

In January 2004, France implemented a radically different census system, accomplished each year by surveying a fraction of the territory, with a five-year rotation of the census zones (called "rotation groups").

This system, named OCTOPUSSE (Organisation coordonnée de tirages optimisés pour une utilisation statistique des échantillons), required the redefinition of the methodology used for the composition of household survey samples.

The primary innovation consisted of benefiting from the «freshness» of the new census, that is, using the lists of lodgings surveyed in year $n$ as a sample frame for surveys conducted in year $n+1$.

As a result, the primary units had to be redefined to consider the freshness principle and to allow an enumerator assigned to a specific zone to conduct a survey without incurring excessive costs for travelling to respondents' homes. Therefore, only the fraction of the primary unit (or EAZ, enumerator action zone) belonging to the last rotation group sburveyed is drawn for a given survey.

The EAZ is constituted by aggregating communes, under minimal constraints, while minimizing their geographical range. An innovative automated solution was implemented to this effect.

A sample EAZ was then taken with appropriate stratification and balancing conditions. The new system created has been operational since mid-2009. It should be noted that most surveys are touched by this new system, except the employment survey that uses a selection system of clusters, which uses sampling frames from tax files.


[^2]
# The National Statistical Registry on Persons in Italy 

Roberta Vivio, Carla Runci ${ }^{1}$


#### Abstract

The "Statistical Registry on Persons in Italy" is at the moment one of the most important projects at Istat. The Register is going to be the first statistical register on the entire Italian population, and it will imply important consequences in the official statistics whether using it alone (by raw data) or together with other archives (by matched data). Its implementation will allow cost reduction, response burden reduction, improvement of quality, and meeting current and new users' needs. This paper reports on the project's working progress.


## 1. Preview

The statistical demographic registers in Italy started later than the business ones. This is because both the General Government's administrative demographic registers - and consequently their statistical potential - have been recently increased; furthermore, the Italian jurisprudence has just become favourable to using administrative data on physical persons for statistical purpose. The new norms regulating the access and use of such registers established the preconditions for cooperating among administrative authorities and Istat. In this regard were defined several specific laws (e.g. privacy law, etc.) in the last decade.

In other countries - on the other hand - the new era of "virtual population census" has already started. To achieve such a goal a precondition is needed: building a solid statistical system based on a strong population register. Istat is developing that strategic line with the "Statistical Registry on Persons in Italy". It is currently one of its most important projects. Istat is also working on some of the Register's applications and developments. Among them, one of the most important is "The undercount list for census", which aims to reduce the undercount rate. The other implementations are:

- extraction of stratified samples on individuals by socio-demographic characteristics,
- timely demographic indicators,
- "Statistical System of Persons".


## 2. Italian Statistical Population Register

### 2.1 Population Registers in Italy

The Statistical Registry ranks among the most significant registers on the Italian Population, briefly below.
Municipality Population Register (Anagrafe della Popolazione residente).
Each Italian municipality issues a Population Register. For each resident in the district, it contains identity, residence, demographic events and concerning changes. It happens sometimes that a person recorded in a municipal register is not living there (e.g.: because of relocating internationally, or in another district): this is well known as "register's over counting" (undercounting in the opposite situation). The main problems are over / under counting, and double counting.

[^3]National Population Index (Indice Nazionale delle Anagrafi).
It is carried out by the Department of the Interior. The aim is to interlink eight thousand local registers. It contains only a short set of variables linked to each person. Compared to the previous register, it is bigger with regard to the territorial level, and smaller regarding the number of variables.

The most important result of this project will be setting to zero the risk of double counting. Up to now, it does not contain address and citizenship.

### 2.2 The National statistical registry on persons: DB's archives

The project started in 2006. It has an ambitious aim: building a statistical register on the resident population suitable for clearing double cases, over and under counting. The basic idea is to build the Italian population (about 60 million people) as a mosaic of thematic archives containing proper information to determine who was resident and where. Twelve archives were chosen on physical persons: economic as well as demographic, social and fiscal. Overall, they amount to about 230 million records (Table 2.2-1).

Choosing the right archives/registers for the Register was particularly difficult, primarily because it was needed to press HW and SW needs. This is the reason why it was strategic to select fewer archives. They met some key conditions:

- Covering all age groups;
- Covering the main categories: employees, retired, students, etc.;
- Covering equally male and female people;
- Covering Italians and foreigners.

In table 2.2-1 are listed the input archives for the Prototype_Register, and some basic information such as dimension, linkage rate, statistical unit.

Table 2.2-1"2010_Prototype_Register's" inputs

| $\mathbf{n}$ | Denomination | Units <br> (millions) | Reference time <br> (point, period) | Link <br> rate | Unit |
| :--- | :--- | :--- | :--- | :--- | :--- |
| $\mathbf{1}$ | Tax Register | 72.4 | 2009 | 100.0 | tax-resident person |
| $\mathbf{2}$ | National Population Index | 59.5 | 2009 | 95.14 | resident person (2001- <br> 2009) |
|  |  |  | 2008 | 97.51 | resident person (2008) |
| $\mathbf{4}$ | Municipality Population register. | 6.7 | 2009 | 98.14 | student at public school |
| $\mathbf{5}$ | Students Archive | Birth Reg._children | 3.0 | $2003-2008$ | 96.75 |
| $\mathbf{6}$ | Birth Reg._mothers | 2.7 | $2003-2008$ | 81.57 | child <br> birther register a child in |
|  |  | 4.4 | $2000-2007$ | 12.36 | died resident person |
| $\mathbf{7}$ | Deaths archive | 16.7 | 2007 | 97.07 | pensioner |
| $\mathbf{8}$ | Retired persons register | 36.2 | 2007 | 97.61 | employee |
| $\mathbf{9}$ | Employee Reg. | 3.8 | 2007 | 93.02 | self-employed |
| $\mathbf{1 0}$ | Self-employed Reg. | 15.4 | 2008 | 99.68 | employee |
| $\mathbf{1 1}$ | Welfare-tax return Reg. | 4.6 | 2008 |  | self-employed |
| $\mathbf{1 2}$ | Self-employed Reg. | 2.9 | 2008 |  | self-employed |
| $\mathbf{1 3}$ | One-persons enterprises R |  |  |  |  |

Mainly, the inputs are administrative nationwide archives; they are managed by Central authorities: e.g., Students Archive. Some others are statistical registers: e.g. One person enterprises.

Such a wide data base must face some challenges such as:

- Putting the fiscal and economic information into demographic meaning;
- Reducing the heterogeneity between administrative units, populations, variables and the statistical ones;
- Managing the different reference periods of each administrative register;
- Sustaining the great effort both of hardware-software technologies and processing networks.

The project started with a wide feasibility study in the framework of demographic registers, which aim was capitalizing on their contents.

At the end of in-depth tests was chosen - as pivot of the system - the Registry Database of physical persons of the Tax Registry's informative system (Tax Register) of Ministry of Finance.
The tax register comes to 70 million records. It includes:

- persons legally resident in Italy,
- persons not legally resident, but having a tax code for administrative reasons ${ }^{2}$.

Each record has a unique identification number: 'tax code'. It is a 16 digit alphanumeric code assigned by the Department of Finance. In the production process it is used as the primary linkage key.

It is important to stress that in the backbone were inserted only a subset of variables, extracted from the much wider original archive. The subset is:

- tax code;
- name, surname;
- place and date of birth;
- tax address;
- check variables (date of insert, etc.).

The first pilot survey took place in 2008. The sample size was about 5 million records. In 2009/10, several additional tests were conducted: substantive and methodological (one among them focused on foreigners). In 2011, the first official output will be released.

At the same time, they are running some derived projects: one of them is a specific register on undercounted persons in the census: "The auxiliary lists for 2011 population Census". These lists should include people who are missing in the municipal population registers, but not in other archives referring to the same municipality. The aim is to control the undercount rate.

Going back to the Register, in the Figure 2.2-1 are showed the distributions of the population from the tax register (outer line), and the Italian Resident Population from an Istat source (inner histogram). The distributions are similar although the tax_population amounts to about 72 million people, and the resident to 60 million.

[^4]Figure 2.2-1 Tax population (July 2009) and Istat Resident population (Dec. 2008)

|  | 7 |  |  |
| :---: | :---: | :---: | :---: |

The basic idea is to extract residents from the tax population using auxiliary information from thematic archives.

### 2.3 Variables

The information content is showed in Fig. 2.3-1. A subset is common to all units (physical persons), other subsets are specialized according to matched archives. A short example is showed briefly below.

Common variables:
Tax code
Name
Surname
Sex
Birth and place of birth
Tax address
Specialized variables for students:
Municipality of residence
Citizenship
Address of residence
Specialized variables for mothers:
Citizenship
Etc.

## 3. The process

The archives in the DB_Register are expected to overlap: either in terms of units as shown in Table 3.3 - 1 (a given person can exist as a student as well as an employee) or in terms of variables (Table 3.3-2). At the end of the process, units must collapse to about 60 million from 240 million present at the beginning in the DB per year, and in
the case of multiplicity for a given variable (Tab. $3.3-2$ ), must be chosen only once.
The process is really complex, either because of dimension or the complexity of algorithms used (e.g. linkage algorithm). In a very few words, the steps are: normalisation of variables chosen to be inserted in the system. Then, linkage between backbone and each of the thematic archives. Third, solving the multiplicity cases. Then, replacement of the tax code with the code ISTAT. Fifth, identifying resident people by using all information available, and validating their place of residence. These steps are briefly described below.

### 3.1 Processing steps for experimental version

The processing for the implementation of the experimental version of the Statistical Registry on Persons included the following steps.

Standardization of geographical variables: all variables related to place of birth, place of residence, place of study and workplace have been harmonized according to official classification of 2010 local administrative structures.

Linkage of various sources with the Registry Database of Physical Persons provided by the Department of Economy and Finance : the variables used for the linkage were first the fiscal code and then for the remaining records, the given name, the first name, the date of birth and the address. For each match, a weight was calculated according to the level of similarities of compared variables. This weight was used to determine the acceptance of the linkage as well as to choose the best match in the case of NxM linkages. At the end of the process, the subsequent processing of deceased persons was excluded as well as the individuals who were not linked to any source and for which there is no indication of a recent interaction with the tax system.

Clusters analysis: the matched sets, consequent to the linkage process, linked to the same individual were analysed to identify possible inconsistencies due to sources errors and linkage errors.

Determination of the probability that a person is a resident: each person was linked to a profile based on information registered in various data sources. For some cases, the available data allows categorical exclusion of a subject from the resident population list (e.g. Retired Persons Register allows identification of person living abroad while tax files allow identification of self-employed workers who work in Italy without living in Italy). In some other cases, depending on occurence indicators in different data sources and information provided by them, it is possible to assign a probability. Figure 2.3-1 presents information available for a individual according to it belonging to one or several sets of data.

Determination of municipality of residence: after determining the place of residence in Italy, we must solve situations where different sources contain contradictory information about the municipality of residence and the address. In this case, in addition to considering the differences between files' reference periods, we could use data on the last date of change of address registered in some sources.

The two last steps of Register processing are still in review as well as the possibilities of determination of families.

## a) Multiplicity reduction process

A main part of the register's quality - which is a residence population list - depends on the efficiency in finding the false positives and false negatives for the place of residence according to a specific territorial level (country, municipality, address): "Territorial inconsistency" between information from different archives related to the "place of residence". It involves eight among 12 archives in the system (Table 3.3-1).

Table 3.3-1 Statistical system and variables' multiplicity

| archive order number |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 1 | 2 | 3 | 4 |  | 5 | 6 | 7 | 8 |  | 9 | 10 | 11 | 12 |  |
| Surname | X | X |  | X |  | X |  |  |  |  |  |  |  |  |  |
| Name | X | X |  | X |  | X |  |  |  |  |  |  |  |  |  |
| Sex | X | X |  | X |  | X |  | X |  |  |  |  |  |  |  |
| birth date | X | X |  | X |  | X |  |  |  |  |  |  |  |  |  |
| place of birth | X | X |  | X |  | X |  | X |  |  | X | X |  |  |  |
| citizenship |  |  | X | X |  | X | X |  |  |  |  |  |  |  |  |
| tax code | X | X | X | X |  |  | X | X | X |  | X | X | X | X |  |
| tax code (spouse) |  |  |  |  |  |  |  |  |  |  | X |  |  |  |  |
| tax code (first child) |  |  |  |  |  |  |  |  |  |  | X |  |  |  |  |
| tax address | X |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| place of residence |  | X | X | X |  | X |  | X | X |  | X |  |  | X |  |
| resident abroad |  |  |  |  |  |  |  |  | X |  |  |  |  |  |  |
| residence address |  |  | X |  |  |  |  |  |  |  | X |  |  |  |  |
| place of work |  |  |  |  |  |  |  |  |  |  |  |  | X |  |  |
| paid weeks in a year |  |  |  |  |  |  |  |  |  |  |  |  | X |  |  |
| death date | X | X |  |  |  |  |  | X |  |  |  |  |  |  |  |
| death place |  |  |  |  |  |  |  | X |  |  |  |  |  |  |  |
| household code |  |  | X |  |  |  |  |  |  |  |  |  |  |  |  |
| number of relatives |  |  | X |  |  |  |  |  |  |  |  |  |  |  |  |
| school code |  |  |  | X |  |  |  |  |  |  |  |  |  |  |  |

Going back to the beginning: "Who are resident and where?"
Who? As the chosen archives are all related to activity of daily living performed in Italy (e.g. studying, working, etc.), it is likely that when a unit of the tax register does not match any of the archives, is not resident or has not been resident more. This is why ten million people in the tax register without any link are "Not Resident or no longer resident in Italy" (Table 3.3-2).

Where? To determine where are resident the residual 62 million, they are focused profiles derived from the edit phase.

Technically speaking, for each case a proper variable was created: it was always a string variable that synthesizes in each byte the result of a single comparison.

$$
\begin{aligned}
& \mathrm{X}=\left[\mathrm{x}_{1}, . . \mathrm{x}_{\mathrm{i}}, . \mathrm{x}_{\mathrm{n}}\right]: \mathrm{x}_{\mathrm{i}}=[\text { Place i_REG } \gamma \text { vs Place i_REG } \lambda] \\
& \left.\mathrm{x}_{\mathrm{i}}=[0,1,2, \mathrm{~b}] \text { according to result [different, identical, similar, absent/not due }\right]
\end{aligned}
$$

Each vector is a profile or stratum not equally probable and not equally frequent. As usual, the most frequent clusters are crucial for the result.

An accurate check plan is needed to individuate the wrong cases. For example, two distinct places of residence can be eligible. When, for example:
a. The two are due to asynchrony in point in time reference of the archives;
b. The two are due to failure updating one archive;
c. Etc.

For the moment, we keep on compiling check rules to fit more and more the edit algorithm.

Table 3.3-2 Tax registers persons by number of links with thematic archives

| Frequency | Number <br> of links |
| :--- | :--- |
| $10,260,790$ | 0 |
| $11,590,378$ | 1 |
| $18,504,745$ | 2 |
| $26,014,139$ | 3 |
| $5,313,713$ | 4 |
| 640,976 | 5 |
| 98,532 | 6 |
| 10,661 | 7 |
| 329 | 8 |
| $\mathbf{7 2 , 4 3 4 , 2 6 3}$ |  |

## 4. The possible uses: The auxiliary list on resident persons for the 2011 Population Census. The Interplay between the Register and Census

The 2011 Italian population census is basically traditional, and at the same time guided by a main list: the Municipal Population Register.

For several different reasons, it happens occasionally that a given register does not include a resident person.
An extensive literature documents the effects on population counting based on those lists. To manage the problem, it is being fitted a statistical sound procedure to estimate ex ante the under counted population.
The challenge is really complex: could the statistical register provide such information? Being a register, could it list the persons?

The tests executed up to date have given good results. In particular, on a sample of 6.7 million records (31 municipalities) the first release for undercount list comes to $3.9 \%$ ( 261 thousands), and the second to $2.5 \%$.

It is useful pointing up that the core variables of the register are very similar to the core census variables for 2011. It is a precondition for future developments as strategic Istat's vision suggest.

## References

Belin, T.R. and Rubin, D.B. (1995). A method for calibrating falsematch rates in record linkage. Journal of the American Statistical Association, 90, 694-707.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage, Journal of the American Statistical Association, 64, 1183-1210.

Herzog, T.N., Scheuren, F.J. and Winkler, W.E. (2007). Data Quality and Record Linkage Techniques. New York: Springer, 2007.

Herzog, T., Scheuren, F. and Winkler, W. Data Quality: A Case Study. WIREs Computational Statistics (forthcoming).

Jabine, T.B. and Scheuren, F. (1986). Record Linkages for Statistical Purposes:Methodological Issues, Journal of Official Statistics, 2, 255-277.

Kestenbaum, B., Lochner, K. and Cox, C. (2010). Record Linkage of NCHS Surveys with SSA Administrative Data: A demonstration of New Record-Matching Methodology That Utilizes Partial SSN in the Enumeration Verification System. A paper presented in the Joint Statistical Meetings, Vancouver, Canada.

Nordbotten, S. (2010). The use of administrative data in official statistics - past, present, and future - with special reference to the Nordic countries Official Statistics in Honour of Daniel Thorburn, 205-223.

Scheuren, F. (1993). Turing administrative systems into information systems, Journal of Official Statistics. Vol. 9, No. 1 pp 109-119. Statistic Sweden.

Wallgren, A. and Wallgren, B. (2007). Register-based Statistics: administrative data for statistical purposes. Wiley Series in Survey Methodology

Winkler, W.E., Yancey, W.E. and Porter, E.H. (2008). Fast Record Linkage of Very Large Files in Support of Decennial and Administrative Records Projects, Proceedings of the Section on Survey Research Methods, American Statistical Association

Vivio, R. (2010). Towards Statistical Registry on Persons in Italy from tax and demographic administrative sources In Atti della Conferenza internazionale Statistics Finland Eurostat "European Conference on quality in officials statistics",Helsinky, 3-6 maggio.

Vivio, R. et al. Towards a national statistical register of persons based on administrative sources. Characteristics and potentials for the official statistics. In Atti della Conferenza internazionale Istat Eurostat "European Conference on quality in officials statistics", http://www.istat.it/istat/eventi/q2008/sessions/presentation/04/S0404Vivio.ppt http://www.istat.it/istat/eventi/q2008/sessions/paper/04Vivio.pdf Roma, 8-11 luglio 2008.

Figure 2.3-1-Informations disponibles pour une personne en fonction de son appartenance à un ou plus ensembles.


# Canadian Health Measures Survey, Cycle 2: Combining census and administrative data to improve survey frame efficiency 

France Labrecque, Suzelle Giroux and Andrew Quigley ${ }^{1}$


#### Abstract

The Canadian Health Measures Survey (CHMS) uses a multi-stage sample design. For each collection site sampled, dwellings are selected from the 2006 Census frame using household composition to identify the target age groups. Since this sample design was successful in Cycle 1, the same approach will be used for Cycle 2. The target population for Cycle 2 is people aged 3 to 79. With data collection taking place between the fall of 2009 and the fall of 2011, the 2006 Census frame is deteriorating. It must be updated to cover new dwellings, and a way must be found to identify dwellings with 3-5 year olds, since they cannot be identified from the census frame. This paper begins with an overview of the CHMS. It then explains how the survey frame was updated using the Address Register and the T1 family file to improve coverage and reach the target population. Lastly, it presents findings regarding the efficiency of this approach, based on the sites completed in Cycle 2.


Keywords: Survey frame; stratification efficiency; survey frame update; Canadian Health Measures Survey.

## 1. Introduction

### 1.1 Canadian Health Measures Survey (CHMS)

The CHMS collects important information about the health of Canadians through in-person interviews with households and direct physical measurements in a mobile clinic. At the interview in respondents' homes, information is collected on topics such as nutrition, smoking habits, alcohol use, medical history, current health status, lifestyle, physical activity, the environment, housing characteristics, and certain demographic and socioeconomic variables. Respondents are then asked to go to a clinic, which is set up in a trailer so that it can be moved from one collection site to the next. At the clinic, measurements such as blood pressure, height, weight and physical fitness are taken, and blood and urine samples are collected to test for chronic and infectious diseases, nutrition problems and environmental markers.

In Cycle 2 of the CHMS, the target population consists of persons aged 3 to 79 living in private households in Canada. Persons living on Indian Reserves or Crown lands, residents of institutions, full-time members of the Canadian Forces and residents of certain remote regions are excluded. Cycle 2 collection began in August 2009 and will end in December 2011. To produce national benchmark estimates for various health indicators, the CHMS is being administered to about 6,300 people divided into age groups (3-5, 6-11, 12-19, 20-39, 40-59 and 60-79). ${ }^{2}$ The CHMS covers at least 500 respondents in each age-sex group at the clinic; at least 600 are required for the 3-to-5 age group.

The Cycle 2 sampling strategy is similar to the strategy used in Cycle 1 (Giroux, 2007): a three-stage sample design. First, 18 collection sites were selected from the 257 sites defined for Canada. A collection site is a geographic area with a population of at least 10,000 and is defined so that the distance survey participants have to travel is 50

[^5]kilometres or less in census metropolitan areas (CMAs) and 100 kilometres or less in other areas. Areas that do not meet these criteria were excluded. The 257 sites cover $96.3 \%$ of the Canadian population. Within each site, a sample of about 500 dwellings is selected. A list of the persons who live in the selected dwellings and their dates of birth is obtained, and one or two persons are selected from each household. Since children must be accompanied at the clinic, and to improve the response rate, if a household contains a child aged 3-11, two members of the household will be selected: one in the 3-11 age group and one in the 12-79 age group. If there are no children aged 3 to 11 in the dwelling, only one person is selected from those between 12 and 79 years.

A strategy was developed to stratify the dwellings by age group. First, age at the start of collection is calculated for all persons at a given collection site, using the dates of birth in the survey frame. Then the dwellings are stratified as follows: if a dwelling is home to at least one child aged 3 to 11 , it is assigned to the " $3-11$ " stratum; otherwise, if there is at least one person aged 12 to 19 in the dwelling, its stratum is "12-19"; and so on for strata "60-79", "20-39" and " $40-59$ ", in that order. This stratification procedure is used because, with their distribution in the Canadian population and the CHMS's sampling strategy, people aged 3 to 11 are more difficult to find than people aged 12 to 19 , who in turn are more difficult to find than people aged 60 to 79 , and so on. In addition, there is an "other" stratum for other dwellings, i.e. dwellings that are out of scope because of the residents' age (for example, a dwelling whose only residents are 80 and over), vacant dwellings and missing stratum dwellings (because there is no information about the dates of birth of the members living in the household).

### 1.2 Reasons for updating the CHMS's frame

It is particularly important to update the frame for the CHMS for two reasons. First, because of the high costs of the mobile clinic and the laboratory analyses, it is especially important to ensure that the required number of respondents is obtained in each age-sex group without increasing the sample size unnecessarily. When compensating for stratification errors and attempting to obtain a minimum number of respondents in each age-sex group, it is particularly important to increase the sample size as little as possible due to the high costs associated with the CHMS. Consequently, stratification efficiency is vital for the CHMS. Stratification is efficient if each dwelling's frame-derived stratum-i.e. the stratum used in the sampling of dwellings-is the same as its "actual" stratum. The more efficient the stratification, the easier it will be to target each of the age groups in the target population.

The second reason is that the survey frame deteriorates as the Census recedes in time. The CHMS's dwelling frame is based on the 2006 Census. ${ }^{3}$ Thus, the information in the frame before the stratification update dates back to May 16, 2006, while CHMS collection takes place from August 2009 to December 2011. For Cycle 2, children born after the Census will be three years old in May 2009 or later. An update of the frame is therefore desirable to take into account new dwellings, moves, deaths and especially births since the Census. This is important since, as mentioned earlier, young children are particularly difficult to find.

## 2. Updating the survey frame

### 2.1 Adding new dwellings

The first step in updating the CHMS's frame is to add new dwellings. For example, a new dwelling may be a new building, a new apartment added to an existing dwelling, or an existing dwelling that was missed in the Census. To add new dwellings, the Address Register (AR) was used as soon as the new version was available after the Census, i.e. for site 11 in Cycle 1 of the CHMS. Since then, the AR has been updated quarterly with information from Canada Mortgage and Housing Corporation and telephone billing records.

For the first 10 sites in Cycle 2, about 7\% of the dwellings in the frame were new dwellings. The percentages of new dwellings for the first 10 collection sites are shown in Table 2.1-1. The collection sites are numbered in order of collection, i.e. in chronological order. It would seem logical to assume that, because of the degradation of the frame,
${ }^{3}$ For more information on the construction of the frame for Cycles 1 and 2 of the CHMS, see S. Giroux, D. Malo, E. Neusy and M. Provost (2009).
the percentage of new dwellings would increase as the Census receded in time. However, Table 2.1-1 shows that the percentage of new dwellings varies substantially from site to site. Hence, the percentage of new dwellings depends not only on the degradation of the frame over time, but also on the population growth at each collection site.

Table 2.1-1
Percentage of new dwellings, first 10 collection sites, CHMS, Cycle 2
Percentage of New Dwellings

| Collection Site |  |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $4.9 \%$ | $10.8 \%$ | $6.2 \%$ | $5.7 \%$ | $6.9 \%$ | $9.7 \%$ | $2.5 \%$ | $6.3 \%$ | $8.0 \%$ | $9.6 \%$ |

### 2.2 Updating the stratification

The stratification is updated with the T 1 family file (T1FF). The T1FF is generated from T 1 and T 4 tax returns, the Child Tax Benefit file and the births file. A new version of the T1FF is produced once a year, with a lag of about a year and a half. For example, the 2007 T1FF, containing information from December 31, 2007, became available in September 2009. The T1FF contains people's dates of birth and dwelling addresses, meaning that with this information, the stratification can be updated to reflect moves, births and deaths. The T1FF also makes it possible to add a stratum for dwellings that were new or vacant for the Census but are no longer new or vacant at the time of CHMS collection. However, the quality of the stratification update depends on the quality of the information in the T1FF and the quality of the linkage between the frame and the T1FF.

## 3. Quality and extent of the stratification update

### 3.1 Quality of the information in the T1FF

Even though the T1FF provides more up-to-date information than the Census, the T1FF by itself is not used as a survey frame for two reasons. First, the T1FF's coverage is not as good as the coverage obtained by combining the Census and the AR. For example, vacant dwellings are not in the T1FF, and the quality of the addresses in the T1FF is generally not as good because of the nature of the file (for the CHMS, a dwelling is dropped from the frame when its address is missing or unusable). Second, the T1FF is built around the concept of "family", which is not the same as the CHMS's concept of "household" (second stage of sampling).

Neither the Census nor the AR can be used to assign new or vacant dwellings to a stratum. It is clearly beneficial, therefore, to assign them to a stratum using the T1FF. Yet for other dwellings that can be assigned to a stratum using the Census, the update strategy depends on the quality of the T1FF.

Before the start of Cycle 2, there was a preliminary study of respondents at the last few sites of Cycle 1 of the CHMS to assess the T1FF's quality in terms of stratification efficiency. Once again, the stratification is efficient if each dwelling's frame-derived stratum is the same as its "actual" stratum. To assess the quality of the T1FF, the stratification efficiency that would have been obtained with the 2006 T1FF was compared with the efficiency that was obtained with the 2006 Census. The study showed that the 2006 T1FF is no better or worse than the 2006 Census for stratification efficiency. It was therefore decided to take a cautious approach for the first few sites in Cycle 2 and confine the stratification update to births (dwellings with at least one child aged 3 to 5 in the T1FF are assigned to the " $3-11$ " stratum).

However, as the Census recedes in time and more up-to-date versions of the T1FF become available, it is reasonable to assume that it is more advantageous to use the T1FF. The 2007 T1FF became available at the time the dwellings were selected for collection site 3 . The stratification efficiency was assessed again using the results of collection sites 3 and 4 . This took until the end of site 4 collection, more than four months later. The results of the study showed that an update based on the 2007 T1FF would have improved the stratification efficiency for sites 3 and 4 not only in the "3-11" stratum, but in all the dwelling strata. The decision was therefore made to carry out a full update of the dwelling strata when the sample was selected for site 7 . More details are provided in the stratification efficiency charts presented in Section 4.

### 3.2 Quality of the linkage between the survey frame and the T1FF

The linkage between the survey frame and the T1FF is a deterministic address-based linkage. It is carried out in three steps to allow for incomplete addresses: an initial link with full address, a link with address minus postal code, and a link with address minus city. The linkage itself is straightforward, but the entire process is rather complex, since it includes the standardization of addresses and the geocoding of T1FF records. To facilitate linkage, the address is standardized with a computer application (P-CODE), and the linkage process is inasmuch as possible limited to one collection site at a time. To confine linkage to a single collection site, one collection site has to be assigned to each record in the T1FF. This involves a complex process of geographic coding (assigning a census block to each T1FF record).

Table 3.2-1 presents the linkage rates, namely the percentage of dwellings in the survey frame (a frame based on the Census with the addition of new dwellings from the AR ) that can be linked with the T1FF. The linkage rate varies considerably from site to site, as it depends heavily on the quality of the addresses. The results are generally not as good for rural sites. Note that when no link with the T1FF is possible, the stratum is not updated and is therefore based solely on census information.

Table 3.2-1
Linkage percentage between the survey frame and the T1FF, first 10 sites, Cycle 2
Linkage Percentage Between the Survey Frame and the T1FF

| Collection Site |  |  |  |  |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $72 \%$ | $75 \%$ | $43 \%$ | $71 \%$ | $62 \%$ | $56 \%$ | $74 \%$ | $57 \%$ | $59 \%$ | $52 \%$ |

### 3.3 Extent of the stratification update

### 3.3.1 Extent of the update for dwellings with a Census-derived stratum

Figure 3.3.1-1 shows the stratification update rates for dwellings with a Census-derived stratum for the first nine collection sites in Cycle 2: "3-11", "12-19", "20-39", " $40-59$ " and " $60-79$ " dwellings and dwellings that are out of scope because of age. The percentage of those dwellings that could not be linked with the T1FF (shown in grey) is $31.1 \%$. Hence, linkage with the T1FF was possible for $69 \%$ of dwellings with a Census-derived stratum. For $56.1 \%$ of the dwellings with a stratum (shown in green), the T1FF-based stratum is the same as the Census-derived stratum. For $1.4 \%$ of the dwellings (shown in pink), the stratum has changed to " $3-11$ " because, according to the T1FF, there is at least one child aged 3 to 5 in the dwelling. Since it was decided not to update the other dwelling strata for the first six sites, the Census-derived stratum was retained even though it did not link the T1FF-based stratum for $11.5 \%$ of the remaining dwellings. A full stratification update was carried out for sites 7 to 9 . For $1.1 \%$ of the dwellings with a Census-derived stratum (shown in yellow), the stratum has changed to "3-11" because, according to the T1FF, there is at least one child aged 6 to 11 in the dwelling. The stratum of $10 \%$ of the dwellings (shown in blue) was changed to " $12-19$ ", "20-39", "40-59" or " $60-79$ ". Lastly, for $0.4 \%$ of the dwellings (shown in orange), the stratum was switched to "other" since they are out of scope because of age, according to the T1FF.

Figure 3.3.1-1
Percentage of stratification updates, dwellings with a Census-derived stratum, first nine collection sites, Cycle 2


### 3.3.2 Extent of the update for new dwellings found in the AR

Figure 3.3.2-1 shows the stratification update rates for new dwellings for the first nine collection sites. There was no link with the T1FF for $71.6 \%$ of the new dwellings (shown in grey) as linkage is difficult for new dwellings because the AR provides more recent information than the T1FF. The 2006 and 2007 T1FFs, used for the first nine sites, contain new dwellings from 2006 to 2007, while the AR provides new dwellings found up to 2010 . For $7.3 \%$ of the new dwellings (shown in pink and yellow), the stratum has changed to " $3-11$ ". The stratum of $19.5 \%$ of the new dwellings (shown in blue) was changed to "12-19", "20-39", "40-59" or "60-79". Lastly, for $1.7 \%$ of the new dwellings (shown in orange), the stratum was switched to "other" since they are out of scope because of age, according to the T1FF.

Figure 3.3.2-1
Percentage of stratification updates, new dwellings, first nine collection sites, Cycle 2


### 3.3.3 Extent of the update for dwellings that were vacant at the time of the Census

Figure 3.3.3-1 presents the stratification update rates for dwellings that were vacant at the time of the Census. There was no link with the T1FF for $63.2 \%$ of the vacant dwellings (shown in grey) as linkage is more difficult for vacant dwellings because they are not in the T1FF. Therefore, the percentage of dwellings that could not be linked includes a portion of dwellings that are still vacant. For the rest of the updates, the results are similar to the results for new dwellings.

Figure 3.3.3-1
Percentage of frame updates, dwellings that were vacant at the time of the Census, first nine collection sites, Cycle 2


## 4. Improvement in stratification efficiency

### 4.1 Improvement in stratification efficiency by collection site

Figure 4.1-1 shows the stratification efficiency by collection site for the first seven sites in Cycle 2. The stratification efficiency rate is the percentage of dwellings, excluding dwellings in the "other" stratum, whose frame-based stratum is the same as their "actual" stratum. Since the sites are numbered chronologically, the chart shows how the efficiency changes over time as collection progresses. It is important to note, however, that the stratification efficiency also depends on the population growth at each site. Since each collection site is different, the data shown are not actually a time series.

The blue line represents the efficiency rate for each collection site in Cycle 2. The broken green line represents the Census-based stratification efficiency rate that would have been obtained without updating the stratification with the T1FF. For the first seven sites in Cycle 2, the average efficiency rate was $76 \%$; it would have been $73 \%$ without the T1FF. Therefore, the increase in stratification efficiency associated with using the T1FF was $3 \%$. In addition, using the most recent T1FF seems to make a significant difference, as the efficiency gain improved with site 3 when the 2007 T1FF became available.

The pink line represents the efficiency rate with a partial update for dwellings in the " $3-11$ " to " $60-79$ " strata and dwellings that were out of scope because of age (the stratification is updated only when there is a person aged 3 to 5 in the dwelling, according to the T1FF). This type of update was performed for the first six collection sites, and as a result, the blue and pink lines coincide for the first six sites. The figures for site 7 show that carrying out a full stratification update with the T1FF was very beneficial, as the efficiency gain was almost $8 \%$.

Finally, it is worth examining what the results would have been if a full update had been performed at the beginning of Cycle 2. The orange line shows the stratification efficiency rate with a full update based on the T1FF. The full update becomes worthwhile only after the 2007 T1FF comes into play, starting with site 3 . However, as explained in section 3.1, that observation could not be made until collection was completed for sites 3 and 4, just prior to the frame update for site 7 .

Figure 4.1-1
Stratification efficiency by collection site, first seven sites, Cycle 2


### 4.2 Improvement in stratification efficiency by dwelling stratum

Figure 4.2-1 shows the stratification efficiency by dwelling stratum for the first seven collection sites in Cycle 2. As in the previous figure, the blue line represents Cycle 2, the green line the Census, and the orange line the full update. A comparison of the blue and green lines shows that updating with the T1FF resulted in an efficiency gain for all strata (even the " $60-79$ " stratum, which had a small $0.2 \%$ gain). The impact on the stratification efficiency was particularly large for dwellings in the " $20-39$ " stratum, which had a $14 \%$ efficiency gain following the stratification update with the T1FF in Cycle 2. The orange line in the chart shows that the full update at the beginning of Cycle 2 would have been beneficial for all strata except perhaps the "40-59" stratum (in part because of sites 1 and 2 , for which the full update would not really have been beneficial).

Figure 4.2-1
Stratification efficiency by dwelling stratum, first seven collection sites, Cycle 2


## 5. Conclusion

In conclusion, stratification efficiency is particularly important for the CHMS because of the high costs associated with the direct measurements at the mobile clinic and the laboratory analyses of blood and urine samples. For that reason, it is preferable to increase the sample size as little as possible in order to reduce stratification errors when we want to reach the required number of respondents in each age-sex group. That is why stratification is so important. Stratification efficiency is clearly improved by the T1FF-based stratification update. In Cycles 1 and 2, it has been
rather easy to achieve the target numbers of respondents attending the mobile clinic for each age-sex group. Consequently, adding new dwellings from the AR and updating the stratification with the T1FF will continue in the future. Updating the stratification with the T1FF becomes more desirable as the Census recedes in time, and it is important to use new versions of the T1FF as soon as they become available.

Finally, the method used to link the T1FF and the survey frame could undoubtedly be improved. For example, consideration should be given to using probabilistic linkage instead of deterministic linkage or employing a linkage application that would increase the link rate.

## References

Giroux, S. (2007). Canadian Health Measures Survey: Sampling strategy overview. Health Reports, vol. 18 (Special Issue), Statistics Canada Catalogue n ${ }^{\circ}$ 82-003-S, pp. 35-40.

Giroux, S., Malo, D., Neusy, E. and Provost, M. (2010). Using the Census as a sampling frame: some recent examples. HSMD Working paper 2010-005E, Ottawa, Canada: Statistics Canada.

SESSION 2B

## POTENTIAL USES

# Accounting for temporal effects, sampling variability, incomplete data and reporting error in the integration of consumer expenditure data from survey and administrative-record sources 

John Eltinge ${ }^{1}$


#### Abstract

A primary goal of the U.S. Consumer Expenditure Survey (CE) is to produce estimates of mean consumer expenditures at a relatively fine level of aggregation defined by the Universal Classification Code. At present, CE produces its estimates based on data collected through diary and interview surveys. As part of the effort to improve the balance cost, burden and quality, one could consider supplementing the abovementioned survey sources with administrative-record data. Two examples of prospective administrative data sources are large retail outlets and financial transaction intermediaries. Evaluation of that balance involves three principal issues: (1) for surveys costs, burden and error sources (including population variability, incomplete data and measurement error); (2) for administrative records - the same factors, plus additional factors related to operational risk; (3) for both - variability of factors in (1) and (2) over time.


This paper presents a general methodological framework for the evaluation of issues (1) through (3). This framework leads to several suggestions regarding components of cost, burden, quality and risk for which it is especially important to collect solid empirical information.

[^6]
# The potential of administrative education data used to measure subnational migration in England 

Stephen Jivraj ${ }^{1}$


#### Abstract

Migration is an inherently difficult phenomenon to measure as people who move are difficult to track. In the absence of a population register in the UK, two main datasets have been used to measure subnational migration. The national Census and National Health Service Central Register (NHSCR) of all patients registered with a general practitioner. This paper introduces and examines the potential of the English School Census, an administrative source of subnational migration data that can provide more up to date information than the Census and more detailed socioeconomic and geographical detail than NHSCR. Since 2002, the School Census has been derived from an electronic form completed by each state school in England to cover all enrolled pupils in January of each year. Through the inclusion of a unique pupil identifier, which remains the same throughout a pupil's school career, the data can be matched over time and the change of home address can indicate migrant pupils. Empirical comparison with the 2001 Census and NHSCR shows that the School Census provides similar levels and patterns of subnational migration over time when allowing for the different way in which migration is measured in each dataset. The paper concludes that the data has potential for monitoring subnational migration reflected by the Office for National Statistics in England and Wales exploring its use for local population estimation.


Key Words: Subnational migration, England, School Census, National Health Service Central Register, 2001 Census.

## 1. Introduction

Migration is an inherently difficult phenomenon to measure as people who move residence are difficult to track. National longitudinal datasets can provide a useful source of subnational migration data, however, such studies often find it difficult to retain respondents over time if they move house. Subnational migration can also be measured by asking people retrospectively in a cross-sectional survey whether they have moved within a given time frame. This type of measurement is not without its own limitations. Respondents to retrospective questions in surveys will often find it difficult to remember whether they moved within the time frame specified.

In the UK, the two main datasets that have been used to measure subnational migration are the decennial Census and the National Health Service Central Register (NHSCR) of all patients registered with a general practitioner (GP). This paper introduces and examines the potential of the English School Census, a relatively new source of subnational migration data that can provide more up to date information than the Census and more detailed socioeconomic and geographical information than the NHSCR.

## 2. School Census

The School Census, which is also known as the Pupil Level Annual School Census records details of all state-school pupils in England, and is updated annually. The School Census data is derived from an electronic administrative form completed by each school to cover all enrolled pupils in January of each year (Machin et al, 2006). It is collated nationally by the Department for Education (DfE) through Local Education Authorities. Completion of the SC has been a statutory requirement for all state maintained primary, secondary and special schools since 2002 under section 537A of the Education Act 1996 (Harland and Stillwell, 2007).

[^7]The data forms part of the National Pupil Database, which is a warehouse of education data for national attainment tests and information relating to schools and their staff. The School Census, which provides a link to other data sources, consists of entries for every pupil on roll including information about home postcode, ethnicity, free school meal eligibility (an indicator of family socio-economic status), age, gender and first language. Through the inclusion of a unique pupil reference the data can be matched between years to form a longitudinal source.

Marquis and Jivraj (2009) have assessed the quality of data as supplied by the DfE and have used appropriate interpolation techniques developed by Harland and Stillwell (2007) to clean errors and omissions. The data is found to be of high quality, though changes over time in the way some data are recorded means that certain variables are more prone to inconsistency than others. For example, a different ethnicity classification was used from 2003 which is consistent with the 2001 Census ethnic group classification. The main conclusion regarding the use of School Census data for any research purpose is that cleaning is considered essential before analysis (Ewens, 2005; Harland and Stillwell, 2007; Marquis and Jivraj, 2009).

## 3. Comparison of migration dataset's characteristics

This section provides a comparison of the measure, coverage (population, spatial and temporal) and attributes recorded for the subnational migration data which can be derived from the School Census, NHSCR and decennial Census. In the School Census, a change of a postcode for a pupil in the matched data can be used to measure subnational migration within England. This could be between two consecutive years or over a longer period for which the data is available providing the pupil is present at both time points. This type of measurement of migration is referred to as a transition because not all movements made will be identified. For example, if a pupil moves more than once during the period between two School Censuses only one movement will be recorded and if a pupil moves away and then returns to their original location between School Censuses they will not be recorded as a migrant. The decennial Census also records migration in this way although it uses a retrospective question which asks whether a respondent had a different usual address one year prior to enumeration. The NHSCR, on the other hand, records all moves an individual makes when they re-register with a GP in a different health authority ${ }^{2}$.

The main limitation of School Census subnational migration data is that it only includes school-aged children attending state schools. Therefore, the data will not reflect the movement of all people as families with school-aged children are less likely to migrate and move to different types of places compared with other households (Bailey and Livingston, 2005; Meen et al, 2005). Moreover, the data do not reflect the movements of all families with schoolaged children, as there is no information for those in private education. In England, approximately $92 \%$ of schoolaged children attend a state school, however, this varies between different parts of the country (ONS, 2004). In contrast, the 2001 Census and NHSCR are not limited to small subsets of the population and, in theory, should include almost all people.

The release of the School Census with a postcode identifier for individual pupils means that data can be aggregated to any higher level geography. This is useful when one is interested in arbitrary geographies or movement between small areas. Small area analysis is also possible with the 2001 Census subnational migration data which was released at output $\operatorname{area}^{3}$ at the lowest geographical level. To protect confidentiality, however, the Office for National Statistics (ONS) in England and Wales have used a procedure called Small Cell Adjustment Mechanism which can lead to data between the same released tables at different geographies not adding to a consistent figure (Duke-Williams and Stillwell, 2007). This can create noise in small area analyses. The NHSCR can only provide subnational migration data between health authorities at the lowest geography. Nonetheless, through the combination of individual health authority patient records, ONS have been able to produce annual subnational migration estimates at local authority district level ${ }^{4}$ since 1999 (ONS, 2007).

The School Census has been collected since 2002. Its annual release means that it can provide a more up-to-date and continuous measure of subnational migration compared with the decennial Census which has asked a question about

[^8]migration since 1961. The School Census has been collected tri-annually since 2006 for secondary schools and since 2007 for primary schools, and as a result future data releases will be able to provide even more detail of the residential movement of school aged pupils (Harland and Stillwell, 2007). The temporal coverage of the NHSCR is much more complete than both the School Census and decennial Census, with migration data available since 1975.

The recorded characteristics of migrant pupils in the School Census include age, gender, ethnicity, first language and free school meal status. These breakdowns are limited in comparison with the variables available in the 2001 Census for subnational migrants including age, sex, ethnicity, family status, limiting long term illness, economic activity, socioeconomic class, tenure, and household composition. All Census variables, however, are recorded at the time of enumeration which makes it difficult to determine, for example, how many migrants were living in rented accommodation one year before the Census. For certain variables, including age and sex it is possible to allocate subnational migrants back to their status one year previously. The contextual information recorded in the NHSCR is more limited. Subnational migrants can only be disaggregated by age and sex.

## 4. Empirical comparison with National Health Service Central Register

This section empirically compares subnational migration data which can be derived from the School Census and NHSCR using breakdowns of migrants by age and space. Figure 4-1 compares the average annual level of subnational migration between local authority districts within England recorded by the School Census and NHSCR during the period 2002 to 2006. The NHSCR records a higher number of moves between districts than the School Census at all ages. This is not surprising because the NHSCR records all moves in a year when a patient re-registers with a GP in a different health authority.

The difference between the two datasets is fairly constant for all primary school aged pupils ( 6 to 10 ), but not for secondary school aged pupils ( 11 to 15). In both datasets, the number of moves gradually declines during school age. The level of moves during secondary school age appears to remain fairly constant in the NHSCR except for slight peaks at age 12 and 14 when there were larger cohorts of school aged people. In the School Census, the number of moves falls markedly after age 11. The different pattern in the number of flows during secondary school age is probably a result of more children attending non-state schools at this age.

Figure 4-1 Average annual between local authority district migration flows for $\mathbf{6}$ to 15 year olds from School Census and NHSCR by year of age, 2002-06


Source: National Health Service Central Register, 2002-06
Figure 4-2 shows the degree of association between inflows and outflows of subnational migration between each individual local authority district in England in the School Census and NHSCR. The charts show the average number of each type of flow for compulsory school aged children ( 6 to 15) between 2002 and 2006. There is a strong positive relationship between the inflows and outflows in each dataset. The correlation coefficients for average inflows and outflows between 2002 and 2006 were all above $97 \%$.

The outliers in each scatterplot for inflows and outflows tend to include districts where the compulsory school aged population is considerably lower in the School Census compared with ONS mid-year population estimates for the area. For example, in-migration flows are greater in the NHSCR than the School Census, even after accounting for the generally higher level of subnational migration recorded in the NHSCR, in Waverley, Windsor and Maidenhead,
and Horsham. These local authority districts contain a relatively high number of private schools which may explain the higher levels of inflows into these areas by people of school age measured by the NHSCR. These flows will not be accounted for in the School Census because it only includes state school pupils.

Figure 4-2 Relationship between average annual inflows and outflows for each local authority district in England for 6 to 15 year olds in School Census and NHSCR, 2002-06


Source: National Health Service Central Register, 2002-06

## 5. Empirical comparison with 2001 Census

This section provides an empirical comparison for subnational migration data derived from the School Census and 2001 Census. This comparison is constrained by the different time points each dataset is available. The 2001 Census records subnational migration between April 2000 and April 2001 whereas the School Census records annual pupil subnational migration from January 2002. Nonetheless, comparison is worthwhile because it can show flows within districts, which accounts for three quarters of total migration recorded by the School Census.

Figure 5-1 compares the migration flows within districts from the School Census between 2002 and 2003 and the 2001 Census in the year leading up to April 2001 by year of age. There are more flows within districts recorded in the School Census compared with the 2001 Census for each year of age. The distribution of migrants is similar across the compulsory school age range in each dataset except at age 11. In the School Census there are a much higher number of subnational migrant pupils at age 11. This is likely to be a result of pupils who did not inform their school at time of moving updating their record when they change school and filed new administrative forms. The greater number of moves at the age when pupils change from primary to secondary school could also be a result of administrative errors. For example, if during the process of wholesale updating of a new secondary school age pupil's address details at age 11 the postcode is entered incorrectly they could be misidentified as a migrant.

Figure 5-1 Within local authority district migration flows for 6 to 15 year olds from School Census and 2001
Census by year of age


Note: School Census data for period 2002 to 2003; Census data for period 2000 to 2001
Source: 2001 Census Commissioned Table C0527
Figure 5-2 compares the number of migration flows for single year of age between local authority districts recorded by the School Census between 2002 and 2003 and the Census between 2000 and 2001. The number of between district moves is very similar for each year of age. The number of between district flows is greatest at the youngest compulsory school ages and there are fewer flows for older children in both datasets. Between the ages of 12 and 15 there are slightly higher flows of between district moves recorded by the 2001 Census. This is likely to be a result of higher levels of secondary school aged children attending non-state schools.

Figure 5-2 Between local authority district migration flows for 6 to 15 year olds from the School Census and the 2001 Census by year of age


Note: School Census data for period 2002 to 2003; Census data for period 2000 to 2001
Source: 2001 Census Commissioned Table C0527
Figure 5-3 shows the relationship between within area flows for each individual local authority district in England recorded by the School Census and 2001 Census. The chart shows there is a very strong positive relationship between the two datasets for within district flows. This suggests that although the School Census records a higher number of subnational migrants within local authority districts the spatial pattern of these moves are very similar to the 2001 Census.

Figure 5-3
Relationship between within local authority district flows for 6 to 15 year olds in School Census and 2001 Census


Note: School Census data for period 2002 to 2003; Census data for period 2000 to 2001
Source: 2001 Census Commissioned Tables C0366a-C0366d

## 6. Conclusions

The School Census has potential to provide new insights into the level and pattern of subnational migration in England. The potential of the data is recognised by ONS. They are currently considering using it as part of their official population estimation (ONS, 2009). It can provide a more up-to-date measure of subnational migration than the decennial Census and more detailed contextual information about subnational migrants than NHSCR. The data are, however, limited to the measure of movement of school aged pupils attending state schools. This means that it is difficult to make inferences about the subnational migration patterns of the population as a whole, because families with children of compulsory school age are less likely to move and move to certain places compared with other family groups.

Comparison of data from the NHSCR and the 2001 Census shows that the School Census records similar patterns and trends of subnational migration. Levels of subnational migration between local authority districts are similar in each dataset when allowing for the different way in which subnational migration is measured in NHSCR. Differences between the School Census and NHSCR appear to be due to the population included in each dataset. For example, subnational migration levels are different in some areas which have a lower proportion of school aged pupils in state schools. The School Census does appear to over count the number of moves which take place at age 11, which is when school children change school as they move between primary and secondary education. The updating of records at this age may account for a higher number of moves being recorded at this age, which actually occurred sometime previously.

## Acknowledgements

The author would like to thank Mark Brown, Nissa Finney, Naomi Marquis, Ludi Simpson and Nikos Tzavidis for their contributions to this paper. I am also grateful to the Department for Education for supply of School Census data, Migration Statistics Unit at the Office for National Statistics for the supply of NHSCR data, and the Centre for Interaction Data Estimation and Research for assistance with the 2001 Census Special Migration Statistics. Stephen Jivraj's research is funded by an Economic and Social Research Council postgraduate studentship.

## References

Bailey, N. and Livingston, M. (2005). Determinant of individual migration: an analysis of SARs data. Scottish Centre for Research on Social Justice Working Paper No. 3.

Duke-Williams, O. and Stillwell, J. (2007). Investigating the potential effects of small cell adjustment on interaction data from the 2001 Census. Environment and Planning A, 39, 1079-1100.

Ewens, D. (2005). The National and London Pupil Datasets: An introductory briefing for researchers and research users. Greater London Authority, DMAG Briefing 2005/8.

Harland ,K. and Stillwell, J. (2007). Using the PLASC data to identify patterns of commuting to school, residential migration and movement between schools in Leeds. School of Geography, University of Leeds, Working Paper 07/03.

Machin , S., Telhaj, S. and Wilson, J. (2006). The Mobility of English School Children, London School of Economics, London.

Marquis, N. and Jivraj, S. (2009). Preparation of Pupil Level Annual School Census data for the analysis of Internal Migration. Centre for Census and Survey Research, The University of Manchester, Working Paper 2009-03.

Meen, G., Gibb, K.., Goody, J., McGrath, T. and Mackinnon, J. (2005). Economic segregation in England: Causes, consequences and policy, (Policy Press, Bristol).

ONS (2004). Pupils and teachers: by type of school, 2002/03. Regional Trends 38.
ONS (2007). Estimating Internal Migration: Customer Guidance Notes, Migration Statistics Unit, Office for National Statistics, http://www.statistics.gov.uk/statbase/EXPODATA/commentary/EstimatinginternalMigration.doc

ONS (2009). Use of School Census data to Improve Population and Migration Statistics, Office for National Statistics, www.statistics.gov.uk/about/data/.../population/.../SC_potential_u

# From Neighbourhood Statistics to Beyond 2011Making Better Use of Administrative Data 

Louise Morris and Minda Phillips ${ }^{1}$


#### Abstract

This paper outlines both the overall approach and specific measures taken by the Office for National Statistics to make better use of data from administrative sources in two key projects. Initial efforts focused on the acquisition, processing and dissemination of aggregate outputs to support the National Strategy for Neighbourhood Renewal. Development of new policies, procedures and statistical tools enabled provision of reliable and consistent information from a local to a national level via the Neighbourhood Statistics website. Since 2008, with new legislative opportunities, priority has been given to accessing record level data to take forward a major programme of improvements to population and migration statistics. As part of this work the Beyond 2011 Project will assess the longer-term feasibility of using administrative data, either alone, or in conjunction with other data sources, to meet ongoing requirements for population and wider socio-demographic statistics. This has created further challenges most notably on data quality, the interplay between administrative and other data sources and the linkage of data in a country without a single personal identifier.


Key Words: Administrative Data; Neighbourhood Statistics; Population Statistics.

## 1. Introduction

### 1.1 Overview

At a time of falling response rates and rising costs, there are significant challenges to meeting user needs for highquality population and wider social statistics on a timely and frequent basis. This is particularly the case for statistical systems, such as that of the UK, largely based on a traditional survey and census approach. The rate of change in the UK population has been increasing over the last few decades and a traditional approach cannot react quickly to reflect changes in society, with estimates produced on this basis quickly becoming out of date. The UK, like many other countries, is therefore assessing the possibility of making better use of administrative data sources for statistical purposes. A key advantage in the use of administrative data being that they are often available on an ongoing basis to provide more frequent and timely data. Use of these sources however brings with it a number of challenges, some of which are particularly significant for a country with no single personal identifier or maintained population register source. Nevertheless there have been a number of developments in the use of administrative data in the production of social statistics.

### 1.2 Use of Administrative Data

In a limited number of areas administrative data has, for some time, provided a key source of information, in particular in the production of statistics on, for example, vital events. Administrative data also are a core component in the production of business statistics providing structural information to develop and update the Business Register and auxiliary information to reduce sample sizes and improve the precision of survey estimates. The potential to make better use of administrative data in the production of population and social statistics has been widely recognised and highlighted in a number of published reports. In particular, a 2008 report of the Treasury Select

[^9]Committee ${ }^{2}$ recommended improved use of government administrative databases in the production of population and migration estimates. However, restrictions on access to data sources have, until recently, limited the scope to assess and make better use of many administrative data sources for statistical purposes.

Whilst early developments at the Office for National Statistics (ONS) focussed largely on use of aggregate outputs to provide information on characteristics of neighbourhoods, more recent developments have considered the scope to make better use of administrative data as a core component of a system to produce population and wider sociodemographic statistics. In both cases use of administrative data has brought with it a number of challenges, particularly in relation to data quality and operational issues and the interplay between administrative and other data sources. This paper discusses some of the challenges and developments in the use of administrative data within the context of two key projects: Neighbourhood Statistics (NeSS) and Beyond 2011. Section 2 provides an overview of NeSS and challenges faced in the use of administrative data in this context. Section 3 describes legislative changes that have provided the opportunity for increased use of administrative data sources for statistical purposes. Finally, section 4 describes current work to consider more use of administrative data in the production of population and socio-demographic statistics within the context of the Beyond 2011 project.

## 2. Neighbourhood Statistics

### 2.1 Background

Whilst current work is concentrated on the use of administrative data for the production of population estimates, earlier developments focussed more on use of administrative data as a key source for the provision of outputs for small geographic areas. Established initially from work undertaken by the government Policy Action Team 18 (PAT18) to support the National Strategy for Neighbourhood Renewal ${ }^{3}$, the NeSS project is an example of the early work on the use of administrative data for official statistics purposes. The PAT18 report ${ }^{4}$ underlined the need for and importance of having comprehensive and up-to-date information about deprived neighbourhoods. NeSS was initially established to provide the statistical information to meet this requirement but now has a much broader range of users and uses. Focussing on the use of low level aggregate data, NeSS successfully brought together census, survey and administrative sources to provide coherent and consistent information about neighbourhoods from very small local areas to a national level.

### 2.2 Use of Administrative Data Sources

Whilst there had been previous uses of administrative data for statistical purposes, NeSS faced a number of new challenges including the requirement for access to data sources outside the direct control of the statistical department (or the national statistical institution). The requirement to combine data from a number of independent sources into an integrated system for the provision of coherent results on a broad range of topics in an easily accessible format provided further challenge. Issues relating to geography, data quality and confidentiality were especially challenging and steps taken to resolve these issues are discussed in more detail below.

### 2.2.1 Geography

There is a wide range of geography classifications in place in the UK used for a variety of purposes (health, electoral, postal, etc.). Differences in coverage of the classifications, frequent boundary changes and variations in unit size (geographic area or population coverage) both within and across geographies make comparisons over time

[^10]and between areas difficult. Consistent outputs could not be produced from datasets compiled on the different geographic bases. Limitations of existing geography classifications led to the development of a statistical geography designed specifically for NeSS purposes based on the 2001 Census Output Area (OA) classification. The Super Output Area (SOA) geography developed consists of two layers: Lower Layer Super Output Area (LSOA) and Middle Layer Super Output Area (MSOA) forming a hierarchy of areas intermediate in size between OAs and local authorities. Within the SOA, classification units are more similar in size of population than is the case for other geography classifications (e.g. electoral wards). SOA classifications are also intended to be more stable over time.

The development and provision of an address matching tool to supplying departments ensured that all source datasets could be geo-coded on a consistent basis to this new SOA statistical geography classification.

### 2.2.2 Quality

Restricting potential data sources to those over which government departments have most control - data sources collected locally or centrally but always centrally maintained - went some way towards safeguarding quality. It was however also necessary to address the more general challenges associated with use of administrative data for statistical purposes. In utilising data sources often collected for very different administrative purposes it was important to have a comprehensive understanding of their scope and content and to identify issues affecting the quality of the data and its suitability for NeSS. As is often the case there were differences in recording practices and definitions applied both between sources and between administrative and statistical definitions. Units, such as households for example, may be recorded differently for statistical and administrative (e.g. tax) purposes and it is important that all differences that may impact on the use of the data are identified and understood. Lack of control over data sources also means that any changes in the content and coverage over time need to be clearly understood.

Policies on data quality, revisions and geography were developed for NeSS, a number of which have subsequently been incorporated into the National Statistics Code of Practice ${ }^{5}$. Potential data sources were assessed against the quality policy to determine suitability and the ongoing requirement to provide detailed metadata to accompany each data source ensures that NeSS users have a clear understanding of the strengths and limitations of the published information as well as change over time.

### 2.2.3 Confidentiality

Preserving the confidentiality of individual records is essential for any statistical organisation. This provided a particular challenge for NeSS where the objective of providing outputs for small geographic areas split by a number of classificatory variables meant that the potential for disclosure of information about an individual could be high. Based on tools developed by Statistics Netherlands, policies and tools were developed for NeSS and provided to supplying departments to ensure that commitments and legal requirements to maintain data confidentiality were adhered to prior to the publication of information on the NeSS website ${ }^{6}$.

## 3. Legislative Changes

### 3.1 Statistics and Registration Service Act 2007

NeSS provided the opportunity to develop an improved understanding of the challenges faced when using administrative data for statistical purposes. Lessons learnt and policies, procedures and tools implemented have since effectively supported use of administrative data in a number of other projects. However, there was little scope to further extend the use of administrative data, and in particular to utilise data at record level until recent legislation made specific provision for the sharing of data for statistical purposes.

The Statistics and Registration Service Act (SRSA) 2007 enables public authorities to share administrative data for statistical purposes. This Act provides the basis for establishing a legal gateway (via the preparation of Information

[^11]Sharing Orders) in cases where access to the data is prohibited or there is no existing legal power which would allow the data to be disclosed. Each Information Sharing Order requires the consent of relevant Ministers and the approval of Parliament. Even when an Information Sharing Order is in place the disclosure of the data is not automatic and can only proceed with the permission of the relevant data owner.

### 3.2 Challenges

Securing access to record level data can be both resource intensive and time-consuming. Early experiences of using the SRSA have identified a number of challenges with legal aspects of the process. Developing, at an early stage, a clear understanding of the scope of the data source of interest including how and when data are collected is important both to ensure that the data are fit for purpose and to assist with development of the required business case. As with aggregate data an understanding of the quality (coverage, completeness, etc) data definitions and classifications is crucial. Early stage feasibility work and collaboration with supplying departments has proved invaluable in developing a sound understanding of a data source, its strengths and limitations.

Issues of data security and confidentiality must be addressed and a secure environment with controlled access established. In addition to legal requirements, ensuring compliance with the supplying department's policies and procedures provides a further challenge. Nevertheless the SRSA provides the scope to significantly broaden the use of administrative data in a statistical context. Improvements to population statistics have already been implemented and further enhancements are being considered as a consequence of securing access to data sources for specific population sub-groups (school children, students and migrants). Administrative data are likely to feature as a key component of any alternative system considered by Beyond 2011 and at an early stage of the project access to priority data sources will be sought.

## 4. Beyond 2011

### 4.1 Current System

As noted above (section 1.1), the current UK social and population statistics system continues to be largely based on a traditional census and survey approach. A census is carried out on a ten yearly basis supported by a large-scale social survey programme to update estimates in the inter-censal years. Population estimates are also based on Census outputs updated annually using a cohort component method. This approach does make some use of administrative data to provide information on population change in the inter-censal years but current systems consist largely of data collected specifically for statistical purposes.

The advantages of this census based approach are well rehearsed: at a given point in time and with regular frequency a census provides a rich and comprehensive picture of the whole population in a country. As with a census, statistical organisations also have control over the content, design and data collected by means of a social survey programme. This approach is however costly and costs are continuing to increase, partly due to a decreasing willingness of the general public to cooperate. Administrative data sources are becoming increasingly available within government and their potential to provide basic statistical information is recognised. A number of countries, in particular Scandinavian countries, utilise rich and well-established register-based systems better suited to counting the population more frequently and efficiently to produce a broad (sometimes complete) range of census outputs. Other countries such as the USA and France have adopted alternative approaches providing population and census statistics through a combination of a short form Census and rolling survey approach (USA) or a rolling census design (France).

The fitness for purpose of the current approach in the UK, where the rate of change in the population has been increasing in the last few decades, has been questioned. Greater population mobility is making it increasingly difficult to provide an accurate count of the population or information on population characteristics on a frequent basis and at low levels of geography based on a traditional approach. However, lack of a unique personal identifier or maintained population register adds significant complexity to implementation of an administrative data approach.

### 4.2 Aims and Approach

A number of short and medium term improvements to population statistics, in particular migration statistics, have been made in recent years. Building on these developments the Beyond 2011 Project has been established to provide a coherent framework to meet the challenges to the continued production of population and broader sociodemographic statistics in the longer-term. The high-level aim of the Beyond 2011 work programme is to investigate the feasibility of improving population statistics in the UK by making use of integrated data sources to replace or complement existing approaches and, if feasible, to develop the longer term programme of work that would be needed to implement the recommended alternative model.

Initial work is focussing on gathering information on user requirements, assessing potential methods that may form the basis of an alternative system and completing an initial assessment of issues that would impact on the implementation of an alternative approach - legislative issues, security, IT infrastructure requirements, etc. A central focus of this work will be to consider how to develop population data of the required coverage and quality as efficiently and effectively as possible, focussing initially on collection of the minimum necessary data about the population at minimum cost whilst meeting public concerns about security, privacy and intrusiveness.

A range of potential alternative designs for producing the required population and social statistics will be assessed including both outputs traditionally provided by a census and the intercensal updates to population estimates. Potential approaches will be assessed against a range of criteria to determine whether alternative designs have the ability to produce the priority statistics identified through consultation with users. If alternative methods prove feasible a longer-term programme of work will be taken forward to implement the recommended alternative model. Designs under consideration range from alternative census designs (for example a short-form census supplemented by an ongoing survey similar to the US approach) to survey based approaches (including a rolling census option) and models based on administrative data sources. Whilst each type of data source (census, survey and administrative) has its own strengths and limitations, it is unlikely that a single source type will be able to meet the full range of priority user requirements and the project will consider how sources can best be combined in an alternative system.

### 4.3 Use of Administrative Data

Administrative data are likely to feature as a core component of any alternative solution, either as the basis for a system in its own right or to support designs based on census and survey approaches where it could, for example, be used in the production of small area estimates under a survey based design. As the basis of an alternative system an assessment will be made of the scope to create an administrative data based population spine (record level data) or to use low level aggregates to provide initial population counts. Administrative data may provide some topic information, estimates that cannot be produced on this basis could, for example, be produced through a household survey approach.

Low level aggregate approaches are likely to be based on development of a model to combine a number of administrative data sources to produce population estimates, supplemented by administrative or survey data for other outputs. This approach could update both the census and inter-censal estimates but scope to provide multivariate estimates would be limited. Unit level approaches are likely to use records from one or more sources to provide a basic population spine, supplemented by address information and other administrative and survey data to produce broader social statistics outputs.

### 4.4 Challenges

The requirement to reduce costs makes administrative data based systems appealing since associated operating costs are generally known to be lower than a census design. However, there are a number of significant challenges that will need to be overcome if an administrative data based system is to be successfully developed. A single data source is unlikely to provide the basic population information required to meet user needs and the lack of a consistent unique identifier across data sources provides a further challenge. It is likely to be a complex and resource intensive process to reconcile individual records across data sources. Under an aggregate data design the models to combine multiple data sources can also be complex to develop and maintain. Administrative sources are not usually able to
provide detailed information on person characteristics and when they do definitions and / or coverage may not be consistent with the definitions required for statistical purposes.

As with NeSS, the challenges in using data collected for administrative rather than statistical purposes will need to be addressed, adapting the data sources to make them fit for producing statistical outputs, developing a sound understanding of the quality, coverage and content of each source. At the early stage of the project there will be a focus on identifying key data sources of interest, assessing the suitability of a number of broad coverage data sources such as those held by the Department for Work and Pensions on benefit claims, HM Revenue and Customs on National Insurance and tax information and data on patient registrations with general practitioners. The scope to supplement such data to provide better quality information for specific population subgroups or population characteristics will also be assessed. Challenges associated with legislative aspects of the process will be a key consideration at this stage. As work progresses technical issues associated with the storage, security and processing of a number of large-scale data sources will be addressed.

## 5. Conclusion

As in many other countries, the UK is carrying out a programme of work to consider the scope to make better use of administrative data in the production of population and broader socio-demographic statistics. Whilst there are a number of significant challenges to be considered during the course of the work, previous developments in the use of administrative data and in combining administrative and statistical data sources for the production of neighbourhood statistics will be relevant to build on. Experience in securing access to record level data for previous projects has also enabled a realistic assessment of the issues, workload and timeframe for this aspect of the project. Initially feasibility work for the Beyond 2011 project itself will provide the opportunity to explore the extent to which it may be possible to exploit administrative and other data sources to meet key user requirements for population statistics in a UK context.

## References

House of Commons Treasury Committee (2008). Counting the Population, Eleventh Report of Session 2007-08, London: The Stationery Office.

National Strategy for Neighbourhood Renewal (2000). Report of Policy Action Team 18: Better Information, London: The Stationery Office.

Report by the Social Exclusion Unit (2001). A New Commitment to Neighbourhood Renewal, National Strategy Action Plan. London: Cabinet Office.

UK Statistics Authority (2009). Code of Practice for Official Statistics. UK Statistics Authority.

# Validation of financial data using tax records 

Philippe Wanner ${ }^{1}$


#### Abstract

This paper describes the use of tax records to validate the financial aspects of information collected in a survey on the living conditions of households in Switzerland. By matching these two sources using the reported values for last name, first name, date of birth and community of residence, information provided by survey participants can be cross-checked with entries in their tax returns. The results reveal that a bias of social desirability motivates individuals to inflate the income levels they declare in the survey and that a survey cannot fully capture the wealth of individuals.


Keywords: matching, tax records, validation.

## 1. Introduction

This paper is based on a feasibility study conducted on behalf of the Swiss Federal Statistical Office (Neuchâtel). In the context of this study, we sought to demonstrate how comprehensive administrative data can be useful for validating survey data. We drew on 2006 tax records from three Swiss cantons (Bern, Vaud and St. Gallen) to be matched with data from the Statistics on Income and Living Conditions (SILC) survey.

The objectives of this study were two-fold: to match records from the two sources, and once the records corresponding to the same individual were identified, to validate the economic data (income, wealth, etc.) collected in the survey by comparing their values with those declared in the tax records.

## 2. Matching the SILC survey with tax records

### 2.1 Method

For the first objective, deterministic and probabilistic matching methods were tested. To achieve this, the variables selected for matching - those common to the different data sources-were harmonized in the three sets of tax records and the SILC data. Specifically, the formats of dates of birth, nationality and marital status were harmonized, and some variables were renamed for consistency across sources. In addition, we tested several ways of accounting for variant spellings (eliminating accents, the Soundex procedure, etc.) and of harmonizing addresses, which often have several variations (Ch. or Chemin, Rte or Route, etc.). In a multilingual country like Switzerland, it can also happen that some names are transcribed according to German spelling conventions in one source and to French ones elsewhere (e.g., the "w" frequently becomes "v"), creating a further need to harmonize spelling. Finally, we had to account for the fact that in tax records, women are often listed under their official name (last name followed by their husband's name, if they opted to keep their maiden name), while typically only maiden names were used in the survey, making harmonization necessary.

Notwithstanding these various harmonizations, matching must make allowance for some errors (errors in dates, such as dates of birth, errors or variations in the spelling of names), so we adopted a probabilistic approach based on last name, first name, date of birth, sex, nationality, marital status, address and community of residence. The chosen
${ }^{1}$ Philippe Wanner, Laboratoire de démographie, université de Genève, Pont D'Arve 40, 1211 Genève 4, Switzerland (Philippe.Wanner@unige.ch)
approach takes potential errors in one variable or another into account. In particular, when matching on the first and last names failed, the Soundex procedure was applied. This procedure translates words into a phonetic code based on the pronunciation of the name.

In total, we find after matching that approximately $94 \%$ of surveyed households in the canton of St. Gallen appear in the tax records (recall that these are comprehensive - Table 1). The corresponding value is $86 \%$ for the canton of Vaud and slighter lower for the canton of Bern (84\%), where there appear to be certain issues involving the spelling of names, in particular among married individuals. The main constraints observed relate to spelling errors, which are difficult to correct or account for during matching. Also, in the case of the canton of Vaud, there is frequent confusion between the principal taxpayers and their spouses in the dates of birth.

## Table 1:

Final results for the canton of St. Gallen

|  | Sample | Percentage |
| :--- | :---: | :---: |
| St. Gallen | $\mathbf{9 2 4}$ |  |
| Number of individuals surveyed by SILC* | 924 |  |
| Number of households surveyed by SILC | 351 | $87.9 \%$ |
| Number of individuals found | 812 | $94.3 \%$ |
| Number of households found | 331 |  |
| Vaud | 1,509 |  |
| Number of individuals surveyed by SILC* | 612 | $42.4 \%$ |
| Number of households surveyed by SILC | 640 | $85.8 \%$ |
| Number of individuals found | 525 |  |
| Number of households found | 2,228 |  |
| Bern | 2,228 | 972 |
| Number of individuals surveyed by SILC* | 972 |  |
| Number of households surveyed by SILC | 1,392 | 814 |
| Number of individuals found |  | $83.7 \%$ |
| Number of households found |  |  |

### 2.2 Factors explaining non-matches

Non-matches occur for a number of reasons. First, the different sources may cover different time periods. The 2007 SILC survey was conducted between March 9 and July 28, 2007, and the socio-demographic data collected (marital status, number of individuals in the household, etc.) refer to the date of the interview. Conversely, income data are in principle from the year 2006. As for tax records, they reflect the situation on December 31, 2006, with total income for the year 2006 serving as the tax base. Individuals having experienced mobility (migration, but also marriage) between December 31, 2006 and the date on which they participated in the survey may not be matched. A second factor involves non-permanent residents (for example, those with temporary work permits) who are subject to payroll deductions (with part of their wages being retained at the source) not appearing in the tax records. Moreover, some individuals were surveyed in their secondary residence, but pay taxes in their principal residence.

## 3. Comparison of economic values and quality tests

### 3.1 Validation tests

Tax records contain economic variables (income, pensions, wealth, etc.) that can be compared with data in the SILC survey, which collects the same variables by means of CATI. Comparisons of the economic variables from the two sources were performed using the raw CATI data, but also reflect corrections incorporated ex post during the data validation phase. The comparison is between the values declared in the survey and those reported in the tax records.

Furthermore, we assessed the extent to which survey participants provide a representative sample of the total population present in the tax records. ${ }^{2}$

The validation tests addressed:

1) the proportion of individuals having declared a positive value for an income source (for example, the proportion of individuals earning employment income, the proportion of individuals receiving a pension);
2) for each individual, the income declared in the survey in comparison with the value reported in the tax records.

Any inconsistencies encountered were almost always verified manually. Thus, for example, if an individual declared income from self-employment to the tax authority but not in the SILC survey, we sought to identify the reason for this inconsistency by examining individual data. On occasion, these manual verifications led to a source of error being identified (for example, two entries in the tax records being transposed). Similarly, discrepancies between the amounts declared in the survey and those reported to the tax authority were systematically analyzed and interpreted.

The following dimensions are common to the two sources: income from employment, income from selfemployment, "first-pillar" benefits (old age, disability, widowhood), "second-pillar" and "third-pillar" benefits (funded occupational pensions), unemployment insurance benefits, income from wealth, and the level of wealth.

### 3.2 Results of the validation

Since we cannot present all the results of the validation here, we will focus on a few dimensions that illustrate the problems that may arise when collecting financial data with surveys.

### 3.2.1 Overrepresentation of self-employment

As shown in Table 2, the proportion of individuals who declared income from self-employment in the telephone interviews is approximately $13.7 \%$ (another $7.2 \%$ were unable to provide their exact status), compared with $10.7 \%$ in the tax records. The latter value is certainly correct because self-employment is subject to specific provisions with regard to taxation, and there are strict controls on the taxpayer's status.

The situation in which some individuals declare income from self-employment, though they are in fact employees, primarily arises if they have the status of employee within their own business or if they exercise an independent activity within a company (for example, as managers) and erroneously consider themselves self-employed. In the survey, the declaration refers to a self-evaluation of their activity rather than to a precise status.

Table 2:
Proportion of individuals with self-employment income, by source

|  | CATI |  |  |  | Tax records |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | No |  |  | Yes |  |  | Don't know |  |  |  |
|  | N | $\%$ | N | $\%$ | N | $\%$ | N | $\%$ | N |  |
| Bern | 75.3 | 502 | 16.0 | 107 | 8.7 | 58 | 86.7 | 578 | 13.3 | 89 |
| St. Gallen | 82.1 | 486 | 10.1 | 60 | 7.8 | 46 | 92.7 | 549 | 7.3 | 43 |
| Vaud | 80.5 | 504 | 14.5 | 91 | 5.0 | 31 | 88.8 | 556 | 11.2 | 70 |
| All cantons | 79.2 | 1,492 | 13.7 | 258 | 7.2 | 135 | 89.3 | 1,683 | 10.7 | 202 |
| Source: 2007 SILC survey and tax records |  |  |  |  |  |  |  |  |  |  |

### 3.2.2 Differences in income from self-employment

For individuals who declare income from self-employment in both sources, Table 3 presents some statistical indicators by source. Owing to tax avoidance, which provides an incentive to the self-employed to reduce their tax burden by deducting some personal expenses, values in the tax records are lower than those declared in the survey.

[^12]Thus, the median income declared by self-employed workers is CHF 48,000 or CHF 4,000 monthly ${ }^{3}$. Their monthly income as reported in the tax records is approximately CHF 400 lower. This is certainly because the amounts received (declared to the survey) were reduced by some personal expenditures (such as automobile expenses, etc.) in order to decrease taxable income.

Table 3:
Distribution of income from self-employment (Swiss francs, CHF)

|  | 1st quartile | Median | 3rd quartile | Mean | N |
| :--- | :---: | :---: | ---: | ---: | ---: |
| SILC | 30,000 | 48,000 | 80,000 | 71,692 | 110 |
| Tax records | 20,501 | 43,356 | 73,151 | 67,394 | 110 |
| Source: 2007 SILC survey and tax records |  |  |  |  |  |

### 3.2.3 Differences in employment income

As Figure 1 suggests, employment income declared in the survey is overall quite close to the corresponding value in the tax records, with a cluster of points closely bunched around the diagonal-which represents equality of the reported values. However, there are some outliers. Those lying below the diagonal represent individuals who declare a higher income in the survey than they do to the tax authority. There are two reasons behind this. On one hand, some individuals (for example, students) do not declare occasional professional income to the tax authority, which does not always detect this omission. On the other hand, a bias of social desirability may motivate some whose income falls below their expectations to inflate the amount they declare to the survey.

Figure 1.
Cross-tabulation of employment income declared to the survey and reported in the tax records


Source: 2007 SILC survey and tax records

[^13]Several points in the graph fall below the diagonal, indicating that the amount reported to the survey is below that declared to the tax authority. This surely means that the individual does not know his or her income level, or that they fail to account for income from a secondary source at the time of the survey.

### 3.2.4 Difficulties reporting the level of wealth

Finally, Figure 2 demonstrates the difficulty encountered by surveyed individuals in correctly declaring their level of wealth. In fact, there is practically no relationship between the value declared in the survey and the estimated level of wealth in the tax records. While survey participants are relatively aware of their income, they struggle to quantify the goods they own and estimate their value. The same observation holds true with regard to income from wealth (result not presented here).

Figure 2.
Cross-tabulation of wealth declared in the survey and reported in the tax records


Source: 2007 SILC survey and tax records

| Valeur registre fiscal | $=$ | Value in tax records |
| :--- | :--- | :--- |
| Valeur cati | $=$ | CATI value |

## 4. Conclusion

This analysis has revealed several somewhat positive elements. On one hand, for some dimensions (such as employment income), the quality of data collected by telephone for the SILC survey may be deemed good. Thus, for each surveyed individual, a significant degree of correspondence is observed between the information in the SILC survey and that reported in the tax records. Moreover, median and mean values from the survey are close to their comprehensive values in the tax records. In contrast, some discrepancies between the two sources suggest that the SILC survey may be more reliable than tax records-for example, in the case of income from self-employment,
which tends to be underestimated in the tax records because of tax avoidance. Conceptual differences may also create inconsistency between the two sources, as in the case of the meaning of the status of self-employed.

However, for other dimensions, a telephone survey presents obvious limitations. In this case, it is useful to seek other paths for collecting information, and the use of complementary sources, such as tax records, may be worth exploring.

## Bibliography

Wanner, P. (2010). Couplage de données de registres administratifs avec l'enquête SILC. Laboratoire de démographie, Université de Genève.

## SESSION 2C

MULTI-MODE DATA COLLECTION

# Using administrative data to find the best medium: Examples of mixed sources and mixed modes 

Jannes Hartkamp and Hans Rutjes ${ }^{1}$


#### Abstract

The use of administrative or register data as a sampling frame for surveys is a common practice in the worlds of both official and non-official statistics. The rise of mixed mode surveys has raised the question of how information from registers can be used optimally to determine what would be the best survey mode or the best combination of options for different groups. Personal characteristics available beforehand can be used to make fairly accurate response estimations for different modes for each group. If, as is the case in some common survey designs, telephone interviews are used as a 'last resort' after approach by post and/or email has failed to yield a response, one needs to estimate in advance for what proportion of a given group a correct postal address, e-mail address and telephone number will be available, not just what the willingness to respond will be among those who can be reached.

Notwithstanding the recent debate about the possible negative effect on response rates offering a choice could have in some cases, on the whole, clever mixed mode designs certainly do optimize response rates. However, exactly what constitutes the best mode mix varies from survey to survey and within a survey from group to group. Moreover, what may have been the best mix three years ago may be no longer the best mix now. In most Western societies, the proportion of the general population for which a telephone number can be found in a publicly available telephone register is steadily decreasing (in the Netherlands, by a few percentage points each year). Internet coverage is still expanding, but now the novelty of web surveys has seen some respondents rather return to paper. E-mail addresses are hard to come by and notoriously volatile. Finally, population characteristics sometimes change quite rapidly in some areas or some sectors. Survey experts often have good reasons for opposing unnecessary changes; however, in survey design, stagnation nowadays soon means decline.


[^14]
# Accounting for the effects of data collection modes in population surveys 

Y. Celia Huang, Mary E. Thompson, Christian Boudreau, and Geoffrey T. Fong ${ }^{1}$


#### Abstract

Increasingly, survey data are being collected by more than one mode. Some surveys use a combination of web (CAWI) and telephone (CATI) collection; others may use a combination of telephone and face-to-face interviewing. For many types of questions, the distribution of response options may be different in the two (or more) mode samples. The difference is partly attributable to selection effects, for example, due to web and telephone respondents being recruited in different ways, or using different frames. Typically, web and telephone samples differ with respect to their distributions of age, sex, education and variables related to personal outlook. Another source of difference in the response option distribution is "technical" in origin, having to do with respondents' tendency to process the options differently depending on whether the options are heard or seen. For example, it is often found that telephone respondents are more likely than web respondents to use the extremes of a Likert scale. The purpose of this presentation is to illustrate an approach to modeling in a mixed-mode survey that takes into account both selection and technical mode effects, using data from the International Tobacco Control (ITC) surveys in the Netherlands and Malaysia. The model uses a propensity score for selection effects, and up to two parameters for technical effects. It can be extended to apply to cross-country comparison data, or to longitudinal data where the collection mode may change from wave to wave.


Key Words: Mixed-mode surveys, Mode effects models, Propensity score.

## 1. Introduction

As obtaining a probability sample from a survey population becomes more and more difficult, survey practitioners are turning increasingly to mixed-mode survey methods (Blyth, 2008). Telephone surveying is ideal for many purposes, because the questionnaire is administered by a person trained to elicit information and to keep the respondent engaged, and because travel costs are eliminated (Roberts, 2007). Face-to-face administration of a survey can have advantages in terms of data quality and higher response rates (Groves, 1979). Self-administered web data collection is particularly attractive because there are no interviewer or data entry costs (Blyth, 2008; Roberts, 2007).

When a survey is conducted by different modes, there may be two types of mode effects. First, what might be called "technical effects" arise from how the mode affects cognitive processing. Respondents in different modes may see or hear the questions and response options differently. For instance, respondents in a face-to-face interview may see flash cards; respondents in a telephone interview hear the questions and response options instead of reading them. Furthermore, whether there is an interviewer present may have some effects. For example, in a web survey, when respondents answer sensitive questions and difficult questions, they may feel freer to express their own thoughts when there is no interviewer. In a telephone survey, when there is an interviewer, respondents tend to give more socially desirable answers (Kreuter et al., 2008), and more recently heard responses (Bishop et al., 1988). They are also more likely to choose the extreme ends of a 5-point Likert scale (Dillman et al., 2009).

The second type of mode effects, namely "selection effects", arise when the sample of respondents in one mode cannot be considered as a random subsample of the whole sample. This happens when respondents in different modes are recruited in different ways, yielding samples differing with respect to demographics (Nagelhout et al., 2010) and other variables. Selection effects can be regarded as encompassing the difference in coverage bias and non-response bias for the two samples.

[^15]This paper will illustrate several methods to account for these two types of mode effects, and apply the methods in population surveys. More comprehensive treatments can be found in Thompson et al. (2010) and Huang (2010).

## 2. Proposed method

### 2.1 Propensity score construction

One method to account for the selection effects is the use of a propensity score, introduced by Rosenbaum and Rubin (1983). Theoretically, this is the probability of responding by a specified mode, given the fact of being in the combined sample, as a function of explanatory variables of interest X , and other covariates W , for example, demographic variables such as age, gender and income. The propensity score is usually estimated using a logistic regression model, regressing an indicator for responding by the specified mode on the covariates X and W , for example, we can estimate propensity for responding by telephone. Since the propensity score quantifies the selection effects, controlling for propensity score in comparing the results from the two sample parts allows us, in principle, to separate technical mode effects from selection effects. That is, if we compare telephone and web respondents with the same propensity score, the average mode differences will not be confounded with the variables X and W , and are therefore more likely to be technical effects.

### 2.2 Latent trait model

To account for technical effects, we propose a latent trait model for survey questions with ordinal responses, where $y$ denotes the coded response:

$$
P(y \leq d \mid u, \text { mode, } x)=\frac{\exp \left(\eta_{d}\right)}{1+\exp \left(\eta_{d}\right)}
$$

where $d$ goes from 1 to $D_{y}-1$, and $D_{y}$ is the number of response option for $y$; the quantity $\eta_{d}$ is given by

$$
\eta_{d}=c_{d}-(\alpha+\beta(d-1)) \times \text { mode }+b_{0} u-B_{1} Z
$$

The parameters $c_{d}$ can be thought of as location parameters for the response option thresholds. The parameters $\alpha$ and $\beta$ are the technical mode effect parameters. For example, if the variable mode takes values 0 and 1 , for web and telephone respectively, $\alpha$ represents the amount by which the telephone mode translates the locations, and $\beta$ represents a factor by which the locations may be spread apart or contracted by the telephone mode. The variable $Z$ is the logit of the individual's estimated propensity to respond by a certain mode. The variable $u$ is a latent variable which is assumed to be $N(0,1)$. Useful references for ordinal response models like the one proposed here include McCullagh and Nelder (1989) and Grilli and Pratesi (2004).

This model for the mode effects can be used to test for the presence of mode effects in the response to certain questions or groups of questions; to estimate the distribution of what an individual's response to one mode might be, given a response in the other mode; or as a component of an ordinal logistic regression model for a response subject to mode effects.

### 2.3 Accounting for mode effects in explanatory variables

In a regression or generalized linear model, it would sometimes be the case that both response and explanatory variables would be subject to mode effects. In such a case, there are at least two possible approaches to modeling. In the first, assuming no selection effects for simplicity, we might consider a kind of structural equation approach, postulate $N(0,1)$ or logistic latent variables $\eta_{y}$ and $\eta_{x}$ and assume:

$$
\eta_{y}=a+b \eta_{x}+\varepsilon
$$

Then

$$
P(y \leq d)=P\left(\eta_{y} \leq c_{y d}-\left[\alpha_{y}+\beta_{y}(d-1)\right] \times \text { mode }\right), \quad d=1, \ldots, D_{y}-1
$$

Similarly,

$$
P(x \leq d)=P\left(\eta_{x} \leq c_{x d}-\left[\alpha_{x}+\beta_{x}(d-1)\right] \times \text { mode }\right), \quad d=1, \ldots, D_{x}-1
$$

The parameters would be $a, b,\left\{c_{y d}\right\},\left\{c_{x d}\right\}, \alpha_{y}, \beta_{y}, \alpha_{x}, \beta_{x}$ and $\sigma_{\varepsilon}^{2}$.

In the second, we consider a functional approach. Suppose it is reasonable to use a latent trait variable for $y$, but to treat $x$ as an observed categorical variable. We might then think of the $x$ values as "correct" for mode 0 , and observed with error for mode 1 . Supposing that $x$ has possible values 1,2 and 3 , the formulation for $y$ could look something like:

$$
P(y \leq d)=P\left(\eta_{y} \leq c_{y d}-\left[\alpha_{y}+\beta_{y}(d-1)\right] \times \operatorname{mode}-\gamma_{1} I\left(x^{*}=1\right)-\gamma_{2} I\left(x^{*}=2\right)\right), \quad d=1, \ldots, D_{y}-1
$$

where $\eta_{y}$ has a standard logistic distribution, and $X^{*}$ is the latent "correct" value of $x$ (under mode 1 ). A model for the "mapping" of $x^{*}$ to the observed $x$ (under mode 1 ) is needed. The parameters of the model would be the parameters for that mapping, along with $\left\{c_{y d}\right\}, \alpha_{y}, \beta_{y}, \gamma_{1}$ and $\gamma_{2}$.

## 3. Applying the method to the ITC Surveys

### 3.1 Description of the ITC Surveys

The proposed models are applied to the International Tobacco Control (ITC) Policy Evaluation Project (ITC Project) surveys. These are longitudinal surveys, mainly of adult smokers, in 20 countries, for the purpose of evaluating policy measures introduced under the Framework Convention on Tobacco Control of the World Health Organization. The data collection modes in the ITC surveys include face-to-face, telephone and web.

### 3.2 Applications and results

In the first application, the latent trait model, as known as an item response theory (IRT) model, is used. In this model, cohort variables are included to examine the time-in-sample effects. The data is from the ITC Malaysia Survey Wave 3 in 2008, including recontact and replenishment. The data collection modes are face-to-face, coded as 1 , for 867 respondents, and telephone, coded as 0 , for 1090 respondents. The response variable is a statement "Malaysia society disapproves of smoking." with five response options: "1-Strongly disagree", "2-Disagree", "3Neither agree nor disagree", "4-Agree" and "5-Strongly agree". The model for this application is given by:

$$
P(y \leq d \mid u, \text { mode, } x)=\frac{\exp \left(\eta_{d}\right)}{1+\exp \left(\eta_{d}\right)}
$$

where $d$ goes from 1 to 4 , and

$$
\eta_{d}=c_{d}-(\alpha+\beta(d-1)) \times \text { mode }+b_{0} u-B_{1} Z-B_{2} \text { Cohort } 1-B_{3} \text { Cohort } 2
$$

The parts for cohort effects are added to the latent trait model shown before in section 2.2 . Cohort 1 is coded as 1 if the respondent was recruited at wave 1 and 0 otherwise. Similarly, Cohort 2 is coded as 1 if the respondent was recruited at wave 2 and 0 otherwise. The application of the model is restricted to areas where both data collection methods were used.

Table 3.2-1
Results for the application of the latent trait model to the statement: "Malaysia society disapproves of smoking."

| Parameter | Estimate | Standard Error | P-value |
| :--- | ---: | ---: | ---: |
| $\boldsymbol{c}_{\mathbf{1}}$ | -2.8078 | 0.1686 |  |
| $\boldsymbol{c}_{\boldsymbol{2}}$ | 0.7122 | 0.1207 |  |
| $\boldsymbol{c}_{\mathbf{3}}$ | 1.3994 | 0.1280 |  |
| $\boldsymbol{c}_{\mathbf{4}}$ | 4.2731 | 0.2513 |  |
| $\boldsymbol{\alpha}$ | 0.2326 | 0.1561 | 0.1363 |
| $\boldsymbol{\beta}$ | -0.2737 | 0.08641 | 0.0016 |
| $\boldsymbol{b}_{\mathbf{0}}$ | 0.000431 | 0.4262 | 0.9992 |
| $\boldsymbol{B}_{\mathbf{1}}$ | -0.1738 | 0.06122 | 0.0046 |
| $\boldsymbol{B}_{\mathbf{2}}$ | 0.5036 | 0.1194 | $<.0001$ |
| $\boldsymbol{B}_{\mathbf{3}}$ | 0.03107 | 0.1413 | 0.8260 |

Looking at the results, $\alpha$ is non-significant, and $\beta$ is significant and negative, which means that face-to-face respondents tend to choose higher agree options for this socially desirable statement. The significant negative $B_{1}$ implies that those with higher telephone propensity have less likelihood of wanting to choose the higher options, and perhaps are less sensitive to social norms. The results suggest that the selection effects and the technical effects are pulling in the same directions. The individual random effect coefficient $b_{0}$ is not significant.

In addition, $\boldsymbol{B}_{\mathbf{2}}$ is significant, indicating a possible time-in-sample effect, but $\boldsymbol{B}_{\mathbf{3}}$ is not significant, with Cohort 3 as the reference level. The higher likelihood of agreement for Cohort 1 may also reflect the policy environment in Malaysia at the time its members were recruited, because there was an anti-smoking quit campaign 'Tak Nak' taking place in Malaysia just before Wave 1, highlighted in the Wave 1 questionnaire.

The second application is a multi-country comparison. The data are from five European countries: the Netherlands, Germany, France, Scotland and the United Kingdom. The mode telephone is coded as 1, and web is coded as 0. Note that only the Netherlands has two modes and other countries all have telephone interviews. Wave 1 data is used for these countries, except for the United Kingdom, where Wave 5 data is used because the United Kingdom Survey started earlier. The response variable is a question: "In the last month, how often, if at all, have you noticed the warning labels on cigarette packages?" The five response options are: "1-Never", "2-Rarely", "3-Sometimes", "4Often" and " 5 -Very often". The model for this application is given by:

$$
P(y \leq d \mid u, \text { mode }, x)=\frac{\exp \left(\eta_{d}\right)}{1+\exp \left(\eta_{d}\right)}
$$

where $d$ goes from 1 to 4 , and

$$
\eta_{d}=c_{d}-(\alpha+\beta(d-1)) \times \operatorname{mode}+b_{0} u-B_{1} Z I_{N e t h}-f(X) B_{2}-C \gamma
$$

In addition to the proposed latent trait model discussed in Section 2, $I_{\text {Neth }}$ is an indicator for the Netherlands, $f(X)$ represents a one-dimensional summary of the demographic variables which is a good predictor for noticing labels, and $C$ is the set of country indicators. Cross-sectional survey weights, scaled to sum to country sample size, have been used in this analysis.

Table 3.2-2
Results for the application of the latent trait model to the question: "In the last month, how often, if at all, have you noticed the warning labels on cigarette packages?"

| Parameter | Estimate | Standard Error | P-value |
| :--- | ---: | ---: | ---: |
| $\boldsymbol{c}_{\mathbf{1}}$ | -4.5416 | 0.5978 |  |
| $\boldsymbol{c}_{\mathbf{2}}$ | 0.9806 | 0.5329 |  |
| $\boldsymbol{c}_{\mathbf{3}}$ | 5.4087 | 0.6142 |  |
| $\boldsymbol{c}_{\mathbf{4}}$ | 11.2396 | 0.8378 |  |
| $\boldsymbol{\alpha}$ | -2.0688 | 0.4524 | $<.0001$ |
| $\boldsymbol{\beta}$ | 2.0010 | 0.1654 | $<.0001$ |
| $\boldsymbol{b}_{\mathbf{0}}$ | 5.6028 | 0.3564 | $<.0001$ |
| $\boldsymbol{B}_{\mathbf{1}}$ | 0.3758 | 0.1606 | 0.0193 |
| $\boldsymbol{B}_{\mathbf{2}}$ | 1.3139 | 0.2285 | $<.0001$ |
| France vs. Germany | 3.1517 | 0.2995 | $<.0001$ |
| Scotland vs. Germany | 3.8546 | 0.4183 | $<.0001$ |
| Rest of UK vs. Germany | 3.2357 | 0.3052 | $<.0001$ |
| Netherlands vs. Germany | 0.7063 | 0.4910 | 0.1504 |

The results show that $\alpha$ is negative and significant, and $\beta$ is positive and significant, consistent with a tendency of telephone respondents to show more extreme responses (Nagelhout et al., 2008). A significant and positive $B_{1}$ suggests that the higher the propensity to respond by telephone in the Netherlands, the higher the label salience. The individual random effect coefficient $b_{0}$ is significant for this weighted analysis.

For the country effects, with Germany as the reference level, label salience is the greatest for Scotland, and about the same for the United Kingdom and France. It is significantly lower in Germany and the Netherlands, but not significantly different between the Netherlands and Germany. The country effects follow the same pattern as seen in the analysis of Hitchman et al. (2010). In a different analysis that treats Netherlands telephone and Netherlands web as two separate countries, and the model contains no terms for the technical or selection mode effects, as a result, Netherlands telephone is not significantly different from Germany, while Netherlands web is significantly different. Therefore, when we account for the mode effects, we see the differences between the Netherlands and Germany diminishing.

## 4. Conclusion

In conclusion, the results illustrate that the modeling approach presented in this paper can describe technical mode effects, at least for five-scale questions. Furthermore, the models can distinguish technical effects from selection effects associated with collection mode by adding a propensity term. Finally, the models allow both response variable and explanatory variable to be subject to technical mode effects.

## Acknowledgements

Core funding for the ITC project is provided by the U.S. National Cancer Institute to the Roswell Park Cancer Center (P01 CA138389), and by the Canadian Institutes of Health Research. Major funding for the ITC Netherlands Survey
is provided by ZonMw. Major funding for the ITC Malaysia Survey Wave 3 was provided by the U.S. National Cancer Institute (P50 CA111236), the Canadian Institutes of Health Research (79551) and the Malaysian Ministry of Health. We acknowledge our colleagues Bas van den Putte (University of Amsterdam), Gera E. Nagelhout, and Marc. C. Willemsen (Maastricht University/CAPHRI, STIVORO for a smoke free future) on the ITC Netherlands team and Ron Borland (Cancer Council Victoria), Rahmat Awang, and Maizurah Omar (National Poison Centre Universiti Sains Malaysia) on the ITC Malaysia team.

## References

Bishop, G.F., Hippler, H.J., Schwarz, N. and Strack, F. (1988). A comparison of response effects in self-administered and telephone surveys in R. M. Groves et al. (eds,) Telephone Survey Methodology, New York: John Wiley.

Blyth, B. (2008). Mixed modes: the only 'fitness' regime? International Journal of Market Research, 50:241-266.
Dillman, D., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. and Messer, B.L. (2009). Response rate and measurement differences in mixed-mode surveys: using mail, telephone, interactive voice response and the internet. Social Science Research, 38(1), pp. 1-18.

Grilli, L. and Pratesi, M. (2004). Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. Survey Methodology, 30, pp. 93-103.

Groves, R. (1979). Actors and questions in telephone and personal interview surveys. Public Opinion Quarterly, 43, pp. 190-205.

Hitchman, S. et al (TBA). (2010). Effectiveness of the standardized European Union text-only warning labels among smokers from four countries: Findings from the International Tobacco Control Policy Evaluation Project. In progress.

Huang, Y.-C. (2010). Accounting for the effects of mode in the ITC Malaysia survey. unpublished report, Waterloo, Canada: the International Tobacco Control Policy Evaluation Project.

Kreuter, F., Stanley, P. and Roger, T. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. Public Opinion Quarterly 2008, 72, pp. 847-865.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, Second Edition. Chapman and Hall.
Nagelhout, G.E., Willemsen, M.C., Thompson, M.E., Fong, G.T., van den Putte, B. and de Vries, H. (2008). Mixed modes in the International Tobacco Control (ITC) Netherlands Survey: Differences between CAWI and CATI on smoking-related questions. Joint Conference of SRNT and SRNT-Europe, April 27-30, 2009, Dublin, Ireland.

Nagelhout, G.E., Willemsen, M.C., Thompson, M.E., Fong, G.T., van den Putte, B. and de Vries, H. (2010). Is web interviewing a good alternative to telephone interviewing? Findings from the International Tobacco Control (ITC) Netherlands Survey. BMC Public Health, 10, pp. 351.

Roberts, C. (2007). Mixing modes of data collection in surveys: A methodological review. NCRM Methods Review Papers. Volume NCRM/008 London: Centre for Comparative Social Surveys, City University.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), pp. 41-55.

Thompson, M.E., Huang, Y.C., Boudreau, C., Fong, G.T., van den Putte, B., Nagelhout, G. and Willemsen M.C. (2010). Accounting for the Effects of Data Collection Method with Application to the ITC Netherlands Survey. In progress.

SESSION 3A

## MICROSIMULATIONS

# How survey data and the LifePaths microsimulation model can interact to produce policy-relevant results 

Jacques Légaré, Yann Décarie, Patrick Charbonneau, Samuel Vézina and Janice Keefe ${ }^{1}$


#### Abstract

The purpose of this paper is to show, on the basis of our research experience, how data from Statistics Canada surveys such as the General Social Survey (GSS) and the health surveys can be beneficial when combined with LifePaths projections to produce useful results for the home care and services industry in Canada. To attain its objectives, our research program needs population projections based on many different characteristics for which conventional projection models are inadequate; we therefore have to use more elaborate models such as microsimulation models. We chose Statistics Canada's LifePaths model because it generally satisfied our objectives. However, some variables that we need are still not in LifePaths, such as living arrangements. To address this problem, we first performed a series of regressions using data from the General Social Survey and other sources. The parameters obtained with the regressions were then combined with the LifePaths projections. We illustrate our application with findings from a study of future senior citizens who are the most vulnerable in terms of health and living arrangements. This study is based on projections up until 2031 and focuses in particular on the Canadian population aged 75 and over with a disability, with no surviving spouse or children, and living alone. A comparison between Canada and a few European countries shows the extent to which their specific demographic histories will have different effects on the size of the population we consider the most vulnerable.


Keywords: Microsimulation; LifePaths; General Social Survey (GSS); Canada; population policies.

## 1. Introduction

The purpose of this paper is to show, on the basis of our research experience, how data from Statistics Canada surveys such as the General Social Survey (GSS) and the health surveys can be beneficial when combined with LifePaths projections to produce useful projections for home care and services for senior citizens in Canada.

To attain its objectives, our research program needs population projections based on many different characteristics for which conventional projection models are inadequate; we therefore have to use more elaborate models such as microsimulation models. We chose Statistics Canada's LifePaths model because it generally satisfied our objectives. However, some variables that we need are still not in LifePaths, such as living arrangements. To address this problem, we first performed a series of regressions using data from the General Social Survey and other sources. The parameters obtained with the regressions were then combined with the LifePaths projections (Carrière et al., 2007; Légaré et al., 2011).

With this application, we were able to conduct a study of the most vulnerable future senior citizens in terms of health and living arrangements. This study is based on projections up until 2031 and focuses in particular on the Canadian population aged 75 and over who live in private households, have a disability, have no surviving spouse or children, and live alone. A comparison between Canada and a few European countries shows the extent to which their specific demographic histories will have different effects on the size of the population we consider the most vulnerable.

[^16]
## 2. Methods

## The microsimulation model

LifePaths is a dynamic longitudinal individual and family microsimulation model developed by Statistics Canada over the past few years. It was written in the Modgen programming language. For each simulation, the overlapping cohort model produces a representative sample of the Canadian population. The oldest cohort represented in LifePaths was born in 1872. As a result, the model can produce a complete, representative dataset for all ages for 1971, the year of the first modern census as far as sociodemographic variables are concerned (Statistics Canada, 2004).

The life course of an individual, known as a case, is simulated in LifePaths as a series of events that occur in continuous time (and therefore not artificially restricted to annual intervals) using behavioural equations based on various microdata sources. LifePaths generates statistically representative samples of complete individual life courses. The behaviour equations produce, at a subannual resolution, discrete events that together constitute the life cycle of an individual.

By virtue of its structure, LifePaths uses many different sources of historical data and databases from current Statistics Canada surveys, such as vital statistics, socioeconomic surveys, health surveys, labour force surveys, and surveys of finances. Throughout the simulation, LifePaths uses information from these sources to continually update the list of outstanding events to ensure that the next event is the one with the shortest waiting time. This approach provides a straightforward solution to the problem of concurrent events. Waiting times provide a unified structure for decision making. Probabilistic decisions can be introduced so that the choice among various alternatives is determined by a comparison between two or more waiting times.

What's more, a LifePaths simulation produces a set of mutually independent cases. Each case contains exactly one dominant first-generation individual. The dominant individual's spouse and children are also simulated as an integral part of the case, using marriage and fertility rate equations.

Using the LifePaths microsimulation model, we projected the Canadian population up until 2031. The microsimulated variables are as follows: age, sex, education, region, marital status, spouse's age, place of birth, number of surviving children and disability (Figure 2.1). It is also important to keep in mind that LifePaths is a dynamic model, which means that variables that are not in the above list may affect the results (Légaré and Décarie, 2010).

Monte Carlo variation is another important concept in the model. All cases simulated in the model use waiting times that include a stochastic component. This is one of the reasons why LifePaths can reproduce the diversity observed in actual populations. On the other hand, this variability affects the reliability of the results in the tables in the same way as aggregate results based on a small sample drawn from a large population. To avoid this type of problem, particularly for rare events, we simulated over 6 million cases for our specific projections.

## Combining the data with cross-sectional data

The projections were complemented by the application of parameters derived from data from Statistics Canada's 2002 General Social Survey.

The derived variables are: the need for assistance, living arrangements, assistance received, and the source of the assistance received.

Clearly, the derived variables are associated with many of the variables included in LifePaths. For example, living arrangements depends on the previously determined "need for assistance" variable and LifePaths variables such as age, marital status and level of education.

Moreover, living arrangements is an important variable to determine the source of assistance received along with many LifePaths variables, such as region and number of surviving children (Figure 2.1).

Figure 2.1:
Variables projected by various methods


## 3. Application

### 3.1 The research question

Based on two health scenarios, how can we estimate the number of elderly people who will be in a very vulnerable situation in the future in terms of health, family context and living arrangements?

### 3.2 The health scenarios

The projections were based on two health scenarios:

- a basic scenario in which we observe both life expectancy gains and increases in life expectancy with and without disability;
- a very optimistic health scenario in which all the additional years of life expectancy are lived in good health (i.e. without disability).


### 3.3 Summary of the research results

The results of applying the methods described above project the size of the most vulnerable elderly population (i.e. people who may have a very high likelihood of being institutionalized). The Canadian results are compared with the data for Europe (Légaré et al., 2011).

The family situation of an elderly person with disabilities can be assigned to one of four mutually exclusive categories and ranked hierarchically from most to least favourable in terms of potential helpers. The elderly person has:

- a spouse and at least one surviving child
- a spouse but no surviving children
- no spouse but at least one surviving child
- no spouse and no surviving children

Projecting a population based on such characteristics requires only data from LifePaths (Gaymu et al., 2009).
We are interested in the population 75 years and over, since age 75 represents a turning point in the probabilities of widowhood and dependence, and we define the most vulnerable persons as those who are 75 or over, have a disability, have no spouse or surviving children and live alone.

To estimate the number of persons who live alone, all the variables are from LifePaths (age, surviving children, etc.) except the living arrangement information that comes from the GSS, which depends on the LifePaths variables and the "need for assistance" variable, which is also derived from the GSS.

If we project the lives of elderly Canadians who have disabilities and no surviving spouse between 2001 and 2031, we find more men than women in a state of vulnerability and little difference based on age, except in the 85 -and-over group, where there is a slight decline. There is relatively little change in the distribution across the three states-not alone, alone with or without surviving children-between 2001 and 2031.

Even though the proportion of vulnerable persons varies little over time, with the exception of a slight decrease in the 85 -and-over group, the absolute numbers rise exceptionally fast: they more than double for both men and women and for both the very elderly and the younger elderly. Despite the substantial percentage decline for women 85 years and over, the number of vulnerable women in this age group is expected to jump from about 25,000 to 40,000 .

When we compare Canada with Europe, we find that in the population aged 75 and over with disabilities in both Europe and Canada, four men and seven women out of 10 have no partner.

The increase in the number of vulnerable elderly people is very large, except in the case of women. In Canada, the increase is $187 \%$ for men and $99 \%$ for women; while in Europe, it is $112 \%$ for men and a mere $4 \%$ for women.

When we analyze the situation on a country-by-country basis, we find that the increases in some European countries are comparable to those in Canada: for men, $222 \%$ in the Netherlands, $304 \%$ in Finland and $204 \%$ in Germany. Among women, only Finnish women ( $65 \%$ ) have a similar, though smaller increase compared with Canadian women; while in some countries (Czech Republic, Portugal and Germany), the number of vulnerable elderly women is expected to decline.

It is also worth noting that even in the most optimistic (and unrealistic) health scenario, the number of the most vulnerable persons would grow substantially in Canada for both men and women; and in Europe for men after 2020. However, this health scenario has more positive effects in European countries than in Canada.

In short, the fact is that the baby boom is having a much greater impact in Canada than in Europe.

## 4. Conclusion

This paper provides a good example of how combining Statistics Canada data from various sources can shed light on the impact that will ensue when Canadian baby-boomers become senior citizens, and accordingly assist in the development of appropriate policies.

## Acknowledgements

Our research program is currently funded by the Canadian Institutes of Health Research (Reference No. 81108). The authors would also like to express their sincere gratitude to their colleagues who, in the initial stages of the research program, contributed to the model of combining various Statistics Canada data sources: primarily Yves Carrière and Geoff Rowe, but also Laurent Martel, L. Xiaofen, S. Rajbhandary, X. Lin, Glenda Hawkins, Marc-Antoine Busque and Guillaume Lefrancois.

## References

Carrière, Y., Keefe, J., Légaré, J., Lin, X., and Rowe, G. (2007). Population Aging and Immediate Family Composition: Implications for Future Home Care Services. Genus, vol. LXIII, n ${ }^{0}$ 1-2, pp. 11-31.

Gaymu, J., Busque, M.A., Légaré, J., Décarie, Y., Vézina, S., and Keefe, J. (2010). What will Family Composition of Older Persons Be Like Tomorrow? A Comparison of Canada and France. Canadian Journal of Aging, SEDAP Special Issue, 29 (1), p. 1-15.

Légaré, J., and Décarie, Y. (2010). Using Statistics Canada LifePaths Microsimulation Model to Project the Health Status of Canadian Elderly. International Journal of Microsimulation (à paraitre).

Légaré, J., Décarie, Y., Charbonneau, P., and Keefe, J. (2011). Parmi les futures personnes âgées canadiennes vivant dans la communauté, estimation de celles qui seront les plus vulnérables en terme de santé et de mode de vie. La Revue canadienne du vieillissementl Canadian Journal of Aging, SEDAP Special Issue, (en évaluation).

Statistics Canada (2004). LifePaths. (http://statcan.ca/english/spsd/LifePaths.htm)

# Using multiple data sources to project U.S. Social Security Finances and the Federal Budget in the long erm 

Jonathan A. Schwabish ${ }^{1}$


#### Abstract

The United States Congressional Budget Office's Long-Term (CBOLT) microsimulation model projects individual emographic and economic behavior of the U.S. population to assess the financial picture of the U.S. Social Security system and the rest of the U.S. federal government. For each individual in the model, CBOLT simulates a wide range of demographic and economic characteristics, including birth, death, immigration and emigration, labor or force participation, earnings, and Social Security taxes paid and benefits received. The core individual-level data used in CBOLT come from administrative data provided by the U.S. Social Security Administration and to augment those records, additional information is taken from a variety of other administrative and survey data.


Key words: microsimulation, Social Security

## 1. Description

The Congressional Budget Office's long-term model, CBOLT, is used to estimate the long-term finances of the United States federal budget and Social Security system. CBOLT is a microsimulation model of the U.S. population, economy, and federal budget, and is used to analyze potential reforms to federal entitlement programs and quantify the nation's long-term fiscal challenges. To estimate costs of Social Security and other long-term programs on the U.S. economy and federal budget, the CBOLT model uses information from a variety of different data sources. ${ }^{2}$ This article provides a brief overview of the Congressional Budget Office (see Box 1), CBO's long-term microsimulation model, and the data sources on which that model is built.

In general, a microsimulation model starts with individual-level data from a representative sample of the population and projects demographic and economic outcomes for that sample through time. In the case of the CBOLT model, for each individual in the sample the model simulates birth, death, immigration and emigration, marital pairings and transitions, fertility, labor force participation, hours worked, earnings, payroll taxes, Social Security benefit claiming, and Social Security benefit levels. A complex actuarial framework wraps around the microsimulation model to provide totals for demographic and economic variables as well as additional information in areas where the microsimulation model has not yet been developed. The model projects individual demographic and economic behavior of the population, the finances of the Social Security system, and the finances of the rest of the federal government more than 75 years into the future. In recent work CBO has added detail on Medicare, Medicaid, and other health care spending to the actuarial framework. CBOLT also includes a macroeconomic model that analyzes the federal sector's role in the larger economy and a repeated-simulation (Monte Carlo) mode that quantifies uncertainty about a variety of outcomes.

[^17]Box 1. The Congressional Budget Office
The Congressional Budget Office (CBO) is a federal agency within the legislative branch of the United States government. Founded in 1974, the agency is charged with two main tasks:

1. Provide the United States Congress with objective, nonpartisan, and timely analyses to aid in economic and budgetary decisions on the wide array of programs covered by the federal budget; and
2. Provide the United States Congress the information and estimates required for the Congressional budget process.
In fiscal year 2009, CBO issued 33 studies and reports, 9 briefs, 11 Monthly Budget Reviews, 38 letters, 8 presentations, and 5 background papers. CBO also testified before the United States Congress 17 times on a variety of issues. In calendar year 2009, CBO completed approximately 480 federal cost estimates as well as about 420 estimates of the impact of legislation on state and local governments, including the identification of any unfunded mandates contained in such legislation, and about 420 estimates of the impact of anv unfunded mandates on the private sector. The CBO currentlv emplovs

## 2. Input Data

The core individual-level data used in CBOLT come from the Continuous Work History Sample (CWHS), an administrative data set provided by the U.S. Social Security Administration (SSA). The data are drawn from a collection of administrative data sets: the Summary Earnings Record, the Detailed Earnings Record, the Numerical Identification System, and the Master Beneficiary Record. ${ }^{3}$ These data contain a history of individual earnings records for a 1 percent random sample of Social Security numbers, beginning in 1951. The data also contain demographic information and Social Security information for each individual. The data on Social Security-for OldAge, Survivors and Disability Insurance (OASDI) -include claiming dates, type of claim (OAI, SI, or DI), the primary insurance amount, the monthly benefit amount, and the reason for disability.

Using administrative data rather than data from surveys has substantial benefits. Administrative earnings records provide a consistent measure of earnings for individuals over many years. In addition, the sample sizes are large: The core data file for CBOLT contains information on approximately 300,000 individuals. (CBO uses a 10 percent sample of the CWHS file in order to facilitate processing the data file.) And because the earnings data are based on administrative records, they are not subject to survey respondents' errors in recalling information correctly or to issues of rounding or nonresponse. ${ }^{1}$

The CWHS also has a number of limitations. The CWHS data capture earnings only from workers in the covered sector-that is, workers who contribute to the Social Security system. ${ }^{2}$ The CWHS does not include earnings of workers who do not have or do not report a valid Social Security number or whose earnings are received "under the table." Perhaps most important, the data contain only limited demographic information: sex; Social Security beneficiary status; date of birth; and, if applicable, date of death. The CWHS also lacks dates of death for many deceased people who had Social Security numbers but did not claim Social Security benefits.

Because the CWHS data include only earnings, age, sex, and OASDI benefit information, CBO uses data from other sources to expand the CWHS record (see Table). Additional demographic and economic data come from other data sets such as the Survey of Income and Program Participation (SIPP), the Current Population Survey (CPS), and the
${ }^{3}$ Here the term CWHS refers to the sampling framework used by SSA to extract data from those sources. The CWHS is a 1 percent samle of Social Security numbers; the 10 percent sample of the CWHS file used by CBO results in a $1-i n-1,000$ random sample of Social Security nubmers. In principle, such a sample should be representative of the U.S.population, although events like death and immigration do not appear to be fully recorded in the data. To adjust the data for some of those potential shortcomings, CBO extensively reweights the CWHS to make sure that the sample is in fact representative along several dimensions, such as age, labor force status, and Social Security beneficiary status.
${ }^{1}$ See, for example, Cristia and Schwabish (2009).
${ }^{2}$ In 1985, 93 percent of paid civilian workers were in the covered sector; by 2002, that share had risen to 96 percent. See House Committee on Ways and Means (2004).

Health and Retirement Survey, all of which can be matched to other administrative data sources from the Social Security Administration. Additional health expenditure data come from the Medical Expenditure Panel Survey (MEPS). Information from those data sources is not merged directly to the individual record in CBOLT. Rather, CBO uses these sources to estimate statistical relationships on the basis of individual characteristics that already exist in CBOLT. Those estimates provide the basis for probabilistically imputing the data that are not available in the CWHS.

Table 1. Data Sources for the CBOLT Model

| Data Type | Number of Individuals (millions) | Periodicit <br> y | Purpose in CBOLT* |
| :---: | :---: | :---: | :---: |
| Administrative Data |  |  |  |
| Continuous Work History Sample | about 3.5 | Annual | administrative earnings, 1978-2007 |
| Detailed Earnings Records | about 3.5 | Annual | administrative earnings, 1978-2007 |
| Summary Earnings Records | about 4.5 | Annual | administrative earnings, 1951-2007 place of birth, date of birth, date of |
| Numerical Identification System | about 4.5 | Annual | death |
| Master Beneficiary Record | about 1.5 | Annual | Social Security benefit information (type of award, claim date, etc.) |
| Survey Data |  |  |  |
| Health and Retirement Survey** | about 22,000 | Bi-Annual | retirement claiming behavior |
| Medical Expenditure Panel Survey | about 30,000 | Annual | health expenditures |
| Survey of Income and Program Participation** | about 100,000 | Monthly | mortality, health status, marital transitions |
| Current Population Survey** | about 100,000 | Annual | earnings, labor force participation, hours worked, unemployment, educational attainment, fertility |
| * Descriptions are not exhaustive. |  |  |  |

CBOLT also incorporates a wide range of aggregate data. The model is calibrated to match certain projections of the United States' Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds. CBOLT matches their projections of the total population by age, sex, and marital status as well as mortality and fertility rates, immigration and emigration counts, and disability incidence and termination rates for each year of the projection period. For the first 10 years of the projections, CBOLT uses economic projections and projections of federal outlays and revenues produced by various other divisions within CBO. To make the transition from the first 10 years to the long-term projection period, CBOLT phases in certain long-term assumptions over a 5to 10 -year period.

## 3. Example: Assigning Labor Force Participation Rates in CBOLT

The method for projecting labor-related economic outcomes for the sample in CBOLT is comprised of separate modules for labor force participation, full-time or part-time employment status, number of hours worked, presence of an unemployment spell, and an idiosyncratic earnings component that helps determine annual earnings. ${ }^{3}$ Equations for each component are designed to generate realistic patterns of behavior in the sample and to operate as part of an overall framework that includes macroeconomic and aggregate budgetary outcomes. Assigning each characteristic to each individual in CBOLT involves both a deterministic, equation-based component and some degree of randomness. For example, the labor force participation equation may generate a 60 percent chance that a particular

[^18]individual will enter the labor force. A model-generated random number (between 0 and 100 percent) for that individual is then compared with that probability to determine whether the person will enter the labor force (in this example, if the random number is less than 60 percent, the individual enters the labor market). Given a large enough sample, the fraction of the population assigned a particular outcome will match the average probability of that outcome.

In the first step of the labor market portion of CBOLT, the decision to enter the paid labor force is predicted for each person using binominal logit regressions with the following set of covariates: age, education level, in-school status, marital status, Social Security beneficiary status, number of children under age 6 (for women), and lagged labor force participation, along with effects of time (trend) and birth cohort. ${ }^{4}$ Separate equations for men and women are estimated using 30 years of pooled data from the Current Population Survey, and are based on an age-centering approach that allows flexibility in the relationship between labor force participation and the underlying determinants across age groups (see Sabelhaus and Walker, 2009). Observed strong persistence in labor force participation behavior is captured directly through the lagged labor force term, an effect that grows with age and is uniformly larger for women than for men.

The equations for predicting labor force participation in CBOLT can be characterized as simply describing patterns of labor force participation across demographic groups and time, as opposed to generating predictions based on economic behavior. Economists often focus on how policy affects labor supply behavior; many models predict behavior on the basis of utility-maximizing principles. The labor force participation equations in CBOLT are not derived from an explicit optimizing model that would include, for example, the effect that tax policy has on the trade-off between leisure and consumption. This is in large part because the primary goal of the microsimulation is to generate a rich data set with realistic demographic and economic heterogeneity. Utility-maximizing models place significant constraints on the extent to which that richness can be generated within the model.

The outcomes of the labor force participation rate (LFPR) regression models from the pooled CPS data and the resulting imputations to the CBOLT model all make sense in the context of historical trends:

- LFPRs are very high, especially for prime-age men, and tend to decline with age;
- Social Security benefit receipt is negatively correlated with LFPR;
- Educational attainment is generally positively correlated with LFPR;
- For women, having children younger than age 6 is negatively correlated with LFPR;
- Last year's LFPR is largely positively correlated with this year's LFPR;
- For men, there is great stability in the predicted distribution of years worked, with fewer than 40 percent working less than 36 years.


## 4. Modeling Uncertainty

To investigate the degree of uncertainty surrounding specific Social Security projections, CBOLT can be run repeatedly, using input assumptions drawn from various distributions, to create probability distributions of outcomes around the central projections (CBO, 2005). Such procedures, known as Monte Carlo simulations, are especially important for long-term projections because uncertainty compounds over time. Moreover, certain types of reforms can increase or decrease uncertainty significantly.

Eleven inputs are varied in the standard Monte Carlo runs. ${ }^{5}$ For runs that include private investments, returns on corporate bonds and equities can also be varied. The model is run hundreds of times, each time with an independently drawn set of input assumptions. The probability distributions of the inputs are based on time-series analyses of historical patterns. In particular, the model assumes that future uncertainty can be described on the basis

[^19]of variation observed in the past. Of course, the distribution is sensitive to the chosen historical time period and the selected timeseries equations.

## 5. Summary

CBO's long-term microsimulation model provides the United States Congress with a valuable resource with which to gauge the long-term fiscal challenges faced by the country. Among its many responsibilities, CBO publishes annual reports documenting these long-term challenges, including its Long-Term Budget Outlook (CBO, 2010c) and the Long-Term Projections for Social Security (CBO, 2010d). CBO has also offered in-depth analysis of a variety of different potential Social Security policy reforms, including proposals Peter Diamond and Peter Orszag (CBO, 2004), Jeffrey Leibman, Maya MacGuiness, and Andrew Samwick (CBO, 2006a), Senator Robert Bennett (CBO, 2006b), and Congressman Paul Ryan (CBO, 2010a), as well as an analysis of 30 different approaches to changing the system (CBO, 2010c).

## References

Congressional Budget Office. (2004). Long-Term Analysis of the Diamond-Orszag Social Security Plan. (December).

Congressional Budget Office. (2005). Quantifying Uncertainty in the Analysis of Long-Term Social Security Projections, Background Paper (November).

Congressional Budget Office. (2006). a. Long-Term Analysis of the Liebman-MacGuineas-Samwick Social Security Proposal. (February 8).

Congressional Budget Office. (2006). b. Long-Term Analysis of S. 2427, the Sustainable Solvency First for Social Security Act of 2006. (April 5).

Congressional Budget Office. (2006). c. Projecting Labor Force Participation and Earnings in CBO's Long-term Microsimulation Model. CBO Background Paper (October).

Congressional Budget Office. (2009). CBO's Long-Term Model: An Overview. CBO Background Paper (June).

Congressional Budget Office. (2010). a. An Analysis of the Roadmap for America's Future Act of 2010. (January 27).

Congressional Budget Office. (2010). b. The Long-Term Budget Outlook. (June).
Congressional Budget Office. (2010). c. Social Security Policy Options. (July).
Congressional Budget Office. (2010). d. CBO's 2010 Long-Term Projections for Social Security: Additional Information. (October).

Cristia, J. and Schwabish. J.A. (2009). Measurement Error in the SIPP: Evidence from Matched Administrative Records, Journal of Economic and Social Measurement, Vol. 34, No. 1.

House Committee on Ways and Means. (2004). 2004 Green Book: Background Material and Data on the Programs Within the Jurisdiction of the Committee on Ways and Means, WMCP: 108-6 (March), Table 1-7 and pp. 1-4.

Sabelhaus, J. and Walker, L. (2009). Econometric Flexibility in Microsimulation: An Age-Centered Regression Approach. International Journal of Microsimulation, Vol. 2, No. 2: pp. 1-14.

# Integrating infectious and chronic disease information to model cervical cancer 

Michael Wolfson ${ }^{1}$


#### Abstract

There is now clear evidence that Human Papilloma Virus (HPV) is causal for cervical cancer. As a result, there is considerable interest in the optimal strategy for screening not only for cervical cancer, but also for HPV infection, in conjunction with the best strategy for HPV immunization. A very powerful method for answering this set of questions is microsimulation modeling, where the simulation model serves as the platform for integrating essential data from diverse sources. Indeed, such data integration requires the construction of a simulation model of some sort.

However, modeling in this case is challenging analytically because the usual approaches to modeling infectious and chronic diseases are quite different. For infectious diseases, it is best to model contacts and transmission of the pathogen between individuals explicitly. This entails a microsimulation model with co-evolving agents. For chronic disease incidence, progression and case fatality, on the other hand, inter-individual interactions are not important. Rather, the emphasis is more appropriately on survival times and sojourn times in various disease states, and their implications for health care utilization.

This paper will describe current work on a model of HPV infection and cervical cancer, being undertaken on behalf of the Canadian Partnership Against Cancer, that not only integrates data from a wide variety of sources, but also involves the construction of two distinct submodels and their integration. The first sub-model simulates the spread of HPV and its control by screening and vaccination, while the second sub-model draws on the outputs of the first and then models cervical cancer incidence, progression and treatment.


[^20]
## SESSION 3B

## SAMPLING FRAMES 1

# Building the Dwelling Frame in Gauteng Province using a combination of statistical methods namely: Census, administrative data and survey data. 

Mahlape Mohale and Harebatho Moletsane ${ }^{1}$


#### Abstract

Statistics South Africa is developing a comprehensive source of information on dwelling units namely the Dwelling Frame (DF), which forms the benchmark for statistical planning in a country. Under stable human settlements conditions the task would not be insurmountable yet in Gauteng Province, which is characterized by a fast growing population, a high rate of informal settlements and a hype of economic activities, the DF project presented serious challenges. The 2001 Census had high undercount and therefore incomplete information on housing, the quality of the administrative data on housing is poor, incomplete and often outdated and the data collected directly from household $s$ in the DF project was incomplete and of poor quality. Lessons learned in collecting data from the three sources and the analysis thereof provided valuable information for future statistical planning such as Census 2011. The contributing factors to poor quality, incomplete data, high data-collection costs and under count from each data source were identified. A review and modification in collection methods were made, which included using cheaper mode of transport, using several teams to sweep through an Enumeration Area (EA) instead of each team per EA, mapping the province and systematically identifying and annotating all the structures in the province, modifying data collection method to the requirements of each of the four settlement types: farms, traditional, formal and informal settlements. By using an integrated data collection method the quality of DF information in Gauteng Province was greatly improved and more efficient data collections techniques were developed.


The aim of the study is to assess the suitability of statistical information obtained from previous censuses, administrative records and Dwelling Frame surveys in building a Dwelling Frame for Gauteng province, which is crucial for:
a) Improving the Dwelling Frame (DF) information, which entails:
i. Identification of most appropriate data collection methods for each settlement type,
ii. Documenting improvements in the collection of DF information from DF survey projects: data collection instruments, procedures and processes.
iii. Annotation on the Enumeration Area (EA) maps indicating place names, EA size, structures, routes to and within EAs.
iv. Informing and adjusting new EA boundaries (demarcation) for Census 2011 and ensure that an EA does not straddle across two or more place names.
v. Updating Dwelling Units (DU) lists as per Census 2001 (EAs) demarcation paying special attention to EA boundaries, addresses, type of DU, number and sequence of DUs, feature classification such as dwelling units, units/rooms within collective living quarters, sports fields, parks, cemeteries, offices, schools, community halls, demolished and semi-demolished structures, filling stations, vacant land /stands.

[^21]b) Development of an integrated program for housing data collection.
c) Establishment of relationships \& partnerships with the main role players.
d) Determination of most appropriate timing and sequencing of data collection activities prior to Census 2011 taking.
e) Identification of most cost effective methods for collection of housing statistics.

Background: Gauteng Province is the smallest of the nine provinces in South Africa accounting for $1.5 \%$ of South Africa's surface area. However it has the largest share of the population at 11.1 million people ( $22 \%$ ) and a high rate of mushrooming informal settlements.

In terms of international obligations and Statistics Act No. 6 of 1999, Statistics South Africa is compelled to develop and maintain registers and lists which may be of use in producing statistics. Currently information on housing in the province is fragmented, incomplete and expensive to collect and maintain due to high rate of urbanization. The resources (human, financial and equipment) are insufficient and the diminishing financial resources against competing interests call for efficacy in the management of statistics. The legal framework for the development and maintenance of the Dwelling Frame (Housing Census) information is inadequate, there is lack of coordination of statistical information on housing and the datasets are not integrated.

There are deficiencies in existing dwelling/housing datasets obtained from previous censuses, Geo-Referenced Dwelling Frame Project and supplementary data obtained from administrative records of external producers. Advanced technologies, which are now able to provide greater coverage and large-scale datasets at fine levels of geographic details, are not gainfully applied. Technological applications by producers of geographic information have been erratic and unsynchronized.

The policymakers are increasingly faced with demands to make decisions at small area levels. There is therefore a dire need for statistical information on human settlements in a congruent form for planning, monitoring and evaluation of the national development program.

Furthermore South Africa will be conducting Census in 2011 and one of the aims is to reduce the under count to below $2 \%$. The two previous censuses done in 1996, 2001 and a Community Survey in 2007 were characterized by high undercounts of $10 \%, 17 \% \& 22 \%$ respectively. A good quality housing database is the bedrock of censuses and surveys hence the need to establish the DF for Gauteng province.

In response to the above mentioned challenges The Geo-Referenced Dwelling Frame (DF) project (2005-2009) was initiated in order to develop a comprehensive dataset on housing, using data from previous Censuses and administrative records as part of the building blocks, however the DF was not completed due to lack of funds. There were problems in DF design, processes and data collection methods; as a result information gap on housing data still exists. Subsequently a rapid test on usefulness of the DF data for Census2011 listing and other application as well as identification of cost effective methods for collection of housing statistics was commissioned in 2009. The 2009 test was conducted on four settlement; formal, informal, traditional and farm types in Gauteng province informs this research paper.

## 1. Methodology

Sampling: A purposive selection of EAs from four different settlement types in Tshwane and Johannesburg Metropolitan Municipalities namely Formal, informal, traditional and farms (see figure 1 below).

Data collection methods: Multidisciplinary project team made up of Gauteng management \& Fieldwork team, Geography and Census divisions collected data from identified areas. The Fieldwork entailed validation and verification of information from previous censuses, Geo-Referenced DF project, and administrative records from
different sources based on availability. A desk-top analysis of the data including aerial photography/ satellite imagery was done to identify data sources, age, key variables, quality dimensions, coverage and ability to integrate within a defined EA boundary. The key variables in each dataset were orientation in place, place name, street name, cadastre number, DU number, specific DU unit identifier, EA boundary, EA size, listing route, name of head of household, GPS coordinates, feature classification, which identifies the main land use. The data quality dimensions considered were relevance, coherence, integrity, accessibility, timeliness, accuracy (completeness and comprehensiveness), and methodological soundness and interpretability.

Data analysis: A qualitative data analysis method was applied to all data sets obtained on the four settlement types using criteria mentioned in table 1 below. Quantitatively analysis was not done due to small sample size. Other aspects evaluated included coordination, collaboration, integration, cost-efficiency and use of technological advances in statistical production.

Figure1: Settlement types


Settlement types: 1A and 1B are racially based formal settlements, 2A \& 2B are informal settlements, $\mathbf{3}$ is a typical traditional settlement and $\mathbf{4}$ is a Farm settlement.

Results: All data sources met the following quality dimensions: Methodological soundness, Relevance, Coherence and Interpretability. Variations existed regarding other quality dimensions. See table 1 below.

Table 1: Key criteria used to evaluate datasets

| Data Sources | Type of <br> data | Key variables <br> and age of <br> data | Quality <br> dimensions | Cost of <br> production | Coverage |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Stats SA | Previous <br> Census | DUs points, <br> Maps, <br> Demarcated <br> EAs (2001) | Comprehensive <br>  <br> Outdated, easily <br> Accessible, <br> Relevant | High | $83 \%$ of SA, <br> $83 \%$ of <br> Gauteng |
|  | Geo- <br> referenced <br> DF project | DUs points <br> 2005-2009 | Incomplete, <br> Outdated, <br> Relevant, easily <br> Accessible | High cost | F0\% of <br> Gauteng |
| Municipalities | Administrativ | Place <br> polygons, <br> Address <br> points, aerial <br> photographs <br> $>3 y e a r s ~ o l d ~$ | Incomplete, <br> Outdated, <br> Relevant, not <br> easily Accessible | Low cost | Excludes <br> Informal |
|  | Administrativ | Place names, <br> Address <br> points, current | Incomplete, <br> Relevant, <br> Accessible | Low cost | Settlements |

## 2. Analysis of results

Coordination and integration of statistical production amongst key producers was lacking. None of the datasets were as yet complete to meet the requirements of Census 2011 listing. Gaps existed in all data sets. The data collection is unsystematic, fragmented, and in most cases outdated. Some of the data producers were uncooperative and the levels of data accessibility varied from easily accessible in the case of Statistics South Africa's data and the private sector, who sold the information to the hard to get data mainly from municipalities and Survey General. Formal data sharing agreements existed only with the Post Office. Other producers provided the information on request.

Application of Statistical support systems, especially technology: there were variations in the use of information technology. Some data were manually produced and stored on paper-based systems while other had a mixture of manual and electronic systems. The private sector used latest technologies and maintained their data well. Poor deployment of new technology by most government departments often caused delays in data accessibility and integration thereof. Legacy systems were prevalent in municipalities and Survey General's Office.

Quality: Data production is not standardized, which may compromise quality. Inadequate data quality assurance and lack of monitoring systems pose major challenges in data integration. Despite the limitations mentioned above Censuses data are found to be comprehensive and form a good baseline for survey data, however they get outdated quickly. Administrative records are largely incomplete and outdated. And survey data have too many errors and gaps.

Settlement types: Data on all four settlement types were incomplete even after integration of the census, administrative records and surveys. Main changes observed across different datasets were on number of EAs, EA size, feature classification, and number of DUs. Data on farms, traditional and formal settlements had fewer changes compared to informal settlements, which were found to be unstable and the developments unpredictable thereby requiring frequent updates of the information. The high rate of urbanization in Gauteng province stretches resources for maintenance of housing information and compels different data collection approaches for each settlement type. Data from informal settlements would only be useful if collected closer to the use e.g. census listing.

Cost efficiency: Escalation of collection costs were observed amongst all data producers. Censuses data are comprehensive and a good baseline for surveys, however the collection and maintenance costs are high. Survey data on the other hand are very expensive to collect, while administrative records are the most cost effective \& sustainable sources of housing data.

Other factors: Census, administrative records and survey data complement each other. Although the study did not quantify the results and the significance of linkages, it provided indicative information regarding the chances of success when using the three data sources are to building a Dwelling Frame in the province. Lack of trust among producers of housing statistics is still high and there is no enabling legal framework to enforce new changes.

## 3. Conclusion

A combination of all three data sources is necessary for development and maintenance of a good Dwelling Frame. There is a need for established of a platform for the development of an integrated program for housing statistics in Gauteng, in order to building trust among other data producers, reduce the escalating cost, implement the necessary infrastructure and networks. Coordination of all role players in the most cost effective manner would result in sustainable Dwelling Frame. Sequencing and harmonization of data production to support Census 2011 is crucial.

Statistics South Africa as the lead agency must develop a new legal framework to facilitate coordination and harmonized of data production by all role players, develop statistical capacity amongst other role players for production of quality information, promote usage of administrative records as main source of data and reduce dependency on surveys, and create a conducive environment that would result in a change in mind set from survey/Census-based DF to a statistical system that is based largely on administrative records. Thus, setting up a tiered geographical information system and database.

Municipalities are well positioned for collection and maintenance of housing data. However the municipalities need to be empowered in order to discharge this responsibility and would need to change their operating systems from manual/semi-manual processes to more efficient user-friendly technologies, which would facilitate assess to external data and metadata through tiered client/server environments. Lastly, quantitative analysis of the data should be done.

## References

The United Nations Handbook of Statistical Organization, $3^{\text {rd }}$ edition, The Operations and Organization of a Statistical Agency.

The United Nations Principles and Recommendations for Population and Housing Censuses. Revision 2. (2006).

The United Nations publication Handbook on Geographic Information System and Digital Mapping for Population and Housing Censuses. (2000).

South Africa, Statistics Act No. 6 of 1999.
PARIS21 A Guide to Designing a National Strategy for the Development of Statistics (November 2004).
Statistics South Africa, Geo-Referenced Dwelling Frame Project charter (2005).
Statistics South Africa, Dwelling Frame Field Test Report, (2009).
Freedman, M., Lane, J. and Roemer, M. New approaches to creating Data for Economic Geographers. Journal of Official Statistics, Vol. 24, No.1, 2008.

Lehtonen, R. and Sarndal, C.E. Research and Development in Official Statistics and Scientifc Cooperation with Universities: A Follow-Up Study. Journal of Official Statistics, Vol.25. No.4, 2009.

# Improving and using the Statistical Administrative Records System 

Leah B. Marshall and Amy B. O'Hara ${ }^{1,2}$


#### Abstract

The U.S. Census Bureau uses an integrated system of administrative records to improve censuses and surveys. We will describe the demographic side of the system, containing data on persons and households, and our plans to improve the data system. We will also describe planned indirect and direct uses of the data. The administrative record research program, established over 15 years ago, benefits analysts developing survey frames and questions and is used to evaluate survey responses. Successful projects include interagency research on public health insurance programs and tax credits for the working poor using linked survey and administrative data. A new large-scale evaluation is the 2010 Census Match Study, which examines person and housing unit coverage of administrative data. All projects rely on the availability of quality source data and record linkage techniques. We will discuss the challenges of acquiring data from federal, state, and private providers. Our assessments of source data timeliness and completeness will also be described. We will provide an overview of our research and production record linkage techniques, noting how our expertise has benefitted other federal agencies. The Census Bureau is poised to assist other agencies, particularly other statistical agencies, with data for program evaluation.


Key Words: Administrative Records; U.S. Census; Record Linkage

## 1. Overview of the Statistical Administrative Records System

### 1.1 Overview

The U.S. Census Bureau's Statistical Administrative Records System (StARS) is a production system created by merging administrative data received from different federal agencies in the U.S. The first StARS database was created in 1999 in response to the 1990 Census Undercount, consistent with the mandate from U.S. Code Title 13 instructing the U.S. Census Bureau to use administrative records data to reduce costs and minimize respondent burden (Farber and Leggieri, 2002). Originally designed to investigate administrative records use for the Decennial Census, StARS has also proven useful as a research tool for other surveys and demographic programs at the U.S. Census Bureau.

### 1.2 Source Files and Processing

The StARS database is created annually using information from the sources given in Table 1.2-1. Each source agency prepares an extract close to April $1^{\text {st }}$ of the given year - for instance, StARS 2009 was based on files cut near April 1, 2009. Most of the files are delivered in May or June of the given year, although tax data from the Internal Revenue Service (IRS) are received much later.

[^22]Table 1.2-1 Sources for StARS

|  | Federal Agency | File Received | What the File Contains |
| :--- | :--- | :--- | :--- |
| 1. | Internal Revenue Service <br> (IRS) | Individual Income Tax Returns | Limited data for individuals who file a tax <br> return (Form 1040). |
| 2. | IRS | Individual Income Information <br> Returns | Name, address, and Taxpayer <br> Identification Number for individuals <br> from financial institutions and employers. |
| 3. | Housing and Urban <br> Development (HUD) | Tenant Rental Assistance <br> Certification System | Persons receiving rental assistance and <br> other assisted housing programs through <br> HUD. |
| 4. | HUD | Public and Indian Housing <br> Information Center | Persons participating in HUD's Public <br> and Indian Housing programs for rental <br> assistance. |
| 5. | Centers for Medicare and <br> Medicaid Services (CMS) | Medicare Enrollment Database | Individuals enrolled in Medicare. |
| 6. | Indian Health Services | Patient Registration System | Registrants in the Indian Health Services <br> system. |
| 7. | U.S. Selective Services | Selective Service System <br> Registration | Registrants with U.S. Selective Services <br> to be eligible for the draft. |
| 8. | Social Security <br> Administration (SSA) | Numerical Identification System <br> (Numident) | Name, date of birth, and Social Security <br> Number (SSN) for persons assigned an <br> SSN by the SSA. |

The source files from IRS, HUD, CMS, Indian Health Services, and U.S. Selective Services are combined to create the StARS database after undergoing extensive address and person processing. This processing includes a validation step to ensure the incoming files contain accurate and complete information before the records are input into StARS. Specifically, the Numident file from SSA is not used as an input into the StARS database but rather is used to help validate the other incoming files.

Before being input into StARS, all the addresses received on the seven input files are matched to the Census Bureau's Decennial Master Address File to pick up the Decennial Master Address File Identifier. The addresses are standardized, and geographic information about the addresses (county codes, block codes, etc) is then retrieved. Finally, the list of addresses is unduplicated to create one file of administrative records addresses called the Master Housing File. Note that although the Decennial Master Address File Identifier is retrieved during address processing, addresses on the Decennial Master Address File may or may not be on the Master Housing File and vice versa, depending upon whether the address existed on one of the seven StARS source files.

Before being input into StARS all the person records received on the seven input files are matched using combinations of name, address, date of birth, and SSN to a reference file based on SSA data from the Numident. Those records that can be validated against the Numident then get assigned a unique identifier called a Protected Identification Key (PIK). This use of the PIK allows SSN and name data to be stripped off of the source files, helping to mitigate concerns in the U.S. about security of the files and privacy of the information. Those records with validated PIKs are then combined to create the Linked Person File. This Linked Person File contains multiple records per validated PIK, common when a record for a given person appears in multiple source files. Records that cannot be validated and assigned a PIK are dropped at this point in the processing and are not included on the Linked Person File. Table 1.2-2 contains an example to make this processing more explicit; in this instance both records 1 and 3 would be included on the Linked Person File but record 2 would not. The Composite Person Record File is
then created by unduplicating the Linked Person File so that there is just one record per validated PIK. Again following the example in Table 1.2-2, records 1 and 3 would be unduplicated so that this particular "John Smith" appears in only one record on the Composite Person Record File. The Person Characteristic File is then created by appending demographic information from other sources onto the Composite Person Record File, such as date of death, modeled race, and modeled ethnicity. To associate each person with an address, the Initial Comprehensive Address File is created. This file contains all the addresses on the administrative records sources for a person. Again following the example given in Table 1.2-2, the Initial Comprehensive Address File would contain the address " 123 Main St" associated with John Smith and "321 First St" associated with John Smith, but the address " 10 Oak St" would not be on the file associated with John Smith.

Table 1.2-2 StARS Validation Example

|  | Name | Date of Birth | SSN | Address | Source File | Validated PIK |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 1. | John D. Smith | $1 / 1 / 90$ | $123-45-6789$ | 123 Main St | IRS 1040 | 999999999 |
| 2. | J. Smith | $? ? / ? ? / 90$ | $122-44-6789$ | 10 Oak St | HUD Tenant Rental <br> Assistance <br> Certification System | None assigned |
| 3. | J.D. Smith | $1 / 1 / 90$ | $123-45-6789$ | 321 First St | U.S. Selective Service | 999999999 |

The creation of these several different files as part of StARS processing allows for different approaches to examine the data, making StARS quite useful as a research tool for surveys and demographic programs at the U.S. Census Bureau.

### 1.3 Validation Results

Each of the seven input files has a very high validation rate, with validation results for StARS 2009 processing ranging from $96.7 \%$ to $99.5 \%$. This means that the U.S. Census Bureau was able to validate and put PIKs on $96.7 \%$ to $99.5 \%$ of the person records received from the source agencies for StARS 2009. The validation rate is believed to be high because StARS source files come from federal databases with SSNs included on many of the records. The federal agencies administer various programs that are dependent upon applicants and users providing accurate information; often applicants will not receive benefits or be given credit for compliance without providing correct SSN, name, and date of birth. Comparatively, commercial data evaluated by the U.S. Census Bureau have validation rates closer to $50 \%$. Commercial data are often compiled from many different places - for example, telephone directories, utility bills, magazine subscriptions, and the like. The information often does not include SSN, and the information provided by a user or customer need not be as accurate or complete to receive services; for instance, a customer will still take home the sweater he just purchased even if he provides no date of birth or an inaccurate zip code to the store. It is for this reason that StARS has historically been composed of federal inputs, though research into uses of commercial data is ongoing.

After all processing and validation steps were complete, the StARS 2009 database contained approximately 300.5 million unique persons. An additional 17.3 million records were included in processing but eventually dropped from StARS 2009 because the SSN could not be located or confirmed. Of these 17.3 million records, approximately $46 \%$ of them had a valid Individual Tax Identification Number from the IRS. It should be noted that because of the reliance of PIK upon SSN during the validation stage each unique person should appear only once in the StARS database.

## 2. Improvement Needs for StARS

Even with such high validation rates for the source files, the StARS database could always benefit from improvements. For example, the StARS database could benefit from improvements related to the data acquisition process and the record linkage algorithms used.

## a. Data Acquisition Improvements

The first improvement regarding the data acquisition process for StARS involves more timely file deliveries. The biggest limitation is on the receipt date of tax data from the IRS. As previously mentioned, StARS is based on files cut near April $1^{\text {st }}$ of the given year with most of the source files delivered in May or June. The IRS tax data, however, are received much later, with the final delivery in January of the year following the intended reference year. To clarify this timing consider StARS 2009 as an example: IRS individual income tax return data for tax year 2008 were due for most of the American public on April $15^{\text {th }}$, 2009. Despite the fact that the income data are for 2008 - the previous calendar year in relation to StARS 2009 - the name, address, and SSN data are from the time of filing. In other words, the name, address, and SSN data have a reference date near April $15^{\text {th }}, 2009$ for the majority of Americans, and hence these data are quite useful for building the StARS 2009 database. Data from IRS Information Returns data is received annually in October - so for StARS 2009, the 2008 tax year Information Returns were received in October of 2009. Data from IRS tax returns for tax year 2008 were received in several installments, the last of which was delivered to the U.S. Census Bureau in January 2010. So the final delivery of IRS tax year 2008 data to be used in creating the StARS 2009 database was not received until January of 2010. Any successful acquisition improvements which could hasten this process, even slightly, would be a vast improvement.

The second improvement need related to the data acquisition process for StARS involves improvements in the negotiation processes. The data received for StARS are subject to negotiations with IRS, HUD, CMS, Indian Health Services, Selective Services, and SSA. Currently the U.S. Census Bureau does not have input into the type of data collected. The data collection is done by the source agency for the source agency's own needs and not for the needs of the U.S. Census Bureau. For instance, a source agency may have no need to collect a current address in order to administer the agency's program, or the source agency may have no need to store first and last name in different variable fields. The source agency also establishes which data fields to provide the U.S. Census Bureau, and for which records to provide the data. As part of the negotiation process each source agency considers their unique perspective on their users' privacy concerns, as well as processing and file transfer constraints for large or complex datasets. Again, any positive interactions and efforts in the negotiation process would be of great benefit to StARS.

Finally, perhaps the most obvious data acquisition improvement need is for additional agreements with new agencies to obtain administrative records sources that may enhance the coverage of StARS. Exploration into data from state and other local governmental sources could be especially beneficial to StARS by filling in coverage gaps in the federal data from the sources listed in Table 1.2-1. For example data from school meal programs, student data systems, or the Children's Health Insurance Program in the U.S. could provide increased coverage of children in StARS. Any increased coverage of the StARS database at the U.S. Census Bureau would be of great benefit to the surveys and demographic programs that utilize this data.

## b. Record Linkage Improvements

Another set of enhancements to benefit StARS are enhancements to the file processing and record linkage algorithms used to create the database. As each file is received by the U.S. Census Bureau, it is put through a set of standardization procedures to format the data into a useable configuration for processing - for example, parsing out name fields and address fields as necessary and checking for missing or invalid values. Not only can these standardization algorithms be improved upon, especially as new information becomes available about the data, but the standardization algorithms must be improved upon to deal with changes in data over time. For instance, the increased use of hyphenated last names must be researched and dealt with appropriately in name standardization algorithms. Continual improvements to the standardization procedures for the input data are necessary to ensure quality in the annual build of the StARS database.

As discussed in section 1.2, each record on each administrative records source file is first validated and assigned a PIK before being added into StARS. This validation is done using data from SSA's Numident file as a reference; hence validation of a record depends upon the existence of an SSN for a given person. There are people in the U.S., however, who do not have SSNs. Methods of validating these people need to be investigated further. The most promising research involves using the IRS's Individual Tax Identification Number - which is similar in format to SSN but assigned only to persons in the U.S. who do not have an SSN - to validate records. Additionally, research
into whether corroborating records in multiple sources could permit assignment of an alternative unique linking identifier may also prove useful (Marshall and O'Hara 2010).

Another reason records fail the validation stage of StARS processing is due to missing or inaccurate information; hence, research into new modules and blocking passes during the StARS validation process could also prove valuable. The current modules employed in the validation system are basic, using essentially name, address, and date of birth in conjunction with SSN to validate records. Additional modules - for example a module which utilizes the knowledge of a relationship between two persons - could be researched to increase validation rates. The current blocking passes employed in the validation system are designed to sort the data before matching in order to cut down on computer processing time. Research into different ways of blocking the data could also result in the validation of more records during this point of the processing. The validation of more records - though not at the cost of the quality of the validation process - may result in a more complete StARS database.

## 3. Uses of the Statistical Administrative Records System

## a. Evaluations and Sampling Frames

The StARS database as well as other sets of administrative records data obtained by the U.S. Census Bureau have shown to be useful research tools for several surveys and programs. Administrative records data can be used as an evaluative tool for surveys. For example, the U.S. Census Bureau's Current Population Survey Annual Social Economic Supplement measures Medicaid coverage in the U.S. Medicaid is a government healthcare program in the U.S. for certain eligible persons. Using administrative records data received from CMS, not only were false-negative reporting rates established, but correlations between respondent characteristics and the Medicaid undercount were established (Klerman et al., 2009).

Another use of administrative records is to establish a sampling frame for surveys, especially surveys which are targeted at a specific portion of the population. For instance, the U.S. Centers for Disease Control and Prevention conducts a National Immunization Survey for children in the U.S. between the ages of one and three years old. The U.S. Census Bureau recently worked with the Centers for Disease Control and Prevention on redesigning parts of the survey but needed a sample of addresses with children in the correct age range with whom to conduct the survey. Rather than using Random Digit Dialing, in person canvassing, or other expensive and/or largely unsuccessful methods to find addresses with children in the correct age-range, the U.S. Census Bureau was able to use commercial administrative records data to construct a sampling frame of addresses (Piani and Stringer 2010). This method of using administrative records data to construct a sampling frame has proven not only useful, but a cost-effective tool to utilize in survey research.

## b. 2010 Census Match Study

The next major research area for the StARS database and other administrative records data is assessing administrative records uses to supplement, conduct, and/or otherwise improve the U.S. Decennial Census. To this end, the U.S. Census Bureau is conducting the 2010 Census Match Study, the purpose of which is to compare the accuracy and coverage of administrative records data to the 2010 U.S. Decennial Census.

The 2010 Census Match Study includes an investigation of new data sources, looking at both federal and commercial sources, and the Census Bureau intends to acquire as many of these sources as possible to fill in existing coverage gaps of StARS. Commercial data may fill in many of the coverage gaps left by utilizing only federal sources. It is hypothesized that individuals who are "off the federal grid" for whatever reason may be contained in commercial files which are often marketing data compiled using utility bills, store customer inventories, real estate records, and the like. After researching other available data sources, the U.S. Census Bureau will undergo negotiations and acquisition processes to obtain the data. Finally, after new data sources have been acquired as part of the 2010 Census Match Study, a new and improved administrative records database will be created for the study. This database will contain StARS at its core, but with the addition of more files than currently in StARS.

The 2010 Census Match Study will provide basic and detailed comparisons of administrative records and 2010 Decennial Census data. A basic count of persons and addresses in the U.S. as determined by the new administrative records database and as determined by the Decennial Census will be provided as part of the research. Then, detailed comparisons will be performed to determine whether the administrative records database and the 2010 Decennial Census actually contain the same addresses, people, and overall demographic characteristics of the people. For instance, while the overall total counts may be similar it is possible and indeed expected that administrative data will contain people who are not included in the 2010 Decennial Census and that 2010 Decennial Census will contain people who are not included in administrative data. The 2010 Census Match Study research seeks to quantify and describe the differences between the newly-created administrative records database and the 2010 Decennial Census.

## References

CMS Medicare Enrollment Database (2010). Type of Data: EDB Customized State File. CMS Website [http://www.cms.gov/PrivProtectedData/08\ _EDBCustomizedStateFile.asp](http://www.cms.gov/PrivProtectedData/08%5C%20_EDBCustomizedStateFile.asp), accessed November 23, 2010.

Farber, J. and Leggieri, C. (2002). Building and Validating a National Administrative Records Database for the United States. New Zealand Conference on Database Integration. U..S. .Census Bureau, March 21, 2002.

HUD TRACS (2010). Additional TRACS Information. HUD Website [http://www.hud.gov/offices/hsg/mfh/trx/trxdesc.cfm](http://www.hud.gov/offices/hsg/mfh/trx/trxdesc.cfm), accessed November 23, 2010.

IRS Information Returns (2010). A Guide to Information Returns. IRS Website [http://www.irs.gov/efile/article/0,,id=98114,00.html](http://www.irs.gov/efile/article/0,,id=98114,00.html), accessed November 23, 2010.

Klerman, J.A., Davern, M., Call, K.T., Lynch, V. and Ringel, J.D. (2009). Understanding the Current Population Survey's Insurance Estimates and the Medicaid 'Undercount'. Health Affairs, November 2009 28, pp. w991-w1001.

Marshall, L. and O’Hara, A. (2010). 2010 Census Evaluations, Experiments and Assessments Study Plan 2010 Census Match Study. U.S. Census Bureau, Draft November 3, 2010.

Piani, A. and Stringer, C. (2010). National Immunization Survey - Evaluation Study Interim Report. U.S. Census Bureau, March 12, 2010.

# Use of the Canada Child Tax Benefit as a frame for the Survey of Young Canadians 

Martin Pantel ${ }^{1}$


#### Abstract

The Canada Child Tax Benefit (CCTB) is a non-taxable amount paid monthly to help eligible families with the cost of raising children under 18 years of age. The Canada Revenue Agency (CRA) manages this program and maintains a file identifying all applicants to the CCTB. Interest in this file as a potential sampling frame for the Survey of Young Canadians (SYC) has grown as few other options are available for this difficult to target population. Given Statistics Canada's limited experience with the CCTB data, two concerns with the use of this file were potential coverage issues (for example, undercoverage of higher income families) and the quality of the contact and auxiliary information. The coverage aspect of the ССТВ will be presented first where population and subpopulation totals from the CCTB and the Canadian Census of Population are compared. Distributions for variables common to both sources are also compared. Then, using probabilistic linkages, profiles of Census families that could not be linked to the CCTB will be drawn. The quality of contact and auxiliary information was assessed via a SYC Contact Test that was carried out in February 2010 with a sample of 1000 children drawn from the CCTB. The presentation will discuss the design, implementation and results of both the probabilistic linkage and the contact test. In light of the encouraging results, final recommendations on the use of the CCTB as a frame for the SYC will be made.


[^23]
# Targeting populations of interest in the sample design using the Census of Population: an example from the Ontario Survey on the Prevalence of Hypertension 

Lori Stratychuk and Jean Dumais ${ }^{1}$


#### Abstract

When designing a survey, there are two traditional options to target specific populations. The first option is two-phase sampling, which tends to be expensive. The other option is to enrich the frame using external administrative data. For the Ontario Survey on the Prevalence of Hypertension (OSPH), the latter method was used. One of the main objectives of the OSPH was to look at the relationship between both the prevalence and correlates of hypertension, and ethnicity. Given the rarity of the ethnic groups of interest, special attention was necessary in the sample design in order to allow comparison of the ethnic groups of interest to the rest of the population. A multi-stage sampling plan was created using traditional area methods. When possible, ethnic strata were created at the second stage of stratification using the information from the most recent Census of Population. One of the challenges in the sample design was creating rules to classify the ethnicity of the primary sampling unit, in this case, the Census Dissemination Area. In order to augment the ethnic diversity of the final sample, the ethnic strata of interest were then over-sampled. Ultimately, the sampling plan succeeded at enhancing the ethnic representation of the sample.


Key Words: Hypertension, Ethnicity, Oversampling, Rare Populations

## 1. Introduction

### 1.1 Description

The Statistical Consultation Group (SCG) at Statistics Canada was engaged by the Heart and Stroke Foundation of Ontario (HSFO) in order to assist the University of Ottawa Heart Institute (OHI) in the creation of a survey to examine both the prevalence and factors related to hypertension in the province of Ontario ${ }^{2}$. In addition to the common challenges in designing a completely new survey, the Ontario Survey on the Prevalence of Hypertension (OSPH) had the additional constraint that the clients desired to investigate the relationship between ethnic origin and both the prevalence and determinates of hypertension.

### 1.2 Ethnic origin

Studies in other countries have confirmed a relationship between ethnic origin and the prevalence, as well as correlates, of hypertension (Colin-Bell, 2002). The three ethnic groups of interest, as determined from previous studies, include the "Black" ethnic group ${ }^{3}$ (i.e., those whose ancestors are from African or Caribbean countries), the "South Asian" ethnic group (i.e., those whose ancestors are from India, Pakistan, Sri Lanka, or Bangladesh), and the "East Asian" ethnic group (i.e., those whose ancestors are from China, Japan, Korea or Thailand). The residual, comprised mostly of Caucasian individuals, constitutes the category of the "Rest." (Figure 1.2-1)

[^24]Figure 1.2-1
Ethnic Origin and Country of Origin


### 1.3 Issues with Ethnic Origin

According to the 2006 Census of Population, the minority of individuals in Ontario reported belonging to any of the ethnic groups of interest. Four percent declared themselves as falling into the Black ethnic group, 5\% declared themselves as East Asian and 7\% declared themselves to be South Asian. Should any form of simple random sample be selected, it is expected that relatively few individuals of the desired ethnic origins would appear in the sample. Given the interest in these groups expressed by the survey clients, it became apparent that special measures needed to be taken in the creation of the sample design in order to enrich the representation of the ethnic groups in the sample.

## 2. Methodological Approach

### 2.1 Oversampling rare populations

There are a multitude of methods that exist in order to study rare populations in social surveys (Kalton, 2009). Two methods that were examined but dismissed include two-phase sampling and unequal probability sampling within households. Two-phase sampling, also known as double sampling, is the process of taking a large initial sample and then only following up with the elements of the populations of interest. This works best when the first stage collection costs are low, and for this reason could not be used since the initial contact costs were quite high. Numerous surveys have used unequal probability sampling in order to oversample, for example, households with young children. However the high correlation between the ethnic backgrounds of individuals within the same household made the additional complexity of unequal probability sampling within households an inadvisable choice.

Based on the available resources, it was decided the most appropriate method to enhance the sample with the ethnic groups of interest was through frame enrichment. Essentially, the method of frame enrichment allows the targeting of certain populations of interest by targeting the neighbourhoods that these populations are known to live in.

The publicly available information from the Census of Population, specifically the Profile of all levels of geography in Canada provides detailed information on an array of variables down to the level of Dissemination Areas (DAs). DAs are the lowest level of geography for which information is published, are usually a few city blocks, and tend to contain between 400 and 700 people. Given the stability of neighbourhood composition, this information was exploited to target the DAs known to be rich in the ethnic groups of interest as of the last Census of Population.

### 2.2 Other Considerations in the Sample Design

Due to the numerous distractions that can occur in a respondent's home, it was necessary for respondents to attend a clinic to have their blood pressure and other anthropometric information recorded. Various clinics could be set up across Ontario, however in the interest of costs it was necessary for the sample to be clustered by geography.

Given that blood pressure was being measured, it was necessary to have trained health professionals to collect the data. Additionally, to maximize the response rate, the initial contact with individuals had to be through personal interview.

Another cost consideration was to exclude the far northern region (approximately north of Sudbury), which contains nearly half of the land mass of Ontario but less than $5 \%$ of the entire population of Ontario.

### 2.3 Sample Design

The OSPH sample design involved a multi-stage complex sampling plan based on traditional area frame methods, with the small variation of frame enrichment using the information from the most recent Census of Population.

A final sample size of 3,000 respondents was desired, where the response rate was expected to be below $50 \%$. Therefore, an initial sample of roughly 7,000 individuals was targeted.

### 3.2.1 First stage of sample selection

The first stage of sampling involved the selection of Census Subdivisions ${ }^{4}$ (CSDs) with probability proportional to size (Cochran, 1977). Some CSDs were selected with certainty for a variety of considerations: Ottawa was selected in order to serve as a pilot for the survey; Toronto was selected due to the large population and ethnic diversity; and finally, the CSD of Sudbury was selected as per the client's specific request. Another 10 CSDs were randomly selected with probability proportional to size: Barrie, Brampton, Guelph, Lakeshore, Markham, Mississauga, Niagara-on-the-Lake, Oshawa, South Stormont and Stratford.

Given the importance of Toronto, it was decided to oversample it. Roughly a quarter of the population of Ontario resides in Toronto, however slightly more than a third of the entire sample (2500) was allocated to Toronto. The rest of the sample was allocated to the other CSDs proportional to their population.

### 3.2.2 Second stage of sample selection

Before proceeding with the second stage of sample selection, it was necessary to classify all the Dissemination Areas (DAs) within the selected CSDs as being Black, South Asian, East Asian, or Rest. After trying a few variations in order to classify the DAs, it was determined that the best method to classify DAs was to base it on the ethnic group that had the highest proportion in that DA. For example, a DA with an ethnic distribution of: $10 \%$ Black, $40 \%$ South Asian, $39 \%$ East Asian and $11 \%$ from the Rest would then be classified as a South Asian DA. One of the unanticipated advantages of this classification scheme was that the DAs rich in one of the three minority ethnic groups of interest would often be rich in one or two of the other ethnic groups of interest.

Using, this definition, it was possible to stratify the DAs by ethnicity for the CSDs of Brampton, Markham, Mississauga and Toronto. The other 9 CSDs were left with no ethnic stratification, where individuals from the ethnic groups of interest would be found, on average, at the same rate as they exist in each of the selected areas.

The number of DAs to be selected in a given CSD was chosen in order to have a sampling fraction of roughly 5\%, which is a compromise between minimizing collection costs and minimizing the impact of intra-class correlation. For CSDs with ethnic stratification, the allocation to each of the ethnic strata was a compromise between a

[^25]proportional and an equal allocation. (Table 3.2.2-1) Using a proportional allocation would be more optimal for estimates for all of Ontario. A completely equal allocation between the ethnic strata would be more optimal for the domain estimates (of each ethnic group). Using the compromise between the equal and proportional allocation was an attempt to balance the objectives of producing estimates for Ontario, as well as estimates for each of the ethnic groups.

Table 3.2.1-1
Sample allocation by CSD

| CD | CSD | Name of CSD | Number of DAs <br> in Sample | Total <br> sample for <br> CSD | Sample size <br> for each DA <br> in CSD |
| :---: | :---: | :--- | :---: | :---: | :---: |
| 43 | 042 | Barrie | 8 | 300 | 38 |
| 21 | 010 | Brampton | 22 | 800 | 36 |
| 23 | 008 | Guelph | 10 | 300 | 30 |
| 37 | 064 | Lakeshore | 4 | 100 | 25 |
| 19 | 036 | Markham | 16 | 600 | 38 |
| 21 | 005 | Mississauga | 34 | 1100 | 32 |
| 26 | 047 | Niagara-on-the-Lake | 4 | 100 | 25 |
| 18 | 013 | Oshawa | 14 | 300 | 21 |
| 06 | 008 | Ottawa | 16 | 400 | 25 |
| 01 | 011 | South Stormont | 2 | 100 | 50 |
| 31 | 011 | Stratford | 4 | 100 | 25 |
| 53 | 005 | Sudbury | 16 | 500 | 31 |
| 20 | 005 | Toronto | 78 | 2500 | 32 |

### 3.2.3 Third stage of sample selection

Using the standard area frame techniques, all dwellings in the selected DAs were listed shortly prior to interviewing. Once there was a list of all the dwellings in the selected DA, dwellings would be chosen using a systematic sample.

### 3.2.4 Fourth and final stage of sample selection

From within each of the selected dwellings, one in-scope adult (aged 16-79) would be selected from all of the household members using the commonly used birthday method, which simply selects the individual whose birthday has occurred most recently.

### 3.2.5 Data Collection

OHI interviewers would first visit the selected dwellings and select the appropriate individual within the dwelling. They would then collect some background information, and then book an appointment for the individual to come to a nearby clinic in order to complete the more detailed questionnaire and also have their blood pressure and other anthropometric information measured and collected.

### 3.2.6 Weighting and Estimation

Design weights were created using the initial selection probabilities at each stage. Final weights were calculated by calibrating to the most recent Census of Population using sex, age group and ethnic group. In order to allow the design-based analysis of the data, both by the SCG and eventually by external researchers, mean bootstrap weights (Yung, 1997) were created and appended to the data file.

## 4. Results

### 4.1 Response Rates

Ultimately, forty-six percent of all in-scope individuals completed the background questionnaire. Of those who responded to the initial questionnaire, a large majority ( $89 \%$ ) attended the blood pressure clinic and provided their blood pressure and other anthropometric information, such as height, weight and waist circumference.

The somewhat low response rates were due in part to the difficulties encountered in making the initial contact. However, monetary incentives to compensate individuals for their travel time appeared to have some positive influence on clinic attendance.

Some initiatives which helped to improve the response rates included: media promotions prior to collection, the availability of translators for the applicable languages (such as Mandarin) and also having the questionnaires translated into various languages. Other initiatives to improve the response rates, which are based on common practices in Statistics Canada, included specific follow-up procedures and scheduling of call-backs over different time slices ${ }^{5}$.

### 4.2 Distribution by ethnic groups

The "ideal" ethnic distribution, from an analytical but somewhat naive perspective, would be to have equal shares in each of the ethnic groups (Table 4.2-1). Although this might appear to be the optimal solution for the analysis, the caveat is that with equal (unweighted) numbers in each ethnic group, there would be extremely high weights for the "Rest" group, which constitutes approximately $85 \%$ of the population in Ontario.

Were a sample to have been done with no special measures to target the ethnic groups of interest, then the distribution of the sample by ethnic group would likely have followed the pattern for the province, with only $3.6 \%$ of the individuals falling into the Black group, $5.2 \%$ falling in the East Asian group and $6.5 \%$ falling in the South Asian group.

Under the assumption of completely random response patterns, the expected distribution by ethnic group was calculated using the ethnic composition of the selected DAs from the most recent Census of Population. This was a point of reference for the improvement expected in the sampling plan, without the influence of field processes. Given the sampling plan, it was expected that $6.9 \%$ of the sample would be from the Black ethnic group, nearly double their proportion in the Ontario population (3.6\%).

Finally, the unweighted distribution for the sample gives an indication of how successful the sampling plans, as well as the field processes, were at targeting the desired ethnic populations. All three minority ethnic groups of interest attained proportions roughly double their proportion in Ontario. For both the Black and East Asian groups, the proportion actually attained in the sample was slightly better than expected; $8.7 \% \mathrm{vs} .6 .9 \%$ for the Black group, and $9.9 \%$ vs. $9.4 \%$ for the East Asian group. For the South Asian group, their proportion in the sample was $15.0 \%$, which is slightly lower than what was expected (15.7\%), but is still more than double their proportion in Ontario (6.5\%).

Table 4.2-2
Response rates by Ethnic Group

| Source | Ethnic Group |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
|  |  | Black | East Asian | South <br> Asian |
| "Ideal" | $25 \%$ | $25 \%$ | $25 \%$ | $25 \%$ |
| Weighted (2006 Census of Population) | $3.6 \%$ | $5.2 \%$ | $6.5 \%$ | $84.7 \%$ |
| Expected (unweighted) | $6.9 \%$ | $9.4 \%$ | $15.7 \%$ | $68.0 \%$ |
| Attained (unweighted) | $8.7 \%$ | $9.9 \%$ | $15.0 \%$ | $66.4 \%$ |

[^26]
## 5. Conclusions

The method of frame enrichment succeeded in targeting the ethnic populations of interest for the Ontario Survey on the Prevalence of Hypertension (Leenan, et. al., 2008)

It would be an interesting project to redesign the same survey using ethnic stratification at the first stage. Another interesting option would be to instead employ two-phase sampling relying on the Canadian Community Health Survey (CCHS); an option that was not possible at the time. It is impossible to know if these methods would have yielded better gains in ethnic diversity in the sample.

## References

Cochran, K. (1977). Sampling Techniques. New York: John Wiley.
Colin-Bell, A., Adair, L. and Popkin, B. (2002). Ethnic differences in the association between body mass index and hypertension. Am J Epidemiology, 155:346-53.

Leenan, F., Dumais J., McInnis, N., Turton, P., Stratychuk, L., Nemeth, K.., Lum-Kwong, M. and Fodor, G. (2008). Results of the Ontario Survey on the Prevalence and Control of Hypertension. Canadian Medical Association Journal, 178(11):1441-9.

Kalton, G. (2009). Methods for oversampling rare subpopulations in social surveys. Vol. 35, No. 2, pp. 125-141.
Lohr, S.L. (2010). Sampling: Design and Analysis. Thomson Publishing, $2^{\text {nd }}$ Ed.
Statistics Canada (2003). Survey Methods and Practices. Catalogue No. 12-587-XPE.
Statistics Canada (2007). Profile of all levels of geography in Canada. Catalogue no. 94-568-XCB.
Stratychuk, L. and Dumais, J. (2007). Ontario Survey on the Prevalence of Hypertension in Adults: Practical Aspects in Sample Allocation and Collection. Proceedings Statistical Society of Canada Meeting, Statistical Society of Canada.

Yung, W. (1997). Variance estimation for public use microdata files. Proceedings of the Statistics Canada Symposium, 91-95.

## SESSION 3C

## POVERTY DATA

# Using a 'community survey' and previous census estimates to produce composite estimators for the South African Index of Multiple Deprivation 2007 

Chris Dibben, Gemma Wright and Michael Noble ${ }^{1}$


#### Abstract

A number of countries carry out large sample 'community surveys' to enable small area estimates to be made for inter-census periods. This paper discusses the use of the South African Community Survey, a survey covering some 274,000 dwelling units, to produce the South African Index of Multiple Deprivation (SAIMD) 2007. The SAIMD, a small area index, was originally census-based and produced by the University of Oxford's Centre for the Analysis of South African Social Policy (CASASP) for the South African Department for Social Development. However, because of the desire to assess change since the census was carried out in 2001, a methodology capable of updating the previous index using information from the South African Community Survey was explored and developed. The potential for using 2007 administrative data was also examined but no suitably robust datasets were found. A composite estimator was, therefore, produced within a multilevel logistic modeling framework using both the original 2001 census variables and the 2007 community survey. Measures of uncertainty were also derived from the model, with some estimators of the different dimensions of the index being stronger than others. The estimators were then combined to produce an overall index. The pattern revealed underlines the continuing problem of inequality across South Africa and especially the clear spatial legacy of apartheid.


[^27]
# Using complete administration data for nonresponse analysis: The PASS Survey of Low-Income Households in Germany 

Rainer Schnell, Tobias Gramlich, Alexander Mosthaf, Stefan Bender ${ }^{1}$


#### Abstract

In this study, nonresponse in a large survey of welfare receiving households in German is studied using complete administrative data of the German social security system on the individual and household level for both, respondents and nonrespondents. Despite the low response rate, only small differences between participating and nonresponding households and individuals could be found. A comparison of the estimates of different statistical nonresponse procedures with known values yield biased estimates for weighting and multiple imputation.


Keywords: refusal, non-contact, nonresponse, administrative data, weighting, imputation, propensity score weighting

## 1. Introduction

Nonresponse poses severe problems to all surveys when there are systematic differences between responding and nonresponding units. However, since nonresponse bias is a function of the response rate and the differences between responding and nonresponding units, a low response rate does not necessarily result in a large nonresponse bias (Groves, 2006; Groves and Peytcheva, 2008). Therefore, for an assessment of nonresponse bias, data on respondents and nonrespondents is needed. Detailed data on different nonresponse mechanisms is rarely available. Typically, there is only little information from the sampling frame, para data (e.g. the number and time of the contact attempts) or interviewer observations of the neighbourhood. In the case of the German household panel survey PASS rich information is available for analysis of nonresponse to the initial wave of the panel. For the first time, linkage of administration data of the German social security system was possible. These data provide detailed information on demographic characteristics and on variables related to the topic of the PASS survey ${ }^{2}$.

## 2. Nonresponse and low income samples

Surveys of low income populations suffer more than general population surveys from cooperation problems and increased non-contact rates. Low-income households are often more hard to find. Some factors are related to a lower level of education which strongly is related to low income and which may cause low literacy, a lower sympathy or understanding for the survey request, shame or fear of consequences when participating. Low educational status also is related to higher fear of crime and lower social trust in strangers (i.e. the interviewer). Due to a lack of social integration or increased alienation, a sense of civic duty or norm of reciprocity may be decreased, leading to lower survey cooperation.

Surveys on low income populations have to deal with the increased nonresponse problems (e.g. Gibson et al., 1999; Frederickson et al., 2005; and especially the book edited by Verploeg et al., 2001). Typically, these surveys achieve lower (compared to general population surveys) response rates. For example, Gallagher et al. (2005, p.7) report

[^28]typical response rates in Medicaid surveys down to $25 \%$ (depending on complexity and mode); Hernandez (1999) reports response rates to the initial wave of the low-income sample ('Survey of Economic Opportunity (SEO)'sample) of the Panel Study of Income Dynamics (PSID) of about 51\% (compared to the `Survey Research Center (SRC)' part of the PSID with an initial response rate of about 77\%).

## 3. Data

### 3.1 PASS Survey

The PASS panel survey (Promberger 2007; Trappmann et al., 2009) focuses on households of welfare recipients and the low income population in general. PASS consists of two subsamples. Sample I is a household sample of persons drawing social security benefits. It has been selected directly from the registry of welfare receivers at the German Federal Employment Agency. For these households administration data is available also on nonrespondents to the initial wave of PASS. Sample II is a stratified address sample of the general population with focus on low socioeconomic status households. We report here only on the sample of welfare recipients (PASS subsample I).

The survey was conducted in mixed mode. In principle, all households should be contacted and interviewed by phone (CATI). In cases no telephone number was available interviewers tried to contact the households in person at the sampled addresses (CAPI). Households contacted by telephone could switch to a personal interview if desired.

A total of 23,812 households have been sampled for the initial wave of the PASS survey. Table 3.1-1 shows the use of the different modes in sample I: in many cases telephone numbers were available for households of the receivers registry sample, therefore $86 \%$ percent of all households could be contacted by telephone (with $11 \%$ of all households switching mode to CAPI), only $14 \%$ of all sample I households were contacted and interviewed only by CAPI.

Table 3.1-1 Household numbers by sample and mode, PASS sample I

| Mode | Count | $\boldsymbol{\%}$ |
| :--- | ---: | ---: |
| CATI | 17,828 | 75 |
| CAPI | 3,320 | 14 |
| CATI/CAPI | 2,664 | 11 |
| Total | 23,812 | 100 |

## Nonresponse in PASS

Overall response rates to the first wave are low: only slightly more than one in four households have been contacted successfully and at least on interview within this household could be conducted. In sample I, overall response rate is $28.7 \%$. Table 3.1-2 shows the condensed and combined response codes from the contact reports of the interviewers. The largest subgroup of nonrespondents is refusals, followed by non-contacts. About $4 \%$ of all households moved since the date of sampling. In $2 \%$ interviewers reported language problems.

Table 3.1-2 Final response codes of households, PASS sample I

| Response Code | Count | \% |
| :--- | ---: | ---: |
| Interview | 6,844 | 28.7 |
| Address not found | 3,486 | 14.6 |
| No contact | 3,091 | 13.0 |
| Phys. or mental problems | 146 | 0.6 |
| Language problems | 469 | 2.0 |
| Refused | 7,006 | 29.4 |
| Unknown/other | 1,544 | 6.5 |
| Hh moved/dissolved | 915 | 3.8 |
| Deceased | 45 | 0.2 |
| Mode switch to CAPI | 206 | 0.9 |
| Not evaluable | 60 | 0.3 |
| Total | 23,812 | 100 |

### 3.2 Administrative data

Complete administration data also for nonrespondents are available for the sample of benefit receiving households. Administration data was linked to response code only and not to survey data. We used social security administration data coming from the German Federal Employment Agency prepared by the Institute for Employment Research (IAB ${ }^{3}$ provided by their research data centre ${ }^{4}$. The database provides spell data on persons receiving social security benefits. It includes sex, age, nationality, marital status, type and size of household and information on sanctions against the household when offending against terms and conditions. Additionally, we used spell data from a database on working histories of all persons ever employed. It contained information on education and vocational training, job position, working time, wages, unemployment episodes. ${ }^{5}$

## 4. Nonresponse analysis using process data

We used administrative data in order to check for nonresponse bias in sample I of welfare receiver households. Figure 4-1 shows the head of household's mean age, the household's size, days since the last job, the cumulative time of unemployment, and the daily wage or benefit rate for the different response outcomes from the overall sample-I mean assuming no nonresponse. All comparisons show no or only small differences between respondents and the full sample. There are also no or only small differences for refusals. Most notably are differences for nonrespondents with language problems (on average, they are older and especially have larger household sizes and less days in unemployment) or who moved shortly before the survey (younger, smaller household sizes, less days since the last job, days in unemployment), as well as ill nonrespondents (older, smaller household size, more days since the last job).

[^29]Figure 4-1 Differences in age, household size, time in unemployment, time since last job and daily wage by nonresponse generating mechanisms


Figures 4-2 and 4-3 show the hazard rates of welfare termination for relocated households and for households where no one was able to participate. Clearly, these are (rather small but) problematic groups, especially with relocated households and ill nonrespondents leaving welfare receipt significantly earlier and faster.

## Figure 4-2 Hazard rates of welfare termination for relocated households



Figure 4-3 Hazard rates of welfare termination (health or language problems)


## 5. Nonresponse correction

We compared 5 different techniques (e.g. see Sikkel et al., 2009; or Bethlehem, 2009). As a baseline we simply ignored nonresponse and didn't adjust the estimates at all. For raking we used the primary sampling unit (PSU), age and sex as control variables. The same variables were used for post-stratification. For propensity weighting we used mode, sex, age, nationality, household composition, education, changes in marital and employment status, categorized time since last job, changes in household size, and details on welfare receipt (administration type, sanctions and end of receipt) as predictors. For multiple hotdeck imputation PSU, age and sex were used with m=10 imputations. ${ }^{6}$

[^30]Table 5-1 Results of different nonresponse correction techniques

| Method | Time since last job <br> (days) | Household size <br> (persons) | Transitions out <br> $(\%)$ |
| :--- | ---: | ---: | ---: |
| Raking (PSU x age x sex) | $1,541.7$ | 2.24 | 15.3 |
| Post-stratification (PSU x age x sex) | $1,548.4$ | 2.23 | 14.6 |
| Propensity weights (13 variables) | $1,555.8$ | 2.20 | 17.7 |
| Multiple imputation (PSU x age x sex), m=10 | $1,524.1$ | 2.24 | 14.6 |
| Respondents (no adjustment) | $1,560.8$ | 2.26 | 14.8 |
| Full sample (no nonresponse) | $1,592.4$ | 2.19 | 17.8 |

All but propensity weights underestimate transitions out of welfare receipt, as well as all but propensity weights overestimate the household size. All of the corrections in our example perform poor in adjusting the time since the last job.

Figures 5-1 and 5-2 show the cumulative hazard rates adjusted for nonresponse. Post-stratification and rake weights both underestimate hazard rates of welfare termination, propensity weighting on the other side, overestimates hazard rates.

## Figure 5-1 Adjusted hazard rates of welfare termination (post-stratification, raking)



Figure 5-2 Adjusted hazard rates of welfare termination (propensity weights)


## 6. Results and Discussion

Despite the low response rate of less than $29 \%$ of all sample I households, overall nonresponse bias in PASS is small. This even is true for refusals. Furthermore, bias is specific to some variables. Of all examined variables (age, time since last job, cumulated time in unemployment, wage, household size, sex marital status, number of children) only household size and cumulated time unemployed show significant although small overall nonresponse effects. Nevertheless, for some nonresponse generating mechanisms we observe differences between respondents and nonrespondents. The largest differences can be shown for relocated households, ill nonrespondents, and for households for which the interviewers reported language problems. Fieldwork methods should concentrate more on these groups than on refusals alone.

In our analyses of different nonresponse compensating techniques, propensity weighting performed best. Like all other techniques, propensity weighting underestimates time since last job. Simple weighting schemes and multiple imputations resulted in biased estimates.

Based on this limited example we would conclude that nonresponse compensating techniques perform better when more and detailed data on nonresponse generating mechanisms can be used.

## Acknowledgements

This work is funded by the German Research Foundation DFG (Priority Program 1292 'Survey Methodology', PPSM) and in cooperation with the Research Data Center (FDZ) of the German Federal Employment Agency at the Institute for Employment Research (IAB). The authors are indebted to Mark Trappmann and the PASS project group at the Institute for Employment Research, Nuremberg.

## References

Bethlehem, J. (2009). Applied Survey Methods. A Statistical Perspective, Hoboken: Wiley.
Fredrickson, D., Jones, T. and Molgaard, C. (2005). Optimal Design Features for Surveying Low-Income Populations, Journal of Health Care for the Poor and Underserved, 16(4), pp. 677--690.

Gallagher, P., Fowler, F. and Stringfellow, V. (2005). The Nature of Nonresponse in a Medicaid Survey: Causes and Consequences, Journal of Official Statistics, 21(1), pp. 73--87.

Gibson, P., Koepsell, T., Diehr, P. and Hale, C. (1999). Increasing Response Rates for Mailed Surveys of Medicaid Clients and Other Low-income Populations, American Journal of Epidemiology, 149(11), pp. 1057--1062.

Groves, R. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys, Public Opinion Quarterly, 70(5), pp. 646--675.

Groves, R. and Couper, M. (1998). Nonresponse in Household Interview Surveys, New York: Wiley.
Groves, R. and Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias. A Meta Analysis,Public Opinion Quarterly, 72(2), pp. 167--189.

Hernandez, D.J. (1999). Comparing Response Rates for SPD, PSID, and NLSY. Memorandum to Daniel Weinberg (Housing and Household Economic Statistics Division), SPD Working Paper 99-2, Washington, D.C.: US Census Bureau. Online available at www.census.gov/spd/workpaper/spd-comp.htm

Jacobebbinghaus, P. and Seth, S. (2007). The German Integrated Employment Biographies Sample IEBS Schmollers Jahrbuch: Journal of Applied Social Science Studies, 127, pp. 335--342.

Kauff, J., Fowler, L., Fraker, T. and Miliner-Waddell, J. (2001). Iowa Families That Left TANF: Why did they leave and how are they faring?, MPR Research Report (Reference No. 8217-318), Washington, DC: Mathematica Policy Research.

Promberger, M. ed. (2007). Neue Daten für die Sozialstaatsforschung. Zur Konzeption der IAB-Panelerhebung Arbeitsmarkt und Soziale Sicherung, IAB-Forschungsbericht 12, Nuremberg: Institute for Employment Research. Online available at http://doku.iab.de/forschungsbericht/2007/fb1207.pdf.

Sikkel, D., Hox, J. and De Leeuw, E. (2009). Using Auxiliary Data for Adjustment in Longitudinal Research, in P. Lynn (ed.) Methodology of Longitudinal Surveys, Chichester: Wiley, pp. 141-156.

Trappmann, M., Christoph, B., Achatz, J., Wenzig, C., Müller, G. and Gebhardt, D. (2009). Design and stratification of PASS. An new Panel Study for Research on Long Term Unemployment, IAB-Discussion Paper 5, Nuremberg: Institute for Employment Research. Online available at http://doku.iab.de/discussionpapers/2009/dp0509.pdf

VerPloeg, M., Moffitt, R. and Citro, C. eds. (2001). Studies of Welfare Populations. Data Collection and Research Issues. Panel on Data and Methods for Measuring the Effects of Changes in Social Welfare Programs, Washington DC: National Research Council. Committee on National Statistics.

## SESSION 4A

## JUSTICE DATA

# Operational needs versus data requirements: Challenges in providing policing data to Statistics Canada 

Tracesandra McDonald, M. A ${ }^{1}$


#### Abstract

The Canadian Centre for Justice Statistics (CCJS), in partnership with Canadian police agencies, collects and disseminates police-reported crime statistics with the ultimate objective of providing academics, governments, the media, and citizens in general, with a local and national picture of the nature and extent of crime in Canada. The data is collected via the Uniform Crime Reporting (UCR) and Homicide Surveys. The data collected from these surveys represent substantiated reports of criminal activity that include descriptive information surrounding the incident, the victim (in cases of crimes against persons), and the offender. At the national, provincial and municipal levels, crime statistics are used to support a variety of purposes including policy direction, program evaluation, legislative development, allocation of criminal justice resources and performance measurement. This also holds true from a law enforcement perspective with Intelligence-Led policing - an approach whereby key intelligence data and statistics are analyzed and used to inform management of specific problem trends within their jurisdiction, as well as the most efficient and effective operational response. That being said, operational conditions in a policing environment are not always conducive to the legislated requirement to provide Statistics Canada with timely and accurate crime data. This paper explores the various issues and challenges to the collection of crime data from the perspective of the police who are at the respondent level. It will also explore ongoing strategies to enhance the quality and relevancy of this very important source of information for both law enforcement and the public at large.


## 1. Defining Data Quality

The concept of high quality data goes beyond simple accuracy and completeness of the data set. While those are arguably the basis for quality information exchange, there are many other key criteria that should be considered when keeping the consumers of the data in mind. These include such attributes as accessibility, timeliness, usability, relevance, and interpretability (Strong, Lee and Wang, 1997). From a strategic management perspective, "data quality is a fundamental success factor for decision-critical information that drives the tactical and strategic initiatives essential to the enterprise's mission." (Harris, 2009).

As we move forward with our quest in the law enforcement industry to become more interoperable, the consequences of poor data, and thusly poor decision making, become a high risk activity that needs to be identified and managed appropriately. Bad data can "go viral", and determining where they will spread to next, or the harm they might cause, can be difficult to predict (Redman, 2008).

## 2. What is operational Policing Data?

Operational police information at its most basic level consists of all the data fields that make up a police case file. These are generally based on a call for service from the general public, but will also include self-generated work, including offences uncovered while on patrol, as well as street checks and traffic stops.

While the academic or business intelligence analyst focus will be on the aggregated crime trends that are derived from this data source, the front-line policing perspective is much different. The focus here is the building of enough information to identify a suspect and solve the case. An officer must offer enough proof or evidence to support ultimate prosecution and conviction in court. And ultimately, the goal of data collection for police agencies is to

[^31]prevent and reduce crime in the communities they serve. This is generally achieved through linkages: in other words, connections between people, vehicles, property and locations.

Intelligence-Led Policing is a much used concept in contemporary policing. It demonstrates the move from anecdotal information to support operational and tactical strategies to evidence-based decision-making. It is most apparent in such policing management models as Compstat, and the growing use of police performance indicators to measure success within an organization. There is an increased demand for accountability for police resources and actions in public forums, and mandatory data collection is now commonly built into legislative changes. There is also a rise in partnerships between the police and academic communities in order to better identify root causes of crime and develop and support effective crime reduction and crime solving initiatives.

## 3. Busting the Mythology of Policing Data

There are a number of myths around data quality in the policing environment. Some are pervasive at the senior management levels, others focused at the front-line, and some that are well held views by all levels of the organization. These myths need to be examined and resolved prior to undertaking a successful data quality initiative.

## Myth: The data doesn't really matter. No one looks at the data. Aside from Statistics Canada, no one cares about statistics.

This view is often held by front-line personnel, not so much by senior managers who are more likely to see the bigger picture, as well as be a part of the organization's performance accountability framework. At the national, provincial and municipal levels, crime statistics are used to support a variety of purposes including policy direction, program evaluation, legislative development, allocation of criminal justice resources and performance measurement.

More importantly, when trying to change attitudes at the front-line, policing organizations use their operational information to make decisions about people, vehicles, locations and how these are approached. This information, at its best improves public and officer safety, and helps solve and prevent criminal activity. Moreover, with advances in information sharing initiatives with law enforcement partners, police are now using one another's data to make more informed decisions. From this perspective, increased access to information can be an incredible asset, but exchanging poor or inaccurate information is a major liability for law enforcement.

## Myth: The serious case files are fine. It's only the small stuff that police are called to where we have errors in our data.

There is no link between severity and probability for error with policing data. Major case files such as homicide and multi-jurisdictional operations can improve opportunities for error and duplication just by their sheer size, numerous database silos, and the amount of file contributors. In addition, data requirements for major investigations are exponentially increased and will have multiple stakeholders whose needs will vary. Data security and investigational integrity become paramount in major cases which can often mean that important quality control personnel simply don't have access to the files.

Even though the collection of organized crime data for the Uniform Crime Reporting survey has been a priority for government and the law enforcement community, the data has yet to be published due to data quality issues. The Police Information and Statistics Committee of the Canadian Association of Chiefs of Police have undertaken to identify and resolve underlying causal factors prior to publishing this very important data. Failing that, the Committee will be forced to decide whether or not the capture and reporting of this information be discontinued on the survey due to lack of available and accurate data.

## Myth: Policy Development and Training is the answer to the data quality problem in our agency.

While policy, documentation, and training are key starting points in the quest to improve data quality, they cannot fix a problem in and of themselves. Standard operation procedures are the backbone of any policing organization. Just because rules are documented doesn't mean that they are communicated effectively or that there will be compliance among the membership. Training is often another overly used fall-back scapegoat for data quality problems. Training
associated to information management and Statistics Canada scoring is complex and thus most often requires an inclass setting in order to achieve a measure of success. In-class training is difficult in an operational policing environment particularly if large numbers of personnel should be in attendance. In addition, training can only go so far in terms of the learning process when it comes to this type of subject matter. The only way to truly assimilate this type of information is via application, trial and error, as well as ongoing support for case-specific assistance.

## Myth: Our data quality issues need to be resolved by our IT staff and/or our vendor.

Amazing things can be done to improve data quality from a technological perspective. Business rules and edits can be built into records management systems to block poor data from getting into the system to begin with. However, in an operational policing setting it is important to ensure that you don't make entering information too restrictive or onerous as key information needs to be collected and disseminated in a timely manner with colleagues back at the office and at the scene. Also, data quality cross-checks can have an impact on overall system performance. System performance is paramount in an operational setting. If the system is too slow or if it goes down it simply won't be used to enter or retrieve information vital to police responders. David Loshin, President of Knowledge Integrity Inc., refers to the reliance on IT to provide a silver bullet solution for data quality issues as Data Quality Mistake No. 1: "This optimistic hope for a 'magic tool' is evidenced by how often people acquire a data quality tool as a first step in setting up their data quality program. Buying software before developing a program is indicative of a reactive environment - and a misguided thought that data quality is a technology-driven solution." (Loshin, June 2007). Given that the main thrust of the problem is garbage in, garbage out - it is important to focus on the people and the culture of the organization, as this is the ultimate source of the data, not the technology itself.

The data is not owned by the IT department of a policing organization. Like other organizations the data in the operational systems is owned by the business - those who are responsible for understanding the purpose of the data contained therein and its use in driving the day to day decision-making activities in the policing universe both at the front-line and senior management levels. In order to be successful in any data quality or data driven initiative, the business and IT must partner effectively. Systems analysts and computer programmers don't generally come with the experience of a policing background. It is important that the solutions to improve quality data are driven by those who understand and rely on the data. Without them, solutions are likely to be flawed and ineffective.

## Myth: Data cleansing or data quality should be handled in project management mode.

Certain data quality issues can be handled in a project management mode so long as a short and long term strategy have been developed to clean up historical records and address the root cause of the problem to avoid future repetition. A large number of temporary resources can be targeted at such a project, a "tiger team" for example, particularly if there is a deadline for the use or publication of the data in question.

That being said, the problem is that senior managers in many organizations are under the misimpression that data quality problems can be identified, rectified, and that the issues are resolved - never to recur. What caused one data quality issue will not be the same catalyst for another. System upgrades, turnover of system users, changes to policy, procedures and legislation can all instigate serious problems with data input and output. Data Quality is a program state activity. If you approach data quality in project management mode, you will always be behind the eight ball, constantly reacting to issues instead of proactively seeking ways and means to better the quality and value of your information.

## 4. Factors Contributing to Data Quality Issues in Policing

There are a number of other factors in the policing environment that contribute to issues of data quality. Attitudes, as exemplified by the many data quality myths described above, play perhaps the most significant role. Traditionally, "paperwork" has been devalued in the policing world. While changes in technology have meant that paperwork has become more and more the flow of electronic information within and outside of the organization, attitudes around data capture remain much the same. The focus is on catching the bad guy and the tools necessary to do that - the chase, the slew of equipment in the patrol car and on the tool belt, the firearm and hand to hand combat training. Multiple silos of data sources are a common problem in the policing community. Various records management and database systems have different architectures, slight variations on data collection requirements and policies
governing their collection and disclosure. It is a major obstacle to the goal of one-time data entry. In addition, it often means that data is duplicated or missing from an essential source or it is conflicting due to a lack of automatic updating of information across silos. This leads to major problems of comparability.

The recent exponential increase in the amount of data to be collected is another major contributor to problems of data quality. Within the past decade most Canadian police agencies have gone through a major technological upgrade of their records management systems. Since 2003, the Royal Canadian Mounted Police (RCMP) has rolled out new records management systems from its legacy system. This meant going from 500+ to 13, 000+ fields of potential data capture. Aside from the immediate effect of overwhelming the users in the field, this also increased the opportunity for error to a very significant degree. From a business perspective, it is important to be an effective gatekeeper with regards to data collection requirements. The most expensive data is that which is not used. Information overload is a major issue in today's society. In August of 2010 Eric Schmidt, CEO of Google, was quoted at a conference as saying "every two days we create as much information as we did from the dawn of civilization until 2003...that is something like five exabytes of data, he says." (Siegler, MG, August 4, 2010). It is therefore important not to make the mistake of collecting data for the sake of having it versus actually using it.

Along the same lines, major upgrades and changes to the technology are a constant impediment to producing quality information. While understandably the goal is to improve the functionality of the software, it also means constant change and confusion for front-end users. New fields and growing demands for data collection and reporting requirements further exacerbate the problem. It is important to mitigate this through well-timed and well-planned releases that involve sufficient notice and well documented communication on the changes that are taking places and the roles and responsibilities of the various key contributors to the data.

## 5. Keys to Success and Quality Information

The chart below reflects the key elements to achieving quality data within a policing organization. It is important that all of these tools be developed and utilized in order to create a successful outcome.

## Policy

There is chaos and anarchy without rules. You cannot have "errors" in your data without establishing the rules around data entry. Policies and procedures are particularly important in a paramilitary organization. Police understand the concept of "by the book". Policies and procedures allow the organization to set clear expectations of what is required and ensure consistency. In an agency with a large geographical jurisdiction this is integral as you need to know that a member in one area is doing the same thing in the same way as another member in another area, and that if needed, these members can be transferred between jurisdictions with little or no effort involved in retraining.

## Training

Like policy, training serves the purpose of maintaining consistency around data entry and reporting. As mentioned earlier, training can present unique problems in an operational policing environment. In-class, while being the most effective medium, is not always feasible due to competing operational priorities and whatever major events that are occurring in the community at any given time. This is where the power of super-users and subject matter experts will be invaluable to the organization, as they are the key to success in allowing new users to apply the knowledge gained in training in an environment where "the rubber meets the road". In addition, organizations should consider some form of certification to occur prior to users being allowed onto the operational systems. This will allow you to identify individuals who did not assimilate the information provided in training so that you can provide them with the opportunity for coaching and retraining in a form more conducive to their learning style. In order to promote training around data collection and reporting requirements, not generally considered the most sought after courses among front-line membership, organizations often include HR incentives. Job postings to promotional opportunities at the more senior ranks can be tied to the successful completion of this type of training.

## Communication

Making sure the right information gets to the right people in a timely manner is key to improving the quality of policing data. Important messages around changes to the system, policies, procedures and legislation, identified problem trends in the data and how to correct and prevent them, as well as documented roles and responsibilities

when it comes to the entering, monitoring and approval of the data contained within the case file need to be disseminated to all levels of the organization. This can present particular challenges as various means must be employed in order to reach all individuals, including email, intranet pages, bulletin boards, and word of mouth. Who the message comes from can also be important to whether or not the communication is assimilated by its audience, meaning that it is necessary for senior command personnel to support and be an integral part of your communication strategy.

## Audit

An organization will not be able to effectively gage the quality of their data, or specifically what overall problem trends need to be addressed, without the ongoing performance of quality assurance audits. Audits should always be conducted and perceived as a learning opportunity instead of a method of measuring error. Command staff should support the process as it allows them to successfully identify and address the risks they are currently running to which they may have been previously unaware. Audits should make recommendations for improvement to business processes, identify areas of focus for refresher instruction and coaching, as well as provide guidance back to main policy centres for enhancements to training modules, communication and policy.

A less formal version of auditing can also come in the form of specific verification exercises. With a more focused scope of monitoring, these types of projects often examine outliers and the surpassing of set thresholds. Verification exercises are a particularly effective tool in the preparation of the data prior to publication in order to increase confidence in the data, ensuring credibility of the organization, and the preparation of the owners of the data for the explanation of trends, outliers or anomalies that are found.

As discussed previously, technology has an important role to play in achieving quality data. System edits, automated auditing programs, and key configuration settings can all be a part of an overall strategy to restrict errors from getting into your key operational databases. Again, it is important to keep in mind that you don't want to overly impede data entry to the point of user frustration, since this has a tendency to either alienate them from using the system altogether, or results in user's making up data in order to get a record opened. It is a tricky balancing act that requires extensive analysis, development, testing and post-implementation evaluation.

Other technologies that will assist police agencies in improving the quality of their information sources are those that provide more seamless information entry, such as swipe card technology. Getting key personal identifiers into the database with the swipe of a card can significantly improve the amount and quality of master name files, avoid the creation of duplicate entities, and reduce administrative burden from front-line personnel. Front-end applications that streamline data entry based on event type can also assist with data quality as it focuses users to the key data elements required for this particular type of file instead of overwhelming them with the entire set of data fields contained within the case management system.

There are also many IT programs that resolve data quality problems through verification analysis and resolution at the Business Intelligence or reporting stage of data usage. While these software applications are great for analysts working on tight deadlines that will want the opportunity to assess and potentially eliminate outliers or anomalies, as well as resolve entity duplication, they are not designed to resolve the problem that is still contained within the source systems. These source systems will continue to propagate that information outward via interfaces, increasing the risk of bad decision-making, both internal and external to the policing agency.

Scripting of data can also be a wonderful tool when there is discovery of large amounts of incorrect data that can be fixed in an automated fashion without user intervention. That being said, it is important to use this type of tool sparingly, as it does not get at the root cause of the problem, which is more often than not people's behaviour.

## Accountability

Lack of supervision and accountability are often cited as key factors to many of the problems facing the policing community at this time. Data quality problems are no different and these have a significant impact on the information being collected. Common attitudes around accountability are: "What is going to happen, are you going to take away my birthday?" and "Someone else will catch the error, complete or update the information, etc."

Many police organizations utilize their civilian personnel for data entry and the reporting of statistical information to Statistics Canada. It reduces the burden on the front-line and allows for a smaller subset of users to monitor and train for quality assurance purposes, thus increasing consistency in data entry, and freeing up police resources for investigation, crime prevention and community policing activities. The drawback is that civilian staff do not have direct access to the scene, or the people involved, and they still rely on police officers to impart this information to them in a timely manner. At the end of the day, police records staff are not accountable in a court of law for the information contained on that case file. It is the lead investigator that is ultimately accountable for his or her file and the data contained therein. It is thus important that clear expectations of roles and responsibilities are documented and accessible, and that consequences for non-compliance carry some weight, or have significance to the individual. This leads us straight into the next element that discusses the use of compliance mechanisms.

## Compliance

It is important that an organization develop clear consequences for poor data quality. At the organizational level these consequences can be significant, resulting in loss of credibility, loss of business, and with police agencies, the most serious data quality errors have the potential for civil litigation and decreased officer and public safety. This should be made clear to members at the front-line as they are often cut-off from the larger picture. Compstat meetings and models increase accountability for the data and the actions of command staff in a public forum. In addition, top ten best and worst lists can engender healthy competition within an organization. It is important to also utilize positive reinforcement, rewarding staff for low error rates and ensuring their best practices are shared with others.

## Buy-in

This is probably the most important element to ensuring quality information, and yet probably the most elusive in a policing organization. Quality management personnel need to ensure the age old question of "what's in it for me? " is answered. At the end of the day, there exists one main factor to solving and preventing crime in the community and that is information. Information is indeed power. The data police collect is ultimately used by public safety stakeholders to make key decisions about changes to legislation (potentially making their job easier), the evaluation of criminal justice programs, and to determine next year's policing budgets. While it may not be considered the "sexiest"" part of the job, police are information managers, and they need to understand the awesome power and responsibility that goes with that role. Much of the attitude change comes through the demonstration of what quality information can do for them, and conversely, how poor data can sabotage major decisions they make at both the front-line and senior command levels.

## References

Harris, J. (June 20, 2009 at 4:46pm). The Data- Information Continuum (blog posting), at http://www.ocdqblog.com/home/the-data-information-continuum.html

Loshin, D. (June 2007). Data quality management pitfalls: Three common mistakes to avoid", online article from Ask the Data Management Expert: Questions and Answers at http://searchdatamanagement.techtarget.com

Redman, T.C. (2008). Data Driven: Profiting from Your Most Important Business Asset, Boston, Massachusetts: Harvard Business Press.

Siegler, M.G. (August 4, 2010), Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003, online article from TechCrunch at http://techcrunch.com/2010/08/04/schmidt-data/

Strong, D.M., Lang, W.L. and Wang, R.Y. (1997). Data Quality in Context", Communications of the ACM, May 1997, Volume 40, No. 5.

# The Impact of changing concepts on estimates from administrative data: The case of the Integrated Criminal Courts Survey 

Anthony Matarazzo ${ }^{1}$


#### Abstract

The Integrated Criminal Court Survey (ICCS) is a national criminal court database containing statistical information on appearances, charges, and cases for youths and adults for all Criminal Code and other federal statute charges heard in criminal courts in Canada. In any given year, approximately 13 million court appearance records are received by the Canadian Centre for Justice Statistics (CCJS) from the provinces and territories. Basic units of count derived by the ICCS from these microdata include court appearances, criminal charges (approximately 1.5 million) and, ultimately, criminal cases (approximately 500,000).

This presentation describes the basic structure of the ICCS and highlights the process by which charge and cases estimates are derived from the microdata. Particular attention is paid to recent changes to the ICCS and its extraction process, which has moved from an "end-date" case to a "snapshot" person-based definition. The impact of these changes on derived estimates as well as the various mechanisms in place to address them are also discussed.


[^32]
# Opportunities and challenges in using administrative justice data for longitudinal and network research 

Peter J. Carrington ${ }^{1}$


#### Abstract

This paper explores ways in which data on the administration of justice held by the Canadian Centre for Justice Statistics (CCJS) have been used, and could be used, for longitudinal and network research on delinquency and crime, and some problems associated with such research. The paper focuses mainly on the Incident-Based Uniform Crime Reporting Survey (UCR2), but also considers the Youth Court Survey (YCS) and Adult Criminal Court Survey (ACCS). Although none of these surveys has a unique identifier for individuals, they have been used to do longitudinal research on the recorded "criminal careers" of accused offenders, by linking the records over multiple years for each accused person, matching on soundex code, sex and date of birth. Some results are presented. Research on delinquent and criminal networks would require linking together incidents involving overlapping sets of co-accused in different incidents.


Key Words: Longitudinal; networks; criminal; delinquent; UCR2.

## 1. Longitudinal research on crime and delinquency

### 1.1 Research questions in life-course criminology

Life-course criminology has become one of the major paradigms for the study of crime and delinquency (Sampson and Laub, 1993; Laub and Sampson, 2003). It considers parameters of the criminal career, or sequence of crimes committed over the life-span, and their causes and consequences. Key parameters include:

- The age of onset and termination of offending.
- Age-specific and cumulative prevalence and rate of offending, and the contribution of each to the agespecific volume of crime.
- Frequency of offending (i.e. the number of offences in the criminal career).
- Specialization versus versatility in the types of crimes committed.
- Escalation/de-escalation/stability in the seriousness of offending.
- Co-offending (i.e. committing crimes with accomplices).


### 1.2 Data sources for life-course research in criminology

Delinquent and criminal behaviour over the life-course are usually studied using custom data sets. Data are collected prospectively, following the sample as they age, and incorporating data from interviews and from administrative records such as police files or school records (Liberman, 2008). Because of the labour-intensiveness of this mode of data collection, samples are usually relatively small, comprising 500-2,000 cases.

Alternately, data may be taken exclusively from administrative records. This strategy permits the use of much larger samples or populations, and obviates the need for prospective following of subjects over a lengthy period of time. The main cost is the restricted range of variables available, especially on the correlates of offending. The author has adopted this strategy to study delinquent criminal careers of Canadian youth, using data from three surveys maintained by the Canadian Centre for Justice Statistics:

[^33]- the Incident-Based Uniform Crime Reporting Survey (UCR2),
- the Youth Court Survey (YCS), and
- the Adult Criminal Court Survey (ACCS).


### 1.3 Selected results

Figure 1.3-1 is taken from a report on delinquent careers of Canadian youth between the ages of 5 and 17 inclusive, using data on police-recorded crime of age cohorts born in 1987 and 1990, from the UCR2 Survey (Carrington, 2007). It shows the percent of the age cohort who were first identified by police as offenders (i.e. the age of onset of recorded offending) at each year of age. For males the peak years of onset are 15 and 16; prevalence of recorded offending is lower in females, and the peak ages of onset are 14 and 15.

Figure 1.3-1
The peak ages of onset of police-recorded offending are 14-16 years of age


Source: P.J. Carrington. The Development of Police-reported Delinquency Among Canadian Youth Born in 1987 and 1990. Crime and Justice Research Paper Series. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada, 2007. Cat. No. 85-561-009.

Figure 1.3-2 shows similar information, but for "court careers" (the sequence of court cases in offenders' lives) over the ages of 12-21 inclusive, based on data from the YCS and ACCS. The peak age of onset for incidents that resulted in youth court or criminal court proceedings, and for cases resulting in a finding of guilt, was 18 at the time of the incident (Carrington, Matarazzo and de Souza, 2005).

Figure 1.3-2
The peak age of onset of court careers is 18
Figure 2
Rates per 100 cohort members of referral to court and conviction, by the age of the accused at the time of the incident


Data source: Statistics Canada, Canadian Centre for Justice Statistics, Youth Court Survey and Adult Criminal Court Survey.

Source: P.J. Carrington, A. Matarazzo, and P. de Souza. Court Careers of a Canadian Birth Cohort. Crime and Justice Research Paper Series. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada, 2007. Cat. No. 85-561-006.

Figure 1.3-3
The "age-crime curve": age-specific volume of crime, by sex


Source: P.J. Carrington. The Development of Police-reported Delinquency Among Canadian Youth Born in 1987 and 1990. Crime and Justice Research Paper Series. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada, 2007. Cat. No. 85-561-009.

Figure 1.3-3, based on UCR2 data, shows the age-specific volume of recorded crime for ages 5-17 inclusive, by sex. There appears to be a rapid increase, or "take-off", in the volume of recorded crime after the age of 11 , followed by a tapering-off after 15 (for females) and 16 (for males). Life-course criminologists have asked whether this is primarily due to:

- An increase in prevalence; i.e. the proportion of the population who participate in crime, or
- An increase in the rate of offending, or
- Neither, but the result of under-recoding by police of crime committed by children under 12, since they are below the age of criminal responsibility.
Figure 1.3-4
Observed and fitted age-specific prevalence, males, combined cohorts


Source: Adapted from P.J. Carrington, The transition from childhood delinquency to teenage youth crime. Conference on Life Course Transitions of Children and Youth, Atlantic Research Data Centre. Dalhousie University, Halifax, October, 2007.

Figure 1.3-5
Comparison of growth in prevalence and rate of offending, both sexes, combined cohorts


Source: Adapted from P.J. Carrington, The transition from childhood delinquency to teenage youth crime. Conference on Life Course Transitions of Children and Youth, Atlantic Research Data Centre. Dalhousie University, Halifax, October, 2007.

Analysis of UCR2 data suggested the first explanation. Figure 1.3-4, fitting the spline curve $\log (y)=0.48(x-5.25)-$ $3.5+(x>11.75)\left[0.30(x-11.75)-0.09(x-11.75)^{2}+0.00176(x-11.75)^{3}\right]$ shows that the age-specific prevalence of recorded offending by males increases exponentially up to the age of 12 , then more than exponentially from the ages of 12 to 15 . Adding a parameter for a step-increase at 12 , testing the third hypothesis, did not improve the fit of the curve. Figure 1.3-5 shows that the rate of offending increases much less rapidly, thus disconfirming the second explanation.

Figure 1.3-6 shows the uneven distribution of recorded delinquent activity among offenders. A few "chronic" offenders, with 5 or more recorded offences during their teenage years, are responsible for almost half of all recorded crime committed by this cohort. On the other hand, more than 60 per cent of offenders have only one recorded offence and are responsible for only one-quarter of recorded crime. This result, first discovered by Wolfgang and associates (1972), motivated the adoption of policies of selective incapacitation and intensive supervision of chromic offenders.

Figure 1.3-6
A disproportionate amount of crime is due to a few chronic offenders


Source: P.J. Carrington. The Development of Police-reported Delinquency Among Canadian Youth Born in 1987 and 1990. Crime and Justice Research Paper Series. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada, 2007. Cat. No. 85-561-009.

Figure 1.3-7 shows the peak in co-offending in childhood and early adolescence, and the decline with age thereafter. This result, which has been replicated by research with other datasets, is susceptible of several different explanations:

- Age-related changes in the type of crime committed, from crimes characterized by cooffending to those characterized by solo offending;
- Selective attrition of offenders with a tendency to co-offend;
- Recruitment versus experience;
- "Maturation": i.e. age-related decline in gregariousness and/or increase in autonomy.

Figure 1.3-7

## Co-offending by the age of the offender



Source: Adapted from P.J. Carrington. Group crime in Canada. Canadian Journal of Criminology 44 (3) (2002): 277-315, Figure 1.

Figure 1.3-8
Percent of incidents involving co-offending, by the age of the offender and selected types of offence


Source: Adapted from P.J. Carrington. Group crime in Canada. Canadian Journal of Criminology 44 (3) (2002): 277-315, Figure 4.

Figure 1.3-8 shows variations in the age/co-offending curve for different types of offences, suggesting that changes with age in the type of offence may be part of the explanation. Figure 1.3-9 shows a complex relationship between co-offending and criminal experience: co-offending decreases with criminal experience (the sequence number in the criminal career of the incident) for the great majority of offenders (those with 7 or fewer incidents in their careers), but is approximately constant with experience for high-activity offenders, and increases with experience for younger offenders.

Figure 1.3-9
Four co-offending trajectories


Source: P.J. Carrington, Co-offending and the development of the delinquent career. Criminology 47(4) (2009): 1295-1329.

## 2. Network research on crime and delinquency

What is the social structure of co-offending? There are two opposed views:

- Co-offending is a manifestation of organized criminal groups; or
- Co-offending arises from the more-or-less random activities of "flexible and transient...microgroups" (LeBlanc and Fréchette, 1989).

Network analysis of longitudinal co-offending patterns could resolve this question. Non-randomness in co-offending would suggest criminal organization; randomness would suggest the converse. Examples of non-randomness in network structure include:

- Repetition: a tendency for co-offending pairs, triplets, etc. to be repeated (i.e. not to be transitory);
- Transitivity (clustering)
- Centrality (a tendency for one or some offenders to be central in the network).

The extent of these types of non-randomness could be estimated using network analysis (Carrington, Scott and Wasserman, 2005). Figure 2-1 shows the network structure among a hypothetical set of 8 offenders involved in 9 cooffending incidents, where the numbers on the lines indicate the number of incidents of co-offending between each pair of offenders. There is evidence of repetition (e.g. Jill co-offenders twice with Jamie and Jim), of transitivity, or clustering (Jill co-offends with Jamie and with John, who also co-offend with each other), and of centrality (Jill cooffends with four of the seven other offenders). While this evidence is visual and impressionistic, there are regression-like methods of network analysis that allow estimation of the strength and statistical significance of such tendencies.

Figure 2-1

## Co-offending network structure among 8 hypothetical offenders



## References

Carrington, P.J. (2002). Group crime in Canada, Canadian Journal of Criminology, 44, pp. 277-315.
Carrington, P.J. (2007). The Development of Police-reported Delinquency among Canadian Youth Born in 1987 and 1990. Crime and Justice Research Paper Series No. 9. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada.

Carrington, P. J. (2009). Co-offending and the development of the delinquent career, Criminology, 47, pp. 12951329.

Carrington, P.J., Matarazzo, A. and de Souza, P. (2005). Court Careers of a Canadian Birth Cohort. Crime and Justice Research Paper Series No. 6. Ottawa: Canadian Centre for Justice Statistics, Statistics Canada.

Carrington, P.J., Scott, J. and Wasserman, S., eds. (2005). Models and Methods in Social Network Analysis. New York: Cambridge University Press.

Laub, J.H. and Sampson, R.J. (2003). Shared Beginnings, Divergent Lives: Delinquent Boys to Age 70. Cambridge, MA: Harvard University Press.

Le Blanc, M. and Fréchette, M. (1989). Male criminal activity from childhood through youth. New York: SpringerVerlag.

Liberman, A.M. (ed.) (2008). The Long View of Crime. New York: Springer.
Sampson, R.J. and Laub, J.H. (1993). Crime in the Making: Pathways and Turning Points Through Life. Cambridge, MA: Harvard U. Press.

Wolfgang, M.E., Figlio, R.M. and Sellin, T. (1972). Delinquency in a Birth Cohort. Chicago: University of Chicago Press.

## SESSION 4B

SAMPLING FRAMES 2

# Use of address databases as a basis for household frames: Confirmation bias in housing unit listing 

Stephanie Eckman and Frauke Kreuter ${ }^{1}$


#### Abstract

Field listing of housing units is an expensive and time-consuming stage of the survey process. In recent years, survey researchers have begun using external data sources, such as administrative records or commercial address databases, as a basis for housing unit listing. Listers update the existing list in the field, adding and deleting units as necessary. This method of listing, which we call dependent listing, is believed to be more accurate and less expensive than listing from scratch, and is used by several survey organizations including the U.S. Census Bureau. This paper uses an experimental repeated listing design to demonstrate the presence of confirmation bias in dependent listing. We find that when provided with a list to update in the field, listers tend not to add missing units or delete inappropriate units. The listers are biased towards confirming the initial list as correct. We call this phenomenon confirmation bias. Frames developed with dependent listing are for this reason much more reliant on the underlying quality of the initial listing than has been thought. Furthermore, if the kinds of units undercovered and overcovered by the input list are different than those which are properly covered, confirmation bias can contribute to coverage bias in survey estimates. Survey researchers should have a good understanding of the source and quality of the records they use as a basis for dependent listing, as listers are not able to fix all of the errors in the field.


[^34]
# Experiences using the census as a sampling frame: The Case of the Access and Support to Education and Training Survey 

Elisabeth Neusy and Yves Bélanger ${ }^{1}$


#### Abstract

This paper provides an overview of the survey design of the 2008 Access and Support to Education and Training Survey. The sampling frame chosen for the survey was a telephone list frame constructed from two sources: telephone numbers from the 2006 Census of Population, supplemented with residential numbers from a Random Digit Dialling frame to improve coverage. The Census portion of the frame was stratified according to the age composition of the households, allowing households containing youths to be oversampled. Interviews were conducted with the households reached by the telephone numbers selected in the sample. Following collection, the survey design was evaluated.


Key Words: Sampling frame; Sample design.

## 1. Introduction

The Access and Support to Education and Training Survey (ASETS) addresses issues relating to antecedents and determinants to access to Post Secondary Education (PSE), including the role of student loans and savings in the financing of PSE, as well as participation in adult education and training. It was sponsored by Human Resources and Skills Development Canada. The target population for the survey consists of all Canadian residents aged less than 65 years old, excluding individuals residing in the three territories in the North and individuals residing in institutions.

The ASETS replaces three previous education surveys that covered specific population groups: 1) the Survey of Approaches to Educational Planning, which focused on 0 to 18 year olds; 2) the Post-Secondary Education Participation Survey, which focused on 18 to 24 year olds; and 3) the Adult Education and Training Survey, which focused on 25 years of age and older. Each of the surveys collected data on specific facets of Canadians' educational experience. They were last conducted in 2002 and 2003 as supplements to the Labour Force Survey (LFS). The surveys were integrated into ASETS in order to expand the analytical potential of the three separate surveys and to reduce costs.

This paper provides an overview of the survey design of the 2008 ASETS. Section 2 describes the sampling frame that was chosen for the survey, and Section 3 the sample design. Section 4 summarizes results of post-collection evaluations undertaken to assess the performance of the survey design.

## 2. Sampling Frame

### 2.1 Survey Requirements

One of the main considerations in choosing a sampling frame for the ASETS was that the sample size requirements were given in terms of three age groups: children aged 0 to 17 , youths aged 18 to 24 , and adults aged 25 to 64 , which correspond to the target populations covered by the three surveys ASETS replaces. The requirements, given in Table 2.1-1, were specified in terms of 'effective sample size', which is the number of respondents divided by the design effect of the sample design. As seen from the table, an oversample of youths was required: they represent $21 \%$ of the

[^35]overall sample but make up only $9 \%$ of the overall population.
Table 2.1-1
Effective sample size requirements by age group

| Age Group | Sample Size | \% Sample | \% Population |
| :--- | ---: | ---: | ---: |
| Children (0-17) | 7,000 | $25 \%$ | $22 \%$ |
| Youths (18-24) | 6,000 | $21 \%$ | $9 \%$ |
| Adults (25-64) | 15,500 | $54 \%$ | $56 \%$ |
| Seniors (65+) | 0 | $0 \%$ | $13 \%$ |
| Total | 28,500 | $100 \%$ | $100 \%$ |

Another consideration in choosing a frame was the data collection method: Computer Assisted Telephone Interviewing (CATI).

### 2.2 Frame Options Considered

Several frame options were considered:

1) Conducting the survey as an LFS supplement: This option was not feasible because too many LFS rotation groups would have been required to meet the sample size requirements for youths.
2) Random Digit Dialling (RDD): This is an attractive option because RDD frames are dynamic and up-to-date. However, a large sample would have been required because households cannot be targeted based on their age composition. The RDD option was abandoned because it would have been too inefficient and costly to implement.
3) Using the 2006 Census of Population as a frame of individuals: The individuals are stratified by their age group. However, because the survey's collection period was in 2008, recent immigrants and births since the 2006 Census would not have been covered. Tracing would have been required for out-of-date Census telephone numbers, which could have led to the undercoverage of individuals who move and are difficult to trace.
4) Using the 2006 Census of Population as a frame of dwellings: The dwellings are stratified based on their Census household age composition. New residences would not have been covered. Tracing would have been required for out-of-date Census telephone numbers, which could have led to the undercoverage of individuals who move to dwellings that are difficult to trace (e.g., rural dwellings with non-standard addresses).
5) Using the 2006 Census of Population as a frame of telephone numbers: The Census telephone numbers are stratified based on their Census household age composition. This option entails no tracing costs. However, there were serious concerns about the undercoverage of such a frame, two years after the Census.

In the end, a sixth option was chosen, that is, option five supplemented with 'Residential' numbers from an RDD frame to improve coverage. The next section provides further details.

### 2.3 ASETS Frame

The sampling frame chosen for ASETS is a telephone list frame, constructed from two sources: the 2006 Census telephone numbers, and telephone numbers from the 'Residential' portion of an RDD frame. It was constructed in several steps.

The construction of the frame began with a 2006 Census extraction file containing person level Census data and Census telephone numbers. Each household in Canada received either a short questionnaire (Form 2A) or a long questionnaire (Form 2B/2D). The extraction file contained both Census 2 A and $2 \mathrm{~B} / 2 \mathrm{D}$ households. Households with no valid telephone number ( $7 \%$ of households), duplicate telephone numbers ( $1.5 \%$ ), and households composed of individuals with imputed birth dates ( $2 \%$ ) were dropped from the Census extraction file. We expected some of these excluded households to be covered by the RDD portion of the frame.

The next step was to add RDD telephone numbers. RDD frames consist of banks of telephone numbers with a status for each number. The telephone numbers have a status of 'Residential', 'Business' or 'Unknown' (status cannot be determined). The usual sample selection process for RDD surveys involves selecting banks of telephone numbers, and then selecting telephone numbers within the selected banks. Telephone numbers of 'Residential' or 'Unknown' status are sent to the field for collection. This was not the selection process used for ASETS. Rather, the RDD frame information was used to extract the 'Residential' telephone numbers from the RDD banks and create a list of telephone numbers. The list of 'Residential' RDD telephone numbers was compared to the Census telephone numbers, and the RDD numbers that were missing from the Census were added to the ASETS frame in a separate stratum. The ASETS frame is therefore composed of distinct telephone numbers with no duplicates.

After the Residential RDD numbers were added to the frame, the Census 2A households were dropped from the frame. The advantage of subsampling from the Census $2 \mathrm{~B} / 2 \mathrm{D}$ households is that more Census variables are available to perform the nonresponse weighting adjustments. Note that the Census 2A households were not removed from the frame until after the match with the RDD numbers in order to ensure that they do not appear in the RDD stratum, as they are covered through the Census $2 B / 2 D$ weights.

## 3. Sample Design

### 3.1 Stratification and Sample Allocation

The telephone numbers on the ASETS frame were stratified by province, and each province was further substratified into six strata, as defined in Table 3.1-1. The first five strata are strata for the Census portion of the frame. The age of each household member was calculated as of August 8, 2008 (assumed midpoint of ASETS collection) based on the Census birth date, and used to classify the households.

Table 3.1-1
Stratification and sample allocation

| Telephone Source | Stratum Name | Household Composition | Allocation | Sampling <br> Fraction |
| :--- | :--- | :--- | ---: | ---: |
|  | Youth Only | Youths (aged 18-24) only | 1,360 | 0.0108 |
|  | Youth Mixed | Youths and other age groups | 19,640 | 0.0108 |
|  | Children Mixed | Children (aged 0-17) and no youths | 17,000 | 0.0072 |
|  | Adults Only | Adults (aged 25-64) only | 14,000 | 0.0030 |
|  | Seniors Only | Seniors (aged 65 and over) only | 0 | 0.0000 |
| Residential RDD | RDD | Residential RDD telephone numbers | 20,000 | 0.0070 |
| Total sample size |  |  |  |  |

The total sample size was 72,000 telephone numbers. Table 3.1-1 also provides the overall sample allocation to the six strata. The sampling fractions were controlled in order to meet the effective sample size requirements given in Table 2.1-1 as efficiently as possible. The households containing youths were oversampled, whereas households composed of Adults only were undersampled. No sample was allocated to the strata for households composed only of persons aged 65 and over according to the Census.

Kish allocation was used to allocate the sample between the provinces. A Kish parameter of 0.67 was used, which gives an allocation which is a compromise between equal allocation and proportional allocation. The sample was then allocated to the province by household composition substrata in such a way that the marginal allocations were respected (i.e. the provincial allocation and the household composition allocation).

### 3.2 Sample Selection

For the Census portion of the frame, the sample of telephone numbers was selected as a second phase to the Census 2B/2D long-form sample. The sample was drawn using systematic, probability proportional to size sampling, with the size measure equal to the Census weight of the household, so that the households within the same stratum end up with equal design weights. For the RDD strata, the sample was drawn using simple random sampling.

The selected telephone numbers were sent to the field for collection. The interview was conducted with the household reached by the selected telephone number, regardless of whether it was the same household as in the frame or still had the same age composition.

Upon contact with a household, a roster listing all household members and their age was collected, and one household member aged between 0 and 64 was randomly selected for the survey. The probability of selection depended on age: youths aged 18 to 24 were eight times more likely to be selected than adults aged 25 to 64 , and five times more likely than children aged 0 to 17 . If all members of the household were aged 65 and over, no household member was selected and the interview ended.

### 3.3 Arriving at the Final Sample Design

The sample allocation scheme and set of selection probabilities for the household roster were determined through an iterative process. Several different options were considered. For each option, the number of respondents and the effective sample size were predicted for each of the three age groups of interest (children, youths, adults), and compared to the survey requirements given in Table 2.1-1.

The predicted results were based on a number of assumptions, including the hit rates and mobility rates obtained from a pre-test of 1,000 Census telephone numbers. As well, design effects due to changes in household age composition (stratum jumpers) and due to the unequal probabilities for selecting a member from the roster, were predicted using the Kish formula given in Kish (1992). The predicted design effects were calculated for each of the three age groups of interest using the Census data. For the $k$ th age group ( $k=1,2,3$ ), the following formula was used:

$$
\operatorname{deff}_{k}=\sum_{i} \operatorname{prob}_{i k} \frac{\sum_{i} \operatorname{prob}_{i k} \mathrm{wgt}_{i k}^{2}}{\left(\sum_{i} \operatorname{prob}_{i k} \mathrm{wgt}_{i k}\right)^{2}}
$$

where $\operatorname{prob}_{i k}$ is the probability that person $i$ belonging to age group $k$ is selected and is a respondent, and wgt ${ }_{i k}$ is the predicted weight for person $i$ belonging to age group $k$. The weights, $\mathrm{wgt}_{i k}$, are the product of three factors: the inverse of the probability of selecting the household, the inverse of the probability of selecting the person from the household roster, and a predicted non-response/calibration adjustment factor.

The sample allocation scheme given in Section 3.1 and the set of roster selection probabilities described in Section 3.2 gave predicted results that were closest to the survey requirements.

## 4. Post-Collection Evaluations

ASETS data collection took place from June to October 2008. After collection, design weights were calculated and adjusted to take nonresponse into account, as well as calibrated to demographic totals. Bootstrap weights were calculated for variance estimation. Once the final survey weights and bootstrap weights were calculated, several evaluations were undertaken in order to assess the performance of the survey design. This section briefly describes the results of these evaluations.

### 4.1 Collection Results

Of the 72,000 telephone numbers sent to the field for collection, reported data was obtained for 31,500 individuals. Of the in-scope telephone numbers, a response rate of $64 \%$ was obtained.

Table 4.1-1 provides weighted stratum-level out-of-scope (OOS) rates. Note that households composed of 'Seniors Only' (all members aged 65+) are excluded from these out-of-scope rates for comparability with other surveys. The out-of-scope rate is highest for households in the 'Youth Only' stratum at $65 \%$. The second column of the table gives the estimated proportion of cellular phones. A standard set of screening questions is used for RDD surveys to flag
cellular phone numbers as out-of-scope. The same set of questions was used for ASETS because there was insufficient time before the start of collection to modify standard procedures.

Table 4.1-1
Estimated ASETS out-of-scope rates (weighted)

| Stratum | Cellular Phones | Out-of-Service | Other OOS | Total OOS |
| :--- | ---: | ---: | ---: | ---: |
| Youth Only | $25 \%$ | $30 \%$ | $10 \%$ | $65 \%$ |
| Youth Mixed | $3 \%$ | $12 \%$ | $3 \%$ | $18 \%$ |
| Children Mixed | $2 \%$ | $14 \%$ | $4 \%$ | $20 \%$ |
| Adults Only | $8 \%$ | $14 \%$ | $4 \%$ | $26 \%$ |
| RDD | $1 \%$ | $21 \%$ | $18 \%$ | $40 \%$ |
| Total | $4 \%$ | $15 \%$ | $9 \%$ | $28 \%$ |

The next table (4.1-2) provides stratum-level information on the stability of household composition. It gives the estimated proportion of in-scope telephone numbers that reach households whose composition based on the ASETS household roster is the same as it was on the frame. The table shows that the household age composition is least stable for 'Youth Only' households, and most stable for 'Children Mixed' households, as might be expected. Changes in household age composition since the 2006 Census are due to household members joining or leaving households, or to households moving and telephone numbers being reassigned to different households.

Table 4.1-2
Household composition: frame vs. collected

| Stratum | \% Same Household Type |
| :--- | ---: |
| Youth Only | $40 \%$ |
| Youth Mixed | $73 \%$ |
| Children Mixed | $90 \%$ |
| Adults Only | $86 \%$ |

### 4.2 Coverage

The coverage of the survey was evaluated by dividing the sum of the pre-calibration weights by the demographic counts used for calibration. Results are given in Table 4.2-1 by age group. The table shows that the RDD stratum helped increase coverage. Coverage is lower for the 22 to 34 age group, which can be explained by the higher mobility rates and lower hit rates observed for persons in this age group.

Table 4.2-1
ASETS coverage by age group

| Age Group | Overall Coverage | Census Strata | RDD Stratum |
| :--- | ---: | ---: | ---: |
| $0-17$ | $80 \%$ | $66 \%$ | $14 \%$ |
| $18-21$ | $79 \%$ | $66 \%$ | $13 \%$ |
| $22-24$ | $57 \%$ | $45 \%$ | $12 \%$ |
| $25-34$ | $61 \%$ | $47 \%$ | $15 \%$ |
| $35-44$ | $77 \%$ | $65 \%$ | $12 \%$ |
| $45-54$ | $85 \%$ | $74 \%$ | $11 \%$ |
| $55-64$ | $89 \%$ | $77 \%$ | $11 \%$ |
| Total | $78 \%$ | $65 \%$ | $13 \%$ |

The ASETS coverage was compared with that of the Canadian Tobacco Use Survey (CTUMS), which is an RDD survey: it is estimated that CTUMS has 7 percentage points better coverage than ASETS for persons aged 15 to 64 .

### 4.3 Minimum Proportions

One of the evaluations performed to assess the design's success, was to calculate minimum proportions the survey should be able to reliably estimate with coefficients of variation (CVs) of $16.5 \%$ based on median design effects. Table 4.3-1 presents the minimum proportions that were obtained, as well as the minimum proportions that had been
predicted at the design stage prior to collection. The table shows that the survey objectives were met for Children, and were fairly close to being met for Adults. The results for Youths are below target partly due to fewer Youths responding than predicted, and also due to additional unforeseen weighting adjustments.

Table 4.3-1
Canada-level minimum proportions for CVs of $16.5 \%$

| Children (0-17) |  | Youths (18-24) |  | Adults (25-64) |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Predicted | Realized | Predicted | Realized | Predicted | Realized |
| $0.7 \%$ | $0.7 \%$ | $0.7 \%$ | $1.2 \%$ | $0.3 \%$ | $0.4 \%$ |

### 4.4 Use of Census data in Weighting

The non-response weighting adjustments were performed by constructing logistic regression models to model propensity to respond and form weighting classes. For the Census portion of the frame, Census long-form variables were available as explanatory variables for the modelling. When the logistic regression models were constructed, several Census variables were significant and were therefore included in the nonresponse models, namely: mother tongue, education level, household income, dwelling type, immigrant status and family type.

After weighting was completed, a study was undertaken to assess the effect of including Census variables in the nonresponse models. A new set of weights was calculated without the aid of Census data, and several tables of estimates were recalculated. Twenty percent of the estimates were found to be significantly different. There is evidence that using the Census data in the nonresponse weighting adjustments helped reduce the nonresponse bias.

## 5. Conclusion

The survey design used was a good choice for ASETS. The Census portion of the frame allowed us to stratify based on Census characteristics and oversample households likely to contain Youths. The Residential RDD portion helped improve coverage. While the design was a good choice, given the survey's particular requirements, we would recommend a regular RDD design for surveys targeting the general population, as RDD designs appear to have better coverage (e.g. CTUMS).

For future surveys, we recommend that the possibility of conducting interviews with cellular phones be investigated. This would improve coverage, especially for the 22 to 34 age group. The coverage could be further improved by considering other sources of telephone numbers in addition to the RDD Residential numbers.

## Acknowledgements

The authors would like to thank Marcelle Tremblay, Suzelle Giroux and Ioana Schiopu-Kratina for providing valuable comments and suggestions.

## Reference

Kish, L. (1992). Weighting for unequal $\mathrm{P}_{\mathrm{i}}$. Journal of Official Statistics, 8, pp. 183-200.

# Using administrative data sources to target sampling in the Wealth and Assets Survey in the UK 

Paul Smith, Karl Ashworth and Charles Lound ${ }^{1}$


#### Abstract

The Wealth and Assets Survey is a recent addition to the UK's social survey portfolio and is a longitudinal survey following a panel of respondents. To over-sample certain categories with high overall wealth, sample areas were first selected from the UK's household survey frame, the Postal Address File, and the selected areas were matched to administrative data to indicate addresses with households likely to have certain characteristics. These were then sampled at different rates. The survey was weighted to population totals (derived from census and administrative sources in a standard way) to compensate for non-response. The paper will describe the survey design and the use of administrative data, and evaluate how well the sampling worked in practice using information from the first wave of the survey. It will also comment on how effective the weighting adjustments were for this type of data.


Key Words: sample design, design evaluation, differential sampling.

## 1. Introduction

There has been a persistent gap in the topic coverage of official social surveys in the United Kingdom, in that there has been no data on the assets and liabilities of households, except for a few attempts based on partial data from tax records. These did not include non-marketable assets other than pensions, as there was no suitable data source. The only way to get an estimate of the total wealth of the UK population was to run a suitable survey which could collect information directly, including the components of wealth not available from other sources. Limited information had been available from two waves of the British Household Panel Survey (BHPS) relating to the savings, assets and debts of households. Two BHPS measurements, five years apart, were used by the Institute for Fiscal Studies to assess the wealth of the population (Banks, Smith \& Wakefield, 2002). A survey also had the advantage that it could easily collect information relating to households, whereas administrative sources generally relate to individuals, so that the combination of values for different individuals into appropriate households is more difficult.

In 2000 the Office for National Statistics (ONS) was tasked by the Prime Minister's Performance and Innovation Unit to "bring together interested government departments to assess whether there was a case for a regular survey of wealth and assets". After an assessment of information collected in existing surveys and administrative sources, ONS produced a business case for a dedicated Wealth and Assets Survey. In 2005 a consortium of Office for National Statistics, Department for Work and Pensions, HM Revenue and Customs, Department for Business Innovation and Skills, Communities and Local Government, Her Majesty's Treasury and the Scottish Government agreed collectively to sponsor a new survey, the Wealth and Assets Survey (WAS) to provide these data to meet the policy needs.

## 2. Survey design

### 2.1 Survey objectives

Some of the objectives of the survey were well-specified at the beginning, and others emerged from discussion over

[^36]how the survey could be implemented and what it should cover. The main objectives were:
Measure all components of wealth: There are many components of wealth, and a wide range of financial instruments for various aspects, and the survey should provide an opportunity to make estimates for them all.

Longitudinal follow-up: It was clear from the beginning that the survey should follow up the same individuals and households, so as to make the measurement of change as accurate as possible and to investigate individual change. Initial expectations were that savings and investments would accumulate relatively slowly, whereas indebtedness could change relatively quickly, and that some adaptability in follow-up might be needed to account for these differences.

Disproportionate sampling of the wealthiest: There is interest in the whole of the wealth distribution, but particular interest near to the ends, and in order to get good estimates for the wealthiest they should be disproportionately included within the sample (with the added benefit of reducing the variance of the estimate of total wealth, see section 2.3). At the same time this oversampling should have a minimal impact on the accuracy of estimates for other parts of the wealth distribution, particularly the poorest group.

### 2.2 Sample size determination

To determine the achieved sample required, the likely precision of estimates of level and change produced by various options had to be predicted. There was a limited amount of survey information about wealth holdings in Great Britain with which to evaluate potential designs, but the BHPS, had collected total savings and total assets for each respondent on two waves ( 5,500 households and 10,000 people in 1995; around 8,500 households in 2000). These data could be used as a basis for a design, but needed some preparation, including imputing average values where only a banded response was obtained and apportioning joint holdings. These were used to estimate standard errors and then a design factor. Several untested approximations and working decisions were made, for example on inclusion of outliers, in order to make the predictions used for the design as robust as possible.

The exploration of various designs focussed on the likely precision for estimates of savings. The average level of savings was estimated to have increased by $18 \%$ over the five year interval between the two BHPS waves where these questions were asked. It is not clear whether this represents a typical movement in this measure over recent years, nor whether such change will continue.

The design factor summarises the overall impact of complex sample design features (largely the multi-stage sampling in this case) by comparing the standard error achieved with the complex design with that which would have been achieved with a simple random sample (Kish, 1995). The clustering on the BHPS is different to that proposed for the WAS, but this difference can be approximately adjusted using a simple model relating the design factor, deft, to the average number of adults interviewed per primary sampling unit (PSU), $\bar{b}$, and the rate of homogeneity, $r$, which captures the impacts of both stratification and multi-stage sampling:

$$
d e f t=\sqrt{1+(\bar{b}-1) r}
$$

Butcher \& Elliot (1992, chapter 4). Using this model, we can estimate $r$ from an existing sample and apply a different value for $\bar{b}$ under a different design. This model is usually applied where there is a single level of clustering, whereas here there is clustering at both the PSU and household level. Since the modifications to deft required in this analysis were small, however, we reasoned that the impact of this approximation would be small.

Estimates of change from longitudinal surveys are much more precise than those from independent samples, because random error from drawing a new sample is excluded. To estimate this precision the observed correlations between measures on the same people in the BHPS, five years apart were used. These correlations vary for different groups in the population. For example, the five-year correlation for savings, $\rho_{\mathrm{s}}$, for those initially aged 16-24 is $29 \%$, but this rises to a maximum of $69 \%$ for those aged $40-49$, and is lower, at $12 \%$, for those aged $50-59$, some of whom will pass through retirement during the five years between waves. These were adjusted for a different intervening time interval by assuming that the correlation falls geometrically, with perfect correlation at year zero and the observed correlation at year five. So, where the observed BHPS correlation after five years was $\rho_{s}=0.29$, we predicted that the observed correlation after one year would be $\rho_{s}^{\frac{1}{5}}=0.79$, and after two years $\rho_{s}^{\frac{2}{5}}=0.61$. Based on these inputs and an interaction with the field design (see section 2.5), the survey sample was set at 32,000 households per wave. The design estimates of the sampling error and the actual observed errors are presented and discussed in section 5.

### 2.3 Disproportionate sampling

The distribution of wealth is well known to be highly skewed, with most people having below the mean level of wealth and a few people having a level of wealth many times larger than the mean. In an equal probability sample selection we would expect 10 per cent of households to fall above the top decile of the wealth distribution in the absence of differential non-response effects. The impact of differential non-response on this survey was difficult to predict. We expected more difficulty gaining response at the top end wealth distribution because the survey is more onerous when there are more assets. However, it is also well-known that a higher level of education, which is correlated with wealth, is indicative of a greater propensity to respond to surveys (Durrant \& Steele 2009), so it was unclear how these two factors would interact. The actual response pattern is evaluated in section 5.1.
The reasons to attempt to increase the achieved sample number of households falling in the top decile group are:

- to allow for more precise, separate analysis of this group so we could look in more detail at more complex arrangements for those with greater assets. The ability to do this detailed analysis was a key requirement for some of the survey customers;
- to improve the precision of total and average wealth estimates from the survey data. Since the distribution of wealth is highly skewed, the variation in wealth between wealthy individuals is much larger than that between less wealthy individuals. This population variation in the level of wealth influences the sampling error of the survey estimates and if relatively more of those showing the greatest variation can be sampled, then this sampling error will be reduced;
- to provide protection against a possibly worse response from the wealthiest group.

In order to have such a design, additional information correlated with the wealth of the population was needed; the source judged to be best was an administrative dataset based on self-assessment tax records.

### 2.4 Using administrative data for design

### 2.4.1 The administrative data

Her Majesty's Revenue and Customs (HMRC) produces extracts of self-assessment tax records for statistical analysis, including deriving estimates of total assets from the yields on investment reported in the self-assessment returns. This imputed wealth can cover only those assets which provide income and notably excludes owneroccupied housing and popular tax-free investments such as Investment Savings Accounts. Comparing the numbers of people on the self-assessment tax records with the size of the UK population, we found that around $18 \%$ of adults were included on the extract; $12 \%$ of adults had imputed wealth over $£ 1$ and $7 \%$ of adults had imputed wealth of $£ 10,000$ or more. For WAS design purposes, the imputed wealth for a household was taken as the sum of the imputed wealths of people at the same address.

This imputed wealth will not match the wealth measured from a survey like Wealth and Assets, because (a) the coverage of both the population and components of wealth are different, (b) imputed wealth relates to people but the WAS measures household wealth, and (c) there is a time lag between the tax year the data for the derivation of imputed wealth refer to and when the survey data are collected, and we can't be sure that the same people will remain in the same part of the wealth distribution or at the same address. However, from the process by which imputed wealth is derived and because we expect wealth to change slowly, we anticipate a high correlation between the imputed wealth of people at an address at an earlier reference period and total household wealth measured by interviewing those present in the current survey period. We can therefore use this imputed wealth as a stratum indicator for the Wealth and Assets Survey, and can then sample those with higher imputed wealth at a higher rate than others.

### 2.4.2 Methods for oversampling the wealthiest

Different ways of approaching this oversampling were considered. One possibility is to characterise each PSU in terms of the proportion of people with higher imputed wealth and then selecting relatively more of those PSUs. This is simple to do, requiring no address matching and results in a sample which has a simple structure. However, this approach was rejected as the indirect targeting by area meant that those with low wealth living in prosperous areas would be over-sampled and vice versa.

Instead, targetting at address level was assumed to be more effective. (The relative effectiveness of alternative approaches could be tested only once adequate data are available.) There are two ways to approach oversampling of the wealthiest addresses within the sampled PSUs. One is to draw a regular sample from the postcode address file (PAF) and top this up with addresses with high imputed wealth sampled from the self assessment records. (We could then subsample these addresses to give a fixed number of addresses for each interviewer.) With this method we would have a dual frame survey, and either need to determine the probabilities of selection for each respondent or use a pseudo maximum likelihood estimator (Skinner \& Rao, 1996). Where the address is sampled from the self assessment tax records and the household is found to be eligible for the survey, it is reasonable to assume that it could have been sampled from the PAF. However, only a proportion of PAF addresses are also on the self assessment tax records, so we would have to determine which of the addresses from the PAF sample could have been sampled from the self assessment tax records. Determining this complex historical information from survey is unrealistic, so the only realistic approach would be by matching these on to the self assessment tax frame. As the address records come from different sources, we could not expect a perfect match, so some of the probabilities of selection and hence the resulting weights would be inappropriate.

A second approach is to extract the self-assessment tax records addresses of households in the sampled PSUs whose imputed wealth exceeds a threshold value. These addresses can then be matched to the PAF and flagged. The sample of twenty six addresses per PSU is then drawn from the PAF, but giving the flagged addresses a higher chance of selection than other addresses. We implemented this by attributing a size measure to each address prior to systematic random sampling, a method which allows close control over the sampling rates. In this approach, even if the address matching process is not perfect, the weights derived will appropriately reflect the true probability of selection so any quality deficit will be with sample efficiency rather than bias. This method was chosen; it meant that only address data were passed from HMRC to ONS (see 2.3.3). With this approach, and a fixed sample size for each interviewer, the probability of selection within the strata defined by the flagging varies according to whether the selected PSU has a high or low proportion of flagged addresses. However, we estimate that the impact of this variation in the weights will be small.

### 2.4.3 Access to administrative data

There are some legal constraints on what the administrative data can be used for, but HMRC resolved that the sample of identified addresses, which was all that was passed from HMRC to ONS did not constitute "personal information". However, the legal situation (this predated the Statistics and Registration Service Act 2007 with its enabling clauses for data sharing for statistical purposes) and the context are constantly changing, and continued access is not guaranteed, which induces an element of uncertainty into long-term plans for the WAS. Security measures were put in place to safeguard the data, and the direction of data flow was restricted - in particular, ONS does not pass identifiable sample information to HMRC.

### 2.5 Fieldwork influences on the survey design

32,000 face-to-face household interviews was a substantial survey to begin from scratch, and therefore some aspects of the design were instituted in order to make the field data collection deliverable. First, it was decided that each wave of the survey should last for two years (so people would in general be followed up on a longitudinal cycle after two years). This fitted with the expected slow change in the levels of wealth, but was not ideal for following the evolution of indebtedness, which was expected to be less stable. Therefore an additional follow-up after one year of the most indebted households was included to meet the need for this information. This sample subset could then be analysed to give information on changes in indebtedness; this aspect of the design is not discussed further here.

The questionnaire development indicated that, particularly for the wealthiest households, the survey interview might be quite long, so in an attempt to reduce the burden, some questions were identified which could be asked of only a subset of the sample, and these were divided into two modules, such that each household would get only one of the modules. This subsampling involved some imputation to provide values for unasked questions as well as dealing with the more normal non-response issues.

## 3. Evaluation of the design

### 3.1 Response

There was some expectation that the wealthiest households would have a lower response rate because of the complexity and length of the questionnaire. The response rates in the design strata which represented the lowest $90 \%$ of imputed wealth and the top $10 \%$ of imputed wealth were $54.0 \%$ and $54.3 \%$ respectively, which are not significantly different using an appropriate Wald test ( $\mathrm{p}=0.603$ ).

### 3.2 Oversampling

The weighted survey data were used to construct estimated deciles of the distribution of total wealth excluding physical wealth. The distribution of the (unweighted) survey responses between these deciles, shown in table 3.2-1, shows the effect of oversampling the top decile group of imputed wealth. This distribution shows total wealth measured in the survey, but prior to imputation for missing item values.

Table 3.2-1
Frequency of sample households in estimated decile groups of the wealth distribution from wave 1 of the
Wealth \& Assets Survey

| Decile group of estimated <br> total wealth | \% cases |
| :---: | ---: |
| Lowest 1 | 9.8 |
| 2 | 8.5 |
| 3 | 8.9 |
| 4 | 9.2 |
| 5 | 9.3 |
| 6 | 9.5 |
| 7 | 9.9 |
| 8 | 10.4 |
| 9 | 11.1 |
| Highest 10 | 13.5 |

This oversampling was intended to meet the HMRC requirement of achieving 4,500 households, or $14.1 \%$ of the total sample of 32,000 households, above the top wealth decile. After the first year of wave 1, when addresses in the top decile group of imputed wealth had 2.5 times the probability of selection in the sample than other addresses, the target was not met based on unimputed totals, but was estimated to be met after imputation. The oversampling rate for the second year of wave 1 was increased to three times for safety, but after the full two years the target was still not (quite) met. The differences in selection probabilities are accounted for in the weighting. Since the relationship between the strata and wealth decile groups is not perfect, we see some oversampling of decile groups neighbouring the top group. This inevitably leads to fewer sample households in other decile groups, particularly the secondlowest, although the lowest group size appears unaffected. Nevertheless, the oversampling procedure seems to have effectively delivered the increased number of cases required by the survey objectives.

### 3.3 Sampling Errors

Table 3.3-1 shows the expected sampling errors based on the design, and how they compare with the actual errors achieved. The achieved errors are larger, but this might be expected for several reasons - the design errors are never achieved in practice, there was extra variability in the selection from the varying proportions of flagged households in each PSU, and the differences in weights were not included in the derivation from the BHPS information.

Table 3.3-1
Expected cross-sectional sampling errors from the sample design work, and the sampling errors achieved in wave 1 of the WAS

|  | Confidence Interval in \% of two-year cross-section |  |
| :--- | :---: | :---: |
|  | Design expectation | Achieved |
| Great Britain | 3.8 | 6.4 |
| Male | 4.1 | 7.6 |
| Female | 6.2 | 10.5 |
| 16 to 29 | 7.0 | 31.3 |
| 30 to 39 | 8.0 | 24.5 |
| 40 to 49 | 8.2 | 15.8 |
| 50 to 59 | 11.9 | 13.7 |
| 60 to 69 | 6.6 | 11.4 |
| 70 and above | 5.2 | 10.6 |

### 3.4 Discussion

Overall the design of the Wealth and Assets Survey worked as anticipated, which is interesting given the relative paucity of information from which the survey was designed. The sampling errors are larger than designed, but for accountable reasons. The use of the administrative data has facilitated the development of efficient sampling for WAS.

## References

Banks, J., Smith, Z. and Wakefield, M. (2002). The distribution of financial wealth in the UK: Evidence from 2000 BHPS data, working paper 0221, London: Institute for Fiscal Studies.

Butcher, B. and Elliot, D. (1992). A sampling errors manual London: Office of Population Censuses and Surveys.
Durrant, G.B. and Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys, Journal of the Royal Statistical Society, Series A, 172, pp. 361-381.

Kish, L. (1995). Methods for design effects, Journal of Official Statistics, 11, pp. 55-77.
Skinner, C.J. and Rao, J.N.K. (1996). Estimation in dual-frame surveys with complex designs, Journal of the American Statistical Association, 91, pp. 349-356.

## SESSION 4C

## LINKAGE OF HEALTH DATA

# Impact of identifying information requested on consent rates 

Jenna Fulton ${ }^{1}$


#### Abstract

As administrative records become more accessible, surveys are increasingly requesting the permission of respondents to access and link their administrative data. Access to these records can only occur with respondents' consent and, in most cases, in the U.S., respondents' agreement to provide linking information such as their Social Security Number. With evidence from several national surveys indicating that the public is becoming less willing to provide personallyidentifying information in the survey context, researchers are developing techniques to link without identifiers. There is potential for bias in the resulting linked dataset if some respondents fail to consent, and if those who consent are not representative of the entire sample. To date, no research has investigated the impact of the specific information requested to facilitate record linkage on consent rates, and how they might differ from surveys that do not request an identifier. Using a meta-analysis, this research will investigate the effect of the specific information requested on the proportion of respondents who give consent. This will compare surveys that request personal identifiers, such as Social Security Number and Medicare Number, as well as those which are able to link without requesting any identifying information. Existing surveys that request respondent consent to link records will be included in the analyses, for example, the Health and Retirement Survey, National Health Interview Survey, and Panel Survey of Income Dynamics. This research will help inform future linkage efforts by informing researchers which requests elicit the highest rates of consent to linkage.


[^37]
# When the Census and administrative data combine to illustrate social inequalities in health in Canada 

Denis Hamel, Robert Pampalon and Philippe Gamache ${ }^{1}$


#### Abstract

Most administrative databases in the health sector in Canada contain no socioeconomic information that allows social inequalities in health to be tracked. One way to get around this problem is to create a deprivation index at the level of the dissemination area. The methodological aspects of constructing such a two-dimensional index will be examined here, along with problems related to matching administrative databases and census data. Finally, we will show how the extent of health inequalities in Canada can vary depending on whether the deprivation is measured on the territorial or individual scale.


Key Words: Deprivation; census; social inequalities; area-based measurement; Canada.

## 1. Introduction

Health authorities at the ministerial level in Canada have always been interested in measuring social inequalities in health, particularly those resulting from individuals' socioeconomic status and place of residence (urban or rural). Since the late 1970s, surveys such as the Canada Health Survey, the National Population Health Survey (NPHS) and the Canadian Community Health Survey (CCHS) have helped to track these inequalities. But it is another matter with the administrative databases that provincial authorities have created to track vital statistics, such as mortality or the use of health services, including hospitalization and primary health care. In these databases, there is no socioeconomic information.

To fill this gap, researchers usually use geographic proxies. A geographic proxy is socioeconomic information based on small areas that is introduced into administrative databases, using keys for moving between the geography of these areas and the geography available in the database. This practice blossomed in Great Britain (Gordon, 2003) and spread to a number of other countries (Kunst et al., 1994; Krieger et al., 2002), including Canada (Mustard et al., 1999; Wilkins et al., 2002).

In this presentation, we propose a Canadian version of the deprivation index developed in Quebec (Pampalon et al., 2009a). We will essentially focus on the conceptual and methodological basis for the index. We will briefly examine the problem of matching between the census and administrative databases in the health sector and then illustrate the usefulness of this index using two examples: premature mortality in Canada in 2001 and a comparison between an individual measure and a geographic measure of deprivation in order to gauge disparities in life expectancy for a cohort of persons aged 25 and over in 1991.

## 2. Construction of a deprivation index

### 2.1 Concept of deprivation

It was Peter Townsend (Townsend, 1987) who, in the mid-1980s, proposed a definition of the concept of deprivation. For him, deprivation is "a state of observable and demonstrable disadvantage relative to the local community or the

[^38]wider society or nation to which an individual, family or group belongs." Townsend distinguishes between two main forms of deprivation: material and social. Whereas the former refers to the goods and conveniences of modern life, such as having an adequate home, a car, a television set or recreational space in the neighbourhood, the latter has to do with social relations within the family, at work and in the community.

### 2.2 Basic geographic unit

The deprivation index is based on a geographic unit of observation. Since the objective of the index is to take the place of an individual measure, the unit chosen must be as small as possible (Krieger et al., 2002) to ensure that there is substantial homogeneity in the socioeconomic conditions that will be imputed to each resident in the unit. The unit chosen is the dissemination area (DA) from the Canadian census, made up of one or more adjacent blocks, with a population of 400 to 700 .

To construct the index in 2006, a two-step process was followed. The first step was to exclude DAs with no population, DAs with a high proportion of collective households or institutionalized persons, DAs with no B profile (a socioeconomic profile) or income data (sparsely populated DAs), and DAs belonging to the territory of Nunavut or Indian reserves. This left 42,430 DAs, covering just over $93 \%$ of the Canadian population, on which to calculate the index. The second step was to project the values of the index to an additional number of DAs: DAs for which it was possible to impute an income value ( $3,572 \mathrm{DAs}$ ); DAs associated with Nunavut and Indian reserves-with a complete B profile or imputed income-( 857 DAs ); and DAs excluded because of their high proportion of collective households or institutionalized persons, but whose B profile (or imputed income) accounted for more than $85 \%$ of the total population ( 605 DAs ). When the projection was completed, a deprivation index could be established for 47,464 DAs, bringing the population covered by the deprivation index to nearly $98 \%$ of the total population of Canada in 2006.

Geographic units were determined in the same way for the previous three censuses: 1991, 1996 and 2001. For the 1991 and 1996 censuses, the concept of dissemination area does not exist, and the unit chosen was the enumeration area; on average, enumeration areas include slightly more individuals than DAs.

### 2.3 Choice of socioeconomic indicators

We chose indicators for constructing the index after reviewing literature and made our choice based on four criteria: having known relationships with various health services, previous use as a geographic proxy, affinities with the material or social dimension of deprivation, and availability by DA (Gordon, 2003; Marmot et al., 2003 ; Tello et al., 2005). This led us to identify six indicators for use in constructing the index. These indicators are the proportion of persons aged 15 and over with no high school diploma; the ratio of employment to the population aged 15 and over; average income of persons aged 15 and over; the proportion of persons aged 15 and over living alone in their home; the proportion of persons aged 15 and over who are separated, divorced or widowed; and the proportion of singleparent families.

The indicators selected sometimes vary greatly with the age and sex of the population. This is the case, for example, with education, which is not necessarily completed for youths under 20 years of age and which is not very high for a large number of seniors. Since the variations being sought are socioeconomic rather than demographic in nature, these indicators, except for Single-parent families, were adjusted according to the age and sex structure of the Canadian population (Salmond et al., 1998). Direct standardization was used. Additionally, some indicators were transformed to normalize their distribution (Gilthorpe, 1995). Thus, the income variable was transformed into its logarithm values and the variable Persons living alone into its arc sin values.

### 2.4 Indicator integration

The integration of the indicators in the form of a deprivation index was carried out using principal component analysis (PCA) (Tello et al., 2005; Salmond et al., 1998). This analysis results in a reduced number of dimensions, reflecting the spatial organization of the socioeconomic indicators (Table 2.4-1). A varimax rotation was applied to those dimensions to increase their readability and make them independent (or orthogonal). To validate the relevance
of this factor structure at the scale of Canada, PCA was repeated in the three largest census metropolitan areas (CMAs) (Toronto, Montréal and Vancouver), various geographic areas (other CMAs, census agglomerations [CAs], small towns and rural areas) and regions of Canada (Atlantic, Quebec, Ontario, the Prairies, and British Columbia). Deprivation measures perform differently in urban and rural areas.

Table 2.4-1
Principal components from the PCA with the six indicators selected in deprivation index according to DA of Canada in 2006

| Indicator | Component |  |
| :--- | ---: | ---: |
|  | Material | Social |
| No high school diploma | -0.82 | 0.03 |
| Employment/population ratio | 0.70 | -0.15 |
| Average personal income | 0.81 | -0.30 |
| Living alone | -0.02 | 0.84 |
| Separated, divorced, widowed | -0.18 | 0.88 |
| Single-parent families | -0.28 | 0.63 |

Variability explained by the components: material $=33 \%$; social $=32 \%$. Above values are factor loadings; they can be interpreted as correlation coefficients between an indicator and a component.

PCA produces a factor score for each component. This score represents the value of the component in each DA. To ensure that there is a degree of statistical precision in the analysis of social inequalities in health, DAs were grouped. They were then ordered on the basis of their factor score, from the most advantaged to the most disadvantaged. Then the distribution of the DAs was fragmented into quintiles, with each quintile representing $20 \%$ of the population. Quintile 1 is the least disadvantaged population, while quintile 5 is the most. These operations were performed separately for each component emerging from the analysis. Finally, since deprivation is seen as a relative disadvantage in relation to the community to which one belongs, different versions of the index were produced by modifying the reference area. Therefore, there is a national version, a version by major CMA, a version by geographic area and a version by region of Canada. These versions are based on the PCA carried out in each environment and over a distribution of factor scores, providing an equal distribution of the population ( $20 \%$ ) per material and social quintile. The use of any of these versions serves to reflect disparities in deprivation as they occur in each environment, and at the same time to compare numbers of similar proportion. Also, this same type of analysis was done for previous censuses: 1991, 1996 and 2001. The analyses revealed the presence of the same components with slight differences in the factor loadings associated with the six indicators.

## 3. Linkage between census and health databases

The deprivation index is determined according to the individual's DA of residence. The next step consists in introducing this deprivation index into the health databases. These do not include the DA, a key piece of information; instead, they contain the postal code as a geographic measure. To make this match, we therefore used Statistics Canada's postal code conversion file-commonly known as the PCCF+-as a tool. However, it should be noted that this tool has limitations and that assigning a dissemination area based on the postal code does not always yield a unique result, especially in rural areas. This can be a major problem if the focus is on a fine level of geography. But in our case, we want to assign the deprivation index in quintiles, and often adjacent DAs have identical deprivation quintile values. We therefore believe that this problem is more marginal at a more macro scale.

## 4. Illustrations with indicators of premature mortality and life expectancy

Approximately $94 \%$ of premature deaths in 2001 received a deprivation index, for a total of 85,614 deaths. Of the deaths that did not receive an index $(\mathrm{n}=5,625), 14 \%$ resulted from erroneous codes and $86 \%$ from DAs with no index, in many cases corresponding to institutional populations.

The adjusted premature mortality rate is 310 deaths per 100,000 persons in Canada in 2001. The mortality rate rises continually with both material and social deprivation (Figure $4-1$ ). The mortality ratio between the extreme groups for material and social deprivation is 2.41 , and the difference in mortality is 302 deaths per 100,000, a value equivalent to that observed in Canada as a whole.

Figure 4-1
Premature mortality rate by quintile of material and social deprivation, Canada, 2001


Since the deprivation index is an ecological-type index and is seen as an individual proxy for deprivation, we wanted to validate it by comparing it with an individual measure of deprivation (Pampalon et al., 2009b). The recent availability of a file matching a sample of the 1991 Canadian census with a follow-up of mortality from 1991 to 2001 provides a unique opportunity to explore how the individual and geographic versions of the deprivation index respectively contribute to health indicators (Wilkins et al., 2008). The next illustration concerns life expectancy at age 25 between the material and social deprivation extremes according to geographic area and version of the index (Figure 4-2).

Regardless of whether our calculations are based on individual or geographic data, all geographic areas exhibit deprivation-related disparities in life expectancy at age 25. However, the magnitude of the inequalities observed with respect to material deprivation varies according to the version of the index used. Thus, disparities in life expectancy at age 25 that are based on the individual version of the index are greater in census agglomerations (8.6 years) as well as in small towns and rural areas ( 7.4 years) than in the large census metropolitan areas (Toronto, Vancouver and Montréal CMAs) ( 5.4 years) and other CMAs ( 6.4 years). However, according to the geographic version of the index (at the enumeration area level), "other CMAs" register the largest gap in life expectancy.

Figure 4-2
Difference in life expectancy at age 25 between highest and lowest quintiles of material and social deprivation, by version of index and geographic setting, Canada, 1991 to 2001

$\dagger$ Material deprivation (Q1 minus Q5)
\# Social deprivation (Q1234 minus Q5)
Note: Major CMA (census metropolitan area): Toronto, Vancouver and Montréal; Other CMAs: population of 100,000 or more; CAs (census agglomeration): population of 10,000 or more.

Not surprisingly, social inequalities in health, at least for life expectancy, are greater with an individual measure. Thus, the deprivation index on a geographic basis appears to explain only a part of the social inequalities due to individuals (composition effect). Other results, not presented here, show that the relationships remain significant for the deprivation index on a geographic basis in the presence of the individual measure, which suggests that the geographic index is also expressing something else (context effect) (Pampalon et al., 2009b).

## 5. Conclusion

Thanks to census data at the dissemination area level, we are able, by principal component analysis, to group six socioeconomic indicators into two components: material and social. These same components appear when the analysis is limited to dissemination areas belonging to a specific geographic area in Canada, and yield a profile of deprivation that is specific to this environment. The same exercise was carried out for four census years, from 1991 to 2006. Therefore, there are several possible versions of the index, depending on the census year and the geographic area.

This index can be assigned to administrative databases in the field of health that contain no socioeconomic information by using Statistics Canada's PCCF+. This tool links the postal code present in the health databases to the dissemination area. The deprivation index may then be seen as a geographic proxy for deprivation.

We found that variations in the deprivation index are closely linked to variations in premature mortality in Canada in 2001. Each form of material and social deprivation contributes independently to mortality, and this contribution gradually rises with the level of deprivation.

The deprivation index has its limitations. It is not an individual measure of socioeconomic conditions, but rather a measure of the conditions encountered in the neighbourhood. It does not take into account how all socioeconomic conditions, such as immigration and Aboriginal status, impact health. It also does not differentiate the respective roles of each of the indicators that comprise it. The index nevertheless proves to be an excellent geographic marker of deprivation. It is linked not only to the characteristics of the population (composition effect), but also to individuals' immediate living environment (context effect).

Finally, it should be noted that there are several products accessible on the bilingual website Santéscope (http://www.inspq.qc.ca/santescope/liens.asp?Lg=en\&NumVol=undefined\&nav=M\&comp=9). Among other things, this website includes, for a given version of the deprivation index, an SAS assignment program for the index, as well as tables of equivalence between DAs and indexes, and articles published on the subject.

## Bibliography

Gilthorpe, M.S. (1995). The importance of normalisation in the construction of deprivation indices, $J$ Epidemiol Community Health, 49, Suppl 2, pp. S45-S50.

Gordon, D. (2003). Area-based deprivation measures: A U.K. perspective, I. Kawachi and L. Berkman (ed.), Neighborhoods and health, New York: Oxford University Press, p. 179-210.

Krieger, N., Chen, J.T., Waterman, P.D., Soobader, M.J., Subramanian, S.V. and Carson , R. (2002). Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of areabased measure and geographic level matter?: the Public Health Disparities Geocoding Project, American Journal of Epidemiology, 156(5), pp. 471-82.

Kunst, A.E. and Mackenbach, J.P. (1994). Measuring socioeconomic inequalities in health, Copenhague, WHO.
Marmot, M. and Wilkinson, R.G. (ed.) (2003). Social Determinants of Health. $2^{\text {ième éd, Copenhague, Organisation }}$ mondiale de la santé.

Mustard, C.A., Derksen, S., Berthelot, J.M. and Wolfson, M. (1999). Assessing ecologic proxies for household income: a comparison of household and neighbourhood level income measures in the study of population health status , Health Place, 5(2), pp. 157-171.

Pampalon, R., Hamel, D., Gamache, P. and Raymond, G. (2009a). Un indice de défavorisation pour la planification de la santé au Canada, Maladies chroniques au Canada, 29(4), pp. 191-208.

Pampalon, R., Hamel , D. and Gamache, P. (2009b). A comparison of individual and area-based socio-economic data for monitoring social inequalities in health, Health Reports, 20(4):85-94.

Salmond, C., Crampton, P. and Sutton, F. (1998). A New Zealand index of deprivation, Aust N Z J Public Health, 22(7) pp. 835-837.

Tello, J.E., Jones, J., Bonizzato, P., Mazzi, M., Amaddeo, F. and Tansella, M. (2005). A census-based socioeconomic status (SES) index as a tool to examine the relationship between mental health services use and deprivation, Social and Science in Medecine, 61(10), pp. 2096-2105.

Townsend, P. (1987). Deprivation, J Soc Pol., 16, pp. 125-146.
Wilkins, R., Berthelot, J.M. and Ng, E. (2002). Tendances de la mortalité selon le revenu du quartier dans les régions urbaines du Canada, 1971-1996, Rapports sur la santé, Supplément, 13, pp. 1-28.

Wilkins, R., Tjepkema, M., Mustard, C. and Choinière, R. (2008). Étude canadienne de suivi de la mortalité selon le recensement, 1991 à 2001, Rapports sur la santé, 19(3) : 27-48

# Indigenous life expectancy using multiple Australian data sources 

Richard Madden, Leonie Tickle, Lisa Jackson Pulver, Ian Ring and Lee Taylor ${ }^{1}$


#### Abstract

In 2009, the Australian Bureau of Statistics (ABS) released new estimates of Indigenous life expectancy for Australia. The estimates, which were substantially higher than previously published estimates, were based on linkage between Indigenous deaths registered in the period from August 7, 2006 to June 30, 2007 and 2006 census records. State estimates were also produced for some states, showing substantial variations between states; the life expectancies are inversely related to the calculated completeness of Indigenous identification of death registrations.

Analysis based on more comprehensive linkage of death records in New South Wales (NSW) over 5 years suggests that the ABS methods have understated Indigenous deaths and so overstated life expectancy. The paper will report the NSW results, based on several linkage algorithms, including a comparison of the algorithms. Resulting changes in life expectancy estimates will be reported. Suggested improvements to ABS methods will be discussed.


[^39]
# 1991-2011 Canadian Census Mortality and Cancer Follow-Up Study 

Paul A. Peters and Michael Tjepkema ${ }^{1}$


#### Abstract

Census mortality linkages are proven to be powerful tools for analysing the mortality differences for numerous population groups. In an recently approved record linkage, the 1991 Census of Population, Canadian Mortality Database, and Canadian Cancer Database will be linked in order to examine cancer incidence and causes of death in conjunction with sociodemographic and neighbourhood characteristics. The linkage of the 1991 Census cohort to these databases will allow for the analysis of mortality using the CMDB in conjunction with the extensive information from the 1991 Census of Population long-forms ( 2 B and 2D), the recording of individual mobility over time using postal codes of tax filers from the Tax Summary Files, and the inclusion of important analysis of cancer morbidity via the CCDB. This presentation overviews the previous census mortality linkage, describe the new linkage, outline the linkage process, and present some initial linkage results.


Key Words: Census; Health; Mortality; Cancer; Data Linkage.

## 1. Introduction

This paper describes the development of and preliminary results from the linkage of the 1991 Canadian Census 2B (long form) with the Canadian Mortality Database (CMDB), Canadian Cancer Database (CCDB), and annual Tax Summary Files (TSF) (record linkage number 052-2009). This linkage is an expansion and extension of the earlier 1991-2001 Canadian Census Mortality Follow-Up Study (Wilkins et. al., 2008). The linkage of the 1991 Census cohort to these three databases allows for the analysis of mortality using the CMDB in conjunction with the extensive information from the 1991 Census, the recording of individual mobility over time, and the inclusion of important analysis of cancer morbidity.
The primary purpose of this expansion and extension is to assess the impact of long-term exposure to air pollution on human health, with the objective of informing the development of Canada-wide standards for key criteria air pollutants. The specific objectives of this expanded linkage are first, to determine whether mortality from all causes combined, from ischaemic heart disease, from cardiopulmonary disease, from respiratory cancer, and from all cancers combined are associated with long-term exposure to ambient air pollutants. Second, to determine whether there are air pollution risks for cancer incidence and the risks for specific types of cancer. Third, to extend the 10year follow-up on the 1991 Census cohort in order to examine cancer incidence and causes of death in conjunction with socio-demographic, disability, and neighbourhood characteristics over an additional 10-year period.

There is a recognised need for more environmental data related to human exposure, with linkages of separate sources of information identified as an important way in which Statistics Canada can meet identified data gaps (Statistics Canada, 2008). In particular, findings from a recent health-environment expert panel report noted that it is critical to investigate the health of Canadians who live, work, or are educated near sources of pollution. This linkage request addresses these key recommendations with the development of a significant baseline cohort that could be used to evaluate the risk factors of environmental exposure on human health outcomes.

[^40]
## 2. Background and Purpose

### 2.1 Background

The 1991-2001 Census Mortality Follow-Up Study forms the basis for the current expanded linkage. The 1991-2001 linkage created a cohort of individuals with the goal of producing indicators of mortality for monitoring health disparities across regions and among socio-economic groups, Aboriginal peoples, and immigrant groups in Canada (record linkage number 012-2001). The linkage included the 1991 Census of population (long forms), 1991 Health and Activity Limitation Survey (HALS), the 1990 and 1991 TSF, and mortality data between 1991 and 2001 from the CMDB. The 1991-2001 linkage yielded a cohort of approximately $15 \%$ of the Canadian non-institutional resident population age 25 and older in 1991.

This sample of $2,735,152$ Census respondents was first linked via probabilistic record linkage to the 1990 and 1991 T1 personal tax files, matching date of birth, sex, marital status and postal code on the Census and tax files to obtain a name from the TSF to use in the linkage to death registrations in the CMDB from 1991 to 2001. To preserve respondent anonymity, names were encrypted on the T1 and the CMDB prior to linkage. The Social Insurance Number (SIN) from the tax file was not used in the linkage and only tax deaths and emigration from the TSF were put into the analysis file and used to calculate person-years at risk.

The intention of original linkage of the 1991 Census cohort to mortality was to create a set of baseline indicators of mortality, in order to monitor health disparities in the Canadian population. Initial work on the health effects of ambient air pollution using the existing linked analysis file is underway via a joint project of Statistics Canada and Health Canada researchers. However, there are several limitations to the 1991 Census cohort linked file. First, it does not include place of residence information for respondents on an annual basis after the 1991 Census. This information is required in studies of air pollution, to measure the residential mobility of the population, as air pollution is known to vary significantly at the local level (Jerrett et. al., 2009).

Second, the analysis file is limited in scope given the number of deaths which occurred in this cohort (260,820 deaths from 1991 to 2001). A larger number of deaths are required in order to obtain sufficient statistical power to accurately measure the long-term effects of air pollution (Krewski et. al. 2009). As such, in order to examine the effect of air pollution on the risk of mortality and cancer incidence a longer follow-up period is required.
Third, the existing linkage does not include the CCDB, so no information is available on cancer incidence in this cohort. Few studies have examined the risks of long-term exposure to ambient air pollution on cancer incidence and mortality. In particular, the proposed linkage to the CCDB will allow for analysis of cancer incidence rather than only deaths due to cancer.

### 2.2 Purpose

The primary purpose for the expanded linkage is to provide a cohesive dataset for analyzing the health effects of long-term exposure to air pollution on Canadians. A secondary use of the linked data base will be to examine the relationship of socio-demographic and neighbourhood characteristics to cancer incidence and causes of death.
Extension of the 1991 Census cohort follow-up linkage will allow for a substantial increase in the number of deaths over the longer time period, and the expansion of the linkage to annual tax data will result in a reduction in potential misclassification of exposure to air pollution. Since air pollution mortality risks are small, on the order of $10 \%$ increased risk over the exposure range seen in Canada, a large cohort size is required to obtain sufficient statistical power to detect such effects. The mortality follow-up on the 1991 Census cohort to 2011 is expected to more than double the number of deaths included in the estimates of years of life lost. This additional power will allow for the identification and analysis of susceptible population sub-groups who may have different levels of risk. Regional risk estimates can also be calculated and potential factors identified that correspond to different levels of risk between geographic regions.

In addition, there has been a considerable improvement in the monitoring of air pollution since 2000. New techniques of measuring exposure using satellite imagery or land-use regression modelling allow for significantlyimproved estimates of exposure. The combination of these methods with the large cohort included in the linked
database allows for a wider geographic coverage for exposure estimates and the ability to link information on a greater number of individuals to air pollution estimates.

A key component of this linkage is the inclusion of place of residence from the Tax Summary Files, permitting the recording of Census respondents' post-censal mobility. This is important because exposure to ambient air pollution is assigned based on the respondent's place of residence. Research in Canada has shown that air pollution risks vary at the local level (Burnett et. al. 1998). Thus, detailed place of residence information is required to accurately attribute exposure values to individuals. Research in the US has shown that air pollution-related mortality risks vary widely across that country (Krewski et. al. 2009). Thus, estimates of regional risk in Canada are also expected to have substantial variation. The TSF provide a means of obtaining the six-digit postal code of Census respondents on an annual basis after 1991.

### 2.3 Benefits / Public Good

It is now widely recognized that exposure to outdoor air pollution generated from combustion sources poses a public health risk to Canadians (Koranteng et. al., 2007; Samoli et. al., 2008). However, much of the scientific evidence is focussed on linking short-term or acute exposure to the exacerbation of health problems. For example, exposure to elevated daily outdoor concentrations of combustion-related pollution, such as ground level ozone and fine particulate matter, have been linked to increased asthma symptoms (Burra et. al., 2009), visits to emergency departments for heart, lung, and circulatory problems (Szyszkowskz, 2008; Szyszkowskz et. al., 2009), hospital admissions (Steib et. al., 2002; Villeneuve et. al., 2006) and mortality (Burnett et. al., 1998; 2000; 2003). Much of the evidence of acute health effects on Canadians has been obtained through studies of hospital admission and mortality data conducted in a continuing collaborative effort between Health Canada and Statistics Canada since 1992. However, detailed cohort studies in Canada have been limited by the data sources available and have often relied on extrapolation to the Canadian context from US sources.

Several cohort studies of mortality and air pollution have been conducted in the United States. The largest American study is the Cancer Prevention Cohort II study conducted by the American Cancer Society (ACS) (Pope et. al., 1995). In that study, 1.1 million subjects were interviewed in 1982 and followed for vital status to the present. The results of that study and the extended re-analysis show links between long-term fine particulate exposure and deaths from ischaemic heart disease and lung cancer (Krewski et. al., 2009). Importantly, other cohort studies in the US show that long-term exposure to fine particulate matter is associated with mortality risks ten times greater than that of short-term exposures, based on time-series studies. A paper published recently in the New England Journal of Medicine using the ACS study demonstrates for the first time that long-term exposure to ozone is linked to increased deaths attributed to respiratory causes (Jerrett et. al., 2009). The results of studies using the ACS study have led the US Environmental Protection Agency to promulgate a national annual average standard, in addition to their 24-hour standard which safeguards against extreme short-term exposures (U.S. EPA, 2004).

To date, Canada-wide standards for annual averages of either fine particulate matter or ozone have not been developed, largely due to lack of evidence from the Canadian population and uncertainties about transferring risks observed in the US to Canada (Raisenne, 2003). A recent multi-national collaborative study between Canada, the US, and Europe supports this concern of transferability, where the association between short-term air pollution exposure and mortality is examined (Samoli et. al., 2008). The results of that study indicate that risk estimates based on Canadian populations are twice as large as those based on US and European studies, highlighting the importance of obtaining direct estimates of risk on Canadians.

In addition, information on the risk of development of cancer in the general population due to ambient exposure to carcinogens from fossil fuel combustion and other sources is not well defined. While fine particulate air pollution from combustion can be linked to deaths from lung cancer, the relationship to incidence of cancer is not well understood (Krewski et. al., 2009). This expanded cohort linkage study provides important information on whether exposure to combustion-related ambient pollution plays a role in developing cancer in the Canadian population. Linkage of the cohort to the CCDB will provide the means to assess the risk of cancer incidence.

## 3. Linkage Methodology

The basis for the linkage is the 1991 Census of Population 2B and 2D. In the 1991 Census of Population, one in five households (20\%) in non-remote areas of Canada received a Census 2B (long form) questionnaire. The Census 2D (long form) "Canvasser questionnaire" was used to enumerate all households in remote northern areas of Canada and on Indian reserves. The 2 B and 2D questionnaires contain all the questions from the Census 2 A (short form) plus additional questions on topics such as education, ethnicity, mobility, income, employment and dwelling characteristics. In 1991, the total population who completed the long-form questionnaires was about 3.6 million individuals. The 1991 Census cohort used for this linkage study includes only individuals aged 25 and older.
To facilitate the linkage, TSF are used as a "name file" to improve data linkage as a bridge between individuals on the Census, the CMDB, and the CCDB (Figure 3.1). Here, the TSF from 1991 to 2007 were also used to create a mobility and migration component. The TSF is an annual file derived from personal tax returns filed for the year of reference. For this study, TSF were used to improve data linkage, provide alive follow-up, and record mobility where available using the postal code.

Figure 3-1
Linkage methodology flowchart for 1991-2011 Canadian Census Mortality and Cancer Follow-Up Study


Mortality information was drawn from the CMDB, which contains demographic and medical (cause of death) information on deaths registered by all provincial and territorial vital statistics registries in Canada. The CMDB comprises all deaths that occur within Canada, as well as deaths of Canadian residents reported by some US states. Cancer incidence was taken from the CCDB, which contains diagnosed incidences of cancer reported for all individuals whose usual place of residence is Canada or who are non-permanent residents. For cancer analysis, it is important to begin the study period (1991 to 2011) with a group of disease-free (no cancer diagnosis) individuals. To accomplish this, record linkage of the cohort to the cancer incidence data from 1969 to 2006 will be used to flag the records of Census respondents who were diagnosed with cancer before 1991.

On the 1991 Census cohort file, Statistics Canada assigned a randomly-generated Statistics Canada number to each respondent and the linkage variables were pre-processed so that they are compatible with variables on the tax, cancer, and mortality databases. The cohort file was then linked to TSF from 1990 to 2007, and variables from the tax files were appended to the cohort file for use in the mortality linkage and for retention in the migration output
file. Each year of the TSF was used to identify additional deaths in the cohort not ascertained by the CMDB. Income data were not retained on this file.

The cohort file will be linked to the 1969 to 2006 CCDB using probabilistic linkage techniques to select the best of several linkages. A cancer output file will be prepared, containing the randomly-generated Statistics Canada respondent number and cancer incidence information for those in the 1991 Census cohort whose records linked to the CCDB. Names and other identifiers, such as cancer registry number, will not be included in this file.

### 3.2 Analytic Methodology

The methodology for analysis of ambient air pollution will be similar to that being used in an on-going project, in which the 1991 Census cohort linked to the 1991 to 2001 mortality data is analysed using ambient air pollution concentrations from various geographic locations across Canada. The analysis planned for the expanded linkage to the 1991 Census cohort will develop time-varying estimates of outdoor pollution concentrations for sulphur dioxide, nitrogen dioxide, and fine and coarse particulate matter across the study period (1991-2011).

The models used to estimate exposure to air pollution will include use of such information as fixed-site ambient monitoring data, land-use characteristics, satellite information, visibility at airports, proximity to major roads, and traffic counts. Fixed-site measurements from the National Air Pollution Surveillance (NAPS) Network are not available for every pollutant for every year between 1991 and 2001, but are available for most areas beyond 2001. Where necessary, exposure values will be estimated via a combination of land use regression modelling, satellite interpolation, and kriging surfaces (Lamsal et. al., 2010; Ontario Medical Association, 2005). Sensitivity analysis will be conducted on each exposure model to evaluate the model strength (Krewski et. al., 2009).

Relating mortality risk factors and longevity requires the use of survival models. Analysis of the 1991 Census cohort will use newly-developed computer software which implements a Cox proportional-hazards survival model with spatially auto correlated random effects at several levels of geographic nesting, such as province, community, and neighbourhood, with each level incorporating a spatial autoregressive error process. A detailed explanation of the methodology can be found in Krewski et. al. (2009).

## 4. Discussion

Census respondents were eligible for the cohort if they were usual, non-institutional residents of Canada on the day of the Census, were in the long-form Census records, and had reached age 25 by Census day. These individuals are those considered "in-scope" in Table 4-1. However, only those linked to a name file could be reliably followed for mortality, where approximately $20 \%$ could not be linked. As shown below, $2,860,244$ individuals were eligible to be part of the cohort. To reduce the final cohort to equal $15 \%$ of the Canadian population, a random sample of $4.4 \%$ were removed, leaving a final cohort size of $2,734,835$ persons.

Of the cohort, 426,979 died during the follow-up period, where 17,268 deaths were ascertained only by tax records and thus do not have any associated information on the cause, location, or nature for mortality. From the TSF, $96.7 \%$ of the cohort was followed for mobility between 1990 and 2007, although if individuals did not file taxes for each year during this period, missing values for place of residence will appear for each corresponding year.

In the absence of large-scale nationally-representative longitudinal health surveys, cohorts datasets such as that described here allow for the study of numerous aspects of the population-health inequalities and outcomes, particularly as related to environmental exposures. The first census mortality linkage has resulted in over a dozen major publications on a range of topics. The expanded cohort created by the addition of cancer incidence, mobility, and additional years of mortality will facilitate numerous additional analyses of key health outcomes by populationbased socio-economic indicators and individual characteristics.

Table 4-1
Derivation of the cohort from 1991 Census records, Canada

|  | Number |
| :--- | ---: |
| Derivation of cohort |  |
| In-scope census records $^{1}$ | $3,576,487$ |
| Not linked to name file | 716,243 |
| Linked to name file (1990 and 1991 TSF) | $2,860,244$ |
| Linked to name file but not followed for deaths | 125,409 |
| Linked to name file and followed for deaths (the cohort) ${ }^{\mathbf{2}}$ | $2,734,835$ |
| Died during the follow-up period | 426,979 |
| Deaths ascertained by CMDB | 409,711 |
| Deaths ascertained by tax only | 17,268 |
| Followed for mobility from tax summary files (1991-2007) | $96.7 \%$ |
|  |  |
| Percentage of population | $18,225,349$ |
| 1991 mid-year population estimate for population aged 25 and older ${ }^{3}$ | $15.0 \%$ |
| Cohort as a percentage of the population aged 25 and older |  |

${ }^{1}$ Non-institutional residents of Canada aged 25 or older with long-form questionnaire
${ }^{2}$ Random sample of $4.4 \%$ of those linked to name file
${ }^{3}$ CANSIM table 051-0001/3604
Source: Canadian Census Mortality and Cancer Follow-Up Study, 1991-2006

## References

Burnett, R.T., Brook, J., Dann, T. et al. (2000). Association between particulate- and gasphase components of urban air pollution and daily mortality in eight Canadian cities. Inhalation \& Toxicology, 12(suppl 4), pp. 15-39.

Burnett, R.T., Cakmak, S. and Brook, J.R. (1998). The effect of the urban ambient air pollution mix on daily mortality rates in 11 Canadian cities. Canadian Journal of Public Health, 89, pp. 152-156.

Burnett, R.T. and Goldberg, M.S. (2003). Size-fractioned particulate mass and daily mortality in eight Canadian cities. Revised analysis of time-series studies of air pollution and health, Boston, MA: Health Effects Institute, pp. 85-89.

Burra, T.A., Moineddin, R., Agha, M.M. and Glazier, R.H. (2009). Social disadvantage, air pollution, and asthma physician visits in Toronto, Canada. Environmental Research, 109, pp. 998-1003.

Statistics Canada (2008), Health environment linkages expert panel report to Statistics Canada, Environment Accounts and Statistics Division, December 19, 2008.

Jerrett, M., Burnett, R.T., Pope, C.A. III et al. (2009). Long-term ozone exposure and mortality, New England Journal of Medicine, 360, pp. 1085-1095.

Jerrett, M., Finkelstein, M., Brook, J.R. et al. (2009). A cohort study of traffic-related air pollution and mortality in Toronto, Ontario, Canada. Environmental Health Perspectives, 117(5), pp. 772-777.

Koranteng, S., Osornio Vargas, A.R. and Buka, I. (2007). Ambient air pollution and children's health: A systematic review of Canadian epidemiological studies. Paediatric and Child Health, 12, pp. 225-233.

Krewski, D., Jerrett, M., Burnett, R.T. et al. (2009). Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Boston, MA: Health Effects Institute.

Lamsal, L.N., Martin, R.V., van Donkelaar A. et. al. (2010). Indirect validation of tropospheric nitrogen dioxide retrieved from the OMI satellite instrument: insight into the seasonal variation of nitrogen oxides at northern midlatitudes. Journal of Geophysics Research, 115.

Ontario Medical Association (2005). The Illness Costs of Air Pollution: 2005-2026, Health and Economic Damage Estimates, Toronto: Ontario Medical Association.

Pope, C.A. III, Thun, M.J., Namboodiri M.M. et al. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. American Journal of Respiratory and Critical Care Medicine, 151, pp. 669674.

Raisenne, M. (2003). Science and regulation - U.S. and Canadian overview. Journal of Toxicology and Environmental Health, Part 4, 66, pp. 1503-1506.

Samoli, E., Peng, R., Ramsay, T., et al. (2008). Acute effects of ambient particulate matter on mortality in Europe and North America: Results from the APHENA study. Environmental Health Perspectives, 116, pp. 14801486.

Stieb, D.M., Smith-Doiron, M., Brook, J.R. et al. (2002). Air pollution and disability days in Toronto: results from the National Population Health Survey. Environmental Research Section A, 89, pp. 210-219.

Szyszkowicz, M., Rowe, B.H. and Kaplan, G.G. (2009). Ambient sulphur dioxide exposure and emergency department visits for migraine in Vancouver, Canada. International Journal of Occupational Medicine and Environmental Health, 22, pp. 7-12.

Szyszkowicz, M. (2008). Ambient air pollution and daily emergency department visits for ischemic stroke in Edmonton, Canada. International Journal of Occupational Medicine and Environmental Health, 21, pp. 295-300.
U.S. Environmental Protection Agency (2004). Air quality criteria for particulate matter EPA 600/P-99/002aF-bF. Washington DC: U.S. Environmental Protection Agency.

Villeneuve, P.J., Chen, L., Stieb, D. and Rowe, B.H. (2006). Associations between outdoor air pollution and emergency department visits for stroke in Edmonton, Canada. European Journal of Epidemiology, 21, pp. 689-700.

Wilkins, R., Tjepkema, M., Mustard, C. and Choinière, R. (2008). The Canadian census mortality follow-up study, 1991 through 2001. Health Reports, 19, pp. 25-43.

## SESSION 5A

## WAKSBERG AWARD WINNER ADDRESS

# The organization of statistical methodology and methodological research in national statistical offices 

Ivan Fellegi ${ }^{1}$


#### Abstract

The paper explores and assesses the approaches used by statistical offices to ensure effective methodological input into their statistical practice. The tension between independence and relevance is a common theme: generally, methodologists have to work closely with the rest of the statistical organisation for their work to be relevant; but they also need to have a degree of independence to question the use of existing methods and to lead the introduction of new ones where needed. And, of course, there is a need for an effective research program which, on the one hand, has a degree of independence needed by any research program, but which, on the other hand, is sufficiently connected so that its work is both motivated by and feeds back into the daily work of the statistical office.

The paper explores alternative modalities of organisation; leadership; planning and funding; the role of project teams; career development; external advisory committees; interaction with the academic community; and research.


Dr. Fellegi's complete paper can be found in Survey Methodology December 2010.

[^41]
## SESSION 6A

NEW DEVELOPMENTS IN RECORD LINKAGE METHODS

# Glink - A Probabilistic Record Linkage System 

Antoine Chevrette and Michael Wenzowski ${ }^{1}$


#### Abstract

It is commonly the case that data originating from disparate sources, and even from the same source over time, exists in incompatible formats. The consequence is that the analysis of such data must often be preceded by some form of record linkage operation in order to assemble the data into a well-structured form. When reliable and invariant identifiers are available, such linkages become a relatively easy task to perform. In other cases, the identifier values may vary, requiring the reformatting of the data and possibly the execution of a fuzzy match. In both of these scenarios, "off the shelf" software is readily available that is well-suited to the task at hand. However, in cases in which a greater degree of variance must be accommodated, especially in cases in which multiple fields must be examined in order to identify a linkage, the task falls to software capable of performing a sophisticated and complex probabilistic record linkage. Unfortunately, software capable of this degree of sophistication is very highly specialized and only sparsely available.


We present the results of a recent Statistics Canada initiative to re-engineer our generalized record linkage system in order to enhance its applicability across a wide range of processing problem and subject matter domains. The software faithfully implements the probabilistic record linkage methodology first described by Fellegi and Sunter ${ }^{2}$, and includes many extensions and enhancements to increase the utility of the application. We will demonstrate how we have improved the software by offering more intuitive controls over managing the complexity of internal processing; by extending and enhancing the software's capabilities; and by simplifying the installation, setup and processing models.

[^42]
# Record linkage methods and techniques as proposed in RELAIS 

Nicoletta Cibella, Marco Fortini, Daniela Ichim, Tiziana Tuoto ${ }^{1}$


#### Abstract

Nowadays in official statistics, bringing together data from different sources is largely widespread in a context of increasing demand of statistical information and stricter budgetary constraints. Record linkage techniques are a multidisciplinary set of methods and practices aiming at identifying the same real world entity at individual micro level. In this paper we propose the RELAIS system. The basic idea of RELAIS is to face the linkage complexity by decomposing the whole problem in its constituting phases and to dynamically adopt the most appropriate technique for each singled step. RELAIS is an open source project. The methodological core of RELAIS is based on the well-known Fellegi-Sunter theory, allowing its usage from both researchers and non-experts.


Key Words: software for record linkage, probabilistic record linkage, open-source software.

## 1. Introduction

Nowadays in official statistics, bringing together for statistical purpose huge amount of data coming from different sources is largely widespread in a context of increasing demand of statistical information, on one side, and stricter budgetary constraints on the other side. Indeed, the joint analyses of statistical and administrative data allow saving time and money by reducing survey costs, response burden, etc. Record linkage techniques are a multidisciplinary set of methods and practices aiming at identifying the same real world entity at individual micro level, even when differently represented in data sources. The complexity of the whole linkage process is due to many issues. For example, the lack of unique identifiers requires sophisticated statistical procedures, the huge amount of data to process involves complex IT solutions, constraints related to a specific application may require the solution of difficult linear programming problems. Record linkage techniques are extensively used in medicine, computer science and, of course, in official statistics. Anyway, despite this proliferation, no particular record linkage technique has emerged as the best solution for all cases. We believe that such a solution does not actually exist, and that an alternative strategy should be found: in this paper we propose the RELAIS (Record Linkage At Istat) system. The basic idea of RELAIS is to face the record linkage complexity by decomposing the whole problem in its constituting phases and to dynamically adopt the most appropriate technique for each step, in order to define the most suitable strategy depending on application and data specific requirements. RELAIS is configured as an open source project, a winning choice for sharing techniques and software. The methodological core of RELAIS is based on the wellknown Fellegi-Sunter theory, allowing its usage by both researchers (who may easily enrich it) and non-experts (who have it embedded in the software).

The paper is organized as follows. In Section 2, we describe the general problem of record linkage, while in Section 3 the idea, the design and the current state of implementation of RELAIS are delineated. In Section 4, several record linkage projects of ISTAT in which RELAIS is adopted are described. Finally, in Section 5 there are some concluding remarks and directions for future works.

[^43]
## 2. The record linkage problem

Record linkage techniques are a multidisciplinary set of methods and practices with the main purpose of accurately recognize the same real world entity at individual micro level, even when differently stored in sources of various type. Record linkage is also know as object identification, record matching, entity matching, entity resolution, reference reconciliation. Furthermore, the overall record linkage workflow could change from user to user, due to different restrictions, such as legal and practical issues, in various fields and countries. Even in a statistical system with shared goals and regulations, as the European Statistical System, different constraints, for instance based on language features, may be present and affect the outcome of the same linkage. There are different purposes to perform a record linkage project and it has recently revealed a powerful support to decisions in large commercial organizations and government institutions. In official statistics data integration procedures are becoming extremely important due to many reasons: the cut of the costs, reduction of response burden and use of information derived from administrative data are some of the most crucial ones; this is a strong incentive to the investigation of new methodologies and instruments to deal with record linkage projects.

Several applications need record linkage techniques, including: enriching and updating the information stored in sources; the elimination of duplicates within a data frame; the creation of a sampling list; improving the data quality of a source; estimating number of units in a population amount by capture-recapture method; assessing the disclosure risk when releasing microdata files; the study of the relationship among variables reported in different sources.
In particular, according to the privacy laws, a disclosure occurs when a unit is identified and/or confidential information about a unit may be retrieved. In statistical disclosure control, record linkage methodologies may be used to derive several measures of disclosure risk based on external registers scenarios, i.e. making assumptions on how an intruder could identify units in a released microdata file. Generally speaking, the disclosure scenario models the uncertainty on intruder's information (data, tools and knowledge). Record linkage techniques may be used to assess both the risk of disclosure of the confidential microdata file and the efficiency of the applied protection method (Shlomo, 2009; Skinner, 2008).

Since the earliest contributions to modern record linkage, dated back to Newcombe et al (1959) and to Fellegi and Sunter (1969), there has been a proliferation of different approaches, that make use also of techniques based on data mining, machine learning, equational theory. However no particular record linkage technique has emerged as the best solution for all cases. Additionally, in some applications, there is no evidence to prefer one method to others or of the fact that different choices, at a linkage stage, could bring to the same results. As an intuitive example, alternative reduction methods will lead to different pair search space and consequently different results; as a further example, if no specific error-rates are required by the application, it can be appropriate the usage of an empirical decision model. Also the choice of which decision model to apply is not immediate: the usage of a probabilistic decision model can be more appropriate for some applications but it can be less appropriate for others, for which an deterministic decision model could prove more successful. We believe that there isn't a best solution for all cases and that an alternative strategy should be adopted: it could be reasonable to dynamically select the most appropriate technique for each phase of the linkage problem and to combine the selected techniques for building an overall strategy. In addition, from the analyst's point of view, it is important to have the possibility to experiment alternative criteria and parameters in the same application scenario.

The unit identification seems very hard to achieve in absence of unique identifiers or when the variables are affected by errors. So record linkage can be seen as a complex process consisting of several distinct phases involving different knowledge areas and the choice of the most appropriate technique does not depend only on the practitioner's skill but it is also application specific.

The main phases in which a record linkage problem can be decomposed are: pre-processing/preparation of the input files; creation-reduction of the search space of link candidate pairs; choice of the common identifying attributesmatching variables; choice of comparison functions; choice of decision model; identification of unique links; record linkage procedures evaluation.
Generally speaking, the complexity of a linking process relies on several aspects. As mentioned above, if unique identifiers are available in the data sources the problem can be quite easily treated but unique identifiers are not always available and more sophisticated statistical procedures are required. Obviously, errors in the linking variables may invalidate the linkage results, thus a big effort for reducing such errors is necessary to prepare input files; such a phase, according to Gill (2001), requires $75 \%$ of the whole effort to implement a record linkage procedure.

In a linkage of two datasets, say $A$ and $B$, all pairs in the cross product $A \times B$ needed to be classified as matches, nonmatches and possible matches. When dealing with large datasets, comparing all the pairs in the cross product of the two datasets is almost impracticable, in fact while the number of expected matches increases linearly, the computational problem raises quadratically (Christen and Goiser, 2005). To reduce this complexity it is necessary making use of many different techniques that can be applied to reduce the search space; blocking and sorted neighbourhood are the two main methods.

After the search space creation/reduction, it is important to pay attention to the selection of matching variables. The matching variables need to be as suitable as possible for the considered linking process that is why they are generally chosen by a domain expert. Unique identifiers can be considered the best link variables; but very strict controls need to be made in case of using numeric identifiers alone. Variables like name, surname, address, can be used jointly instead of using each of them separately; in such a way, one can overcome problems like the wide variations of the name spelling. It is evident that the more heterogeneous are the items of a variable, the higher is its identification power; moreover, if missing cases are relevant it's not useful to choose the variable as a matching one.
Comparison function is used to calculate the distance between records that are compared respect to the values of the selected matching variables (see, for a review of comparison functions, Koudas and Srivastava (2005)).

Starting from the pairs in the cross product or in the reduced search space, different decision models can be applied in order to classify them into the set of matches, the set of non-matches or in the set of possible matches. The decision rule can be deterministic or probabilistic: the former considered a pair as a true match if it agrees completely on all the chosen matching variables or if it satisfies a defined rule-base system, that is if it reaches a score which is besides a threshold when applying the comparison function. The probabilistic approach, based on the Fellegi and Sunter model, requires an estimation of the model parameters which can be performed via the EM algorithm, Bayesian methods, etc.
Additionally, a linkage process can be classified as: (i) one-to-one problem, if one record in the set A links to only one record in B and also the other way around, (ii) many-to-one problem if a record in a set can be matched with more than one of the compared file, (iii) many-to-many problem allows more than one record in each file to be matched with more than one record in the other. The latter two problems may imply the existence of duplicate records in the linkable data sources.

Finally, as not every record matched in the linkage process refers to the same identity, in the record linkage procedure evaluation, it is necessary to classify records as true link or true non link, minimizing the two types of possible errors, false matches and false non-matches respectively. The first type of error refers to matched records which do not represent the same entity, while the latter indicates unmatched records not correctly classified, that imply truly matched entities were not linked. Generally, false non-matches of matching cases are the most critical ones because of the difficulty of checking and detecting them (Ding and Fienberg, 1994). In general, it's not easy to find automatic procedures to estimate these types of errors so as to evaluate the quality of record linkage procedures.

In the next section we describe the RELAIS (REcord Linkage At IStat) toolkit; this software allows to dynamically select the most appropriate technique for each of the record linkage phases and to combine the selected techniques so that the resulting workflow is actually built on the basis of application and data specific requirements (Fortini et al (2006), Tuoto et al (2007)).

## 3.RELAIS: a solution for record linkage

The main use of the record linkage techniques in official statistics produced many software and tools both in the academic and private sectors, like BigMatch (Yancey, 2007), GRLS (Fair, 2001), Febrl (http://www.sourceforge.net/projects/febrl), Link Plus (http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm), Tailor (Elfeky et al. 2002) but, in our opinion, any of these tools provides the flexibility of multiple choices for each of the record linkage phase.

In Italian National Statistical Office (ISTAT) there is a wide use of record linkage techniques in different production processes; the first experiences date back to ' 80 s but the common practice was to develop ad hoc linkage procedures for each project, basically via deterministic approaches. Unfortunately, there was little awareness of linkage errors in further analyses of linked data and only a few official experiences with probabilistic method.

However the decennial studies on the Fellegi-Sunter theory and the belief that there is no a unique solution to record linkage problem, led an Italian team of IT researchers and statistical methodologists to design and implement the RELAIS system. RELAIS 1.0 was released in February 2008 on the Istat website with the implementation of a probabilistic model based on Fellegi-Sunter theory, the EM algorithm for the parameters estimation and file architectural structure. Due to the profitable experiences on data integration as coordinator of Essnet DI (Data Integration), the consciousness of the common nature of problems and needs of NSIs in data integration projects and also thanks to the profitable experiences in cooperation with NSIs in sharing the same software tools (Cibella et al, 2009), the RELAIS software was improved and the current version 2.1 is dramatically enriched compared to the first release.

The main ambitious goal of this software is to allow the dynamic selection of the most appropriate technique for each of the record linkage phases and the combination of the selected techniques so that the resulting workflow is actually built on the basis of application and data specific requirements. We believe that the choice of the most appropriate technique not only depends on the practitioner's skill but, most of all, it is application specific. The inspiring principle of the tool is to allow combining the most convenient techniques for each phase and also to provide a library of patterns that could support the definition of the most appropriate workflow, in both cases taking into account the specific features of the data and the requirements of the current application. In addition, RELAIS aims at joining specifically the statistical and computational essences of the linkage issue. Moreover, in order to re-use the several solutions already available for record linkage in the scientific community and to gain the several experiences in different fields, we started to develop the RELAIS project as an open source project in order to provide, in the shortest possible time, a generalized toolkit for dynamically building record linkage workflows. In line with the open source philosophy, RELAIS has been implemented by using two languages based on different paradigms: Java, an object-oriented language, and R, a functional language. This choice depends on our belief that a record linkage process is composed of techniques for manipulating data, for which Java is more appropriate, and of calculationoriented techniques for which R is a preferable choice. Form the version 2.0 RELAIS has been implemented using a relational database architecture, in particular it is based on a mySql environment. In particular, the relational database architecture optimizes the performances with respect to the management of huge amount of data through the whole record linkage project (input, intermediate phase and output).It is also added the explicit management of the output and residual files to iterate several processes and back-up management.

Thanks to the modular approach and the open source, adding new techniques to the pool already available is really easy. The RELAIS project aims to provide record linkage techniques easily accessible to not-expert users. Indeed, the developed system has a GUI (Graphical User Interface) that, on the one hand, permits to build record linkage workflows with a good flexibility; on the other hand, it checks the execution order among the different provided techniques whereas precedence rules must be controlled. The idea of decomposing the record linkage process in its phases is the core of the RELAIS toolkit and makes the whole process easier to manage; each phase has its own windows. Each of the phases described in section 2 can be performed according to different techniques; depending on specific applications and features of the data it can be suitable to iterate and/or omit some phases, as well as it could be better to choose some techniques rather than others; in the current version, RELAIS provides some of the most widespread methods and techniques for the record linkage phases.

In the Relais 2.1 version is possible the:

- Reading of input files both in text format and from database (mysql or Oracle) tables;
- Data profiling to guide the choice of matching and blocking variables;
- Creation of the search space of pairs candidate to link by means of the "cross product", "blocking" method or "sorted neighborhood" method;
- Choice of matching variables;
- Set of comparison functions (with several string distances);
- Probabilistic record linkage: estimation of the F-S model parameters via the EM algorithm;
- Deterministic record linkage: both exact and rule based;
- Reduction from $\mathrm{N}: \mathrm{M}$ to 1:1 matching solution with optimal or greedy methods.

The dataset acquisition phase permits to read two input datasets from both in text format and from database (mysql or Oracle) tables. The datasets must have the same names for the common variables that are the ones considered by the system in the subsequent phases. The database architecture allows both to start new project and to continue working to a previous one, saved as back-up. After the acquisition phase, it is possible to pass directly to the search
space creation/reduction phase or to the data profiling phase. The data profiling phase permits to characterize available variables with respect to some quality features that are used to support two critical tasks, that is blocking variables choice and matching variables selection. To give the opportunity to the user of designing the more appropriate record linkage workflow for its own application, RELAIS 2.1 supplies quality metadata, calculated starting from real data provided as input. Moreover, in order to go towards needs of less-expert users, RELAIS proposes a set quality metadata, coming from our experience to help the decision-making stages. In this phase, the metadata of quality are: Completeness, Accuracy, Consistency, Entropy, Correlation and Frequency Distributions. The search space creation/reduction phase allows to build the set of the candidate pairs to be linked. Besides the complete cross product of the file to link, two methods for space reduction are implemented, namely blocking and sorted neighborhood method.

A set of comparison functions is available in order to compare strings according to an exact or an approximate procedure. The comparison functions provided by RELAIS 2.1 are: Equality, Numeric Comparison, 3Grams, Dice, Jaro, Jaro-Winkler, Levenshtein, Soundex (http://www.dcs.shef.ac.uk/~sam/stringmetrics.html). As far as the choice of the decision model to determine the matching status on candidate pairs is concerned, the current version of RELAIS implements two kind of models, deterministic and probabilistic. Deterministic approach allows two options. Actually, according to some authors, deterministic record linkage is defined as the method that individuates links if and only if there is a full agreement of unique identifiers or a set of common identifiers, i.e. the matching variables. This corresponds to the Exact Match option in RELAIS. Other authors backed up that in deterministic context a pair can be linked also if some specific and pre-defined criteria are satisfied. Being not exact in the strict sense this kind of linkage is assumed as almost-exact and RELAIS defines it as Rule-based Match. The matching rules are defined by the users throughout the selection of matching variables and related comparison function in a disjunctive format proposed by RELAIS. The probabilistic model currently available in RELAIS consists of an implementation of the Fellegi-Sunter decision model, assuming latent dichotomous variable for the linkage status and conditional independence model for the manifest variables. The EM algorithm is used so to estimate the parameters. When blocking method is performed to reduce the search space of pairs, RELAIS allows the users to choose between two different ways of applying the probabilistic model: it can be applied in a one-shot way to all the blocks or a specific block can be selected. On the results of the model assignment, it is possible to produce an $\mathrm{N}: \mathrm{M}$ matching result or a 1:1 matching result, applying a dedicated reduction phase (Jaro, 1989). The latter phase can be applied by resolving a linear programming problem on the $\mathrm{N}: \mathrm{M}$ output by means of the simplex algorithm (optimal solution) or by a greedy algorithm, when the amount of data prevents from applying the simplex method due to its complexity. The reduction from a matching $\mathrm{M}: \mathrm{N}$ to a matching $1: 1$ is available for both probabilistic and deterministic matching. Finally, the output of the linkage process consists of several disjoint datasets: match, non-match pairs, possible match and residuals. For the Possible matches no decision is taken and they need to be processed by clerical review or by further linkage process. Also residual non-matched records resulting from the two starting files can be submitted to further analyses, that is a new record linkage process can be started by processing the residuals directly or, as an alternative, later by means of a residual back-up. Also intermediate outputs can be saved, such as blocking summary, contingency tables and parameter estimate tables.

## 4.The Censuses and other challenges

An important part of our activities is represented by some data integration projects in which we link different data files using the software RELAIS. This gives us the opportunity to test the software on real datasets and to develop, implement and update RELAIS with respect to different production needs. In this section several Istat projects relying on record linkage will be briefly described. These projects concern linkage of censuses, administrative data sources and surveys.

In the near future, the key challenges for the national statistical offices are the censuses. The next Italian Census will be assisted, for the first time, by a population register and other administrative archives. In some more details, every municipality will contact households using their own population register, deliver and then collect a census form either by mail or through alternative modes such as web form.

Alternative rosters will be used in order to deal with undercoverage of population registers i.e. people who actually dwell in a given municipality without being enlisted into the corresponding population register. Such alternative lists are based on fiscal data for the whole population, residence permits for foreigners as well as municipal lists of postal
addresses especially reviewed for census occurrence. Nevertheless, in order to successfully include people who would not be otherwise enumerated by census merely through population register, these alternative sources of information need to be previously integrated each other and together with population registers. Given that a personal identification number (fiscal code or CF in what follows) is extensively available for Italian population and allows for the use of deterministic linkage procedures, a probabilistic approach is however useful in order to handle cases in which CF are corrupted or absent.

Another important use of record linkage techniques in censuses is the detection of duplicates, i.e. people that could have filled the census form more than once. Apart from the obvious circumstances of people that send the questionnaire more than once from the same address, a less straightforward example is represented by those dwelling at two addresses during the year because of work or study reasons. These cases are more difficult to detect and probabilistic record linkage techniques need to be considered for this situation. Furthermore, record linkage techniques will be adopted for the evaluation of the census coverage by means of capture-record models. The quality of the Population and Agricultural Census will be verified adopting the Petersen model; linkage between census and post enumeration surveys data will be performed in order to estimate the coverage rate of the Census.

RELAIS is adopted also for applying record linkage procedures in socio-demographic studies concerning, from one hand, the analysis of fertility practices within the marriages couples both of the Italian population and of foreigners, especially long-term migrants, in mixed couples. Data on marriages could be linked with the registers of births for the years following the marriage, via both deterministic and probabilistic approach. From the other hand, longitudinal studies on the characteristics of the subpopulation of legal foreigners and the evolution of the needs of some specific communities and their settlement on the Italian territory can be performed by linking the permits registers through time. The deterministic approach is not suitable for this particular analysis since unique identifiers are not available for the administrative data sources taken into account. Additionally, several blocking procedures are needed due to the large amount of the data to be linked. RELAIS is the proper instrument for designing a probabilistic procedure in this complex context.
In the business survey context, there is a currently running project aiming at integrating data from two surveys: Community Innovation Survey (CIS) and Information and Communication Technologies (ICT). The main goal is to provide an integrated dataset to stimulate and support research focusing on the analysis of ICT as a key driver of innovation and analysis of investments in ICT by innovating firms. Indeed, the internal variability of productive systems and industries, firms' heterogeneity and variety can be detected just by looking at firm-level data. The ICTinnovation bidirectional relationships and their causality issues can be addressed in a comprehensive manner only by integrating different micro-level statistics. Broadly speaking, existence or not of such bidirectional relationships is a feature to be evaluated at the beginning of any integration process. First of all, such causalities might guide the linking process; for example, one could know in advance if one-to-one or many-to-many linkage are possible. Moreover, ignoring causalities might diminish the analytical potential of the final integrated dataset. When linking two surveys, designing a data integration strategy is a very complex process. First, the definition of a common target population has to be tackled, since structural differences may be taken for granted. For example, one has to consider that the surveys do not refer to exactly the same population, they have different reference years and different demographic events on enterprises might happen in the time between the surveys reference periods, the moment of sample selections and estimation phases. Second, a careful choice of the blocking and linking variables should be made. If possible, the reduction of the search space (or blocking) should incorporate information about the occurred demographic events. For example, in an enterprise setting, a too detailed principal activity classification could not be a suitable as blocking variable. As for the comparison variables, for business surveys, except for few structural variables, the common shared variables, hence candidate comparison variables, are continuous variables. For such variables, the choice of a numerical distance between variables is almost compulsory.

To allow rigorous analyses of bidirectional relationships, the integrated dataset should guarantee representativeness of both surveys. When producing the integrated dataset, differences in the design strategies have to be considered. Special attention has to be paid to the non-respondent units and to the non-response adjustment procedures. This is a particularly challenging task since the known totals of both surveys should be simultaneously preserved. Moreover, coherence with structural information on enterprises or population should be achieved by further integrating information stemming from registers. It should be observed that the issue of structural information on statistical units in the integrated dataset is directly related to the target population definition. As a first consequence of this weights re-adjustment, the original surveys weights would be modified; researchers should be well informed about the right way to conduct their analyses. Finally, every unit of the integrated output dataset should be supplied with a linkage
probability, representing the probability of being correctly linked. In this way, any analysis based on the integrated data could explicitly report the linkage errors and the related outputs could be refined with respect to the linkage process. For the CIS-ICT data, we will rely as much as possible on RELAIS. This will be a good opportunity to test its applicability when integrating two surveys data in business context.

## 5. Future research projects and conclusions

In this paper, we have illustrated the RELAIS project, an open source toolkit for building record linkage workflows; RELAIS is composed by a collection of techniques for each record linkage phase that can be dynamically combined in order to build the best record linkage strategy, given a set of application constraints and data features provided as input. The toolkit aims at offering multiple techniques for record linkage, both deterministic and probabilistic, with the possibility of building $a d$-hoc solution combining each module. In fact, RELAIS proposes solutions based on a set of choices and parameters, i.e. which variables have to be used for blocking and matching respectively, the choice of the decision model, whose parameter settings is left to the user's choices.
The RELAIS project has been enriched also thanks to the cooperation with many researchers and users in international context (Cibella et al, 2008). Several remarks come from the profitable share of knowledge and solutions among different institutes and countries in dealing with 'real-world' tasks: first of all, the awareness of the common nature of the faced problems; then, the advantages in designing standardized answers to specific, though widespread, applications.

In the next future work, we have planned to extend the current functionalities of RELAIS and to optimize its performances, especially to deal with very large amount of data (the millions of record of the next Italian Censuses).

In particular we would consider:

- the explicit application for de-duplication;
- a procedure for nesting the blocking methods;
- the possibility to set probabilities by the users;
- the implementation of a new method for optimal 1:1 reduction based on the Hungarian algorithm;
- the improvement of GUI functionalities for output management and user interactions (manual review);
- a summary output on linkage results;
- the batch execution;
- the interfaces for clerical review.

Also some efforts will be dedicated to implement some standardized functionalities for the pre-processing phase, as character conversions, schema reconciliation, standardization, etc.; the modification of the probabilistic approach; the interactions between matching variables; the Bayesian approach and the graphical analysis on the model fitting. In the paper, we have described some studies as a proof of concept of the inherent complexity and variety of record linkage processes. Indeed, due to such complexity and variety, the great modularity and flexibility of RELAIS are necessary in order to properly build application specific record linkage workflows.

Furthermore, besides the enrichment of the set of available techniques, future work for RELAIS will include several patterns that can guide the design of record linkage workflows. Indeed, we believe that the design stage of a record linkage workflow could usefully exploit patterns extracted from previous knowledge and experiences. In this way, the toolkit can assist non-expert users in designing their specific record linkage workflows. Each technique could be characterized in terms of pre-conditions that must be respected in order to be part of a record linkage workflow. We will study the possibility of using formal languages for the specification of such preconditions, in order to check properties like consistency and completeness of a proposed workflow solution with respect to given application and data requirements.

## References

Christen, P. and Goiser, K. (2005). Assessing deduplication and data linkage quality: What to measure?, Proceedings of the fourth Australasian Data Mining Conference (AusDM 2005)', Sydney.

Cibella, N., Fortini, M., Scannapieco, M., Tosco, L. and Tuoto, T., (2008). Theory and practice of developing a record linkage software. Proceedings of the Combination of surveys and administrative data Workshop of the CENEX Statistical Methodology Project Area, Integration of survey and administrative data, Vienna, Austria.

Cibella, N., Fernandez, G.L., Fortini, M., Guigò, M., Hernandez, F., Scannapieco, M., Tosco, L. and Tuoto, T. (2009). Sharing Solutions for Record Linkage: the RELAIS Software and the Italian and Spanish Experiences. Proceedings of the New Techniques and Technologies for Statistics (NTTS) Conference, Bruxelles, Belgium.

Ding Y. and Fienberg S.E. (1994). Dual system estimation of Census undercount in the presence of matching Error. Survey Methodology, 20, pp. 149-158.

Elfeky, M., Verykios, V. and Elmagarmid, A.K.: Tailor (2002). A Record Linkage Toolbox. Proceedings of the 18th International Conference on Data Engineering IEEE Computer Society, San Jose, CA, USA.

Fair, M. (2001). Recent developments at Statistics Canada in the linking of complex health files. Federal Committee on Statistical Methodology, Washington D.C., USA.

Fellegi, I.P. and Sunter, A.B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, pp. 1183-1210.

Fortini, M., Scannapieco, M., Tosco, L. and Tuoto, T. (2006). Towards an Open Source Toolkit for Building Record Linkage Workflows. Proceedings of SIGMOD 2006 Workshop on Information Quality in Information Systems (IQIS), Chicago, USA.

Gill, L. (2001). Methods for Automatic Record Matching and Linkage and their Use in National Statistics.National Statistics Methodological Series. 25, HMSO Norwich, UK.

Koudas, N. and Srivastava, D. (2005). Approximate joins: Concepts and techniques. Proceedings of VLDB 2005.
Jaro, M.A. (1989). Advances in record linkage methodology as applied to the 1985 census of Tampa Florida, Journal of the American Statistical Society, 84 (406), pp.414-20.

Newcombe, H., Kennedy, J., Axford, S. and James, A. (1959). Automatic Linkage of Vital Records. Science, 130, pp. 954-959.

Shlomo, N. (2009). Assessing disclosure risk under misclassification for microdata. UNECE/EUROSTAT Work session on Statistical Confidentiality, Bilbao. Available at http://www.unece.org/stats/documents/2009.12.confidentiality.htm.

Skinner, C. (2008). Assessing disclosure risk for record linkage. Domingo-Ferrer, Josep and Saygin, Yücel (eds.) Privacy in Statistical Databases. Lecture Notes in Computer Science 5262 Springer, Berlin, Germany, pp. 166-176.

Tuoto, T., Cibella, N., Fortini, M., Scannapieco, M. and Tosco, L. (2007). RELAIS: Don't Get Lost in a Record Linkage Project. Proceedings of the Federal Committee on Statistical Methodologies (FCSM) Research Conference, Arlington, VA, USA.

Yancey, W. (2007). BigMatch: A Program for Extracting Probable Matches from a Large File. Technical report Statistical Research Division U.S. Bureau of the Census, Washington D.C. Research Report Series Computing, 2007-01.

# Obtaining estimates of parameters for probabilistic record linkage using the EM algorithm 

William E.Yancey ${ }^{1}$


#### Abstract

Following Fellegi and Sunter, probabilistic record linkage uses the conditional probabilities for agreement patterns. Since these probabilities are conditioned on the true match status of the agreement pattern types, they are presumably unknown, so that implementation of agreement weight calculations requires estimating these probability parameters. The EM algorithm enables us to calculate maximum likelihood estimates for these parameters conditioned on the latent classes of true matches and true non-matches without the use of training data. If a sufficiently large proportion of matching record pairs is present in the total sample of record pairs, then under the conditional independence assumption, the EM algorithm is straightforward to implement, converges rapidly, and provides useful parameter estimates that are specific to the particular set of record pairs to be linked. In the case of census data person records, improved parameter estimates can be obtained by using a three latent class model. It is possible to include frequency data to adjust the parameter values. The conditional independence assumption can also be replaced by including interactions between the matching fields within each of the latent classes.


[^44]
## SESSION 6B

## MULTIPLE FRAMES

# The optimal estimator in dual frame surveys using indirect sampling 

Manuela Maia and Paula Vicente ${ }^{1}$


#### Abstract

Undercoverage is one of the most common problems with sampling frames and a likely cause of coverage error. To reduce the impact of undercoverage on survey estimates, several frames can be combined in order to achieve a complete coverage of the target population. Multiple frame estimators have been developed to be used in the context of multiple frame surveys. Sampling frames may overlap in some situations, which is the case when some units from the different sampling frames are related to the same units of the target population. Indirect sampling (Lavallée, 1995) is an alternative approach to classical sampling theory in dealing with the overlapping issue of sampling frames on survey estimates.

In this paper, a new class of estimators is presented that is the result of merging dual frame estimators with indirect sampling estimators in order to bring together in a single estimator the effect of several frames on survey estimates. These estimators are compared with the optimal estimator from Deville and Lavallée (2006), and we try to obtain the general formula for the variance of these estimators for different kinds of links. Additionally, a practical case of estimation in the context of the Eurobarometer is presented.


[^45]
# Demographic analyses in 2010: Methodological challenges 

Kirsten West, J. Gregory Robinson, Jason Devine and Renuka Bhaskar ${ }^{1}$


#### Abstract

In the United States, demographic analysis methods have historically been used to develop estimates of the population for comparison with decennial census counts. The estimates are developed from various types of demographic data independent of the census, such as administrative statistics (births, deaths, and Medicare data) and estimates of immigration and emigration. The paper focuses on the methodological challenges when following birth cohorts since 1945 to create an estimate of the size of the U.S. population on April 1, 2010. Specifically, the paper addresses the uncertainties in the administrative birth and death records, the reliance on survey data to estimate immigration, and the use of the Medicare enrollment database to estimate the elderly population. Also, we discuss the key issues associated with assigning race to the multiple data sources so that the estimates are consistent with the race categories developed for the decennial census.


Key Words: Demographic Analysis; Methodological challenges.

## 1. Introduction

In the United States, demographic analysis methods have historically been used around the time of the decennial census to develop estimates of the population for comparison with the census counts. For the 2010 Census, the U.S. Census Bureau is again using Demographic Analysis (DA) estimates to assess the quality of the census. The estimates are produced at the national level by age, sex, and two broad race categories, Black and non-Black.

The 2010 DA estimates are built from components of population change. They start with the births that occurred in 1945 and annually thereafter to get the population base. Immigrants are added to this component of natural population growth. Attritions from the population are produced from deaths and emigration statistics.

Administrative records on births and deaths combined with survey-based estimates of international migration are used to estimate the population under age $65 .{ }^{2}$

Finally, the demographic component estimates for the population under age 65 are combined with the Medicare-based estimate for the population 65 years of age or older to produce the total DA population estimate.

The sections below focus on the methodological aspects of the estimates production. Specifically, we show the challenges associated with combining data from administrative records, censuses, and surveys to create estimates of the U.S. population at a given point in time. Five methodological areas illustrate these challenges:

1. Selection of a data source
2. Vintage of the data source
3. Selection of the universe
4. Coverage of the selected universe
5. Consistency of data items across data sources
[^46]
## 2. Methodological challenges

### 2.1 Selection of a data source

To create the DA estimate we need data to measure population growth at the national level. The basic requirement is births, deaths, and net international migration. Because national birth records do not go back any further than 1933, and because Medicare data did not become available until the 1960s, different estimation strategies have been employed over time for different age groups.

For example, for the 1970 DA estimates, the components of change method (births, deaths, net international migration) anchored the DA estimates for the population under age 35, Medicare provided the basis for the population aged 65 and older, and indirect estimation methods such as mathematical modeling techniques were used for the population in the age group 35 to 64 .

The DA methodology has continued to evolve, new data sources have been incorporated, and the estimated components of change have been improved. The growing importance of immigration as a component of change has placed greater demands on the accuracy of this component.

Administrative data from the Department of Homeland Security's Office of Immigration Statistics (DHS/OIS) can be used to estimate the number of immigrants who arrived in the United States during a given time period and adjusted to legal permanent resident (LPR) status. The data include records of new arrivals admitted to the country as legal permanent residents and records of "adjustees," or those immigrants who were already present in the United States when they adjusted to LPR status. This data source has shortcomings due to missing information on many data items such as year of arrival. Also, the database does not include temporary residents who depart the country or die before obtaining legal permanent residence, permanent foreign-born residents who would never receive immigrant visas, or unauthorized migrants who would never appear in immigrant data as legal permanent residents.

It is also increasingly problematic to equate legal immigration, as defined by DHS, with actual moves to the United States, as would be defined by the Census. Above all, rapid changes in the level of immigration and large fluctuations in temporary migration (especially in the 1990s) belie the assumption that difficult-to-measure temporary or undocumented arrivals can be approximated by the more measurable adjustments of status.

Thus, over the past decade, the U.S. Census Bureau has undertaken a major initiative to improve its ability to measure net international migration. The implementation of the American Community (ACS) provides critical demographic information between decennial censuses. This new data source has led to new methods of estimating components of migration. One methodology, referred to as the Residence One Year Ago (ROYA) method, is the basis for one of the DA 2010 estimates for the 2000 to 2010 time period. ${ }^{3}$ An alternative to the ROYA-based estimate, data collected in the ACS on year of entry (YOE) is also used. With this approach, the foreign born whose year of entry is the year prior to the survey year are considered immigrants.

Estimates of foreign-born emigration are developed from estimated rates of emigration which are then applied to estimates of the foreign-born population from the ACS. The foreign-born household population in Census 2000 is aged forward using the National Center for Health Statistics (NCHS) life tables to obtain the expected population in 2005, 2006, 2007, 2008, and 2009. The expected foreign-born population estimates for those years are then compared to the foreign-born population estimated by the ACS. Subtracting the estimated from the expected population produces a residual, which serves as the basis for emigration rates. This residual methodology with decennial census data has been used in the past to measure emigration. The incorporation of data from the ACS provides a more updated measure of emigration. However, there are several important limitations to this method, including an assumption that coverage differentials between Census 2000 and the ACS are minimal; a lack of

[^47]foreign-born specific death rates; and a reliance on the accuracy and consistency of year of entry and nativity reporting.

The net-native born emigration is another difficult component to estimate because of a lack of adequate data. For DA 2010, we built on work by Schachter (2008) that examined census data from other countries. The estimates are again developed using a residual method. The number of those native to the United States is ascertained from country specific census data, aged forward, and then compared to the number of those native to the United States in a later census from the same country. Assumptions about the capture in the census of the population that is born abroad of U.S. citizen parents are built into the measure.

In summary, the DA method is an accounting of population change through the use of various data sources. Vital statistics data are readily available to estimate the birth and death components of population growth. Medicare enrollment data are also considered a good source for estimating the size of the elderly population. Net international migration, however, continues to be a difficult component to estimate. There is not a good single source. The U.S. Census Bureau has elected to follow a strategy that relies on survey data supplemented with checks from administrative sources.

### 2.2 Vintage of the data sources

The data files customarily lag one to two years and sometimes three years behind the desired target date for the estimation. This is the case for the vital statistics data, the administrative enrollment database utilized to produce the DA estimate and the ACS survey data. The vital statistics data are generally less current.

The National Center for Health Statistics (NCHS) provides the U.S. Census Bureau with individual level data on births and deaths without name and address information. The data are compiled and released with some delay. Each new DA estimate covers an additional decade. For the majority of years in a decade, the data are available in full detail. As you get close to the date of the estimate, the data are available in either provisional or preliminary detail and decisions have to be made whether to use the less complete provisional and preliminary data, or to project old values to future dates. The differences between the preliminary and final data are usually very small, but the reliance on data that are not "final" for the last year introduces some error. ${ }^{4}$

The Centers for Medicare and Medicaid Services (CMS) provide the database used to estimate the population aged 65 and over. The file is current and different cut dates can be requested. For preliminary DA estimates the file lags behind by 9 months, but since the final DA estimates are produced in the fall of the estimates year to cover only the period up to April 1 (Census Day), the cut date for the file can be requested to be close enough to coincide with the population on April 1 of the estimates year and no further projections or extrapolations are needed.

In summary, the selected data source may not cover the date of the estimates. To achieve that goal it is often necessary to project the data to the desired date using suitable projection methodologies.

### 2.3 Universe selection

Once a suitable data source has been identified, the proper universe must be selected. The DA estimates pertain to the U.S. resident population in the 50 states and the District of Columbia. The estimates exclude residents of the Commonwealth of Puerto Rico and residents of the Island areas under U.S. sovereignty or jurisdiction (principally American Samoa, Guam Island, United States Virgin Islands, and the Commonwealth of the Northern Mariana Islands).

The definition of resident population conforms to the criteria used in the decennial census, which defines a resident of a specified area as a person "usually resident" in that area. Estimates of the resident population exclude the U.S. Armed Forces overseas, as well as civilian U.S. citizens whose usual place of residence is outside the United States. It includes all persons regardless of their legal status.

[^48]In order to be consistent with these universe rules, it might be necessary to select only eligible cases or to make modifications to the administrative database. Slight adjustments must be made to the birth data. Demographic analysis includes births to mothers who resided in the United States at the time of the birth. Births to non-resident mothers are excluded. While these births would be considered native born, the children may not reside in the United States. There may also be births that occur to resident mothers who leave before the estimates date. These births may or may not be accounted for in the estimates of emigration.

Similarly, not all records on the Medicare database meet the DA universe criteria. It is necessary to exclude individuals not residing in the United States and those who are not 65 years of age or older as of April 1 of the estimate year (some individuals under age 65 may be eligible for Medicare if diagnosed with permanent disability or permanent kidney failure that requires dialysis or a kidney transplant). Records missing a date of birth also have to be excluded. Furthermore, all persons on the file must have been alive on the estimates date. Deceased persons must be excluded. Finally, as an additional quality assurance check, efforts are made to eliminate duplicate records.

### 2.4 Coverage of the selected universe

Administrative records are specific to the purpose they serve. Coverage errors might not be an impediment to that purpose. It may, however, be of concern when the file is used for other purposes such as the estimation of the entire population.

In the United States, birth registration has been considered complete since 1985. Tests of birth registration completeness were conducted for 1940, 1950, and 1964-1968 to provide correction factors for those years; correction factors for other years were derived by interpolation and extrapolation.

The methods to interpolate and extrapolate the birth registration adjustment factors were modified over the years. In the 1980 DA, a curvilinear interpolation of the 1940 and 1950 test results was incorporated to allow for a more rapid improvement in registration completeness in the early 1940s. In the 1990 DA, the adjustment factors for Blacks for 1935 to 1940 were revised to account for a bias in the 1940 birth registration results (Passel, 1992).

Though the coverage of birth records is generally accepted as complete, there may still be issues that need to be addressed to derive the total number of births for a specific year. State-specific issues may affect the total number of reported births. For example, a micro-level file provided by NCHS for Vermont in 2005 was based on incomplete data and underreported resident births by about 3 percent of resident births (U.S. Department of Health and Human Services, 2008). In addition, during the past decade, natural disasters, such as Hurricane Katrina, may have disrupted reporting of births and deaths for large reporting areas (Hamilton et al., 2009).

Administrative data on deaths also have shortcomings, but there is little information available on which to quantify empirically the possible extent of the underregistration (U.S. Bureau of the Census, 1988, 1991). Therefore, the magnitude of underregistration of deaths must be based on assumptions. Up to and including 1959, it was assumed that infant deaths were underregistered at one-half the rate of the underregistration of births. No adjustment for infant deaths is made for years since 1960 (U.S. Bureau of the Census, 1988). The registration of deaths for ages 1 and over has been considered 100 percent complete for the entire DA time series starting in 1935.

The American Community Survey (ACS) serves as the data source for estimates of foreign-born immigration. It is a main assumption that the potential under-representation of foreign-born immigrants in the ACS is accounted for by controlling the ACS to Census 2000-based population estimates (U.S. Census Bureau, 2009). The ACS data are controlled so that the number of housing units and people by age, sex, race, and Hispanic origin match the Census Bureau's official population estimates. Differential coverage of those who have immigrated in the last year within these control categories or imprecision in the population controls may result in variation in the survey-based estimates of foreign-born immigration.

Medicare enrollment is generally presumed to be a nearly complete source of information on the size of the U.S. resident population 65 years of age or older. In the late 1990s, the Social Security Administration estimated the Medicare coverage of the population to be about 96 percent (U.S. House of Representatives Committee on Ways and Means, 1998).

Information from survey data is used to get an estimate of potential under-enrollment by sex and race. The Current Population Survey (CPS) Annual Social and Economic Supplement (ASEC)-Health Insurance data have been determined to be a good source for the development of the never-enrolled adjustment factor. ${ }^{5}$ The CPS ASEC respondents are asked whether they had any medical insurance coverage during the previous calendar year based on interviews conducted in February-April of the survey year. A specific question on the survey inquires about enrollment in Medicare.

### 2.5 Consistency of data items across data sources

The demographic characteristics on the administrative file may not matter to the specific program purpose of the file, but it matters if the data source is used to estimate a specific characteristic for the entire U.S. population. Race falls in this category. Race classification is a complex and substantial challenge for the DA estimation. Production is limited to the race information available on the historical data files and the way it was collected. DA estimates are generally produced for two broad race categories: Black and All Other Races Combined (the latter is referred to as "non-Black").

Decisions about the assignment of race to the vital events can have an impact on the results obtained from DA, especially births because births serve as the primary source for determining the size of each native-born birth cohort. The information recorded on the birth certificate on the race(s) of the mother and father is used to assign race to each birth. Race can be ascribed based on the mother's race, the father's race, or a combination of both.

Often, the administrative records do not have consistent race information. For example, race classification is a challenge when using the Medicare database. While only a small percentage of records have missing, unknown, or invalid race codes, a larger issue is the treatment of Hispanic origin as an additional race category, rather than as an ethnicity. The database also includes an "Other" category. Thus, a strategy for redistributing these cases to the two DA race categories is needed. Over the last decade, the U.S. Census Bureau has built a person-level file from administrative records primarily for application to decennial census research and development. The research file offers an opportunity to use a new approach for DA to assigning race to the Medicare data.

The research file starts with the Social Security Administration's (SSA's) Numident file. The SSNs are replaced by randomly assigned Protected Identity Keys (PIKs) to maintain confidentiality (Farber and Miller, 2002) and then matched to the person record from Census 2000. The result is the so-called 100 percent Census Numident file. This file is enhanced so that each record has either the census race or a model-based race and Hispanic origin code. The enhanced Census Numident is known as the Person Characteristic File (PCF). The race information on this file was used to redistribute records to the DA race categories of Black and non-Black.

If the immigration data come from surveys, race is usually readily available. For other data sources, it may be necessary to use proxy universes that are assumed to be representative of the universe. The race distribution in the decennial census is a frequently used source for the proxy universes.

In summary, when using multiple data sources to produce estimates by characteristics such as race, the absence of the information, the quality of the information, and the conceptual inconsistencies in reporting the information become methodological challenges.

## 3. Summary and conclusion

In this paper, we describe how Demographic Analysis combines censuses, surveys, and administrative data to create population estimates. Demographic Analysis has a long history at the U.S. Census Bureau. The DA analytic methods and underlying assumptions have evolved over time as new data or information became available. The logical consistency and interrelationships of the underlying demographic variables and the data used to measure them are strengths of the demographic method.

[^49]In the previous sections, we described the methodological challenges that are encountered by a demographic analysis approach to create an estimate of the size of the population on Census Day. The data source must measure the estimated component; the vintage of the administrative records and the survey data must cover the time period of the DA estimates; the administrative records and the DA estimates universe must be in agreement; and allowances must be made for incompleteness in the administrative records or the survey universe.

Perhaps the biggest methodological challenge is to find good data sources for capturing net international migration. The impact of this component on population change is increasing and must be carefully monitored from different data sources throughout the decade between the censuses to ensure that it is captured.

The production of estimates by race characteristics is also a big challenge. The DA estimates are produced by single year of age and sex, and for Blacks and non-Blacks. Even with this limited set of categories, it is a challenge to achieve a race distribution that is uniform across the various data sources and consistent with the census. It requires a number of manipulations and the utilization of information across data sources.

## References

Farber, J. and Miller, E. (2002). Matching Census 2000 to Administrative Records, Proceedings of the 2002 Joint Statistical Meeting, Alexandria, VA.

Hamilton, B.E., Martin, J.A. and Ventura, S.J. (2009). Births: Preliminary Data for 2007. National Center for Health Statistics. National Vital Statistics Reports 57(12).

Passel, J.S. (1992). Alternative Estimates of Black Births Corrected for Underregistration: 1935-1980. The Urban Institute, Washington, D.C.

Schachter, J. (2008). Estimating Native Emigration from the United States, Memorandum dated December 24, developed during contract work for the U.S. Census Bureau.
U.S. Census Bureau (2009), Design and Methodology, American Community Survey, U.S. Government Printing Office, Washington, DC.

$$
0
$$

(1991), Uncertainty Measure for Other Components. DA Evaluation Project D7. PERM No.80, by Robinson, J.-G., Woodrow, K.-A., and Ahmed, B.
$\qquad$ (1988), Methodology for Developing Estimates of Coverage in the 1980 Census Based on Demographic Analysis: Birth and Death Statistics, 1935-1980, by Passel, J.-S., and Robinson, J.-G. 1980 Census Preliminary Evaluation Results Memorandum Series 112.
U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics (2008). Detailed Technical Notes, 2006. Accessed on December 17, 2009 (http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm).
U.S. House of Representatives Committee on Ways and Means (1998). 1998 Green Book-Background Material and Data on Programs within the Jurisdiction of the Committee on Ways and Means. Washington, DC: U.S. Government Printing Office.

# Dual-frame weighting and estimation challenges for the Canadian Community Health Survey 

Kate Wilder and Steven Thomas ${ }^{1}$


#### Abstract

The Canadian Community Health Survey (CCHS) is a continuous-collection cross-sectional survey with an annual sample of 65,000 respondents. The CCHS selects its sample from two sources in most health regions: an area frame and a telephone list frame. This dual-frame design creates several challenges in weighting and estimation, including the creation of a representative sample when frame memberships are unknown for some units, and dealing with the effect of collection mode on the survey responses. This article describes the CCHS's efforts to create consistent and reliable estimates in the face of these challenges.


Key Words: Multiple Frames; Weighting; Mode Effect.

## 1. Introduction

### 1.1 Background

The Canadian Community Health Survey (CCHS) is an important survey at Statistics Canada that focuses on the health status, risk factors and access to health care of all Canadians aged 12 and older living in private dwellings. The survey is cross-sectional, with data being collected on a continuous basis. Results are released on an annual or sub-annual basis. The target domain for estimation is the health region (HR), of which there are approximately 121 across Canada. In most health regions, the CCHS uses two sources to select the sample of dwellings: an area frame and a telephone list frame (see Béland et. al., 2005). The details for each frame, including the sampling methodologies, are outlined in the next two sub-sections. Dual-frame specific challenges and solutions will be discussed in section 2 while the conclusions will be discussed in section 3 .

### 1.1.1 Telephone Frame

Slightly more than half of the final CCHS sample of respondents is selected from the CCHS telephone list frame. This frame is an external database of names, telephone numbers and addresses in Canada. It consists of publicly listed landline telephone numbers only, so all households with an unlisted telephone number or those with a cellular phone only are not covered by this frame. Overall, it covers approximately $70 \%$ of Canadian households in the target population of the CCHS. To select a sample from this frame, the $N_{T}$ telephone numbers are first allocated to the different HRs according to their postal code. Next, $n_{T}$ telephone numbers are selected using stratified simple random sampling, where the strata are health regions. One advantage of using the telephone list frame is this simple sampling design. The main advantage, though, is the relatively low collection costs, since the interviews for respondents selected from this frame are conducted through computer-assisted telephone interviewing (CATI).

[^50]
### 1.1.2 Area Frame

The area frame consists of a list of geographic clusters of dwellings. This frame was created primarily for the Canadian Labour Force Survey (LFS) but other household surveys also use it. Unlike the telephone list frame, the area frame covers $98 \%$ of the target population. To select a sample of $n_{A}$ households using this frame, the CCHS must use a complex, multi-stage design based on that of the LFS (see Statistics Canada, 2008), where the first stage is a selection of geographical clusters and the second stage is a selection of a systematic sample of dwellings within each cluster. The area frame also uses health regions as stratification. This more complex sampling design is one of the disadvantages of the frame. The main disadvantage, though, is the relatively high collection costs, since the interviews for respondents selected from this frame are conducted through computer-assisted personal interviewing (CAPI) instead of by telephone. The collection costs associated with personal interviews are so high that a CAPIonly survey for the CCHS is not feasible.

## a. Dual-frame sampling

To understand the survey's sampling and weighting methodology, some notation specific to its dual-frame design may be helpful. Note that there is no 'telephone-only' domain as the area frame has assumed $100 \%$ coverage. The notation in Table 1.2-1 is based on Hartley (1962):

Table 1.2-1
Dual frame notation

|  | Frame |  | Domain |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Area (A) | Telephone (T) | Area frame only $(a)$ | Overlap <br> $(a t)$ |
| Population count <br> (households) | $N_{A}$ | $N_{T}$ | $N_{a}$ | $N_{a t}$ |
| Sample count <br> (households) | $n_{A}$ | $n_{T}$ | $n_{a}$ | $n^{A}{ }_{a t}, n^{T}{ }_{a t}$ |
| Population total | $Y_{A}$ | $Y_{T}$ | $Y_{a}$ | $Y_{a t}$ |
| Sample total | $y_{A}$ | $y_{T}$ | $y_{a}$ | $y^{A}{ }_{a t}, y^{T}{ }_{a t}$ |

For the CCHS, sampling is performed independently for each frame. In order to ensure that the collection is continuous and equally representative of the entire year, the telephone and area frame samples are allocated equally to six two-month collection periods throughout the year. From each responding household, one individual is selected with unequal selection probabilities to complete the questionnaire.

The samples of households drawn from each frame can be visualized using Figure 1.2-1. The main aspect to notice is that there is a certain part of the population that is only covered by the area frame and represented by the sample $n_{a}$. The remainder of the population, the overlap of the area and telephone frames, is represented by two samples $n^{T}{ }_{a t}$ and $n^{A}{ }_{a t}$. These dwellings have roughly twice the probability of selection compared to a dwelling selected from the noncommon portion since they have a chance of selection from either frame.

Figure 1.2-1
Area frame and Telephone frame samples of households


As will be further explained in section 2.3, not only are the populations covered by each frame different, but responses obtained through CATI may differ from responses obtained through CAPI for certain CCHS questions. This leads to some extra considerations during the weighting process. Overall, the use of the two sampling frames means the CCHS can have a more efficient sample for a fixed budget. From 2005 to 2008, approximately $50 \%$ of the final CCHS respondents were selected from the telephone frame and $50 \%$ were selected from the area frame. Due to increasing collection costs, the proportion of CATI interviews in the sample has been increasing since 2009.

### 1.3 Weighting

On the CCHS, survey weights must correctly take into account individual selection probabilities, household and person nonresponse, and calibration to known population totals. For the CCHS, there are over 13 weight adjustments involved in the derivation of the final person-level weight. For more information on the CCHS weighting methodology see Sarafin (2007).

First, design weights are calculated separately for the dwellings selected from each frame. These weights are adjusted to account for household nonresponse leading to household-level weights $w_{A}$ and $w_{T}$. These weights do not take into account the increased selection probabilities for households in domain at. The next step in the weighting process is to combine all in-scope sampled households into one sample that is representative of the target population of households. The CCHS refers to this process as integration.

## 2. Dual-frame Challenges and Solutions

### 2.1 Integrating the Units from Each Frame

To adjust the weights and create one representative sample, the basic method of Hartley (1962) is used. In the general case, the weight of each observation is adjusted by a factor $\alpha$ for units selected from the first frame and by a factor $1-\alpha$ for units selected from the second frame. This leads to a composite estimator of the form:

$$
\hat{Y}_{\mathrm{int}}=\alpha \hat{Y}^{A}+(1-\alpha) \hat{Y}^{T}
$$

The estimator is unbiased for any choice of $\alpha$ as long as both $\hat{Y}^{A}$ and $\hat{Y}^{T}$ are unbiased. In other words, both frames must entirely cover the target population and both sets of respondents must respond the same way. In the case of the CCHS, a portion of the target population is not covered by the telephone frame. After household weights are calculated for each frame, we observe the following: $\sum w_{a}=N^{A} \approx N$ and $\sum w_{t}=N^{T}$. Therefore, this composite estimator will lead to biased estimates unless there is an adjustment for the undercoverage of the telephone frame and will continue to be biased if the households not covered by the telephone frame have different characteristics than the households covered. In the case of the CCHS, there is reason to believe that this is the case, as households with no landline have been associated with a generally younger, single-person demographic due to the increasing reliance on cellular phones only.

To take the undercoverage into consideration, the CCHS integrates only the sampled households that are common to both frames, or, in other words, are part of domain $a t$. The weights of the households that belong to domain $a$, or the
non-common portion of the area frame, do not have their weights adjusted. Because of this, these households can represent other similar unlisted households in the population. A composite estimator for a total thus takes the form:

$$
\hat{Y}_{\mathrm{int}}=\hat{Y}_{a}^{A}+\alpha \hat{Y}_{a t}^{A}+(1-\alpha) \hat{Y}_{a t}^{T}
$$

The expected value will be an unbiased estimate of the population total $Y=Y_{a}+Y_{a t}$. In the case of a proportion, if the estimates coming from the common portion of each frame are unbiased, it can be shown that the expected value of this estimator is $p Y_{a}+(1-p) Y_{a t}$, where $p$ is the proportion of units not covered by the telephone frame.

### 2.2 Identifying Units in the Common Domain at

In order to adjust the weights of the units in the common portion only, we must first determine frame membership for all households. This is straightforward for most of the households in the sample. Households selected from the telephone frame are, by design, also on the area frame since the area frame has complete coverage. In order to determine frame membership for the households selected from the area frame, they must be linked to the telephone list frame. This is done through the use of questions asked of all final respondents. One set of questions determines the landline telephone number(s) of the respondent: "Q1, Q2: What is your main (second) telephone number?" The other question attempts to determine, for the respondents who do not provide a telephone number, whether they have a listed number: "Q3: Do you have a landline telephone number that is listed in a paper or Internet telephone book?" Respondents with no listed telephone number are considered to fall in the "non-common portion", whereas respondents whose telephone numbers are linked to the CCHS telephone list frame are considered to be in the "common portion". Frame membership remains unknown for person nonrespondents and the respondents who neither provide a telephone number nor state whether they have a listed landline. For the 2009 annual CCHS sample, the breakdown for households in applicable health regions can be seen in Table 2.2-1.

Table 2.2-1
Domain membership for all applicable households

|  | Considered in <br> domain $a t$ | Considered in <br> domain $a$ | Domain <br> unknown | Total |
| ---: | ---: | ---: | ---: | ---: |
| Tel. number obtained (Q1,Q2) | 21,065 | 5,610 | 0 | $26,675(84 \%)$ |
| 'Listed status' obtained (Q3) | 239 | 1,962 | 0 | $2,201(7 \%)$ |
| Nonresponse \& Refusals | 0 | 0 | 3,051 | $3,051(9 \%)$ |
| Total | $\mathbf{2 1 , 3 0 4}(\mathbf{6 7 \%})$ | $\mathbf{7 , 5 7 2 ( 2 4 \% )}$ | $\mathbf{3 , 0 5 1 ( \mathbf { 9 \% } )}$ | $\mathbf{3 1 , 9 2 7 ( 1 0 0 \% )}$ |

For the $9 \%$ of cases for which neither a telephone number nor a 'listed status' is obtained, domain membership remains unknown. There are several ways to treat these households. One way is to impute frame status using related auxiliary information. This strategy could preserve the distribution of frame status, but a more precise method is preferable. Instead, logistic regression is used to model the probability of being in the "area-frame only" domain, $a$. For this model, information obtained during the initial household contact is used, including geographic information and the age, gender and number of household members. An indicator $p$ is created with $p=1$ for households in domain $a$ and $p=0$ for those in the common domain $a t$. For the "unknowns", the modeled value of $p$ will be between 0 and 1 . The average value of $p$ represents the proportion of units in the population covered by the telephone frame.

Using this indicator $p$, we can now view the composite estimator as:

During integration, the weight of each household selected from the telephone frame is adjusted by ( $1-\alpha$ ) and the weight of each household selected from the area frame is adjusted by $\left[p_{i}+\alpha\left(1-p_{i}\right)\right]$. In a way, the households in the unknown portion contribute to both the common portion and the non-common portion of the estimator.

### 2.3 Collection Mode Effect

The final challenge with the dual-frame design of the CCHS is the use of multiple collection modes and the effect this has on the choice of $\alpha$. As mentioned previously, if estimates coming from each frame are the same and unbiased, then any choice of integration factor will ensure an unbiased estimate. In the case of the CCHS, the estimates coming from each frame may differ. The main reason is that the mode of collection differs for the two frames. An extensive study conducted in 2004 (Béland and St-Pierre, 2008) showed that the effect of the collection mode on the survey responses is not ignorable. Quigley (2008) looked at mode effect in the context of extra CATI samples (also called "buy-ins") added to certain health regions in the province of Quebec. In both cases, certain key indicators are affected, such as self-reported height and weight and level of physical activity. Table 2.3-1 shows the magnitude of some significant differences in estimates that were produced during the 2004 study.

Table 2.3-1
Some CCHS mode-affected variables during a 2004 study (Béland, 2008) ( $* *=\mathrm{p}<0.01$ )

| Health Indicator | CAPI (\%) | CATI (\%) | Difference (\%) |
| :--- | :---: | :---: | :---: |
| Obesity (self-reported height and weight) | 17.9 | 13.2 | $4.7^{* *}$ |
| Physical inactivity | 42.3 | 34.4 | $7.9^{* *}$ |
| Contact with medical doctors in past 12 months | 83.5 | 78.4 | $5.1^{* *}$ |

Overall, 15 of the 70 health indicators examined showed significant differences between the two collection modes. These differences may be due to the "social desirability" phenomenon, wherein respondents may modify their responses in order to be viewed in what they feel to be a more positive light. The authors of the study believe that another factor may be the "interviewer" effect, which means interviewers may change their behaviour according to the monitoring practices for each mode (as of 2009, CATI interviews are monitored while CAPI interviews are not), and this could have an impact on the responses. It is important to note that while even some area-frame interviews are conducted over the telephone (about $25 \%$ ) they may end up more similar to face-to-face interviews than to the actual CATI interviews due to the differing interview and monitoring practices.

The difference in responses collected from the two frames and two modes suggests that it is not safe to assume that the estimator $\alpha \hat{Y}_{a t}^{A}+(1-\alpha) \hat{Y}^{T}{ }_{a t}$ is an unbiased estimator for the common portion of the population since the expected value of $\hat{Y}_{a t}^{T} \neq \hat{Y}^{A}{ }_{a t} \neq Y_{a t}$. We actually end up estimating for a weighted average of the area and telephone frame responses: $Y_{a t}=\alpha Y^{A}{ }_{a t}+(1-\alpha) Y^{T}{ }_{a t}$.

### 2.3.1 Calculating the Integration Factor $\alpha$

The estimates calculated for the common portion of the frames is a function of the values reported under the telephone and area frames as well as the value of $\alpha$. So the question becomes: how to choose $\alpha$ ? Theoretically, $\alpha$ can be chosen to minimize the variance for a fixed cost, or vice versa for a particular $y$-variable as mentioned in Hartley (1962). Like most household surveys, the CCHS only creates one set of weights, which can be applied to every variable on the file. For the CCHS, there are hundreds of collected variables, of which over 25 of them are key indicators. For this reason, choosing $\alpha$ based on a single variable does not make sense as there is no particular variable more important than the others. So $\alpha$ must be chosen in another way.

Up until 2008, the value $\alpha$ varied by health region. Originally, it was based on effective sample sizes as follows:

$$
\alpha_{l}=\frac{n_{\text {atl }}^{*^{A}}}{n_{\text {atl }}^{*_{A}}+n_{\text {atl }}^{* T}}
$$

where $n_{a t l}^{* i}$ (for $i=A$ or $T$ ) was the effective sample size for HR $l$, defined as the sample size divided by the design effect for the frame at the HR level. The drawback of this method was that the time-consuming process of calculating bootstrap weights and estimates had to be performed in order to determine the design effects. A closer look at the design effects obtained from each frame at this point in the survey process suggested that the design effects from both frames were similar, implying that a simplification to a method based on raw sample sizes was sufficient (see Korn and Graubard, 1999). Thus, $\alpha$ changed to be equal to the proportion of area-frame households in the domain at as follows:

$$
\alpha_{l}=\frac{n_{\text {atl }}^{A}}{n_{\text {atl }}^{A}+n_{\text {atl }}^{T}}
$$

In theory, basing $\alpha$ on sample sizes reduces the variability of the sample weights and hence the estimates. Unfortunately, it does mean that $\alpha$ changes from one health region to another and also fluctuates over time. Letting $\alpha$ fluctuate due to changes in sample allocation or response rates may create "false" changes in estimates (or mask real changes in the population). Take the following example in Table 2.3-1, where the estimate coming from the area frame differs from the estimate coming from the telephone frame.

Table 2.3-1
Changes in estimates due to mode effect and a varying integration factor

| Year | Integration factor $(\alpha)$ | $\hat{\mathrm{Y}}_{\text {at }}^{\mathrm{A}}$ | $\hat{\mathrm{Y}}_{\text {at }}^{\mathrm{T}}$ | $\hat{\mathrm{Y}}_{\text {at }}$ |
| :---: | :---: | :---: | :---: | :---: |
| 2007 | 0.3 | 0.8 | 0.9 | 0.87 |
| 2008 | 0.5 | 0.8 | 0.9 | 0.85 |

In this hypothetical example, despite the fact that the estimates from the telephone and area frame samples remain consistent from year to year, the overall estimate coming from the common portion changes due to the changing integration factor, $\alpha$. This effect is undesirable since it does not necessarily reflect reality.

The other consideration is the fact that there is no gold standard for many of the mode-affected variables. We expect that for some variables, such as self-reported weight, the CAPI responses are closer to the truth due to the presence of an interviewer, while in other cases the anonymity of a telephone interview may yield a more truthful response. Thus, we aim to estimate an average for the two frames. This can be achieved by fixing $\alpha$ (across regions, throughout time) in order to control the proportion of weight coming from each frame regardless of the fluctuations in sample allocation. In this way, we consistently estimate the same quantity. Since $2008, \alpha$ has been fixed at 0.4 , a value equal to the overall proportion of area frame households in domain at in 2008. Empirical studies have shown that the effect of a fixed $\alpha$ on the variance of the estimates is minimal.

## 3. Conclusions

The use of a dual-frame design poses several challenges. The CCHS has adopted strategies to deal with these challenges and is constantly improving its methods. First, only the weights of the households in the overlap domain are adjusted, in order to avoid introducing bias during the integration of the samples from the two frames. Second, logistic regression is used to model the probability of having no listed telephone number, for all households with an unknown frame status. This precision also helps minimize bias. Third, an integration factor that is fixed from region to region and throughout time has been used to adjust the weights of the households in each sample. This ensures that the same quantity, which is an average of the values that would have been obtained using CAPI and CATI, is being estimated across regions and through time. With this methodology, we ensure that reliable, coherent estimates are being produced. In the end of this process we accept that we are estimating a population parameter that is a function of our telephone frame undercoverage $p$, the integration factor $\alpha$ and the population values $Y_{a}, Y_{a t}^{T}$, and $Y_{a t}^{A}$.

## Acknowledgements

The author would like to thank Yves Béland, Leon Jang and Michelle Simard for their helpful suggestions.

## References

Béland, Y. and St-Pierre, M. (2008). Mode Effects in the Canadian Community Health Survey: A Comparison of CATI and CAPI. In J. M. Lepkowski et al. (eds.) Advances in Telephone Survey Methodology, New York: Wiley, pp. 297-314.

Béland, Y., Dale, V., Dufour, J. and Hamel, M. (2005). The Canadian Community Health Survey: Building on the Success from the Past. Proceedings of the Survey Research Section, American Statistical Association, pp. 2738-2746.

Hartley, H.O. (1962). Multiple Frame Surveys. Proceedings of the Social Statistics Section, American Statistical Association, pp. 203-206.

Korn, E.L. and Graubard, B.I. (1999). Analysis of Health Surveys, New York: Wiley.
Quigley, A. (2008). Adjustments for Mode Effect Bias for the Canadian Community Health Survey. Proceedings of the Section on Government Statistics, American Statistical Association, pp. 875-882.

Sarafin, C., Thomas, S. and Simard, M. (2007). A Review of the Weighting Strategy for the Redesigned Canadian Community Health Survey. Proceedings of the Survey Methods Section, Statistical Society of Canada.

Statistics Canada (2008). Methodology of the Canadian Labour Force Survey. Statistics Canada. Cat. No. 71-526-X.

## SESSION 6C

DATABASES 1

# The Research Data Centre-in-Research Data Centre approach 

Stefan Bender and Jörg Heining ${ }^{1}$


#### Abstract

Remote data access, defined as the possibility for a researcher to access and evaluate even weakly anonymised data via a secure Internet connection from his home desktop computer at any time, has not been implemented by a German RDC so far. Privacy regulations and especially the problem of admission control are reasons why German RDCs are not able to offer their data via remote data access to the research community. Therefore, weakly anonymised data may still only be accessed through on-site use with the consequence of time-consuming and costly guest stays at a RDC.

In order to facilitate data accessibility, the Research Data Centre of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) in Nuremberg, Germany, in cooperation with the Research Data Centres of the statistical offices of the Länder, developed the so-called Research Data Centre-in-Research Data Centre (RDC-in-RDC) approach. The basic idea of the RDC-in-RDC approach is to offer researchers the ability to access Research Data Centre via a secure Internet connection from locations outside of Nuremberg. Vice versa, the microdata of the statistical offices of the Länder may be accessed through a guest stay at the Research Data Centre. Moreover, a branch of the Research Data Centre at the Institute for Social Research (ISR) of the University of Michigan in Ann Arbor is planned to enable more researchers from North America to access Research Data Centre data.


[^51]
# The Canadian Forces Portrait: A Case study on analyzing administrative and survey data to develop a personnel management resource 

Karen Daley ${ }^{1}$


#### Abstract

The Canadian Forces (CF) has a unique workforce, with its own governance, policy, and occupational requirements. As such, the personnel management system must place the right person, with the right qualifications, in the right place at the right time. In order to do this, decision-makers, planners and policy developers must have accurate, timely, and appropriate personnel data at their disposal. The CF Portrait is designed to provide this information in a single-source, high-level, sociodemographic summary. The CF Portrait is based on the analysis of administrative and survey data on Regular Force members in a broad spectrum of personnel dimensions aiming to inform policy and program development. Census data are also used to provide the larger context for internal demographic trends. This paper discusses the CF Portrait as a case study of using both administrative and survey data to create a practical descriptive resource in an applied setting.


## 1. Introduction

The objective of personnel management in the Canadian Forces (CF) is deceptively simple: to develop policy that puts the right sailor, soldier, airman and airwoman with the right qualifications in the right place at the right time (R4; Chief Military Personnel, 2009). The mechanisms that work to realize this objective are complex and interconnected, a network of policies, programs, and organizations that orchestrate and deliver the personnel management systems essential to achieving R4. In this context, a demographic profile of the CF, complemented by survey and census findings, will support senior leaders and policy developers as they strive to implement R4. Thus, the CF Portrait is designed to collate and summarize key sociodemographic information from multiple sources. To this end, information was obtained from personnel databases and departmental research in a wide variety of personnel domains. Comparison information on the Canadian population is used where available to provide a wider context for understanding internal demographic trends. Developing the portrait is a long-term project and this first iteration summarizes existing information for Regular Force personnel. Subsequent iterations will expand the portrait and identify gaps. Future CF Portraits will also include Reserve Force personnel, with a focus on identifying sources of comparable demographic information. This current CF Portrait represents a proof of concept of this approach, and was presented to the Statistics Canada International Methodology Symposium as a case study of using administrative and survey data to conduct sociodemographic analysis in support of effective personnel management. Canadian census of population information is also used to provide larger context to examine internal demographic trends.

## 2. Data Sources

### 2.1 Administrative Data

The data used in the analyses that are reported here are taken from the Human Resources Management System (HRMS) maintained by the Director Human Resources Information Management (DHRIM) and from databases maintained by the Workforce Modelling and Analysis team within Director General Military Personnel Research and Analysis. Analyses using administrative data are complemented with information from survey sources. The survey samples used for the portrait can be considered generally representative of the populations from which they were drawn. However, sub-group membership may not necessarily be similarly representative.

[^52]Reported population data were reflective of the entire enrolled CF Regular Force effective 29 January 2009 and were extracted from two sources: HRMS $(\mathrm{n}=65,219)$ and the database maintained by the Workforce Modelling and Analysis team ( $\mathrm{n}=65,252$ ).

The administrative data have several shortcomings, in that some variables are unreliable or inconsistent for analytical purposes. Because the purpose of the database is for human resource practitioners' use, the data are not necessarily ideally suited for research purposes. For example, the education field represents the course or program most recently completed rather than the highest level of education attained. To use administrative data most effectively, a deep level of familiarity with the fields and definitions is required.

### 2.2 Survey Data

Survey data are drawn from research conducted within Director General Military Personnel Research and Analysis (DGMPRA) within the Department of National Defence (DND). Because DGMPRA provides most of the personnel research for the CF, it is relatively easy to obtain survey data for the CF Portrait. Surveys are generally conducted with random samples of participants that are representative of the population overall (with some exceptions). Surveys are available on a wide range of topics, including work-life balance, retention, exit factors, etc.

Like administrative data, survey data have shortcomings. For example, the surveys are generally cross-sectional so it is difficult to compare results over time. Surveys are not always randomly sampled and there are some groups that are difficult to reach with surveys. Nevertheless, because survey data are based on attitudes and perceptions, they provide an informative and important complement to administrative data.

### 2.3 Census Data

Census data were obtained from Statistics Canada using sources available to the public. Census data are helpful for providing the larger context for internal demographic trends review. However, these data may not necessarily be comparable in terms of the population of interest without time-consuming and costly custom analyses.

## 3. Sociodemographic Analysis

The current CF Portrait is in the proof-of-concept stage of development. As such, it comprises demographic information currently available. For this symposium, four dimensions were selected to illustrate the approach: age, years of service, education, and diversity.

### 3.1 Age

With respect to age, $65 \%$ of Regular Force personnel are between 25 and 44 years of age, with a mean and median age of 35 years. The median age for officers is 37 and for non-commissioned members is 34 . The Regular Force population is older than it was in 1985, when the median age of officers and non-commissioned members was 33 and 29, respectively (Workforce Analysis and Modelling Team, 2005). The census (Statistics Canada, 2009) reveals that the Canadian population is ageing as well, with a median age of 39.5 years and only $18 \%$ of the population under the age of 15. Canada has never had so many people close to retirement and the current "replacement" ratio is 1.1 compared to 2.3 in 1976. The implication for the CF is that the recruitable population is ageing, which will have an impact on the future availability of recruits.

### 3.2 Years of Service

In terms of years of service, $33 \%$ of Regular Force members have between 5 and 14 years of service, followed by $29 \%$ with 4 years of service or fewer. The average length of service is 12.5 years. The years-of-service profile has been influenced by two main events: the Force Reduction Program of the 1990s where recruiting was severely restricted and the recruiting surge of the early 2000s (see Figures 3.1-1 and 3.1-2). The post-reduction "valley" can be seen for both officer and non-commissioned member populations, but particularly in the latter group. The
recruiting surge of the early 2000s can also be seen in the spike of personnel with less than five years of service (Workforce Modelling and Analysis Team, 2010).

Years of service is a critical factor in personnel management as it has implications for rank level, promotion and career management. The CF has a closed personnel system, which means that intake normally occurs at the entry level. As a result, the CF must grow its own leaders, and so a close watch must be kept on managing personnel as they advance through the system to ensure that sufficient personnel have the training and experience to take on leadership roles throughout the forces. When there are extremes in the years-of-service profile, such as the peaks and valleys in Figures 3.2-1 and 3.2-2, retention of personnel becomes critical to reduce the impact of personnel shortages in the 12 to 18 years-of-service range.

Figure 3.2-1 - Officer population profile by years of service and rank at 31 March 2009 (from Workforce Modelling and Analysis Team, 2010)


Figure 3.2-2
Non-commissioned member population profile by years of service and rank at 31 March 2009 (from Workforce Modelling and Analysis Team, 2010)


Given the context of an ageing force and an ageing Canadian population combined with the years-of-service profile described above, retaining personnel in the 12-18 years of service range is very important. A recent survey looked at work satisfaction and intention to leave for Regular Force members was conducted (Yelle \& Daley, 2010). Highlights of findings include $89 \%$ reporting satisfaction with the military way of life, yet mixed results for intention to leave. Specifically, when asked about their intentions to leave the CF as soon as they have completed their current terms of service, $41 \%$ agreed yet $43 \%$ disagreed. Similarly, when given the item, "I intend to stay in the Canadian Forces as long as I can," $46 \%$ agreed but $37 \%$ disagreed.

### 3.3 Education

The administrative data on educational attainment are difficult to obtain. However, the Workforce Modelling and Analysis Team has done historical analyses of administrative data to determine that education level is increasing in the CF. For example, in 2001, $13 \%$ of CF personnel had a university degree compared to $16 \%$ in 2007-08 and the proportion of personnel with less than a high school diploma dropped from $32 \%$ in 2001 to $24 \%$ in 2007-08 (Directorate Strategic Human Resources, 2003, Workforce Modelling and Analysis Team, 2009). In general, there is a difference between officers and non-commissioned members with respect to highest level of education achieved. For example, in the 2007-08 Annual Report (Workforce Modelling and Analysis Team, 2009), 56\% of commissioned officers had a Bachelor's degree and $16 \%$ had a Master's or higher. This is a reflection of the CF's decision to move towards a degreed officer corps. In contrast, $11 \%$ of non-commissioned members had a college or CEGEP diploma, $48 \%$ had a high school diploma, and $29 \%$ had less than high school.

The Canadian population has a somewhat higher level of education. In 2006, Canada was sixth among all member countries of the Organisation for Economic Co-operation and Development (OECD) in terms of the proportion of adults with a university degree, and the highest proportion of combined university and college graduates (Statistics Canada, 2009). Recent Census results (Statistics Canada, 2009) showed that $24 \%$ of adults 25 to 64 years had a high school diploma as their highest level, while $15 \%$ had less than a high school education. Young adults had a higher level of educational attainment than their older counterparts. For example, the percentage of people who have not completed high school and had no further education has declined in $2006,11 \%$ of those aged 25 to 34 had no formal
education credentials compared to $23 \%$ of those aged 55 to 64 . Similarly, $23 \%$ of people aged 25 to 34 had a college diploma, compared to $16 \%$ of 55 to 64 year olds. Furthermore, $29 \%$ of those aged 25 to 34 years had a university degree in 2006 compared to $18 \%$ aged 55 to 64 .

Statistics Canada (2009) reported that young adults are studying different trades than older generations, with shifts evident from Mechanic and Repair Technologies/Technicians to Personal and Culinary Services. What is not known is to what extent this is tied to the career aspirations of young people over labour market influences such as the decline in domestic manufacturing and the growth of the hospitality and service sector. Furthermore, a much higher percentage of women than men aged 25 to $34(33 \%$ vs. $25 \%)$ had a university degree compared to those aged 55 to $64(16 \%$ vs. $21 \%)$. Over half of recent immigrants had a university degree, which is more than twice the proportion of degree holders among the Canadian population ( $51 \%$ vs. $20 \%$ ), and also much higher than the proportion ( $28 \%$ ) among immigrants who arrived before 2001. This shift is partly due to a change of requirements implemented in 2003, which dramatically increased the number of people eligible to immigrate in the "skilled worker" category, requiring, among other things, a two-year postsecondary degree or diploma.

### 3.4 Diversity

The Canadian Forces is committed to increasing the diversity of the Forces to reflect the Canadian population. This is a complex issue, but generally speaking, visible minorities, Aboriginals, and women (groups designated by the Employment Equity Act) are underrepresented in the Canadian Forces relative to the Canadian population. The recruitable population is increasingly diverse and well-educated. For example, visible minorities were $16 \%$ of the population in the 2006 Census, up from $11 \%$ in 1996 (Statistics Canada, 2009). Women comprise a substantial portion of the labour market and are increasingly obtaining university degrees ( $33 \%$ of women vs. $25 \%$ of men aged 25 to 34, 2006 Census; Statistics Canada, 2009). The CF is researching the issue and has several programs in place to recruit and retain members of these designated groups.

## 4. Conclusion

This CF Portrait represents a proof of concept of creating a sociodemographic profile of the Regular Force using administrative data, complemented by survey and census findings. This approach and highlighted results were presented at the Statistics Canada International Methodology Symposium 2010. Combining findings from the Canadian census, survey analysis and administrative sources can help users of the Portrait to understand the results in a more meaningful way by providing a broader context for interpreting internal demographic trends. Future directions of the project aim to identify gaps in the sociodemographic profile and develop a process for the systematic production of the CF Portrait on a cyclical basis to provide policy developers and senior leaders with a useful resource that could assist evidence-based decision making in support of R4.

## References

Chief Military Personnel (in custody of Director Military Personnel Planning and Programme Coordination).(2009). The Fight of Today: Military Personnel Planning and Management Document for Fiscal Year 2009-2010, Department of National Defence, Government of Canada.

Directorate Strategic Human Resources. (2003). D Strat HR News, Strategic Outlook Series, Vol. 1, Assistant Deputy Minister (Human Resources - Military), Department of National Defence, Government of Canada.

Statistics Canada. (2009). 2006 Census Analysis Series, Statistics Canada, Government of Canada.
Workforce Analysis and Modelling Team. (2005). Annual Report on Regular Force Personnel 2003-2004, Directorate Strategic Human Resources analysis, Chief of Staff - Assistant Deputy Minister (Human Resources - Military), Department of National Defence, Government of Canada.

Workforce Analysis and Modelling Team. (2009). Annual Report on Regular Force Personnel 2007-2008, Directorate Strategic Military Personnel Research and Analysis, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

Workforce Analysis and Modelling Team. (2010). Annual Report on Regular Force Personnel 2008-2009, Directorate Research - Personnel Generation Research, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

Yelle, M. and Daley, K. (2010). Top Line Results from the Fall 2009 Your-Say Survey Core Section on Life Satisfaction, Career Management, Work-Life Balance, Imposed Restriction, Supervisors, Canadian Forces as a Whole, Career Intentions and Willingness to Deploy, Directorate Research - Organizational and Operational Dynamics, Director General Military Personnel Research and Analysis, Department of National Defence, Government of Canada.

SESSION 7 A
TREATMENT OF ITEM NONRESPONSE IN SOCIAL SURVEYS

# Imputing underreported treatments using multiple sources of treatment information in a Cancer Services Study 

Alan M. Zaslavsky, and Yulei $\mathrm{He}^{1}$


#### Abstract

Cancer registry records, patient surveys, and administrative systems record adjuvant therapies (chemotherapy and radiation) for cancer patients; however, subject to underreporting, this could bias analyses. We propose to impute true treatment status using sample validation data from medical records and analyze the imputed data. We extend earlier studies with a single outcome (provision of chemotherapy) and base data system (the registry), to multiple measures (provision of chemotherapy and radiation therapy) and multiple data systems (the registry, a patient survey, and Medicare claims). Bayesian hierarchical models for provision and reporting of multiple cancer therapies take into account their associations and multilevel structure, using related multivariate probit models for reporting of each therapy. The methodology is applied to data for patients with colorectal cancer in California.


[^53]
# Proxy pattern-mixture analysis for survey nonresponse 

Rebecca R. Andridge ${ }^{1}$ and Roderick J. A. Little ${ }^{2}$


#### Abstract

Proxy pattern-mixture analysis (PPMA) is a simple method for assessing nonresponse bias for the mean of a survey variable $Y$ subject to nonresponse, when there is a set of covariates observed for nonrespondents and respondents. PPMA combines all three key aspects of nonresponse -- the nonresponse rate, the strength of the covariates in predicting $Y$, and the difference in covariate distributions for respondents and nonrespondents -- in an intuitively reasonable way. The method does not assume the missing-data mechanism is missing at random (MAR), providing a sensitivity analysis for deviations from MAR. Maximum likelihood, Bayesian and multiple imputation versions of PPMA are considered. Properties are examined through simulation, and extensions of the analysis are outlined.


Key words: missing data, nonignorable nonresponse, sensitivity analysis

## 1. Introduction

Missing data is a common problem in sample surveys. In this paper we focus on measurement of, and adjustment for, nonresponse in a single variable $Y$ subject to missing values, when a set of variables $X$ are measured for both respondents and nonrespondents. With unit nonresponse, this set of variables is generally restricted to survey design variables, except in longitudinal surveys where variables are measured prior to dropout. With item nonresponse, the set of observed variables can include survey items not subject to nonresponse, and hence is potentially more extensive. With a set of variables $Y$ subject to nonresponse, our methods could be applied separately for each variable, but we do not consider here methods for multivariate missing data where variables are missing for different sets of cases.

Limiting the impact of nonresponse is an important design goal in survey research, and how to measure and adjust for nonresponse is an important issue for statistical agencies and other data collectors, particularly since response rates are on the decline. However, specific analysis recommendations are lacking, focusing on methods for accurately calculating response rates. While the response rate is clearly an important feature of the problem, there is a tension between increasing response rates and increasing response error by including respondents with no inclination to respond accurately. Indeed, some studies have shown that response rates are a poor measure of nonresponse bias (Curtain et al., 2000, Keeter et al., 2000).

There are three major components to consider in evaluating nonresponse: the amount of missing data, differences between respondents and nonrespondents on characteristics that are observed for the entire sample, and the relationship between these fully observed covariates and the survey outcome of interest. Each facet provides some information about the impact of nonresponse, but no single component completely tells the story.

Historically the amount of missing data, as measured by the response rate, has been the most oft-used metric for evaluating survey quality. However, response rates ignore the information contained in auxiliary covariates observed for nonrespondents. Federal reports have recommended the second component, evaluating nonresponse based on differences between respondents and nonrespondents (Federal Committee on Statistical Methodology, 2001). A related approach is to focus on measures based on the response propensity, the estimated probability of response given the covariates, which is the auxiliary variable that is most different between respondents and nonrespondents. In particular, Schouten, Cobben, and Bethlehem (2009) propose the use of R- indicators to assess the

[^54]"representativeness" of respondents with respect to the complete sample, and Sarndal and Lundstrom (2005, 2008) propose a q2 measure based on the variability of the inverse estimated response propensity. Though response propensity-based measures are appealing, nonresponse bias depends on the strength of the correlation between the survey variable of interest and the probability of response, and bias will vary across items in a single survey (Bethlehem, 2002, Groves, 2006).

The final component is the value of the auxiliary information in predicting survey outcomes, as measured by the reduction in variance of regression estimate relative to the sample mean (see Eq. 7.14 in Little and Rubin, 2002). This approach is attractive since it gives appropriate credit to the availability of good predictors of $Y$ in the auxiliary data as well as a high response rate, and arguably good prediction of the survey outcomes is a key feature of good covariates; in particular, conditioning on a covariate $Z$ that is a good predictor of nonresponse but is unrelated to survey outcomes simply results in increased variance without any reduction in bias (Little and Vartivarian, 2005). A limitation with this approach is that it is more focused on precision than bias, and it assumes the data are missing at random (MAR); that is, missingness of $Y$ is independent of $Y$ after conditioning on the covariates $Z$ (Rubin, 1976). This approach cannot provide a single measure of the impact of nonresponse, since by definition measures are outcome-specific.

Previous work has focused on distinct measures based on these considerations, but in our view has not integrated them in a satisfactory way. We propose a new method for nonresponse bias measurement and adjustment that takes account all three aspects, in a way which we find intuitive and satisfying. Section 2 described the pattern-mixture model which is the basis for the proposed analysis, and maximum likelihood (ML), Bayes and multiple imputation under the model. Section 3 describes a set of simulation studies to demonstrate the assessment of nonresponse bias using these methods. Section 4 presents discussion, including extensions of the proposed method. A more comprehensive discussion of the approach is given in Andridge and Little (2010).

## 2. Proxy Pattern-Mixture Model

We consider the problem of assessing nonresponse bias for estimating the mean of a survey variable $Y$ subject to nonresponse. For simplicity, we initially consider an infinite population with a sample of size $n$ drawn by simple random sampling. Let $Y_{i}$ denote the value of a continuous survey outcome and $Z_{i}=\left(Z_{i 1}, Z_{i 2}, \ldots, Z_{i p}\right)$ denote the values of $p$ covariates for unit $i$ in the sample. Only $r$ of the $n$ sampled units respond, so observed data consist of $\left(Y_{i}, Z_{i}\right)$ for $i=1, \ldots, r$ and $Z_{i}$ for $i=r+1, \ldots, n$. Let $M$ denote the missingness indicator, such that for unit $i$ $M_{i}=0$ if $Y_{i}$ is observed and $M_{i}=1$ if $Y_{i}$ is missing.

To reduce dimensionality, we replace $Z$ by a single proxy variable $X$ that has the highest correlation with $Y$. This proxy variable can be estimated by regressing $Y$ on $Z$ using the respondent data, including important predictors of $Y$, as well as interactions and nonlinear terms where appropriate. Specifically, we assume the regression model $E(Y \mid Z, M=0)=\alpha_{0}+\alpha Z$, and let $X=\alpha Z$. The regression coefficients $\alpha$ are subject to sampling error, so in practice $X$ is estimated rather than known, but we address this complication later.

We model the joint distribution of $X, Y$ and $M$ using the following proxy pattern-mixture model, similar in form to that discussed in Little (1994):

$$
\begin{aligned}
& (X, Y \mid M=m) \sim N_{2}\left(\left(\mu_{x}^{(m)}, \mu_{y}^{(m)}\right), \Sigma^{(m)}\right) \\
& M \sim \operatorname{Bernoulli}(1-\pi) \\
& \Sigma^{(m)}=\left(\begin{array}{cc}
\sigma_{x x}^{(m)} & \rho^{(m)} \sqrt{\sigma_{x x}^{(m)} \sigma_{y y}^{(m)}} \\
\rho^{(m)} \sqrt{\sigma_{x x}^{(m)} \sigma_{y y}^{(m)}} & \sigma_{y y}^{(m)}
\end{array}\right)
\end{aligned}
$$

where $N_{2}$ denotes the bivariate normal distribution. Our interest is the marginal mean of $Y$, which can be expressed as $\mu_{y}=\pi \mu_{y}^{(0)}+(1-\pi) \mu_{y}^{(1)}$. This model is under-identified, since there is no information on the conditional normal
distribution for $Y$ given $X$ for nonrespondents ( $\boldsymbol{M}=\mathbf{1}$ ). However, Little (1994) shows that the model can be identified by making assumptions about how missingness of $Y$ depends on $Y$ and $X$. Specifically, we assume that

$$
\operatorname{Pr}(M=1 \mid Y, X)=f\left(X^{*}+\lambda Y\right),
$$

for some unspecified function $f$ and known constant $\lambda$. Here $X^{*}=X \sqrt{\sigma_{y y}^{(0)} / \sigma_{x x}^{(0)}}$ is the proxy $X$ scaled to have the same variance as $Y$, which aids the interpretation of $\lambda$ by putting $X$ and $Y$ on the same scale. This mechanism is MAR when $\lambda=0$, and deviates increasingly from MAR as $\lambda$ increases. With this assumption, the parameters are just identified by the condition that

$$
\left((Y, X) \perp M \mid X^{*}+\lambda Y\right)
$$

where $\perp$ denotes independence. Arguments similar to Little (2004) yield explicit form ML estimates of the parameters for any given value of $\lambda$. In particular, the ML estimate of the mean of $Y$, averaging over patterns, as

$$
\hat{\mu}_{y}=\bar{y}_{R}+\left(\frac{n-r}{n}\right)\left(\frac{\lambda+\hat{\rho}}{\lambda \hat{\rho}+1}\right)\left(\bar{x}_{N R}-\bar{x}_{R}\right),
$$

where $\hat{\rho}$ is the respondent sample correlation. Note that the adjustment of the respondent mean $\bar{y}_{R}$ in this estimate incorporates the three key nonresponse factors we mentioned in the introduction - the nonresponse rate $(n-r) / n$, the correlation $\hat{\rho}$ between $X$ and $Y$, and the deviation of $\bar{x}_{N R}$ from $\bar{x}_{R}$, in a very simple and intuitive way. The adjustment increases with the nonresponse rate and the deviation of the mean of $X$ for respondents and nonrespondents.

The middle factor depends on the choice of $\lambda$. We would like to be able to estimate this parameter, but unfortunately there is no information about it in the data -- this is generally true for methods that model deviations from MAR. Following Little (1994), we propose a sensitivity analysis, where estimates are generated for a range of values of $\lambda$ between 0 and infinity, specifically 0,1 and infinity; the choice of 0 corresponds to MAR, the intermediate choice of 1 implies the bias of $Y$ is the same as the bias of $X^{*}$, and the choice of infinity is the most extreme deviation from MAR; estimates for this case have the highest variance. As $\lambda$ varies between 0 and infinity, the middle factor $(\lambda+\hat{\rho}) /(\lambda \hat{\rho}+1)$ varies between $\hat{\rho}$ (when $\hat{\mu}_{y}$ is the standard regression estimator of the mean) and $1 / \hat{\rho}$ (when $\hat{\mu}_{y}$ is the inverse regression estimator proposed by Brown (1990)). The sensitivity of the estimate to the choice of $\lambda$ is small when $\hat{\rho}$ is close to 1 , that is we have a strong proxy variable, and large when $\hat{\rho}$ is close to 0 , that is we have a weak proxy variable.

The estimator described here is maximum likelihood (ML) for the pattern-mixture model. Large-sample variances are given by Taylor series calculations as in Little (1994), though this approximation may not be appropriate for small samples. Additionally, the ML estimate and corresponding inference do not take into account the fact that the regression coefficients that determine $X$ are subject to sampling error. Better methods incorporate this uncertainty, such as Bayesian methods and multiple imputation. The latter creates $K$ complete data sets by filling in missing $Y$ values with draws from the posterior distribution, based on the pattern-mixture model. Draws from the posterior distribution of $Y$ for a particular choice of $\lambda$ are obtained by first drawing the parameters from their posterior distributions, and then drawing the missing values of $Y$ based on the conditional distribution of $Y$ given $X$ for nonrespondents $(M=1)$. Details of these methods are given in Andridge and Little (2010).

An advantage of the multiple imputation approach is that complex design features like clustering, stratification and unequal sampling probabilities can be incorporated in the within-imputation variance component of the multiple imputation inference. Once the imputation process has created complete data sets, design-based methods can be used to estimate $\mu_{y}$ and its variance. Incorporating complex design features into the model and applying maximum
likelihood or Bayesian methods is less straightforward, though arguably more principled. See for example Little (2004) for more discussion.

## 3. Simulation Studies

We now describe a simulation study designed to (1) illustrate the PPMA sensitivity analysis for a range of values of the parameters, and (2) assess confidence coverage of ML, Bayes and MI inferences. Other simulations in Andridge and Little (2010) demonstrate robustness of the PPM model when data arise from a selection model with a range of nonresponse mechanisms. All simulations and data analysis were performed using the software package R (CRAN).

We created a total of eighteen artificial data sets in a $3 \times 3 \times 2$ factorial design. A single data set was generated for each combination of $\rho=\{0.8,0.5,0.2), d^{*} \equiv\left(\mu_{x}-\mu_{x R}\right) / \sqrt{\sigma_{x x}}=(0.1,0.3,0.5)$ and $n=(100,400)$ as follows. A single covariate $Z$ was generated for both respondents and nonrespondents with the outcome $Y$ generated only for respondents. Respondent data were created as pairs $\left(z_{i}, y_{i}\right), i=1 \ldots r$ with $z_{i} \sim N\left(0, \rho^{2}\right)$ and $y_{i}=1+z_{i}+e_{i}$, where $e_{i} \sim N\left(0,1-\rho^{2}\right)$. Nonrespondent data were Z's only, generated from $z_{i} \sim N\left(2 \rho d^{*}, \rho^{2}\right)$ for $i=r+1 \ldots n$. The nonresponse rate was fixed at $50 \%$. This data structure was chosen so that the variance of the complete case mean would be constant (and equal to one) across different choices of $\rho$ and $d^{*}$, and so that varying $\rho$ would not affect $d^{*}$ and vice-versa. $R^{2}$ values that corresponded to the selected $\rho$ were $64 \%, 25 \%$, and $4 \%$, covering a range likely to be encountered in practice.

For each of the eighteen data sets, estimates of the mean of $Y$ and its precision were obtained for $\lambda=(0,1, \infty)$. For each value of $\lambda$, three $95 \%$ intervals were calculated:
(A) ML: the maximum likelihood estimate $+/-2$ standard errors (large-sample approximation)
(B) PD: the posterior median and 2.5 th to 97.5 th posterior interval based on 5000 draws from the posterior distribution of $\mu_{y}$
(C) MI: mean +/- 2 standard errors from 20 multiply imputed data sets.

Posterior median and quantiles were used because initial evaluations showed that the posterior distribution of $\mu_{y}$ was skewed and had extreme outliers for small $\rho$ and large $\lambda$. The complete-case estimate $+/-2$ standard errors was also computed for each data set; note that the expected value of the respondent mean and corresponding confidence interval is constant across all values of $\rho$ and $d^{*}$ for each $n$.

Results from applying the three estimation methods to each of the nine data sets with $n=100$ are displayed in Figure 1. The complete-case estimate is shown alongside $95 \%$ intervals estimated by maximum likelihood, multiple imputation, and the posterior distribution, for $\lambda=(0,1, \infty)$. For each population the PD intervals are longer than the ML and MI intervals for all choices of $\rho$ and $\lambda$, especially for weak proxies and $\lambda=\infty$. Results for $n=400$ were similar and are not shown.

Populations with a strong proxy $(\rho=0.8)$ do not show much variation across values of $\lambda$; there is evidence that nonresponse bias is small for small $d^{*}$ and there is good information to correct the potential bias for larger values of $d$. For moderately strong proxies $(\rho=0.5)$ the intervals increase in length, with differences between PD and ML becoming more exaggerated as $d^{*}$ increases. As expected, when the proxy is weak ( $\rho=0.2$ ) we see large intervals for models that assume missingness is not at random $(\lambda \neq 0)$; this reflects the fact that we are in the worst-case scenario where there is not much information in the proxy to estimate the nonresponse bias. Notice that in this simulation the true mean of $Y$ is not known; we simply illustrate the effect of various values of $\rho$ and $\lambda$ on the sensitivity analysis.

The second objective of the simulation was to assess coverage properties for each of the three estimation methods. We generated 500 replicate data sets as before for each of the eighteen population designs and computed the actual coverage of a nominal $95 \%$ interval and median interval length. The Bayesian intervals were based on 1000 draws from the posterior distribution. Coverage is based on the unreasonable assumption that the assumed value of $\lambda$ equals the actual value of $\lambda$. This is unrealistic, but coverages are clearly not valid when the value of $\lambda$ is misspecified, and uncertainty in the choice of $\lambda$ is captured by the sensitivity analysis.

Table 1 displays the nominal coverage and median CI width for each of the eighteen populations. For populations with a strong or moderately strong proxy $(\rho=0.5,0.8)$ coverage is at or above nominal levels for all three methods, for both the smaller and larger sample sizes and for all levels of $d$. For these populations, PD inference is slightly more conservative; intervals are larger than ML for most populations. However, when the proxy is weak, ML coverage is below nominal levels for larger values of $\lambda$, while both PD and MI have coverage close to nominal levels. With small sample size and weak proxies, taking $\lambda=\infty$ leads to large confidence intervals. The $\lambda=\infty$ model requires a strong proxy or large sample size to provide reliable estimates of $\mu_{y}$.

## 4. Discussion

The PPM analysis for survey nonresponse has the following attractive features: it integrates all the various components of nonresponse noted in the introduction into a single sensitivity analysis. It is easy to implement, since the ML form is simple to compute, and the Bayesian implementation is noniterative, not requiring iterative Markov Chain Monte Carlo methods that pervade more complex Bayesian methods and might deter survey practitioners; the MI method is also non-iterative, and it readily allows complex design features to be incorporated in the withinimputation component of variance. The MI model for the between-imputation component proposed here does allow survey design variables to be included as predictors in the regression for the proxy measure, but does not include random effects to model clustering of the sample -- a more principled extension would incorporate such design features directly into the imputation model.

Since most nonresponse adjustments applied in the survey setting assume MAR, and MAR is often a strong and questionable assumption with unit nonresponse, an important advantage of PPM analysis is that it does not assume MAR. We believe that it provides a picture of the potential nonresponse bias under a reasonable range of MAR and non-MAR mechanisms. It gives appropriate credit to the existence of good predictors of the observed outcomes, since the sensitivity analysis is less variable in this case. The analysis reinforces the idea that emphasis at the design stage should be on collection of strong auxiliary data to help evaluate and adjust for potential nonresponse, not solely on obtaining the highest possible response rate.

The PPM method employs a sensitivity analysis to assess deviations from MAR, in contrast with some selection model approaches that attempt to use the data to estimate parameters that capture deviations from MAR (Heckman, 1976). These models are technically identified in situations where pattern-mixture models are not, but estimation of the NMAR parameters is still based on strong and unverifiable structural and distributional assumptions, and a substantial number of researchers believe that a sensitivity analysis is the right approach. The assumptions about deviations from MAR are more transparent in the pattern-mixture factorization, since differences between respondents and nonrespondents are directly modeled (Little and Rubin, 2002). The PPM sensitivity analysis only varies one sensitivity parameter, $\lambda$, but still manages to capture a range of assumptions on the missing data mechanism. Both the standard and inverse regression estimators are contained in the PPM analysis framework. A fraction of missing information can be defined based on the PPM model to measure the impact of nonresponse this aspect is described in Andridge and Little (2010).

A limitation of PPM analysis is that by reducing the auxiliary data to the single proxy $X$, the coefficient $\lambda$ is not associated with any particular covariate and hence is difficult to interpret, since the effects on missingness on individual covariates $Z_{j}$ are lost. The pattern-mixture model proposed by Daniels and Hogan (2000) in the context of longitudinal data, uses a location-scale parameterization to model differences in the marginal distribution of $Y, Z$ for respondents and nonrespondents. This model is more readily interpretable than our approach, but it is very
underidentified, even with a single $Z$ it has three unidentified parameters, and additional specification is needed to limit the number of parameters to be varied in a sensitivity analysis. Our approach trades off interpretability for parsimony, allowing a single parameter to model deviations from MAR.

Another limitation of our analysis is that it focuses only on the mean of a particular outcome $Y$, so it is outcomespecific. Thus, in a typical survey with many outcomes, the analysis needs to be repeated on each of the key outcomes of interest and then integrated in some way that reflects the relative importance of these outcomes. This complication seems unavoidable, since nonresponse bias is small for variables unrelated to nonresponse, and potentially larger for variables related to nonresponse. Measures that do not incorporate relationships with outcomes, like measures based on the nonresponse weights, cannot capture this dimension of the problem.

The pattern-mixture model that justifies the proposed analysis strictly only applies to continuous survey variables, where normality is reasonable, although we feel it is still informative when applied to non-normal outcomes. Extensions to categorical variables are possible via probit models, and many other extensions can be envisaged, including extensions to other generalized linear models. PPM analysis can be applied to handle item nonresponse by treating each item subject to missing data separately, and restricting the covariates to variables that are fully observed. However, this approach does not condition fully on the observed information, and extensions for general patterns of missing data would be preferable. Our future work on PPM analysis will focus on developing these extensions.

## References

Andridge, R.H. and Little, R.J. (2010). Proxy pattern-mixture analysis for survey nonresponse, submitted to Journal of Official Statistics.

Bethlehem, J. (2002). Weighting Nonresponse Adjustments Based on Auxiliary Information, Chapter 18 in Survey Nonresponse, eds. Groves, R., Dillman, D., Eltinge, J., and Little, R., New York: Wiley, pp. 275-287.

Brown, C.H. (1990). Protecting Against Nonrandomly Missing Data in Longitudinal Studies, Biometrics, 46, pp. 143-155.

Curtain, R., Presser, S. and Singer, E. (2000). The Effects of Response Rate Changes on the Index of Consumer Sentiment, Public Opinion Quarterly, 64, pp. 413-428. \}

Daniels, M.J. and Hogan, J.W. (2000). Reparameterizing the Pattern Mixture Model for Sensitivity Analyses under Informative Dropout, Biometrics, 56, pp. 1241-1248.

Federal Committee on Statistical Methodology (2001). Measuring and Reporting Sources of Error in Surveys, Statistical Policy Working Paper 31, Executive Office of the President of the United States of America.

Groves, R.M. (2006). Nonresponse Rates and Nonresponse Bias in Household Surveys, Public Opinion Quarterly, 70, pp. 646-675.

Heckman, J.J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models, Annals of Economic and Social Measurement, 5, pp. 475-492.

Keeter, S., Miller, C., Kohut, A., Groves, R.-M., and Presser, S. (2000). Consequences of Reducing Nonresponse in a National Telephone Survey, Public Opinion Quarterly, 64, pp. 125-148.

Little, R.J. and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means?, Survey Methodology, 31, pp. 161-168.

Little, R.J.A. (1994). A Class of Pattern-Mixture Models for Normal Incomplete Data, Biometrika, 81, pp. 471-483.

Little, R.J. (2004). To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, Journal of the American Statistical Association, 99, pp. 546-556. \}

Little, R.J.A. and Rubin, D.-B. (2002). Statistical Analysis with Missing Data, 2nd ed., Wiley: New York.
Rubin, D.B. (1976). Inference and Missing Data (with Discussion), Biometrika, 63, pp. 581-592.

Sarndal, C.E. and Lundstrom, S. (2005). Estimation in Surveys with Nonresponse. Wiley: New York.
Sarndal, C.E. and Lundstrom, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator, Journal of Official Statistics, 24, pp. 167-191.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the Representativeness or Survey Response, Survey Methodology, 35, pp. 101-113.

Table 1. Coverage and median confidence interval length for eighteen artificial populations. ML: Maximum likelihood; PD: Posterior distribution; MI: 20 multiply imputed data sets. Results over 500 replicates.

| Population |  |  | $\mathrm{n}=100$ |  |  |  |  |  | $\mathrm{n}=400$ |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | Coverage |  |  | CI Width |  |  | Coverage |  |  | CI Width |  |  |
| $\rho$ | $d$ | $\lambda$ | ML | PD | MI | ML | PD | MI | ML | PD | MI | ML | PD | MI |
| 0.8 | 0.1 | 0 | 93 | 94 | 93 | 0.46 | 0.47 | 0.47 | 95 | 94 | 94 | 0.23 | 0.23 | 0.23 |
|  |  | 1 | 95 | 95 | 95 | 0.47 | 0.48 | 0.48 | 95 | 95 | 95 | 0.24 | 0.24 | 0.24 |
|  |  | $\infty$ | 95 | 95 | 96 | 0.51 | 0.52 | 0.52 | 95 | 95 | 94 | 0.25 | 0.25 | 0.25 |
| 0.8 | 0.3 | 0 | 94 | 94 | 94 | 0.48 | 0.49 | 0.49 | 96 | 95 | 95 | 0.24 | 0.24 | 0.24 |
|  |  | 1 | 96 | 96 | 96 | 0.50 | 0.51 | 0.51 | 96 | 95 | 96 | 0.25 | 0.25 | 0.25 |
|  |  | $\infty$ | 96 | 95 | 96 | 0.55 | 0.56 | 0.56 | 96 | 95 | 96 | 0.27 | 0.27 | 0.27 |
| 0.8 | 0.5 | 0 | 95 | 96 | 95 | 0.52 | 0.53 | 0.53 | 96 | 95 | 95 | 0.26 | 0.26 | 0.26 |
|  |  | 1 | 96 | 97 | 96 | 0.54 | 0.56 | 0.55 | 96 | 95 | 97 | 0.27 | 0.27 | 0.27 |
|  |  | $\infty$ | 97 | 96 | 97 | 0.62 | 0.64 | 0.64 | 97 | 96 | 96 | 0.31 | 0.31 | 0.31 |
| 0.5 | 0.1 | 0 | 93 | 93 | 93 | 0.52 | 0.53 | 0.54 | 94 | 93 | 94 | 0.26 | 0.26 | 0.27 |
|  |  | 1 | 95 | 96 | 95 | 0.56 | 0.59 | 0.59 | 95 | 95 | 96 | 0.29 | 0.28 | 0.29 |
|  |  | $\infty$ | 97 | 95 | 97 | 0.84 | 0.98 | 0.96 | 96 | 95 | 95 | 0.41 | 0.42 | 0.43 |
| 0.5 | 0.3 | 0 | 93 | 94 | 94 | 0.54 | 0.56 | 0.56 | 94 | 94 | 94 | 0.27 | 0.27 | 0.28 |
|  |  | 1 | 96 | 97 | 96 | 0.59 | 0.64 | 0.64 | 95 | 95 | 96 | 0.3 | 0.31 | 0.31 |
|  |  | $\infty$ | 96 | 96 | 97 | 1.0 | 1.2 | 1.2 | 95 | 95 | 96 | 0.51 | 0.52 | 0.53 |
| 0.5 | 0.5 | 0 | 95 | 95 | 95 | 0.58 | 0.6 | 0.61 | 94 | 94 | 95 | 0.29 | 0.29 | 0.3 |
|  |  | 1 | 97 | 97 | 98 | 0.64 | 0.73 | 0.72 | 95 | 96 | 97 | 0.33 | 0.35 | 0.35 |
|  |  | $\infty$ | 96 | 97 | 96 | 1.3 | 1.6 | 1.6 | 97 | 96 | 96 | 0.66 | 0.68 | 0.69 |
| 0.2 | 0.1 | 0 | 93 | 94 | 94 | 0.55 | 0.56 | 0.57 | 94 | 93 | 93 | 0.28 | 0.27 | 0.28 |
|  |  | 1 | 94 | 96 | 96 | 0.64 | 0.72 | 0.72 | 95 | 95 | 95 | 0.33 | 0.33 | 0.34 |
|  |  | $\infty$ | 94 | 97 | 97 | 2.5 | 9.9 | 9.0 | 94 | 97 | 96 | 1.2 | 1.7 | 1.6 |
| 0.2 | 0.3 | 0 | 94 | 95 | 94 | 0.57 | 0.59 | 0.6 | 95 | 94 | 94 | 0.29 | 0.29 | 0.29 |
|  |  | 1 | 87 | 96 | 94 | 0.66 | 0.98 | 0.97 | 95 | 96 | 97 | 0.34 | 0.38 | 0.38 |
|  |  | $\infty$ | 87 | 96 | 93 | 4.7 | 23 | 19 | 90 | 97 | 94 | 2.3 | 3.4 | 3.3 |
| 0.2 | 0.5 | 0 | 95 | 95 | 95 | 0.62 | 0.63 | 0.65 | 96 | 95 | 94 | 0.31 | 0.31 | 0.32 |
|  |  | 1 | 86 | 98 | 97 | 0.73 | 1.7 | 1.4 | 95 | 97 | 98 | 0.36 | 0.45 | 0.45 |
|  |  | $\infty$ | 85 | 96 | 94 | 7.4 | 39 | 32 | 90 | 97 | 96 | 3.5 | 5.6 | 5.3 |

Bolded coverages are below 1.96 simulation standard errors.

Figure 1. $95 \%$ confidence intervals for nine generated data sets $(\mathrm{n}=100)$ for $\lambda=(0,1, \infty)$ and $d^{*} \equiv\left(\mu_{x}-\mu_{x R}\right) / \sqrt{\sigma_{x x}}=(0.1,0.3,0.5)$. Numbers below intervals are the interval length. CC: complete-case analysis; ML: maximum likelihood; PD: posterior distribution; MI: 20 multiply-imputed data sets.


Proceedings of Statistics Canada Symposium 2010
Social Statistics: The Interplay among Censuses, Surveys and Administrative Data

# Calibrated robust imputation in surveys 

David Haziza and Pierre Duchesne ${ }^{1}$


#### Abstract

Deterministic regression imputation within classes that include ratio and mean imputation within classes as special cases is widely used in surveys. It consists of replacing a missing value by its predicted value obtained under an assumed linear regression model. However, in the presence of outliers, deterministic regression imputation will potentially lead to very unstable estimators. To overcome this problem, we propose to derive a set of imputed values such that the imputed estimator (defined as the weighted sum of the observed and imputed values) is calibrated on an estimator that is known to have good properties in the presence of outliers. More precisely, the idea is to start with initial imputed values and find a final set of imputed values as close as possible to the initial ones so that the imputed estimator is calibrated on an appropriate estimator. This is closely related to the reverse calibration proposed by Chambers and Ren (2003). Results from an empirical study on the performance of the resulting estimator will be shown.


[^55]
## SESSION 7B

## ADMINISTRATIVE CENSUSES

# Methodology in the Swedish register-based census 

Martin Axelson, Dan Hedlin, Anders Holmberg, Ingegerd Jansson ${ }^{1}$


#### Abstract

All European Union members will conduct a census with a reference day in 2011. Statistics Sweden faces the challenge of conducting Sweden's first fully register-based census. The existing population register and the real property register and the new register of dwellings, which is not yet complete, will be matched to allow us to estimate, for example, distributions of variables such as the ratio of living space to the number of occupants. Issues include statistical matching, disclosure control, effective editing methods for categorical data and 'unit editing' (i.e. whether the right units have been identified), and evaluation of model assumptions and other quality aspects of the register-based methodology.


Key Words: administrative register, statistical register, register of dwellings.

## 1. Introduction

In many European countries, including Sweden, conducting a traditional census is politically sensitive due to large costs and perceived response burden. The last traditional census in Sweden was conducted in 1990, with data being collected by a self-administered mail-out mail-back questionnaire. At that time, a register-based census was not feasible owing to the lack of a link between the Population register and the Real Property register. As early as in 1995, the Swedish parliament decided that the next census should be completely register-based. For several reasons, among them political concerns of privacy, the necessary legal regulation was not in place until more than ten years later. In 2007, the Swedish parliament passed a bill on the creation of a new register of dwellings (flat and houses) which will be linked to both the Population and the Real Property registers.

For the first time all European member states are going to conduct a census in the same year and with the same variables. The reference year is 2011, and the member states may chose any day during the year as reference day. Statistics Sweden has chosen 31 December as the reference day, thus allowing time for the necessary registers to be properly updated during 2011. The deadline for data to be sent to Eurostat, the statistical office of the European Union, is end of March 2014.

Although the Census 2011 is an important task, the main advantage for Statistics Sweden of building a complete system of registers on individuals and real property is the possibility to improve the official statistics on households and housing. Statistics Sweden will be able to provide official statistics in areas where there has been considerable shortage so far.

## 2. Register-based statistics

Although vastly experienced in register-based statistics, Statistics Sweden faces new challenges. Register-based statistics differ from traditional surveys in many respects. A register-based survey utilises a register created and maintained for administrative purposes. In a register-based survey a 'statistical register' is first created where the administrative data are edited and transformed to best meet the aims of multiple surveys. Other typical processes

[^56]include merger of several administrative registers where issues such as statistical matching and derivation of new variables have to be addressed.

A main difference between register-based surveys and traditional surveys is the lack of control of the data collection process. The statistical agency has to rely on another authority for data collection, and hence the effects might include inadequate definitions of variables and reference periods, reporting delays, or lack of relevance and validity. As of yet, there is no unifying theory for administrative data as a source of data collection, but there is a clear notion of the necessity of such theory. Some recent references are Wallgren and Wallgren (2007) and Zhang (2009; in press).

The common way to gain insight into the quality of the data and the data collection process is to conduct a population enumeration survey (PES), thereby converting the solely register-based survey to a partly conventional sample survey. Statistics Sweden will opt out from a PES since with a PES most of the gain with a register-based census will be lost. The challenge is to assess quality without a PES.

## 3. Input of the census

### 3.1 Existing registers

The statistical registers of Statistics Sweden that will be used in the census are mainly utilizing administrative data from the Swedish Tax Agency and the Swedish mapping, cadastral and land surveying authority. Data on businesses and schools will also be made use of. The system of statistical registers relies on three core registers: the Business register, the Population register, and the Real Property register. The core registers are linked to various subject matter registers such as registers of employment, occupation, education, and buildings.

### 3.2 The missing link

The system of statistical registers is dependent on the possibility to uniquely link information to objects. Comprehensive identification of persons and businesses is an essential part of the system. Every individual on the Swedish population register has a personal identification number, widely used in all major administrative registers. The personal identification number links an individual to data on for example occupation and education. The personal identification number and the identification number of businesses give the prerequisites for creating statistics on employment. Businesses and individuals are linked to the Real Property register by the address of the house or building where they reside. For those living in houses the address is unique. For apartment buildings, the address until recently would only tell the entrance and possibly the floor, but not the flat. With the new register of dwellings, a unique identification number is given to all dwellings (flats and houses). The dwelling number will give Statistics Sweden new opportunities to create statistics on households and housing conditions. Since cohabiting without being married is common in Sweden, there has previously not been any administrative data covering households.

### 3.3 Creating the register of dwellings

The relevant register building process can very briefly be outlined in the following steps.

1. The local administration checks and updates all addresses within a municipality and issues formal addresses for example to dwellings that have only traditional, informal names or to houses sharing one address.
2. The property owner labels flats in a block of flats and submits the 'flat numbers' to the land surveying authority where the administrative register of addresses and dwellings is kept. The labelling should be done according to specific rules since the numbers carry information about the ordering of flats on a floor.
3. The property owner informs the residents of their flat numbers.
4. By a mailed form to every adult in the country, the Tax Agency asks individuals about their address of residence, including flat number.
5. Individuals living in apartments are expected to be aware of their flat number and to give this information to the Tax Agency where the administrative population register is updated.

Steps 1-5 are subject to error. The apartment number is expected to be particularly prone to be missing or incorrect.
The registers will only be joined by Statistics Sweden in the instance when needed. The final, merged register will not be kept as such.

## 4. Disclosure control

Another challenge is disclosure control. Data are to be delivered in "hypercubes" (multidimensional tables) defined by Eurostat. Feasible measures for risk assessment and methods of protection must be decided on, taking into account the multidimensional structure of the tables. In addition, there are practical and technical issues. Which tool to use (or develop), how to ensure consistency of tables, and when to perform disclosure control (in national databases or "on the fly") are examples of such issues (Lanzieri \& Schulte Nordholt 2009). Eurostat requires that data are protected from disclosure, but each country has its own regulation on confidentiality and secrecy which also must be respected.

## 5. Error sources

Many of the variables in the census do not depend on the merging of registers since they are derived from a single register. Challenging variables include overcrowded living conditions, which is defined in terms of the ratio of number of occupants and living space. As an example, let $m$ be the number of people living in the accommodation minus 1 and $v$ the number of rooms minus one, excluding the kitchen. When $m / v>1$ for a household with a cohabiting couple, the household is said to live in overcrowded conditions. For households without a cohabiting couple $m$ is the number of people living in the accommodation, and the limit of $m / v$ for overcrowdedness is still 1 . The numerator $m$ will be taken from the Population register through the Register of dwellings and the denominator $v$ from the Real estate register.

In particular the number of occupants will be subject to missingness and error due to missing or faulty flat numbers in the population register. That is, there will be people who according to the administrative register do not belong to any dwelling despite their address (house number and street) being known. Thus, the task is to allocate these people to flats given auxiliary information on number and sizes of flats at the relevant addresses, and the marital status, surnames, ages, etc of the people involved. There will also be flats without registered occupants.

Some of the errors in, for example, number of occupants may be easily localised if they are unusual in combination with other data. 'Inliers', that is, data that are incorrect but in all respects similar in their multivariate distribution to other data will, as always, be hard to identify. However, even if errors are localised, correcting them by making callbacks may not be cost-efficient. For continuous data selective editing has been proven effective. For categorical data however, there is less research. A similar and rather little researched issue is what may be referred to as 'unit editing', that is, localising missing or misidentified units.

Unmarried individuals with or without children and without valid flat numbers can be randomly allocated to flats with probabilities matching frequencies of individuals with valid flat numbers. A non-settled issue is how to define areas to compute frequencies within. An alternative is to repeat the random allocation $M$ times and view it as multiple imputation of individuals and thus allow for an estimate of imputation variance.

Eurostat requires the member states to submit a description and evaluation of the quality of the census data. Different aspects of quality in official statistics, such as accuracy, efficiency, and consistency are defined in terms of sampling, and there is a need of a corresponding set of concepts for the description of quality in register-based statistics. Statistics Sweden has initiated some work on this part. A thorough description of the data collection process will be part of the documentation. By describing the process in detail, sources of errors are detected and methods for handling them can be applied. The time aspect is important. Data will be merged at different stages of the process and edits and other checks of consistency, confidentiality, etc. must be carried out when it is most efficient in order to avoid overlapping work or other mistakes. In particular it is important to monitor the output from item 5 above. In the early stages of 2011 there are chances to identify and correct errors in the building process of the new registers.

An opportunity under consideration is to utilize the data collection in ongoing surveys e.g. the LFS for comparisons and checks against the register data.

In summary, the way Sweden has decided to conduct our part of the Eurocensus means many new methodological challenges. In this work it makes sense to not only build for the Census-task itself but also plan for and look for methodological solutions to future statistics. In times when National Statistical Organisations face budget cuts this is even more important.

## References

Lanzieri, G. and Schulte Nordholt, E. (2009). Confidentiality Issues in the EU Population and Housing Censuses of 2011. Paper presented at a joint UNECE/Eurostat work session on statistical data confidentiality, Bilbao, Spain.

Wallgren, A. and Wallgren, B. (2007). Register-based Statistics - Administrative Data for Statistical Purposes. New York: John Wiley.

Zhang, L.C. (2009). A triple-goal imputation method for statistical registers. Paper presented at the UNECE Work Session on Statistical Data Editing, Neuchatel, Switzerland.

Zhang, L.C. (in press). A unit-error theory for register-based household statistics. To appear in Journal of Official Statistics.

# Making better use of administrative data: how far can New Zealand go? 

Christine Bycroft ${ }^{1}$


#### Abstract

The traditional model of census and household surveys supported by administrative data is coming under increasing funding pressure. Statistics New Zealand is developing a long-term view of the overall design or architecture of data sources as a key step towards securing ongoing sustainable funding for social and population statistics. A major theme of the proposed architecture is making more use of data we already have, especially using administrative data to enhance or supplement the information from census and social surveys. Census is the major budget item. While there is no Population Register in NZ, we are challenged as to why we cannot use the several administrative registers or lists that do exist (e g from tax, health and education systems) in place of census. The paper uses the tax-based Linked Employer-Employee Data as an example of the potential uses of administrative data for social and population statistics in the New Zealand situation.


Key Words: administrative data; census; Linked-Employer-Employee Data;

## 1. Introduction

Official social statistics in New Zealand are based on a model of census and household sample surveys supported by administrative data. However this traditional model is coming under increasing funding pressure. Statistics New Zealand is developing a long-term view of the overall design or architecture of data sources as a key step towards securing ongoing and sustainable funding for social and population statistics. A major theme of the proposed architecture is making more use of data we already have, especially through making better use of administrative data.

Administrative data can be statistically useful in itself, especially if source agencies recognise its strategic importance for improving their own services. But it is through combining census, survey and administrative data in various ways that we gain most from administrative data. Use of statistical models that combine aggregate data from various sources and record linkage techniques that combine data at a micro level can create a more effective system than is possible from using single data sources in isolation.

Using administrative and survey data in combination is hardly a new idea. Since 1851, when the first New Zealand Census of Population and Dwellings was held, population statistics have been based on a combination of the census and administrative data from birth and death registers and records of external migration. Census provides the detailed basis of the population and dwelling counts every five years. These census counts are updated during the inter-censal period using the administrative sources to measure population change.

It is the census however, that is the major budget item. The key determinant of social statistics architecture is the ability to find an administrative replacement for the detailed population and dwelling information currently derived from a full enumeration census. Some countries have moved to a register-based system (Walgren and Walgren, 2007) where an administrative Population Register linked to an Address Register serves as a central reference frame for the population in scope. Statistics New Zealand business statistics already have an administrative-derived Business Register, and are increasingly making effective use of administrative data to reduce survey costs and to increase research potential. While there is no Population Register in NZ, (nor an official address register) we are

[^57]being challenged as to why we cannot use the several administrative registers or lists that do exist (e g from tax, health, electoral and education systems) in place of the census.

This paper will give an overview of the interactions between census, surveys and administrative data in New Zealand at present and look at opportunities for further use of administrative data. The paper concludes with a discussion of the challenges we face in New Zealand in addressing the question of whether a paradigm shift to a register-based system is desirable or feasible.

## 2. The role of administrative data now

The census, sample surveys and administrative data make up the three primary data sources for official social and population statistics. The main strengths and limitations of each type of source data are well known. The Census of Population and Dwellings sits at the heart of social and population statistics. The census provides the basis for population estimates and a range of socio-economic information for the whole population, for local communities and for small sub-groups. However, the census is periodic (every five years in New Zealand) and the length and complexity of questions are limited. Sample surveys play a key role in producing statistics more frequently than is possible with a census (e g Household Labour Force Survey), or can explore subjects in far more depth (e g Household Expenditure Survey, General Social Survey). The small size of sample surveys limits their ability to produce detailed breakdowns for sub-groups and small areas.

A wealth of administrative data associated with government activities exists which could potentially be used for statistical purposes. Some official social statistics sourced directly from administrative data are currently produced by Statistics NZ (e g Vital Statistics, Police and Justice, job creation and destruction). Other agencies produce statistics from their own administrative sources, (e g Health and Education sectors). Government agencies are improving their own information systems, and some have introduced central indexes (National Health Index, National Student Index) or have combined internal datasets for better analysis (Benefit Dynamics Database).

Administrative data incurs no additional respondent burden or direct collection costs and does not carry the uncertainty associated with sample error, but there is no free lunch. Administrative data collected by government reflect the interaction of government with its citizens, and are specific to the population of interest and activities of each government sector. The main issues encountered with using administrative data are due to this specificity of the administrative purpose. Often there is a lack of reference to the statistically defined resident population, reporting units may be defined as events rather than statistical units, the available variables are limited in scope and definitions may differ from statistical concepts.

So while administrative data can be very useful when the focus is on providing information directly related to the administrative purpose, producing statistics with sufficient context and comparability to the statistically defined resident population can be difficult. Put another way, administrative data is good for understanding government inputs and outputs, but is weak in understanding outcomes for people.

The solutions lie in using these individual data sources together. There is already a complex inter-dependency between all three primary data sources. To make "better use" of administrative data, we will focus here on the opportunities provided by good use of statistical methodologies. Existing examples of the use of explicit statistical models include the calibration of sample survey estimates to population totals, and use of a cohort component model of population change for producing inter-censal population estimates. Several record linkage projects have proven to be of high value, for example linkage of census to death records, and of census to the Cancer Register. One survey has linked survey responses to health records with the consent of respondents. Record linkage proposals have been assessed on a case by case basis, balancing the benefits against privacy concerns.

We will use administrative data derived from the taxation system to illustrate the opportunities for extending our use of record linkage and modelling techniques. We want to look at ways of making better use of the taxation-sourced data to enhance our statistical system, and to identify the real barriers which will limit how far we can go.

## 1. Opportunities using the Linked Employer-Employee Data

### 1.1 Linked Employer-Employee Data (LEED)

The Linked Employer-Employee Database ${ }^{2}$ (LEED) is created by linking a longitudinal employer series from Statistics New Zealand's Business Frame to a longitudinal series of Employer Monthly Schedule payroll data from Inland Revenue (New Zealand's tax collection agency). The smallest reporting unit is a unique employer-employee combination (or job), placing LEED in a unique position at the interface between person-level and business statistics. Its primary purpose is to provide data on the dynamics of the labour market. LEED data covers all taxable income (except for investment income) and so includes beneficiaries and the self-employed as well as wage and salary earners. The original base tax data has been extended to include person-level links to benefit data and to tertiary student data. Government services (transfer payments, education) can now be related to outcomes (employment, income), and this can be studied at a longitudinal micro level.

### 1.2 LEED supporting sample surveys

Because LEED includes the employed and identifies that receiving unemployment benefit, there is a strong relationship with the Household Labour Force Survey (HLFS). It seems unlikely that LEED could replace the sample survey because of scope, definitional and timing issues. LEED does not include all of those "not in the labour force", tax residency is not the same as the statistical usual residence definition and ILO unemployment definitions are different from the receipt of unemployment benefit. Lastly, collection through the administration system is slower than direct surveying, an important aspect because timeliness is critical for the leading economic indicators on labour force status.

Despite these differences, we should be able to exploit the LEED data to support and enhance the labour force survey. One advantage LEED has is the availability of sub-national geographic information. Use of small area estimation is an obvious route. The small area model makes best use of the survey information and the geographic patterns present in related auxiliary data to produce survey estimates at sub-national levels. Very promising results for sub-national unemployment estimates have been achieved using a fairly standard EBLUP area-level model and cross-sectional HLFS data. Future work will investigate models that make use of the longitudinal nature of the HLFS panels, and will consider whether joint employment and unemployment distributions at sub-national levels can be produced.

At a micro-level, the relationship between the survey labour force status and LEED employment and benefit receipt data is unknown. By linking the HLFS to LEED at unit record level, we will improve our understanding of both these measures of labour market activity, and improve survey quality assurance.

LEED also includes income from most income sources, and does not suffer from respondent recall errors or misunderstanding. Again, there is an opportunity for comparison and validation of income survey responses if we link at unit record level. Going further, data substitution, where linked administrative data replaces survey questions, has the potential to reduce respondent burden and improve data quality. The feasibility of replacing some income survey questions by linking the survey response to LEED data is being investigated. Generally the income variables match conceptually, although LEED does not include investment income or non-taxable income such as the government accommodation supplement.

A feasibility study to test probabilistic linking between HLFS and LEED (without using the unique tax number) is underway. Technical linkage issues and the potential to improve HLFS statistics will be investigated. While there are no privacy concerns at this early feasibility stage, in the long-term we are unsure about the best ways to address the privacy issues associated with linking survey responses to administrative data. We should ask respondents for their consent, but what is the best approach? Should we ask the respondent for their tax number? Is there a risk to response

[^58]rates? How should we manage historical linkages? Providers are also concerned about linkage of their data with other sources, and we must manage provider concerns over security, access and confidentiality.

### 1.3 Sample surveys supporting LEED

LEED is a longitudinal dataset, following individuals over time via the unique (anonymised) IRD number identifier. The key limiting factors for LEED are the lack of socio-demographic variables available on the tax data (only age and sex) and the uncertain relationship of LEED to the usual resident population. Both can be addressed by linking LEED to a sample survey. The HLFS, for example, includes ethnicity, qualifications, occupation, hours worked and family and household structures, and is a nationally representative sample. The HLFS-LEED linkage will broaden the opportunities for labour market analysis and productivity measures, and improve understanding of the consistency between survey measures and administrative reporting.

### 1.4 A systematic approach to record linkage based on LEED

The linkages outlined above fit within the existing Statistics NZ Data Integration Policy and continue the approach of assessing new data linkages on a case by case basis. There are advantages in a more systematic approach to data linkage. Van Tuinen (2009) recommends that statistical agencies collect as much microdata as possible, make it standardised and harmonised so that it can be linked at unit record level. The linked data provides the flexibility to respond rapidly to new information needs and is particularly advantageous for longitudinal information.

Extending the existing LEED linkages across multiple data sources (to include for example, health, justice, births, deaths and migration) is likely to be technically feasible in New Zealand if the aim is to produce statistical outputs and to support research. The issue is whether systematic linkage of personal data fits within the public's tolerance over privacy. There is a choice of models for managing privacy, and the challenge is to develop appropriate privacyfriendly protocols for holding linked data. Providing access to researchers is also a major barrier. At present, access to the tax-based microdata is permitted only for Statistics NZ employees, and secondment arrangements are made to accommodate government researchers.

## 2. LEED and population statistics

Social statistics are statistics about people. The story of social statistics starts with population statistics - how many people are there, where do they live, what is their age, sex and ethnic group, and how are these distributions changing? The way in which population statistics are produced is the key to understanding how far we can take the use of administrative data in providing statistical information about people.

### 2.1 LEED supporting inter-censal population statistics

Most adults resident in New Zealand earn taxable income and so one might expect a high degree of overlap between the people in LEED and those in the usual resident population. Ryan (2009) however shows significant differences between the LEED population and the official Estimated Resident Population (ERP). In contrast to the point-in-time ERP definition, LEED records events over a period of time, and the numbers counted by LEED change dramatically depending on the criteria set and whether one considers those paying tax in a given month, over a year, or all those registered for tax. The relationship of LEED to ERP also varies by age group, by sex and by region. On the positive side, these relationships do show strong consistency over time. Also LEED is a national system and at least some changes of address are captured. This opens up the possibility of using LEED to assist in estimating internal migration.

Inter-censal population estimates use births, deaths and external migration data as high quality direct measures of population change. There are no direct measures of internal migration. The current process for sub-national population estimates relies on subjective judgment. LEED and some other administrative datasets (Electoral Roll, school enrolments) do capture some movement between areas. A modelling approach would combine the partial information from different sources to produce sub-national population estimates.

Work is underway on a Bayesian hierarchical model for gross migration rates. The model will combine historical migration patterns from censuses with the information from LEED and other sources that partially reflect internal migration in real time. The aim is to use the administrative data in sound quantitative models to improve the quality and efficiency of sub-national population estimates between censuses, as well as providing measures of uncertainty.

### 2.2 LEED and counting the population

The next question is whether LEED could produce population statistics without a full enumeration census. For LEED on its own, the answer is clearly "no". The nature of the tax system means that LEED has the characteristics of an excellent register, for example a well-controlled unique identifier, and complete coverage of the target population of tax-payers. However, its purpose is quite different from a Population Register, and there are major issues of undercoverage and over-coverage in relation to the usual resident population.

Could we produce good population estimates by linking LEED to other administrative sources? Could we fill the gaps and identify the deaths and out-migration? If external migration data were linked, those no longer resident in New Zealand could be flagged. Linking to health data, the electoral roll or education data would include children and others without taxable income. Perhaps commercial data sources could help? This might sound plausible until one considers the methodological issues.

First, there is no unique identifier available across administrative data sources. Linkages would be through probabilistic matching and would incur associated linkage errors. Probabilistic matching has been applied successfully when the aim is to investigate relationships. Linkage errors are far more problematic when the aim is to estimate the joint population. For example, when a match that should be made between records for the same person is missed, perhaps because the name is recorded differently, then that person will be counted twice in the union of the two data sources. Linking more datasets to LEED in the hope of filling gaps might grossly inflate the population. Accurate address information required for sub-national estimates will be another problem.

Even if another good data source could be accurately linked to LEED, there are still likely to be people missing. Capture-recapture models estimate those missing, but require an independence assumption. This is clearly not the case when linking administrative data (Asher and Fienberg, 2001). Could some equivalent to a census coverage survey be used to estimate undercount and overcount? Aside from the methodology questions, again there would be privacy concerns to manage, as well as consideration of the need for population estimates to be independent of the electoral system and free of influence by agencies reliant on a population-based funding allocation.

## 3. Discussion

All primary data sources have their limitations. The census reaches the whole population but is infrequent and must be simple enough for self-completion. Sample surveys can provide depth, but not detailed breakdowns for subgroups. Asking respondents directly produces data about complex concepts and high quality information, but imposes costs on both the survey agency and the respondent's time. Administrative data incurs no new respondent burden, but is limited to information about the sub-populations and variables related to the specific administrative purpose.

The solutions lie in combining administrative and survey sources using statistical techniques that make up for the weaknesses of one source through the strengths of the other. Most of the above uses of LEED will add value to our system of social and population statistics at relatively low cost, but do not replace the census or sample surveys. Statistical models have an advantage because they do not require record linkage, and are able to make use of data which does not fit neatly into statistical concepts and definitions. The main limitations are likely to be in maintaining a sufficient level of expertise, and also perhaps in convincing decision-makers and data users of the validity of modelling approaches that they do not fully understand.

Privacy concerns may impose limits on the extent of unit record linkage, but we are unsure where the boundaries lie. We need to better understand public attitudes to data sharing for statistical purposes, and to ensure that privacy-
friendly protocols are in place if we shift to a more systematic approach to linking administrative data sources. Providing suitable access for researchers is another challenge.

The major barriers arise when we look for replacements to the census. There is no Population Register in New Zealand. Producing population estimates from existing administrative data poses a formidable challenge, and seems unlikely to be feasible with our present understanding. Perhaps solutions will be found in a combination of new linkage models for population size estimation (Manrique-Vallier, 2008) or aggregate level models of the kind outlined above and some kind of reduced census or large coverage survey.

Conceptually, "counting the population" may sound simple, but this impression is deceptive. Total population is relatively easy to measure in New Zealand since we have good administrative systems for births, deaths and external migration, but in practice, it is the detailed distributions at sub-national geographies and by age, sex and ethnicity which are both critically important and difficult to measure accurately. Purpose-built systems are required. It seems that for some time at least, a periodic full enumeration census will form the basis for population statistics in New Zealand, and will be the main information source for small areas and sub-populations.

Better use of administrative data should enhance our survey outputs and inter-censal population statistics and help meet expectations for more detailed and more flexible information through re-use of existing sources. It may help manage escalating costs especially if we can reduce the need for longitudinal surveys. The administrative data plays an important, but still supporting role - there is no fundamental change to our survey-based system.

## References

Asher, J. and Fienberg, S. (2001). Statistical variations on an administrative records census Proceedings of the annual meeting of the American Statistical Association, August 5-9, 2001

Manrique-Vallier, D. and Fienberg, S. (2008). Population size estimation using individual level mixture models Biometrical Journal, 50, 6, 1051-1063

Ryan, M. (2009). Comparing LEED to the 30 June 2001 and 30 June 2006 ERP bases Unpublished Statistics New Zealand report

Van Tuinen, H.K. (2009). Innovative statistics to improve our notion of reality. Journal of Official Statistics Vol 25 No 42009

Walgren and Walgren (2007). Register-based Statistics: administrative data for statistical purposes Wiley, Chichester, UK

# First Register-based census in Slovenia: How to convert administrative sources into statistics 

Danilo Dolenc ${ }^{1}$


#### Abstract

The first register-based census in Slovenia (reference date 1 January 2011) depends on direct linking of data from about 30 administrative and statistical sources. Three administrative registers form the framework of the register-based system (Central Population Register, Household Register and Real Estate Register). The last two are used in the statistical census process for the first time. The statistical definitions differ significantly from the administrative ones in all framework registers. Several methodological solutions are being prepared to overcome the administrative concepts. Besides that, improving the quality of the administrative sources is one of the priorities in the preparatory phase. The new approach to census-taking also implies the intensified use of statistical methods in the process and overall quality assurance.


Key words: Register-based Census; Administrative sources; Statistical concept; Data quality.

## 1. Introduction

Although a part of the former Yugoslav statistical system, the Slovenian Statistical Office already in the 1970s started to follow the experience of the Nordic countries about using administrative data for statistical purposes. As there were no similar initiatives in the other governmental bodies, the Statistical Office itself as a producer established four basic registers in close cooperation with corresponding authorities:

1. The first attempt to establish the Central Population Register (hereinafter CPR) was prior to the 1971 Population Census. More successful was the second attempt when new PIN numbers where delivered prior to the 1981 Population Census. The 1981 Census data were also besides the local registers of permanent population the second source of data for the establishment of the CPR. The statistical data on population were for the first time published from the CPR in 1986. The CPR has been kept by the Ministry of the Interior since 1998.
2. The Register of Areas of Territorial Units and the Record of House Numbers (predecessors of the present Register of Spatial Units) were set up in the 1980s together with the Surveying and Mapping Authority of Slovenia, which has also been the only keeper since 1995.
3. The Statistical Register on Employment (the only one still kept by the Statistical Office) was set up on the basis of the Census of Employees in 1986 and it is updated by the registration data for pension, disability and health insurance, parental protection insurance, unemployment insurance and records on employment provided by the Health Insurance Institute of Slovenia.
4. The Business Register has been available since 1976 and has been kept by the Agency for Public Legal Records and Related Services since 2002.

The data from above mentioned registers were used already at the 1991 Census and to a greater extent at the 2002 Census. The main features where (Dolenc, 2003):

- Pre-print of data (name, surname, address, date of birth) on questionnaire in 1991 and 2002;
- Pre-print of identifications on the basis of the CPR data intended for electronic data capture in 2002;
- Marking which contents on the questionnaires for person have already been obtained from the sources and exist in the pre-census database in 2002 (no field collecting);
- Use of the Statistical Register on Employment data for persons in paid employment in case of existing link (PIN) (1991 Census);

[^59]- Some contents were entirely taken over from the registers at the 2002 Census and therefore for the first time in censuses in Slovenia not included in the questionnaire and not collected in the field (e.g. place of birth, last migration, citizenship, marital status, activity status, occupation, industry, place of work). The imputation was the methodological solution for entries not existing in databases but found in the field enumeration.

The use of data from different administrative and statistical sources enabled more efficient and simple data collection in the field, reduced response burden and gave the opportunity for shorter, better and more cost effective data processing (Dolenc, 2009).

## 2. New Developments since 2002

General prerequisites for undertaking a register-based census:

- Legislation which enables the linkage of data from different sources (Articles 32 and 33 of the National Statistics Act give the right to the Statistical Office to collect, use and link data from different administrative sources and other collections for statistical purposes only);
- Establishment of appropriate administrative and/or statistical sources with unique identifier (e.g. PIN, address) and corresponding agreements with the data keepers for the regular transmission of data to the Statistical Office;
- Appropriate topics in the sources which cover demands of users and obligatory EU legislation topics and are also comparable (on a large scale) with the previous censuses.

Some 30 different databases from 10 data providers (including the Statistical Office) will be linked together. The most important data on population, households and dwellings will be extracted from single sources which were developed or completed in the meantime (Dolenc, 2010):

- Establishment of the Real Estate Register in 2007 on the basis of the field census of buildings and dwellings (Real Estate Census), which took place in 2006 (partly 2007, too). From the register-based census point of view one of the most important elements was the introduction of a unique system of labeling dwelling numbers in multi-dwelling buildings;
- Supplementing of addresses in the CPR with dwelling numbers in multi-dwellings buildings as a previous missing link;
- Informatization of paper household files (which have in fact existed for several years but have not been used in statistics) and establishment of the Household Register kept by the Ministry of the Interior. The most important advantage of the Household Register compared to registers of other countries is the availability of data on relation to the reference person of the household. The same classification was used at our previous 2002 Census, so it is in fact a statistical classification taking into account the most common relations in households which enable also to derive data on families in the household.


## 3. Administrative versus Statistical

The administrative and the statistical concepts are not always the same, so the main methodological challenge in the register-based census is to produce output statistical information from input administrative data. Despite the harmonization of methodology presented for example in the form of international advisory Recommendations for the 2010 Censuses of Population and Housing (Recommendations, 2006) or obligatory regulations for the EU Member States (the Regulation on Population and Housing Censuses from $2008^{2}$ or the Regulation on Community Statistics on Migration and International Protection from $2007^{2}$ ), the adoption of detailed census methodology is anyway the domain of every country itself depending on its national needs and in case of register-based approach also on the availability of data. Besides the unavailability of data, dependence on existing administrative sources, their content, management, methodology and quality of the data are the main restrictions.

[^60]
### 3.1 Schematic Review of the Basic 2011 Register-based Census Concepts

One of the most important features of the 2011 Register-based Census in Slovenia is the fact that only one administrative source is used in whole for compiling data on the essential topics of every modern census operation. Besides that, the Slovenian particularity is also the method of using administrative data. There are no so called statistical registers as an intermediate step in the statistical process, but statistical data are derived directly from the original administrative output based on independent statistical methodology. However, every source mentioned below has its own characteristics which had to be taken into consideration when the statistical concepts are applied. As in the classical field enumeration, the main and in our experience the most essential characteristics of the registerbased approach is linking together the persons with their households and connecting household(s) to the appropriate dwelling.

Table 3.1-1
Schematic Review of the Population, Household and Dwellings Concepts in the Administrative Sources and Applicability of the Administrative Concepts in the 2011 Register-based Census in Slovenia

| Topic | Administrative concepts | Statistical concepts |
| :--- | :--- | :--- |
| Population | More than one registered residence at the <br> same time possible: <br> one permanent and/or one or more <br> temporary residence | Only one (usual) residence at the same <br> time: <br> Temporarily absence abroad (but still with <br> length and intention of stay (one- <br> year criterion) <br> priority rules in the case of two <br> administrative registered residence |
| Households | Household data available only for <br> permanent residence <br> Private households only <br> Self-declared reference person <br> relations in household refer to the <br> reference person only | No data on families available |
| Generating families: <br> directly from the household <br> relations <br> indirectly from the relations <br> between biological parents and <br> children and between spouses |  |  |
| Dwellings | Institutional households |  |
| Type of building <br> Construction type | Actual use of building / part of building | Conventional dwellings <br> Other housing units |

### 3.2 Population

According to the new definition, the population of Slovenia consists of persons (regardless of citizenship) with registered residence in Slovenia who live or intend to live in Slovenia for one year or more and are not temporarily absent from Slovenia for a year or more. The new statistical definition of population is harmonized with the definition of population and migrants in the Regulation on Community Statistics on Migration and was introduced in 2009 (and is valid from 2008 data onwards). The population definition at the 2011 Register-based Census will be for the first time in censuses the same as the definition used in our regular quarterly population stock statistics. At the 2002 Census the difference between census and regular statistics data was 30,000 persons $(1.6 \%)$.

The direct consequences of the new population definition are lower number of population itself (mostly because of excluding short-term immigrants from the stock) and the territorial redistribution of the population (statistical migration in the case of permanent and temporary residence). The last one has also an important impact in the
household generation and census statistical process itself as there is no administrative data on belonging to the households in case of temporary residence considered as usual residence.

### 3.2.1 Quality Obstacles in the Administrative Source (Central Population Register)

In general, the CPR is a high-quality source of data in terms of coherence of data as it is in fact a technical integration of several administrative registers regarding the Internal Administrative Affairs (e.g. Civil Status Register, Register of Permanent Population, Register of Foreigners, Citizenship Register). It contains well documented and verified data on demographic and geographic characteristics of persons living (lived) in Slovenia and of persons without residence in Slovenia but with any kind of ties to Slovenia (e.g. citizenship, property).

Formally correct data in the CPR do not necessarily correspond to the real field situation of population and their households. The quality from this point of view (relevance) depends also on how strictly people respect the legislation on registration of residence. According to the law, the registration must be done in eight days after settlement. But we well know from the previous census and from other statistical household surveys that mostly in larger towns around $10 \%$ of people live somewhere else and not at the registered address. There are four main reasons:

- Some personal benefit (mostly of financial nature);
- Owners of private dwellings do not register their tenants or subtenants in order to avoid paying taxes;
- Many rights (in administrative sense) are connected to the permanent residence;
- Administrative burden (time-consuming) and costs (change of personal documents).

From the statistical processing point of view the main problem in the CPR was (is) missing dwellings number as an obligatory supplement to the address for persons living in multi-dwellings buildings. Dwelling numbers should be collected in the Real Estate Census and transmitted to the administrative keeper. At the beginning of 2010 only 20\% of respective records in the CPR contained also the dwelling number but a few weeks before the census reference day (1 January 2011) less than $10 \%$ of required dwelling numbers are still missing. There were two successful actions undertaken in cooperation between the Ministry of the Interior and the Statistical Office:

- Linkage of data on ownership of dwellings, registered residence of owners and their households. Dwelling numbers were automatically allocated to the CPR records in the case of complete matching (the presumption was that most of the owners live in their properties);
- Mailing out of almost 50,000 letters to the reference persons of the households after finishing the first action with request to return (free of charge via the enclosed envelope) the information on their dwelling number. The response rate was almost $75 \%$.


### 3.3 Households

The concept of the household in the register-based approach results from the changed statistical definition of population and is closely connected to the availability of relevant data in basic administrative sources. The former concept of sharing income for covering the basic costs of living is not suitable anymore. The dimensions of the new household definition are: usual residence derived from the CPR, living in the same dwelling, declared household in the Household Register (persons with the same household number at the address); relation to the reference person of the household, which shall be stated for every person. The new definition of the household in the Register-based Census 2011 is therefore a group of persons (or a person living alone) living in the same dwelling with the same household number.

Recommendations (2006) distinguish two concepts of private households (paragraphs 479-481): the householddwelling concept (exactly one household per occupied housing unit) and the housekeeping concept (two or more households can share the same dwelling). In most register-based countries the first concept is applied but in Slovenia the existing Household Register enables us to implement the housekeeping concept, which is in fact more appropriate as cohabitation of more households in the same dwelling (mostly in detached houses) is very common in Slovenia.

### 3.3.1 Household Register

The Household Register is a part of the Register of Permanent Population. Data on household members are collected at the time of registration or de-registration of permanent residence by the statement of being a household member. Relevant changes of household members which are not the result of a change of address (for example death, divorce, formation of a new household at the same address) are updated by authorized person by official duty. The household identifier in the Household Register is the serial number of the household running from 1 to NNN at the same address. An important advantage of the Household Register is also the availability of data on relation to the reference person of the household. The same classification of relations was used at our previous 2002 Census, so it is in fact a statistical classification taking into account the most common relations in households. By using data on relations to the reference person, for most of the population generation of households and families could be done automatically without using complicated statistical methods. For this purpose for every single combination of relations in the household a unique matrix is being prepared which enables us to derive all obligatory household and family topics from EU Regulation (size of household / family, household / family status, type of household / family) directly from the Household Register.

### 3.3.2 Quality Obstacles in the Administrative Source (Household Register)

As the new source has not been used before for any statistical purposes, we found many obvious errors which could be very easy to solve with more exact guidelines and simple rules when entering data into electronic form at the administrative body. It is very important that errors of this kind are corrected already in the original administrative files and not later in the statistical process. The permanent and long-continued use of the administrative data in the statistical process is one of the most important prerequisites for better quality.

The Household Register is in fact not a classical register because there are no transactions on entries or exits but only momentary and non-repeated cross-section of data for specific point in time is available. Besides that, it is not completely simultaneous with the CPR data (although it should be) but in some cases there are delays in updating. The main reason is that by now the Household Register has been considered as non-priority in the administrative procedure. The data provider (Ministry of the Interior) is now aware of its responsibility for the success of the register-based census and has already significantly improved the quality of the data. It has been proved once again that close cooperation between data providers and data users is of great importance and the feedback (even in aggregate form) can improve the quality with a little effort by both sides.

Anyway, from the statistical point of view the share of the inconsistencies in the first release of household data to the Statistical Office was negligible (below 1\%). Most of the errors were found in the combination of variables age and relation to the reference person (most common inverted relation - e.g. mother younger than child). Another aspect of the quality of the household data is linked to the possibility of generating families directly from the relation to the reference person data. Around $3 \%$ of records are not useful as relations to the reference person are not sufficiently identifying.

### 3.3.3 Some Methodological Solutions for Generating Statistical Households

In most cases (for some $90 \%$ of the population) the administrative and statistical households are the same. For the rest of the population methodological solutions and the use of statistical methods (according to the Nordic countries experience) are foreseen to overcome administrative concepts:

- For foreigners with temporary residence only and for citizens with usual residence at the temporary address the households have to be generated in the statistical process on the basis of living in the same dwelling, including biological and marital relations (parents-children, spouses);
- The same method is also used for the family generation in case of non-identifying household relations as we have the possibility to link household and personal data for almost all citizens. In recent years the administrative bodies have put great effort into supplementing the CPR with data on biological parents (predominantly for generations with still living parents);
- In case of the statistical relocation of the reference person from the household to the temporary residence, matrixes have been prepared to transform old relations to the new ones. The basic rule is the same hierarchical level of the new reference person (for example a spouse becomes the reference person);
- Institutional households were identified in advance on the basis of the number of persons without household number living at the address. More than 900 addresses with 60,000 persons have been determined and precisely checked in different public databases. An inventory of types of living quarters has been elaborated and classified into 7 basic categories (secondary and tertiary student homes, old people's homes, welfare institutions, correctional and penal institutions, religious institutions).


### 3.4 Dwellings

Dwelling definition in the Register-based Census 2011 is the same as it was in the 2002 Census, but it is derived from several variables of the Real Estate Register. The statistical procedure for deriving housing census units is in fact rather complex as the content of the Real Estate Register is much broader than it is used in the census (including also data on parcel of land, non-residential buildings, infrastructure and other construction objects).

### 3.4.1 Quality Obstacles in the Administrative Source (Real Estate Register)

When we are talking in terms of the quality of the input administrative data, the Real Estate Register is the weakest administrative source. The main problems we are facing are: distinguishing multi-dwelling buildings and defining the conventional dwellings in the building; unreliable data on some housing topics; missing units; occupied buildings without address; missing otherwise obligatory numbering of dwellings in multi-dwelling buildings, duplication of dwellings (same dwelling number appears more than once at the address).

Fortunately, the data from the Real Estate Register will be used for the taxation and a special Mass Real Estate Valuation Act has been adopted which obliges the owners of the real estates to check the data and provide correct data to the administrative body (Surveying and Mapping Authority of the Republic of Slovenia). The process of the evaluation and correction of the data is in progress at the time being. A special web application is also available for direct correction of data. By the last review of the time schedule the action will be finished at the end of January just the right time for the census purposes as the transfer of data from the administrative source to the Statistical Office is foreseen for March 2011.

### 3.5 Conclusion

The reference date (January 1st) is in fact in the register-based census approach 3 months before the real start of the statistical process with the integration of data from administrative and statistical sources. The framework register for the whole statistical operation is the CPR database. The statistical population is derived from the CPR database according to the new statistical definition. Data from all other sources will be gradually (according to the availability of data) linked to the census population. The new approach to census-taking implies the intensified use of statistical methods in the process and overall quality assurance. The statisticians have a privilege to modify administrative data according to the statistical concepts. The register-based census is not a simple push on the button but a very complex and demanding process with several advantages in comparison to the previous field enumeration.

## References

Dolenc, D. (2003). Register based 2002 Census of Population, Households and Housing in Slovenia and New Solutions in Data Processing. Paper presented at the Joint ECE-EUROSTAT Work Session on Population and Housing Censuses, Ohrid, The Former Yugoslav Republic of Macedonia.

Dolenc, D. (2009). Register based Census 2011 - a New Challenge for the Slovenian National Statistics. Paper presented at the $19^{\text {th }}$ Statistical days 2009, Radenci, Slovenia.

Dolenc, D. (2010). Quality Assessment in Register-based Census - Administrative versus Statistical Concepts in the Case of Households. Paper presented at the European Conference on Quality in Official Statistics, Helsinki, Finland.

Oblak Flander, A. and Seljak, R. (2008). Quality Assessment of the Register-based Slovenian Census 2011. Paper presented at the Joint UNECE/Eurostat Meeting on Population and Housing Censuses, Geneva, Switzerland. Recommendations for the 2010 Censuses of Population and Housing (2006), New York and Geneva: United Nations.

## SESSION 7C

DATABASES II

# Historical data linkage of tax records on labour and income: The case of the Living in Canada Survey pilot 

Andrew Heisz, Manon Langevin and Jeffrey Randle ${ }^{1}$


#### Abstract

Data matching is a common practice used to reduce the response burden of respondents and to improve the quality of the information collected from respondents when the linkage method does not introduce bias. However, historical linkage, which consists in linking external records from previous years to the year of the initial wave of a survey, is relatively rare and, until now, had not been used at Statistics Canada. The present paper describes the method used to link the records from the Living in Canada Survey pilot to historical tax data on income and labour (T1 and T4 files). It presents the evolution of the linkage rate going back over time and compares earnings data collected from personal income tax returns with those collected from employer files. To illustrate the new possibilities of analysis offered by this type of linkage, the study concludes with an earnings profile by age and sex for different cohorts based on year of birth.


Key words: Retrospective linkage; tax data; matching; linkage rate; historical linkage; administrative data.

## 1. Introduction

Data linkage is a common practice at Statistics Canada and in several other statistical agencies around the world. It is a good way to reduce the costs associated with survey activities and to enhance the analytical strength from existing data sources. Certain types of information will always be hard to collect by survey, either because it requires significant recall by respondents (e.g. monthly calendar of employment status) or because the nature of the subject may be embarrassing to discuss with a stranger (e.g. victim of sexual assault). For this reason, analysts in several fields are regularly confronted with insufficient data, which limits the type of analysis that can be conducted.

Data linkage can also be used, in the context of social surveys, to replace questions related to income, significantly reducing the response burden of respondents who consent to this type of procedure and increasing the accuracy of the data collected. The Survey of Labour and Income Dynamics is a good example of this. Generally, $80 \%$ of respondents of this survey consent each year to data matching. The consent is valid for the duration of the respondents' participation in the survey and the linkage of records is carried out solely between data files belonging to the same collection year. However, the linkage carried out with the pilot of the Living in Canada Survey (LCS) is somewhat different because it involves historical linkage. More specifically, the LCS records were linked to tax files from years prior to the LCS collection year; this type of linkage had never been done before with a Statistics Canada social survey.

The purpose of this study is to explain how the historical linkage was done and to present the benefits in terms of analytical potential. In particular, the study examines (1) the degree to which the linkage rate diminishes going back in time, (2) the accuracy of the information contained in the different linked tax files, and (3) the potential of the retrospectively linked information in analyzing phenomena that require a long data series.

[^61]
## 2. Linkage

The Living in Canada Survey pilot data were linked with different tax files of individuals and businesses: (i) the personal income tax form (T1 file), (ii) the statement and summary of compensation paid by employers (T4 file), and (iii) the Pension Plans in Canada Survey file. ${ }^{2}$ For these files, two different types of linkage were carried out: (i) yearly linkage (renewable for each new wave of the survey) and (ii) historical linkage of tax data for the years going back to 1990.

Despite the fact that this second type of matching had never been done before at Statistics Canada, there is some emerging literature on the topic (Reimer and Künster, 2004; Roemer, M., 2002; Sears and Rupp, 2003). A team of German researchers had carried out an exercise similar to ours by comparing administrative data on employment status to employment data from a survey (Huber and Schmucker, 2009). One of their main findings was that the probability of the two files corresponding was negatively correlated to the number of events. In other words, they found that respondents who experienced several changes in employment status during the period analyzed were significantly more likely to forget events and to produce an incorrect statement (memory bias). This appears to indicate that the historical administrative data improved the quality of the information for this type of individual in particular.

The consent of LCS respondents to yearly and historical linkage was requested completely separately. The first question allowed respondents to give permission to access their tax file for the pilot's reference year and for the duration of their participation in the survey. Permission to access the same files, but for all calendar years going back to 1990 , was asked in a second question and solely of those respondents who had previously consented to linking their tax files for 2007 (Table 2.0-1).

Table 2.0-1
Permission for yearly and historical linkage from the LCS pilot questionnaire
In order to increase the accuracy of the data and reduce the time of the interview, you can give permission for Statistics Canada to access your income tax records.

1. Statistics Canada would obtain information from: the T1 Income Tax and Benefit Return, the T4 file of records from employers, and the employer pension plan file. The information we obtain would be used for statistical purposes only, and would be kept confidential. Do we have this permission?
2. Statistics Canada would also like your permission to access information contained in your past tax records back to 1990. This will enhance the quality of the results from the survey. The information we obtain would be used for statistical purposes only, and would be kept confidential. Do we have this permission?

For confidentiality reasons, respondents could not be asked directly for their social insurance number (SIN) in the LCS interview. Consequently, respondents' SIN was identified by probabilistic record linkage and by using auxiliary variables such as last name, first name, sex, date of birth, marital status, address and postal code. It should be noted that for technical reasons, data from T4 files were not available for years prior to 2001 at the time of this analysis.

## 3. Data Source

The main purpose of the LCS was to provide Canada with a longitudinal survey with panels of indefinite length and content covering several topics of interest for the development of public policy. More specifically, the survey's content should enable researchers to identify the links between the four main areas of an individual's life, namely, work, education, family life and health status. Similar surveys are available in several countries and have proven useful in developing better policies and programs by all levels of government.

[^62]The pilot was a means to test the survey questionnaire. Sample coverage was limited to four of Canada's ten provinces: New Brunswick, Quebec, Ontario and Saskatchewan. The pilot covered the population of these four provinces, excluding regular members of the Canadian Forces, individuals living in institutions and on a native reserve. In total, the file contains data from 2,881 respondents, aged 15 years and older, $79 \%$ of whom consented to linking their data to the reference year. Of that number, $94 \%$ also consented to the historical linkage of their tax data going back to 1990 . Since the analysis focused only on respondents who consented to both types of linkage, the sample available included 2,137 observations.

## 4. Analytical Results

### 4.1 Linkage rate between 1990 and 2007

Data linkage fails when the unique matching key does not find a match in the linked file. There may be several reasons for this failure, and in the case of historical linkage, the potential sources are even more numerous. Other than errors arising from the processing of records, the extent of which is difficult to quantify, some of the reasons for a failed link are more easily identifiable and can be divided into two main sources: (1) the person did not file an income tax return or did not work for an employer during the taxation year; or (2) the person's matching key was not constant over time.

This second type of reason will introduce bias in the longitudinal analysis of data since the linkage failure cannot be attributed to the absence of an income tax return during those years. Although the SIN is a relatively stable key over time, it can change, notably in the case of immigrants who are assigned a temporary SIN on their arrival in Canada and who are then assigned a permanent SIN. ${ }^{3}$

The linkage rate for the T1 data files was calculated for the period from 1990 to 2007 to determine the degree to which it declines going back over time. Three different types of rates were calculated: (1) a gross rate using all of the available sample, (2) an adjusted rate using a sample excluding respondents aged 15 years or under and immigrants landed in, or in the year prior to, the fiscal year, and (3) a second adjusted rate based on a sample excluding respondents aged 20 years or under and immigrants landed in the three years preceding the fiscal year. The restrictions on age and immigrant status reflect the fact that these two groups are less likely to produce an income tax return during a given year.

As Figure 4.1-1 shows, the results indicate that the linkage rate drops going back in time regardless of the sample used for the calculation. However, the most significant reduction in the linkage rate occurs when the calculation is based on a sample without exclusions, in which case it drops from $96.7 \%$ in 2007 to only $60.5 \%$ in 1990. When we exclude those respondents for whom linkage was a priori more likely to fail, that is, those less likely to file an income tax return, the rate drops much less quickly and remains at over $80 \%$ in 1990 for both samples with exclusions.

[^63]Figure 4.1-1
Linkage rate of the personal income tax return file (T1) from 1990 to 2007


Source: LCS (2007) and linked data from T1 file (1990 to 2007)

### 4.2 Comparison of earnings from T1 and T4 files

One way to verify the accuracy of the information reported by a respondent on a survey is to compare it to information contained in administrative files. Administrative files are considered a more reliable data source and many studies use this type of methodology to assess the impact of non-response and attrition on estimate quality (Pyy-Martikainen and Rendtel, 2003; Roemer, 2002; Johnson and McMahon, 2002). However, very few studies make this type of comparison among different administrative files in order to determine the correspondence between the amounts reported in each source.

Earnings amounts in T1 data files and in T4 data files were compared for the period from $2001^{4}$ to 2007. The results show that the majority of cases-approximately $98 \%$ each year-present a similar earnings situation in both the personal file (T1) and the employer file (T4). In other words, when earnings are found in the T1 file, they are also found in the T4 file and vice-versa. In 2007, for example, there are only 33 cases with earnings from a single data source. Of that number, the majority of linked information, some 21 cases, comes from the T1 file and the median earnings associated with those cases is only one dollar (Table 4.2-2).

Table 4.2-1
Difference in earnings reported in the T1 file and the T4 file

|  | $\mathbf{2 0 0 7}$ | $\mathbf{2 0 0 1}$ |
| :--- | :---: | :---: |
| T1 earnings between (T4 earnings $-\$ 1000.00$ and T4 earnings $+\$ 1000.00$ ) | $98.66 \%$ | $98.83 \%$ |
| T1 earnings between (T4 earnings $-\$ 1.00$ and T4 earnings $+\$ 1.00$ ) | $95.17 \%$ | $87.59 \%$ |
| T1 earnings less than (T4 earnings $-\$ 1000.00$ ) | $0.79 \%$ | $0.90 \%$ |
| T1 earnings greater than (T4 earnings $+\$ 1000.00$ ) | $0.55 \%$ | $0.27 \%$ |

Source: LCS (2007) and linked data from T1 and T4 files (2001-2007)
When employment earnings are present in both files, the difference between the amounts reported is no more than one dollar in $95 \%$ of cases (Table 4.2-1). The reason for this finding is quite simple and is due to the fact that earnings amounts reported in the T 1 files are rounded to the nearest dollar, which is not done with the amounts reported in the T4 files. Referring to the results in Table 4.2-2, we find that the median earnings are very similar

[^64]from one data source to the other. From 2001 to 2007, the difference in median employment earnings calculated from the two data sources is on average $\$ 150.00$.

Table 4.2-2
Median employment earnings ${ }^{5}$ according to statements in the T1 and T4 files

|  | $\mathbf{2 0 0 7}$ |  |  | $\mathbf{2 0 0 1}$ |
| :--- | :---: | :---: | :---: | :---: |
|  | $\mathbf{T 1}$ | $\mathbf{T 4}$ | $\mathbf{T 1}$ | $\mathbf{T 4}$ |
|  | $\$ 31,745$ | $\$ 31,886$ | $\$ 30,774$ | $\$ 30,930$ |
| Statement of earnings in T1 and T4 files | $\$ 1.00$ | $\mathrm{n} / \mathrm{a}$ | $\$ 1.14$ | $\mathrm{n} / \mathrm{a}$ |
| Statement of earnings in T1 file only | $\mathrm{n} / \mathrm{a}$ | $\$ 278$ | $\mathrm{n} / \mathrm{a}$ | $\$ 462$ |

Source: LCS (2007) and linked data from T1 file and T4 file (2001-2007)

### 4.3 Profiles of earnings by age and sex

The lack of longitudinal data on certain types of research topics leads researchers to use more approximate analytical methods to examine certain phenomena. For example, creating synthetic cohorts is a common practice to monitor the evolution in earnings over time when data comes from cross-sectional surveys. Retrospectively linked data could overcome this data limitation and also allow analysis of any issue requiring use of a long data series.

To fully exploit the potential of linked information and to verify whether such information can produce analytical results comparable to those achieved using other data sources, an earnings profile by age and sex was developed for different cohorts based on birth year. Given that the earnings amounts reported were similar regardless of the source, the earnings reported in personal tax return files were used to develop the earnings profile, thus producing a longer data series than the data drawn from the employer files. The sample was divided into seven birth year groups, each with a 10-year interval, for which the change in employment earnings for different age groups was tracked.

As Figure 4.3-1 shows, the earnings profile developed using historical data reveals general trends comparable to those from other data sources, specifically: (i) an earnings profile increasing with age and beginning to decline around age 50 years as individuals retire from the workforce, and (ii) a higher earnings profile for men than for women for all age groups analyzed.

The earnings profile also shows trends by cohort similar to those found in related literature (Beach and Finnie, 2004; Boudarbat, Lemieux and Riddell, 2003; Burbidge, Magee and Robb, 1997). First, there is a trend of lower earnings at the beginning of their career for workers in the most recent cohorts (e.g. 1961-70 cohort versus 1971-80 cohort) and second, a faster growth in earnings ${ }^{6}$ for workers in the more recent cohorts. As a result, there is a trend toward higher earnings at career end for workers in the more recent cohorts than for the older cohorts (e.g. 1951-60 cohort versus 1941-50 cohort).

[^65]Figure 4.3-1

## Earnings profile by age and sex



Source: LCS (2007) and linked data from T1 files from 1990 to 2007

## 5. Conclusion

The results of this study represent only the first step in our research. At present, the preliminary findings indicate that the information provided by retrospectively linked data could be of sufficiently good quality to be used in analyzing phenomena that require the availability of a long data series. According to several studies, this type of linkage could also gather more reliable data for respondents likely to provide incorrect responses on their historical information (e.g. workers with several jobs in recent years).

In the specific case of the LCS, the historical linkage of tax data also made it possible to add complementary information to the survey data about the respondents' past (education, family life, work and health). The addition of this historical information allows analysis of longitudinal phenomena as of the first wave of the survey, an operation that would not normally be possible until after several waves. It also offers new tools to analysts without increasing the response burden of respondents or the costs associated with the survey's activities.

## References

Beach, C. and Finnie, R. (2004). A Longitudinal Analysis of Earnings Change in Canada. Analytical Studies BranchResearch Paper, Statistics Canada. http://www.statcan.gc.ca/pub/11f0019m/11f0019m2004227-eng.pdf

Burbidge, J.B., Magee, L. and Robb, A.L. (1997). Cohort, year and age effects in Canadian wage data.Independence and Economic Security of the Older Population (IESOP), McMaster University, Research Paper no.19. http://www.ciln.mcmaster.ca/papers/cilnwp13.pdf

Boudarbat, B., Lemieux, T. and Riddell, W.C. (2003). Recent Trends in Wage Inequality and the Wage Structure in Canada. University of British Columbia, TARGET, Research Paper no.6. http://faculty.arts.ubc.ca/criddell/rectrend.pdf

Huber, M. and Schmucker, A. (2009). Identifying and Explaining Inconsistencies in Linked Administrative and Survey Data: The Case of German Employment Biographies. Historical Social Research, Vol. 34-2009No. 3, 230-241.

Johnson, B.W. and McMahon, P.B. (2002). Using auxiliary information to adjust for non-response in weighting a linked sample of administrative records. Internal Revenue Service, Presentation to the 2002 American Statistical Association. http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-001045.pdf

Pyy-Martikainen, M. and Rendtel, U. (2003). The Effects of Panel Attrition on the Analysis of Unemployment spells. CHINTEX, Research Paper no. 10 .
http://www.destat.net/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/P rojekt/Downloads/WorkingPaper6 042003, property=file.pdf

Reimer, M. and Künster, R. (2004). Linking Job Episodes from Retrospective Surveys and Social Security Data: Specific Challenges, Feasibility and Quality of Outcome. Berlin, Max-Planck-Institut für Bildungsforschung. http://www.mpib-berlin.mpg.de/en/institut/dok/full/reimer/linkingj/Arbeitsber8.pdf

Roemer, M. (2002). Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation. U.S. Census Bureau, LEHD Program, Technical Paper no. TP-2002-22. http://lehd.did.census.gov/led/library/techpapers/tp-2002-22.pdf

Sears, J. and Rupp, K. (2003). Exploring Social Security Payment History Matched with the Survey of Income and Program Participation. Presented on the $18^{\text {th }}$ of November at the Federal Committee on Statistical Methodology of 2003. http://www.oecd.org/dataoecd/15/36/36232603.pdf

# Creation of a child-centric inter-agency data warehouse: The Longitudinal Study of Early Development 

Melissa R. Pfeiffer, Maushumi Mavinkurve, Meredith E. Slopen, Slavenka Sedlar, Allison E. Curry, Jisen Ho and Katharine H. McVeigh ${ }^{1}$


#### Abstract

Government agencies increasingly leverage administrative data for research and policy. However, linkage methodologies are rarely published. We utilized probabilistic technology to link children and siblings across five sources from two New York City agencies. We describe the process (data preparation, threshold setting, and quality assurance) and results (match rates, comparisons to expectations, and cohort characteristics). The Longitudinal Study of Early Development data warehouse contains data for $1,942,942$ children born $1994-2004$. The estimated false match rate is $0.6 \%$. Over half of the children were identified in multiple sources, and $20 \%$ have siblings. These techniques may be replicated to link data for other populations.


Key Words: Probabilistic linkage, administrative data, registries, linkage methods

## 1. Introduction

The Longitudinal Study of Early Development (LSED) data warehouse links children from five New York City (NYC) registries and administrative databases held by two local government agencies: administrative data from the Department of Education (DOE) and four sources from the Department of Health and Mental Hygiene (DOHMH) the birth, death and Lead Poisoning Prevention Program registries, and administrative data from the Early Intervention Program (EI) for children with developmental delays. This data warehouse will enable the investigation of questions pertaining to educational outcomes among children who received EI services for developmental delays as well as among those who had an elevated blood lead test, in addition to other topics. We found that details of linkage methodologies or guidelines on how to conduct and evaluate linkages were rarely published. Consequently, to construct a warehouse capable of addressing our research priorities, we needed to devise a strategy for executing the linkage. The dissemination of the techniques established for this multi-file linkage can aid others seeking to conduct this type of work.

## 2. Pre-linkage Process

### 2.1 Linkage preparation

Our pre-linkage process included standardizing variables, such as converting names of "Bob" and "Robbie" to "Robert". We also needed to define the unique entities that we would be identifying. For this project, the linkage was aimed at identifying unique children as well as sets of siblings. Ideally, we would have determined the criteria for assessing the linkage and our targets at this stage. Two evaluation measures that we used were false match rates and expectations of the yield of the linkages between sources. These measures will be presented below.

### 2.2 Methodological challenges

We encountered several specific challenges that others may experience in their linkage projects. For instance, two data sets had multiple observations per child (i.e., more than one observation per identification number). We also assumed there to be unresolved duplication (i.e., one child given more than one identification number) in three data

[^66]sets. Further, slightly different identifiers were available from each source, requiring a strategy for capturing partial linkages between records for a single child. Additionally, we wanted the final data warehouse to allow for analyses with various referent populations, such that some would include all children born in NYC, whereas others would include all children in the DOE, regardless of whether they were born in NYC. This necessitated an approach other than a series of two-file linkages. Finally, we needed to establish methods to evaluate the linkage, which included assessing the level of error in the linkage.

### 2.3 Data sets characteristics and identifiers for linkage

The data sets contained children born 1994 through 2004, except for DOE, in which the latest year of birth was 2001, as shown in Table 2.3-1. This was done so that all children included were at least in kindergarten; those born beyond 2001 were not old enough for enrolment when the data were prepared. Each of the five sources had a child identification number; in EI and Lead there could be more than one record per identification number, with different identifiers in each. Consequently, the number of observations is greater than the number of children. The EI, Lead and DOE data sets were expected to contain unresolved duplicates. Birth and Death, on the other hand, were both believed to contain unique children or only one identification number per child. This became a central assumption in our methodology.

The variables available for linkage were not consistent across data sources. Child's first name, last name, date of birth, sex and address were present in all sources. Child's full middle name was present in the Birth, Lead and Death data sets, but was recorded only as an initial in EI and DOE. Social security number was available only in the Birth, EI and Death data sets. There was greater variation in the availability of parental identifiers. The Birth and EI data sets contained mother's and father's first and last names and mother's date of birth; Lead provided only mother's information (full name and date of birth) for this linkage. DOE had guardian information, but it was unclear if that corresponded to the mother, father or someone else. Death had only father's last name.

Table 2.3-1
Characteristics of data sets linked to construct the Longitudinal Study of Early Development warehouse

| Source | Birth | EI | Lead | DOE | Death |
| ---: | ---: | ---: | ---: | ---: | ---: |
| Birth years <br> included | $1994-2004$ | $1994-2004$ | $1994-2004$ | $1994-2001$ | $1994-2004$ |
| Number of <br> observations | $1,380,608$ | 251,757 | $7,772,305$ | 617,934 | 8,331 |
| Number of <br> children | $1,380,608$ | 156,834 | $1,469,265$ | 617,934 | 8,331 |
| Unique children | Yes | No | No | No | Yes |

## 3. Linkage process

### 3.1 General linkage process

In general terms, linkage was conducted by executing multiple passes to link records together. Each pass grouped records on the basis of identical values on defined criteria and then linked records with comparable information. For example, one pass grouped records with identical values on components of the child's first name, last name, address, month and year of birth and then linked records within each group on the following variables: child's name, full birth date, gender, address, social security number, father's name, and mother's name and date of birth. The passes generated sets of potential matches, in which one record became the primary record and the others were considered secondary records for the same person. Each secondary record was given a weight, or an indication of how well it matched the primary record. A threshold was then established to separate potential matches. The threshold was the weight above which potential matches were accepted and below which they were rejected.

### 3.2 Phase 1: Birth and Death

We conducted all linkages using IBM's QualityStage software and used a two phase process to identify unique children. For the first phase, we included the two data sets, Birth and Death, that each had unique children (i.e., only one identification number per child). This allowed us to leverage the assumption that these data sets had no duplicates into a strategy to find the best match for each child. We used 10 passes for this phase. All of the potential matches were reviewed independently by two investigators. When those two investigators arrived at discordant decisions about whether the potential match was truly a match, the tie was broken by a third investigator.

During this phase, we linked 1,806 death records against $1,380,608$ birth certificates. We found 1,357 potential matches and accepted $96 \%(n=1,299)$ of them. One of our quality assurance measures was to see how often the reviewers came to the same decision and found that they were in agreement more than $98 \%$ of the time.

### 3.3 Phase 2: Multi-file linkage

For the next phase of the linkage, we created one data set that combined records from all five of the source data sets and then identified sets of records for each unique child within that one large data set. Using all five data sources at once allowed us to maximize linkages between records where only some of the identifiers were the same. For instance, a record could be brought into a potential match set because it had an address in common with one record and a nickname in common with another. Together those three records represent one child with a formal name, a nickname and two addresses. For this phase, we conducted nine passes, executing each one sequentially.

### 3.4 Sampling potential matches

The results of these passes were too large to allow for review of all of the potential matches as we did for the Birth-Death linkage, so we developed a sampling scheme to select and review a proportion of potential matches. We took a random sample of approximately 800 secondary records that had the same weight distribution as the entire pass. The samples were reviewed by two independent reviewers to decide which potential matches were true matches and which were false matches. After review, we calculated the cumulative false match rate at each weight as the proportion of false matches among all of the potential matches with the same or higher weight. We set our threshold at the lowest weight that had a cumulative false match rate below $1 \%$, having decided that we would accept that at most $1 \%$ of our matches were errors. As there is no standard for the ideal error rate, investigators must choose a level based on the objectives of the project and the importance of errors. Our decision was based in part on these data being used for research rather than surveillance or outreach; for example, the consequences of false matches are greater when patients are contacted regarding linked data, therefore the acceptable proportion of false matches may be lower when data are used for outreach. There is a balance in selecting the false match tolerance: accepting fewer false match errors would have meant increasing the number of true matches that were missed. An example of this process is shown in Table 3.4-1. At weight 49, the cumulative false match rate was $0.88 \%$ (six false matches among 682 potential matches with a weight of 49 or higher). Although the false match rate rose over $1 \%$ at weight 48 with seven false matches among 684 potential matches, it fell back below $1 \%$ as we added more true matches to the denominator. When we got to eight false matches out of 711 potential matches, the false match rate climbed above $1 \%$ and stayed there. So for this pass, our threshold was 43: any potential match with a weight 43 or higher was accepted, anything below that was rejected. The thresholds were applied to the entire pass, including potential matches reviewed during the threshold setting process. False matches above the threshold weight identified during the threshold setting process were retained and became part of our acceptable level of errors.

Table 3.4-1
Example of cumulative false match rates for threshold setting

| Weight | Number of <br> potential matches <br> reviewed | Number <br> determined to <br> be false matches | Cumulative number of <br> potential matches reviewed <br> (true matches + false matches) | Cumulative <br> number of <br> false matches | \% false <br> matches |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 49 | 2 | 0 | 682 | 6 | 0.880 |
| 48 | 2 | 1 | 684 | 7 | 1.023 |
| 47 | 3 | 0 | 687 | 7 | 1.019 |
| 46 | 3 | 0 | 690 | 7 | 1.014 |
| 45 | 4 | 0 | 694 | 7 | 1.009 |
| 44 | 6 | 0 | 700 | 7 | 1.000 |
| 43 | 6 | 706 | 7 | 0.992 |  |
| 42 | 1 | 711 | 8 | 1.125 |  |
| 41 | 5 | 0 | 713 | 8 | 1.122 |

### 3.5 Combine linked records

After thresholds were set for all passes, all of the accepted matches for a single child were combined, providing another opportunity to create linkages. This process again allowed records that were not directly linked to be identified as belonging to a single child. For instance, records $\mathrm{A}, \mathrm{B}$ and C all represented a single child if records A and B, though not linked directly with each other, were both linked with record C. This sometimes created violations of our assumption that all birth records represented unique children, as did death records. Sets of records with these violations, such as sets of records for a single child with more than one birth certificate, were hand reviewed and separated into distinct children.

### 3.6 Sibling linkage

After the child linkage was completed, we conducted a similar linkage process using mother's information in order to create sibling sets, or children born to the same mother. Death and DOE were not included in this linkage because neither had information specifically on the mother, but death and DOE records are part of a final sibling set when they are connected with a birth, EI, or Lead record for a given child. The sibling linkage was done in three passes using the threshold setting process described above.

## 4. Linkage evaluation

### 4.1 Overall false match rate

One of our evaluation measures was the overall false match rate. Based on hand-review determination of a final sample of 1,390 children, we estimated that $0.4 \%$ of all children are false matches. Among the children for whom more than one original identification number was brought together, meaning that at least two different records came together as a unique child, the false match rate was estimated at $0.6 \%$. Source-specific false match rates varied from $0.0 \%$ to $0.9 \%$. When we evaluated the false match rate among siblings, we found that $1.6 \%$ of sibling sets are children with different mothers (i.e., false matches). As a further evaluation measure, we looked at how often the children's dates of birth are implausibly close together. This was defined as children who were born more than two days apart but less than seven months apart. Among the sets with more than one child, less than $1 \%$ have an invalid birth interval.

### 4.2 Comparison with expectations

We compared the results of the linkage with the proportion of matches we expected to find. Our expectations were derived from previous linkages conducted with different software, somewhat different data sets and for different purposes, such as a previous EI-Birth linkage and Lead-Birth linkages. We also used available information, such as the proportion of foreign-born children in DOE and in- and out-migration of families with children. The former indicated the minimum proportion of children in DOE who would not be expected to have a record in the birth
certificate registry. Table 4.2-1 shows the proportion of children in the source data sets (the columns) who were linked with a record in one of the other sources (the rows). For instance, $85 \%$ of the children in the EI data set had a birth certificate, and fewer than $1 \%$ of the children in the EI data set were linked with another child in the EI data set, meaning that fewer than $1 \%$ were unresolved duplicates. We compared the yield of the linkage with what we expected in 10 cells. In the cells with diagonal lines, the yield of the linkage was within our expected range, the light grey cells had a lower yield, and dark grey indicates a higher yield than expected.

Table 4.2-1
Comparison of linkage results with expectations of the linkage yield

|  |  | Among children in these data sets |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Birth | EI | Lead | DOE | Death |
| Percent that matched with these data sets | Birth | 0.0 | 85,0 | 65.9 | 74.3 | 86.5 |
|  | EI | 9.6 | 0.6 | 8.8 | 12.3 | 12.4 |
|  | Lead | 67.7 | 79.7 | 3.4 | 77.7 | 10.2 |
|  | DOE | 44.8 | 62.0 | 44.9 | 1.1 .3 | 3.0 |
|  | Death | 0.5 | 0.7 | 0.1 | 0.0 | 0.0 |

## 5. Results

### 5.1 Demographic characteristics

As a result of the linkages, we identified $1,942,942$ unique children (Table 5.1-1) who constitute the full cohort of the LSED data warehouse. The proportion of males and females is nearly equal, with a small proportion with gender unknown. Over 200,000 children (11.2\%) were born in 1994. The number born in each year drops by $0.4 \%$ to $8.7 \%$ until 2004, in which year about 150,000 children ( $7.6 \%$ ) were born. Race/ethnicity for the child was derived from data in all of the four sources in which it was recorded (Lead, EI, DOE and Death) and assigned according to the following hierarchy: Hispanic, if Hispanic in any source, regardless of other values; multiple race, if more than one non-Hispanic value (i.e. black, white, Asian/Pacific Islander, other) was recorded; black, white, Asian/Pacific Islander or other if that value was the only one recorded; unknown if no race/ethnicity information was found in any source. We observe that the majority are Hispanic, but more notably, over half of the children have unknown race/ethnicity. When we instead look at mother's race/ethnicity from the birth certificate, we have fewer unknowns. Consequently, mother's race/ethnicity might be used instead of child's for some analyses, or analysts may choose to use child's with unknown values imputed with mother's race/ethnicity.

Table 5.1-1
Demographic characteristics of children with a record in the Longitudinal Study of Early Development warehouse ( $\mathrm{N}=1,942,942$ )

| Characteristic | Number | Percent |
| ---: | ---: | ---: |
| Child gender |  |  |
| Male | 985,619 | 50.7 |
| Female | 934,050 | 48.1 |
| Unknown | 23,273 | 1.2 |
|  |  |  |
| Child's race/ethnicity |  |  |
| Hispanic | 314,405 | 16.2 |
| Black, non-Hispanic | 265,194 | 13.7 |
| White, non-Hispanic | 144,041 | 7.4 |
| Asian/Pacific Islander | 88,751 | 4.6 |
| Other/multiple | 14,536 | 0.7 |
| Unknown | $1,116,015$ | 57.4 |
| Mother's race/ethnicity |  |  |
| Hispanic | 446,069 | 23.0 |
| Black, non-Hispanic | 364,707 | 17.8 |
| White, non-Hispanic | 407,281 | 21.0 |
| Asian/Pacific Islander | 154,104 | 7.9 |
| Other/multiple | 4,311 | 0.2 |
| Unknown | 566,470 | 29.2 |

### 5.2 Record source

There are multiple reasons that children do not have a record in all sources. Not all children experience record-generating events; for instance, only a small proportion are referred to EI, and some are not tested for lead poisoning despite mandated screening at one and two years of age. Additionally, children may not have been residing in NYC when events occurred, and thus, their birth or lead test may be recorded in another jurisdiction. Additionally, all school children do not appear in DOE, since that includes only those enrolled in public school. However, most children $(71.1 \%)$ in the cohort have a birth record (Table 5.2-1). A slightly higher proportion of children (73.0\%) have a Lead record, and nearly a third have a DOE record. Nearly half of the children have data from just one source, but one-third has data from two sources and one-fifth have data from three sources.

Table 5.2-1
Source characteristics of children with a record in the Longitudinal Study of Early Development warehouse

| Characteristic | $\mathbf{N}$ | Percent |
| ---: | ---: | ---: |
| Sources |  |  |
| Birth | $1,380,535$ | 71.1 |
| EI | 155,880 | 8.0 |
| Lead | $1,417,655$ | 73.0 |
| DOE | 609,321 | 31.4 |
| Death | 8,331 | 0.4 |
|  |  |  |
| Number of sources |  | 42.5 |
| One | 826,294 | 34.2 |
| Two | 664,144 | 20.2 |
| Three | 392,955 | 3.1 |
| Four | 59,470 | 0.0 |
| Five | 79 |  |

### 5.3 Siblings

Most children ( $63.2 \%$ ) do not have a sibling in the data warehouse, but we identified 306,240 sibling sets. Most sibling sets $(75 \%)$ have two children, with up to 14 siblings in a set. We expect that we did not find more or larger sibling sets for two reasons. One is that the mothers are still in their reproductive years, and the other is that some families move out of NYC after a child is born, and thus, younger siblings are not included in our cohort.

## 6. Conclusion and recommendations

### 6.1 Conclusion

In creating the LSED data warehouse, we developed a unique methodology that differed from similar linkage projects in several ways. No single source was the focal population, and consequently, we created linkages between all combinations of the five original sources, each of which can be used as the base population. To create these linkages, we conducted a simultaneous multi-file linkage and determined thresholds for acceptable matches using a false match rate. The simultaneous process allowed us to leverage incomplete linkages that might not have been identified in a sequential linkage approach. We invite researchers to adapt these methods to their own projects. Additionally, external investigators are welcome to apply to use the data and may contact us at LSED@health.nyc.gov to request more information.

### 6.2 Recommendations

There are elements that we would change if conducting the linkage again and recommendations that we would like to share with others beginning linkage projects. First, several sources contained more than one record per identification number, which complicated the linkage and the evaluation. We recommend allowing only one record per unique identification number per source. Supplementary variables containing alias information can be used rather than additional observations (i.e., shifting the data from a row orientation to a column orientation). Second, it is important to establish expectations of the linkage outcomes early, so that the results do not influence your expectations. Furthermore, the expectations will inform whether the process is proceeding too slowly because the linkage criteria are too conservative. Third, a critical step is the establishment of evaluation measures, their calculations and targets. We primarily used a false match rate, but did not assess the false non-match rate. The DOHMH has since established methods for assessing sensitivity and specificity as alternate measures, which use as the denominator record-based, hand-reviewed gold standards instead of person-based linkage results. Whichever assessment tools are used, acceptable error rates should be established before the linkage begins. These values will be influenced by the purpose of the project. The fact that this linkage was a research project rather than routine health department surveillance influenced our error tolerance. Finally, we encourage researchers to document and share their linkage methods to improve the collective expertise of the field.

# The Interplay between Administrative Data and Survey Data in the Rotterdam Early School Leavers Monitor 

Jannes Hartkamp ${ }^{1}$


#### Abstract

At the 2000 Lisbon Summit of the European Union the ministers of education of the member states agreed to aim for a 50 percent reduction of the number of early school leavers in a period of ten years. By a standard definition, early school leavers are youngsters who are not enrolled in school and have not (yet) completed upper secondary education. The 'Lisbon target' has been adopted by national, regional and local governments throughout the EU. The City of Rotterdam, the Netherlands, incorporated the target in its policy goals and launched an annual early school leaver monitor in 2000. This ESL-Monitor combines administrative data from the municipal pupil register and data from an additional survey among youngsters whom the register does not cover sufficiently. Interplay between administrative data and survey data in the Monitor is threefold: 1. analysis of the (lack of) data in the register reveals which youngsters should be contacted for the additional survey; 2. administrative data and (weighted) survey data are combined into one integrated research dataset; and 3. information from the survey is used to update the register. Because of this feedback process, the scale of the additional survey has continuously decreased. In addition to this, the survey has also helped to attach the problem it measures.


Key Words: early school leavers, education, register data, survey data, Lisbon target.

## 1. Introduction

### 1.1 Early School Leavers

Early school leavers, also known as lower level leavers, are youngsters who are not in education and have not yet obtained sufficient qualifications. Definitions vary, but in international comparisons 'sufficient' qualifications usually start at upper secondary, the International Standard Classification of Education (ISCED) Level 3. ${ }^{2}$ Upper secondary qualifications are regarded here as the minimum basis for a start at the labour market and in general for successful economic, social and personal development later in life. In general early school leavers are economically and socially clearly more vulnerable than those with higher qualifications.

### 1.2 The Lisbon Target and Babylonian Confusion

At the European Union Summit in Lisbon in 2000, the ministers of education of the EU member states agreed to aim for a fifty percent reduction of the number of early school leavers in a period of ten years. This 'Lisbon target' has since been adopted by national, regional and local governments throughout the EU.

How exactly the number of early school leavers should be monitored was not specified when the Lisbon target was set. As a result, definitions and measurement methods have varied widely between countries and over the years. One important source of variation has been the national definition of 'upper secondary' qualifications. As the term fits some educational systems better than others, the border between lower and upper secondary education may sometimes be rather disputable. Also, changes in the educational system with simultaneous changes in the definition

[^67]of upper secondary qualifications may bring about rather artificial sudden changes in the number of early school leavers from one year to another.

A fundamental difference also exists between dynamic and static approaches to measuring early school leavers. A dynamic approach attempts to count the number of new leavers within a certain period (in the rule a school year): those who within this period have left school without sufficient qualifications. From this gross number of new leavers one may or may not attempt to arrive at a net figure of new early leavers by detracting the re-entrants during the same period: those who had left school early in a previous period but started to follow education again now. Measuring such flows has proven to be a complex task. Results of this approach - adopted for instance by the Dutch education department - have shown large fluctuations heavily influenced by changes in registration procedures and practices. The alternative, static approach attempts to measure the number of early school leavers at a certain point in time: all who are not enrolled and have not yet obtained a certain level of qualification count as early leavers, regardless of when they have left school. The results of this approach - adopted for instance by Eurostat - have turned out to be far more robust over the years. A last example of differences in definition and measurement concern the age group. Eurostat figures concern the number of early school leavers as a percentage of the total population aged between 18 and 24 years, Organisation for Economic Co-operation and Development (OECD) figures for example apply to the age group 20-24.

### 1.3 The Rotterdam Target and Monitor

Rotterdam is the second-largest city in The Netherlands, with a population of 593.000 as of January $1^{\text {st }} 2010 .{ }^{3}$ Compared to the national average, the population of Rotterdam has a relatively high proportion of youngsters, a relatively high proportion of immigrants and a relatively low average level of education.
The City of Rotterdam was an early adopter of the 'Lisbon target'. June 2001 the City launched an Early School Leaver Attack Plan, comprising various projects tackling early leaving. The 2002 programme of the municipal executive of mayor and aldermen included as a main goal a thirty percent reduction in the number of early school leavers aged 17 through 22 between 2000 and 2006, five percent per year. The following municipal programme similarly included a twenty percent reduction target for the 2004-2008 period.

Since 2000 specific research is carried out annually under responsibility of the Youth and Education Office of the City of Rotterdam, in order to track the ESL-developments and to establish whether the policy targets are being reached. This ESL-Monitor is based on analysis of data from the municipal pupil administration system and an additional survey among youngsters for whom the information available from the register is insufficient. The Monitor has been carried out by DESAN Research Solutions since 2005.

## 2. The Rotterdam Early School Leavers Monitor

### 2.1 Population and ESL definition

The primary purpose of the Rotterdam Early School Leavers Monitor is to monitor the development of the number of early school leavers as a percentage of the total youth population living in Rotterdam. The population of the monitor consists of all inhabitants of the Rotterdam municipality aged 17 through 22. Reference date for population, age and education status is July $31^{\text {st }}$ of the given year, which counts as the last day of the school year. Anyone who is not in school or training ${ }^{4}$ at that date and has not yet obtained a qualification at minimum upper-secondary level (ISCED$3)^{5}$ at that date is regarded as early school leaver at that date. Whereas the Rotterdam early school leavers percentage

[^68]is the prime figure resulting from the ESL Monitor and the main indicator in the policy targets, the monitor also provides data on the percentage of youngsters with upper-secondary qualifications, the percentage of youngsters in education and the percentage of early leavers, broken down by age, sex, district, ethnicity and other variables.

### 2.2 Data sources

The Rotterdam ESL Monitor is based on analysis of data from the municipal pupil administration system and an additional survey among youngsters for whom the information available from the register is insufficient. The main source of data is the municipal pupil administration system. For each member of the youth population the system contains data on all education stages from primary school through secondary school, along with personal details (name, sex, date of birth, country of birth, current address, date of immigration if applicable, parental countries of birth). The 'school career' data include, for each stage, start and end date of enrolment, school details, level and type of education and 'exit level' (qualification or else last class if the stage ended without a qualification). The school career information is supplied to the municipality by primary and secondary schools inside and outside Rotterdam, for all pupils attending school in Rotterdam or living in Rotterdam.

For the majority of the Rotterdam youth population, the data in the municipal administration system - some minor exceptions left aside - would be sufficient as a basis for the ESL Monitor. For youngsters who have moved to the Rotterdam region during their youth however, education data for the period until they moved to or started to go to school in Rotterdam are missing from the system. If they attend primary or secondary school, they are by definition no early leavers. But if they do not attend school, they may or may not be early leavers, depending on previously obtained qualifications.

Newcomers are 'blanks' in the register, until they or their parents or their school provide the missing information. Especially newcomers who have passed schooling age, are not attending secondary school anymore and do not live with their parents, often fail to provide information on their school career despite repeated requests of the education department of the municipality. For the administration system a 'guilty until proven innocent' approach is used: youngsters for which no information is available count as 'unknown early leavers'. Some of these may indeed have no upper secondary school qualification, but others may have sufficient, but still unknown, qualifications.

In order to provide the information which is missing from the municipal administration system, an additional postal survey is carried out annually among the 'unknown early leavers'. Core questions concern previous and current education and qualifications.

### 2.3 The Monitor process

The process of the ESL Monitor, illustrated in the figure below, starts with an analysis of the administrative data (1). The analysis (A) identifies the cases (youngsters) for which insufficient data are available to establish whether they count as early school leavers (2 and 3).

Figure 2.3-1
Schematic representation of the ESL Monitor process


These youngsters are kindly requested to provide the missing core data in a one-page postal questionnaire or via internet (A). A minority responds (B). The information provided by the respondents (4) is then (C) processed, matrix-weighted (simultaneously on sex, age, ethnic background, district and other background characteristics) to account for differences in response rates and extrapolated ('blown up') to represent the operational gross sample
again (5). The weighted and extrapolated data from the additional survey are merged back (D) with the known administrative data. The resulting integrated research dataset (6) is the basis for all analysis (E), providing the ESL indicator (X) and other figures.

### 2.4 Outcomes

Once the additional survey has been carried out and processed and the final integrated dataset has been constructed, the analysis for the main indicator - the percentage of early school leavers - is fairly simple: anyone who is not enrolled in education at the reference date and has not yet obtained an upper secondary qualification at that date is regarded as an early school leaver. For the period 2000-2008 the monitor has shown a strong decrease in the number of early school leavers as a percentage of the total Rotterdam population aged 17 through 22: from 28,0 percent in 2000 to 15,2 percent in 2008 (Figure 2.4-1). The 2008 result means a 46 percent reduction in comparison to 2000 and a 37 percent reduction since 2004, the year the 2006-2010 municipal executive set as its reference point (the dotted line in Figure 2.4-1 represents the policy target). The reduction in percentage of early leavers in Rotterdam over the 2000-2008 period which the ESL Monitor shows (for age group 17-22) is much stronger than the reduction Eurostat figures show for the Netherlands or the European Union as a whole (for the age group 18-24). ${ }^{6}$

Figure 2.4-1
Percentage of early school leavers within the Rotterdam population aged 17-22, 2000-2008 (July $\left.31^{\text {st }}\right)^{7}$


## 3. Interplay

### 3.1 Interplay between register and survey

The Rotterdam ESL Monitor is a clear instance of strong interplay between registers (in this case the municipal pupil administration system) and surveys (the additional survey among 'unknown early leavers'). The gaps in the register define the sample for the additional survey and the existing administrative data and new survey data are combined into one integrated research dataset. But the results from the survey do not just serve the monitor, they are also used

[^69]to update the register. Once it has been established that a respondent has obtained sufficient qualifications, he or she does not have to be contacted again. Hence the additional survey has helped to decrease the size of the additional survey year after year: for the 2004 measurement 14375 youngsters were approached (about one quarter of all 17 to 22 year olds) and 3463 responded ( $24,1 \%$ ), for the 2008 measurement only 2985 were approached (about five percent of the target population) and 487 responded ( $16,3 \%$ ). The decrease in response rates is explained by the fact that an ever growing proportion of the sample is formed by the non-respondents of previous years, who are most likely to not respond again. Together with a significant improvement of the exchange of data between regional and national education registers reducing the number of 'unknowns' in the administration, the survey itself has, through the feedback of data into the register, helped to make itself more or less redundant.

### 3.2 Interplay between survey and subjects

Beside the interplay between register data and survey data, there is another type of interactivity characterising the ESL Monitor: the additional survey on early leaving has actually served, albeit perhaps to a limited degree, to decrease the number of early school leavers. Under different conditions such interference may be regarded as a deadly sin against sane statistics, but the Monitor has indeed influenced what it measured. Merely asking youngsters about their school career and qualifications may already stimulate some early leavers to consider returning to education or training. But the accompanying letter has also pointed the youngsters to sources of information and advice on schooling and in the questionnaire respondents could indicate if they would like the local Youth Counter to contact them. Every year about one third of the respondents indeed answered they would appreciate this. The first time this option was included, in 2005, the Youth Counter in fact got heavily overburdened for months.

## 4. Postscript

Despite and partly thanks to its success, the Rotterdam ESL Monitor as described here has been discontinued after the 2008 measurement and will be replaced in 2011 by a new monitor. The reference date probably moves to October 1st, some definitions are revised, the additional survey - which had already dwindled in scale - has been abandoned and the focus of the monitor will also shift due to reformulation of policy goals and targets. Despite the changes one will strive for a continuation of the core trend figures, which may be regarded as unique at least inside the Netherlands.

## References

Hartkamp, J. (2009). Monitor Voortijdig Schoolverlaten Rotterdam 2009, 8-meting, eind schooljaar 2007/2008, Amsterdam: DESAN Research Solutions.

European Commission Directorate General for Education and Culture (2005), Study on Access to Education and Training, Basic Skills and Early School Leavers (Ref. European Commission DG EAC 38/04). Lot 3: Early School Leavers. Final Report. Brussels: European Commission.

## SESSION 8A

USE OF AUXILIARY SOURCES IN WEIGHTING

# Combining information from two independent surveys through the Pseudo Empirical Likelihood Method 

Changbao Wu and Jae-kwang Kim ${ }^{1}$


#### Abstract

Recent work on using a pseudo empirical likelihood (PEL) approach to finite population inferences with complex survey data focused primarily on a single survey sample, non-stratified or stratified, with considerable effort devoted to computational procedures. This paper presents PEL methods for combining information from two independent surveys. We consider two scenarios commonly encountered in practice: (1) the study variable y and the vector of auxiliary variables $x$ are observed from the sampled units for both surveys; and (2) the vector of auxiliary variables $x$ is observed from a large survey while both the study variable $y$ and auxiliary variables $x$ are observed from a smaller survey. Our main focus is on the optimal point estimation of finite population parameters and the construction of PEL ratio confidence intervals, using either a chi-square approximation or a bootstrap calibration method. The proposed approach has several advantages over the traditional approach and also provides a general tool for survey weight construction using auxiliary sources of information. Simulation results on the performance of the PEL method relative to the traditional method will be reported.


This is joint work with Jae-kwang Kim of Iowa State University and J.N.K. Rao of Carleton University.

[^70]
# Estimation of correlations between cross-sectional estimates from repeated surveys - An application to the variance of change 

Yves G. Berger ${ }^{1}$ and Rodolphe Priam ${ }^{2}$


#### Abstract

Measuring change over time is a central problem for many users of social, economic and demographic data and is of interest in many areas of economics and social sciences. Smith et al. (2003) recognised that assessing change is one of the most important challenges in survey statistics. The primary interest of many users is often in changes or trends from one time period to another. A common problem is to compare two cross-sectional estimates for the same study variable taken on two different waves or occasions, and to judge whether the observed change is statistically significant. This involves the estimation of the sampling variance of the estimator of change. Estimation of variance of change would be relatively straightforward if cross-sectional estimates were based upon the same sample. Unfortunately, samples from different waves are usually not completely overlapping sets of units, because of rotations used in repeated surveys. This implies that crosssectional estimates are not independent. Correlation plays an important role in estimating the variance of a change between the cross-sectional estimates. The unbiasedness of an estimator of a correlation is crucial, because a small bias can significantly over-estimate or under-estimate the variance of change (Berger, 2004). Several methods can be used to estimate correlations, some of which use re-sampling and/or Taylor linearization. We propose to use a multivariate linear regression approach to estimate the correlation. The proposed estimator is not a model-based estimator, as this estimator is valid even if the model does not fit the data. We show that the regression approach gives design-consistent estimator for the correlation when the finite population corrections are negligible. We show how the proposed estimator can accommodate stratified and two-stage sampling designs. We also show how the proposed estimator can be used for estimator of correlation between complex estimators of change.


Key Words: Finite population corrections, Linearization, Multivariate regression, Rotation,Stratification, Two-stage sampling, Unequal probabilities

## 1. Introduction

Suppose, we wish to estimate the absolute change $\Delta=\tau_{2}-\tau_{1}$ between two population totals $\tau_{1}=\Sigma_{i \in U} y_{1 ; i}$ and $\tau_{2}=\Sigma_{i \in U} y_{2 ; i}$, from wave 1 and wave 2 respectively; where $U$ denotes the population which is assumed to be the same at wave 1 and 2 . The quantities $y_{1 ; i}$ and $y_{2 ; i}$ are respectively the value of the variable of interest for wave 1 and 2. Suppose that $\Delta$ is estimated by $\hat{\Delta}=\hat{\tau}_{2}-\hat{\tau}_{1}$; where $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ are two cross-sectional estimators which can be Horvitz-Thompson (1952) estimators defined by (2). The design-based variance of the change $\hat{\Delta}$ is given by

$$
\begin{equation*}
\operatorname{var}(\hat{\Delta})=\operatorname{var}\left(\hat{\tau}_{1}\right)+\operatorname{var}\left(\hat{\tau}_{2}\right)-2 \operatorname{corr}\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right) \sqrt{\operatorname{var}\left(\hat{\tau}_{1}\right) \operatorname{var}\left(\hat{\tau}_{2}\right)} \tag{1}
\end{equation*}
$$

Standard design-based estimators can be used to estimate the variances $\operatorname{var}\left(\hat{\tau}_{1}\right)$ and $\operatorname{var}\left(\hat{\tau}_{2}\right)$. The correlation $\operatorname{corr}\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)$ is the most difficult part to estimate because $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ can be estimated from different samples.

In § 2, we defined the class of rotation schemes considered. The proposed estimator for the correlation is defined in § 3. In §4, we show how the statistical software R (The R-project) can be used to implement the proposed estimator. In §5, we define the rotating conditional Poisson sampling (RCPS) design which can be used to justify the proposed estimator defined in §3. We show how the proposed estimator can be used to accommodate stratification in § 7 and

[^71]two stage designs in § 9. In § 8, we give the result of a brief simulation study. In § 10, we show how linearization can be used to estimate the variance of complex measures of change.

## 2. Rotation scheme

Let $s_{1}$ and $s_{2}$ denote respectively the first and second wave samples. Consider that $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ are respectively the cross-sectional estimators based respectively on the data of $s_{1}$ and $s_{2}$.

The estimation of the correlation would be relatively straightforward if $s_{1}$ and $s_{2}$ were the same sample. Unfortunately, $s_{1}$ and $s_{2}$ are usually not completely overlapping sets of units, because repeated surveys use rotating designs which consist in selecting new units to replace old units that have been in the survey for a specified number of waves. The units sampled on wave 1 and 2 represent usually a large fraction of the sample $s_{2}$. This fraction varies from the surveys to surveys.

Assume that $s_{1}$ is a probability sample without replacement with first-order inclusion probabilities $\pi_{1 ; i}$. Suppose that $s_{2}$ is a simple random sample without replacement (SRSWOR) sample of $n_{12}$ units selected without replacement from $s_{1}$ combined with a sample of $n_{1 \mid 2}$ units selected without replacement from $U / s_{1}$ with probabilities $q_{i}$; where $U / s_{1}$ is the set of units not selected at wave 1 and $n_{1 \mid 2}=n_{2}-n_{12}$. Tam (1984) studied this scheme when $\pi_{1 ; i}=n_{1} / N$ and $q_{i}=n_{1 \mid 2} /\left(N-n_{1}\right)$.

The conditional probability of selecting $i$ given $s_{1}$ is $\pi_{2 ; i}\left(s_{1}\right)=g \delta\left\{i \in s_{1}\right\}+q_{i} \delta\left\{i \notin s_{1}\right\}$, where $g=n_{12} / n_{1}$ denotes the fraction of the common sample. The function $\delta\{A\}$ is the indicator function which is equal to one when $A$ is true and zero otherwise. The first-order inclusion probabilities of $s_{2}$ are given by $\pi_{2 ; i}=E_{1}\left(\pi_{2 ; i}\left(s_{1}\right)\right)=g \pi_{1 ; i}+q_{i}\left(1-\pi_{1 ; i}\right)$; where $E_{1}(\cdot)$ denote the design expectation with respect to the first wave design. Thus the first and second wave Horvitz-Thompson (1952) estimators are given by

$$
\begin{equation*}
\hat{\tau}_{1}=\sum_{i \in s_{1}} y_{1 ; i} / \pi_{1 ; i} \quad \text { and } \quad \hat{\tau}_{2}=\sum_{i \in s_{2}} y_{2 ; i} / \pi_{2 ; i} \tag{2}
\end{equation*}
$$

Note that $\pi_{2 ; i}=\pi_{1 ; i}$, when $q_{i}=\pi_{1 ; i}(1-g) /\left(1-\pi_{1 ; i}\right)$.

## 3. Estimation of the correlation by the multivariate regression approach

Several methods can be used to estimate the correlation in (1) (e.g. Kish, 1965; Tam, 1984, Holmes \& Skinner, 2000, Berger, 2004; Qualité \& Tillé, 2008; Wood, 2008). We propose to use the following linear regression model to estimate the correlation.

$$
\begin{equation*}
\binom{\breve{y}_{1 ; i}}{\breve{y}_{2 ; i}}=\binom{\beta_{1}^{(1)} z_{1 ; i}+\beta_{2}^{(1)} z_{2 ; i}+\beta_{12}^{(1)} z_{1 ; i} \times z_{2 ; i}}{\beta_{1}^{(2)} z_{1 ; i}+\beta_{2}^{(2)} z_{2 ; i}+\beta_{12}^{(2)} z_{1 ; i} \times z_{2 ; i}}+\boldsymbol{\varepsilon}_{i} ; \tag{3}
\end{equation*}
$$

where $i \in s=s_{1} \cup s_{2}$ and the residuals $\boldsymbol{\varepsilon}_{\boldsymbol{i}}$ have a bivariate distribution with mean $\mathbf{0}$ and an unknown variancecovariance matrix $\boldsymbol{\Sigma}$. The dependent variables in the model are

$$
\begin{equation*}
\bar{y}_{1 ; i}=y_{1 ; i} / \pi_{1 ; i} \delta\left\{i \in s_{1}\right\}, \quad \bar{y}_{2 ; i}=y_{2 ; i} / \pi_{2 ; i} \delta\left\{i \in s_{2}\right\} . \tag{4}
\end{equation*}
$$

The independent variables $z_{1 ; i}$ and $z_{2 ; i}$ are design variables defined by

$$
\begin{equation*}
z_{1 ; i}=\delta\left\{i \in s_{1}\right\}, \quad z_{2 ; i}=\delta\left\{i \in s_{2}\right\} \tag{5}
\end{equation*}
$$

Note that there is no intercept in the model (3).
It is important to note that we have the following fixed size constraints $\sum_{s} z_{1 ; i}=n_{1}, \Sigma_{s} z_{2 ; i}=n_{2}$ and $\sum_{s} z_{1 ; i} z_{2 ; i}=n_{12}$, because only samples with the required sample sizes can be selected. Thus, by using the design variables as independent variables, we are conditioning of them. This takes into account the fixed size constraints in
the estimation of the correlation. This can be justified by using a rotating conditional Poisson sampling design (see § $5 \& 6$ ). Note that this model includes interactions between the variables $z_{1 ; i}$ and $z_{2 ; i}$. These interactions capture the rotation of the sampling design which is represented by the constraint $\Sigma_{s} z_{1 ; i} z_{2 ; i}=n_{12}$.

Let $\hat{\boldsymbol{\Sigma}}$ be the ordinary least squares estimate of residuals variance-covariance matrix $\boldsymbol{\Sigma}$. An estimate for the correlation between $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ is given by (see Theorem $2 \& 3$ in $\S 6$ )

$$
\begin{equation*}
\operatorname{côrr}\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)=\hat{\Sigma}_{12} / \sqrt{\hat{\Sigma}_{11} \hat{\Sigma}_{22}} \tag{6}
\end{equation*}
$$

where $\hat{\Sigma}_{i j}$ is the component $(i, j)$ of the matrix $\hat{\boldsymbol{\Sigma}}$. Note that the estimator of the variance of change is always positive, because $\left|\operatorname{corrr}\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)\right|<1$, as $\hat{\boldsymbol{\Sigma}}$ is non-negative definite. Note that the parameters $\beta_{1}^{(0)}, \beta_{2}^{(0)}, \beta_{12}^{(0)}, \beta_{1}^{(1)}, \beta_{2}^{(1)}$ and $\beta_{12}^{(1)}$ are nuisance parameters, as we are only interested in the estimate $\hat{\boldsymbol{\Sigma}}$.

In the following $\S \S$, we will show that the multivariate linear regression approach is suitable for estimating the correlation, and how this approach could be extended to include stratification, two-stage sampling and more complex measures of change.

The multivariate regression approach is not based on a super-population model, as the model (3) does not need to fit the data. This approach gives a design-based consistent estimator for the correlation whether or not the model fits the data (see Theorem 3 in § 6). However, it relies on several assumptions such that the finite population corrections are negligible, the samples are selected with high entropy, and that the sampling distribution of $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ is normal. These conditions are clearly stated and discussed in $\S 6$. The multivariate regression approach is an extension of Berger (2004) methods when the finite population corrections are negligible, which is usually the case for social surveys.

Note that when $\pi_{1 ; i}=n_{1} / N$ and $q_{i}=n_{1 \mid 2} /\left(N-n_{1}\right)$, the quantity $\hat{\Sigma}_{12}$ reduces to the Tam (1984) estimator

$$
\hat{\Sigma}_{12}=N^{2} n_{21}\left(n_{1} n_{2}\right)^{-1} \operatorname{cov}\left(y_{1 i}, y_{2 i}\right)
$$

where $\operatorname{cov}\left(y_{1 i}, y_{2 i}\right)=n_{21}^{-1} \Sigma_{i \in s_{12}} y_{1 i} y_{2 i}-\bar{y}_{1} \bar{y}_{2}$ and $\bar{y}_{\ell}=n_{21}^{-1} \Sigma_{i \in s_{12}} y_{\ell i} \quad(\ell=1,2)$. The proposed estimator is more general, because it takes the unequal probabilities into account as well as stratification (see § 7). The proposed estimator is also easier to implement because it is based on a multivariate regression approach which can be implemented with most statistical softwares (see §4).

## 4. Estimation of the correlation with the software $\mathbf{R}$ (The R-project)

The multivariate regression approach is simple to implement as it can be easily implement is most statistical softwares. It is only necessary to create the design variables $z_{1 ; i}$ and $z_{2 ; i}$. For example, with the statistical software R (The R-project), the following command can be used to fit the multivariate regression model

Fit <- lm(formula=Y~-1+Z1+Z2+Z1*Z2)
where Z1 and Z2 are $n \times 1$ vectors containing the values of $z_{1 ; i}$ and $z_{2 ; i}$; where $n=n_{12}+2 \times n_{1 \mid 2}$. Note that there is no intercept in the model. The $n \times 2$ matrix Y contains the values of $\bar{y}_{1 ; i}$ and $\bar{y}_{2 ; i}$. The estimator $\hat{\boldsymbol{\Sigma}}$ is given by the command estVar (Fit).

## 5. Rotating conditional Poisson sampling (RCPS) design

Because we consider unequal inclusion probabilities, it is necessary to make assumptions about the rotating sampling design used to select the samples. We consider that the sample $s_{1}$ and $s_{2}$ are selected with unequal probabilities using a high entropy Rotating Conditional Poisson Sampling (RCPS) design (Berger 2004) defined by

$$
p_{1}\left(s_{1}\right)=P_{1}\left(s_{1}\right) / P_{1}\left(s_{1} \in \Omega_{1}\right) \delta\left\{s_{1} \in \Omega_{1}\right\}, \quad p_{2}\left(s_{2} \mid s_{1}\right)=P_{2}\left(s_{2} \mid s_{1}\right) / P_{2}\left(s_{2} \in \Omega_{2} \mid s_{1}\right) \delta\left\{s_{2} \in \Omega_{2}\right\}
$$

where $p_{1}\left(s_{1}\right)$ is the probability of selecting the first wave sample $s_{1}$, and $p_{2}\left(s_{2} \mid s_{1}\right)$ is the probability of selecting the second wave sample $s_{2}$ given the first wave sample $s_{1}$. The sets of fixed sizes samples $\Omega_{1}$ and $\Omega_{2}$ are defined by

$$
\Omega_{1}=\left\{s_{1}: v_{1}(s)=n_{1}\right\}, \quad \Omega_{2}=\left\{s_{2}: v_{2}(s)=n_{2} \text { and } v_{12}(s)=n_{12}\right\}
$$

where $v_{1}(s), v_{2}(s)$ and $v_{12}(s)$ denote respectively the number of units in $s_{1}, s_{2}$ and $s_{12}$; and $s=s_{1} \cup s_{2}$. The probabilities $P_{1}\left(s_{1}\right)$ and $P_{2}\left(s_{2} \mid s_{1}\right)$ are respectively the probabilities of the unconditional Poisson sampling designs which have inclusion probabilities given by $\pi_{1 ; i}$ and $\pi_{2 ; i}\left(s_{1}\right)$. The quantity $P_{1}\left(s_{1} \in \Omega_{1}\right)$ and $P_{2}\left(s_{2} \in \Omega_{2} \mid s_{1}\right)$ are respectively the probabilities of selecting a sample $s_{1}$ in $\Omega_{1}$ and a sample $s_{2}$ in $\Omega_{2}$ under unconditional Poisson sampling.

The RCPS design can be implemented by using a rejective approach which consists in selecting two samples with replacement with probabilities of selection proportional to $\pi_{1 ; i} /\left(1-\pi_{1 ; i}\right)$ for the first wave sample and with probabilities of selection proportional to $\pi_{2 ; i}(s) /\left(1-\pi_{2 ; i}(s)\right)$ for the second wave sample (see Hájek, 1964). If the $s_{1} \notin \Omega_{1}$ or $s_{2} \notin \Omega_{2}$, we reject both samples and start again until we select two samples $s_{1}$ and $s_{2}$ such that $s_{1} \in \Omega_{1}$ and $s_{2} \in \Omega_{2}$. Note that when $\pi_{1 ; i}=n_{1} / N$ and $q_{i}=n_{1 \mid 2} /\left(N-n_{1}\right)$, the rejective approach is equivalent in selecting a SRSWOR sample for the first wave and SRSWOR sample of $n_{12}$ units from $s_{1}$ combined with a SRSWOR sample of $n_{1 \mid 2}$ units from $U / s_{1}$ for the second wave.

Because of the rejective nature of the RCPS design, the first-order inclusion probabilities of $s_{1}$ and $s_{2}$ are not exactly equal to $\pi_{1 ; i}$ and $\pi_{2 ; i}\left(s_{1}\right)$ (Hájek, 1964). However, under conditions specified in the following Theorem, the first-order inclusion probabilities can be approximated by $\pi_{1 ; i}$ and $\pi_{2 ; i}\left(s_{1}\right)$.

## Theorem 1

Assuming $n_{1} \rightarrow \infty$ and $\left(N-n_{1}\right) q_{\min }\left(1-q_{\max }\right) \rightarrow \infty$ and $1-\pi_{1 ; \max } \rightarrow 1$, we have that $\tilde{\pi}_{1 ; i} / \pi_{1 ; i} \rightarrow 1$, and $\tilde{\pi}_{2 ; i}\left(s_{1}\right) / \pi_{2 ; i}\left(s_{1}\right) \rightarrow 1$; where $\pi_{1 ; \max }=\max \left\{\pi_{1 ; i}: i \in U\right\}, q_{\max }=\max \left\{q_{i}: i \in U\right\}$ and $q_{\min }=\min \left\{q_{i}: i \in U\right\}$. The quantities $\tilde{\pi}_{1 ; i}$ and $\tilde{\pi}_{2 ; i}\left(s_{1}\right)$ denote the first-order inclusion probabilities of the RCPS design.

The proof is given in Appendix A. In this paper, we suppose that the conditions of Theorem 1 hold. Hence, we consider that the first-order inclusion probabilities of the RCPS design are given by $\pi_{1 ; i}$ and $\pi_{2 ; i}\left(s_{1}\right)$ asymptotically.

There are other rotating sampling designs used in practice such that the rotation groups sampling design and the rotating systematic sampling design. By assuming that the rotation groups are randomly constructed, the rotation group sampling design and the RCPS design are equivalent when $\pi_{1 ; i}=n_{1} / N$ and $q_{i}=n_{1 \mid 2} /\left(N-n_{1}\right)$. The rotating systematic sampling is approximately equivalent to the RCPS design, if we assume that the population is arranged randomly.

## 6. Justification of the regression approach

Note that under the Poisson sampling designs $P_{1}\left(s_{1}\right)$ and $P_{2}\left(s_{2} \mid s_{1}\right)$, the quantities $v_{1}(s), v_{2}(s)$ and $v_{12}(s)$ are random. We suppose that the distribution of $\left(\hat{\tau}_{1}, \hat{\tau}_{2}, v_{1}(s), v_{2}(s), v_{12}(s)\right)^{\prime}$ is normal (see Assumption 1) under Poisson sampling.

Assumption 1: We assume that the distribution of $\mathbf{u}=\left(\hat{\tau}_{1}, \hat{\tau}_{2}, v_{1}(s), v_{2}(s), v_{12}(s)\right)^{\prime}$ is normal under Poisson sampling $P_{1}\left(s_{1}\right)$ and $P_{2}\left(s_{2} \mid s_{1}\right)$; that is, $\mathbf{u} \sim N\left((\boldsymbol{\tau}, \mathbf{n})^{\prime}, \boldsymbol{\Sigma}_{P}\right)$; where $\boldsymbol{\Sigma}_{P}$ is the variance-covariance matrix of $\mathbf{u}$ under Poisson sampling and $(\boldsymbol{\tau}, \mathbf{n})=\left(\tau_{1}, \tau_{2}, n_{1}, n_{2}, n_{12}\right)$. The expression of $\boldsymbol{\Sigma}_{P}$ is given by (7).

Assumption 1 can be justified by the fact that the distributions of the random variables $\hat{\tau}_{1}, \hat{\tau}_{2}, v_{1}(s), v_{2}(s)$ and $v_{12}(s)$ are asymptotically normal under Poisson sampling under regularity conditions proposed by Hájek (1964). Note that the two Poisson samples are not independent because the probability of selecting $s_{2}$ depends on $s_{1}$.

Berger (2004) shows that

$$
\Sigma_{P}=\left(\begin{array}{ll}
\boldsymbol{\Sigma}_{\boldsymbol{\tau} \boldsymbol{\tau}} & \boldsymbol{\Sigma}_{\boldsymbol{\tau} \mathbf{n}}  \tag{7}\\
\boldsymbol{\Sigma}_{\boldsymbol{\tau} \mathbf{n}}^{\prime} & \boldsymbol{\Sigma}_{\mathbf{n n}}
\end{array}\right)
$$

with

$$
\begin{aligned}
& \boldsymbol{\Sigma}_{\boldsymbol{\tau} \boldsymbol{\tau}}=\left(\begin{array}{lll}
\sum_{i \in U} \pi_{1 ; i}\left(1-\pi_{1 ; i}\right)\left(y_{1 ; i} \pi_{1 ; i}^{-1}\right)^{2} & \sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) y_{1 ; i} \pi_{1 ; i}^{-1} y_{2 ; i} \pi_{2 ; i}^{-1} \\
\sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) y_{1 ; i} \pi_{1 ; i}^{-1} y_{2 ; i} \pi_{2 ; i}^{-1} & \sum_{i \in U} \pi_{2 ; i}\left(1-\pi_{2 ; i}\right)\left(y_{2 ; i} \pi_{2 ; i}^{-1}\right)^{2}
\end{array}\right),\left(\begin{array}{ccc}
\sum_{i \in U} \pi_{1 ; i}\left(1-\pi_{1 ; i}\right) & \sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) & \sum_{i \in U} g \pi_{1 ; i}\left(g-\pi_{1 ; i}\right) \\
\sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) & \sum_{i \in U} \pi_{2 ; i}\left(1-\pi_{2 ; i}\right) & \sum_{i \in U} g \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) \\
\boldsymbol{\Sigma}_{\mathbf{n n}} g \pi_{1 ; i}\left(g-\pi_{1 ; i}\right) & \sum_{i \in U} g \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) & \sum_{i \in U} g \pi_{1 ; i}\left(1-g \pi_{1 ; i}\right)
\end{array}\right) \\
& \boldsymbol{\Sigma}_{\boldsymbol{\tau} \mathbf{n}}=\left(\begin{array}{lll}
\sum_{i \in U} \pi_{1 ; i}\left(1-\pi_{1 ; i}\right) y_{1 ; i} \pi_{1 ; i}^{-1} & \sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) y_{1 ; i} \pi_{1 ; i}^{-1} & \sum_{i \in U} g \pi_{1 ; i}\left(g-\pi_{1 ; i}\right) y_{1 ; i} \pi_{1 ; i}^{-1} \\
\sum_{i \in U} \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) y_{2 ; i} \pi_{2 ; i}^{-1} & \sum_{i \in U} \pi_{2 ; i}\left(1-\pi_{2 ; i}\right) y_{2 ; i} \pi_{2 ; i}^{-1} & \sum_{i \in U} g \pi_{1 ; i}\left(g-\pi_{2 ; i}\right) y_{2 ; i} \pi_{2 ; i}^{-1}
\end{array}\right)
\end{aligned}
$$

where the 2 x 2 matrix $\boldsymbol{\Sigma}_{\boldsymbol{\tau} \boldsymbol{\tau}}$ is the variance-covariance matrix of the vector $\hat{\boldsymbol{\tau}}=\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)^{\prime}$, The $3 \times 3$ matrix $\boldsymbol{\Sigma}_{\mathbf{n n}}$ is the variance-covariance matrix of the vector $\mathbf{v}(s)=\left(v_{1}(s), v_{2}(s), v_{12}(s)\right)^{\prime}$, and the $2 \times 3$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{\tau} \mathbf{n}}$ is the covariance matrix between $\hat{\tau}$ and $\mathbf{v}(s)$.

## Theorem 2

Under Assumption 1, the variance-covariance matrix of $\hat{\boldsymbol{\tau}}=\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)^{\prime}$ under RCPS is given by the Schur complement (Harville, 1999, page 100) of $\mathbf{\Sigma}_{\mathbf{n n}}$; that is,

$$
\begin{equation*}
\Sigma=\Sigma_{\tau \tau}-\Sigma_{\tau \mathrm{n}} \Sigma_{\mathrm{nn}}^{-1} \Sigma_{\tau \mathrm{n}}^{\prime} \tag{8}
\end{equation*}
$$

The proof is given in Appendix B.
In order to proof Theorem 3, we need to make the following assumption.
Assumption 2: We assume that the finite population corrections are negligible; that is, for all $i \in U$, we have that $1-\pi_{1 ; i} \approx 1,1-\pi_{2 ; i} \approx 1,1-\pi_{2 ; i} g^{-1} \approx 1$ and $1-\pi_{1 ; i} g \approx 1$.

Note this assumption implicitly assume that $\pi_{2 ; i} g^{-1}$ is small. Note that $g=0$ when there is no overlap between $s_{1}$ and $s_{2}$. In this situation, the regression approach is not suitable. Nevertheless, if $g=0$, we are likely to have two independent samples, and the correlation equal zero in this situation.

## Theorem 3

Under Assumption 2, a design-consistent estimator (under RCPS) of $\boldsymbol{\Sigma}$ defined in (8) is given by $\hat{\boldsymbol{\Sigma}}=\left(\breve{\mathbf{Y}}_{s}-\mathbf{Z}_{s} \hat{\boldsymbol{\beta}}\right)^{\prime}\left(\breve{\mathbf{Y}}_{s}-\mathbf{Z}_{s} \hat{\boldsymbol{\beta}}\right)$ where $\hat{\boldsymbol{\beta}}=\left(\mathbf{Z}_{s}^{\prime} \mathbf{Z}_{s}\right)^{-1} \mathbf{Z}_{s}^{\prime} \breve{\mathbf{Y}}_{s}$ is the ordinary least squares estimate of the parameter

$$
\begin{equation*}
\boldsymbol{\beta}=\left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\right)^{\prime} \tag{9}
\end{equation*}
$$

where $\boldsymbol{\beta}^{(1)}=\left(\beta_{1}^{(1)}, \beta_{2}^{(1)}, \beta_{12}^{(1)}\right)^{\prime}$ and $\boldsymbol{\beta}^{(2)}=\left(\beta_{1}^{(2)}, \beta_{2}^{(2)}, \beta_{12}^{(2)}\right)^{\prime}$ are parameters of the multivariate regression model (3), $\quad \breve{\mathbf{Y}}_{s}=\left(\breve{\mathbf{y}}_{1}, \breve{\mathbf{y}}_{2}\right)^{\prime} \quad$ and $\quad \mathbf{z}_{s}=\left(\mathbf{z}_{1}, \mathbf{z}_{2}, \mathbf{z}_{12}\right) \quad$ with $\quad \breve{\mathbf{y}}_{1}=\left(\breve{y}_{1 ; 1}, \breve{y}_{1 ; 2}, \cdots, \breve{y}_{1 ; n}\right)^{\prime}, \quad \breve{\mathbf{y}}_{2}=\left(\breve{y}_{2 ; 1}, \breve{y}_{2 ; 2}, \cdots, \breve{y}_{2 ; n}\right)^{\prime}$, $\mathbf{z}_{1}=\left(z_{1 ; 1}, z_{1 ; 2}, \cdots, z_{1 ; n}\right)^{\prime}, \mathbf{z}_{2}=\left(z_{2 ; 1}, z_{2 ; 2}, \cdots, z_{2 ; n}\right)^{\prime}, \mathbf{z}_{12}=\left(z_{1 ; 1} z_{2 ; 1}, z_{1 ; 2} z_{2 ; 2}, \cdots, z_{1 ; n} z_{2 ; n}\right)^{\prime}$ and $n=n_{12}+2 \times n_{1 \mid 2}$.

The proof is given in Appendix C. Theorem 3 implies that the ordinary least squares estimates $\hat{\boldsymbol{\Sigma}}$ of the residual covariance matrix obtained by fitting the model (3) is a design-consistent estimator of the variance-covariance matrix
$\boldsymbol{\Sigma}$ (see (8)) of the vector $\hat{\boldsymbol{\tau}}=\left(\hat{\tau}_{1}, \hat{\tau}_{2}\right)^{\prime}$ (see Theorem $2 \& 3$ ). Thus, we have that (6) is a consistent estimator for the correlation.

## 7. Extension to stratified sampling design

The proposed estimator can be easily extended to accommodate stratification. Suppose that we have $H$ strata $U_{1}$, $U_{2}, \ldots, U_{H}$ such that $\bigcup_{h=1}^{H} U_{h}=U$. Let $s_{1 h}$ and $s_{2 h}$ denote respectively the samples of $U_{h}$ for wave 1 and 2 . Let $n_{1 h}, n_{2 h}$ and $n_{12 h}$ be respectively the sample size of $s_{1 h}, s_{2 h}$ and $s_{12 h}=s_{1 h} \cap s_{2 h}$. The following sample sizes are fixed: $n_{1 h}$ and $n_{2 h}$. Hence the following independent variables $z_{1 h ; i}=\delta\left\{i \in s_{1 h}\right\}$ and $z_{h ; i}=\delta\left\{i \in s_{2 h}\right\}$ specify the stratification. If we have a rotation within each stratum, the following sample sizes $n_{12 h}$ are fixed and we need to include the interactions $z_{1 h ; i} \times z_{2 h ; i}$ into the multivariate model which is given by

$$
\begin{equation*}
\binom{\breve{y}_{1 ; i}}{\bar{y}_{2 ; i}}=\sum_{h=1}^{H}\binom{\beta_{1 h}^{(1)} z_{1 h ; i}+\beta_{2 h}^{(1)} z_{1 h ; i}+\beta_{12 h}^{(1)} z_{1 h ; i} \times z_{2 h ; i}}{\beta_{1 h}^{(2)} z_{1 h ; i}+\beta_{2 h}^{(2)} z_{2 h ; i}+\beta_{12 h}^{(2)} z_{1 h ; i} \times z_{2 h ; i}}+\boldsymbol{\varepsilon}_{i}, \tag{10}
\end{equation*}
$$

with the dependent variables defined by (4). The ordinary least squares estimate $\hat{\boldsymbol{\Sigma}}$ of residuals variance-covariance matrix is used to estimate the correlation between $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ (see (6)). If the rotation is across stratum, we would need to use the interaction $z_{1 ; i} \times z_{2 ; i}$ instead of the interaction $z_{1 ; i} \times z_{2 ; i}$; where $z_{1 ; i}$ and $z_{2 ; i}$ are defined by (5).

## 8. A brief simulation study

The dataset contains 2732 units from the British Labour Force Survey. We consider two waves: October-December 2007 and October-December 2008. The population is stratified in 5 strata which are consecutive number of stints. The dataset is duplicated 10 times in order to create a large dataset of size $N=27320$. We consider the following samples sizes of $n_{1}=250, n_{2}=275$ and different fraction $g$ of the common sample $(40 \%, 54 \%, 68 \%, 82 \%$ and $96 \%$ ). The sampling fraction is equal to $n_{1} / N=1 \%$. We assume that the first-order inclusion probabilities are equal $\left(\pi_{1 ; i}=n_{1} / N\right.$ and $q_{i}=n_{1 \mid 2} /\left(N-n_{1}\right)$ ). Table 1, gives the relative biases (\%) and relative mean squared errors (\%) for the variance of $\hat{\Delta}=\hat{\tau}_{2}-\hat{\tau}_{1}$, where $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ denote the Horvitz-Thompson estimator of total incomes for wave 1 and 2.

|  |  |  | $g=$ | $g=$ |  |  |
| :--- | :--- | :---: | :---: | :---: | :---: | :---: |
|  |  | $g=40 \%$ | $54 \%$ | $68 \%$ | $g=82 \%$ | $g=96 \%$ |
| Relative | Regression Approach | 0.13 | 0.06 | 0.00 | 0.03 | -0.01 |
| Bias (\%) | Tam (1984) | 0.13 | 0.04 | -0.02 | 0.02 | -0.03 |
|  | Qualité \& Tillé (2008) | 0.12 | 0.05 | -0.02 | 0.01 | -0.03 |
|  | Wood (2008) | 0.12 | 0.04 | -0.02 | 0.01 | -0.03 |
| Relative | Regression Approach | 0.22 | 0.24 | 0.27 | 0.29 | 0.33 |
| Mean | Tam (1984) | 0.21 | 0.22 | 0.23 | 0.25 | 0.31 |
| Squared | Qualité \& Tillé (2008) | 0.22 | 0.24 | 0.27 | 0.28 | 0.34 |
| error (\%) | Wood (2008) | 0.23 | 0.25 | 0.28 | 0.30 | 0.35 |

Table 1: Empirical Relative biases and Relative mean squared errors for the measure of change
In Table 1, the proposed regression approach is compared with its competitor estimators proposed by Tam (1984), Qualité \& Tillé (2008) and Wood (2008). Note that the proposed estimator is as precise as its competitors. The advantage of the proposed estimator is its generality, simplicity and flexibility.

## 9. Extension to two stage sampling design

Suppose that we have stratified samples of Primary Sampling Units (PSUs), and that the rotation consists in rotating PSUs rather than Secondary Sampling Units (SSUs). We suggest using an ultimate cluster strategy to estimate the correlation. Let $z_{1 h ; i}$ and $z_{2 h ; i}$ be the variables that specify the stratification at PSU level; that is, $z_{\ell h ; i}=1$ if the $i$ ${ }^{\text {th }}$ PSU is selected in stratum $U_{h}$ at wave $\ell=1,2$. The estimators $\hat{\tau}_{1}$ and $\hat{\tau}_{2}$ are now given by $\hat{\tau}_{1}=\Sigma_{i \in s_{1}} \hat{\tau}_{1 ; i} \pi_{1 ; i}^{-1}$ and $\hat{\tau}_{2}=\Sigma_{i \in s_{2}} \hat{\tau}_{2 ; i} \pi_{2 ; i}^{-1}$; where $s_{1}$ and $s_{2}$ denote the first and the second wave sample of PSUs. The quantities $\pi_{1 ; i}$ and $\pi_{2 ; i}$ are the first-order inclusion probabilities of the $i^{\text {th }}$ PSU for the first and the second wave. The quantities $\hat{\tau}_{1 ; i}$ and $\hat{\tau}_{2 ; i}$ are the Horvitz-Thompson estimators of $\tau_{1 ; i}$ and $\tau_{2 ; i}$ with respect to the second stage; where $\tau_{1 ; i}$ and $\tau_{2 ; i}$ denote the PSU totals of the first and the second wave.

If the PSUs rotate within each stratum, the multivariate model (10) could be used with the dependent variables $\bar{y}_{1 ; i}=\hat{\tau}_{1 ; i} / \pi_{1 ; i} \delta\left\{i \in s_{1}\right\}$ and $\bar{y}_{2 ; i}=\hat{\tau}_{2 ; i} / \pi_{2 ; i} \delta\left\{i \in s_{2}\right\}$. The independent variables are defined at the beginning of this §.

## 10. Extension to more complex measure of change

Consider that the parameter of interest is a function of $P$ totals (e.g. ratio, index, ...); that is, $\hat{\theta}_{1}=g\left(\hat{\tau}_{11}, \hat{\tau}_{12}, \cdots, \hat{\tau}_{1 P}\right)$ and $\hat{\theta}_{2}=g\left(\hat{\tau}_{21}, \hat{\tau}_{22}, \cdots, \hat{\tau}_{2 P}\right) ;$ where $\hat{\tau}_{\ell p}=\sum_{i \in s_{\ell}} y_{\ell p ; i} \pi_{\ell ; i}^{-1} \quad(\ell=1,2$ and $p=1,2, \cdots, P)$. Thus, the change is a function of these totals; that is, $\hat{\theta}_{2}-\hat{\theta}_{1}=f(\hat{\boldsymbol{\tau}})$ where $\hat{\boldsymbol{\tau}}=\left(\hat{\tau}_{11}, \hat{\tau}_{12}, \cdots, \hat{\tau}_{1 P}, \hat{\tau}_{21}, \hat{\tau}_{22}, \cdots, \hat{\tau}_{2 P}\right)^{\prime}$. Using Taylor linearization, we have that

$$
\left(\hat{\theta}_{2}-\hat{\theta}_{1}\right)-\left(\theta_{2}-\theta_{1}\right) \approx \operatorname{grad}(\boldsymbol{\tau})^{\prime}(\hat{\tau}-\tau),
$$

where $\operatorname{grad}(\hat{\boldsymbol{\tau}})^{T}$ is the gradient of $f(\hat{\boldsymbol{\tau}})$ at $\boldsymbol{\tau}$. Therefore, and estimator for the MSE is

$$
\operatorname{MSE}\left(\hat{\theta}_{2}-\hat{\theta}_{1}\right)=\operatorname{grad}(\hat{\boldsymbol{\tau}})^{\prime} \operatorname{var}(\hat{\boldsymbol{\tau}}) \operatorname{grad}(\hat{\boldsymbol{\tau}}) .
$$

The matrix var $(\hat{\boldsymbol{\tau}})$ can be estimated using the multivariate regression approach. We have now $2 P$ dependent variables $y_{\ell p ; i}=y_{\ell p ; i} \pi_{\ell ; i}^{-1} \delta\left\{i \in s_{\ell}\right\} \quad(\ell=1,2, p=1,2, \cdots, P$ and $i \in s)$. Depending on the design, the model (3) or (10) can be used to estimate the residual variance-covariance matrix. Let $\hat{\boldsymbol{\Sigma}}$ be the ordinary least squares estimate of the $2 P \times 2 P$ residuals variance-covariance matrix. The matrix $\operatorname{var}(\hat{\boldsymbol{\tau}})$ is given by vâr $(\hat{\boldsymbol{\tau}})=\mathbf{D} \dot{\mathbf{\Sigma}} \mathbf{D}$ with $\mathbf{D}$ the $2 P$ diagonal matrix defined by

$$
\mathbf{D}=\operatorname{diag}\left\{\sqrt{\operatorname{vâr}\left(\hat{\tau}_{11}\right) \hat{\Sigma}_{11}^{-1}}, \sqrt{\operatorname{vâ}\left(\hat{\tau}_{12}\right) \hat{\Sigma}_{22}^{-1}}, \cdots, \sqrt{\operatorname{vâr}\left(\hat{\tau}_{1 P}\right) \hat{\Sigma}_{P P}^{-1}}, \sqrt{\operatorname{vâr}\left(\hat{\tau}_{21}\right) \hat{\Sigma}_{(P+1)(P+1)}^{-1}}, \cdots, \sqrt{\operatorname{var}\left(\hat{\tau}_{2 P}\right) \hat{\Sigma}_{(2 P)(2 P)}^{-1}}\right\} .
$$

Note that $\operatorname{MSE}\left(\hat{\theta}_{2}-\hat{\theta}_{1}\right)>0$ because var $(\hat{\boldsymbol{\tau}})$ is non-negative definite.

Acknowledgment: This work was supported by grant RES-000-22-3045 of the Economic and Social Research Council (UK).

## Appendix A (Proof of Theorem 1)

Applying Hájek (1964, page 1510) result, we have that $\tilde{\pi}_{1 ; i} / \pi_{1 ; i} \rightarrow 1$ when $d_{1}=\Sigma_{i \in U} \pi_{1 ; i}\left(1-\pi_{1 ; i}\right) \rightarrow \infty$. As $1-\pi_{1 ; \text { max }} \approx 1$, we have that $d_{1}>\left(1-\pi_{1 ; \text { max }}\right) n_{1} \approx n_{1}$. Thus $d_{1} \rightarrow \infty$ holds when $n_{1} \rightarrow \infty$. Applying Hájek (1964, page 1510 ) result, we have that $\tilde{\pi}_{2 ; i}\left(s_{1}\right) / \pi_{2 ; i}\left(s_{1}\right) \rightarrow 1$ when $d_{2}\left(s_{1}\right)=\Sigma_{i \in U} \pi_{2 ; i}\left(s_{1}\right)\left(1-\pi_{2 ; i}\left(s_{1}\right)\right) \rightarrow \infty$. It can be shown that $d_{2}\left(s_{1}\right)>n_{12}(1-g)+\left(N-n_{1}\right) q_{\min }\left(1-q_{\max }\right)>\left(N-n_{1}\right) q_{\min }\left(1-q_{\max }\right)$. Hence, $d_{2}\left(s_{1}\right) \rightarrow \infty$ when $\left(N-n_{1}\right) q_{\min }\left(1-q_{\max }\right) \rightarrow \infty$. The Theorem follows.

## Appendix B (Proof of Theorem 2)

Using (7), we have that

$$
(\hat{\boldsymbol{\tau}}, \mathbf{v}(s)) \sim N\left(\binom{\boldsymbol{\tau}}{\mathbf{n}},\left(\begin{array}{cc}
\boldsymbol{\Sigma}_{\boldsymbol{\tau}} & \boldsymbol{\Sigma}_{\mathrm{\tau n}}  \tag{11}\\
\boldsymbol{\Sigma}_{\boldsymbol{\tau}}^{\prime} & \boldsymbol{\Sigma}_{\mathrm{nn}}
\end{array}\right)\right)
$$

where $\boldsymbol{\tau}=\left(\tau_{1}, \tau_{2}\right)^{\prime}$ and $\mathbf{v}(s)=\left(v_{1}(s), v_{2}(s), v_{12}(s)\right)^{\prime}$. Under the RCPS design, the sizes $\left(v_{1}(s), v_{2}(s), v_{12}(s)\right)$ are fixed, as $s_{1} \in \Omega_{1}$ and $s_{2} \in \Omega_{2}$. This implies that $v_{1}(s)=n_{1}, v_{2}(s)=n_{2}$ and $v_{12}(s)=n_{12}$. Hence $\mathbf{v}(s)=\mathbf{n}=\left(n_{1}, n_{2}, n_{12}\right)$. Thus the sampling distribution of $\hat{\tau}$ under RCPS, is the following bivariate conditional distribution $(\hat{\boldsymbol{\tau}} \mid \mathbf{v}(s)=\mathbf{n}) \sim N(\boldsymbol{\tau}, \boldsymbol{\Sigma})$; where $\boldsymbol{\Sigma}$ is defined by (8). Hence $\boldsymbol{\Sigma}$ is the variance-covariance matrix of the vector $\hat{\tau}$ under RCPS. The Theorem follows.

## Appendix C (Proof of Theorem 3)

Berger (2004) showed that the design-consistent Horvitz-Thompson estimators (under RCPS) of the matrices $\boldsymbol{\Sigma}_{\boldsymbol{\tau} \boldsymbol{\tau}}$, $\boldsymbol{\Sigma}_{\mathbf{n n}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\tau} \mathbf{n}}$ that are given by

$$
\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\tau} \tau}=\left(\begin{array}{ll}
\sum_{i \in s}\left(1-\pi_{1 ; i} \breve{y}_{1 ; i}^{2}\right. & \sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) \breve{y}_{1 ; i} \breve{y}_{2 ; i} \\
\sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) \breve{y}_{1 ; i} \breve{y}_{2 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i} \breve{y}_{2 ; i}^{2}\right.
\end{array}\right),
$$

$$
\begin{aligned}
& \hat{\mathbf{\Sigma}}_{\mathbf{n n}}=\left(\begin{array}{lll}
\sum_{i \in s}\left(1-\pi_{1 ; i}\right) z_{1 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) z_{1 ; i} z_{2 ; i} & \sum_{i \in s}\left(1-\pi_{1 ; i}\right) z_{1 ; i} z_{2 ; i} \\
\sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) & \sum_{i \in S}\left(1-\pi_{2 ; i}\right) z_{2 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i}\right) z_{1 ; i} z_{2 ; i} \\
\sum_{i \in s}\left(1-\pi_{1 ; i}\right) z_{1 ; i} z_{2 ; i} & \sum_{i \in S}\left(1-\pi_{2 ; i}\right) z_{1 ; i} z_{2 ; i} & \sum_{i \in s}\left(1-\pi_{1 ; i} g\right) z_{1 ; i} z_{2 ; i}
\end{array}\right), \\
& \hat{\mathbf{\Sigma}}_{\boldsymbol{\tau n}}=\left(\begin{array}{lll}
\sum_{i \in s}\left(1-\pi_{1 ; i}\right) \bar{y}_{1 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) \check{y}_{1 ; i} z_{2 ; i} & \sum_{i \in s}\left(1-\pi_{1 ; i}\right)_{1 ; i ;} z_{2 ; i} \\
\sum_{i \in s}\left(1-\pi_{2 ; i} g^{-1}\right) \breve{y}_{2 ; i} z_{1 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i}\right) \breve{y}_{2 ; i} & \sum_{i \in s}\left(1-\pi_{2 ; i}\right) \breve{y}_{2 ; i} z_{1 ; i}
\end{array}\right) ;
\end{aligned}
$$

where $s=s_{1} \cup s_{2}$. Thus a design-consistent Horvitz-Thompson estimator of $\boldsymbol{\Sigma}$ is given by

$$
\begin{equation*}
\hat{\Sigma}=\hat{\Sigma}_{\tau \tau}-\hat{\Sigma}_{\tau n} \hat{\Sigma}_{\mathrm{nn}}^{-1} \hat{\Sigma}_{\tau \mathrm{n}}^{\prime} \tag{12}
\end{equation*}
$$

Assumption 2 implies

$$
\begin{equation*}
\hat{\Sigma}_{\tau \tau} \approx \breve{\mathbf{Y}}_{s}^{\prime} \breve{\mathbf{Y}}_{s}, \quad \hat{\Sigma}_{\mathrm{nn}} \approx \mathbf{Z}_{s}^{\prime} \mathbf{Z}_{s} \quad \text { and } \quad \hat{\mathbf{\Sigma}}_{\tau \mathrm{n}} \approx \breve{\mathbf{Y}}_{s}^{\prime} \mathbf{Z}_{s} \tag{13}
\end{equation*}
$$

Finally by substituting (13) into (12), we obtain

$$
\hat{\boldsymbol{\Sigma}} \approx \breve{\mathbf{Y}}_{s}^{\prime} \breve{\mathbf{Y}}_{s}-\breve{\mathbf{Y}}_{s}^{\prime} \mathbf{Z}_{s}\left(\mathbf{Z}_{s}^{\prime} \mathbf{Z}_{s}\right)^{-1} \mathbf{Z}_{s}^{\prime} \breve{\mathbf{Y}}_{s}=\left(\breve{\mathbf{Y}}_{s}-\mathbf{Z}_{s} \hat{\boldsymbol{\beta}}^{\prime}\right)^{\prime}\left(\breve{\mathbf{Y}}_{s}-\mathbf{Z}_{s} \hat{\boldsymbol{\beta}}\right) ;
$$

where $\hat{\boldsymbol{\beta}}=\left(\mathbf{Z}_{s}^{\prime} \mathbf{Z}_{s}\right)^{-1} \mathbf{Z}_{s}^{\prime} \breve{\mathbf{Y}}_{s}$ is the ordinary least squares estimate of the parameter $\boldsymbol{\beta}$ defined by (9). The Theorem follows.

## References

Berger, Y.G. (2004). Variance Estimation for Measures of Change in Probability Sampling, Canadian Journal of Statistics, 32, pp. 451-467.

Hájek, J. (1964). Asymptotic Theory of Rejective Sampling With Varying Probabilities from a Finite Population, Ann. Math. Statist., 35, pp. 1491-1523.

Harville, D.A. (1999). Matrix Algebra from a Statistician's perspective: Springer.
Holmes, D.J. and Skinner, C.J. (2000). Variance Estimation for Labour Force Survey Estimates of Level and Change, Government Statistical Service Methodology Series, 21, London, England.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe, Journal of the American Statistical Association, 47, 663-685.

Qualité, L. and Tillé, Y. (2008). Variance Estimation of Changes in Repeated Surveys and its Application to the Swiss Survey of Value Added, Survey Methodology, 34, pp. 173-181.

Kish, L. (1965). Survey Sampling, New York:Wiley.
Smith, P., Pont, M. and Jones, T. (2003). Developments in Business Survey Methodology in the Office for National Statistics, 1994-2000, Journal of the Royal Statistical Society, Ser. D, 52, pp. 1-30.

Tam, S.M. (1984). On Covariances from Overlapping Samples, The American Statistician, 38, pp. 288-289.

The R-project. http://www.r-project.org

Wood, J. (2008). On the Covariance Between Related Horvitz-Thompson Estimators, Journal of Official Statistics, 24, pp. 53-78.

# Use of auxiliary sources in weighting: The case of integration of social survey data 

Takis Merkouris ${ }^{1}$


#### Abstract

Various forms of integration of social survey data involve some combination of information from different survey sources that are used as auxiliary to each other. A compilation of examples of integration of social surveys is presented, highlighting in each case the motivation, the sampling setting and the type of auxiliary information. Incorporation of auxiliary information into the weighting structure of the integrated survey sources may give substantially improved estimates of interest. Micro-integration through an adjustment of the survey weights can be accomplished by suitable calibration schemes, which are equivalent to regression procedures based on the principle of best linear unbiased estimation.


Key Words: Auxiliary Information; Combining Information; Composite Estimation; Calibration; Best Linear Unbiased Estimation; Generalized Regression Estimation.

## 1. Introduction

Integration of social survey data may be generally defined as some combination of information from various social survey sources, which may serve as auxiliary to each other. Such auxiliary survey sources may be other social (or household) surveys, sub-samples of a single survey, or past data from a repeated survey. The auxiliary information used in data integration is meant to be sample data on target or auxiliary variables in the sources involved. For such auxiliary variables, no information exists beyond what is provided by the sample. Auxiliary information may be available in the form of sample micro-data or in the form of sample estimates.

Possibilities for integration of social survey data are on the increase in contemporary survey practice. This is because there is a growing interest in integrating field work and survey functions of various survey sources for diverse reasons, such as, reduced cost and efficient survey operations, reduced response burden and improved data quality, harmonized survey content and data consistency, improved estimation and analysis. We may classify the various forms of such integration as planned integration with pre-specified objectives, and as unplanned integration of existing sources exploiting the possibility to achieve certain of the aforementioned objectives. Evidently, planned integration usually serves more purposes than unplanned integration. A hybrid of these two types may also be encountered in survey practice.

In the context of this paper, the primary objective in the integration of survey sources is the improvement of estimation of population totals. This can be achieved by pooling data on variables that are common between the various sources, and by exploiting the correlation between variables surveyed in these sources. Such an approach to efficient estimation is pertinent to all the traditional and new schemes of data integration presented in this paper.

Incorporation of auxiliary information into the weighting structure of the integrated survey data may give substantially improved estimates of interest. Micro-integration of data through an adjustment of the survey weights can be accomplished by suitable calibration schemes, which are equivalent to regression procedures based on the principle of best linear unbiased estimation.

Underlying the integration of auxiliary sources for the improvement of estimation, is the assumption that combined data on common variables are comparable, or, as usually called, harmonized. In planned integration, harmonization

[^72]of concepts, questions and survey functions is one of the key objectives, and thus comparability of data is inbuilt by design, whereas in unplanned integration comparability may not be always ensured. It is to be noted that harmonization of data is increasingly happening in national statistical agencies.

In Section 2, a compilation of examples of integration of social surveys is presented, highlighting in each case the motivation, the sampling setting and the type of auxiliary information. In Section 3, a general weighting and estimation methodology based on generalized regression is demonstrated for two of these examples. Concluding remarks are made in Section 4.

## 2. Paradigms of Social-Survey Data Integration

A broad array of paradigms of data integration is presented in this section, arranged according to whether integration is planned or not, and according to the presence or absence of sample independence.

## Unplanned integration, independent samples

1. Two or more separate surveys of the same population with some variables in common. These common variables may be primary target variables for these surveys, or auxiliary variables commonly included in social surveys, such as tenure, household composition or ethnicity. Examples include various income surveys or consumer expenditure surveys conducted by many national statistical agencies, such as the Survey of Labour and Income Dynamics, Survey of Household Spending and Survey of Financial Security of Statistics Canada, and the Diary and Interview components of the Consumer Expenditure Survey of the US Bureau of Labour Statistics. The auxiliary sources are the different surveys, and the auxiliary information is the data on the common variables. The objectives of data integration may include improved estimations for all variables in all surveys - primarily and more profitably for the common target variables - and consistency of the estimated totals of the common target variables.
2. Two surveys of the same population with some auxiliary variables in common. One of the surveys is a large-scale household survey, typically the Labour Force Survey, and the other survey, called a "supplement", uses a sub-sample of the first to collect data on a different set of target variables. Such supplement surveys are conducted regularly or on ad-hoc basis by several national statistical agencies. In this case the auxiliary source is the part of the large-scale sample not used by the supplement survey. Since the two complementary samples are typically independent, this case can be viewed as a special form of double sampling (Merkouris, 2008). The auxiliary information is the data on the common auxiliary variables, and the objective of data integration is improved estimation for the variables of the supplement, through borrowing strength from the rest of the large-scale sample.

## Planned integration, independent samples

3. A single survey consisting of a core and a booster sample, where the booster sample has either the same or some of the target variables of the core sample. The auxiliary source is the booster sample, and the auxiliary information is the data on the target variables that are common between the two samples. The objective of using the booster sample is to improve the efficiency of small domain estimation for the common target variables. Two important cases involve:
(a). An independent booster sample from the same entire frame, for targeting small domains cutting across strata. This sampling scheme may also be effective in surveying rare populations (e.g., a population of rare crops, or a population of aboriginal people).
(b). An independent booster sample from the frames of strata containing or coinciding with the domains of interest (typically small areas).
4.A single survey consisting of a core and a supplementary sample, with the latter collecting data on the same auxiliary variables as in the core sample. An example of such a survey is the Canadian Survey of Employment, Payroll and Hours (Hidiroglou, 2001). This sampling scheme, known as "non-nested" double sampling, differs from that of paradigm 3 in that the common variables are auxiliary. Here the auxiliary source is the supplementary sample, and the auxiliary information is the data on the common auxiliary variables. The objective of integration of the core and the supplementary samples is improved estimation for the target variables of the core sample.
4. A single survey in which a sub-sample is used to collect information on a set of additional target variables. Collection of all data is done simultaneously from the entire sample, but the module with the additional questions is administered only to the sub-sample in order to reduce response burden and cost. The motivating example is the multiple-panel Labour Force Survey in the European Union, which uses a sub-sample to collect information on additional variables. This scheme too might be regarded as a special form of double sampling (Merkouris, 2008). Here the auxiliary source is the complement of the sub-sample, and the auxiliary information is the data on selected common target variables. The objective of data integration is improved estimation for the additional variables.

## Planned integration, dependent samples

6. A single survey with double (two-phase) sampling design. In this case the auxiliary source is the complement of the second-phase sample, and the auxiliary information is the data on the auxiliary variables. The objective of integrating the second-phase sample with its complement is improved estimation for the variables surveyed in the second-phase sample. This scheme differs from the non-nested double sampling (paradigm 4) in that here the second-phase sample is a sub-sample of the first-phase sample, and thus the second-phase sample and its complement are dependent.
7. A single survey with matrix sampling design, which sometimes is referred to as split questionnaire. In the context of data integration we consider three principal variants of matrix sampling designs:
(a). A large questionnaire is divided into subsets of questions, and these subsets are then administered to different random sub-samples of an initial sample. An additional core set of questions is administered to all sub-samples. Each sub-sample is an auxiliary source for the other sub-samples, and the auxiliary information is the data on the core set of questions. The objective in integrating the data from the various sub-samples is the improvement of the estimation for all variables by using the core items to borrow strength across the sub-questionnaires.
(b). Different sub-samples receive different (non-overlapping) sub-questionnaires, and an additional sub-sample receives the full questionnaire. This may be viewed as a generalization of double sampling.
All sub-samples are auxiliary sources for the other sub-samples, and the auxiliary information is the data on all items. The objective of data integration is as in case (a).
(c). This variant of design (b), in which the core set of questions is administered to all sub-samples, embodies all features of designs (a) and (b). Auxiliary sources and auxiliary information are as in (b), and the objective of data integration is more improved estimation for all variables.
8. A survey with non-nested matrix sampling design. In analogy with non-nested double sampling, in this form of matrix sampling the various sub-questionnaires are administered to independent samples. It is to be noted that the advantages of matrix sampling are not always contingent on using sub-samples (necessarily dependent) of an initial sample. On the contrary, it may be more practical in certain situations to use independent samples. The three variants of the ordinary matrix sampling can be used also in this setting. Recent uses of matrix sampling in various statistical agencies (e.g., British Office of National Statistics, Australian Bureau of Statistics), marking current trends and offering an outlook for future survey practice, involve the integration of a number of existing independent household surveys for the benefit of streamlined survey operations, harmonized survey content and consistency of outputs. In this non-ordinary matrix sampling setting, the distinct surveys may use sub-samples of a large master sample or independent (and non-overlapping) samples from the same population.
9. Three other special cases of social-survey data integration may be considered as fitting into the present setting. These cases differ from the preceding ones in that the auxiliary survey sources do not represent exactly the same population, or in that the integrated data are measurements of the same concept but at different reference periods:
(a). A survey with multiple-frame design. The integrated auxiliary sources are the parts of the samples from the different frames that cover the same population.
(b). A repeated survey with partially overlapping rotating panels. This is the typical sampling design of the Labour Force Survey in many countries. The auxiliary sources are the rotating panels, and the integrated data in this case is past (from previous survey period) data and present data on the same survey items.
(c). A multiple-panel longitudinal survey used for cross-sectional estimation. The integrated auxiliary sources are the parts of the panels that represent the population at the same reference period.

## 3. Weighting and Estimation in Social Survey Data Integration

In this section, a general estimation method, applicable to paradigms of data integration 1 to 8 described above, will be demonstrated for the particular cases 1 and 8(b).

### 3.1 Weighting in paradigm 1

Paradigm 1 is quite generic, covering, with the appropriate adaptations, the next four. The way to improve the efficiency of estimation is to combine in some manner the information on common variables. For the simplest case involving two surveys with some variables in common, methodologies for such combination have been described in Zieschang (1990), Renssen and Nieuwenbroek (1997), Wu (2004) and Merkouris (2004, 2010a). These methodologies include empirical likelihood estimation, by Wu (2004), and generalized regression estimation and calibration by the other authors. Variations of generalized regression and calibration methods for paradigms 2, 3, and 5 are described in Merkouris (2008, 2010a), and for paradigm 4 in Hidiroglou (2001) and Merkouris (2008a).

We assume that we have samples $S_{1}$ and $S_{2}$ of sizes $n_{1}$ and $n_{2}$ respectively from two separate surveys of the same population, with vectors of target variables $y_{1}$ and $y_{2}$ associated with $S_{1}$ and $S_{2}$, and with a vector of common variables $z$. We denote by $w_{i}$ the vector of design weights for sample $S_{i}, \mathrm{i}=1,2$, and by $Y_{i}$ and $Z_{i}$ the sample matrices of $y_{i}$ and $z$ - the subscripts indicating the sample. We also denote by $\hat{Y}_{i}\left(=Y_{i}^{\prime} w_{i}\right)$ the HorvitzThompson (HT) estimate of the population total $T_{y_{i}}$ of $y_{i}$, and by $\hat{Z}_{i}$ the HT estimate of $T_{z}$ of $z$, based on sample $S_{i}$. It is assumed that the data on $z$ from the two samples are harmonized, so that the two estimates $\hat{Z}_{1}$ and $\hat{Z}_{2}$ have the same expectation $T_{z}$. To produce a more efficient estimator of $T_{z}$, data on $z$ from the two samples are combined by using a special calibration scheme for the combined sample $S=S_{1} \cup S_{2}$, whereby the two calibrated estimates of $T_{z}$ are equated. Thus, a vector of calibrated weights $c$ is constructed to satisfy the constraint $Z_{1}^{\prime} c_{1}=Z_{2}^{\prime} c_{2}$ while minimizing the generalized least-squares distance $(c-w)^{\prime} \Lambda^{-1}(c-w)$, where $w$ is the vector of design weights for $S$ and $\Lambda$ is the diagonal matrix whose entries are the design weights of $S$. This vector $c$ is given by

$$
\begin{equation*}
c=w+\Lambda Z\left(Z^{\prime} \Lambda Z\right)^{-1}\left(0-Z^{\prime} w\right) \tag{1}
\end{equation*}
$$

where $Z=\left(Z_{1},-Z_{2}\right)^{\prime}$ is the matrix producing the alignment of the two estimates of $T_{z}$. Then, using the calibrated weights of samples $S_{i}$, we can obtain two calibration estimators $\hat{Z}_{i}^{c}=Z_{i}^{\prime} c_{i}$ of $T_{z}$, which are equal by construction. Using these sets of weights, we can also construct the calibration estimator $\hat{Y}_{i}^{c}=Y_{i}^{\prime} c_{i}$ of $T_{y_{i}}$. In view of (1), these composite calibration estimators take the form of generalized regression (GREG) estimators

$$
\begin{align*}
& \hat{Z}_{1}^{C}=\hat{Z}_{1}+\left(I-\hat{B}_{z}\right)\left(\hat{Z}_{2}-\hat{Z}_{1}\right), \\
& \hat{Z}_{2}^{C}=\hat{Z}_{2}-\hat{B}_{z}\left(\hat{Z}_{2}-\hat{Z}_{1}\right) \tag{2}
\end{align*}
$$

where $\hat{B}_{z}=Z_{2}^{\prime} \Lambda_{2} Z_{2}\left(Z_{1}^{\prime} \Lambda_{1} Z_{1}+Z_{2}^{\prime} \Lambda_{2} Z_{2}\right)^{-1}$, and

$$
\begin{align*}
& \hat{Y}_{1}^{C}=\hat{Y}_{1}+\hat{B}_{y 1}\left(\hat{Z}_{2}-\hat{Z}_{1}\right) \\
& \hat{Y}_{2}^{C}=\hat{Y}_{2}-\hat{B}_{y 2}\left(\hat{Z}_{2}-\hat{Z}_{1}\right), \tag{3}
\end{align*}
$$

where $\hat{B}_{y i}=Y_{i}^{\prime} \Lambda_{i} Z_{i}\left(Z_{1}^{\prime} \Lambda_{1} Z_{1}+Z_{2}^{\prime} \Lambda_{2} Z_{2}\right)^{-1}$. Especially for the estimators in (3), we can write

$$
\begin{equation*}
\hat{Z}_{1}^{C}=\hat{Z}_{2}^{C}=\hat{B}_{z} \hat{Z}_{1}+\left(I-\hat{B}_{z}\right) \hat{Z}_{2}, \tag{4}
\end{equation*}
$$

which shows that the composite estimator of $T_{z}$ is obtained as a linear combination of the two elementary estimators $\hat{Z}_{1}$ and $\hat{Z}_{2}$. Although the composite estimators of $T_{y_{i}}$ and $T_{z}$ are constructed using weighted data of only one of the samples, they are more efficient than the estimators based on a single survey because each sample's calibrated weights incorporate auxiliary information on $z$ from the other sample.

Data on vectors of auxiliary variables $x_{1}$ and $x_{2}$ with known population totals $T_{x_{1}}$ and $T_{x_{2}}$ may also be collected in the samples $S_{1}$ and $S_{2}$, respectively. These are socio-demographic variables typically used in ordinary calibration in household surveys. Then, the composite GREG estimators (2) and (3) may include the ordinary GREG terms $\hat{B}_{x_{1}}\left(T_{x_{1}}-\hat{X}_{1}\right)+\hat{B}_{x_{2}}\left(T_{x_{2}}-\hat{X}_{2}\right)$. These extended GREG estimators have improved efficiency, as they incorporate additional information, and are generated by a calibration procedure that includes the additional two constraints $X_{i}^{C}=T_{x_{i}}, \mathrm{i}=1,2$, giving a vector of calibrated weights $c$ as in (1), but with augmented design matrix $(\mathcal{Z}, \mathrm{X})$, where X is the block-diagonal matrix $\mathrm{X}=\operatorname{diag}\left(X_{i}\right)$. These composite GREG estimators can be written in the form

$$
\begin{equation*}
\hat{Z}^{C}=\hat{B}_{z} \hat{Z}_{1}^{G R}+\left(I-\hat{B}_{z}\right) \hat{Z}_{2}^{G R} \tag{5}
\end{equation*}
$$

and

$$
\begin{align*}
& \hat{Y}_{1}^{C}=\hat{Y}_{1}^{G R}+\hat{B}_{y 1}\left(\hat{Z}_{2}^{G R}-\hat{Z}_{1}^{G R}\right),  \tag{6}\\
& \hat{Y}_{2}^{C}=\hat{Y}_{2}^{G R}-\hat{B}_{y 2}\left(\hat{Z}_{2}^{G R}-\hat{Z}_{1}^{G R}\right),
\end{align*}
$$

where $\hat{Z}_{i}^{G R}$ and $\hat{Y}_{i}^{G R}$ are GREG estimators incorporating only the auxiliary information on $x_{i}$, and the matrix $\Lambda_{i}$ in $\hat{B}_{z}$ and $\hat{B}_{y_{i}}$ is now replaced by $\Lambda_{i}\left(I-P_{X_{i}}\right)$ with $P_{X_{i}}=X_{i}\left(X_{i}^{\prime} \Lambda_{i} X_{i}\right)^{-1} X_{i}^{\prime} \Lambda_{i}$; see Merkouris (2004).

The special GREG estimators (5) and (6) may be viewed as approximate forms of best linear unbiased estimators (BLUE), i.e., minimum-variance linear unbiased combinations of the four estimators $\hat{Y}_{1}^{G R}, \hat{Y}_{2}^{G R}, \hat{Z}_{1}^{G R}$ and $\hat{Z}_{2}^{G R}$. In matrix form the BLUE are given by

$$
\begin{equation*}
\left(\hat{Y}_{1}^{B}, \hat{Y}_{2}^{B}, \hat{Z}^{B}\right)^{\prime}=P\left(\hat{Y}_{1}^{G R}, \hat{Y}_{2}^{G R}, \hat{Z}_{1}^{G R}, \hat{Z}_{2}^{G R}\right)^{\prime}, \tag{7}
\end{equation*}
$$

where $P=\left(W^{\prime} V^{-1} W\right)^{-1} W^{\prime} V^{-1}$, the matrix $W$ expressing the condition of unbiasedness is such that $P W=I$, and $V$ is the variance-covariance matrix of ( $\hat{Y}_{1}^{G R}, \hat{Y}_{2}^{G R}, \hat{Z}_{1}^{G R}, \hat{Z}_{2}^{G R}$ )'. This is an adaptation of the form of BLUE found in the literature on composite estimation in a different context of survey sampling; see, for example, Wolter (1979) and Jones (1980). It can be shown that the BLUE $\hat{Y}_{1}^{B}, \hat{Y}_{2}^{B}, \hat{Z}^{B}$ have the regression form (6) and (5), and that

$$
P=\left(\begin{array}{ccc}
I & 0 & B_{y_{1}}  \tag{8}\\
0 & I & B_{y_{2}} \\
0 & 0 & B_{z}
\end{array}\right),
$$

where the matrices $B_{y i}=\operatorname{Cov}\left(\hat{Y}_{i}^{G R}, \hat{Z}_{i}^{G R}\right)\left[\mathrm{V}\left(\hat{Z}_{1}^{G R}\right)+\mathrm{V}\left(\hat{Z}_{2}^{G R}\right)\right]^{-1}$ and $B_{z}=\mathrm{V}\left(\hat{Z}_{2}^{G R}\right)\left[\mathrm{V}\left(\hat{Z}_{1}^{G R}\right)+\mathrm{V}\left(\hat{Z}_{2}^{G R}\right)\right]^{-1}$ minimize the variances of $\hat{Y}_{i}^{G R} \pm B_{y i}\left(\hat{Z}_{2}^{G R}-\hat{Z}_{1}^{G R}\right)$ and $B_{z} \hat{Z}_{1}^{G R}+\left(I-B_{z}\right) \hat{Z}_{2}^{G R}$, respectively. Thus, estimated versions of $B_{y_{i}}$ and $B_{z}$ render the components of BLUE in (7) optimal regression estimators. The variances of these estimators are

$$
\begin{aligned}
& \hat{V}\left(\hat{Y}_{i}^{B}\right)=\hat{V}\left(\hat{Y}_{i}^{G R}\right)-\operatorname{Cov}\left(\hat{Y}_{i}^{G R}, \hat{Z}_{i}^{G R}\right)\left[\hat{\mathrm{V}}\left(\hat{Z}_{1}^{G R}\right)+\hat{\mathrm{V}}\left(\hat{Z}_{2}^{G R}\right)\right]^{-1}\left(\operatorname{Cov}\left(\hat{Y}_{i}^{G R}, \hat{Z}_{i}^{G R}\right)\right)^{\prime} \\
& \hat{V}\left(\hat{Z}^{B}\right)=\hat{V}\left(Z_{1}^{G R}\right) \hat{V}\left(Z_{2}^{G R}\right)\left[\hat{\mathrm{V}}\left(\hat{Z}_{1}^{G R}\right)+\hat{\mathrm{V}}\left(\hat{Z}_{2}^{G R}\right)\right]^{-1}
\end{aligned}
$$

These expressions show the efficiency of $\hat{Z}^{B}$ relative to the $\hat{Z}_{1}^{G R}$ and $\hat{Z}_{2}^{G R}$, and the dependence of the efficiency of $\hat{Y}_{i}^{B}$ relative to $\hat{Y}_{i}^{G R}$ on the strength of the correlation between the variables $z$ and $y$.

In general, computation of the optimal regression estimators for all variables and at any population level is not at all practical because they require the computation of the variance-covariance matrix $V$ of all required GREG estimators based on $x_{1}$ and $x_{2}$. The approximation of the optimal regression estimators by the composite GREG estimators (5) and (6) is shown by expressing the estimated covariances and variances in the optimal $B_{y_{i}}$ and $\boldsymbol{B}_{z}$ as

$$
\operatorname{Cov}\left(\hat{Y}_{i}^{G R}, \hat{Z}_{i}^{G R}\right)=Y_{i}^{\prime}\left(I-P_{X_{i}}\right) \Lambda_{i}^{0}\left(I-P_{X_{i}}\right)^{\prime} Z_{i} \quad \text { and } \quad \hat{\mathrm{V}}\left(\hat{Z}_{i}^{G R}\right)=Z_{i}^{\prime}\left(I-P_{X_{i}}\right) \Lambda_{i}^{0}\left(I-P_{X_{i}}\right)^{\prime} Z_{i}, \quad \text { with }
$$ $\Lambda_{i}^{0}=\left\{\left(\pi_{k l}-\pi_{k} \pi_{k}\right) / \pi_{k} \pi_{k} \pi_{k l}\right\}$ (in terms of first-and-second order probabilities of selection). Replacing then the matrix $\Lambda_{i}^{0}$ by $\Lambda_{i}$ in these expressions gives the suboptimal matrices $\hat{B}_{z}$ and $\hat{B}_{y_{i}}$ in (5) and (6), but makes the implied GREG estimators somewhat less efficient than the optimal regression estimators. However, for certain sampling designs and under certain conditions on the auxiliary vectors $x_{i}, \hat{\boldsymbol{B}}_{z}$ and $\hat{B}_{y_{i}}$ matrix are optimal and the composite GREG estimators are optimal; see Merkouris (2004, 2010a). Optimality requires also the adjustment of the k-th entry of the weighting matrix $\Lambda_{i}$ by a proper calibration constant which is a simple function of the probability of selection of the k-th element. For general designs, a simpler adjustment involves multiplying the entries of $\Lambda_{1}$ and $\Lambda_{2}$ by $1-\phi$ and $\phi$ respectively, with $\phi=n_{1} /\left(n_{1}+n_{2}\right)$, to account for the differential in sample size between the samples $S_{1}$ and $S_{2}$. In any case, as regression coefficients the matrices $\hat{B}_{z}$ and $\hat{B}_{y_{i}}$ are optimal in the sense of generalized least-squares.

### 3.2 Weighting in paradigm 8 (b)

The application of the general estimation method to paradigm 8 (b) is illustrated for the simplest setting involving three independent samples $S_{1}, S_{2}$ and $S_{3}$, representing the same population, with vectors of variables $x$ and $y$ surveyed in $S_{1}$ and $S_{2}$, respectively, and both vectors surveyed in $S_{3}$. Obviously, we can obtain simple HT estimates $\hat{X}_{1}$ and $\hat{X}_{3}$ of the population total $T_{x}$ of $x$, using $S_{1}$ and $S_{3}$, respectively, and HT estimates $\hat{Y}_{2}$ and $\hat{Y}_{3}$ of the total $T_{y}$ of $y$, using $S_{2}$ and $S_{3}$. Micro-integration of the data from the three samples for more efficient estimation of the totals $T_{x}$ and $T_{y}$ can be achieved by a weight calibration procedure applied to the combined sample $S=S_{1} \cup S_{2} \cup S_{3}$ and satisfying the constraints that estimates of the same total from two different samples be calibrated to each other. The resulting vector of calibrated weights $c$ is given by

$$
\begin{equation*}
c=w+\Lambda \mathcal{X}\left(\mathcal{X}^{\prime} \Lambda \mathcal{X}\right)^{-1}\left(0-\mathcal{X}^{\prime} w\right) \tag{9}
\end{equation*}
$$

where $w$ is the vector of design weights for $S, \Lambda$ is the diagonal matrix whose entries are the design weights of $S$, and

$$
\chi=\left(\begin{array}{cc}
-X_{1} & 0 \\
0 & -Y_{2} \\
X_{3} & Y_{3}
\end{array}\right)
$$

is the matrix producing the alignment of the two estimates of $T_{x}$ and the two estimates of $T_{y}$. Then, using the calibrated weights of either sample $S_{1}$ or $S_{3}$, we can obtain the same composite estimator $\hat{X}^{c}$ of $T_{x}$ that incorporates information from all three samples $S_{1}, S_{2}$ and $S_{3}$. Thus, using the sub-vector of calibrated weights $c_{3}$ of sample $S_{3}$ we obtain $\hat{X}^{C}$ in the simple linear form $\hat{X}^{c}=X_{3}^{\prime} c_{3}$. Similarly we obtain a composite estimator
$\hat{Y}^{C}=Y_{3}^{\prime} c_{3}$ of $T_{y}$. It is instructive to express these composite estimators equivalently as GREG estimators. Thus, in view of (9), the estimator $\hat{X}^{c}$ can be expressed as

$$
\begin{align*}
\hat{X}^{C} & =\hat{X}_{3}+\hat{B}_{1 x}\left(\hat{X}_{1}-\hat{X}_{3}\right)+\hat{B}_{2 x}\left(\hat{Y}_{2}-\hat{Y}_{3}\right) \\
& =\hat{B}_{1 x} \hat{X}_{1}+\left(I-\hat{B}_{1 x}\right) \hat{X}_{3}+\hat{B}_{2 x}\left(\hat{Y}_{2}-\hat{Y}_{3}\right) . \tag{10}
\end{align*}
$$

It is seen from (10) that the composite GREG estimator $\hat{X}^{C}$ is approximately (for large samples) unbiased, and derives its efficiency from combining the two elementary estimators $\hat{X}_{1}$ and $\hat{X}_{3}$ (pooling information from samples $S_{1}$ and $S_{3}$ ) and from borrowing strength from sample $S_{2}$ through the correlation between $x$ and $y$. Another expression for $\hat{X}^{C}$ is

$$
\begin{align*}
\hat{X}^{C} & =\hat{X}_{3}^{G R}+B_{1 x}\left(\hat{X}_{1}-\hat{X}_{3}^{G R}\right) \\
& =B_{1 x} \hat{X}_{1}+\left(I-B_{1 x}\right) \hat{X}_{3}^{G R}, \tag{11}
\end{align*}
$$

where $\hat{X}^{G R}=\hat{X}_{3}+X_{3}^{\prime} \Lambda \Psi\left(\Psi^{\prime} \Lambda \Psi\right)^{-1}\left(\hat{Y}_{2}-\hat{Y}_{3}\right)$, with $\Psi$ being the second column of $\mathcal{X}$, is the GREG estimator of $T_{x}$ incorporating the regression effect of $y$ on $x$. The derivation of (11) and analytical expressions of the partial correlation coefficients $\hat{B}_{1 x}$ and $\hat{B}_{2 x}$ can be found in Merkouris (2010b). Expression (11) suggests that the composite GREG estimator $\hat{X}^{c}$ is more efficient than each of its two components $\hat{X}_{1}$ and $\hat{X}_{3}^{G R}$, and more efficient than the composite regression estimator which does not incorporate the information on $y$ (by not borrowing strength from sample $S_{2}$ ).

In case that the three samples have some common auxiliary variables with known totals, an extended calibration procedure, similar to that described in Section 3.1, can be used to incorporate this additional auxiliary information into the weighting structure of the three samples.

Details on the described estimation methodology, including discussion of the efficiency of the composite estimators $\hat{X}^{C}, \hat{Y}^{C}$ and their connection with the associated BLUE, are given in Merkouris (2010b). It should be noted that the BLUE was proposed for the present paradigm by Chipperfield and Steel (2009), who provided analytic expressions of the estimators for scalars $x$ and $y$ assuming simple random sampling and known variance covariance matrix of the vector ( $\left.\hat{X}_{1}, \hat{X}_{3}, \hat{Y}_{2}, \hat{Y}_{3}\right)^{\prime}$. Application of the proposed general estimation method to paradigm 8(a) is as in paradigm 1, and application to paradigm 8 (c) is a mixture of the techniques used in paradigms 8 (a) and 8 (b).

## 4. Conclusion

There is a variety of possible schemes of data integration in social surveys, with a wealth of related auxiliary sources. Suitable weighting of the combined samples can produce efficient composite estimators incorporating information from these sources. Considering the added practical complexity in processing multiple sources of data, it is important that any weighting procedure applied to the combined sample be as statistically efficient as possible under the constraint of operational efficiency.

The proposed general weighting method produces composite estimators of totals which are approximately (for large samples) BLUE, that is, they are approximately unbiased, as special regression estimators, and approximately of minimum variance - in certain sampling settings they are exactly BLUE. The method is computationally very convenient, requiring only a simple adaptation of the generalized regression procedure commonly used in statistical agencies. It involves a single-step calibration of the weights of the combined sample. Thus, using a single set of calibrated weights that incorporate the available information from all the combined samples, estimated totals for all variables at any population level can be obtained simply as weighted sums of values for the relevant units from only
one of the samples containing the variables of interest. Furthermore, carrying out the described calibration procedure on the combined sample greatly facilitates variance estimation by replication methods, such as the Jackknife.

The domain of application of the proposed weighting method, as described in Sections 3.1 and 3.2, covers paradigms 1 to 5 and 8. A generalization of the estimation method for sampling settings involving more than two auxiliary sources is straightforward, making more evident the operational power of the calibration procedure. An extension of the method to paradigms 6 and 7, with the appropriate modifications to handle the dependence of the samples, awaits further study and will be discussed elsewhere. For paradigm 7 (double sampling), in particular, there is an extensive literature on alternative weighting and estimation methods. The more traditional settings of data integration of paradigm 9, already much investigated, are outside the scope of the present study.

## References

Chipperfield and Steel (2009). Design and Estimation for Split Questionnaire Surveys, Journal of Official Statistics, 25, pp. 227-244.

Hidiroglou, M.A. (2001). Double sampling, Survey Methodology, 27, pp. 143-154.
Jones, R.G. (1980). Best Linear Unbiased estimators for Repeated Surveys, Journal of the Royal Statistical Society, Ser. B, 42, pp. 221-226.

Merkouris, T. (2004). Combining Independent Regression Estimators from Multiple Surveys, Journal of the American Statistical Association, 99, pp. 1131-1139.

Merkouris, T. (2008). Efficient Calibration for Some Variants of Double Sampling, American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 3329-3335

Merkouris, T. (2010a). Combining Information from Multiple Surveys by Using regression for More Efficient Small Domain Estimation. Journal of the Royal Statistical Society, Ser. B, 72, pp. 27-48.

Merkouris, T. (2010b). An Estimation Method for Matrix Survey Sampling, American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 4880-4886.

Renssen, R.H. and Nieuwenbroek, N.J. (1997). Aligning Estimates for Common Variables in Two or More Sample Surveys, Journal of the American Statistical Association, 92, pp. 368-375.

Wolter, K.M. (1979). Composite Estimation in Finite Populations, Journal of the American Statistical Association, 74, pp. 604-613.

Wu, C. (2004). Combining Information from Multiple Surveys through the Empirical Likelihood Method, Canadian Journal of Statistics, 32, pp. 15-26.

Zieschang, K.D. (1990). Sample Weighting Methods and Estimation of Totals in the Consumer Expenditures Survey, Journal of the American Statistical Association, 85, pp. 986-1001.

## SESSION 8B

## THE CANADIAN CENSUS

# Integrating Census and Household Survey Listing Operations 

Tom Haymes and Jean-Luc Bernier ${ }^{1}$


#### Abstract

The 2011 Census will once again utilize a mailout methodology and the Address Register will be the source of the list of addresses. The major change over the 2006 Census is that a $100 \%$ field verification of the address lists will not be performed. Instead, field verification will be performed on only a portion of the listing units. A new targeting methodology was developed to select the listing units that will undergo field verification with the goal to minimize dwelling undercoverage. The main component of the score function developed is the growth in addresses resulting from updating the Address Register with administrative sources. Another change for 2011 is that the listing activity will be conducted over a two year period prior to the Census instead of the one-time approach utilized for 2006. The Household Survey Program also conducts a field activity to create and/or verify address lists for in-sample clusters. In order to avoid both programs conducting listing activities in the same area, a coordination of listing activities has been done and the verified address lists are shared between the programs. This has involved promoting in time Census listing units so the results can be used by Household Survey, and avoiding clusters that have been verified by Household Survey and updating the Address Register with the cluster address lists. In addition, refinements to the listing units were required to handle the spatial intersection between listing units and clusters.


Key Words: Census, Address Register, Targeted Listing, Household Survey, Administrative Sources

## 1. Introduction

Statistics Canada conducts a census of population every five years. The census covers the entire population of Canada providing key information regarding population demographics, influencing economic and social policies throughout the country. The Labour Force Survey (LFS) is one of the major surveys conducted by the Household Survey Program. It is a monthly survey of approximately 56,000 households which provides a detailed, current picture of the labour market across the country. Many of the other Household Survey projects also use the LFS frame to create samples as well.

Both of these operations require an accurate dwelling frame in order to ensure proper representation, which is integral in producing valid and accurate statistics. To ensure an accurate and up to date dwelling frame, these operations employ a methodology combining the use of the Address Register (AR) and manual field verification. With similar methods and goals, the Census and Household Survey Listing Operations have been integrated by coordinating listing activites and sharing listing results.

## 2. Census Listing Operation

### 2.12006 vs. 2011 Methodology

The 2006 Census was the first Canadian census where the majority of questionnaire delivery was conducted by mailout. A mailout region was defined containing approximatly $67 \%$ of all dwellings. To be able to mail census questionnaires to these dwellings, a complete list of mailing addresses in the mailout region was required. To achieve this, $100 \%$ field listing was conducted in the fall of 2005 to update the existing address frame. The initial address frame divided into collection units (CU) was provided by the AR, to which field staff verified,

[^73]updated, added or deleted dwellings within the CU. The AR proved to be accurate in some areas while lacking in others. Well established areas proved to change slowly, if at all. The AR was excellent in these areas, however, in many developing/changing areas the field verification proved very useful in identifying missing or superfluous dwellings on the AR. In these areas, characterized by established neighbourhoods being modernized or new subdivisions being built in the urban fringes, the listing was most beneficial.

For the 2011 Census, listing is performed on a lesser scale than 2006. Targeted listing will be performed in listing units (LU) where it is deemed to be most beneficial. These LU are defined very similarly to CU. Areas not listed will use the AR as a frame without field verification. The objective is to optimize the balance between reducing dwelling coverage errors and costs. For the most part, listing will only be performed in LU with similar characteristics to the problematic areas observed in the 2006 listing, avoiding well established neighbourhoods and focusing mainly on the areas of high growth. Approximately one third of the mailout area will be listed in quarterly cycles from the spring of 2009 into early 2011.

Also new for the 2011 Census is an increase in the size of the mailout region, now containing $79 \%$ of all dwellings, and a quarterly update of the AR using administrative growth sources.

### 2.2 Targeting LU

While overcoverage is not desirable, the principle focus of targeted listing will be reducing dwelling undercoverage. However, an alternate method known as Non-Response Follow-Up (NRFU) is put into place during census to obtain completed questionnaires from non-respondents through a combination of telephone calls and field visits. During this process field staff will also cancel non-existent or non-applicable dwellings, reducing the overcoverage on the AR.

A score function was created to determine which LU will be targeted for field verification. The algorithm utilizes the counts of AR growth records since the last Census, geocoding inconsistencies, and the discrepancy between current valid 2006 dwellings on the AR compared to 2006 dissemination counts.

Growth is identified through the quarterly update of the AR using administrative source files, matching these files against the AR to identify new addresses. Areas of high growth suggest that a lot of changes have occurred since the 2006 listing and thus positively influence the score function. Also these sources are often incomplete, meaning that we can expect that the sources only identify a subset of the growth existing in these areas. Testing has shown that listing areas with a large number of dwellings added since the 2006 Census often results in a large return of added dwellings.

Discrepancies existing between the 2006 dissemination counts and the current AR counts of valid 2006 dwellings produce an obvious red flag. For example, if an area produced a 2006 dissemination count of 150 dwellings, and currently in the same area the AR only has addresses for 75 of the 2006 valid dwellings, then clearly this area is a prime suspect of undercoverage. These situations often occur due to non-civic style addressing in the 2006 mailout region or the field use of descriptive addresses in the 2006 Visitation Records which were used to extend the mailout region for 2011.

Geocoding inconsistencies between the 2006 field verified block and the block determined through geocoding the AR to the Statistics Canada geographic frame are also a source of potential coverage problems. These can suggest an inaccuracy in the address itself, or if an abundance of dwellings are not geocoding within the same area it could suggest that the existing geographic frame is not up to date.

### 2.3 Quality Verification

As a quality assurance measure, each cycle $5 \%$ of LU selected is non-targeted LU selected as control units. These have been used to measure the validity of the score function through comparison to targeted LU . The control LU also provides an estimate of the coverage in non-targeted areas. Control LU can be any LU below the targeting threshold that are not already being listed by the Household Survey Operation and are also within a reasonable geographic distance from a targeted LU, which reduces the cost associated with its listing.

## 3. Household Survey Listing Operation

### 3.1 The Use of the Address Register

Similar to the Census Operation, the Household Survey Operation conducts a listing activity to create and verify address lists for in-sample clusters. Clusters are similar to the LU used in the Census methodology, both being a set of listing blocks, however they do not necessarily cover the same area. A cluster can intersect several LU, and conversely an LU can intersect several clusters.

Prior to a redesign in 2004, the sole source of addresses used to build the second stage sampling frame for the LFS were listing results of all sampled clusters. However, since then the methodology has included the use of the AR, allowing the dwelling frame to be built in a less costly manner. To accommodate this use, clusters are now maintained on the geographic database so AR address lists geocoded to clusters can be provided to Household Survey.

### 3.2 Initial Listing and Listing Maintenance

Depending on estimated AR quality, clusters may be listed when joining the sample, known as initial listing, and may be listed on a periodic basis while in the sample, known as listing maintenance. The quality of the AR is assessed using a similar approach to the Census Operation, taking into account growth, coverage, geocoding inconsistencies, as well as the count of structures containing two or three dwellings as these are known to be suspect of both undercoverage and overcoverage. Clusters are then classified into one of three categories:
1.) AR Group 1 - The AR is deemed to be of excellent quality with few geocoding uncertainties, little growth and an estimated coverage rate between $97.75 \%$ and $103 \%$. The AR is used to build the address frame, without initial listing.
2.) AR Group 2 -The AR is deemed to be of good quality with an acceptable amount of geocoding uncertainties and a coverage rate above $90 \%$. The AR is used to build the address frame which is then field verified through initial listing.
3.) AR Group 3 - The AR is deemed to be of poor quality or there are a large number of geocoding uncertainties. The address frame is created from scratch through initial listing.

For AR Group 2 and AR Group 3 clusters, initial listing is performed two to four months prior to the first interview.
Clusters can also be listed on a periodic basis while in the sample through listing maintenance. This usually occurs during survey week of the LFS. In principle, listing maintenance ensures an up to date address frame for all insample clusters, regardless of the AR Group when introduced.

## 4. Integrating Listing Operations

The Census and Household Survey Listing Operations have been collaborating to share the listing results and responsibilities. To better utilize resources, instead of having the mutually targeted areas listed by both projects, measures are taken whenever possible to have the area listed by only one project with the results provided to both.

### 4.1 Integrated Listing Application

A new listing application was required to incorporate the needs of both operations. The developed application allows for a fully automated data collection, entry and capture system using a laptop in the field. In the past the Census Operation used paper booklets in the field followed by a data entry operation at head office, while the Household Survey Operation was using a similar electronic application. However, this application was growing out of date and did not meet the needs of the Census, most importantly the block number was not captured. It was concluded that it was more feasible to create a new application encompassing all aspects required rather than update the existing one. The new application provides consistency in the listing results allowing for the data collected to be easily interchangeable between both operations. It also provides a consistent application for listers regardless of which
operation they are currently working for and a more efficient electronic data capture for the Census Operation. Currently the new application is only being used by the Census Operation but will be employed by the Household Survey Operation in the near future. Until the implementation, a block must be assigned to the cluster listing results before applying it to the AR, which is accomplished through geocoding methods.

### 4.2 Coordinating Listing Activities

Census and Household Survey listing schedules are compared each cycle. The spatial overlap of in-sample clusters and targeted LU are examined to determine possible areas of cooperation. The Household Survey Operation will provide the Census Operation with the listing results of every cluster listed. For this reason, LU will only be targeted based on the blocks that are not covered by a cluster that has been in the sample in the prior three months or will be in the sample in the following nine months. By not applying the score function to in-sample blocks, the Census Operation avoids targeting an LU based on the blocks they will already be receiving listing results for.

In addition, if at least $85 \%$ of the dwellings within an LU also fall within an in-sample cluster then the LU will not be listed by the Census Operation, instead the AR will be updated using the results of the cluster listing. If $15 \%-85 \%$ of the dwellings within an LU also fall within an in-sample cluster then the LU can be refined to not include the cluster. The LU will be restructured to remove the blocks within the in-sample cluster, with the new groups of blocks joining with neighbouring LU or possibly forming new LU on their own. If less than $15 \%$ of the dwellings within an LU are covered by an in-sample cluster then the LU is treated as a normal LU. If this LU were to be targeted the entire LU, including the overlapping region, would be listed.

If a targeted LU or group of targeted LU completely span a cluster that the Household Survey Operation is planning on listing in the future, then the Census Operation will list the LU during the upcoming cycle and provide the results to the Household Survey Operation. The AR is updated with the listing results within six months of the start of the listing cycle.

### 4.3 Benefits

By using the results provided by the Census Operation, the Household Survey Operation is able to obtain the listing results for in-sample clusters without actually having to spend the time and money required to list it themselves. As long as the Census Operation plans to list all LU associated to the cluster, it can be listed within the time restraints required to meet the survey deadlines.

Similarly, the Census Operation is able to obtain more listing results without actually having to spend the resources to list them. This provides an increase in the accuracy of the overall dwelling coverage for the Census. Any LU that is targeted but also covered by an in-sample cluster allows the Census Operation to list a different LU, while still receiving the results for both. Also provided are the listing results for non-targeted LU, while not expected to dramatically increase coverage they still provide a marginal increase. And teamed with the control LU listed by the Census Operation, they provide more information to critique the score function with a better estimate of the coverage in non-targeted areas.

## 5. Conclusion

In conclusion, the integration of the Census and Household Survey Listing Operations has created a mutually beneficial relationship between the two projects. Through sharing listing results the Household Survey Listing Operation is provided an opportunity to reduce field costs without the often associated sacrifice in quality, while the Census Operation is able to increase quality without the often associated increase in field costs.

### 5.1 Future Work

The integration of the two operations is still a young project, and like most new projects there is always room for improvement. Specific areas that will be investigated include:

- Additional administrative growth sources used to update the AR. This includes but is not limited to a Point of Call file from Canada Post Corporation currently being evaluated. Coupled with coverage estimates for the 2011 Census, the strength of future administrative sources will weigh heavily on the listing methodology for the 2016 Census.
- The geographic compatibility between Census and Household Survey definitions. This will increase the level of efficiency employed during interaction and could include an easier transition between LU and clusters.
- Similarly, LU could be refined based on problematic areas. This would reduce the amount of stable areas being listed.
- The newly developed listing application will be employed by Census listing, LFS initial listing and LFS listing maintenance.


## References

Chen, E. and Meyer, S. (2010). The Labour Force Survey Frame Creation, Sample Control and Listing Operations, unpublished report, Ottawa, Canada: Statistics Canada.

Post, S. (2007). Partial Block Canvass for the 2011 Census: Risks and Rewards, unpublished report, Ottawa, Canada: Statistics Canada.

Rodrigue, J.F., Thompson, G. and Turmelle, C. (2005). Using the Canadian Address Register in the Labour Force Survey: Implementation results and lessons learned, unpublished report, Ottawa, Canada: Statistics Canada.

# Overview of the Canadian Postcensal Survey Program 

Éric Langlet ${ }^{1}$


#### Abstract

Since 1986, a postcensal survey program has been put in place in Canada to conduct follow-up interviews on specific themes with a sample of respondents to the census long-form questionnaire. This long form is administered to one in five households in Canada, except in remote areas and on Indian reserves where all households must complete it. Following this first phase, a stratified sample of long-form respondents is selected based on their characteristics observed in the first phase. Using the census as a survey frame makes it possible to study rare populations belonging to small domains and provides a rich set of characteristics that can be used in sample selection, non-response adjustment and post-stratification. However, this two-phase sampling plan where the first-phase sampling fraction is not negligible presents particular challenges when estimating variance. This paper gives an overview of the postcensal survey program since its creation and provides a more detailed look at the methodological issues associated with Aboriginal postcensal surveys, including the on-reserve survey.


Keywords: Postcensal survey; two-phase design; generalized bootstrap method; on-reserve telephone survey.

## 1. Principle of postcensal surveys

A postcensal survey is a survey where the sample is selected from the responses to the census, usually the long form or detailed questionnaire. On reserves and in remote areas, all households receive version 2D of that form while elsewhere in Canada, a systematic one-in-five sample of households receives version 2B of that form. Version 2D of the long form is similar to version 2 B , except that it is designed to be completed by personal interviews rather than by self-enumeration, and the examples given are adapted for remote areas and Indian reserves. A postcensal survey sample design is thus a two-phase design. The long-form sample constitutes the first-phase sample, while in the second phase, a stratified sample of respondents is selected based the characteristics observed in the first phase.

Postcensal surveys are conducted shortly after the census, thus allowing certain census and survey operations to be integrated. The postcensal approach is particularly suited to collecting information on rare and dispersed populations since a very large first-phase sample is available. Moreover, postcensal surveys have a wealth of characteristics for stratification, non-response adjustment and calibration, or all long-form information. Once the survey is completed, a very rich database is created by combining survey and long-form data.

## 2. History

### 2.11986 and 1991 HALS

The first Canadian postcensal survey, the Health and Activity Limitation Survey (HALS), was created in 1986 and repeated in 1991. Its target population was children and adults with activity limitations of varying degrees of severity. It included a household component, covering private dwellings and certain types of non-institutional collective dwellings, as well as an institutional component. The survey's household component collected information on the nature and severity of disabilities and on barriers disabled persons face in their activities of daily living. .

[^74]The 1986 and 1991 surveys used two Census filter questions on activity limitations. A sample of individuals who answered "YES" to those questions was drawn with a large sampling fraction, while a sample of individuals who answered "NO" was drawn with a small sampling fraction. The survey questionnaire began with a set of filter questions on activity limitations defining the survey's target population. A detailed follow-up was administered to the individuals selected based on those questions. In the 1991 HALS, $80 \%$ of people who answered "YES" to the Census filter questions were limited in the survey, compared to only $10 \%$ for those who answered "NO". Since the "NO" answers represented $90 \%$ of the total population, activity-limited individuals in the survey of that group comprised $53 \%$ of the HALS target population. The disability rate obtained in the 1991 HALS was $17 \%$. These figures are summarized in Table 2.1-1.

Table 2.1-1
Relationship between limited individuals in the Census and the 1991 HALS

| Adult population |  |  | 1991 HALS |  |
| :---: | :---: | ---: | ---: | ---: |
|  |  |  |  |  |
|  |  |  |  |  |
| 1991 Census | "YES" (limited) | $\mathbf{8 0 \%}$ | Not limited | Total |
|  |  | $(47 \%)$ | $(30 \%$ | $10 \%$ |
|  |  | $\mathbf{1 0 \%}$ |  |  |
|  | "NO" (not limited) | $90 \%$ | $\mathbf{9 0 \%}$ |  |
|  | $\mathbf{( 5 3 \%})$ | $(97 \%)$ |  |  |
|  |  | $\mathbf{1 7 \%}$ | $83 \%$ | $100 \%$ |

### 2.22001 and 2006 PALS

In 2001, the survey was renamed Participation and Activity Limitation Survey (PALS); it was repeated in 2006. In tests conducted before the 2001 PALS, a study showed that the disability rate would be close to $40 \%$ with the "YES"/"NO" sampling strategy used in 1986 and 1991. This was essentially due to a very large increase in mild disabilities, resulting to a large degree from the "NO" sample. Following this test, the Census filter questions were modified to make them more inclusive and the "NO" sample was eliminated. From that point on, emphasis was put on a more severely limited population, i.e. individuals limited according to both the Census and survey filter questions.

Table 2.2-1 presents the changes in disability rates between 1986 and 2006. A drop was naturally observed between 1991 and 2001 due to the change in methodology. The rate continued to increase after that. In addition to aging of the population, the increased rate is due to the fact that disability is more socially accepted these days. People are therefore less reluctant to report a disability than in the past.

Table 2.2-1
HALS and PALS disability rates between 1986 and 2006

| HALS 1986 | HALS 1991 | PALS 2001 | PALS 2006 |
| ---: | ---: | ---: | ---: |
| $12.5 \%$ | $17.0 \%$ | $14.6 \%$ | $16.6 \%$ |

### 2.31991 and 2001 APS

The first Aboriginal Peoples Survey (APS) was conducted in 1991. It targeted individuals with Aboriginal identity living both on and off reserve and concerned Aboriginal lifestyles and living conditions. In 1991, individuals were included in the Aboriginal identity population if they reported being North American Indian (NAI), Métis or Inuit (called Aboriginal self-reporting), or if they reported being a registered Indian under the Indian Act. Moreover, individuals belonged to the Aboriginal ancestry population if they reported Aboriginal ethnic origin among their ancestors. It should be noted that individuals belonging to the Aboriginal ancestry population may or may not be included in the Aboriginal identity population; they can have Aboriginal ancestors, yet not consider themselves as Aboriginal, nor be a registered Indian. Since there were no questions on Aboriginal self-reporting in the 1991 Census, the sample was drawn from those who reported having Aboriginal ancestry or being a registered Indian. In the interview, detailed follow-up was done solely on individuals with Aboriginal identity according to the survey.

The questions on Aboriginal self-reporting and Indian band membership were introduced in the 1996 Census. From then on, individuals were included in the Aboriginal identity population if they reported being Aboriginal, a registered Indian or an Indian band member. The survey drew its sample from those who reported either Aboriginal identity or ancestry. The two populations were also part of the survey target population. Other than off-reserve regions, only large reserves in each province were covered. Estimates referred only to the reserves selected and no aggregated estimates were produced for reserves.

## 3. 2006 Aboriginal postcensal surveys and variance estimation

A new Aboriginal survey dealing with early childhood development among young Aboriginal children aged 0 to 5, the Aboriginal Children's Survey (ACS), was introduced in 2006 while the APS, on its side, was again repeated. The target population for these two surveys included both Aboriginal identity and Aboriginal ancestry individuals. Only areas outside reserves were covered. The domains of estimation for these surveys were made up of cross-tabulations of geographic regions, Aboriginal groups and age groups.

As with other postcensal surveys, the Aboriginal surveys' sample design is a two-phase design. As mentioned above, the long-form sample comprises the first-phase sample. It is a systematic household sample stratified by collection unit. A collection unit is a small geographic region of about 200 to 500 households used for collection. Recall that in regions where version 2B of the long form is distributed, the sampling fraction is one in five households, while it is $100 \%$ where version 2 D is distributed. In the second phase, a stratified sample of Aboriginals in the census is drawn in each domain of estimation by substratifying by region type (2B or 2D). An optimal allocation is used between 2B and 2D components of each domain of estimation.

This sample design presents many challenges for variance estimation. First, the sampling units are not the same in the two phases since households are selected in the first phase and individuals in the second. Stratification is different in each of the two phases, and the second-phase strata are not nested within the first-phase strata. Moreover, since the first-phase sampling fraction is one-fifth in regions that receive version 2 B of the long form, it is not negligible. The sampling fraction in the second phase is typically very high, particularly for the ACS where the domains of estimation are often very small. Many types of estimates are produced for these surveys. In addition to the descriptive statistics produced such as proportions, means, totals and ratios, an analytical file is sent to research data centres. With this file, users can perform logistic regressions, estimate distribution quantiles, etc. The method must therefore be simple and flexible to be applicable to all these different types of estimation.

The method selected for estimating variance is the generalized bootstrap method for two-phase designs (Langlet, Beaumont and Lavallee, 2008). The idea of the generalized bootstrap method is described in Beaumont and Patak (2010). Application of any bootstrap method is simplified by using bootstrap weights. It can be shown that whatever the bootstrap method used, the initial bootstrap weight is the product of the sampling weight multiplied by a random adjustment factor. For the with-replacement Rao-Wu bootstrap (Rao and Wu, 1988), for example, the random adjustment is a function of the multiplicity of the unit in a given bootstrap sample, i.e. the number of times the unit is found in that bootstrap sample.

The method used can be described as follows. Let $w_{k}$ be the survey weight associated with unit $k$ combining the two sampling phases, $w_{k}^{*}$ the bootstrap weight associated with this unit for a particular bootstrap sample, and $a_{k}$ the corresponding random adjustment factor. Then, $w_{k}^{*}=w_{k} a_{k}=w_{k} a_{1 k} a_{2 k}$, where $a_{1 k}$ and $a_{2 k}$ are random adjustment factors associated with each of the two phases. In principle, any specific distribution can be used to generate the adjustment factors, provided that those distributions meet certain specific conditions described in Langlet, Beaumont and Lavallée (2008).

Having two sets of random correction factors has a major advantage. The first set can be used for estimates based on the first phase only, i.e. estimates based on the census long form. These estimates are used to calibrate weights to the census total estimates. This produces variable census totals for each bootstrap sample, which takes into account the fact that the census totals used are based on a sample and are not fixed known totals.

Disadvantages of this approach include the possibility of generating negative bootstrap weights. To remedy this, the bootstrap weights were transformed to reduce their variability. As a result, the variance calculated using the modified bootstrap weights had to be multiplied by a factor that is a function of a parameter called phi. The value of the parameter is selected to be the smallest integer which results in all bootstrap weights being positive. The method used is also slightly biased in the sense that it overestimates variance a bit. The magnitude of this overestimation is considered negligible for Aboriginal postcensal surveys.

## 4. On-reserve surveys

No aggregated on-reserve estimates have been produced from Aboriginal postcensal surveys since 1991. Various Aboriginal survey clients and users have identified a need to conduct a cyclical thematic survey similar to the Statistics Canada's General Social Survey. Thus, a particular theme would be chosen for each survey occasion and would be repeated regularly every 2,3 or 5 years, for example. Each thematic survey would also include a set of indicators common to all themes in order to regularly measure their evolution over time. A major problem with conducting such thematic surveys is that a particular survey occasion would not necessarily fall in a census year. Obviously, the census frame quickly becomes outdated after a year or two.

The possibility of conducting intercensal on-reserve telephone surveys was examined using a test. This test, carried out in 2009, three years after the 2006 Census, was covering four groups of reserves, including the three Treaty regions in Alberta (which covers all Alberta reserves) and the Nishnawbe-Aski Nation group of reserves in Northern Ontario. Estimates were targeted for the four groups of reserves, not the reserves themselves.

In previous Aboriginal postcensal surveys, reserves were always covered with personal interviews. The main advantage of telephone surveys for on-reserve surveys, other than their significantly lower cost, is that a sample of households or individuals can be distributed over all the reserves of a group without additional costs to cover more reserves. Since the personal interview method is much more expensive, an additional sampling stage is necessary and consists of first drawing a sample of reserves in each group, then drawing a sample of households or individuals within the selected reserves. This reduces the efficiency of the sample design. Since there can be a significant cluster effect at the reserve level, a substantially larger sample size is necessary to achieve the same accuracy as a survey drawing a random sample of households or individuals within a group of reserves. In the 2009 test, two collection methods were compared using two independent samples: telephone interview collection and personal interview collection.

Since no up-to-date telephone number survey frame was available, a random direct dialling (RDD) method, known as "elimination of non-working banks" by Statistics Canada, was used. A sample of telephone numbers was drawn within each group of reserves, which is equivalent to a simple random sample of households in the first stage. In the second stage, one Aboriginal adult aged 15 years and over was drawn at random for each household selected in the first stage.

For personal interviews, a sample of reserves stratified by size was drawn in the first stage in each group of reserves. All dwellings on the reserves selected in the first stage were then listed. A dwelling sample was then drawn in the second stage, then one Aboriginal adult aged 15 years and over was drawn at random in the third stage.

A number of problems were experienced with the RDD approach. Among other things, it is very hard to target geographic units as small as groups of reserves with this method. In Alberta, this led to many calls ending up outside the target reserves. Since telephone number banks (first eight digits of telephone numbers) have very low density in Ontario especially, many invalid numbers were generated. Good coverage of existing telephone numbers was obtained with RDD, but telephone penetration rates (proportion of households with telephone service) were low for many reserves. Since the characteristics of individuals with a telephone are different from those without, considerable coverage bias was observed. Moreover, since the response rate was significantly lower for the RDD survey than for personal interviews, greater non-response bias was also noted.

The effect of those two sources of bias was that the telephone approach gave a better picture of the Aboriginal population for some characteristics, and a worse picture for others. In particular, the telephone approach underestimates the proportion of Aboriginals with strong Aboriginal characteristics (such as speaking an Aboriginal
language or being heavily involved in community activities). Since such Aboriginals are generally more mobile, they are harder to contact and will have a greater tendency to become non-respondents. In this particular case, this would be more related to the effect of higher non-response bias for telephone interviews than for personal interviews.

## 5. To come in 2011

In terms of postcensal surveys, only the Aboriginal surveys are planned for 2011, providing that funding can be obtained. Those surveys should cover the off-reserve part only; no on-reserve survey is planned at the moment. It has not been confirmed, but the content of the 2011 APS will probably focus on education and employment rather than a variety of topics as in the past. The content of the 2011 ACS should be based on that of the new Survey of Young Canadians, which has replaced Statistics Canada's National Longitudinal Survey of Children and Youth.

The sample designs for those two surveys will have to be changed radically to take into account the new voluntary survey that replaces the mandatory census long form, i.e. the National Household Survey (NHS). Since it is a voluntary survey, a substantially lower response rate than in previous censuses with the long form is expected. Moreover, the average sampling fraction outside remote areas and reserves will go from one in five households in 2006 to about one in three in 2011, which will increase the number of non-respondents even more. For budget reasons, it will therefore not be possible to follow up on all non-respondents. The strategy used will be rather to draw a sample of non-respondents for non-response follow-up. The sample for Aboriginal postcensal surveys should be drawn from NHS respondents who responded without follow-up and from respondents selected in the nonrespondent follow-up sample. Thus, the sample design for postcensal surveys will go from two phases to three phases, which will pose a number of challenges. Contingency plans should also be considered in case the number of available Aboriginal respondents is too low in some strata or non-response bias for Aboriginals in the NHS is too high. To this end, sampling strategies for the non-respondents remaining from the NHS follow-up sample are being considered.

## References

Beaumont, J.F. and Patak, Z. (2010). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling, paper submitted for publication.

Langlet, É., Beaumont, J.F. and Lavallee, P. (2008). Bootstrap Methods for Two-Phase Sampling Applicable to Postcensal Surveys, paper presented to Statistics Canada's Advisory Committee on Statistical Methods, April 2008, Ottawa.

Rao, J.N.K. and Wu, C.F.J. (1988). Resampling inference with complex survey data, Journal of the American Statistical Association, 83, pp. 231-241.

# An Integrated approach to collection of Census data in Canada 

Pamela Tallon and Linda Ramsey ${ }^{1}$


#### Abstract

The 2006 Canadian Census saw the successful implementation of the biggest methodological changes since 1971. The 2011 Census will largely repeat the 2006 approach; however, methodologies must continue to be enhanced to address the challenges that arise and to work towards new efficiencies and streamlining of processes. Firstly, to further reduce the reliance on a large and decentralized workforce, the plan is to increase the target for mail-out from the $70 \%$ of dwellings achieved in 2006 to $80 \%$ in 2011. Secondly, based on the strength of the Internet responses in 2006, the plan is to reduce printing and handling of paper questionnaires by targeting a $40 \%$ Internet take-up rate in 2011 compared to the worldleading $18 \%$ achieved in 2006. This will be facilitated by mailing a letter instead of a questionnaire to areas indicating a high probability of internet connectivity, as well as implementing a "Wave Methodology" whereby questionnaires and letters are sent out at predefined times coordinated tightly with a corresponding communication strategy. A number of new tools and strategies are being implemented to help achieve these goals, and all of this has meant an integrated approach to collecting Census data in Canada. This paper presents an overview of the 2011 Canadian Census. It focuses on the various response channels for collecting data, how they are integrated into a single strategy, and the innovative tools and methodologies that are being introduced to conduct the first Canadian Census where the majority of Cana dians may never receive a paper.


Key Words: Canadian Census; Wave Methodology; Internet response channel; field management system

## 1. Introduction

### 1.1 Background

The last Census conducted in Canada in 2006 saw some ground breaking changes not seen in census taking in Canada in 35 years. These included the option of completing the questionnaire on the internet, a Master Control System (MCS) reflecting the current collection status of each dwelling in Canada, mail-out to $70 \%$ of all dwellings, the return of all responses to one centralized data operations centre (DOC), automated data capture, automated editing and failed-edit follow-up from one of 3 offices using a computer assisted telephone interview application.

There were some lessons learned coming out of the 2006 Census. A number involved communication with field staff particularly related to the work that had been completed and what was left to be collected as well as the timeliness of this communication. In addition, there were some processes that were inefficient that caused delays in some of the administrative functions related to the productivity of field staff and their remuneration.

### 1.2 Collection Objectives

Relevance and maintaining a high level of data quality and cost effectiveness continue to be significant drivers in the development of any Census. For 2011, the basic approach is to leverage on the processes and technological advancements that were so successfully implemented in 2006.

There are some specific objectives that are being pursued for the 2011 Census. Firstly, in order to continue to reduce the reliance on a large and decentralized workforce for the operations in the field, the target is to increase the mailout to dwellings from the $73 \%$ figure achieved in 2006 to $80 \%$. Secondly, based on the strength of the Internet response in 2006 as well as the research conducted live during the 2006 Census, the target for internet responses has more than doubled from the $18 \%$ achieved in 2006 to $40 \%$ in 2011. A large internet take-up rate will result in

[^75]increased overall item responses, as well as potential efficiencies in paper processing and follow-up efforts for missing information.

Thirdly, communication between field staff, their supervisors and Head Office will be improved with the introduction of a Field Management System (FMS), a web-based application that will be accessible by all field staff. Lastly for the purposes of this paper, in order to improve the timeliness of knowing a questionnaire has been completed and mailed back to Statistics Canada, Canada Post will read the barcodes of all returned questionnaire packages in one of 16 decentralized sites across Canada, and then the list of barcodes will be automatically transmitted to the DOC on an hourly basis for registration. Notifications of these returns will be automatically sent to the field staff within 2 hours of receipt at the DOC through the FMS so that enumerators stop collection for householders who have already responded.

## 2. Collection Approaches

### 2.1 Delivery

There are three main delivery methods for the 2011 Census:

1. Mail-out to areas where Statistics Canada maintains an Address Register of mailable addresses. These will cover about $79 \%$ of all dwellings in Canada. These addresses are used to initialize the mail-out areas on the Master Control System (MCS).
2. List/Leave where a more traditional approach of census collection is used. Questionnaire packages are dropped off at each dwelling and the address listed in a visitation record book at the same time. These areas are in the more rural areas and will cover a little less than $20 \%$ of all dwellings in Canada. The MCS is initialized with an estimated number of dwellings in list/leave areas. The actual number is confirmed when the information from the visitation records is data captured and transmitted to the MCS.
3. Canvasser approach where enumerators complete the questionnaire through interview. This approach is used in the extreme north, Indian Reserves and small communities where it is expensive to travel. Canvasser areas include about $1 \%$ of all dwellings in Canada.

### 2.2 Response Channels

Canadians living in mail-out or list/leave areas can respond to the Census through a number of response channels:

1. They can mail-back the paper questionnaire. It will be mailed directly to a centralized Data Operations Centre (DOC) as was done for the first time in 2006.
2. They can use a secure access code to complete the questionnaire on-line. The secure access code and URL appear on the front of the questionnaire as well as on the respondent letter.
3. If a completed response is not received by a specific date, an enumerator will contact them by telephone or at their door and complete the questionnaire through interview (called Non-Response Follow-up (NRFU)).
4. After the start of NRFU, a householder will be able to call the census help line and complete the questionnaire through interview on the phone provided that the census operator can identify the specific dwelling on the Master Control system (called Direct Response).

## 3. Tools and Methodology

This section describes some of the new tools and methodologies being implemented for the 2011 Census.

### 3.1 Expanding Mail-out

As noted earlier, Statistics Canada maintains an Address Register which includes all known dwellings with an address in Canada. Only the addresses which Canada Post uses for delivering mail are used to initialize the MCS (the Address Register has addresses that Canada Post does not recognize, e.g. the emergency numbers assigned in the rural areas where Canada Post uses name and rural route type delivery).

The goal was to increase the number of dwellings to which we could mail a letter or questionnaire from $70 \%$ in 2006 to $80 \%$ of all dwellings in Canada. Starting with the 2006 mail-out areas, the following steps were used to delineate the 2011 mail-out areas.

1. The visitation records that were used in the 2006 list/leave areas were data captured. Areas where there were civic style addresses were candidates for inclusion in the 2011 mail-out area. (A civic style address has a civic number, street name and apartment number if applicable).
2. The Address Register was examined to determine where it had good coverage of the area. To do this, a comparison was made of the number of civic style addresses on the Address Register to the 2006 published dwelling count. Where the coverage was poor, it was assumed that there were additional dwellings for which Canada Post Corporation (CPC) would not deliver based on address alone (CPC required the resident's name with the rural route address for delivery).
3. For high coverage areas, the addresses were examined to determine if CPC would recognize them for delivery by running the addresses through software that attaches CPC municipality and postal codes. Where successful, the area was included in the 2011 mail-out.
4. Some areas were removed based on operational requirements, particularly if it was a very small area surrounded by list/leave area.

Subsequently, coverage is ensured within the defined mail-out areas by identifying new dwellings from administrative sources every quarter (e.g. InfoDirect, a list of published residential telephone listings; GST-HST new Housing Rebate file; and the Canada Mortgage and Housing Corporation monthly starts and completion data). In addition, quarterly listing operations were conducted where field staff updates address lists of pre-selected small areas through field checks. To select the areas for listing, the primary trigger is where the administrative growth is being added to the Address Register. Other factors considered are existing problems linking the address to Statistics Canada's geographic base and areas where the Address Register coverage is weaker than in other mail-out areas.

### 3.2 Wave Methodology

Research conducted live during the 2006 Census demonstrated that sending out a letter with a secure access code rather than a paper questionnaire to targeted areas can increase online responses. It was also observed that if nothing else is done, the non-response follow-up workload will also increase. Since non-response follow-up is the most expensive field activity, the goal is to decrease it, not increase it. However, it is known from the literature that multiple contacts using various stimuli are essential for maximising response. One of the most innovative new methodologies for the 2011 Census is the introduction of a "wave approach" to increase the returns of questionnaires by internet without impacting the non-response follow-up workload. The goal is a $40 \%$ internet take-up rate for 2011, up from $18.3 \%$ in 2006.

With the "wave approach", the initial delivery of questionnaires or letters, followed by one or two reminders of the benefits and obligations of responding to the Census, and finally the non-response field follow-up for missing responses during the collection period, can be viewed as consecutive waves of events to motivate response. Four waves of events will be used in mail-out areas and three in list/leave areas for the 2011 Census.

### 3.2.1 Wave 1

Wave 1 will start a week before Census Day and will involve mailing out a letter instead of a questionnaire to areas selected based on their internet response rate in 2006. The letter will be mailed to about $60 \%$ of all dwellings in Canada. It will provide a secure access code for accessing the online questionnaire as well as a toll free telephone number for requesting a paper questionnaire if that is the householder's preference. Sending the letter without a paper questionnaire will encourage respondents to complete their questionnaire online. The letter will ask that the census questionnaire be completed within the next ten days. In the remaining mail-out areas (approximately $20 \%$ of all dwellings in Canada), the traditional questionnaire package will be mailed with the secure access code printed on the front page. In List/Leave areas (the remaining $20 \%$ of all dwellings), Wave 1 will be the traditional drop-off of the questionnaire package, which will also have the secure access code printed on the front page.

### 3.2.2 Questionnaire Request System

Testing during the 2006 Census demonstrated that if a letter is mailed instead of a questionnaire, there must be an easy way for the householder to get a paper questionnaire if that is their preference. For this, an automated telephone service called the Questionnaire Request System (QRS) is being introduced. The householder calls the toll free phone number printed on the letter and enters the 15-digit secure access code also from the letter. Once validated, a questionnaire package will be automatically mailed within 24 hours and received at the dwelling within 2 to 4 days.

For mail-out areas, all of the census questionnaires will be barcoded with the dwelling identifier and addressed months before they are actually delivered. In order to implement a QRS, a "just-in-time" addressing process was required. To do this, a certain number of questionnaires to satisfy an expected volume of questionnaire requests are pre-printed with "alternate" dwelling identifiers and barcodes; the address portion is left blank. The alternate barcode is positioned on the questionnaire to show through the envelop window on both the mail-out and mail-back envelops. When calls come in for paper questionnaires, Statistics Canada uses the secure access code that the caller enters into the telephone to identify the real dwelling identifier and the address for the dwelling on the MCS. These are transmitted in a secure fashion to Canada Post. Canada Post prints the real dwelling identifier and address on the mail-out questionnaire envelop. Both the alternate barcode showing through the envelop window and the real one on the questionnaire envelop are then scanned to create a correspondence file that is transmitted to Statistics Canada. When the completed questionnaire is mailed back showing only the alternate barcode through the window of the mail-back envelop (the real identifier was on the mail-out envelop), the real identifier is found through the correspondence file.

### 3.2.3 Wave 2

As the response from the Wave 1 letters and questionnaires starts to decline, the concept is to send out a reminder letter in Wave 2 to motivate non-responding householders to complete their questionnaire either online or on paper. It is important to give them enough time to react to Wave 1 before starting the Wave 2, but it can't be too long as there is only a limited time to conduct a Census.

Wave 2 will begin on Census Day when non-respondents in mail-out areas will be mailed a reminder letter. The reminder letter will again include a secure access code to complete online or the number for the QRS to request a paper questionnaire. In the List/Leave areas, Wave 2 will begin on Census Day but instead of an addressed letter, a generic reminder card will be dropped off by Canada Post at all list/leave dwellings. The letter will remind householders of their legal obligation to complete a questionnaire, and it will also thank people who may have already responded.

Wave 2 will stimulate an increase in response and again, as the returns from the reminder letter start to decline, another reminder is sent out to motivate even more response.

### 3.2.4 Wave 3

Wave 3 will start eight days after Census Day. Rather than send out another letter, the reminder will be a questionnaire package for those who have not yet received a questionnaire (they received a letter in Wave 1 and did not request a paper questionnaire through QRS or from the Census Help Line). It has been determined that it is not cost effective to send out a second questionnaire to dwellings that have not responded. However, if the dwelling received a questionnaire in Wave 1 and is still non-responding, and there is a valid phone number on the MCS, they will receive a voice broadcast message by telephone reminding them of their obligations to complete a census questionnaire.

The questionnaire package for Wave 3 will contain a letter that will remind householders that it is their legal obligation to complete a questionnaire and that the deadline to return it is May 31, 2011. If one is not received by the deadline, an enumerator will call or visit them to obtain a completed questionnaire. There will also be a notice that refusing to complete a census questionnaire may result in information being referred to the Public Prosecution Service of Canada for further action.

There is no specific Wave 3 in list/leave areas since a second drop-off of a questionnaire by field staff would be too costly.

### 3.2.5 Wave 4

Wave 4 is the traditional non-response follow-up (NRFU). This begins 22 days after Census Day for the mail-out areas and 10 days after Census day for the list/leave - earlier in list/leave areas since there is no Wave 3. The wait period is 22 days after Census Day to ensure that householders have enough time to react to each of the first 3 waves. The start of NRFU should not be delayed any longer than that since it would mean most of the follow-up would occur in the summer months when many households are on holidays. At the start of non-response follow-up, nonresponding householders in mail-out areas may receive a telephone voice broadcast message to remind them of their legal obligation to respond to the census.

### 3.3 De-centralized Registration

When a dwelling returns a questionnaire through the mail, the questionnaire is considered "registered" when Statistics Canada is notified even though the questionnaire may not have been received physically at the DOC. The objective then is to get that registration information to the field staff as quickly as possible so that collection can be stopped on these cases. In 2006, all questionnaires that were returned by mail had to be physically returned to Canada Post in Ottawa. Canada Post would scan the questionnaire's barcode through a window in the envelope. The list of scanned barcodes were written to CD and transported to the data operations centre with the returned questionnaires for registration onto the Statistics Canada's systems. This resulted in a delay of up to a week or more for returns from the extremes of Canada causing increased respondent burden and unnecessary costs contacting households for non-response follow-up after they had actually mailed back the questionnaire. For 2011, a tremendous improvement in timeliness is expected since the mailed-back questionnaires will be scanned in one of sixteen decentralized Canada Post sites across the country. Canada Post will transmit the questionnaire identifiers electronically back to the DOC every hour for registration onto Statistics Canada's systems, and the field will be notified electronically within two hours from that.

### 3.4 Unoccupied and Cancelled dwellings

At any point in time, there are around one million dwellings on the MCS that are unoccupied and about 500,000 dwellings that will subsequently be "cancelled" because they are not considered valid dwellings for Census purposes (they could be businesses, or dwellings still under construction or even burnt down). The goal for the Census is to create a portrait of people living in Canada and the housing stock as of Census Day including the occupancy status. The further from Census Day, the more difficult it is to classify a dwelling as being occupied or unoccupied on Census Day and unnecessary resources can be expended during non-response follow-up when the primary goal is to collect completed questionnaires from occupied dwellings.

For the 2011 Census, a new activity is being introduced to identify occupancy status in mail-out areas as close as possible to Census Day - called Dwelling Occupancy Verification or DOV. Areas will be pre-selected where there is an expectation of large concentrations of unoccupied dwellings. Field staff will focus on these areas during the first two weeks after Census Day so that any unoccupied and cancelled dwellings that are found can be removed from the non-response workload. This will also improve the quality of the identification (a dwelling may be unoccupied on Census Day but occupied by the time non-response follow-up begins).

Also new for 2011 is allowing respondents to declare dwellings as unoccupied or cancelled directly within the online questionnaire (e.g. cottages, $2^{\text {nd }}$ homes, unoccupied rentals, private homes converted to a business, etc.)

### 3.5 Field Management System (FMS)

Activities in the field are impacted by the tools and methodologies discussed in this paper (collection approach, registrations from the 4 response channels, the wave methodology, Dwelling Occupancy Verification, as well as all of the other processes to be conducted in the field and not mentioned here). These must be integrated into a single collection strategy for the people conducting operations in the field as well as for Head Office overseeing the activities. Field collection for 2011 will involve about 35,000 temporary field staff hired for a limited period of time across Canada. The Field Management System (FMS) is the tool that has been designed to specifically address many of the issues that arose during collection in 2006 as well as providing the 35,000 field staff a one-stop shop for conducting and managing their work.

The FMS is a web-based application that will be used by all collection staff, down to the individual enumerator level. Its main goal is to enable timely communication among the field enumerators, field supervisors and Head Office, thus improving the effectiveness and efficiency of the field activities. Communication in 2006 was mainly through faxing and telephone calls and one can imagine what problems can ensue with those vehicles and that volume. For the 2011 Census, notifications of registered questionnaires will be sent to the field through FMS on an hourly basis so that the enumerators will know within $2-3$ hours of a questionnaire being received at one of the 16 decentralized Canada Post sites or submitted through internet. Vice versa, enumerators will inform Head Office on a daily basis which questionnaires they have collected during non-response follow-up so that management will be working with the latest possible return figures. In 2006, this could have taken up to 2 weeks for the field completed questionnaires to be shipped to the data operations centre for subsequent registration. In addition, the FMS will facilitate the creation of work assignment lists for the enumerators for DOV and non-response follow-up, and give them the vehicle for entering their productivity and pay information so that management information numbers to manage workloads and control costs are available in a very timely fashion.

## 4. Final Observations

This paper has presented many of the tools and methodologies that will be utilized to meet the objectives for the 2011 Census in Canada. The overall collection strategy for 2011 will take into consideration the three delivery mechanisms and four response channels. The results from the de-centralized registration of questionnaires that are mailed-back as well as the new DOV activity will be utilized within the new Field Management System (FMS) that will be the core tool used in the field by all personnel including the 35,000 enumerators. A Wave Methodology will be implemented for the first time. It has been designed to encourage response through internet without impacting the workload of the non-response follow-up activities. Sixty percent of the dwellings in Canada will receive a letter instead of a questionnaire to encourage them to go online to complete the questionnaire. Non-respondents from this group will receive a letter reminding them to respond, and if they do not, they will subsequently receive a questionnaire package (if they have not already requested one). The other $20 \%$ of dwellings in mail-out areas will receive a questionnaire package followed by a reminder letter. If they do not respond to the reminder, they may subsequently receive a voice broadcast message by telephone reminding them of their obligations. In List/Leave areas, enumerators will drop-off a questionnaire package at each dwelling. Canada Post will then deliver a reminder card to all dwellings. Non-response follow-up by an enumerator will be the last Wave to garner a response. FMS will facilitate the enumerators work and keep them informed as new registrations continue to be received on which dwellings no longer require follow-up. FMS will also facilitate reporting in a timely fashion on which dwellings are resolved by Field staff so that the workforce can be assigned where needed to complete the census on time. The goal is that at least $40 \%$ of all dwellings will respond to the 2011 Census by internet and never have to receive a paper questionnaire. This will be the majority of dwellings if things go as planned for the 2011 Census.

## References

Hamel, M., Hamilton, G. and Gilmour, G (2008). Changes in the management of data collection operations for the Canadian Census in 2011, Proceedings of Statistics Canada Symposium 2008.

Taylor, J. (2009). 2009 Census Test: Analysis of the supplementary sample, Internal document, Statistics Canada.

Lebrasseur, D., Morin, J.P., Rodrigue, J.F. and Taylor, J. (2010). Evaluation of the innovations implemented in the 2009 Canadian Census Test, paper presented at the Joint Statistical Meetings, 2010.

Dillman, D.A., Smyth, J.D. and Christian, L.M. (2009). Internet, Mail, and Mixed-Mode Surveys, The Tailored Design Method, John Wiley \& Sons.

# The 2009 Census test supplementary study of the new Census collection methodology 

Jennifer Taylor ${ }^{1}$


#### Abstract

A new collection methodology is planned for Canada's 2011 Census of Population in order to increase the response by Internet without increasing the amount of non-response. Under this plan, the majority of dwellings will receive a letter with a unique access code for online completion of the questionnaire. Non-responding households will receive reminders and, if needed, a paper questionnaire. In addition to the sites of the 2009 Census Test, a supplementary sample of 25,000 dwellings was selected across Canada to evaluate the new collection methodology and Internet promotion letter. The study compared two applications of the collection methodology and two versions of the first wave letter. This paper presents an overview of the new collection methodology, a description of the five panels of the sample and the results of the test.


Key Words: Census, Internet, Collection

## 1. Study context

The Canadian Census of Population is held every five years. The last census was held in 2006 and was the first Canadian census to offer the option of Internet response to almost all Canadians. A secure access code (SAC) was provided on each paper questionnaire delivered, allowing the household to access and complete the questionnaire online. Households could also respond by completing the paper questionnaire and mailing it back or by phoning the Census Help Line for assistance. The use of Internet as a response mode was very successful. Among the respondents, $18.3 \%$ completed their census questionnaire online, the best rate obtained so far for a census in any country (Côté and Laroche, 2008). As well, the postal system was used for the first time to deliver census questionnaires to about two thirds of the dwellings ${ }^{2}$. This was made possible using Statistics Canada's address register, a list of addresses that is updated using census results and administrative sources.

It is planned to extend the use of mail delivery of questionnaires and the use of the Internet in the 2011 Census. The mailing of census questionnaires using the postal system allows for efficient, coordinated delivery without having to hire and train a large workforce. Although the costs of developing the Internet application are high, there are many advantages to providing this option. The data obtained by this mode of response have higher item response rates and require less follow-up (Grondin and Laroche, 2008). Built-in functionality helps the respondents to give complete and consistent answers by providing reminders and explanations, and automating skip patterns. The notification of an Internet response is also much faster and data is available much earlier. In 2006 and for the 2009 study, a response on a paper questionnaire was registered only once it arrived in Ottawa, location of the head office of Statistics Canada. In 2011, Canada Post will notify Statistics Canada of questionnaires received at 16 mail processing plants located across Canada. This will reduce the delay in registering a return by as much as four business days for the delivery of questionnaires to Ottawa from these sites. Then there is the time required for the data capture operation which experiences backlogs during census collection. For Internet responses, all of this occurs almost instantaneously and allows for a more current picture of progress and more efficient use of resources.

In preparation for each census, Statistics Canada conducts an operational test of its systems and procedures. The 2009 Census Test offered the opportunity to study the new collection methodology planned for the 2011 Census, as described in the following section, in order to encourage greater use of the Internet response mode.

[^76]
## 2. New collection methodology

During the 2006 Census, a small study was carried out to see the impact of sending a letter with a SAC rather than a paper questionnaire. Such a letter increased the use of the Internet but showed a potential for increasing the amount of non-response (Côté, 2008). In order to both encourage response by Internet while mitigating the risk of decreased overall response, a new collection methodology is planned for 2011. Called the wave methodology, it involves a series of communications, or waves, with households. Households are contacted at key times in order to remind and motivate them to participate in the census. Each wave's communication provides the information needed to respond. About $60 \%$ of dwellings will receive by mail an Internet promotion letter in 2011. The waves for this group are the focus of this paper and are described in the following paragraph. The wave methodology for areas with hand delivery or mail delivery of questionnaires is described by Lebrasseur et al. (2010).

The first wave is an Internet promotion letter that asks respondents to complete the questionnaire online using the SAC provided or to use an automated system accessed by a toll-free number to request that a paper questionnaire be sent in the mail. This wave started one week before Census Test day, May 12, 2009, when letters were put into the mail stream ${ }^{3}$. The second wave consists of a reminder letter sent to all non-responding households. It contains the SAC and the toll-free number, similar to the first wave letter. It was sent starting two days after Census Test day. In the third wave, starting 10 days after Census Test day, non-responding households were sent a paper questionnaire along with a letter requesting that they respond. Households could still respond online using the SAC printed on the questionnaire. The letter mentioned that after May 31, 2009, a census enumerator might contact them to complete the census questionnaire. This fourth wave, the non-response follow-up (NRFU) operation, would have started on June 1,2009 , but was not performed since households were too spread out to use typical census procedures.

## 3. Study objectives and design

The study was designed to observe the impacts of the Internet promotion letter and the wave methodology on the response by Internet and the NRFU workload. It permitted the comparison between two types of deadlines that were considered for the Wave 1 letter: a timeframe deadline ("Within the next 10 days...") or a date deadline ("By Census Test day..."). It also examined an alternate wave methodology whereby the Wave 2 reminder letter was skipped in favour of sending the Wave 3 questionnaire earlier. In this scenario, the questionnaire package was sent starting five days after Census Test day; thus five days earlier than Wave 3. This option, which we called Wave 2.5, allowed more time before the start of NRFU to receive completed paper questionnaires from the respondents. It also saved the costs of printing and mailing the Wave 2 letter which was expected to have a small impact on response.

A sample of 25,000 dwellings was selected across Canada. These dwellings were supplementary to the 110,000 dwellings used for the main 2009 Census Test of the planned systems and procedures which is described in Lebrasseur et al. (2010). The sample was selected from among dwellings receiving census materials through the mail and identified to receive the Internet promotion letter. Areas are targeted to receive the letter based on Internet and total response rate in the last census and characteristics linked to Internet use. The sample was selected randomly, stratified by province, and split evenly into the following five panels:

Panel 1: Wave methodology with Wave 1 timeframe deadline letter
Panel 2: Wave methodology with Wave 1 date deadline letter
Panel 3: Alternate wave methodology with Wave 1 timeframe deadline letter
Panel 4: Alternate wave methodology with Wave 1 date deadline letter
Panel 5: Paper questionnaire followed by a reminder letter
As in the 2006 Census, one in five households was assigned the census long form questionnaire. The fifth panel was used as a benchmark as it was similar to the 2006 methodology, except that the reminder in 2006 was a postcard that
${ }^{3}$ For the 2011 Census, Canada Post will deliver the Wave 1 census material to all households in mail-out areas on the same day. Because of the small volume of the test, such an arrangement was not made and material was delivered according to the regular mail schedule. Delivery typically takes two to four business days.
did not include the SAC. Due to an error, the Wave 2.5 questionnaire package was not sent to non-responding households in the third panel; thus the third panel is not included in the discussion of the results.

## 4. Results

The study obtained higher than expected return rates given that, unlike the census, it was voluntary and there was no publicity campaign. From Figure 4-1 we note that the panels that received the three waves of the original wave methodology had the highest return rate at the end of collection. Panel 1 had $58.2 \%$ of dwellings returning either a paper questionnaire or responding online and Panel 2 had $57.0 \%{ }^{4}$. This was followed by $51.7 \%$ for the benchmark panel and $49.3 \%$ for the alternate wave methodology of Panel 4.

Figure 4-1
Internet and paper return rates by panel on July 30, 2009


### 4.1 Impact on Internet use

The Internet promotion letter coupled with the wave methodology was successful in generating many Internet returns. We note in Figure 4-1 that the proportion of Internet use was much greater for panels initially receiving the letter rather than a questionnaire, as was the case for the benchmark panel. The original wave methodology in Panels 1 and 2 had about $70 \%$ of returns received online compared to $57 \%$ for the alternate wave methodology and $29 \%$ for the benchmark. The alternate wave methodology in Panel 4 was not as successful in generating Internet returns as the three waves used in Panels 1 and 2 which indicates that the Wave 2 reminder letter was beneficial in this regard.

### 4.2 Impact on NRFU

NRFU is an expensive operation involving a large workforce contacting households in person and over the telephone. It is thus imperative to obtain as many responses as possible before its start. However on June 1, 2009, at the start of NRFU, the benchmark panel had a higher return rate than panels subjected to the Internet promotion letter and wave methodology as seen at the NRFU mark in Figure 4.2-1. This figure depicts the total returns by panel as well as the date where we expect to start to see the impact of each wave. The start of the wave's impact does not mean that all households have received the material at this point since material for each wave is prepared over several days and requires varying lengths of time for delivery depending on the location. The planned start date of NRFU is indicated although no NRFU was performed for the study. The benchmark panel had the highest return rate at $44.5 \%$ just before the start of NRFU on May 31, 2009. This is four percentage points greater than Panel 1 ( $40.5 \%$ ) and six greater than Panel $2(38.5 \%)$. The panels subject to a version of the wave methodology received many returns

[^77]after this date, whereas the return rate began to level off for the benchmark panel. This indicates that there is not enough time between Wave 3 and even Wave 2.5 and the start of NRFU to receive the questionnaires and that the wave calendar should be modified. In 2011, we will also experience quicker registration of questionnaires with the help of Canada Post which will give a more up-to-date picture of the collection progress.

## Figure 4.2-1

Total Internet and paper returns by panel


### 4.3 Wave 1 letter

The small difference between Panels 1 and 2 in Figure 4.2-1 is the effect of the version of the Wave 1 letter used. The timeframe letter of Panel 1 is seen to be slightly better than the date letter in terms of return rate except right around Census Test day (May 12) when the date letter experienced higher daily returns. However statistical tests did not conclude that the final return rate of $58.2 \%$ for Panel 1 and $57.0 \%$ for Panel 2 were different. Before the impact of Wave 2, we can combine the timeframe letter panels together (Panels 1 and 3) and the date letter panels together (Panels 2 and 4). With these larger groups, we reject the hypothesis that the rates are equal just before and after Census Test day, on May 9, 10 and 11 and on May 16 and 17. This is corroborated by comments of some participants of qualitative tests who thought that they needed to wait until the specific date before responding, thus the higher number of daily returns around the given date, May 12. Some also felt that once the date had passed it was too late to respond. This corresponds to the plateau that starts around May 15 for Panels 2 and 4 and continues until the Wave 2 and Wave 2.5 in Figure 4.2-1. The returns also slow down for Panel 1, but not to the same extent; for example, Panel 2 goes from $19.1 \%$ on May 15 to $20.0 \%$ on May 18, an increase of less than a percentage point in the return rate, whereas Panel 1, with the timeframe letter, experiences an increase of 2.0 percentage points.

### 4.4 Alternate wave methodology

In Figure 4.2-1, the difference between Panels 2 and 4 is the difference between receiving three waves and receiving the two waves of the alternate wave methodology. Starting with the impact of the Wave 2 reminder letter, Panel 2 begins to record higher returns and Panel 4 never catches up, finishing nearly eight percentage points behind. The Wave 2 reminder letter thus plays an important role in increasing the return rate. Sending the questionnaire package earlier at Wave 2.5 did result in slightly fewer returns after the start of NRFU for Panel 4 compared to Panel 2, but this was a small benefit compared to the lower overall return rate.

Figure 4.4-1, Total Internet returns by panel, shows that the Wave 2 reminder letter generated a lot of Internet returns before the start of the Wave 3 questionnaire impact. Internet returns continued to be received after the start of NRFU for Panel 2. The questionnaire package of Wave 2.5 does generate Internet returns as seen by the increase in Panel 4 some time after its receipt, but not nearly as many. At the end of collection, the Internet return rate for Panel 4 was 12 percentage points lower than that of Panel 2. The opposite is true of paper questionnaire returns. The earlier receipt of the questionnaire package in Panel 4 meant that responses shifted from Internet to paper earlier than for Panel 2 as seen in Figure 4.4-2. More paper questionnaires were returned for Panel 4, but this is not large enough to compensate for the lower rate of Internet returns.

Figure 4.4-1
Total Internet returns by panel


Figure 4.4-2
Total paper returns by panel ${ }^{5}$

${ }^{5}$ The start for the wave impacts are adjusted by two days compared to the previous charts to partially reflect the delay in registering a paper return.

### 4.5 Other observations

Prior to Wave 2.5 for Panel 4 and Wave 3 for Panels 1 and 2, there were paper returns for panels receiving the Internet promotion letter as seen in Figure 4.4-2. These are cases where the respondent requested a paper questionnaire using the automated telephone system. Although most waited rather than actively requesting a questionnaire, we expect that the publicity and mandatory status of the census will result in more extensive use of this option in 2011.

As seen in Figure 4.2-1, before the impact of the Wave 2 reminder letter, the number of returns levelled off for Panel 2 and slowed down for Panel 1. This indicates that the reminder letter could be sent earlier and that Wave 2.5 is too late. Furthermore, Wave 2 had a definite impact for Panels 1 and 2 and the benchmark as we started to obtain more returns after that point. Comparing Figures $4.4-1$ and $4.4-2$, we note that the mode of response after Wave 2 depended heavily on the options readily available to the household. For Panels 1 and 2, the impact was primarily Internet returns whereas the impact for the benchmark sample was an increase in both Internet and paper returns, but with more paper returns. This is similar to the effect of the questionnaire packages sent to Panels 1,2 and 4 . Once a paper questionnaire was provided, it became the more dominant mode of response.

## 5. Towards the 2011 Census

Since the 2009 Census Test, a change has been made to the census. The collection of the long form questionnaire data has been removed from the 2011 Census and will instead be collected by the voluntary National Household Survey. The results of the study still apply to the census and can be adapted to the context of the new survey to assist in its planning as well.

The study showed that the Internet promotion letter along with the wave methodology is successful in encouraging response by Internet. The Wave 1 Internet promotion letter with the timeframe deadline will be used in the 2011 Census because it resulted in more returns overall and did not experience the same slow down in returns after Census Test day as was observed for the date letter. The Wave 2 reminder letter and Wave 3 questionnaire package will be sent to non-responding households since this combination resulted in the highest return rate. The study showed, however, the potential for a greater workload at the start of NRFU compared to initially sending a paper questionnaire as was done in 2006. In an effort to receive more responses before the start of NRFU, the Wave 3 questionnaire will be sent out earlier in 2011; the Wave 3 will start at 8 rather than 10 days after Census Day. Since the returns started to slow down before the Wave 2 reminder letter was received by households, Wave 2 will be moved earlier in 2011 to start on Census Day so as to keep the momentum in returns.

The study also allowed for refinements to be made to the planning assumptions for 2011. It indicated that fewer households than first expected will request a paper questionnaire using the automated telephone system. We learned that the Wave 2 reminder letter has a significant impact on the return rates and that the Internet promotion letter, when sent to the right areas, can yield a higher percentage of online responses than originally planned.

## Acknowledgments

The author wishes to thank Mario Ghafari, Limei Sun and Lucie Veilleux for their work on the study and Laurent Roy and Jean-François Rodrigue for their comments on the paper.

## References

Côté, A.M. and Laroche, D. (2008). The Internet: A New Collection Method for the Census, Proceedings of Statistics Canada Symposium 2008 on Data Collection: Challenges, Achievements and New Directions.

Grondin, C. and Laroche , D. (2008). Impact of Online Edits and Internet Features in the 2006 Canadian Census, Working Paper No. 5 of the Work Session on Statistical Data Editing of the Conference of European Statisticians, United Nations Statistical Commission and Economic Commission for Europe, Vienna, Austria.

Lebrasseur, D., Morin, J.P., Rodrigue, J.F. and Taylor, J. (2010). Evaluation of the innovations implemented in the 2009 Canadian Census Test, Proceedings of the Survey Research Methods Section of the Joint Statistical Meetings, American Statistical Association, pp. 4089-4097.

Taylor, J. (2009). 2009 Census Test: Analysis of the Supplementary Sample, unpublished report, Ottawa, Canada: Statistics Canada.

## SESSION 8C

## LINKAGE METHODS

# An Empirical comparison of approaches to approximate string matching in private record linkage 

Tobias Bachteler, Rainer Schnell, and Jörg Reiher ${ }^{1}$


#### Abstract

Due to the frequency of spelling and typographical errors in practical applications, record linkage algorithms have to use string similarity functions. In many legal contexts, identifiers such as names have to be encrypted before a record linkage can be attempted. Therefore, algorithms for computing string similarity functions with encrypted identifiers are essential for approximating string matching in private record linkage. This study reports an empirical evaluation of three promising approaches to compute similarities between identifiers in a privacy preserving manner.


Key Words: Approximate String Matching; Privacy-preserving Record Linkage; Private Record Linkage

## 1. Introduction

Combining multiple databases with additional information on the same person is increasingly occurring throughout research. Whenever feasible, the databases are merged using a unique identification number. Otherwise, probabilistic record linkage is most frequently applied for the identification of matching record pairs. However, in many applications the identifiers have to be encrypted due to privacy concerns, which is problematic because linking encrypted identifiers can result in serious complications. The problem of finding records that represent the same individual in separate databases without revealing the identity of the individuals is called the "privacy-preserving record linkage", "blind data linkage", or "private record linkage" problem.

Due to the frequency of spelling and typographical errors, in practical applications record linkage algorithms have to use string similarity functions. In addition, since records with variations of identifiers may have different characteristics than records with exact matching identifiers, restricting the linkage in this manner is not an option. Therefore, algorithms for computing string similarity functions with encrypted identifiers are essential for approximate string matching in private record linkage. This study reports an empirical evaluation of three promising approaches to compute similarities of identifiers in a privacy preserving manner. The performances of these protocols were compared with each other and those of an edit-distance applied to unencrypted identifiers.

### 1.1 Problem Statement

The problem of how to accomplish approximate string matching in private record linkage can be compactly stated as follows: Consider two database owners $A$ and $B$ holding lists of strings (e. g. names) $S_{a}=\left\{a_{1}, \ldots, a_{n}\right\}$ and $S_{b}=\left\{b_{1}, \ldots, b_{m}\right\}$. Compute string similarities of all pairs $\left(a_{i}, b_{j}\right) \in S_{a} \times S_{b}$ such that all the names remain private to the database holders.

### 1.2 Candidate Solutions

Since the early 1990s several intriguing approaches to approximate string matching in private record linkage were proposed. Most of these (e.g. Bouzelat et al., 1996, Atallah et al., 2003, Ravikumar et al., 2004, and Churches and Christen, 2004) face problems, for instance very high computing demands or high rates of false positive links. These

[^78]problems make them unfeasible in realistic record linkage settings. There are however three approaches to compute similarities of identifiers in a privacy preserving manner, which prima facie are promising candidate solutions to the problem, namely the protocols proposed by Pang and Hansen (2006), Scannapieco et al. (2007), and Schnell et al. (2009).

Pang and Hansen (2006) suggested a protocol based on a table of reference strings, $R$, common to $A$ and $B$. The elements of $R$ should reflect the domain of the strings to be matched. For a given identifier, both database holders compute the edit-distances, $d$, between each identifier string and all reference strings in the set. If $d$ is equal or less than a threshold $\delta$, the respective reference string is encrypted using a key previously agreed on by $A$ and $B$. For each identifier string, the resulting set of encrypted reference strings along with their distances, $d$, and an ID number form a tuple. Both database holders send their tuples to a third party $C$. For every pair of ID numbers where the encrypted reference strings agree, $C$ sums the distances, $d$, and finds the minimum of this sum, which serves as an approximation to the distance between the plain text identifiers.

In the protocol of Scannapieco et al. (2007), two data holders, holding lists of names, build an embedding space from random strings and embed their respective strings therein using the SparseMap method (Hristescu and FarachColton, 1999). Then, each data holder sends the embedded strings to a third party who determines their similarity. To create the embedding space, data holder $A$ generates $n$ random strings of length $l$ and builds $k$ reference sets from them. These $k$ reference sets are used to embed the names in a $k$-dimensional space. The coordinates for a given name are approximations of the distances between the name to the closest random string in each of the $k$ reference sets in terms of the edit-distance. As a result, for each name $A$ receives a $k$-dimensional vector. After receiving the $k$ reference sets from $A, B$ embeds his names in the same way. Finally, both data holders send their vectors to a third party, $C$, who compares them using the Euclidean distance between them.
Schnell et al. (2009) proposed to take advantage of the cryptographic properties of Bloom filters (Bloom, 1970) to determine string similarities in a privacy preserving manner. A Bloom filter is a bit array with all bits initially set to 0 . To store a string in a Bloom filter, each of its constituent $n$-grams are hashed using $k$ HMACs (Bellare et al., 1996) and all bit positions that correspond to a hash value are set to 1 . The Bloom filters of two similar strings will share a good proportion of bit positions set to 1 . First, data holders $A$ and $B$ store each name in a separate Bloom filter and transfer them to a third party $C$. Then, $C$ is able to approximate the $n$-gram similarity of the original names by computing the Dice coefficient $D=2 h /(m+n)$ where $h$ is the number of bit positions conjointly set to 1 in both filters, $m$ is the number of bit positions set to 1 in the first and $b$ the number of bit positions set to 1 in the second Bloom filter.

## 2. Methods

### 2.1 Test Data

For empirical testing we used experimentally generated human data. 1,500 surnames were randomly selected from a town register. Then, 15 audio tapes with 100 different names each were recorded with a female voice. Each tape was played to one of 15 groups of students with an average of 19 students per group. Each student wrote the name heard, 13 groups by hand, two groups typed the names. Finally, all student notes were typed by a single typist resulting in 1,286 unique original names (file A) and 27,296 potentially erroneous names (file B).

To compare the three candidate solutions to the problem of approximate string matching in private record linkage we determined the similarity of each pair of names $(a, b) \in A \times B$ using each method in turn and determined their performance by examining precision-recall plots. As an external standard of comparison we determined the Damerau-Levenshtein distance from the plain text names.

### 2.2 Precision-Recall Plots

As is the norm in the information retrieval literature, the criteria of recall and precision were used to determine the linking effectiveness of the method. For a given level of similarity $\varphi$, a pair of records is considered as a match if the pair is actually a true pair, all other pairs are called non-matches. Based on the common classification for true positive $(T P)$, false positive $(F P)$, false negative $(F N)$ and true negative $(T N)$ pairs, the comparison criteria are defined as

$$
\begin{align*}
\text { recall } & =\frac{\sum T P}{\sum T P+\sum F N}  \tag{1}\\
\text { precision } & =\frac{\sum T P}{\sum T P+\sum F P} \tag{2}
\end{align*}
$$

Plotting precision and recall for different similarity values $\varphi$ as a curve in a precision-recall-plot shows the performance of a string comparison method. A procedure with a better performance will have a curve in the upper right of the plot.

## 3. Results

### 3.1 Protocol 1 (Pang and Hansen)

To evaluate the protocol of Pang and Hansen (2006) we varied two parameters: the threshold $\delta$ and the size of the reference table $R$. To obtain suitable reference strings, we sampled surnames from a German phone book.

Figure 3.1-1
Protocol 1: Precision-recall curves using thresholds 2 and 3 with sample sizes 5,000 and 10,000


Figure 3.1-2
Protocol 1: Precision-recall curves using threshold 2 and various sample sizes


Figure 3.1-1 displays precision and recall for reference tables of size 5,000 and 10,000 with $\delta=2$ and $\delta=3$ respectively. Due to protocol design, there are just a few precision-recall points obtainable, which are marked by dots and connected by dashed lines. All curves indicate that the respective parameter settings are performing fairly poor. Starting at a recall of virtually null, precision drops abruptly and reaches a low level of precision before recall starts to rise significantly. That is, the protocol is not able to differentiate between matching and not matching pairs of names at all. This finding is reinforced by the huge gap between the precision-recall points and the reference line of the Damerau-Levenshtein distance.

As figure 3.1-2 shows, increasing the size of the reference table improves the situation to some extent. However, as the sample size gets larger, the additional gain diminishes significantly. Choosing a size beyond 50,000 does not improve the situation further. The use of $\delta \geq 3$ along with the increased sample sizes is impractical since this leads to very long run times. In all, the best performance is obtained with the parameter combination $\delta=2$ and $R=50,000$.

### 3.2 Protocol 2 (Scannapieco et al.)

In the protocol of Scannapieco et al. (2007), two parameters are of interest, namely the length of the random strings $l$ and the number of dimensions $k$. As a starting point we used recommendations given by Scannapieco et al. They suggested a dimensionality of about 20 and string length corresponding to length of input names, which on average is about 7 in our test data base. Because there is randomness built in the protocol via the random strings in the reference sets, we replicated each individual test ten times. The precision-recall curves show the of average recall at 10 fixed precision values. Minimum and maximum recall are marked by caps at each precision value.

Figure 3.2-1
Protocol 2: Precision-recall curves using 25 dimensions and various string lengths


Figure 3.2-2
Protocol 2: Precision-recall curves using string length 20 and various numbers of dimensions


Figure 3.2-1 shows precision and recall depending on different string lengths (number of dimensions fixed at $k=25$ ). All curves have some distance from the reference line indicating a rather poor performance of the parameter settings. However, contrary to the expectations, the string lengths 10 and 20 do definitely better than a string length of 7 .

In a second experiment, we varied the number of dimensions, $k$, at a fixed string length of 20. As shown in Figure 3.2-2, using more dimensions generally leads to improved performances of the protocol. However, as can be seen from the distances between the individual curves, the additional gain attained is diminishing with increasing numbers of dimensions. As the best performing parameter combination we identified a string length of $l=20$ along with a dimensionality of $k=64$.

### 3.3 Protocol 3 (Schnell et al.)

For an evaluation of the protocol of Schnell et al. (2009), the parameter of main interest is the number of hash functions, $k$. We compared performances of different $n$-gram variants. In all experiments, the lengths of the bit arrays were fixed at 1,000 bits. In Figure 3.3-1, the precision-recall curves resulting from 2-grams and different numbers of hash functions are displayed. The best performance is attained with 10 hash functions. Although increasing this number generally results in inferior performances, even the curve resulting from 70 hash functions indicates an acceptable result.

Figure 3.3-1
Protocol 3: Precision-recall curves using 2-grams and various numbers of hash functions


Figure 3.3-2
Protocol 3: Precision-recall curves using 50 hash functions and various variants of $\boldsymbol{n}$-grams


Figure 3.3-2 shows the precision-recall curves of 2-, 3-, and 4-grams with $k$ fixed at 50 hash functions. The best performing variant are 2 -grams, followed by 3 -grams. In all, the best of the evaluated parameter combinations in terms of precision and recall is 2-grams with $k=10$. However, since the use of more hash functions provides more security (Schnell et al., 2009), we selected 2-grams with $k=50$ for the inter-protocol comparison.

### 3.4 Comparison

For comparison we used the best performing variants of Pang and Hansen and Scannapieco et al. along with the 2grams, $k=50$ variant of Schnell et al. The primary comparison criterion is the effectiveness of the protocols in terms of precision and recall. Additionally, we compared the run times.

Figure 3.4-1
Comparison of precision and recall


As shown in Figure 3.4-1, the Bloom filter method outperforms the protocol of Scannapieco et al. and both outperform the protocol of Pang and Hansen substantially. The small gap between the curves of the Bloom filter method and Damerau-Levenshtein indicates that the Bloom filter method performs quite well compared with a similarity function applied to unencrypted strings.

The run times of the various methods are given by table 3.4-1. The protocol of Pang and Hansen requires substantial run times both to encrypt and compare the names. The total run time of nearly 16 hours implies that the protocol will be virtually impractical for most matching tasks. In contrast, the run times of the Scannapieco method and the Bloom filter method are quite reasonable. Due to very efficient encryption, the Bloom filter method did not require substantially more time to compare the names than Damerau-Levenshtein.

Table 3.4-1
Comparison of run times (minutes)

| Method | Encryption | Matching | Total |
| ---: | ---: | ---: | ---: |
| Damerau-Levenshtein | - | $47: 15$ | $47: 15$ |
| Bloom | $0: 04$ | $52: 56$ | $53: 00$ |
| Scannapieco (means) | $19: 25$ | $65: 25$ | $84: 50$ |
| Pang and Hansen | $470: 14$ | $494: 13$ | $964: 27$ |

## 4. Conclusions

In this article, we report the results of an empirical evaluation of three promising approaches to compute similarities between identifiers in a privacy preserving manner. For empirical testing we used experimentally generated human data. We identified the best parameter settings of the three methods. The performance of these methods was compared. The protocol proposed by Pang and Hansen shows poor precision and recall and requires long run times. The protocol of Scannapieco et al. yielded good linkage results and a moderate run time. The protocol proposed by Schnell et al. showed the best linkage results and lowest run time.

## References

Atallah, M. J., F. Kerschbaum, and W. Du (2003), "Secure and Private Sequence Comparisons", Proceedings of the ACM Workshop on Privacy in the Electronic Society, pp. 39-44.

Bellare, M., R. Canetti, and H. Krawczyk (1996), "Keying Hash Functions for Message Authentication", Proceedings of the 16th Annual International Cryptology Conference, pp. 1-15.

Bloom, B. H. (1970), "Space/time Trade-offs in Hash Coding with Allowable Errors", Communications of the ACM, 13, pp. 422-426.

Bouzelat, H., C. Quantin, and L. Dusserre (1996), "Extraction and Anonymity Protocol of Medical File", Proceedings of the 1996 AMIA Annual Fall Symposium, pp. 323-327.

Churches, T. and P. Christen (2004), "Some Methods for Blindfolded Record Linkage", BMC Medical Informatics and Decision Making, 4.

Hristescu, G., and M. Farach-Colton (1999), "Cluster-preserving Embedding of Proteins", DIMACS Technical Report 99-50, DIMACS Center for Discrete Mathematics \& Theoretical Computer Science.

Pang, C., and D. Hansen (2006), "Improved Record Linkage for Encrypted Identifying Data", Proceedings of the 14th Annual Health Informatics Conference, pp. 164-168.

Ravikumar, P., W. W. Cohen, and S. E. Fienberg (2004), "A Secure Protocol for Computing String Distance Metrics", paper presented at the Workshop on Privacy and Security Aspects of Data Mining, Brighton, England.

Scannapieco, M., I. Figotin, E. Bertino, and A. K. Elmagarmid (1996), "Privacy Preserving Schema and Data Matching", Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 653664.

Schnell, R., T. Bachteler, and J. Reiher (2009), "Privacy-preserving Record Linkage using Bloom Filters", BMC Medical Informatics and Decision Making, 9.

# G-LINK: Constructing an Avatar 

Antoine Chevrette ${ }^{1}$


#### Abstract

At Statistics Canada, matching data without unique identifiers is a common practice. The probabilistic record linkage method developed by Ivan Fellegi and Allan Sunter ${ }^{2}$ is the primary method recommended by Statistics Canada for this type of matching. In recent decades, work began to generalize the Fellegi-Sunter algorithm in order to offer our community the opportunity touse this methodology within a computer application. The most recent version of this application is called G-LINK and is part of Statistics Canada's package of generalized systems. By definition, a generalized system must be user-friendly, robust, fast, highly flexible and responsive to user demands. It will be interesting to discover from reading this article how it was possible to meet these criteria using the latest user interface and development technologies.


## Introduction

To fully meet the challenges involved in making the transition from mathematical theory to computer application, it is first necessary to lay the theoretical groundwork. To do this, a general review of the Fellegi-Sunter algorithm will be provided to show the algorithm in relation to its computer avatar G-LINK.

Certain critical methodological elements pose a challenge with respect to both technology and the user interface, and these will be examined in detail. The solution used in G-LINK for each of those challenges will also be described in detail.

Finally, there will be a discussion of future computational developments in G-LINK to carry it to even greater heights.

## Methodological review

What is probabilistic linkage? This is a linkage with no unique identifier, for which we estimate the likelihood that the records correspond to the same entity. From this definition, it can easily be inferred that there is no point in applying probabilistic linkage to data that have a unique identifier (e.g. a social insurance number); this type of matching can easily be done using direct linkage.

The purpose of G-LINK lies in its ability to provide both internal and external probabilistic linkage. Internal linkage is used to find groups of records that refer to the same entity within the same file, while external linkage is used to find groups of records in two different files referring to the same entity.

Probabilistic record linkage generally consists of six separate steps:

[^79]1. Generate potential pairs using a selection criterion
2. Generate rules and apply them to the potential pairs to derive probability ratios
3. Assign a status to the pairs using the probability ratios
4. Apply frequency probabilities (weighting factors)
5. Form groups
6. Resolve conflicts using mapping.

## Generate potential pairs

The sole purpose of this step is to limit the number of potential pairs generated. In a perfect world, it would be essential to consider the set of all possible pairs. For example, in a linkage of two files, if the first file contains 50,000 records and the second contains 20,000 records, then by taking the Cartesian product of these two files we obtain $20,000 \times 50,000=1,000,000,000$ possible pairs to evaluate.

Unfortunately, since this is not a perfect world, and since the computers available to us cannot yet be said to have Herculean power, it is unimaginable to generate all possible pairs resulting from a Cartesian product of files of even modest size. Just imagine a Cartesian product of 30 million times 30 million.

To get around this problem, it is crucial to use the first step of the algorithm, which is to generate potential pairs using a selection criterion based on direct mapping. A user might decide to generate all potential pairs resulting from mapping in which the first three characters of the postal code match and the sex is the same.

## Generate and apply rules

How do we determine whether a pair corresponds to the same entity? To do this, we need to be able to measure the level of agreement in the information characterizing the pairs. This level of agreement is actually a probability ratio that is assigned to the pair by means of rules and levels of agreement.

Predefined character, numerical and date-type rules exist within G-LINK. A rule can be constructed on the basis of several types of comparisons called levels of agreement.

To clarify the concept of rules, there is nothing better than a concrete example.

| TABLE A |  |  | TABLE B |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Surname | Given <br> Name | Birth Year | Surname | Given <br> Name | Birth Year |
| SMITH | SUSAN | 1940 | SMITH | S | 1939 |

(Table 1)
To determine whether the records in Table A and Table B represent the same person (Table 1), it is advantageous to use different rules. For this example, use of a character-type rule on the given name and surname and a date-type rule on the year of birth will serve the purpose.

| Rule | Outcome level |  | Result | Comparison |
| :--- | :--- | :--- | :--- | :--- |
| Surname | 1 | Complete match | Agreement | SMITH=SMITH |
|  | 2 | Partial match (Nysiis) | Not evaluated because the <br> preceding outcome level <br> was true | NA |
|  | 1 | Complete match | Disagreement | SUSAN $\neq$ S |
|  | 2 | Partial match (first character) | Agreement | S=S |
| Birth Year | 1 | Complete match | Disagreement | $1940 \neq 1939$ |
|  | 2 | Partial match <br> (Year minus one) | Agreement | $1940-1=1939$ |

(Table 2)

From Table 2, it is easy to see that the first level of the Surname rule is in agreement and that the second level of the Given Name rule and the Birth Year rule are also in agreement.

An outcome level can yield only three different values: an agreement, a disagreement or a missing (value missing on one or both of the sides compared). When a probability is associated with an outcome level, its probability ratio can be calculated.

How do we assign a probability? It is first necessary to give the probability that the outcome level is true when the pair belongs to the set of related pairs (and is thus a good pair). It is also necessary to give the probability that the outcome level is true knowing that the pair belongs to the set of unrelated pairs (and is thus a bad pair).

Using the rules in Table 2, we have

| Outcome level probabilities |  |  |  |  |  |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Rule: | Surname |  |  |  | Given Name |  |  |  | Birth Year |  |  |  |
|  | 1 | 2 | M | D | 1 | 2 | M | D | 1 | 2 | M | D |
| Linked sets | $\begin{aligned} & .70 \\ & .01 \end{aligned}$ | $\begin{aligned} & .20 \\ & .04 \end{aligned}$ | $\begin{aligned} & .05 \\ & .05 \end{aligned}$ | $\begin{aligned} & .05 \\ & .90 \end{aligned}$ | .80.02 | .10.05 | .05.05 | .05.88 | .87.01 | .08.06 | . 02 | . 03 |
| Non-linked sets |  |  |  |  |  |  |  |  |  |  | . 02 | . 91 |
| Prob. ratio | 70.00 | 5.00 | 1.00 | . 06 | 40.00 | 2.00 | 1.00 | . 06 | 87.00 | 1.33 | 1.00 | . 03 |

(Table 3)
For example, level 1 of the Surname rule (complete comparison) is true $70 \%$ of the time when the pair belongs to the set of linked pairs, while it is true $1 \%$ of the time when the pair belongs to the set of non-linked pairs. The probability ratio is calculated by dividing the probability for the linked set by the probability for the non-linked sets. Below is the mathematical formula describing the probability that the records will be linked for a particular outcome level (probability ratio):

$$
P R_{i}\left[r_{i}(a, b)\right]=\frac{P\left(r_{i}(a, b) \mid(a, b) \in L\right)}{P\left(r_{i}(a, b) \mid(a, b) \in N\right)}
$$

It should be noted that the higher the probability ratio, the greater the probability of having a linked pair; the inverse is also true.

To find the probability ratio of the pair and not of the rule, we need only multiply together the probability ratios of the rules. However, to be valid, this multiplication requires that the rules used be independent.

$$
P R(R)=P R\left(r_{1}\right) \times P R\left(r_{2}\right) \times \ldots P R\left(r_{n}\right)
$$

| Pair |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Rule |  |  |  |
|  | Surname | Given Name | Birth Year |  |
| Outcome level | C (Agreement) | PA (Partial agreement) | PA (Partial agreement) |  |
|  |  |  |  |  |
|  |  | ilities |  |  |
| Linked set | 0.7 | 0.1 | 0.08 |  |
| Non-linked set | 0.01 | 0.05 | 0.06 |  |
| Probability ratio | 70.00 | 2.00 | 1.33 | PR: 186.67 |

(Table 4)
In our example, the probability ratio for the pair is 186.67 . Thus, this pair has 186.7 times the chance of being in the set of linked pairs than of being in the set of non-linked pairs.

## Assign a status to pairs

The status of a pair is a value that is assigned to it in order to categorize it. A pair can be categorized as definitive (good pair), possible (to be considered) and rejected (not considered). There is also a fourth status, "excluded," which is used in applying the rules. If in that process, the probability ratio fails to reach a certain value (called the cut-off threshold), the pair is excluded and becomes isolated from the rest of the process.

In order to assign a status to pairs, thresholds are needed. A lower threshold and an upper threshold are used:

- $T_{i}$ Lower threshold
- $T_{s}$ Upper threshold

The pairs are initialized as follows:

- Weight $(\mathrm{a}, \mathrm{b})<T_{i} \quad$ Status $=\mathrm{R}$ (Rejected)
- $\quad T_{i}<=$ Weight $(\mathrm{a}, \mathrm{b})<T_{s} \quad$ Status $=\mathrm{P}($ Possible $)$
- Weight $(\mathrm{a}, \mathrm{b})>=T_{s} \quad$ Status $=\mathrm{D}($ Definitive $)$


## Apply frequency probabilities

To refine the probability ratios, frequency weights can be applied for all outcome levels of a rule that has a result. A non-linked frequency weight will replace the non-linked portion of the outcome level weight. A linked frequency weight will replace the linked portion of the outcome level weight. Finally, a frequency weight having both linked and non-linked portions will replace the corresponding two components of the outcome level weight.

Non-linked frequency weights can be calculated using the input tables (Table A and Table B). For example, say that we have agreement on the surname Smith and agreement on the surname Aardvark. Since Smith is a much more common value than Aardvark, it is more useful to refine the weights by assigning more importance to agreement on Aardvark (a rarer value). Below is the formula used to calculate this type of frequency weight:

The frequency weights can also be calculated based on the outcome levels (linked frequency weights). For example, say that an outcome level uses the similarity algorithm (developed by William Winkler) returning as a value a percentage of exactness between two strings of characters. More weight should be assigned to rare values ( $99 \%$ ) and less weight to common values $(90 \%)$. The formula used to calculate this type of weight is similar to the previous formula:

$$
\text { LFreqWeight }_{i}=10 * \log _{2}\left[\frac{\text { NumbLinkedPairs }_{i}}{\text { NumbLinkedPairs }}\right]
$$

It is also possible to calculate and apply frequency weights based on the outcome level results but also on the nonlinked portion of the pairs. This non-linked portion, commonly known as the random portion, is constructed by randomly selecting pairs from Table A and Table B.

$$
\text { NFreqWeight }_{i}=-10 * \log _{2}\left[\frac{\text { NumbNonlinkedPairs }_{i}}{\text { NumbNonlinkedPairs }}\right]
$$

## Form groups

Why is it necessary to form groups within pairs? The answer is simple: a number of pairs may be inter-related among themselves. An example of this is where record 1 in Table A is related to record 4 in Table B, but record 4 in Table B is also related to record 10 in Table A, etc.

More specifically, groups are generated from pairs according to their status. There are two types of groups: weak groups and strong groups. Weak groups are made up of records with possible and definitive links. Strong groups are internal to weak groups and contain only records linked together by definitive links.

In summary, this step in the process serves to organize pairs in such a way as to be able to easily resolve conflicts by mapping.

## Resolve conflicts using mapping

Conflict resolution is the last step before exporting the linkage results. It can happen that the returned pairs referring to the same entity are multiple. Mapping is used to specify the type of results sought in record matching. There are four types of mapping: multiple to multiple (output of the creation of groups), one to multiple, multiple to one and one to one.

## Challenges

The above brief review of the methodology will make it easier to understand the challenges that arose in constructing the computer avatar of the Fellegi-Sunter algorithm. Don't worry; no advanced knowledge of computer science is required in order to understand the following sections.

## Synthesizing the information (using a status tree)

A generalized system must be clear and user-friendly; above all, it must not cause the user to feel confused. The methodology review presented the six main steps in the algorithm used, and it presented them in a specific order. That order is not required when creating and implementing a project using G-LINK. For example, a user could first calculate the frequency weights based on the tables (non-linked), then formulate the rules and create potential pairs. Moreover, at any stage in the project, a user could decide to re-apply the previous steps. Why does G-LINK offer this
flexibility? Because as a generalized system, it must meet the needs of all its users, and depending on the project, the order of operations may vary.

What is a clear and straightforward way to present the information on the stages of a project? By showing the steps in tree form. The use of a tree (Figure 1) provides users with a visual cue enabling them to determine instantly what stage their project is at and to determine everything that they can or cannot do. This way of presenting the information enables even novice users to intuitively know and apply the steps involved in record linkage.

(Figure 1)
The tree is composed of nodes, with each node representing a step in the algorithm. The nodes can be expanded to show the sub-steps. Each node has an image attached to it. There are three images: a check mark indicating that the step has been completed; a runner, indicating that the step can be carried out; and a pair of feet, indicating that the step cannot be carried out at this time (since it depends on a later step).

The internal architecture of the system lends itself to possible changes to paths and allows the programmers of the software to add nodes. If a step had to be added in the future, this would have only a minor impact on implementation.

## Accessing the data

Accessing data outside the G-LINK system presented the development team with two completely different challenges: importing the data and displaying them.

If users cannot import the data to compare in G-LINK, they will be unable to use the software, and this is unacceptable for a generalized system. Since the majority of our users use SAS, one of the importation priorities is the importing of SAS files. So how can SAS data be imported in conjunction with Visual Basic.net (the technology used to develop the software interfaces)?

By chance, my previous work experience has led me to examine this quite interesting question. If you turn to the article 'SAS® Integration Technologies, UNIX and Visual Basic .Net Integration Procedure,' ${ }^{3}$ you will find the solution to the problem of combining the two technologies. In brief, the interface (Visual Basic .net) uses the built-in

[^80]features of SAS to create an SAS session invisible to the user to access import data. It should be noted that to use this option, SAS must be installed on the user's computer. G-LINK also offers the possibility of importing data in the format of a space-delimited text file (flat).

After data from an SAS file or a flat file are imported, the user's first impulse will be to check that they are the right data. G-LINK therefore provides a tool for viewing the data. This simple idea may seem trivial, but it isn't.

How can a table containing millions of records be displayed on the screen? The first attempt was simply to load the table into memory, and the result was a monumental failure caused by a lack of memory in the systems. A voluminous table can extend over several GB of spaces, and this often exceeds the total memory available in our operating systems. So what is the solution to this problem?

As it happens, Microsoft has developed a technique called Just In Time Data Loading, whereby only the records that are to be displayed on the screen are kept in memory. The scroll bar of the active window directly controls the data refresh process - for example, 100 records at a time. Obviously, loading 100 records into memory is much more stable than loading millions of records!

## Query to generate initial pairs with indexation

The task of generating the pairs is automated, but generating the query is not. The pleasure of constructing this query falls to the user, and a query to generate pairs can range anywhere from a trivial equality to a complex, multi-level query.

Since G-LINK is a generalized system, one of its main features is user-friendliness. The interface for generating the pairs query was built to meet this criterion. Accordingly, a user with no knowledge of PL-SQL can generate the PLSLQ query with a few mouse clicks. The interface was built (Figure 6) to allow automatic selection of the operators and fields used for comparison. The interface also allows automatic generation of sub-strings. However, is the user knows the P-SQL language, he or she can simply edit or generate a query in the box designed for this purpose.


## Speeding up the process: divide and conquer

It quickly became clear that this process was inefficient for large-scale projects. The "normal" architecture used to create modest- to medium-sized projects cannot be applied to large projects. To give an idea of size, a project is considered "large" when it involves millions of records and pairs. For example, a project that has 20 million records
for Table A and Table B and generates some 300 million possible pairs is considered large. The G-LINK team therefore looked into the problem and found a solution to this complex challenge. The old saying "divide and conquer" was the perfect approach to developing that solution.

Normally, all the data are imported into a Table A and a Table B. The more information there is in a table, the longer the time required for search and comparison. It was therefore decided to give the user the choice of separating the input data into a number of sub-tables. Of course, the task of separating a table is automatically taken on by the system; the user need only tell the system how much data a sub-table should contain.

Potential pairs are then generated using the sub-tables, and the rules are applied to these potential pairs. This technique eliminates the bother of managing voluminous tables.

For example, say that we have a Table A containing 20 million records and a Table B containing 4 million records. If it is decided to separate the tables into sub-tables of 250,000 records, there will be 80 sub-tables representing Table A and 16 sub-tables representing Table B. The process of generating pairs and applying the rules will have to be repeated 480 times ( $80 * 16$ ).

When Table A and Table B are separated into a number of sub-tables, an element of table independence is introduced, allowing the processes of pair generation and rule application to take place simultaneously in the different sub-tables. Thus, if G-LINK is executed on a server with a number of CPUs, several table combinations can be evaluated at the same time. For example, if eight CPUs are used, the system can execute eight pair generation/rule application processes simultaneously. Processing time will be reduced eightfold.

Finally, the user can choose not to retain excluded pairs (pairs that do not reach a predetermined threshold). This means that the application has the potential not to keep superfluous information and therefore it can, once again, avoid having to manage voluminous tables.

## Conclusion

This article shows how the world of methodology and the world of computers have been brought together in the GLINK system. Merely by reading this document, one can gain a good understanding of the methodology and the system. As you have seen, simple methodological ideas can become formidable challenges when it comes to implementing them on a computer.

Currently, G-LINK reproduces the methodology of the Fellegi-Sunter algorithm. Over time, however, other highly useful features will be added to the application.

Batch processing is an important feature that should become available in March 2011. This will make it possible to execute G-LINK from a command line by providing it with an XML configuration file. Accordingly, it will be possible to create and execute a new project without the assistance of the G-LINK interface. This feature will be greatly appreciated in a production environment where tasks are repetitive. Also, since G-LINK will be executed from a command line, it can be used from an SAS environment or any environment that can execute system (DOS) commands.

In the near future, it will be possible to request matching through a web service. A user will be able to submit a query for a specific record and do the matching of it on a complete table.

G-LINK provides an effective solution to complex record-matching problems. The very nature of our work here at Statistics Canada makes G-LINK an indispensable tool. The planned improvements will help make it even more valuable.

# Private record linkage with Bloom filters 

Rainer Schnell, Tobias Bachteler and Jörg Reiher ${ }^{1}$


#### Abstract

In many record linkage applications, identifiers have to be encrypted to preserve privacy. Therefore, a method for approximate string comparison in private record linkage is needed. We describe a new method of approximate string comparison in private record linkage. The main idea is to store q -grams sets derived from identifier values in Bloom filters and compare them bitwise across databases. This exploits the cryptographic features of Bloom filters while nevertheless allowing the calculation of string similarities. We show that the proposed method compares quite well to evaluating string comparison functions with plain text values of identifiers.


Key Words: Privacy-preserving Record Linkage; Private Record Linkage; Bloom Filter

## 1. Introduction

Combining multiple databases with disjunctive or additional information on the same person is occurring increasingly throughout social research. The availability of large databases and unique person identifier (ID) numbers has made widespread use of record linkage possible. But in many research applications not all databases contain a unique ID number. In such situations, probabilistic record linkage is most frequently applied for the identification of matching record pairs. However, in many applications the identifiers have to be encrypted due to privacy concerns. Linking with encrypted identifiers can result in biased estimates. Although some intriguing approaches for private record linkage have been proposed in the literature, each involves either very high computing demands or high rates of false positives or false negatives. Hence we developed a new procedure that addresses these problems.

### 1.1. The Private Record Linkage Problem

Databases of people usually contain identifiers like surnames, given names, date of birth, and address information. Distribution of scientific files containing such information is legally restricted in most countries. The problem of finding records that represent the same individual in separate databases without revealing the identity of the individuals is called the "private record linkage", "blind data linkage", or "privacy-preserving record linkage" problem.

An obvious solution for private record linkage seems to be the encryption of the identifiers with a standard cryptographic procedure. An example is the Keyed-Hash Message Authentication Code (HMAC). However, since these protocols require exact matching of identifiers, they do not tolerate any errors in these identifiers. Applying probabilistic record linkage (Herzog et al., 2007) improves the situation considerably since it does not require exact agreement in all (or even most) identifiers. Rather, agreements in better differentiating identifiers might balance disagreements in other identifiers. However, using string similarity functions within a probabilistic record linkage system will improve the linkage quality considerably. In addition, since records with variations of identifiers may have different characteristics than records with exact matching identifiers, restricting the linkage in this manner is not an option. Therefore, a method for approximate string matching in private record linkage is required.

[^81]
### 1.2 Previous Solutions

Some protocols rely on exact matching of encrypted keys based on phonetically transformed identifiers by a third party. In the proposal of Bouzelat et al. (1996) identifiers are transformed according to phonetic rules and subsequently encrypted with a oneway hash function. The hash values are transferred to a third party who performs exact matching on the resulting hash values. Despite exact matching, the linkage allows for some errors in identifiers, because hash values of phonetic encodings are matched. However, string comparison using phonetic encodings usually yields more false positive links than string similarity functions (Camps and Daude, 2003).

Churches and Christen (2004) suggested a protocol based on hashed values of sets of consecutive letters ( $q$-grams). For each string, the database holders $A$ and $B$ create for each record the power set of the $q$-grams of their identifiers. Each subset of the power set is hashed by an HMAC algorithm using a common secret key of the database owners. $A$ and $B$ form tuples containing the hash values, the number of $q$-grams in the hashed subset and the total number of $q$ grams and an encryption of the identifiers to a third party $C$. To calculate the string similarity between two strings $a$ and $b, C$ computes a similarity measure based on the information in the tuples. $C$ is able to determine a similarity measure of $a$ and $b$ by selecting the highest similarity coefficient of the tuples associated with $a$ and $b$. Apart from an increase of computational and communication costs (Verykios et al., 2009), the protocol is prone to frequency attacks on the hashes of the $q$-gram subsets with just one $q$-gram (Trepetin, 2008).
In the protocol of Scannapieco et al. (2007), two data holders, holding lists of names, build an embedding space from random strings and embed their respective strings therein using the SparseMap method (Hristescu and FarachColton, 1999). Then, each data holder sends the embedded strings to a third party who determines their similarity. To create the embedding space, data holder $A$ generates $n$ random strings and builds $z$ reference sets from them. Next, $A$ reduces the number of reference sets by the greedy resampling heuristic of SparseMap to the best $k<z$ reference sets. These $k$ reference sets are used to embed the names in a $k$-dimensional space. The coordinates for a given name are approximations of the distances between the name to the closest random string in each of the $k$ reference sets in terms of the edit distance. As a result, for each name $A$ receives a $k$-dimensional vector. After receiving the $k$ reference sets from $A, B$ embeds his names in the same way. Finally, both data holders send their vectors to a third party, $C$, who compares them using the standard Euclidean distance between them.

### 1.3 Objectives of Article

The previous solutions transform the identifiers in a manner that allows consideration of string similarities in a probabilistic record linkage procedure despite encrypting them. However, most existing protocols have a number of problems, e.g. they involve very high computing demands or high rates of classification errors. Therefore, we developed a new method for the calculation of the similarity between two encrypted strings for use in probabilistic record linkage procedures.

## 2. Methods

The core problem of a private record linkage protocol is the calculation of the similarity of two encrypted strings. We suggested the use of Bloom filters for solving this problem (Schnell et al., 2009). A Bloom filter is a data structure proposed by Bloom (1970) for checking set membership efficiently. Bloom filters can also be used to determine whether two sets approximately match (Jain et al., 2005).

### 2.1 Bloom filter

A Bloom filter is a bit array of length $l$ with all bits initially set to 0 . Furthermore, $k$ independent hash functions $h_{1}, \ldots, h_{k}$ are defined, each mapping on the domain between 0 and $l-1$. In order to store the set $S=\left\{x_{1}, x_{2}, \ldots, x_{n}\right\}$ in the Bloom filter, each element $x_{i} \in S$ is hash coded using the $k$ hash functions and all bits having indices $h_{j}\left(x_{i}\right)$ for $1 \leq j \leq k$ are set to 1 . If a bit was set to 1 before, no change is made.

In general, set membership can be checked by hashing the candidate element $y$ using the same $k$ hash functions. If all bits having indices $h_{i}(y)$ in the Bloom filter are already set to $1, y$ is presumably a member of the set $S$. There is a
probability that the check indicates membership of $y$ in $S$ when in fact it is not. It is obvious that the probability of false positive cases depends on the bit array length $l$, the number of hash functions $k$, and the number of elements in $S$ denoted by $n$. On the other hand, if at least one of the bits is found to be $0, y$ is definitely not a member of the set $S$.

### 2.2 Illustrating Example

Suppose the similarity of two surnames should be computed. At first, both surnames are split into sets of $q$-grams. Using 2-grams (usually called bigrams), the 2-gram similarity between the input strings _SMITH_ and _SMYTH_ (padded on both sides with blanks) can be computed with the Dice coefficient as $D_{A, B}=\frac{2 \cdot 4}{(6+6)}=\frac{2}{3}$. If we want to compute the similarity between those strings without revealing the bigrams, we must use an encryption. To accomplish this, we store the $q$-grams of each name in a separate Bloom filter using $k$ HMACs. Then we compare the Bloom filters bit by bit and calculate a similarity coefficient.

Figure 2.2-1

## Example of the use of two Bloom filters for the privacy-preserving computation of string similarities



In all, 8 identical bit positions are set to 1 in both Bloom filters. In total, 11 bits in A and 10 bits in B are set to 1 . Using the Dice coefficient, the similarity of the two Bloom filters is $\frac{2 \cdot 8}{(10+11)} \approx .762$ (plain text bigrams: .66). Therefore, the similarity between two strings can be approximated by using the Bloom filters alone. Using Bloom filters, approximate string matching is possible by a third party $C$ despite the privacy of the initial identifiers.

### 2.3 Simulation study

In order to test the effect of different numbers $(k=5,10,25,50)$ of hash functions on the performance of Bloom filters, similarities based on Bloom filters with a fixed filter length of 1,000 bits were compared with similarities based on unencrypted 3-grams (trigrams) of simulated data. For the simulation study, 1,000 surnames were sampled from the electronic German phone book. Lines containing just a single character were removed, umlauts and the German " $\beta$ " converted and blanks, non alphabetic characters and most common surname components like "von" were deleted. A second list of names to be matched with the original surnames was generated in a copy of the file by changing exactly one character per name with probability $p=.2$ at a randomly chosen position in the name. Therefore two files with 1,000 surnames $20 \%$ of which differ at one position were used for computing string similarities.

As is the norm in the information retrieval literature, the criteria of recall and precision were used to determine the linking effectiveness of the method. For a given level of similarity, $\varphi$, a pair of records is considered as a match if the pair is actually a true pair, all other pairs are called non-matches. Based on the common classification for true
positive $(T P)$, false positive $(F P)$, false negative $(F N)$ and true negative $(T N)$ pairs, the comparison criteria are defined as

$$
\begin{align*}
& \text { recall }=\frac{\sum_{T P} \sum_{T P+} \sum_{F N}}{\text { precision }=\frac{\sum_{T P}}{\sum_{T P+} \sum^{F P}}} \tag{1}
\end{align*}
$$

Plotting precision and recall for different similarity values $\varphi$ as a curve in a precision-recall-plot shows the performance of a string comparison method. A procedure with a better performance will have a curve in the upper right of the plot.

## 3. Results

Figure 3-1
Comparison of precision and recall for Bloom filters with unencrypted trigrams using simulated data


Figure 3-1 shows the results for the first simulation study. The precision versus recall curves in subfigure A are very similar. This is due to the fact that the probability of different trigrams being mapped to the same bit is very low (this probability is given by equation 1), since just five hash functions map the trigrams of a name onto 1,000 bits. When 10 hash functions are being used, the performance of the Bloom filter method is still very similar to the unencrypted trigrams (subfigure B). A small difference between the methods can be seen with 25 hash functions (subfigure C). However, even for 50 hash functions, the difference is not large (subfigure D ). To summarize: Inspection of the precision versus recall plots in figure 2 shows that if the number of hash functions is increased, the difference between the curve of the Bloom filter method and the curve of unencrypted trigrams also increases. However, at least up to 25 hash functions the Bloom filter method performs quite well compared with the unencrypted trigrams.

## 4. Further Evaluation

In Schnell et al. (2009), we described a full protocol for probabilistic record linkage using Bloom filters in detail. In the context of an evaluation of different probabilistic record linkage procedures for research purposes, we conducted a test of the proposed procedure on two German private administration databases. The performance of the private record linkage protocol using the Bloom filter method for computing string similarities was compared with the performances of unencrypted bigrams and a German phonetic encoding within the same probabilistic record linkage setting. To accomplish the probabilistic record linkage runs, the "Merge ToolBox" ${ }^{2}$ was used. The performance of the Bloom filter method was quite comparable to the performance of the unencrypted trigrams and outperformed the phonetic encoding.

Durham et al. (2010) also evaluated the Bloom filter method within probabilistic record linkage. Based on their tests with data containing simulated errors they conclude that the method is applicable with medical data files and that run times are practical for real world use.

In the present proceedings, we also report on a comparison of the Bloom filter method with other approaches to determine the similarity between two encrypted strings (Bachteler et al., 2010). Using experimentally generated human data, the performance of the methods proposed by Pang and Hansen (2004) and Scannapieco et al. (2007) were compared to the Bloom filter method by examining precision-recall plots and run times. The results shows poor precision and recall and long run times by the protocol of Pang and Hansen and good linkage results and a moderate run times by the protocol of Scannapieco et al. However, the Bloom filter method yielded the best linkage results and lowest run times.

## 5. Conclusions

In this article, we demonstrate a solution for the private record linkage problem with encrypted identifiers allowing for errors in identifiers. The solution is based on similarity computations using Bloom filters with HMACs on $q$ grams. This method has been tested successfully on simulated test data. We were able to show that the linkage results are comparable to non-encrypted identifiers.

Since the method can be easily enhanced and has a low computational burden, the protocol might be useful for many applications requiring privacy-preserving record linkage. The method can be readily incorporated in existing record linkage software by adding a simple subroutine that computes the Dice coefficient from two Bloom filters.

## References

Bachteler, T., Schnell, R. and Reiher, J. (2010). An Empirical Comparison of Approaches to Approximate String Matching in Private Record Linkage, Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data.

Bloom, B.H. (1970). Space/time Trade-offs in Hash Coding with Allowable Errors, Communications of the ACM, 13, pp. 422-426.

Bouzelat, H., Quantin, C. and Dusserre, L. (1996). Extraction and Anonymity Protocol of Medical File, Proceedings of the 1996 AMIA Annual Fall Symposium, pp. 323-327.

Camps, R. and Daude, J. (2003). Improving the Efficacy of Approximate Searching by Personal-Name, Proceedings of the 8th International Conference on Applications of Natural Language to Information Systems, pp. 70-76.

[^82]Churches, T. and Christen, P. (2004). Some Methods for Blindfolded Record Linkage, BMC Medical Informatics and Decision Making, 4.

Durham, E., Xue, Y., Kantarcioglu, M. and Malin, B. (2010). Private Medical Record Linkage with Approximate Matching, Proceedings of the 2010 American Medical Informatics Association Annual Symposium, pp. 182186.

Herzog, T.N., Scheuren, F. and Winkler, W.E. (2007). Data Quality and Record Linkage Techniques, New York: Springer.

Hristescu, G. and Farach-Colton, M. (1999). Cluster-preserving Embedding of Proteins, DIMACS Technical Report 99-50, DIMACS Center for Discrete Mathematics \& Theoretical Computer Science.

Jain, N., Dahlin, M. and Tewari, R. (2005). Using Bloom Filters to Refine Web Search Results, Proceedings of the Eight International Workshop on the Web \& Databases, pp. 25-30.

Pang, C. and Hansen, D. (2006). Improved Record Linkage for Encrypted Identifying Data, Proceedings of the 14th Annual Health Informatics Conference, pp. 164-168.

Schnell, R., Bachteler, T. and Bender, S. (2004). A Toolbox for Record Linkage, Austrian Journal of Statistics 33, pp. 125-133.

Schnell, R., Bachteler, T. and Reiher, J. (2009). Privacy-preserving Record Linkage using Bloom Filters, BMC Medical Informatics and Decision Making, 9.

Scannapieco, M., Figotin, I., Bertino, E. and Elmagarmid, A.K. (1996). Privacy Preserving Schema and Data Matching, Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 653664.

Trepetin, S. (2008). Privacy-Preserving String Comparisons in Record Linkage Systems: A Review, Information Security Journal, 17, pp. 253-266.

Verykios, V.S., Karakasidis, A. and Mitrogiannis, V.K. (2009). Privacy Preserving Record Linkage Approaches, International Journal of Data Mining, Modelling and Management, 1, pp. 206-221.

## SESSION 9A

## SMALL AREA ESTIMATION

# Estimation of average design-based mean squared error of synthetic small area estimators 

Partha Lahiri and Santanu Pramanik ${ }^{1}$


#### Abstract

Many survey organizations use synthetic methods to produce small area estimates because of their simplicity and wide applicability. In this paper, we propose a class of design-based average mean square error (AMSE) estimators of a set of synthetic estimators of small area means. Using data analyses and Monte Carlo simulations, we compare a particular AMSE estimator in the class with two other AMSE estimators available in the literature. The results suggest that the proposed AMSE estimator is a sensible choice.


Key Words: Bias; Borrow strength; Estimating equation.

## 1. Introduction

Sample surveys designed to produce reliable estimates for a population cannot guarantee estimates of similar precision for small subgroups of the surveyed population, known as small areas or domains. For example, in the National Health and Nutrition Examination Survey (NHANES) III, certain minority groups residing predominantly in certain states (e.g., California and Texas) were oversampled. This design strategy resulted in small or no samples for the states that do not have large populations for these minority groups. This problem of small or zero sample size in small areas prevents the use of the standard direct survey estimate for small area parameters since the estimates are likely to be highly unreliable or unavailable.

In the absence of adequate direct information for small areas, it is customary to borrow strength from related sources to form indirect estimators that increase the effective sample size and hence reduce the sampling errors of the estimators. Such indirect estimators are usually based on implicit or explicit models that combine information from the sample survey, various administrative/census records, or previous surveys. There is a large volume of literature on indirect estimation for small areas. We refer to Rao (2003) and Jiang and Lahiri (2006) for a detailed account of small area estimation.

Synthetic estimation, which applies to any probability and non-probability sample design, is a simple indirect method of borrowing strength from similar areas. The basic assumption underlying a synthetic method is that small areas have same characteristics as the large area (e.g., unemployment rate for a given demographic group remains the same across different states). Synthetic methods are often employed in practice for their simplicity and ability to produce estimates for areas with no sample from the sample survey.

Hansen et al. (1953) presents an early example of synthetic regression method in estimating the median number of radio stations heard during the day for over 500 counties of the United States. Although Hansen et al. (1953) generated synthetic estimates using a simple linear regression models on survey estimates, more sopohisticated models have been used to produce synthetic estimates in other applications. For example, Stasny et al. (1991) developed a regression-synthetic method for estimating county acreage of wheat using a non-probability sample of farms along with auxiliary data on planted acreage and district indicators. Elbers et al. (2003) developed a micro-
${ }^{1}$ Partha Lahiri, Joint Program of Survey Methodology, University of Maryland, College Park, MD 20742,(email:plahiri@ survey.umd.edu); Santanu Pramanik, NORC at the University of Chicago, (e-mail:pramaniksantanu@norc.org). The authors thank Vincent T. Mule and Patrick J. Cantwell of the U.S. Census Bureau for making some useful comments related to the U.S. Census CCM program. The first authors research was supported in part by National Science Foundation grant SES-0851001 and a U.S. Census Bureau contract YA132309CN0057.
simulation synthetic method for poverty mapping, commonly referred to as the ELL method or the World Bank method, using a more sophisticated mixed model to capture different design features of the survey design. More recently researchers at the US Census Bureau are planning on producing dual system estimates of the household population for the 2010 Census coverage using a synthetic method that uses logistic regression.

In official statistics, a major advance on synthetic estimation, based on implicit models, came with an experiment conducted by researchers at the U.S. National Center for Health Statistics (NCHS) during the mid-1960s to estimate disability and other health characteristics for small areas (Rao, 2003). The U.S. Census Bureau followed the NCHS's lead in producing county level synthetic estimates of unemployment rates (see Gonzalez and Hoza (1978)). Researchers at the U.S. Census Bureau is currently exploring the possibility of producing estimates of different components of the 2010 decennial census coverage errors using a synthetic method based on logistic regression. The reweighting method, described in Schirm and Zaslavsky (1997), can be also viewed as a synthetic method where the survey weights are synthetically determined to a match certain known controls. A major advantage of such reweighting method is that for a given small area a single set of weights can be applied to estimate small area characteristics for a large number of variables. Although a number of synthetic estimators have been suggested in the small area estimation literature, their evaluation received relatively less attention. The variance of a synthetic estimator, whether design-based or model-based, as a measure of evaluation could be misleading as it does not incorporate the bias component of the synthetic estimator, which could be more prominent than the variance component and does not vanish even for large samples. The mean square error (MSE), which incorporates both the variance and bias components, may be a more acceptable in evaluating synthetic estimators. The variance of a synthetic estimator can be estimated using any standard variance estimation technique - design-based or model-based - since this involves variance estimation of large area parameters such as common regression coefficients. While it is possible to obtain an unbiased or a nearly unbiased estimator of the bias of the synthetic estimator, such estimators are highly unreliable since they are based on small sample sizes.

Recognizing the problem associated with the individual MSE estimation, Gonzalez and Waksberg (1973) introduced the concept of the average mean square error (AMSE) of a set of synthetic estimators and suggested an estimator of AMSE. In a recent study related to the coverage measurement program of the U.S. Census Bureau, the GonzalezWaksberg method frequently produced negative AMSE estimates, which motivates an exploration of alternative estimators of AMSE.

The outline of this paper is as follows. In Section 2, we introduce a few notations. In Section 3, we discuss synthetic estimation. In Section 4, we define individual and average MSE. In this section, we first revisit the GonzalezWaksberg (GW) AMSE estimator and then introduce a class of AMSE estimators. We compare our new AMSE estimator with two rival AMSE estimators using real life data analyses (Section 5) and a Monte Carlo simulation (Section 6). The paper ends with a concluding remarks section (Section 7) where we summarized our research findings and discuss future research in this area.

## 2. A Table of Notations

We use the following notations throughout the paper:
$m$ : number of small areas,
$U:$ set of units in the finite population of interest, which contains $m$ small areas $U_{i}$,
$N_{i}:$ size of $U_{i},\left(N=\sum_{i=1}^{m} N_{i}\right)$,
$y_{k}$ : value of a characteristic (discrete or continuous) of interest for the $k$ th unit in $U$,
$Y i=\sum_{k \in U,} y_{k}: i$ th area population total,
$\bar{Y}_{i}=\frac{Y_{i}}{N_{i}}, i$ th area population mean,
$s$ : set of units in the sample,
$s_{i}$ : set of units in $s$ that belong to area $i\left(\right.$ size $\left.n_{i}\right)$,
$w_{k}:$ survey weight associated with unit $k \in s$.
Using the notations above, we define survey-weighted direct estimators of $\bar{Y}_{i}$ as $\hat{\bar{Y}}_{i}^{D}=\frac{\sum_{k \in s_{i}} w_{k} y_{k}}{\sum_{k \in s_{i}} w_{k}}$. Throughout the paper, we assume that sampling variances of $\hat{\bar{Y}}_{i}{ }^{D}$ are known. In practice, smoothed estimates $V_{i}^{D}$ are obtained using Generalized Variance Function (GVF) method. We refer to Otto et Bell (1995) for further details.

## 3. Synthetic Estimation

As noted in the introduction, synthetic estimators can be produced using either an implicit or explicit model. In this section, we describe two different synthetic estimators - one based on an area leave model and the other based on a unit level model. In either case, there is no unknown area specific effects present in the model and both methods provide synthetic estimates even when there is no sample data in the area.

### 3.1 Area level Model

We first fit a model that relates survey-weighted direct estimates to a set of auxiliary variables at the small area level and then use fitted values for the areas as synthetic estimates of the small area parameters.

Example: In obtaining synthetic estimators of small area means $\bar{Y}_{i}$, we follow the following two steps:
Step 1: Fit a multiple regression model : $\hat{\bar{Y}}_{i}{ }^{D}=\mathbf{X}_{i}^{T} \beta+\xi_{i}$, where $\mathbf{X}_{i}$ is a vector of known auxiliary variables for area $i ; \beta$ is a vector of unknown de regression coefficients; $\left\{\xi_{i}, i=1, \ldots, m\right\}$ are uncorrelated errors with means 0 and known variances $V_{i}^{D}$.

Step 2 : A synthetic estimator of $\bar{Y}_{i}$ is then given by $\hat{\bar{Y}}_{i}=X_{i}^{T} \hat{\beta}$, where
$\hat{\beta}=\left(\sum_{j} X_{j} X_{j}^{T} / V_{j}^{D}\right)^{-1} \sum_{j} X_{j} \hat{\bar{Y}}_{j}^{D} / / V_{j}^{D}$ is the weighted least squares estimator of $\beta$.
Note that we do not need to estimate any area specific parameters. The regression coefficients $\beta$ are the same across areas.

### 3.2 Unit Level Model

We fit a unit level model with no small area specific parameter to the entire survey data in order to establish a relationship between the target variable and a set of auxiliary variables and then obtain the synthetic estimators of the small area means as $\hat{\bar{Y}}_{i}=N_{i}^{-1} \sum_{k \in U}, \hat{y}_{k}$, where $\hat{y}_{k}$ is the fitted value of $y_{k}$ from the model. Depending of the model, we may or may not need the auxiliary variables for all units in the population. For example, if we fit a
multiple linear regression model, we just need the small area population means for these auxiliary variables to estimate $\overline{Y_{i}}$. On the other hand, for a non-linear model like the logistic model, we need the auxiliary variables for all units of the population.

Example : Let $y_{k}$ be a binary variable associated with an attribute and consider estimation of the total number of units in the $i$ th small areas satisfying the attribute. Assume the following logistic synthetic model:

$$
\operatorname{logit}\left[P\left(y_{k}=1\right)\right]=x_{k}^{T} \beta+\xi_{k},
$$

where $\xi_{k}$ are iid with means 0 and variance $\sigma^{2}$. Note that the above logistic model is synthetic since the model parameters (i.e., $\beta$ et $\sigma^{2}$ ) are assumed to be identical across all small areas. A synthetic estimator of $\bar{Y}_{i}$ is then obtained using the following two simple steps:

Step 1 : Obtain the usual survey-weighted estimate of $\beta$, say $\hat{\beta}$, of the logistic model using the entire survey data. Step 2 : A synthetic estimator of $\bar{Y}_{i}$ is then given by $\hat{\bar{Y}}_{i}=N_{i}^{-1} \sum_{k \in U_{i}} \hat{y}_{k}$, where $\hat{y}_{k}=\frac{\exp \left(x_{k}^{T} \hat{\beta}\right)}{1+\exp \left(x_{k}^{T} \hat{\beta}\right)}$, the fitted value of $y_{k}$ from the logistic model.

In the U.S. Census Coverage Measurement (CCM) program, logistic models will be used in order to estimate the three necessary probabilities of the logistic regression dual system estimator (Mule, 2010). Before examining this more complex estimator, the CCM program evaluated the performance of average mean square error estimators for the synthetic estimate of the number of erroneous enumeration using 2000 coverage data. The logistic model used contains main effects due to different factors such as census region, place size grouping, tenure, race/origin domain, sex, etc. as well as different interactions such as domain by tenure, tenure by sex, etc.

## 4. Design-based Mean Squared Error of Synthetic Estimators

We define the design-based mean square error (MSE) of a synthetic estimator $\hat{\bar{Y}}_{i}$ of the $i$ th small area mean $\bar{Y}_{i}$ as

$$
\begin{equation*}
\operatorname{MSE}\left(\hat{\bar{Y}}_{i}\right) \equiv M_{i}=E\left(\hat{\bar{Y}}_{i}-\bar{Y}_{i}\right)^{2}=V_{i}+B_{i}^{2} \tag{1}
\end{equation*}
$$

where $V_{i}=V\left(\hat{\bar{Y}}_{i}\right)$, variance of $\hat{\bar{Y}}_{i}, B_{i}=E\left(\hat{\bar{Y}}_{i}\right)-\bar{Y}_{i}$, bias of $\hat{\bar{Y}}_{i}$, and the expectations and variances are with respect to the sample design $(i=1, \ldots, m)$.

The variances $V_{i}$ are generally small as they are based on the total sample size, which is generally large. But, $B_{i}^{2}$ is a finite population parameter and does not depend on the sample size. Its magnitude depends on the synthetic assumption that generates the synthetic estimators.

## An Example: Stratified Simple Random Sampling

Consider a stratified simple random sampling design in which a simple random sample of same size $n$ is selected from each stratum (same as small area). In this case, a direct design-unbiased estimator of $\bar{Y}_{i}$ is given by $\hat{\bar{Y}}_{i} D=\bar{y}_{i}$, the sample mean for area $i$. For illustration, we consider the overall sample mean, i.e.
$\hat{\bar{Y}}_{i}=\bar{y}=m^{-1} \sum_{i=1}^{m} \bar{y}_{i}$, as a synthetic estimator of $\bar{Y}_{i}$. This synthetic estimator is reasonable when the synthetic assumption $\bar{Y}_{i}=\bar{Y}(i=1, \ldots, m)$, where $\bar{Y}$ is the overall population mean, is approximately true.

Ignoring the finite population correction (fpc), we have

$$
V_{i}^{D}=\frac{S_{i}^{2}}{n}=O\left(\frac{1}{n}\right), V_{i}=\frac{\overline{S^{2}}}{n m}=O\left(\frac{1}{m n}\right), B_{i}=\bar{Y}-\bar{Y}_{i},
$$

where $S_{i}^{2}$ is the population variance for area $i$ and $\overline{S^{2}}=m^{-1} \sum_{i=1}^{m} S_{i}^{2}$. Notice that the sample size has no impact on $B_{i}$, but $V_{i}$ can be reduced by increasing the sample size. The magnitude of $B_{i}$ depends solely on the validity of the synthetic assumption. A design-consistent estimator of $V_{i}$ is $\hat{V}_{i}=\hat{\bar{S}}^{2} /(n m)$, where $\hat{\bar{S}}^{2}=m^{-1} \sum_{i=1}^{m} s_{i}^{2}$, and $s_{i}^{2}$ is the sample variance. We can estimate $B_{i}$ by an unbiased estimator $\hat{B}_{i}=\bar{y}-\bar{y}_{i}$, but this estimator is very unreliable due to the high variability of $\bar{y}_{i}$.

Following (Gonzalez and Waksberg, 1973), we define the average mean square error (AMSE) of a set of $m$ synthetic estimators of small area population means as

$$
\begin{equation*}
\mathrm{AMSE} \equiv M=\bar{V}+\eta \tag{2}
\end{equation*}
$$

where $\bar{V}=m^{-1} \sum_{i=1}^{m} V_{i}$ and $\eta=m^{-1} \sum_{i=1}^{m} B_{i}^{2}$. A naïve Estimator of AMSE is given by $\hat{M}^{\text {naïe }}=m^{-1} \sum_{i=1}^{m} \hat{V}_{i}=\hat{\bar{V}}$, say, where $\hat{V}_{i}$ is a design-consistent estimator of $V_{i}(i=1, \ldots, m)$. For large $m$ and standard regularity conditions, $E\left[\hat{M}^{\text {naive }}\right] \approx M-\eta$, and thus $\hat{M}^{\text {naive }}$ underestimates the true AMSE, irrespective of the sample size, and the extent of the underestimation depends on the accuracy of the synthetic assumption that generates the synthetic estimators.

Under the assumptions $\operatorname{Cov}\left(\hat{\bar{Y}}_{i}, \hat{\bar{Y}}_{i}{ }^{D}\right) \approx 0$ and $\mathrm{E}\left(\hat{\bar{Y}}_{i}^{D}\right) \approx \hat{Y}_{i}$, one gets

$$
\operatorname{MSE}\left(\hat{\bar{Y}}_{i}\right) \approx E\left(\hat{\bar{Y}}_{i}-\hat{\bar{Y}}_{i}^{D}\right)^{2}-V_{i}^{D},
$$

which motivates the Gonzales-Waksberg (GW) AMSE estimator:

$$
\hat{M}^{\mathrm{GW}}=m^{-1} \sum_{i=1}^{m}\left(\hat{Y}_{i}-\hat{Y}_{i}^{D}\right)^{2}-m^{-1} \sum_{i=1}^{m} V_{i}^{D},
$$

see Gonzalez and Waksberg (1973). Marker (1995) obtained an area specific MSE estimator for small area total estimation. Such a formula can be easily extended when the parameters of interest are the small area means. It is interesting to note that the AMSE formula based on such area specific formula for small area means is identical to the GW AMSE estimator $\hat{M}^{\mathrm{GW}}$. The GW estimator is approximately design-unbiased. However, as Rao (2003), noted, it can produce negative estimates of AMSE, which is highly undesirable.

To obtain a new AMSE estimator, we use (2). First, $\bar{V}$ is replaced by $\hat{\bar{V}}$. To estimate $\eta$, we first find an approximately design-unbiased optimal estimating equation. To this end, first note that an approximately design unbiased estimator of $\quad B_{i}$ is given by $\hat{B}_{i}=\hat{\bar{Y}}_{i}-\hat{\bar{Y}}_{i}^{D}(i=1, \ldots, m)$. Define $\in_{i} \hat{B}_{i}^{2}-\left(\eta+V_{i}^{D}\right), i=1, \ldots, m$. Under
the assumptions $B_{i}^{2}=\eta, \mathrm{V}\left(\hat{Y}_{i}\right)=O\left(m^{-1}\right), \operatorname{Cov}\left(\hat{Y}_{i}, \hat{Y}_{i}^{D}\right)=O\left(m^{-1}\right)(i=1, \cdots, m)$, it can be shown that $E\left(\epsilon_{i}\right) \approx O\left(m^{-1}\right)$, for large $m$.

Consider the following general class of estimating equations :

$$
\begin{equation*}
q(\eta ; \mathbf{l})=\sum_{j=1}^{m} l_{j} \epsilon_{j}+h(\eta)=0, \tag{3}
\end{equation*}
$$

where $\mathbf{1}=\left\{l_{j} \geq 0, j=1, \ldots, m\right\}$ is a vector of constants (possible depending on $\eta$ ) with $\sum_{j=1}^{m} l_{j}=1$ and $h(\eta)=O\left(m^{-1}\right)$. Different choices of the weights $l_{j}$ and adjustment factor $h(\eta)$ yield different estimators of $\eta$. First note that $E[q(\eta: \mathbf{1})]=O\left(m^{-1}\right)$, and hence equation (3) yields an approximately design-unbiased estimating equation. It is interesting to note that the choice $l_{j}=m^{-1}, h(\eta)=0$ in (3) yields an estimator of $\eta$ that is identical to the GW AMSE estimator. So for this choice, our AMSE estimator would be larger than the GW AMSE estimator by an amount $\hat{\bar{V}}$.

Next, we obtain an approximately optimal estimating equation in the class of approximately design-unbiased estimating equations. We call such an estimating equation approximately design-unbiased optimal estimating equation. First note that, ignoring the covariances, we obtain $V[q(\eta: 1)] \approx \sum_{j=1}^{m} l_{j}^{2} a_{j}$, where $a_{j}=V\left(\epsilon_{j}\right)=V\left(\hat{B}_{j}^{2}\right)$. We minimize $\sum_{j=1}^{m} l_{j}^{2} a_{j}$ with respect to $l_{j}(j=1, \ldots, m)$, such that $\sum_{j=1}^{m} l_{j}=1$. To this end, note that

$$
\begin{equation*}
\sum_{j=1}^{m} l_{j}^{2} a_{j}=\sum_{j=1}^{m}\left(l_{j}-\frac{a_{j}^{-1}}{\sum_{j=1}^{m} a_{j}^{-1}}\right)^{2} a_{j}+\frac{1}{\sum_{j=1}^{m} a_{j}^{--}} . \tag{4}
\end{equation*}
$$

Since the second term in the right hand side of (4) is non-negative, $\sum_{j=1}^{m} l_{j}^{2} a_{j}$ is minimized subject to $\sum_{j=1}^{m} l_{j}=1$ if $l j=\frac{a_{j}^{-1}}{\sum_{j=1}^{m} a_{j}^{-1}}(j=1, \ldots, m)$. Moreover, if we assume $V\left(\hat{B}_{j}^{2}\right) \approx k\left(\eta+V_{j}^{D}\right)^{2}$, for some constant $k$, an approximately design-unbiased optimum estimating equation is given by

$$
\begin{equation*}
q(\eta ; \mathbf{l})=\sum_{j=1}^{m} \frac{\hat{B}_{j}^{2}-\left(\eta+V_{j}^{D}\right)}{\left(\eta+V_{j}^{D}\right)^{2}}+\tilde{h}(\eta)=0 \tag{5}
\end{equation*}
$$

where

$$
\tilde{h}(\eta)=h(\eta) \sum_{j=1}^{m} \frac{1}{\left(\eta+V_{j}^{D}\right)^{2}}
$$

We propose to choose $\bar{h}(\eta)$ that performs well in terms of the bias of $\hat{\eta}$.

## 5. SAIPE Data Analyses

In this section, we use state level data for the years 1991, 1993 and 1997 from the Small Area Income and Poverty Estimates (SAIPE) program of the U.S. Census Bureau in order to compare our new AMSE estimator with the naïve
and GW AMSE estimators. For details about the program we refer to (Citro and Kalton, 2000). For this data analyses, the parameters of interest are the true proportions of 5-17 year old children in poverty from the fifty states and the District of Columbia for the years1991, 1993 and 1997. The survey-weighted proportions $\hat{\bar{Y}}_{i}$ are obtained using the Current Population Survey (CPS) data. The sampling variance estimates $\hat{V}_{i}$ are obtained using a Generalized Variance Function (GVF) method, following the lines of Otto and Bell (1995). The Census Bureau now switched from the CPS to the American Community Survey (ACS) data as the direct source of survey data in the production of the state level SAIPE estimates. Also, the Census Bureau does not use synthetic estimators for state level SAIPE. Nonetheless, we find it convenient to illustrate the performances of different AMSE estimators using the old SAIPE data.

In this data analyses, we obtain synthetic estimates using an area level multiple regression synthetic model described in Section 3.1. As for the auxiliary variables $\mathbf{X} i$, we use Internal Revenue Service (IRS) data, food stamp data and census residuals, which are used in the SAIPE program. First we form four distinct groups of states and the District de Columbia, so that the number of states within a group is about the same ( 13 for three groups and 12 for one). The small areas in a given group are similar in terms of the sampling variance estimates $V_{i}^{D}$. In other words, for a given year the reliability of direct estimates varies across groups. For example, direct estimates for the states within the first group are much more reliable than those in the second group.

We present our data analyses in Table 1. Note that the GW AMSE estimates produce negative estimates in all but one of the cases. The performance of the GW method seems to deteriorate as we consider groups with progressively larger sampling variances. The estimates of $\eta$ are smaller than the corresponding naïve AMSE estimates in all the cases, but they contribute considerable (sometimes about a half of the corresponding naïve AMSE estimate) to the new AMSE estimates. In terms of the estimated coefficient of variation, the synthetic estimates are doing much better than the direct estimates for states with highly variable direct estimates. For states with relatively more reliable direct estimates, the synthetic estimates still improve on the direct estimates, but by a smaller margin compared to the states with highly unreliable direct estimates. The use of the new average CV estimates associated with the synthetic estimates appears to be more sensible compared to the naïve average CV estimates for the same synthetic estimates. For example, for the bigger states where the direct estimates may be quite reliable, the new average CV estimates narrows down the gap between the estimated CVs for the synthetic and direct estimates.

Table 1: Evaluation of Different AMSE Estimator

|  | Average MSE Estimates of Synthetic |  | Coefficient of Variation |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Grouping | GW | Naïve | Bias $^{2}$ | New | Direct | Synthetic | Synthetic |
| States by |  |  | $(\hat{\eta})$ |  |  |  |  |
| $V_{i}^{D}$ |  |  |  |  |  |  |  |

Results based on 1991 SAIPE Data

| $[1,94,8,2)$ | $\mathbf{1 . 0 4 3}$ | $\mathbf{2 . 5 2 5}$ | $\mathbf{1 . 5 3 5}$ | $\mathbf{4 . 0 6 0}$ | $\mathbf{0 . 1 1 4}$ | $\mathbf{0 . 0 8 9}$ | $\mathbf{0 . 1 1 3}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $[8,2,11,8)$ | $\mathbf{- 5 . 8 4 9}$ | $\mathbf{3 . 3 8 5}$ | $\mathbf{1 . 6 5 1}$ | $\mathbf{5 . 0 3 5}$ | $\mathbf{0 . 1 9 1}$ | $\mathbf{0 . 1 0 8}$ | $\mathbf{0 . 1 3 2}$ |
| $[11,8,14,1)$ | -7.119 | 3.757 | $\mathbf{1 . 9 6 3}$ | $\mathbf{5 . 7 2 0}$ | $\mathbf{0 . 2 2 0}$ | $\mathbf{0 . 1 1 8}$ | $\mathbf{0 . 1 4 6}$ |
| $[14,7,30)$ | $\mathbf{- 1 0 . 1 9 8}$ | $\mathbf{4 . 0 3 4}$ | $\mathbf{2 . 0 9 5}$ | $\mathbf{6 . 1 2 9}$ | $\mathbf{0 . 2 1 2}$ | $\mathbf{0 . 1 0 3}$ | $\mathbf{0 . 1 2 7}$ |

Results based on 1993 SAIPE Data

| $[2,16,7,97)$ | $\mathbf{- 0 . 1 8 7}$ | $\mathbf{2 . 1 0 5}$ | $\mathbf{1 . 0 8 7}$ | $\mathbf{3 . 1 9 2}$ | $\mathbf{0 . 1 1 1}$ | $\mathbf{0 . 0 7 7}$ | $\mathbf{0 . 0 9 5}$ |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $[7,97,12,2)$ | $\mathbf{- 2 . 5 7 6}$ | $\mathbf{5 . 3 4 5}$ | $\mathbf{2 . 4 3 3}$ | $\mathbf{7 . 7 7 7}$ | $\mathbf{0 . 2 3 3}$ | $\mathbf{0 . 1 6 8}$ | $\mathbf{0 . 2 0 2}$ |
| $[12,2,14,7)$ | $\mathbf{- 5 . 8 7 0}$ | $\mathbf{4 . 6 8 8}$ | $\mathbf{2 . 2 8 9}$ | $\mathbf{6 . 9 7 8}$ | $\mathbf{0 . 2 2 2}$ | $\mathbf{0 . 1 3 1}$ | $\mathbf{0 . 1 6 0}$ |
| $[14,7,38,2)$ | $\mathbf{- 2 . 8 1 4}$ | $\mathbf{8 . 6 5 0}$ | $\mathbf{3 . 7 4 7}$ | $\mathbf{1 2 . 3 9 7}$ | $\mathbf{0 . 1 8 6}$ | $\mathbf{0 . 1 2 8}$ | $\mathbf{0 . 1 5 4}$ |

Results based on 1997 SAIPE Data

| $[2,34,7,87)$ | $\mathbf{- 2 . 0 8 5}$ | $\mathbf{1 . 4 3 5}$ | $\mathbf{0 . 6 9 7}$ | $\mathbf{2 . 1 3 2}$ | $\mathbf{0 . 1 3 1}$ | $\mathbf{0 . 0 7 2}$ | $\mathbf{0 . 0 8 7}$ |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| $[7,87,11)$ | $\mathbf{- 3 . 0 0 6}$ | $\mathbf{4 . 5 5 3}$ | $\mathbf{2 . 1 1 8}$ | $\mathbf{6 . 6 7 1}$ | $\mathbf{0 . 2 4 4}$ | $\mathbf{0 . 1 7 0}$ | $\mathbf{0 . 2 0 6}$ |
| $[11,13,7)$ | $\mathbf{- 9 . 6 6 7}$ | $\mathbf{1 . 4 9 9}$ | $\mathbf{1 . 2 4 6}$ | $\mathbf{2 . 7 4 5}$ | $\mathbf{0 . 2 0 0}$ | $\mathbf{0 . 0 7 0}$ | $\mathbf{0 . 0 9 5}$ |
| $[13,7,30,8)$ | $\mathbf{- 6 . 5 7 7}$ | $\mathbf{6 . 5 6 0}$ | $\mathbf{2 . 9 7 0}$ | $\mathbf{9 . 5 3 0}$ | $\mathbf{0 . 2 0 9}$ | $\mathbf{0 . 1 3 3}$ | $\mathbf{0 . 1 6 0}$ |

## 6. Simulation Study

For our simulation, we consider $m=20$ small areas, each containing $N=400$ units. We generate the finite population based on the following nested error regression model (Battese et al., 1988):

Level
Level

For the generation, we assume: $\mu=10, \sigma_{v}^{2}=10, \sigma_{e}^{2}=100$. Once we generate the population, we assume it to be fixed, as common in a design-based framework. To estimate the small area means $\bar{Y}_{\boldsymbol{i}}$, we draw a stratified simple random sample, considering the small areas to be strata, of size $n=8$ from each stratum. For the evaluation purpose, we repeat the sampling procedure 2000 times. We consider $\bar{h}(\eta)=\frac{1}{2 \eta(\sqrt{n+1})}$.


Figure 1: Boxplot of AMSE estimates
Figure 1 is the boxplot of AMSE estimates obtained after applying three different methods mentioned above. The GW method yields a certain percentage ( $3 \%$ ) of negative estimates, as evident from the boxplot. The solid red line in the graph indicates the zero line. The naïve MSE estimate severely underestimates the true AMSE of the synthetic estimators as it ignores the bias part completely. The new AMSE estimator slightly overestimates the true AMSE of the synthetic estimators, whereas the GW MSE estimate slightly under-estimates the true AMSE. In terms of the variability, the GW and new AMSE estimators are comparable, although the new AMSE estimator has smaller variance (48.57) as compared to 50.74 of the GW estimator (see Table 2). In summary, the new AMSE estimator excludes the possibility of producing negative estimates and in terms of AMSE it is comparable with the GW method as we noticed in (Table 2).

Table 2: Evaluation of AMSE Estimators

| Summary Measures | GW | Naïve | New |
| :--- | :---: | :---: | :---: |
| Bias | -1.21 | -12.67 | 1.63 |
| Variance | 50.74 | 0.005 | 48.57 |
| MSE | 52.21 | 160.60 | 51.23 |

## 7. Concluding Remarks

According to our limited data analyses and simulation results, the new AMSE estimator appears to be more sensible than the naïve and GW AMSE estimators. Our preliminary analytical results (not reported here) suggest that the new AMSE estimator is strictly positive and design-consistent, under certain regularity conditions. We shall report these theoretical results in a separate paper. Extension of the methodology to cover the more general composite estimator is currently under investigation.

## Bibliography

Battese, G., Harter, R. and Fuller, W. (1988). An error-components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistical Association, 83 28-36.

Citro, C. and Kalton, G. (eds.) (2000). Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Panel on Estimates of Poverty for Small Geographic Area, Committee on National Statistics, Washington, DC: National Academy Press.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro-Level Estimation of Poverty and Inequality, Econometrica, 71 355-364.

Gonzalez, M., and Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. Journal of the American Statistical Association, 73 7-15.

Gonzalez, M. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. In first meeting of the international Association of Survey Statisticians, Vienna, Austria, vol. 18. 25.

Hansen, M., Hurwitz, W. and Madow, W. (1953). Sample survey methods and theory, vol 1.
Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. Test, 15 1-96.
Otto, M. and Bell, W. (1995). Sampling error modelling of poverty and income statistics for states. In American Statistical Association, Proceedings of the Section on Government Statistics. 160-165.

Rao, J.N.K. (2003). Small area estimation. Wiley Series in Survey Methodology, Wiley-Interscience [John Wiley \& Sons].

Schirm, A.L. and Zaslavsky, A.M. (1997). Reweighting households to develop microsimulation estimates for states, American Statistical Association, Proceedings of the Survey Research Methods Section.

Stasny, E., Goel, P. and Rumsey, D. (1991). Estimation de la production de blé par comté, Téchniques d'enquête, 17, 211-225.

# Estimation of poverty measures for small areas 

J. N. K. Rao ${ }^{1}$


#### Abstract

Eradication of extreme poverty and hunger is the first of the Millennium Development Goals established by the United Nations. Availability of reliable statistics on people's living conditions is a basic requirement for the achievement of this goal. However, information collected from national surveys is often limited and allows reliable estimation only for large regions or large population subgroups. For areas with small (or even zero) sample sizes it is necessary to employ indirect estimation methods that can lead to reliable estimates by borrowing information across related areas through linking models. In this talk, our focus is on small area estimation of poverty measures that are complex non-linear functions of the values of a welfare variable, in particular on a class of measures used by the World Bank that covers poverty incidence, poverty gap and poverty severity as special cases. We use the empirical Bayes (EB) method to obtain efficient EB estimators of small area poverty measures, based on a nested error linear regression linking model relating a transformed welfare variable to auxiliary variables available for all population units. These estimators are calculated through simulation methods, as they do not admit closed form expressions. The mean squared prediction error (MSPE) of the EB estimators is estimated through a parametric bootstrap method. We study the performance of the proposed estimators relative to direct area-specific estimators and widely used synthetic estimators, based on a "simulated census" approach, using model-based and design-based simulation studies. We also apply the proposed method to estimate poverty incidences and poverty gaps in Spanish provinces by gender. The proposed methodology is applicable to general non-linear parameters.


This talk is based on joint work with Isabel Molina, Universidad Carlos III de Madrid, Getafe (Madrid), Spain.

[^83]
# Alternative forms of small area models for using survey, census and administrative records data 

William R. Bell and Elizabeth T. Huang ${ }^{1}$


#### Abstract

We start with the basic Fay-Herriot area level model for small area estimation and then review its generalizations to bivariate, multivariate, and measurement error models. A primary motivation for using a more general model may be the potential for achieving greater improvements in small area estimators by making use of additional data. Our focus, however, is more on reviewing (1) the assumptions made by these models about the various data sources (survey, census and administrative records) being used, (2) how the models thus use the various types of data, and (3) what assumptions seem generally appropriate for which types of data sources. Conversely, we also consider how choosing a model whose assumptions do not match the data could lead to problems, or at least to suboptimal data usage. We provide illustrative examples using data from the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program.


[^84]
## SESSION 9B

EMPLOYMENT DATA

# Combining administrative and survey data in a German research data centre: Linkage, quality and future developments 

Stefan Bender, Jörg Heining, Tanja Hethey and Patrycja Scioch ${ }^{1}$


#### Abstract

One of the challenges in data production is to get richer data sets by linking different data sources. The Research Data Centre (FDZ) of the German Federal Employment Agency at the Institute for Employment Research tries to link administrative data to other data sources like survey data. Linkage has been done at the individual, establishment and regional level.

Jointly with the RWI Essen and the SOEP (Socio Economic Panel) group, the FDZ currently works on the linkage of labour market microdata to the SOEP on a disaggregated level, i.e. postal code areas. One aim of this project is the evaluation of non-market interactions and neighborhood effects.

One of the most visible data sets is the Linked-Employer-Employee data set called LIAB. In this data set, administrative individual employment history data are linked to an establishment survey. Because neither the survey nor the administrative data are able to fulfill all of the researchers' needs, the FDZ constructed a double linked employer employee data set for further training where both dimensions (individuals and establishment) are measured with survey and administrative data at the same time (WeLL). By comparing the same variables stemming from different data generating processes, we also learn more about the quality of data.

Future developments will not end at linking survey with administrative data. In the near future, German administrative data will be linked across different data producers. For example, the KombiFiD project aims to combine firm data from different public data producers. Another project focuses on the linkage of patent data (innovators) with administrative employment data.


[^85]
# Employee flows to study firm and employment dynamics 

Karen Geurts, Monique Ramioul, and Peter Vets ${ }^{1}$


#### Abstract

Statistics on labour market dynamics which are derived from administrative firm-level data sets are biased due to failures in the longitudinal linking of firm records. This paper presents a method for record linking based on information on the continuity of the firm's work force. Our point of departure is a linked employer-employee data set covering all private employment in Belgium. We describe the linkage algorithm and how it can be used to obtain more accurate statistics. The aim of this paper is to demonstrate that a substantial quality improvement of statistics on labour market dynamics can be reached by using simple criteria. We illustrate this by showing the impact of the method on commonly used measures of firm and employment dynamics. The method is the result of close collaboration with the Statistics Department of the Belgian Social Security Office.


Key Words: Firm dynamics; Job creation and destruction; Linked employer-employee data; Firm-level micro data; Longitudinal record linkage.

## 1. Introduction

Overestimation in measures of firm and employment dynamics, based on large administrative data sets, is well understood (Spletzer, 1998; Brandt, 2004; Abowd and Vilhuber, 2005). Missing links between different identifiers of one and the same employer, and between firms involved in a restructuring, results in an upward bias in dynamics statistics. Commonly applied methods to address these problems use probabilistic matching based on similarities in partial firm identifiers. Such matching techniques, however, require careful manual review and do not fully capture firm restructuring events such as mergers, takeovers, split-offs, and other forms of restructuring. Furthermore, calculating accurate job creation and destruction statistics of firms in restructuring causes additional difficulties (Pinkston and Spletzer, 2004; Eurostat/OECD, 2007, p. 26; Ahmad, 2008, p. 132). This is where the need for an alternative approach comes in. Recently, linkage methods are being developed which are based on employee flows between firms (Persson, 1999; Korkeamäki and Kyyrä, 2000; Baldwin et al., 2002; Mikkelson et al., 2006; Benedetto et al., 2007). These methods are in an experimental stage but yield promising results: the main advantages are that the results are easily reproducible and, even more important, that these methods allow the observation of changes in firm structure that correspond to real economic events. In keeping with this novel approach, this paper presents a method for the longitudinal linking of firm records by using employee flow information. The aim is to demonstrate that a substantial quality improvement of statistics on labour market dynamics can be reached by using simple criteria. We illustrate this by showing the impact of the method on commonly used measures of firm and employment dynamics. The study relies on a linked employer-employee data set which is maintained by the Belgian National Social Security Office (NSSO). The NSSO data set covers $99 \%$ of Belgian employers and $90 \%$ of total employment (only part of public employment is not covered).

[^86]
## 2. Employee flows to link longitudinal data of firms

### 2.1 Linking firm records by using employee flows

The idea of the employee flow method is simple: a firm is considered continuous when one of the main production factors, i.e. the work force, is largely the same at two points in time. A similar definition of firm continuity can be found in Eurostat/OECD (2007, p. 26) and Ahmad (2008, p. 132). Thus, if firm $A$ at time $t-1$ and firm $B$ at time $t$ employ (partially) the same work force, then $A$ and $B$ relate to (parts of) the same firm. Hence, continuity of employment is used as the main criterion to establish links between unmatched records of the same, or parts of the same firm. Technically, this is observed in the data set as a 'flow of a cluster of employees' from firm record $A$, at time $t-1$, to firm record $B$, at time $t$. In this case, a relationship between a 'predecessor' $A$ and a 'successor' $B$ can be established. The continuity of (part of) the work force between a predecessor and successor firm gives a strong indication that they relate to the same, or parts of the same firm. The predecessor and successor might be one and the same firm that changed its identification code, or they might be different firms involved in a restructuring event such as a merger or a split-off.

### 2.2 Significant clusters of employee flows

What the minimum 'significant' size of the cluster of employees should be in order to establish a link between such a predecessor and successor is, however, the subject of academic discussion. Depending on the aim of the study, different relative or absolute thresholds are proposed. The drawback of a low cut-off level is that it risks including a considerable amount of mobility of individual employees, which are simply job changes. A high threshold, on the other hand, risks failing to capture restructuring events involving small firms.

In this study, we take all transitions of at least 10 employees as a point of departure. These are considered 'significant' employee flows on the basis of which a link between a predecessor and successor firm is established. As to movements of less than 10 employees, we believe there is a high probability that the results of an employee flow approach are affected by individual employee mobility. Therefore, we suggest that the present approach is complemented with other methods. We also restrict the analysis to predecessor-successor relationships which involve at least one firm entry or exit. Relationships between two continuous firms are not included.

## 3. Identifying spurious firm openings and closings

Once predecessor-successor relationships have been established, it is possible to identify spurious firm openings and closings. The latter are removed from the total population of firm entries and exits, hence allowing the more accurate estimation of firm dynamic measures.

Spurious openings and closings are defined as following. Let $N_{i q}$ be the number of employees working at firm $i$ in quarter $q$, and $N_{j q-1}$ the number of employees working at firm $j$ in the previous quarter $q-1$. Let now $N_{j q-1, i q}$ be the number of employees which are working at at firm $j$ in quarter $q-1$ and also at firm $i$ in quarter $q$. Then, according to or definition, firm $i$ is a successor of a predecessor $j$ if $N_{j q-1, i q} \geq 10$. Hence, firm $i$ is a spurious opening in quarter $q$ if $N_{j q-I}=0$ and $N_{j q-1, i q} \geq 10$. In other words, new entrants of firms into the population which do not start from scratch but with a 'significant' part of the work force coming from another firm in the previous quarter, are considered as spurious openings.

Similarly, spurious firm closings are exits from the population of which a 'significant' part of the workforce makes a transition to another firm. Here, an additional condition is imposed to prevent bankruptcies, which are true closings entailing real job destruction, to be defined as spurious closings. In case of a bankruptcy, an important share of the work force may be recruited by local competitors, certainly in industries facing labour market shortages. NSSO experience indicates that $75 \%$ is a natural upper cut-off value for this kind of employee transitions in two successive quarters. Hence, a spurious closing is technically defined as an exit of firm $i$ from the population, of which a 'significant cluster' of employees moves to another firm $j$, and of which at least $75 \%$ of the employees is employed by one or more other firms $k$ in the next quarter.

In line with international definitions on firm demography, we are now able to distinguish between real firm openings or closings, and spurious ones. The latter are entries into or exits from the population which are not the result of the creation of new factors of production or the destruction of existing ones, but which emerge from changes in firm identifiers or firm structure. In the period of observation, 2003/04 to 2007/08, a stable annual share of about $0.8 \%$ of the firms in the NSSO register coincides with a spurious opening or closing, or with a predecessor or successor firm linked to one of these. This is a very small amount of firms, but as we will see below, they represent an important share of total entries and exits of firms with more than 10 employees. What is more, they are responsible for a disproportionately large share of apparent job creation and destruction.

## 4. Adjusting measures of job creation and destruction

The quality of job reallocation measures depends on the correct identification of expanding and opening firms (involving job creation) and of contracting and closing units (involving job destruction). Mistaking a spurious opening for real results in an overestimation of job creation by the opening firm, and of job destruction by the predecessor(s). Similarly, taking a spurious closing for real leads to an overestimation of job destruction by the exiting firm, and of job creation by the successor(s).

Although it is well understood how failures in the longitudinal linking of firm records induces a bias in statistics on firm and employment dynamics (Abowd and Vilhuber, 2005; Brandt, 2004), most national and comparative studies do not take into account this information when constructing job creation and destruction measures.

In the full version of this paper, we present the formulas for the correction of firm level job creation and destruction. Correction of firm level job creation and destruction is achieved for every pair of years, $t-1$ and $t$, by correcting the number of jobs at firm level in $t-1$. Employment levels in $t$ are left untouched; these are considered as the correct point of departure for the calculation of job reallocation measures.

The basic assumption of the correction is simple: the jobs of the predecessor are assigned to the successor. When, for example, firm $A$ changes identification number into firm $B$, correction implies that the original jobs of $A$ in year $t-1$ are assigned to $B$ in $t-1$. In case of multiple successors, or of a predecessor which continues in year $t$, a decision has to be made about the proportional distribution of the predecessor jobs in $t-1$. This is necessary because we do not know the exact size of the labour reallocation that has taken place between firms, since the observed clustered employee flow may be affected by individual mobility of employees. We apply the most natural decision rule in each of the cases, and start from the relative sizes of the firms in year $t$ to decide about the proportional distribution of the corrected number of jobs in $t-1$.

## 5. Impacts on measures of firm dynamics and of job creation and destruction

In this section we illustrate the impact of the employee flow approach on common measures of firm and employment dynamics by means of two examples. More examples can be found in the full version of the paper.

### 5.1 Firm dynamics

As explained before, we follow the Eurostat / OECD recommendations for the definition of firm openings and closings ('births and deaths'). Table 5.1-1 reports the numbers of openings and closings before and after correcting for spurious events. In order to sharply illustrate the impact of our approach, the data in Table 1 are restricted to openings and closings of firms with more than 10 employees. Since 10 is the minimum threshold that is set for identifying clustered employee flows between firms, the correction has no impact on firms with less than 10 employees in the opening or closing year. Because our data set does not contain information on firm employment in 2002 and 2009, openings can only be defined from 2004/05 on and closings only until 2006/07

Table 5.1-1
Number of firm openings and closings before and after correction

|  | Firm openings |  |  | Firm closings |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Uncorrected n | Corrected <br> n | Difference <br> \% |  | Uncorrected n | Corrected <br> n | Difference <br> \% |
| 2004-05 | 847 | 519 | 38.7 | 2003-04 | 1016 | 629 | 38.1 |
| 2005-06 | 868 | 471 | 45.7 | 2004-05 | 949 | 613 | 35.4 |
| 2006-07 | 915 | 461 | 49.6 | 2005-06 | 977 | 536 | 45.1 |
| 2007-08 | 804 | 436 | 45.8 | 2006-07 | 961 | 527 | 45.2 |

Belgium, employer firms with more then 10 employees, NACE Rev. 1.1 C to K, 2003-2008
Source: NSSO - HIVA-K.U. Leuven
In all years of observation, the number of firm openings is strongly reduced after correcting for spurious events: up to $49.6 \%$ of registered firm entries with more than 10 employees do not correspond to the real opening of an employer firm, but results from a change in firm identifier or a restructuring event. Similarly, up to $45.2 \%$ of registered firm exits are not real closings, but firms which change identifier or which are taken over by another firm.

Although spurious openings make up only a small share of total firm entries, they represent a disproportionately large share of total employment created by opening firms and thus induce a strong upward bias in job creation measures. The same goes for job destruction by closing firms. This is where the importance of the correction method becomes clear. Figure 5.1-1 shows annual estimates of aggregated job creation and destruction by opening and closing firms, before and after correction.

Figure 5.1-1
Job creation and destruction by opening and closing firms, before and after correction


Belgium, NACE Rev. 1.1 C to K, 2003-2008
Source: NSSO - HIVA-K.U. Leuven
Annual job creation by new firms on the basis of uncorrected data ranges from 54000 to 68000 jobs in the period of observation. Between $35.6 \%$ and $48.5 \%$ of this apparent job creation is, however, associated with spurious openings. After correction, annual job creation by new firms is revised downward to a range of 35000 to 36500 jobs. Not only
is this a considerable decrease, the impact of the correction is also that the large annual differences in job creation by firm openings are flattened out. The result is a relatively stable number of jobs created by new firms every year, which certainly is a more realistic picture of the contribution of firm openings to employment in the period considered. The effect of correcting for spurious closings on job destruction by exiting firms is similar.

### 5.2 Job creation and destruction

Finally, we turn to the impact of the employee flow approach on measures of job creation and destruction. In Table 5.2-1, uncorrected and corrected measures are compared: both absolute numbers and rates of job creation and destruction are reported.

Table 5.2-1
Total job creation and destruction, before and after correction

|  | Job creation |  |  | Job destruction |  |  | $\begin{gathered} \text { Net } \\ \text { growth } \end{gathered}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Uncorrected <br> (a) | Correcte d <br> (b) | $\begin{gathered} \text { Diff. } \\ (\mathrm{a}-\mathrm{b}) /(\mathrm{a}) \end{gathered}$ | Uncorrected <br> (a) | Corrected <br> (b) | $\begin{gathered} \text { Diff. } \\ (\mathrm{a}-\mathrm{b}) /(\mathrm{a}) \end{gathered}$ |  |
|  | n | n | \% | n | n | \% | n |
| 2003-04 | 241000 | 210500 | 12.7 | 199200 | 168600 | 15.4 | +41 900 |
| 2004-05 | 228800 | 196700 | 14.0 | 178500 | 151200 | 15.3 | +45 500 |
| 2005-06 | 224500 | 198100 | 11.8 | 183200 | 156600 | 14.5 | +41 400 |
| 2006-07 | 260500 | 217300 | 16.6 | 204500 | 161500 | 21.1 | +55 800 |
| 2007-08 | 258400 | 232600 | 10.0 | 175100 | 149500 | 14.7 | +83 100 |


|  | Job creation rate |  |  | Job destruction rate |  |  | Net |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | Uncorrected | Corrected | Differ ence | Un-corrected | Corrected | renc e | rate |
|  | (a) | (b) | (a-b) | (a) | (b) | (a-b) |  |
|  | \% | \% | ppt | \% | \% | $p p t$ | \% |
| 2003-04 | 8.4 | 7.4 | 1.1 | 7.0 | 5.9 | 1.1 | +1.5 |
| 2004-05 | 7.9 | 6.8 | 1.1 | 6.1 | 5.2 | 0.9 | +1.6 |
| 2005-06 | 7.6 | 6.7 | 0.9 | 6.2 | 5.3 | 0.9 | +1.4 |
| 2006-07 | 8.7 | 7.2 | 1.4 | 6.8 | 5.4 | 1.4 | +1.9 |
| 2007-08 | 8.4 | 7.5 | 0.8 | 5.7 | 4.8 | 0.8 | +2.7 |

Belgium, all private employment except NACE Rev.1.1 A-B-L-P-Q, 2003-2008
Source: NSSO - HIVA-K.U. Leuven

Once again, the correction method reveals a strong upward bias in measures based on raw data. During the five years of observation, on average $13 \%$ of registered total job creation and $16 \%$ of registered total job destruction does not reflect real employment dynamics, but is the result of clustered employee flows between firms. A large part of this is accounted for by spurious firm openings and closings, as illustrated in the previous paragraphs. The other part is due to firm restructurings, such as firms taking over (parts of) other firms, mergers, or split-offs of parts of existing firms.

In line with the conclusions in the previous section/paragraphs, annual fluctuations are reduced after correction. The period of observation (2003-2008) refers to five years of stable and moderate employment growth in Belgium
without any important aggregate fluctuations. Uncorrected estimates, however, show strong annual variation in job creation and destruction levels. After correction, annual variation in job creation and especially job destruction is reduced (by $10 \%$ and $40 \%$ respectively). Hence, it is found that a relatively stable number of jobs is created and destroyed every year, which certainly is a more realistic picture of firm and employment dynamics in the years under consideration. In other words, strong peaks in annual flow measures appear to be administrative artefacts.

Rates of job creation and destruction allow the comparison of results across sectors and countries (for definitions see full paper). These rates are reduced by 1.1 to 1.8 percentage points after correcting for employee flows, as is shown in Table 5.2-1. This results in real average job creation and job destruction rates of $7.1 \%$ and $5.3 \%$ respectively in the period under consideration.

## References

Abowd, J.M. and Vilhuber, L. (2005). The sensitivity of economic statistics to coding errors in personal identifiers, Journal of Business and Economic Statistics, 23:2, pp. 133-152.

Ahmad, N. (2008). A Proposed Framework for Business Demography Statistics, in E. Congregado (ed.) Measuring Entrepreneurship: Building a Statistical System, New York: Springer, pp. 113-174.

Baldwin, J.R., Beckstead, D. and Girard, A. (2002). The importance of entry to Canadian manufacturing, STI Working Paper 2002/3, Paris: OECD.

Benedetto, G., Haltiwanger, J.C., Lane, J. and McKinney, K. (2007). Using Worker Flows to Measure Firm Dynamics, Journal of Business and Economic Statistics, 25:3, pp. 299-313.

Brandt, N. (2004). Business dynamics in Europe, STI Working Paper 2004/1, Paris: OECD.
Eurostat / OECD (2007), Eurostat-OECD Manual on Business Demography Statistics, Luxembourg: Office for Official Publications of the European Communities.

Korkeamäki, O. and Kyyrä, T. (2000). Integrated panel of Finnish companies and workers, VATT Discussion Paper 226, Helsinki: Government Institute for Economic Research.

Mikkelson, G., Unger, L.I. and LeBel, D. (2006). Identifying and accounting for mergers and acquisitions in measuring employment, BLS Statistical Survey Papers 2006, Washington: Bureau of Labor Statistics.

Persson, H. (1999). Job flows and worker flows in Sweden 1986-1995, in H. Persson (ed.) Essays on Labour Demand and Career Mobility, Stockholm: Swedish Institute for Social Research, pp. 1-84.

Pinkston, J.C. and Spletzer, J.R. (2004). Annual measures of gross job gains and gross job losses, Monthly Labor Review, 127:11, pp. 3-13.

Schrör, H. (2008). Business demography in Europe: employers and job creation, Eurostat Statistics in focus 100/2008, Luxembourg: Office for Official Publications of the European Communities.

Spletzer, J.R. (1998). The contribution of establishment births and deaths to employment growth, BLS Economic Working Paper 310, Washington: Bureau of Labor Statistics.

# Dealing with inconsistencies in linking administrative data 

Daniela Hochfellner, Axel Voigt ${ }^{1}$


#### Abstract

Using administrative data for research gets more and more popular because of their richness of information on individuals. Due to the rising interest we have started a project to merge German administrative data of two social security agencies, the Federal Employment Agency and the Pension Insurance, to improve the quality of register-based data. In the case of Germany these agencies collect data on employment histories of individuals which are relevant for their own field of activity. The linkage will lead to a unique employment biography dataset for researchers worldwide. Although there is a uniform identifier for the linkage, serious specific problems occur. There are inconsistencies, which result either of the fact of different editing procedures at the agencies or possible mistakes during the data collection that lead to the existence of multiple states in the data. The cleansing procedures for the first mentioned inconsistencies can be easily executed with mathematical conversion algorithms, whereas the second mentioned inconsistencies are much harder to deal with. To cope with these multiple states we analyse sequences of all possible states to identify misfits in the combined data. Finally these identified misfits are corrected with heuristics based on certain assumptions regarding the Code of Social Law. Our paper will deal with how we cope with these various types of inconsistencies and give information of the cleansing procedures we use to improve the content of administrative data.


Key Words: Data Linkage; Administrative Data; Data Cleansing; Process Generated Data.

## 1. Introduction

In the project "Biographical data of selected social insurance agencies in Germany" (BASiD), assisted by the Federal Ministry of Education and Research, German administrative data on individuals of two social security agencies, namely the Federal Employment Agency and the German Pension Insurance, are merged. The aim of the project is to generate a combined longitudinal biography dataset through filling up gaps in the single data sources by using the information of the other data source. This requires a multitude of information. The project combines four different data sources, which are described in the next section. The joint research project was started in the beginning of 2009 with a current time of three years. At the end of the project the developed dataset will be provided to the scientific community as Scientific Use File as well as a weak anonymous dataset accessible by on-site use.

The collection of the different data sources is based on the same process, namely the notification process of the social security. Therefore a unique identifier is available for the data integration. It can be assumed that there is no problem when merging the various data sources with the unique identifier. Unfortunately there are still difficulties which occur during the linkage. In our case the information is stored in different formats in the multiple data sources. Furthermore we discovered inconsistencies which can lead to an existence of implausibel multiple states in the final data. The mentioned problems occur because each agency has its own way of collecting and processing the data. Also only the relevant information for their own field of activity is collected. In terms of this setting it is not remarkable that only ten percent of the observations match perfectly when doing the data intergration. In order to get more identical matches we developed strategies and heuristics to find the identical observations in the various data sources. This paper will give an overview of the strategies we used to combine the different administrative databases.

[^87]
## 2. Data Sources

This section briefly introduces the collection and the structure of the data the German Pension Insurance and the Federal Employment Agency hold. Afterwards the characteristics of the different registers we used for our linkage are described.

### 2.1 The German Pension Insurance

To have a pension insurance account is obligatory for all employed persons in the private and public sector. Pension contributions mainly depend on earning points individuals get for their employment histories. For every employment notification of the social security system individuals get earning points dependent on the corresponding wage. The higher the achieved earning points, the higher the pension payments. Additionally, in the case of unemployment, pension contributions are paid out of the unemployment insurance. Hence individuals get little pension payments even for times in which they are not employed. Furthermore times of long term illnesses are associated to pension contributions. During this time the health insurance is responsible for the payments. These illustrates that various states in life are taken into account for annuity computation. Consequently $90 \%$ of the German population have an account at the German Pension Insurance (Himmelreicher and Stegmann, 2008).

### 2.1.1 Sample of insured persons and their insurance accounts

The Sample of insured persons and their insurance accounts (VSKT) is an annually survey of the German Pension Insurance. The sample is drawn from all persons with an active insurance account at the end of each year. It provides information about every circumstance that is relevant for pension computations of the insured individuals. This means that the life situation of an insured person can be reconstructed at different points in time. Normally the recorded time span starts at the age of 17 and lasts until the date when a person gets her first pension payment. The information on the lives of the individuals is available since 1938.

### 2.2 The Federal Employment Agency

The Federal Employment Agency (BA) is the labour market's biggest service provider. One area of responsibility is the administration of the compulsory unemployment insurance. The calculations of the amount of benefit an unemployed individual gets are drawn of the social security notifications. Another area of responsibility is the placement offer and the consultation of the unemployed. During these administrative processes a lot of records are generated. The Institute for Employment Research (IAB), which is part of the BA, is allowed to generate and hold historical datasets out of these records.

### 2.2.1 The Benefit and Recipient History

The Employment History and Benefit Recipient History (BLH) is an individual dataset that includes information of different data sources. In the first place the data contains information about times of employment, which is stored in the form of a history dataset. It covers the time span from 1975 on. Since 1st April 1999 notifications about marginal part-time employment are recorded additionally. These records regarding the employment history of individuals liable to social security are supplemented with information of the internal procedures of the Federal Employment Agency. All deregistration notifications of the receipt of unemployment benefit, unemployment assistance or maintenance benefit since 1975 until the 1st January 2005 are added.

### 2.2.2 The Integrated Employment Biographies

On 1st January 2005 the receipt of unemployment assistance and maintenance benefit was pooled together and is now called unemployment benefit II. From that time on this information is stored in the Integrated Employment Biographies (IEB). Additionally the dataset contains references to times of job-seeking and times of participation in programmes of active labour market policy. However one has to take into consideration that information we use from the IEB is initially available since 1st January 2000 (Jacobebbinghaus and Seth, 2007).

### 2.2.3 The Establishment History Panel

The Establishment History Panel (BHP) is generated by the aggregation of the single social security notifications to the establishment level at 30th June each year. Therefore it contains every establishment in Germany that employs at least one person liable to social security at that point in time. Since the 1st January 1999 this is also true for establishments with at least one marginal part-time employee. The BHP is constructed by yearly cross-sections since 1975 in the case of establishments in West Germany and since 1992 for establishments in East Germany (Spengler, 2009).

## 3. Data Linkage

The linking and cleansing process of the different data sources was done in successively arranged steps. The executed routines to merge the different data sources are described in the following sections.

### 3.1 Data integration

The data of the Federal Employment Agency as well as the data of the Pension Insurance have a uniform identifier available: the social security number. Unfortunately this identifier is not enough for a successful data merging. The linkage of both data sources was done via the identifier, begin and end date of the episode, the actual state in the employment history of the individual and the daily wage. During the data merging, there were inconsistencies turning up. Due to this matter of facts it was not remarkable that only 10 percent of the observations match perfectly. Therefore we had to develop strategies to find the identical information in the different data sources.

### 3.2 Data Cleansing

One of the found inconsistencies can be explained through different observation periods. An episode splitting to construct identical observation periods was executed. After the episode splitting, the information in the data has to be adjusted to the new observation period. Another explanation why we could not find the identical observations during the data merging is that the information is stored in different formats in the single data sources. Because the daily wage is an identifier in the integration routine the format of this variable has to be adjusted, too. Three reasons apply why the daily wage differs between the two data sources. This makes mathematical algorithms necessary to adjust the information. First of all the calculation of the daily wage is based on working days as well as calendar days in the data sources of the Federal Employment Agency, while the German Pension Insurance uses calendar-days continuously. Second the data of the German Pension Insurance do not display the wage that is really earned by a person, but the wage that is relevant for the calculation of pensions. Because of that, the daily wages employees in Eastern Germany earn have to be converted corresponding to annex 10 SGB VI § 256. Finally the adjustment of the currency change has to be executed.

### 3.3 Comparison of simultaneous observations

At this time of the data integration process all the identifiers, we use to find identical observations, show the same structure and format. The next step is to compare the identifiers of simultaneous observations in the different data sources with regards to find identical information, in the following called "twin-spell". One "twin-spell" can be equated with the existence of two identical observations. The searching routine is done by dividing the merged data in subsamples which are contemplated separately. The executed searching routine is described in Figure 3.3-1.

Figure 3.3-1
Comparison of the simultaneous observations


The data of the Pension Insurance has one crucial advantage, namely the account clarification. This information shows whether the reported notification is proofed. From the age of 30 on, employees liable to social security get a regularly information writing containing the employment times that are relevant for their annuity computation. This way originated mistakes are recognized and corrected. The data of the Federal Employment Agency do not have this correction loop. Accordingly we did not compare simultaneous observations regarding the employment state and the daily wage if the insurance accounts are clarified. We assumed that the information of the Pension Insurance is the correct one for all considered combinations and did overwrite the corresponding information of the Federal Employment Agency. So, all observations that are clarified are recognized as "twin-spells". For the simultaneous observations, which are not clarified, the employment state and the daily wage were compared. If the employment state was identical, we declared the simultaneous observations as "twin-spells" regardless of the daily wage, because even if the daily wage differs, the considered observations can be seen as "twin-spells" concerning the employment state. More difficult was the searching for "twin-spells" in the case of no identical employment state. In the executed routine we differentiated between the attribute identical wage and no identical wage. If the wage was identical the employment state was corrected. The final decision for the correct employment state was made by comparing the information on the considered observations with the code of social law. Additionally plausibility checks within the dataset were executed. The only simultaneous observations which could not be identified as "twin-spell" were differing in the employment state as well as in the daily wage.

### 3.4 Information transmission

The aim of this step is to transfer information for the located "twin-spells" in the data. For each of these spells the information is compressed onto one single observation. In the developed routine the information on the observation of the German Pension Insurance has been transferred onto the observation of the Federal Employment Agency. After the transmission the duplicate is deleted. The transmission loop has to be executed several times, because the existence of multiple parallel episodes is possible.

### 3.5 Comparison with the Code of Social Law

For the existence of multiple employment states, the so called "no twin spells" has to be proven, if their combination is possible. The checking was done by analysing sequences of different states that appear at the same time by comparing their existence with the code of social law. The following example shall illustrate the procedure: In the merged dataset there may be a person that first receives unemployment benefit and secondly is a job-seeker in the data of Federal Employment Agency. At the same time this person is also a marginal part-time employee. While marginal part-time employment and job-seeking is compatible with both the receipt of unemployment benefit and unemployment assistance, it is not continuously in the case of simultaneous receipt of unemployment benefit and unemployment assistance. After the comparison with the code of social law most of the analysed sequences can be seen as conform regarding the Code of Social Law and will therefore stay in the data.

## 4. The final BASiD data

The combined BASiD data will differ in certain characteristics from the previous existing datasets. It contains a variety of characteristics, which allows the researchers to deal with research questions that could be answered for Germany only less precise in the past. Another benefit is that the dataset contains complete employment biographies of individuals (Hochfellner and Voigt, 2010). The information the data contains is exemplified displayed in Table 4-1.

Table 4-1
Information in the final BASiD data

| Information on... | original data source |  |
| :--- | :---: | :---: |
|  | Federal Employment <br> Agency | German Pension <br> Insurance |
| Employment and benefit history | $\checkmark$ | $\checkmark$ |
| Education (military and civil service) |  | $\checkmark$ |
| Times of illness |  | $\checkmark$ |
| Information on occupation | $\checkmark$ |  |
| Job seeking \& training measures | $\checkmark$ | $\checkmark$ |
| Job payments | $\checkmark$ | $\checkmark$ |
| Earning points |  | $\checkmark$ |
| Motherhood \& number of children |  |  |
| Regional \& establishment information | $\checkmark$ | $\checkmark$ |
| Sociodemographic information | $\checkmark$ |  |

Hence, analyses with regard to birth-rates and employment histories of women, the influence of military or civil service on the employment histories, life-income and earnings points for the pension or influence of start-up conditions on the career can be arranged.

## 5. Conclusion and further developments

To sum up, the developed cleansing procedures and heuristics are very promising. At the end there are about one percent of all observations left, which cannot be cleared with the routines. For these cases we assume mistakes in the data collections which can neither be defined nor corrected. These episodes will be deleted of the BASiD data. In this case the person will have an information lack for the respective time. The project is currently work in progress. The future agenda is to test the combined data and generate a Scientific Use File for the scientific community. The next step therefore is the anonymization of the data. This will be done by implementing the concept of factual anonymization.

## References

Himmelreicher, R. and Stegmann, M. (2008). New Possibilities for Socio-Economic Research through Longitudinal Data from the Research Data Centre of the German Federal Pension Insurance (FDZ-RV)", Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 4, pp. 647-660.

Hochfellner D. and Voigt A. (2010). Improving the content of administrative data by linking different register-based data sources, unpublished report, Nuremberg, German: Institute for Employment Research.

Hochfellner, D., Voigt, A., Budzak, U. and Steppich, B. (2010). Das Projekt BASiD: Biographiedaten ausgewählter Sozialversicherungsträger in Deutschland. Projektinhalte, aktueller Stand der Arbeiten und Analysemöglichkeiten, in B. G. Cox et al. (eds.) DRV Schriften, 55/2009, New York: Wiley, pp. 74-84.

Jacobebbinghaus, P. and Seth, S. (2007). The German integrated employment biographies sample IEBS, Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 2, pp. 335-342.

Spengler, A. (2009). The Establishment History Panel, Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften, 3, 501-509.

# You can match my data! Biasing effect in the use of linked administrative and survey data 

Martina Huber, Alexandra Schmucker and Stefan Bender ${ }^{1}$


#### Abstract

Surveys often cope with special problems: gaps in retrospection appear or respondents could not provide details. Sometimes these problems can be solved by using additional information from process-generated data. These administrative data offer valid and exact information, but also include potentially less valid variables with a higher share of missing values. By linking the data, the data quality can be improved by creating a dataset that balances the disadvantages of the administrative and survey data using the advantages of these types of data.


The survey data contain information about 6400 employees from 150 establishments and are sampled from and linked to administrative data via social security records.

The following topics will be analyzed:

1. Unit non-response: information from the administrative data for all non-participants is available.
2. Selectivity of the match: Because of privacy concerns, the respondents have to agree to the match.
3. Deviations: After the comparison of sex and age in both data sets, we identify deviations and correct them.
4. Differences in variables: Some variables that are "weak" in the administrative data are available in both data sets.
5. Retrospective information: Both data sets contain employment histories. We take a closer look at an overlapping period. Missing episodes in the administrative data can be explained by reported information in the survey data. Vice versa recall errors or missing data in the survey data can be corrected by linking administrative data. Furthermore, we show some hypothetical examples of deviations and identify determinants that have an influence.
[^88]
## SESSION 9C

## NON-RESPONSE

# Toward a more robust non-response propensity model: The interplay among NAEP restricted-use data, contact history information and school administrative data 

Asaph Young Chun and Fritz Scheuren ${ }^{1}$


#### Abstract

The U.S. National Assessment of Educational Progress (NAEP) conducted by the National Center for Education Statistics (NCES) is the only nationally representative and continuing assessment of what USA students know and can do in various subject areas, including mathematics and science. NAEP data at grade 12, in particular, is subject to potential non-response bias due to both a low participation rate and the correlation between NAEP survey variables of interest and non-participation likelihood.

The purpose of this paper is to evaluate the impact of nonparticipation bias on estimates of student performance in NAEP. To address this purpose we will utilize a non-response propensity model inspired by Groves and Couper (1998) and an analytical approach by Abraham, Maitland and Bianchi (2006).

We construct non-response propensity models by taking advantage of the interplay among the three datasets: NAEP restricted-use data at grade 12 , NAEP survey contact history data and school administrative data from the High School Transcript Studies (HSTS). We merge them all at an individual student level. Because the transcripts for the HSTS are collected from all students in the same NAEP sample of schools, regardless of the individual student's participation status in NAEP, the data merged between NAEP and HSTS provide an extended list of key correlates of non-response (including course-taking pattern and academic background). Thus, a more robust modeling of nonresponse propensity is feasible. We demonstrate how alternative non-response weighing adjustments derived from non-response propensity models affect NAEP estimates of student performance, and evaluate potential improvement over the current practice of NCES adjustment for non-response, relying, as now, just on sampling frame variables in NAEP.


[^89]
# Multivariate fractional imputation in survey sampling 

Jae-kwang Kim and Wayne Fuller ${ }^{1}$


#### Abstract

Imputation is a frequently used technique for handling missing data in survey sampling. Fractional imputation is proposed as a way of achieving efficient estimates and efficient variance estimation. We consider multivariate data with arbitrary missing patterns and show that calibration techniques can be used to create fractional weights. The proposed imputation method provides efficient estimates for the parameters of the imputation model and also provides reasonable estimates for other parameters. The proposed method is applicable to two-phase sampling where the second phase can be treated as the respondents. Variance estimation using a replication method is discussed and results from simulation studies are presented.


[^90]
# Low response rate and sample representativeness in an online survey of college students 

M. Louise Lawson ${ }^{1}$, Erin O'Connor ${ }^{2}$ Crystal Rouse ${ }^{2}$, Daniel Street ${ }^{2}$, and Svetlana Kulneva ${ }^{1}$


#### Abstract

We conducted an email invitation online survey of college students with a response rate of $2 \%$. The distribution of respondents was very similar to our college's distribution in terms of classification (freshman-graduate) and gender. We calculated the Schouten/Cobben/Bethlehem R-indicator based on known response propensities. As expected for a low response rate, the R-indicator was close to 1 (.9901 for class level and .9984 for gender). Responses to a classroom version of the survey were significantly different from the volunteer sample. Our results indicate that extremely low rates and volunteer response can lead to severe non-response bias even when the sample appears to match demographically to the source population.


Key Words: Non-response bias, Survey, Survey representativeness, Volunteer sample, E-mail survey

## 1. Introduction

### 1.1 Background

The use of surveys is one of the fastest-growing and most pervasive trends on college campuses, with administrations increasingly using surveys to implement shared government. With rising demands for systematic evidence and rapidly shrinking costs to gather such data, colleges and universities are increasingly surveying students, employees, and alumni to measure all kinds of things-from engagement in campus life to satisfaction with employment to opinions about implementation of new fees.

With all those surveys under way, higher-education officials are understandably worried about whether, when, and how much they can trust the results. Many studies have attempted to determine if there is a difference between respondents and non-respondents. Some researchers have reported that people who respond to surveys answer questions differently than those who do not (Benson, Booman, and Clark, 1951; Gough and Hall, 1977). Others have found that late responders answer differently than early responders, and that the differences may be due to the different levels of interest in the subject matter (Bauer, 1947; Brown and Wilkins, 1978; Reid, 1942; Speer and Zold, 1971).

Logically, a perfect 100-percent response rate would eliminate nonresponse bias, but since universal participation is rarely if ever attainable in practice, many colleges and universities focus instead on maximizing response rates. Recent research shows that the actual effect of response rate on survey accuracy is generally small and inconsistent (Radwin, 2009). It is generally agreed that a high response rate does not necessarily guarantee a representative sample, nor does a low rate indicate a non-representative sample. This agreement has been for sample response rates above $50 \%$, however (Schouten, 2009). We are concerned with the much lower response rate found in the typical campus survey, particularly of students. For student email surveys on our campus, for example, response rates are often less than $5 \%$.

[^91]
### 1.2 Impetus for Research

During Fall Semester of 2009 the Kennesaw State University (KSU) administration requested approval of a new \$35 student fee to develop an 88 acre sports park with a stadium. Prior to giving approval, the Student Government Association (SGA) wanted student input on this and other KSU strategic goals, including creation of football exploratory committee and expansion of Student Recreation and Wellness Center. Five discussion forums (called the KSU Vision) were held for the university president to present KSU's strategic goals to students in an effort to gain student support for these goals. Approximately 450 people attended live, and another 200 views occurred online from recorded version of the Vision. Of the 22,841 student body, $1.6 \%$ completed the emailed survey from SGA. These volunteer email respondents were supportive of all initiatives, including the new fee to build the sports park and stadium.

After hearing the survey results, the SGA voted unanimously to recommend a $\$ 35$ fee increase to the Student Fee Committee. During the Student Fee Committee meeting the next month, students and faculty voted unanimously to approve the $\$ 35$ fee. The Board of Regents approved the fee increase during its January, 2010 meeting. Currently, KSU women's soccer is being played in the new stadium and the sports park has been completed.

As we see, policy makers in our university often rely upon voluntary responses to emailed or online surveys to make critical decisions for the university. We usually have very low response rates on such surveys. The student government wanted to know whether its surveys could be justifiably relied upon. The university president wanted to know if the Vision presentation impacted student opinion.

## 2. Methods

### 2.1 Data collection

In November, 2009 we sent an email to all KSU faculty indicating that we would come to their class in the event that they could not attend so that we could administer the survey. We volunteered to take up homework, take roll, and give exams for the faculty member in exchange for the opportunity to administer the survey and show the Vision presentation. Approximately 20 faculty members responded, of whom 4 were specifically interested in having us conduct and discuss the survey with their research methods classes. Two faculty members requested that we give exams for them. All other faculty members were willing to give us access to their classes during times when they would be absent and had no requirements for in class activities.

For these classes, they were randomized if they were the first class surveyed within their college to either see the Vision survey before or after taking the survey. After this, the next class within that college was assigned to the opposite condition. For example, the first class from the College of Science and Mathematics was randomized to see the Vision presentation before taking the survey, so the next class from CSM took the survey first (to represent those students who took the survey without having seen the Vision presentation). Following this, classes were alternately assigned the before or after condition. For most colleges, we did not have a large number of classes, so full randomization could have led to missing data for certain colleges.

The survey was administered following distribution of an informed consent document approved by the KSU Institutional Review Board. The document informed students that this was a survey to determine whether the earlier email survey represented KSU students, and that participation was voluntary. Refusal to participate in the survey was extremely rare, except in the 3 classes in which we were giving a test (several students in each of these classes elected to "cram" rather than taking the survey). Prior to administering the survey (or showing the Vision presentation, whichever came first), we explained to the students the purpose of the research. Following completion of the survey, we took questions from the students on any topics of interest, which varied by researcher. The SGA officers (CR, DS) generally received questions about the SGA policies, while the faculty member (MLL) generally received questions about university or Board of Regents policies.

In March of 2010, while data collection was ongoing, the budget committee of the Georgia State Legislature told the Chancellor of our university system to plan for a $\$ 300$ million dollar cut in funding for the next fiscal year. The

Chancellor responded that to cover this cost in tuition would require a $75 \%$ tuition, which was widely reported in the media as actual plans to increase tuition. Following this and additional political upheaval, student opinions were significantly different about all matters pertaining to fees. Therefore, only data collected prior to the budget discussion are included in this report.

### 2.2 Data Analysis

Data from the original online survey were provided to us in an Excel document and imported into SAS, version 9.2 (Cary, NC) for analysis. The raw data was particularly unkempt, and required a great deal of recoding. Determination of college affiliation was done by the principal investigator (MLL) based on her knowledge of the college of each major or department for individuals who wrote in a major or department rather than a particular college. If it was not clear which college the individual belonged to, the information was left blank. In addition, the original design of the online survey put the question about college at the end of the survey after an open ended comment section, so a significant portion of respondents did not answer the question, presumably because they exited the survey after the comment section.

Data from the in class survey was collected on paper and entered into an Access database by one of the authors (EC). The data was imported into SAS for analysis. The results were tabulated using frequency tables with chi-square and logistic regression to predict the outcome "agree with the proposed fee," the major outcome of interest. Students were classified as agreeing if they answered "agree" or "strongly agree" to the question, "I support a $\$ 35$ student fee to complete the KSU recreation and sports park" We also did weighted frequency and regression analyses (weighting by college and major, since these were the largest differences observed between the two samples).

R-indicator calculations were made first for online survey response data, and then for the summarized online and inclass survey response data (Schouten, 2009). This was done because we wanted to keep the sample size the same. Students were categorized by college they belong to and their school year. Because all KSU students were samples in the online survey, the sample size N is the total number of KSU students.
$\mathrm{p}_{\mathrm{i}}$ - the probability of response within the category (for example, number of College of Arts juniors that are responded divided by total number of College of Arts juniors)
$\bar{p}$ - average population propensity (total number of responds divided by N )
Standard deviation of the response probabilities is

R -indicator is

$$
S(p)=\sqrt{\frac{1}{N-1} \sum_{i=1}^{N}\left(p_{i}-\bar{p}\right)^{2}}
$$

$$
R(p)=1-2 S(p)
$$

To compare the values of R-indicator for different data collection ways, we calculate their standard errors by nonparametric bootstrapping. Employing the standard errors, we determined confidence intervals for R-indicators.

## 3. Results

### 3.1 Respondent Demographics

Respondent demographics were different from the school as a whole and each other, as shown in Table 3.1.1-1. Although between 60 and $65 \%$ of our students are female, we have more $4^{\text {th }}$ year students, the most common college is Social Sciences, and we currently have significantly fewer than a quarter of our students living on campus.

### 3.2 Primary outcome responses

Table 3.2.1-1 shows the number of students who support a $\$ 35$ student fee to complete the KSU recreation and sports park separate for those students who saw Vision Presentation and those who didn't.

Table 3.1.1-1
Characteristics of respondents to the online and in class versions of the survey

|  | Online | In class |
| :--- | :--- | :--- |
| \% Female | $60 \%$ | $69 \%$ |
| Predominant Year in School | $3^{\text {rd }}$ years (26\%) | $2^{\text {nd }}$ year (28\%) |
| Predominant Major | Business (34\%) | Education (26\%) |
| \% Traditional | $74 \%$ (many missings) | $85 \%$ |
| Living situation | On campus (25\%) | On campus (27\%) |
|  | $<5$ miles (29\%) | $<5$ miles (25\%) |
|  | $>5$ miles (47\%) | $>5$ miles (48\%) |
| Graduating by Summer 2010 | $22 \%$ | $11 \%$ |

Table 3.2.1-1
Differences between online and in class responses to questions about fee increase

|  | Saw Vision Presentation |  |
| :--- | :--- | :--- |
|  | Agreed <br> (Strongly Agree + Agree) | Did Not Agree <br> (Neutral, Disagree, and Strongly Disagree) |
| Online Respondents | $89(72 \%)$ | $34(28 \%)$ |
| Classroom Respondents | $50(55 \%)$ | $47(45 \%)$ |
|  | Did Not See Vision Presentation |  |
|  | Agreed <br> (Strongly Agree + Agree) | Did Not Agree <br> (Neutral, Disagree, and Strongly Disagree) |
| Online Respondents | $129(55 \%)$ | $107(45 \%)$ |
| Classroom Respondents | $84(33 \%)$ | $170(67 \%)$ |

Of the students who saw the Vision, those who volunteered to take the survey were 2.5 times more likely to agree to the fee than those recruited in the classroom (95\% CI 1.4-4.3). Of the students who did not see the Vision, those who volunteered to take the survey were 2.4 times more likely to agree to the fee than those recruited in the classroom (95\% CI 1.6-3.5).

Of the students who volunteered to take the survey, $85 \%$ agree ( $95 \%$ CI $80-87 \%$ ) agreed that expansion of the existing Student Recreation and Wellness Center would be good for KSU. Of the students who were recruited in the classroom, only $66 \%$ ( $95 \%$ CI $61-71 \%$ ) agreed. Overall, students support the expansion of the Recreation Center. Those who volunteered online were almost 3 times more likely to agree with this question ( $95 \%$ CI 1.9-3.9.) Seeing the Vision did not impact student's opinion on this question.

Of the students who volunteered to take the survey, $73 \%$ ( $95 \%$ CI $68-78 \%$ ) agreed that a football exploratory committee should be created at KSU. Of the students who were recruited in the classroom, only $58 \%$ ( $95 \%$ CI 53$64 \%$ ) agreed. Overall, students support the development of a football exploratory committee. Those who volunteered online were about 2 times more likely to agree with forming a football committee ( $95 \%$ CI 1.9-3.9). Seeing the Vision did not impact student's opinion on the football committee.

There appears to be a difference between students who saw the Vision in terms of support of the new fee, but not support of improving the Recreation Center or the football committee. Students who volunteered on line appear to have different opinions, and overall, be more in support of the Vision initiatives than those recruited in class.

### 3.3 Sample representativeness

### 3.3.1 Weighted analysis

We divided the survey data to 40 clusters by the college student attend and the student's college year. We continue to have major differences for classroom respondents who saw Vision Presentation and online respondents who did not see Vision Presentation, so these are presented separately. There were some differences in the weighted and unweighted analyses (see Tables 3.3.1.1-1 and 3.3.1.1-2) although they varied between the online and in class respondents, and were not consistent. Overall, the weighting was not particularly informative about the
representativeness of the samples. The largest difference was in the logistic models, where weighting emphasized the large difference between traditional and non-traditional students in terms of supporting the fee among the in class students. This is most likely due to the fact that non-traditional students tend to be working parents in our population, and thus unlikely to respond to an online survey unless they feel strongly about the issue (in this case, supporting the fee, perceived by most to be about eventually having a football team). Thus, it is probable that the in class students more accurately reflect the opinions of the non-traditional students.

Table 3.3.1.1-1
Primary outcome and group weighted by college and year

|  | Saw Vision Presentation |  |
| :--- | :--- | :--- |
|  | Agreed <br> (Unweighted percent) | Did Not Agree <br> (Unweighted percent) |
| Online Respondents | $77 \%(72 \%)$ | $23 \%(28 \%)$ |
| Classroom Respondents | $40 \%(55 \%)$ | $60 \%(45 \%)$ |
|  |  | Did Not See Vision Presentation <br> Did Not Agree <br> (Unweighted percent) |
|  | Agreed <br> (Unweighted percent) | $55 \%(45 \%)$ |
| Online Respondents | $45 \%(55 \%)$ | $68 \%(67 \%)$ |
| Classroom Respondents | $32 \%(33 \%)$ |  |

We also built weighted and unweighted logistic regression models where dependent variable is student's support for a $\$ 35$ student fee to complete the KSU recreation and sports park and explanatory variables are Saw Vision (whether student saw Vision or not) and Traditional Student (whether student under 25 years old or not) separate for those students who volunteered to take the survey and those who were recruited in the classroom.

Table 3.3.1.1-2
Logistic regression models for agreeing with new fee, weighting by college and year in school

|  | Variable | Odds Ratio | Lower CL | Upper CL |
| :---: | :---: | :---: | :---: | :---: |
| Online Respondents |  |  |  |  |
| Unweighted |  |  |  |  |
|  | Saw Vision | 2.11 | 1.18 | 3.80 |
|  | Traditional Student | 2.07 | 1.10 | 3.91 |
| Weighted |  |  |  |  |
|  | Saw Vision | 2.67 | 2.46 | 2.91 |
|  | Traditional Student | 1.92 | 1.75 | 2.12 |
| In class Respondents |  |  |  |  |
| Unweighted |  |  |  |  |
|  | Saw Vision | 2.30 | 1.39 | 3.80 |
|  | Traditional Student | 1.85 | . 90 | 3,81 |
| Weighted |  |  |  |  |
|  | Saw Vision | 1.69 | 1.53 | 1.87 |
|  | Traditional Student | 4.11 | 3.63 | 4.66 |

Outcome=Agreeing to fee

### 3.3.2 R indicators

It appears that holding all things equal, the R indicator would go down as the response rate increases. We are unsure about the correct interpretation of a reduced R indicator, although our results confirm the authors' figure indicating a lack of usefulness of the R indicator in extremely small samples. As expected, R -indicator value is very optimistic because the response rate is so low. As the response rate increases a little, R-indicator drops down (Table 3.3.2.1-1).

Table 3.3.2.1-1
$R$ Indicators for online survey with in class responses added

| Response | n | Rate | R | $C I_{0.05}^{B T}$ |
| :--- | :--- | :--- | :--- | :---: |
| Online | 22508 | $1.8 \%$ | $95.68 \%$ | $(95.67-95.69)$ |
| Online+inclass | 22508 | $3.8 \%$ | $92.46 \%$ | $(92.45-92.46)$ |

## 4. Conclusions

The two different samples resulted in clearly different outcomes, although predictors remained the same within the groups. Standardizing samples to match the population on key variables gave mixed results. The R-indicator was not helpful in this low response situation. If pressed, we would conclude that the online sample was slightly more representative, but this could be misleading due to much of our in-class sample being lost after the budget crisis. We don't believe that either sample was particularly representative, but we are unable to quantify this.

As an epilogue to this research, we presented the findings to our university administration, and as a result were asked to help develop a representative sample of student opinions about willingness to pay a $\$ 100$ fee in order to develop a football team and increase funding for women's sports under US Title IX. We suggested that instead of a representative survey, the administration and SGA consider a vote, which they did. This resulted in a $33 \%$ response rate, which is higher than any survey response we have ever had among our students (to our knowledge). This was a successful alternative to trying for a representative sample of this difficult population, because in the US we are not required to have a representative sample of voters for the vote to count. The caveat to this is that administration must be willing to abide by the results of the vote for the vote itself to be considered valid.

## References

Bauer, E. (1947-48). Response bias in a mail survey. Public Opinion Quarterly 11:594-600.
Benson, S., Booman, W., and Clark, K. (1951). A study of interview refusals. Journal of Applied Psychology 35:116-119.

Brown, T., and B. Wilkins, B. (1978). Clues to reasons for nonresponse, and its effect upon variable estimates. Journal of Leisure Research 10:226-231.

Gough, H., and Hall, W. (1977). A comparison of physicians who did not respond to a postal questionnaire. Journal of Applied Psychology 62:777-780.

Radwin, D. (2009). High Response Rate Does Not Ensure Survey Accuracy. The Chronicle of Higher Education. http://chronicle.com/article/High-Response-Rates-Dont/48642/

Reid, S. (1942). Respondents and non-respondents to mail questionnaires. Educational Research Bulletin 21:87-96.
Schouten B,, Cobbe, F., and Bethlehem, J. Indicators for the representativeness of survey response. Survey Methodology, June 2009, Vol 35, No. 1, pp 101-113

Speer, D., and Zold, A. (1971). An example of self-selection bias in follow-up research, Journal of Clinical Psychology 27:64-68.

SESSION 10A

## REALTIME ACCESS TO DATA FROM MULTIPLE SOURCES

# A service platform providing remote access to data for research purposes 

Annie Bélanger ${ }^{1}$ and Carol Gilbert ${ }^{2}$


#### Abstract

In recent years, access to data obtained by linking administrative or survey files has been a major issue for both research and governance. These data have great analytical potential, but they are underutilized owing to the complexity involved in processing them, the methods of access and the resulting delays. The data access services for research purposes (DAS) at the Institut de la statistique du Québec meet these needs of researchers, and one way they do so is by facilitating remote access in real time. This access, granted subject to various rules, is secure, efficient and in accordance with confidentiality and privacy standards.


Key Words: Research data; Matching; Remote access; File masked against inadvertent identification (FMII).

## 1. Data access services for research purposes (DAS)

## Background

The Institut de la statistique du Québec (ISQ) established data access services for research purposes (DAS) in order to manage requests for access to microdata. Originally, DAS came into being through co-operation with the research community. It was in 2000 that the ISQ established its Centre d'accès aux données de recherche (CADRISQ). In 2005, a partnership with the Centre interuniversitaire québécois de statistiques sociales (CIQSS) enabled it to continue developing data access in the academic sphere. The partnership with the CIQSS not only made DAS known within the academic research community, but also enabled the ISQ to establish valuable collaborations by providing infrastructure to conduct training (classroom, computers, software, audio-visual equipment, etc.) and by providing financial support to student users.

In 2004, a partnership was established with researchers in the child health and welfare network affiliated with the Fonds de la recherche en santé du Québec (FRSQ) in order to expand the data access services for research and governance available in Quebec. It is in this context that EPSEBE (Environnement pour la promotion de la santé et du bien-être) was created, to facilitate access to data held by various Quebec departments and public agencies, to link such data and to retain them as long as required for analysis, while adhering to the principles of protecting personal information and data security. In addition to the data held by departments and public agencies, there are also data from surveys conducted by the ISQ and even data held by researchers. EPSEBE is intended for researchers in all fields, both in universities and government, in Quebec and potentially in Canada or other countries.

[^92]
## Management of requests for access to microdata

Microdata are detailed information-namely individual and not aggregated data-on persons, businesses, associations, etc. They usually consist of the responses given to questions in a survey or census, collected with a view to statistical processing. They may also be data which are collected for other purposes, such as for administering public programs, and which are used secondarily for statistical purposes.

The ISQ makes the following data available to third parties:
(1) Data from surveys whose results are prepared and published by the ISQ itself;
(2) Data from surveys conducted for a third party, generally a government body that sponsors the survey;
(3) Some statistical data that are used under licence from an external supplier, such as Statistics Canada;
(4) Records from the Registre des événements démographiques (births, deaths, stillbirths, marriages);
(5) Data resulting from the linking of survey files, the Registre des événements démographiques, statistical data used under licence and administrative files from Quebec departments and agencies.

In Quebec, microdata are confidential but may be accessible under the Access to Information Act or, in some cases, under other laws. Obtaining access is a relatively simple process when the files are owned by a single body. On the other hand, accessing cross-tabulated data or data with no unique identifier is a complex process. Not only are there the technical hurdles of linking files, identifying the relevant information and evaluating its quality, but there are also problems of legal, ethical, administrative and institutional nature. In short, this is a long and painstaking process, with no guarantee of success as to the quality of the results obtained.

The ISQ has an obligation to maintain a framework for managing researchers' requests to access data. This framework must first ensure that the use of microdata does not contravene either the applicable legislation or the commitments that the ISQ or the providing bodies made when the data were collected, and that the ultimate aim of the researchers justifies using these microdata. Secondly, this framework is necessary because the rules on access to microdata vary depending on the aim of the users, the methods of access, the consent required during collection and the sensitivity of the information.

With DAS, the ISQ ensures both that clients are served quickly and that the rules of privacy protection and confidentiality are followed. In accordance with the principles of confidentiality management, the supervision of access to microdata by DAS is carried out in three stages:

- Admissibility and feasibility: this preparatory stage consists of delimiting the request and consulting the branches or agencies concerned about the survey data or matching data. In particular, this entails considering the advice of the Commission d'accès à l'information and of scientific and ethical committees, including those of the ISQ. It is at this stage that the question on use of the data for statistical purposes is answered;
- Provisions to guarantee confidentiality: with all the necessary information in hand, it is then possible to develop an agreement with the requesting party based on the details, including licence requirements. There is a three-step procedure for ensuring that confidentiality is respected;
- Fulfilling the request in accordance with the conditions set out: the process of delivering files to users can vary, depending on the type of access, i.e. on site or remote. This process involves a number of parties performing various operations.


## Management of confidentiality

The ISQ has developed a three-step procedure for guaranteeing that confidentiality is respected:

- Individuals' confidentiality obligations;
- Analysis of risk of disclosure according to type of access;
- Analysis of risk of disclosure of release tables.

In exchange for having access to the microdata file, users make various commitments regarding confidentiality with which they and any persons working with or for them must comply. For this purpose, they sign an information processing and access agreement that includes a confidentiality agreement form. This agreement commits the ISQ to deliver the requested information under the conditions set out, while it commits the user to comply with the terms
relating to confidentiality, copyright and intellectual property, where applicable. The user maintains ownership of all copyrights associated with use of the microdata file and the products resulting from it. On the other hand, the ISQ maintains ownership of all copyrights with respect to the methodology and techniques used to produce the file.

The ISQ representative gives the user directives and instructions on physical and IT security and confidentiality. In regard to this, all analysts, researchers and assistants who are likely to disseminate results tables are considered "person[s] whose services are used by the [ISQ] director general in the exercise of the director general's functions"; as such they are subject to the confidentiality obligations set out in section 25 of the Act respecting the Institut de la statistique du Québec, and they sign a confidentiality agreement form. Furthermore, the agreements and contracts contain specific provisions to ensure that only authorized analysts or researchers and their assistants have access to the file.

Users may request to access their file:

- at the ISQ's research data access centre (CADRISQ);
- in a remote access environment.

The ISQ assesses disclosure risk according to the level of protection afforded by the place where the files will be consulted. Since the ISQ holds, handles and manages confidential data, it is essential for it to be able to protect those data against any intrusion. The security level of each area is determined by what it contains, that is, by the level of risk and harm associated with the disclosure of confidential information.

At CADRISQ, the environment is completely secured in accordance with the policy on securing premises. This policy incorporates recognized best practices in security. Also, it indicates the guidelines to follow to protect ISQ assets at all times. Many administrative, mechanical and electronic means are employed to achieve this objective.

The remote access environment created on the user's workstation includes security rules appropriate to this type of access. In particular, the workstation must be put in a place with restricted access, and it must be locked when no authorized person is present. Also, it must be equipped with an up-to-date browser and a functional secured HTTPS port. It cannot be subject to remote control and must use logical access controls (passwords) at both the computer and network level.

Even if a set of security precautions is taken, the risk of disclosure in this environment is higher than at CADRISQ. It is accordingly necessary to apply a method of controlling disclosure risk, such as masking or grouping information by data type. We will examine this in more detail in the section on remote access.

The last provision for guaranteeing the confidentiality of microdata concerns tables of results to be released. To control the risk of disclosure of confidential information through these tables, persons who may be releasing data must first determine which tables entail risks of disclosing confidential information, using a procedure suited to this purpose.

In addition to not releasing any result that could link a piece of information, even indirectly, to an individual, household or business, the user must have the ISQ representative check any result before taking it outside the ISQ technological environment by any means whatsoever and must accordingly provide the representative with all the necessary information.

In fact, the risk of disclosure varies depending on the situation. The riskiest situations relate to:

- tables produced from a survey on race or ethnicity ${ }^{3}$;
- tables cross-tabulating a variable associated with race or ethnicity;
- tables produced from an identifiable subpopulation that would be considered very small or very visible;
- tables cross-tabulating several variables.

[^93]When there is an obvious risk of disclosure, two options are available:

- do not release the table;
- apply an adequate masking technique, using the procedure established for the type of data file on which the table is based.

The ISQ representative may choose to apply stricter measures than those that this procedure recommends. If necessary, he/she may turn to methodologists in the methodology branch for support in applying the procedure.

## 2. Proven processes that meet research needs

Chart 1 summarizes, in diagram form, the path of a request to DAS made by a researcher seeking access to data for linking purposes. As noted above, the first two steps are concerned with the submission of the request and its

## Chart 1 <br> Flow chart of a request


evaluation. Once the necessary authorizations are obtained from the different data suppliers and the Commission d'accès à l'information (CAI), we can proceed with securing data exchange using a well-defined process. This secure exchange takes place in two stages so that the identifying data can never be linked to the research data. The first exchange serves to prepare the cohort or the population targeted by the researcher. The data suppliers send the ISQ the identifying data for this population and an internal key for each individual. The ISQ will perform the matching on these identifying data. In a second stage, the research data can be added. At this point, the ISQ sends back to the suppliers an anonymous key for each individual. It is only at this point that the suppliers send the ISQ the research data on the individuals identified. In the fourth stage, the ISQ links the files so that it finally has a single file with the research variables for each individual with identifying information removed. Step 5 lays the groundwork for access to the data and focuses on confidentiality management (Section 1-c), which varies depending on the mode of data access, i.e. on site or remote. As will be explained in detail below (Section 4), all the tables and analyses that the researcher wants to take out of the secure environment for release purposes must be analysed by the DAS pilot.

Lastly, at the end of the contract, all secure work areas used by the researcher are deactivated and the data are destroyed or archived.

## 3. Remote access

## Introduction

In order to provide remote access in real time, the ISQ has adopted a procedure that may seem complex and demanding in some respects, but it has the advantage of meeting requirements regarding confidentiality. In fact, between the production of linked or survey files and the output of release tables in the remote access environment, a number of actions are needed. In particular, these actions have to do with producing documents, involving researchers in the process and meeting security and confidentiality requirements. They also have to do with the parties involved (DAS pilot, person responsible for disclosure control, person responsible for operational security, etc.) who work together to deliver the product requested by the researcher as soon as possible, following the rules on confidentiality and privacy protection. Without reviewing in detail the subject of confidentiality management (Section 3), the next section will describe the methods used to meet these requirements in a remote access environment. The final section will then describe the implications for the researcher.

## Administrative and IT protections

The identification stage is preliminary to creating a remote access environment for the researcher. This stage is essential for the follow-ups to be carried out, and it begins with completion of the form "Request for remote access." This form is divided into several sections. The "user" part enables us to associate a user name with the researcher's contact information and to set up a schedule for use. This ensures that the user can be located, if necessary, and that the time frame set out in the agreement will be adhered to. The agreement is checked to ensure that it is signed so that the access process can continue. The contract number certifies it. The researcher must also sign a confidentiality agreement and take an oath. If these procedures are followed, the researcher is assigned a token number that enables him/her to enter the remote access environment in order to process the file using the applications requested.

Once the administrative procedures are completed, assurance that the physical and computer environment meet the criteria set by the ISQ is required. The necessary guarantees are obtained from the security officer designated by the researcher in the agreement on information handling and access. This person, who has also taken an oath, has the required qualifications and authority to implement, monitor and control compliance with the requirements regarding security and confidentiality in using the file masked against inadvertent identification (FMII).

In brief, this person verifies whether the workstation has the usual protections, namely a firewall to prevent it from being controlled remotely and access via password. The ISQ prefers that the workstation be located in an enclosed space where access is limited to the persons named in the agreement, and that it be dedicated to the authorized person and locked when that person is not present.

With respect to the IT environment, the terms of access require at minimum the web browser Internet Explorer, Version 6. Security is based on the CITRIX software, which provides concurrent management of a number of accesses. It ensures that only computers within the virtual private network created on the Internet can access the data. Thus, more than 50 researchers could have simultaneous and undisturbed access to the files needed for their research.

Once the file is made available in real time, the researcher must be able to work on it with complete security. The system has been designed to allow the work session to be ended in various ways. First, users may end the session themselves. However, if the user is away from the workstation, the system intervenes after a certain period of inactivity. Thus, after 20 minutes of inaction by the user, the system automatically terminates the session. However, it generates a warning five minutes beforehand, in which time the session can be extended. Finally, in the case of processing that takes a number of minutes, researchers can terminate the connection in such a way as to able to return within 24 hours and recover their work.

## Protection against inadvertent identification

Once the administrative and IT protections have been put in place, the researcher can be given valid statistical information without risk of disclosure. The agreement concluded with the researcher specifies that the ISQ will apply its privacy protection procedures, including the guidelines of its policy on the confidentiality of release tables and its policy on access to individual or household survey data. This means that the remotely accessible file must be protected against inadvertent identification. This requires processing the file to detect any individuals at high risk of being identified and then applying a masking technique to reduce this risk. This technique yields a file masked against inadvertent identification, or FMII. In the case of data on businesses, file masking is likely to pose a greater challenge. First, some information present in the file is considered strategic for businesses, and second, a few businesses may account for a sizable share of their industry's output, which makes masking more difficult. Thus, for some files, CADRISQ is the access mode to be preferred in order to protect the sensitive nature of these microdata.

Inadvertent identification is defined as the unforeseen identification, by a user, of an individual included in a microdata file with identifying information removed. This can occur when the user looks through the data and observes unusual characteristics that remind the user of a known individual. A more detailed analysis of various other characteristics may enable the user to confirm the individual's identity. Note that the chances of inadvertent identification are greater if the individual has rare characteristics that are known to the user.

The variables in a microdata file are grouped into three categories: direct identifiers, indirect identifiers and sensitive variables.

Direct identifiers are those variables that clearly identify individuals (e.g. name, address, telephone number, social insurance number). Indirect identifiers describe an observable or known characteristic of a respondent (e.g. municipality of residence, age, sex, marital status, number of children, visible minority status). All other variables are referred to as sensitive variables; they belong to people's private life (e.g. sexual behaviour, criminal past, mental health, alcohol use, drug use, suicide attempt).

Direct identifiers are never included in a file to be distributed, while sensitive variables can be distributed in their entirety.

To study disclosure risk, a set of indirect identifiers is usually selected; they are generally categorical variables. These variables are the identifiers that a user may have with respect to an individual. Indirect identifiers may be provided in a file for distribution if the values of a rare combination of indirect identifiers are not revealed.

## 4. Creating a file masked against inadvertent identification (FMII)

When we are ready to finalize the file requested by the researcher pursuant to the agreement on information handling and access, we inform him/her first of the steps to follow to ultimately have remote access to the file, and second of the changes to make to the file in order for it to be accessible remotely.

During the researcher's first visit to CADRISQ, the DAS pilot provides him/her with a preliminary analysis of the variables likely to pose a risk of inadvertent identification. This document, produced by a professional from the ISQ methodology and quality branch, will be useful for creating indexes for targeting the variables of interest and retaining only the relevant information in a form that is more concise but still useful for research.

If researchers are unable to come to CADRISQ, the DAS pilot will e-mail them a preliminary analysis of their file as well as a file description that may include frequencies of variables. These frequencies may be sent provided they meet ISQ confidentiality standards.

Researchers can take the time to carefully study their file or its description and to consider the variables at risk, so that they can make the collapsings required to create the FMII. The latter is a subset of the researcher's base file, enabling him/her to perform analyses remotely. This subset is based on groupings of variables that are useful for research but are not at risk of inadvertent identification. Researchers should clearly understand that special efforts
will be made to keep their most important variables intact but that collapsings will have to be made of the other variables.

At the stage of creating indexes, a meeting-on site, by video conferencing or by telephone-is organized between the researcher, the DAS pilot and the professional responsible for masking. Together, they examine the indirect identifiers and the groupings proposed by the researcher. The ISQ representatives explain to the researcher any risks associated with those identifiers and the proposed groupings. With the researcher, they try to determine how to modify the different identifiers and groupings (by creating indexes) so as to adequately ensure confidentiality while still meeting the researcher's analytical needs.

When the researcher has finished creating indexes and the file is ready for the subsequent masking stage, the ISQ and the researcher agree on a timetable for making the FMII available. The time required will be longer if the researcher has made few groupings, since the more identifying variables included in the FMII requested by the researcher, the more demanding and therefore more time-consuming the masking task will be. Note that researchers can ask the ISQ to do the index creation for them. However, the time period will be increased and expenses will be charged.

The masking phase consists of applying a confidentiality treatment to the remaining variables (indexes and other variables) to reduce the risk of disclosure of confidential information. To protect a combination at risk, masking techniques are used. Two sets of masking methods may be used jointly: data reduction methods and data modification methods.

Data reduction methods include grouping certain outliers, comprehensive recategorization (combining selected categories of an indirect identifier) and locally suppressing certain data, indirect identifiers or individuals (imputing a missing value). Data modification methods include adding random noise, rounding, data swapping and replacement of some values.

The professional responsible for masking contacts the researcher as needed, to inform him/her of how the procedure is going and to advise of problems encountered and solutions considered. This stage, whereby the FMII is finalized, takes an average of two months.

## 5. Release of output tables

For the release of output tables, which is the final stage of verifying information, a procedure has been established. This procedure applies to the output requests both of users working at CADRISQ and of those working remotely. First, so that survey estimates do not include biases, it is necessary to use survey weights (weights averaging 1). Thus, only weighted results are authorized to leave the ISQ secure environment, except in the case of longitudinal analyses and multi-level analyses for which some software products do not lend themselves to the use of weighting.

All research results are subjected to disclosure risk analysis by an ISQ employee before they leave the secure environment. Insofar as possible, we ask researchers to limit intermediate outputs (mainly cross-tabulations and frequencies) arising from the exploratory phase of data analysis.

It is necessary to indicate, at the beginning of the output files, the population targeted by the results (for example, the population between 15 and 64 years of age, Quebec as a whole). Subsequently, it is important to produce for a given analysis, in separate and well-identified files, both weighted and unweighted results depending on the type of statistics produced.

Methodology has developed a number of rules to limit disclosure risks. For example, no table containing fewer than five units in a cell is authorized.

Tables containing variables associated with race, ethnicity or language must be produced separately and be clearly identified. In addition, the researcher should provide all documentation (e.g. the program for constructing variables, a detailed description of the variables included in tables) that may assist the ISQ employee to understand the analyses and results.

Finally, the files to be submitted for disclosure risk analysis are placed in an output directory. The printing of intermediate products (such as detailed tables of descriptive statistics) is not permitted. The researcher then communicates by e-mail with DAS personnel, who will analyse the said files for disclosure risk. The files analyzed should be available within 72 hours (business days). When informed by e-mail, the researcher can recover them in the ISQ secure environment.

## References

Gagnon, É. and Baulne, J. (2008). Proposition d'éléments pour une politique d'accès aux fichiers de microdonnées d'entreprises. Direction de la méthodologie et de la qualité. Institut de la statistique du Québec. Document non publié.

Gagnon, É. and Des Groseilliers, L. (2006). Contrôle de la divulgation pour la preuve de concept de la plateforme EPSEBE. Direction de la méthodologie et de la qualité. Institut de la statistique du Québec. Document non publié.

Institut de la statistique du Québec. (2010). Guide utilisateur : Accès à distance aux données de l'Institut de la statistique du Québec.

Institut de la statistique du Québec. (2010). Guide de diffusion des tableaux de résultats : Environnement de l'accès à distance.

Institut de la statistique du Québec. (2006). Politique d'accès aux microdonnées d'enquêtes : individus ou ménages.

Institut de la statistique du Québec. (2004). Politique relative à la confidentialité des tableaux de résultats pour diffusion.

Institut de la statistique du Québec. (2004). Politique de sécurisation des locaux.

# The Real Time Remote Access at Statistics Canada 

David Price and Martin Lessard ${ }^{1}$


#### Abstract

Like many national statistical organizations (NSOs), Statistics Canada is facing increasing national and international demands from researchers for access to detailed microdata. Statistics Canada has recently put in place a new service that will improve access while at the same time protecting the confidentiality of the various data. The Real Time Remote Access system is essentially an on-line remote access facility that allows users to create tables, more or less in real time, on microdata or lightly masked microdata sets kept in a central and secure location. There are many challenges in the development of an on-line remote access tool, particularly regarding how to maintain confidentiality knowing the various output products released with their different disclosure rules.


[^94]
# Methods for permitting differential access to sensitive data while maintaining confidentiality 

Peter Meyer ${ }^{1}$


#### Abstract

The Research Data Center (RDC) at the National Center for Health Statistics (NCHS) was developed to provide researchers access to data that are too sensitive to be released publicly because they carry a degree of disclosure risk that is considered too dangerous by the owner of the data system or data product. The source of this comes from various characteristics of the data that include the potential for reidentification of subjects, respondents, or institutions. This paper describes the major data systems at NCHS, identifies some of the confidentiality issues associated with each system, and then focuses on the methods developed to provide researchers access to these data while maintaining confidentiality. We also introduce the data product and explain the difference between linking and merging data. For instance, linked data products, e.g., National Health Interview Survey Mortality files, are a probabilistic match between survey and administrative data to provide more information at the unit level. Alternately, merged data are typically survey data that have been appended with community or some other contextual variables. Finally, we will discuss the interplay between mathematical, sociological, and technological dimensions in protecting different types of sensitive data.


[^95]
## SESSION 10B

## SMALL AREA ESTIMATION

# Feasibility study report for measuring the shadow population in northern Alberta (2010) 

Bradley Brooks, Michael Haan, and Judy Lee ${ }^{1}$


#### Abstract

This paper describes a feasibility study report that was prepared by Statistics Canada, under a letter of agreement with the Alberta Office of Statistics and Information (OSI). The work was carried out in consultation with representatives from the Government of Alberta's OSI and the University of Alberta's Population Research Laboratory (PRL) who, working in collaboration with the Demography Working Group (DWG), a cross-ministry body of the Government of Alberta, provided the groundwork and conceptual framework which serves as a starting point for this study.


Key Words: Labour Migration; Alberta; Oil Sands.

## 1. Introduction

The purpose of this study is to examine particular developmental issues associated with undertaking the collection of information on the mobile work force in Northern Alberta - what is referred to in this report as the "shadow population". It aims to define the data, the analytical requirements, and the concepts involved in measuring this shadow population. While the principal intention of this study is to examine approaches available for undertaking the collection of information on the shadow population, and to provide an outline for the much of the work ahead, it also articulates constraints that may be present in providing the requisite information. The study builds on the work of the Government of Alberta's OSI and the University of Alberta's PRL to inform the development process to ensure that the tools designed to collect information will provide data relevant to the research objectives; that is, information that provides benchmark data on the nature and extent of the shadow population in Northern Alberta. The report presents options for the estimation of this population using survey methodologies. It presents a range of options for survey frame and sample design, data collection methods, data processing, and dissemination of the data and analytical products.

## 2. Objectives and Data Requirements

Alberta has long had 'shadow populations', or a non-permanent portion of the population that spends a significant amount of time in the province but is not counted in official headcounts of usual residents, in its midst. The presence of this population confounds planning and development and creates a gap between existing infrastructures and the people who use and depend on them. This problem exists in part because there is considerable uncertainty around how big the shadow population is.

The Alberta OSI has undertaken the investigation of this phenomenon and initiated a two-phase shadow population project. Phase 1 of the project involved a literature review, data and methodological reviews; Phase 2 assessed current methodologies, their feasibilities and possible recommendations of best practices and options to be applied to the Alberta landscape. This is in preparation of what will hopefully be an implementation of the recommended approach in the Fall of 2011.

The OSI approached Statistics Canada (STC) in the summer of 2009 to elaborate on phase 2 and provide additional options for the measurement of the mobile labour population. This phase enlisted StatCan's methodological expertise

[^96]and leveraged their existing infrastructure, while building on the work that had already been completed in phase 1 by the PRL.

This paper describes the data and analytical requirements and the concepts involved in conducting a special survey to measure the shadow population in Northern Alberta. Its main objective is to present options for the estimation of this population, including a range of options for a survey frame and sample design, data collection methods, data processing, and dissemination of the data and analytical products.

The target population comprises temporary workers in Northern Alberta whose citizenship is Canadian or foreign; whose de facto location is Alberta, but whose de jure location is elsewhere; whose employment status indicates full time employment.

The survey options presented in this report aim not only to provide overall population estimates for the target group, but to also provide estimates on these defining characteristics of the target population (age, sex, citizenship, de facto and de jure location, duration in municipality, industry and occupation of employment).

## 1. Survey Design Considerations

The consideration of survey options was guided by various analytical requirements set out by OSI. For the purpose of this study, the target population is confined to Economic Region 4880 (Wood Buffalo/Cold Lake), which is referred to hereafter as the Wood Buffalo region. The prospective survey of the shadow population is cross-sectional in nature, and contains sampling weights.

This study will produce estimates at two levels, an aggregate and a more detailed level. The aggregate level consists of the region as a whole and the detailed level consists of a split into two sub-population groups - one for the portion of the population living in residential dwellings in the Wood Buffalo region, and one for the portion of the population living in collective dwellings, including work camps, in the Wood Buffalo region.

The sample size for the recommended option is 8223 persons. The ability to meet OSI's analytical requirements will depend on the quality and coverage of the sample frame chosen for the prospective survey of the shadow population. Given that the prospective survey would produce estimates of the shadow population, which is by definition a nonpermanent part of the population, the question may arise whether the shadow population estimates could simply be added to permanent population numbers. For example, if the prospective survey is conducted in 2011, shortly after the federal census, could the shadow population estimates simply be added to the Census counts for the Wood Buffalo region? After consulting with Statistics Canada's methodologists, there is no known reason why this could not be done; however, the nature of the resulting aggregates would carry any limitations of their component data sources, in this case the Census.

## 2. Methodological Framework

A major step in measuring the shadow population in Northern Alberta is the identification of an appropriate survey frame. A good sampling frame should cover all of the population of interest and contain all of the necessary auxiliary information (e.g., collective dwelling lists; returning data from other surveys that might inform pre-screening of inscope dwellings). Two sampling frames were examined by the Special Surveys Division of Statistics Canada during phase 2 of the project and both were deemed to be acceptable for measuring the Shadow Population feasibility study. The first is the Labour Force Survey (LFS) frame combined with the Canadian Community Health Survey (CCHS) frame for the Wood Buffalo region, and the second was the Census of Population platform. Given that the Census is only conducted every five years, while the other surveys are ongoing, OSI's final decision with respect to the timelines for the prospective survey will influence which frame would be most suitable given the objectives and data requirements outlined in this report. While the LFS/CCHS option has been deemed unsatisfactory to serve as a survey frame that would yield a sample that could be used to derive estimates of the shadow population in both collective dwellings and residential dwellings, it is acceptable for producing aggregate estimates. The Census has been deemed acceptable to produce estimates at both the detailed and aggregate levels and is the frame source for the recommended survey option.

Given the nature of the census sampling frame, Special Surveys Division recommended that the Census be used to develop a frame for the prospective survey of the shadow population. Regarding timing, in the event that a survey would be launched in an inter-Censal year, an acceptable alternative would be to use the LFS in combination with CCHS to provide estimates for the residential dwellings, and to supplement this with Census information for collective dwellings, including work camps. While this alternative option would offer some marginal cost efficiencies, and would yield high quality estimates of the shadow population across the entire target area, it would not be conducive to the production of such estimates broken down by residential and collective dwellings, respectively - estimates available under the Census-based option.

## 3. Data Collection Methodology, Survey Schedule and Costs

Either option requires a mixed mode collection design, in which computer assisted telephone interviewing (CATI) is employed for residential dwellings, with a computer assisted personal interviewing (CAPI) follow-up if necessary. For collective dwellings (e.g., industry work camps, motels) each option assumes a mixed mode collection strategy, in which personal interviews are conducted for selected respondents and these are supplemented by drop-off/mailback questionnaires distributed to the remaining residents in the collective dwellings, to produce the required estimates.

Other assumptions shared by the survey options include:

- The survey will be cross-sectional.
- The content of the questionnaire will reflect the indicators of the shadow population (i.e., duration in municipality; de jure location; de facto location; citizenship status; employment status; industry of employment; age; see section 4.3).
- Interviews will be conducted by Statistics Canada interviewers.
- The interview is budgeted at 5 minutes (on average).
- Information from the Census of Population will be used to provide a survey frame for collective dwellings (section 5.3.2).
- A "census" of collective dwellings (e.g. work camps, motels) will be carried out, in which managers of the collectives will be contacted and asked to provide counts for residents who fit the target population profile. This will be followed up by administration of the survey questionnaire a sample drawn from half ( $50 \%$ ) of the collective dwellings on the survey frame, for which the managers will be asked to provide lists of such residents, or rooms housing such persons, which could then be used for sampling (sections 5.4.1, 6.1).
- Managers/Administrators of collective dwellings, including work camps, will be cooperative in participating in the survey and providing information; OSI will play an active role in facilitating the relationship between industry (e.g., Oil Sands Developers Group), local stakeholders (e.g., Regional municipality of Wood Buffalo) and the survey team (section 7.5).
- A sample of non-collective dwellings will also be collected. The target response rate for the survey will be $60 \%$ in both the collective and non-collective (private) dwellings (section 5.3).

The recommended, Census-based, option could go to field in the autumn of 2011, with a tentative six week collection window, running from the beginning of November 2011. . The alternative, LFS/CCHS option offers the possibility of a springtime collection, and could be undertaken in inter-Censal years without concerns for 'aging' of the information on the frame.
Before conducting a survey based on either of the options presented above it is recommended that the collection methods be field-tested to assess their efficiency, effectiveness and the assumptions they are based on.

## 4. Conclusion

This paper describes a cross-sectional survey in the Wood Buffalo region to collect information on temporary workers who are members of what is referred to in Alberta as the "shadow population." The prospective survey will cover a range of indicators, including how long members have lived in the Wood Buffalo region (e.g., less than 30 days or 30+ days), their citizenship, de facto and de jure locations, employment status and industrial classification,
and their ages. This information will be provided to OSI and is intended to inform the planning and delivery of government services and infrastructure in the Wood Buffalo region.

This study discusses issues concerning the objectives and data requirements, the survey design considerations, the methodological framework, the data collection methodology and it has outlined a development schedule and cost estimates based on the proposed November 2011 start date for collection activities for the recommended survey option, the design of which is based on the Census of Population survey platform and which would employ a mixed mode collection approach.

While the methods in this report have been developed for estimating the shadow population in northern Alberta, it should be noted that there is no reason that these methods could not be adapted to inform similar projects or future exercises, in other municipalities in Alberta, or even in other provinces or territories (e.g., Newfoundland, Yukon, Northwest Territories, etc.), where there are currently substantial mobile labour populations, or where such populations arise in the future. In other words, although Alberta's shadow population represents a unique form of labour mobility, it is by no means the only population for which the described methodology can be applied.

# More for less: Using statistical modelling to combine survey data and administrative sources. 

Alasdair Noble ${ }^{1}$, Stephen Haslett ${ }^{1}$, Geoff Jones ${ }^{1}$, and Dimitris Ballas ${ }^{2}$


#### Abstract

This project uses simulated data to investigate three methods for combining data sources. The methods were small area estimation, spatial microsimulation and mass imputation. In developing the simulations some insights into the three methods were gained which were not the focus of the original project but helped to shape the work and its conclusions.


Key Words: Simulations, Combining data, Small area estimation, Spatial microsimulation, Mass imputation

## 1. Introduction

The combination of administrative and survey data is becoming increasingly important in official statistics for a variety of reasons. It has the potential to reduce costs, provide more fine level statistics, reduce response burden, and ameliorate some types of privacy issues. This paper reports the computational aspects of a research project carried out for Statistics New Zealand that considered three statistical techniques which have developed separately, to assess their underlying similarities. The three techniques were small area estimation, spatial microsimulation and mass imputation. The full research report is given in Haslett, Jones, Noble and Ballas (2010).
Our main conclusions are that small area estimation and spatial microsimulation, at least in the form that we considered while expressed differently were essentially the same. Although both use linear models, microsimulation has not traditionally included estimation of standard errors. Recognising the parallels means this deficiency in spatial microsimulation methods can be overcome via the link with the known theory for small area estimation. Secondly, we concluded that the mass imputation was generally rather less effective than the other two methods.
This paper focuses primarily on the process involved in coming to these conclusions.

## 2. Methodology

We considered three methodologies that are commonly used for augmenting survey data using administrative data

- Small Area Estimation
- Spatial Microsimulation
- Mass Imputation

A description of each of these follows, which includes details of how we applied them

### 2.1 Small Area Estimation

Small area estimation includes a collection of statistical techniques designed to improve estimates from survey data through the use of auxiliary information Rao (2003). This research project restricted these techniques to the ELL method (Elbers, Lanjouw and Lanjouw, 2003) used by the World Bank in poverty estimation. In this method survey data is available for the variable of interest as well as a number of auxiliary variables. The auxiliary variables are also collected in a census. A regression model is developed from the survey data for the variable of interest and the auxiliary variables, and this is then used to predict the variable of interest for those members of the census not included in the survey. Small area estimates may then be found by aggregating the predictions for each of the census

[^97]data respondents to the required level of detail. More often than not the process is more complicated as the survey data are collected using a complex sample survey design so that specialized software routines are required that allow for the complex design when model fitting. The modelling process also needs to be handled very carefully as with large data sets many spurious relationships may appear and these may cause severe underestimation of the standard errors. These issues are explained further in Haslett, Jones, Noble and Ballas (2010).

### 2.2 Spatial Microsimulation

There is an extensive literature on microsimulation models beginning with Orcutt (1957) and Orcutt, Greenberger, Korbel and Rivlin (1961). A comprehensive review is given in O'Donoghue (2001). A more detailed list of references is included in Haslett, Jones, Noble and Ballas (2010).
A number of approaches are used for spatial microsimulation but in essence all of them involve reweighting of cross tabulations to fit margins constructed from more recent data. Generally however the choice of margins is ad hoc; there is no attempt to construct a statistically defensible model and standard errors of the estimates are neither reported nor calculated.

### 2.3 Mass Imputation

Statistics Canada (e.g. Colledge, Johnson Paré and Sande 1978; Kovar and Whitridge, 1995) and Statistics Netherlands (Kooiman, 1998; de Waal, 2000) have used mass imputation but the technique has been superseded by other methods. Once again there is a wide range of techniques used and an associated literature which is discussed in Haslett, Jones, Noble and Ballas (2010). A feature of most mass imputation techniques is that a high percentage of the data are imputed and the resulting statistics are usually not particularly accurate.
Our research compared these three techniques to try to understand any links between them and to guide practitioners in assessing which is more appropriate for a given context.

## 3. Missing Data and a Theoretical Framework

The problem all three techniques are designed to address can be viewed as a missing data problem. In the diagram below we see a small part of a large, incomplete census data set in which there is an assortment of missing cells.

In the cases of small area estimation and spatial microsimulation, the missing data are predominantly the unsurveyed observations for the variable of interest - in the case of Figure 3-1 below the missing values of the variable "lnexp". These missing values are predicted using an implicit (spatial microsimulation) or explicit (small area estimation) statistical model.

For mass imputation, variables (which may include the variable of interest) and in some cases complete records are missing for some observations. This way of viewing of the different types of missing data led to the theoretical framework below, which in turn helped in better understanding the modelling process involved in each of the techniques.

Denote the full census data by $\mathbf{C}$ and denote the area-level quantities of interest by $\varphi_{a}(\mathbf{C})$ where $\varphi_{a}$ operates on rows of census data (individual households) to produce values that are then aggregated to area level. We can then partition the census into observed and unobserved partitions where the unobserved denotes the missing data. Hence we can write

$$
\mathbf{C}=\mathbf{C}_{\mathbf{0}}+\mathbf{C}_{\mathbf{u}} \text { where subscripts o and } u \text { denote observed and unobserved data. }
$$

We now assume that $\mathbf{C}_{\mathbf{u}}$ is "like" $\mathbf{C}_{\mathbf{0}}$ in some sense and that this "likeness" (i.e. a "model") is used to infer $\mathbf{C}_{\mathbf{u}}{ }^{*}$.

Figure 3-1
Part of a data set showing different types of missing data.

| ic | ucode | tcode | div | Inexp | urban | num_hh | f_hh | electric | agind |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | 40905 | 409 | 1 | 7.194 | 1 | 8 | 0 | 1 | 0 |
| 2 | 40905 | 409 | 1 | I | 1 | 4 | 0 | 1 | 1 |
| 3 | 40905 | 409 | 1 | , | 1 | 3 | 1 | $1 / 1$ | 1 |
| 4 | 40905 | 409 | 1 |  | 1 | 4 | 0 | 1 | 0 |
| 5 | 40905 | 409 |  | / | 1 | 4 | 0 | 0 | 0 |
| 6 | 40905 | 409 | 1 | / | 1 | 4 | 0 | 0 | 0 |
| 7 | 40905 | 409 | 1 | 6.6678 | 1 | 4 | 0 | 0 | 0 |
| 8 | 40905 | 409 | 1 | / | 1 | 5 | 0 | 0 | 0 |
| 9 | 40905 | 409 |  | , | 1 | 3 | 0 | 0 | 0 |
| 10 | 40905 | 409 | 1 | / \% | 1 | $1{ }^{1}$ |  | 0 | 0 |
| 11 | 40905 | 409 | 1 | 6.0834 | 1 | 7 | 0 | 1 | 1 |
| 12 | 40905 | 409 | 1 | (1) | 1 | 7 | 1 | 1 | 0 |
| 13 | 40905 | 409 | 1 | - | 1 | 3 | 0 | 0 | 1 |
| 14 | 40905 | 409 |  | , | 1 | 6 | 0 | 0 | 1 |
| 15 | 40905 | 409 |  | , | 1 | 1 | 0 | 1 | 1 |
| 16 | 40905 | 409 |  | , | 1 | M | M | M | 1 |
| 17 | 40905 | 409 |  |  | 1 | 5 | 0 | 1 | 0 |
| 18 | 40905 | 409 | 1 | / | 1 | 3 | 0 | 1 | 0 |
| 19 | 40905 | 409 | 1 | 6.2621 | 1 | 4 | 0 | 1 | 0 |
| ... | ... | ... | ... | ... | ... | ... | .. | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | .. | ... | ... |
| $\ldots$ | ... | ... | ... | ... | ... | $\ldots$ | .. | ... | ... |
| 97 | 40905 | 409 | 1 | 6.2838 | 1 | 5 | 0 | 1 | 0 |
| 98 | 40905 | 409 |  |  | 1 | 3 | 1 | 0 | 0 |
| 99 | 40905 | 409 | 1 | /, | 1 | 5 | 0 | 0 | 1 |
| 100 | 40905 | 409 | 1 | 6.2901 | 1 | 7 | 0 | 0 | 1 |

On this basis we can now define $\mathbf{C}_{\mathbf{u}} *$ for each of our techniques

- in small-area estimation, this model is explicit: $\mathbf{C}_{\mathbf{u}}{ }^{*}=\mathrm{E}\left[\mathbf{C}_{\mathbf{u}} \mid \mathbf{C}_{\mathbf{0}}\right]$
- in microsimulation, the model is implicit: $\mathbf{C}_{\mathbf{u}} * \sim \mathbf{C}_{\mathbf{u}} \mid \mathbf{C}_{\mathbf{0}}$ and random draws of households from the observed data are made according to some criteria of similarity based on the auxiliary variables.
- in mass imputation, the model is implicit: $\mathbf{C}_{\mathbf{u}}{ }^{*} \sim \mathbf{C}_{\mathbf{u}} \mid \mathbf{C}_{\mathbf{o}}$ and once again random draws of households that match the records with missing values are made.
In all techniques the estimate of the area-level summary is then:

$$
\varphi_{a}=\varphi_{a}\left(\mathbf{C}_{\mathbf{o}}+\mathbf{C}_{\mathbf{u}}^{*}\right)
$$

This framework was then used for extensive simulations to investigate the techniques' properties in more detail.

## 4. The Simulations

Two population censuses were constructed. The design structure of the censuses had the following characteristics:

| $\mathrm{N}=4$ million |  |
| :--- | :--- |
| Census 1 | 25 Areas, 250 Clusters per area, 200 Households per Cluster |
| Census 2 | 25 Areas, 25 Clusters per area, 2000 Households per Cluster |

These two census versions allowed us to compare the performance of the estimates under those two different situations, with different sized clusters. We also constructed a variable of interest $(\mathrm{Y})$ and four correlated categorical (X) auxiliary variables.

Table 4-1
Definition of categorical Auxiliary Variables

| Variable | A | B | C | D |
| :--- | :--- | :--- | :--- | :--- |
| Levels | 2 | 2 | 3 | 5 |
| Proportions | $0.4: 0.6$ | $0.2: 0.8$ | $0.2: 0.3: 0.5$ | $0.3: 0.3: 0.2: 0.1: 0.1$ |
| Example | Gender | Education | Age Groups | Ethnicity |
|  | Male/Female | Tertiary/Not |  | Pakeha/Maori/Pacific <br> Island/Asian/Other |

Y values were generated under the model:

$$
E(Y)=A+B+C+D+A: B+A: C+A: D
$$

This model ensured that the microsimulation could be carried out using three two way tables AxB, AxC and AxD.
We then considered a variety of censuses and sampling schemes in a factorial design with two levels of each factor for the two census structures, with both census and sample reselected many times, so that the comparisons could be more easily made. The following could vary in both the census and sample:

- total census size
- total sample size
- numbers of areas,
- numbers households per cluster

Using this setup it was not possible to constrain certain variables, such as the total sample size, while allowing others to change. Consequently the following combinations were used in the final simulations.

Table 4-2
Census and sampling designs for the Simulations

|  | Population <br> Areas | Clusters | Households | Sample <br> Clusters | Households | Sample |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| A1 | 200 | 50 | 100 | 2 | 10 | 4000 |
| A2 | 200 | 50 | 100 | 5 | 10 | 10000 |
| A3 | 200 | 50 | 100 | 5 | 4 | 4000 |
| A4 | 200 | 50 | 100 | 2 | 25 | 10000 |
| B1 | 50 | 200 | 100 | 5 | 10 | 2500 |
| B2 | 50 | 200 | 100 | 20 | 10 | 10000 |
| B3 | 50 | 200 | 100 | 2 | 25 | 2500 |
| B4 | 50 | 200 | 100 | 5 | 40 | 10000 |

Finally we added random noise to the variable of interest, Y, in our model.
There were no area effects in the model. This could be seen as unrealistic but we reasoned that if area effects were present in the variable of interest and were not included in the model, the estimates would be degraded but it should not affect the comparison between the techniques. However this caused some initial confusion, as when we looked at coverage of $95 \%$ confidence intervals for a single realisation of the census, the random error caused some areas to be significantly different, in effect introducing area effects. Over many realisations of the census, that is, in the superpopulation, the effect was not present.

Figure 4-1
Coverage versus bias (estimate - true value) for a single realisation of the two censuses


In Figure 4-1 we can see that the when the bias was larger the coverage fell which is as we would expect, and that the bias was lower when more clusters were sampled in each area. What is happening is that although the (superpopulation) mechanism has no area effects, for a particular realisation the census is fixed so that when repeated samples are drawn there can be (and often are) some significant area effects (e.g. around 1 in 20 at a $5 \%$ level). For a fixed census, the areas that are significant remain the same no matter which sample is drawn. For the superpopulation (i.e. many realisations of the census) no particular area effect is significantly different from zero.

## 5. The Modelling

When applying the small area estimation technique an explicit model was fitted and as we knew the model which had been used to generate the census we began with that model. The microsimulation technique also implicitly depends on a model which is defined by the particular tables that are chosen to update the census data. As described in section 4 , the variable of interest in the census was generated using a model with three two way interaction effects which defined the tables that would be used for the updating. The small area estimation technique allowed us instead to consider the best model using the AIC as a fitting criterion. We repeated the model fitting a 100 times and the following models were selected with the proportion of times for each model.

Table 5-1
Model selections

| Selected Model | Proportion of times <br> selected $\%$ |  |
| :--- | :--- | :--- |
| A+B+C+D+A:B+A:C+A:D* | 21 |  |
| A+B+C+D+A:B+A:C | 19 |  |
| A+B+C+D+A:C+A:D | 15 |  |
| A+B+C+D+A:B+A:C+A:D+B:C | 6 |  |
| A+B+C+D+A:B+A:C+B:C | 5 |  |
| A+B+C+D+A:C+A:D+B:C | 5 |  |
| A+B+C+D+A:C | 4 |  |
| A+B+C+D+A:B+A:C+B:D | 3 |  |
| A+B+C+D+A:B+A:D | 3 |  |
| A+B+C+D+A:C+B:C | 3 |  |
| A+B+C+D+A:B+A:C+A:D+B:D | 2 |  |
| A+B+C+D+A:B+A:C+A:D+B:D+B:C | 2 |  |
| A+B+C+D+A:B+A:C+A:D+C:D | 2 |  |
| A+B+C+D+A:B+A:C+C:D | 2 |  |
| A+B+C+D+A:B+A:C+A:D+B:C+A:B:C | 1 |  |
| A+B+C+D+A:B+A:C+A:D+B:C+C:D | 1 |  |
| A+B+C+D+A:B+A:C+A:D+B:D+A:B:D | 1 |  |
| A+B+C+D+A:B+A:C+B:C+B:D+A:B:C | 1 |  |
| A+B+C+D+A:B+A:C+C:D+B:D | 1 |  |
| $A+B+C+D+A: B+B: C$ | 1 |  |
| $A+B+C+D+A: C+A: D+C: D$ | 1 |  |
| A+B+C+D+A:C+B:D+B:C | 1 |  |

As can been seen in the table below the choice of model made little difference to the estimates. In the table .tr denotes use of the known true model and .ch denotes model that was chosen based on the AIC.

Table 5-2
Mean square error, coverage and $\mathbf{9 5 \%}$ confidence interval length for $\mathbf{2}$ censuses and $\mathbf{4}$ sampling schemes

| Census | Census A |  |  | Census B |  |  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :---: | :---: |
| Sample | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |  |  |
| MSE.tr | 0.0933 | 0.0839 | 0.0825 | 0.0892 | 0.0338 | 0.0249 | 0.0598 | 0.0332 |  |  |
| MSE.ch | 0.0934 | 0.0839 | 0.0825 | 0.0892 | 0.0337 | 0.0249 | 0.0599 | 0.0332 |  |  |
| Cov.tr | 0.930 | 0.938 | 0.944 | 0.931 | 0.877 | 0.937 | 0.757 | 0.860 |  |  |
| Cov.ch | 0.936 | 0.957 | 0.956 | 0.945 | 0.910 | 0.946 | 0.781 | 0.874 |  |  |
| Len.tr | 1.1390 | 1.1350 | 1.1580 | 1.1270 | 0.5880 | 0.5740 | 0.5810 | 0.5640 |  |  |
| Len.ch | 1.1370 | 1.1360 | 1.1620 | 1.1230 | 0.5880 | 0.5740 | 0.5810 | 0.5670 |  |  |

## 6. Conclusions

The principal conclusions of this work are all detailed in the report by Haslett, Jones, Noble and Ballas (2010). The main report is intended to be a useful document for practitioners interested in the links between the three techniques: spatial microsimulation, small area estimation and mass imputation. This paper has been more concerned with the simulation design and computational matters important to the development of the project and to its eventual success.

Some general conclusions can also be drawn about simulations and links between apparently different research topics.

Simulation studies seem simple and useful but:

- They rely on a model to generate the data
- If the model is too tightly specified results can be self fulfilling prophecies
- They cannot replace mathematical proofs
- Consider populations versus superpopulations (since estimates may be unbiased in the superpopulation but biased in a single realisation) and what are significant effects can change depending on context.

Simulation studies are however very useful for exploration of ideas, and can lead to a more formal specification of research problems. In our case they have allowed us to elucidate whether different academic areas are using apparently different methods (Geography - spatial microsimulation and Statistics - small area estimation) when these may in fact be the same or strongly linked.

However, the initial step we required was not computational, but was instead to better recognize that crossdisciplinary and inter-disciplinary research has an important role to play. To get to the point where studies such as ours are even possible we conclude that we need to take time to talk to each other across disciplinary boundaries.

## References

Colledge, M.J., Johnson, J.H., Paré, R., and Sande, I.G. (1978). Large scale imputation of survey data, Survey Methodology, 4, 203-224.

Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality, Econometrica, 71, 355-364.

Haslett, S., Jones, G., Noble, A., and Ballas, D. (2010). More for less? Using existing statistical modeling to combine existing data sources to produce sounder, more detailed, and less expensive Official Statistics, Official Statistics Report Series, Vol 3, http://www.statisphere.govt.nz/official-statisticsresearch/series/2010/page1.aspx ISSN 1177-5017, ISBN 978-0-478-31520-2, 2010, 75 pages.

Kooiman, P. (1998). Mass imputation: Why not!?, Research paper, BPA-no. 8792-98-RSM, Statistics Netherlands, Voorburg, 1998 [in Dutch].

Kovar, J.G. and Whitridge, P.J. (1995). Imputation of business survey data, Chapter 22 in Business Survey Methods, eds. Cox, B., Binder, D. A., Nanjamma Chinnappa, B., Christianson, A., College, M. J., Kott, P. S., John Wiley and Sons, New York.

O’Donoghue C. (2001). Dynamic Microsimulation: A Methodological Survey, Brazilian Electronic Journal of Economics, 4(2) [on-line journal], available from: http://www.beje.decon.ufpe.br/v4n2/cathal.htm

Orcutt, G.H. (1957), A new type of socio-economic system, The Review of Economics and Statistics, 39, 116-123.
Orcutt, G.H., Greenberger, M., Korbel, J., and Rivlin, A. (1961). Microanalysis of Socioeconomic Systems: A Simulation Study, Harper and Row, New York.

Rao, J.N.K. (2003). Small Area Estimation. Wiley, New York.
de Waal, T. (2000). A brief overview of imputation methods applied at Statistics Netherlands, Netherlands Official Statistics, 15, Autumn 2000, 23-27.

# Estimation of monthly unemployment figures in a rotating panel: The use of auxiliary series in a structural time series model 

Jan van den Brakel and Sabine Krieg ${ }^{1}$


#### Abstract

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. Recently an estimation procedure that is based on a multivariate structural time series model is adopted to produce monthly official statistics about the labour force. This approach handles problems with rotation group bias and small sample sizes in an effective way and enables Statistics Netherlands to produce timely and accurate estimates about the situation on the labour market. In this paper the time series model is extended by incorporating an auxiliary series about the registered unemployed labour force. It appears that the trend of the series of people registered as unemployed in the register of the Office for Employment and Income is cointegrated or almost cointegrated with the trend of the estimated unemployed labour force of the LFS for several domains. This results in a considerable decrease of the standard errors for the monthly unemployed labour force.


Key Words: Cointegration, Kalman filter, Rotation group bias, Survey error, Small area estimation.

## 1. Introduction

Official monthly figures about the labour force in the Netherlands are based on a multivariate structural time series model. Statistics Netherlands implemented this estimation procedure, originally proposed by Pfeffermann (1991), in 2010 to deal with the relatively small monthly sample sizes of the LFS to account for the rotating panel design of the sample survey. This estimation procedure is reviewed in Section 2. This paper explores the possibilities to further improve this estimation technique by incorporating available auxiliary information about the number of people registered as unemployed. To this end, the trends of this auxiliary series and the LFS are compared in Section 3. The time series model, as described in Section 2, is extended in Section 4 to incorporate the auxiliary series information. Estimation results are presented in Section 5. The paper concludes with a discussion in Section 6.

## 2. Estimation of monthly figures about the labour force

The Dutch Labour Force Survey (LFS) is based on a rotating panel design. Each month a sample of addresses is drawn and data are collected by means of computer assisted personal interviewing (CAPI) of the residing households. The sampled households are re-interviewed by telephone (CATI) four times at quarterly intervals.

The estimation procedure of the LFS for monthly figures starts with the generalized regression (GREG) estimator, developed by Särndal and all.. (1992). There are two major problems with the rotating panel design of this survey and the way that the GREG estimator is applied in the estimation procedure. First there are substantial systematic differences between the subsequent waves of the panel due to mode- and panel effects. This is a well-known problem for rotating panel designs, and is in the literature referred to as rotation group bias (RGB). A second problem is that the monthly sample size of the LFS is too small to rely on the GREG estimator to produce timely official statistics about the monthly employed and unemployed labour force, since GREG estimators have relatively large design variances in the case of small sample sizes. To handle both problems in an effective way, a multivariate structural time series model is used to estimate official monthly statistics about the labour force.

[^98]Let $Y_{t}^{t-j}$ denote the GREG estimator for the unknown population parameter, say $\theta_{t}$, based on the panel observed at month $t$, which entered the survey for the first time at month $t-j$. Due to the applied rotation pattern, each month a vector $\mathbf{Y}_{t}=\left(Y_{t}^{t} Y_{t}^{t-3} Y_{t}^{t-6} Y_{t}^{t-9} Y_{t}^{t-12}\right)^{T}$ is observed. As a result, a five dimensional time series with GREG estimates for the unknown population parameter is obtained. According to Pfeffermann (1991), this vector can be modeled as

$$
\begin{equation*}
\mathbf{Y}_{t}=\mathbf{1}_{5} \theta_{t}+\lambda_{t}+\mathbf{e}_{t} \tag{1}
\end{equation*}
$$

with $\mathbf{1}_{5}$ a five dimensional vector with each element equal to one, $\lambda_{t}=\left(\lambda_{t}^{0} \lambda_{t}^{3} \lambda_{t}^{6} \lambda_{t}^{9} \lambda_{t}^{12}\right)^{T}$ a vector with time dependent components that account for the RGB, and $\mathbf{e}_{t}=\left(e_{t}^{t} e_{t}^{t-3} e_{t}^{t-6} e_{t}^{t-9} e_{t}^{t-12}\right)^{T}$ the corresponding survey errors for each panel estimate. More details about the application of this model to the Dutch LFS, including some possible extensions, is provided by Van den Brakel and Krieg (2009).

The population parameter $\theta_{t}$ in (1) can be decomposed in a trend component, a seasonal component, and an irregular component, i.e. $\theta_{t}=L_{t}+S_{t}+\varepsilon_{t}$, where $L_{t}$ denotes a stochastic trend component, using the so-called smooth trend model, $S_{t}$ a trigonometric stochastic seasonal component, and $\varepsilon_{t}$ the irregular component, see for example Durbin and Koopman (2001) for details about these time series components.

The systematic differences between the subsequent waves are modeled with $\boldsymbol{\lambda}_{t}$ in (1). Additional restrictions for the elements of $\lambda_{t}$ are required to identify the model. Here it is assumed that an unbiased estimate for $\theta_{t}$ is obtained with the first wave, which is observed by CAPI, i.e. $Y_{t}^{t}$. This implies that the first component of $\lambda_{t}$ equals zero, i.e. $\lambda_{t}^{0}=0$ for all $t$. The other elements of $\lambda_{t}$ measure the time dependent differences with respect to the first wave. To this end $\lambda_{t}^{j}$ are modeled as random walks for $j=3,6,9$, and 12, see Van den Brakel and Krieg (2009).

Finally a time series model for the survey errors $\mathbf{e}_{t}$ in (1) is developed. The rotating panel design implies sample overlap with panels observed in the past. Particularly the survey errors of the second, third, fourth and fifth wave are correlated with survey errors of preceding periods. The autocorrelations between the survey errors of the subsequent waves are estimated from the survey data, using the approach proposed by Pfeffermann and al. (1998). In this application it appears that the autocorrelation structure for the second, third, fourth and fifth wave can be modeled conveniently with an AR(1) model, Van den Brakel and Krieg (2009). Direct estimates for the variance of the series of the GREG estimators are used as prior information in the time series model using the survey error modelling approach, proposed by Binder and Dick (1990).

Time series model (1) makes optimal use of the available sample information from preceding periods to improve the GREG estimates for the monthly labour force. Furthermore, the model accounts for the rotating panel design of the LFS by modeling the RGB and the autocorrelation between the survey errors. The general way to proceed is to express the model in the so-called state space representation and apply the Kalman filter to obtain optimal estimates for the state variables, see e.g. Durbin and Koopman (2001). The state space representation of this model is given by Van den Brakel and Krieg (2009b). The software for the analysis and estimation of the time series model is developed in Ox in combination with the subroutines of SsfPack 3.0, see Doornik (1998) and Koopman, Shephard and Doornik (2008).

Since June 2010 this model is used by Statistics Netherlands to produce monthly official statistics about the employed and unemployed labour force for the Netherlands and a break down for age and gender in six domains. For these variables a filtered trend $\left(L_{t}\right)$ and signal ( $L_{t}+S_{t}$ ) is published.

## 3. Auxiliary information for unemployed labour force

Auxiliary information about the unemployed labour force is available from the register of the Office for Employment and Income. The Dutch abbreviation is CWI, which stands for Centre for Work and Income. This name recently
changed to UWV (Dutch abbreviation for Uitvoeringsinstituut voor Werknemersverzekeringen). In Figure 3.1, the filtered trend from the series of the number of people registered as unemployed in the CWI, using a univariate basic structural time series model, and the filtered trend of the unemployed labour force from the time series model of the LFS are displayed for the national level and for men in three age classes.

Figure 3.1

## Filtered trends unemployed labour force Dutch LFS and people registered as unemployed in the CWI for the Netherlands and men in three age classes






It is immediately clear that the trends of the CWI and the LFS have different levels but display more or less the same development. With the exception of the younger men, the level of the CWI is larger compared to the LFS. It can be concluded that the trends of CWI and the LFS are correlated and that it is worthwhile to investigate the possibilities to use the CWI as auxiliary information in the time series modeling approach, described in Section 2, to improve the precision of the estimated monthly unemployed labour force.

## 4. Structural time series model for a rotating panel and auxiliary series

In this section, the time series model describes in Section 2 is extended to incorporate the series of the number of people registered as unemployed in the CWI. This can be accomplished in different ways. One straightforward possibility is to extend the time series model (1) for the population parameter of the LFS with a regression component for the CWI series, i.e. $\theta_{t}=L_{t}+S_{t}+\beta X_{t}+\varepsilon_{t}$, where $X_{t}$ denotes the series available from the CWI and $\beta$ the regression coefficient. The major drawback of this approach is that with this model $L_{t}$ and $S_{t}$ contains a residual trend and seasonal effect, because a main part of the trend and the seasonal effects in $\theta_{t}$ are explained by the series of CWI. This hampers the estimation of a filtered trend for the unemployed labour force of the LFS. An alternative approach, that allows the direct estimation of a filtered trend for $\theta_{t}$, is to extend model (1) with a series for the CWI and model the correlation between the trend of the series of the LFS and the series of the CWI. This gives rise to the following model:

$$
\begin{equation*}
\binom{\mathbf{Y}_{t}}{X_{t}}=\binom{\mathbf{1}_{5} \theta_{t}^{\text {LFS }}}{\theta_{t}^{C W I}}+\binom{\boldsymbol{\lambda}_{t}}{0}+\binom{\mathbf{e}_{t}}{0} . \tag{2}
\end{equation*}
$$

The series of the LFS and the CWI both have their own population parameter that can be modeled with two separate time series models, i.e. $\theta_{t}^{z}=L_{t}^{z}+S_{t}^{z}+\varepsilon_{t}^{z}$, where $z=L F S$ or $z=C W I, L_{t}^{z}$ a smooth trend model, $S_{t}^{z}$ a trigonometric seasonal component and $\varepsilon_{t}^{z}$ white noise to model the unexplained variation. The smooth trend model
for the LFS and the CWI is defined as $L_{t}^{z}=L_{t-1}^{z}+R_{t-1}^{z}$, with $R_{t}^{z}=R_{t-1}^{z}+\eta_{t}^{z}$, with $\eta_{t}^{z} \cong N\left(0, \sigma_{\eta_{z}}^{2}\right)$. $L_{t}$ and $R_{t}$ are often referred to as the level and slope parameter. The model allows for correlation between the noise of the trend of the LFS and the CWI, by assuming that $\operatorname{cov}\left(\eta_{t}^{C W I}, \eta_{t}^{L F S}\right)=\rho \sigma_{\eta C W I} \sigma_{\eta L F S}$. The trigonometric seasonal components and the disturbances for CWI and LFS are assumed to be uncorrelated. Since CWI is based on a registration, this series does not have a RGB or a survey error component.

## 5. Results

Time series models (1) and (2) are used to estimate the monthly unemployed labour force and the results are compared to quantify the effect of the CWI information on the point estimates and the standard errors. Figure 5.1 compares the filtered trend and the standard errors of the unemployed labour force for the national level and the domains of men in three age classes based on a model with and without CWI. In Table 5.1, the correlation $\rho$ between the noise of the trends of the LFS and the CWI is specified. Also the mean of the relative absolute difference (MARD) of the filtered trend and the mean of the relative difference (MRD) of the standard errors between model (1) and (2) for the monthly unemployed labour force are specified for the national level and the series for age and gender in six domains. The mean absolute relative difference is defined as:

$$
\begin{equation*}
\operatorname{MARD}(z)=\frac{1}{T} \sum_{i=1}^{T} \frac{\left|z_{t}^{(1)}-z_{t}^{(2)}\right|}{z_{t}^{(1)}} \times 100, \operatorname{MRD}(x)=\frac{1}{T} \sum_{i=1}^{T} \frac{x_{t}^{(1)}-x_{t}^{(2)}}{x_{t}^{(1)}} \times 100 \tag{3}
\end{equation*}
$$

where $z$ is the filtered trend and $x$ the standard error of the filtered trend at month $t$ and the superscript refers the formula number of the applied time series model.

Striking is the difference between the effect of the CWI series at the national level on the one hand and the domains on the other hand. The filtered trend for the Netherlands is hardly influenced by the CWI series and the reduction of the standard errors is relatively small. The impact of CWI on the filtered trends as well as the standard errors for the domains is, on the other hand, substantial. In these cases the CWI also has a stabilizing effect on the filtered trend estimates. Indeed, the filtered trends for the domains based on the model with CWI are more smoothed compared to the filtered trend based on a model for the LFS data only.

The difference of the effect of CWI at the national level and a breakdown in six domains is also reflected by the estimated correlation between the trend of the CWI and the trend of the LFS. For the domains both trends are cointegrated or almost cointegrated, resulting in substantial standard error reductions, with the exception of the women $45-64$. The correlation between trends of CWI and the LFS at the national level is strong, but substantially smaller compared to the domains.

Figure 5.1
Filtered trend with standard errors of the unemployed labour force at the national level based on times series model (1) and (2)








Table 5.1
Estimated correlation between the trend of CWI and the LFS and the MARD for the filtered trend and the standard error of the unemployed labour force over the last 30 months at the national level and a classification into gender and age in six domains

| Domain | Correlation trend | MARD <br> filtered trend | MRD <br> standard error |
| ---: | ---: | ---: | ---: |
| Netherlands | 0.87 | 0.63 | 7.77 |
| Men $-15-24$ | 0.99 | 4.91 | 43.54 |
| Men $-25-44$ | 1.00 | 4.96 | 51.26 |
| Men $-45-64$ | 1.00 | 5.17 | 37.46 |
| Women $-15-24$ | 0.99 | 6.10 | 32.38 |
| Women $-25-44$ | 1.00 | 4.80 | 50.16 |
| Women $-45-64$ | 0.94 | 1.44 | 9.24 |

## 6. Conclusions

The Dutch LFS employs a multivariate structural time series model to publish official statistics about the labour force on a monthly basis. Extending this model with the series of registered unemployment, which is available from the CWI, clearly improves the estimates for the monthly unemployed labour force. The gains are substantial for the series of domains based on a classification into age and gender in six categories. The gains for the series at the national level are smaller. One reason is that the correlation between trend of the CWI and the LFS in the domains is substantially larger than at the national level. In most domains the trends are cointegrated or almost cointegrated. Another possible explanation is that the sampling error is substantially smaller in the LFS series at the national level compared to the LFS series for the six domains. As a result the series of CWI add more information to the LFS series for the domains.

Amendments of the law with respect to unemployment benefits and social benefits or sudden changes in the mode of operation of the CWI might give rise to discontinuities in the series. As a result, the evolution of the CWI series does not reflect the real development of the unemployed labour force. A consequence of extending the time series model with the auxiliary series of the CWI is that discontinuities in this auxiliary series affect the monthly estimates of the unemployed labour force and will influence the evolution of these series incorrectly. Such situations should not hamper the application of the CWI series as auxiliary information, since this kind of discontinuities can be modeled via an appropriate intervention variable, see e.g. Van den Brakel and Roels (2010) for details.

Currently the use of the CWI series in the estimation approach would slightly delay the timeliness of the monthly unemployment figures since the CWI figures about the latest month becomes available a few days before Statistics Netherlands publish the monthly labour force figures. Another possibility is to use the CWI information to produce revised unemployment figures.

## References

Binder, D.A. and Dick, J.P. (1990). A method for the analysis of seasonal ARIMA models, Survey Methodology, vol. 16, pp. 239-253.

Doornik, J.A. (1998). Object-oriented matrix programming using Ox 2.0., London: Timberlake Consultants Press.
Durbin, J. and Koopman, S.J. (2001). Time series analysis by state space methods. Oxford: Oxford University Press.
Koopman, S.J., Shephard, N, and Doornik, J.A. (2008). SsfPack 3.0: Statistical algorithms for models in state space form, London: Timberlake Consultants Press.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys, Journal of Business \& Economic Statistics, 9, pp. 163-175.

Pfeffermann, D., Feder, M. and Signorelli, D. (1998). Estimation of Autocorrelations of Survey Errors with Application to Trend Estimation in Small Areas, Journal of Business \& Economic Statistics, 16, pp. 339348.

Särndal, CE., Swensson, B. and Wretman, J. (1992). Model Assisted Survey Sampling, New York: Springer Verlag.
Van den Brakel, J.A. and Krieg, S. (2009). Estimation of the monthly unemployment rate through structural time series modelling in a rotating panel design, Survey Methodology, vol. 35, pp. 177-190.

Van den Brakel, J.A. and Krieg, S. (2009). Structural time series modelling of the monthly unemployment rate in a rotating panel design, unpublished report, discussion paper 09031, Statistics Netherlands, Heerlen.

Van den Brakel, J.A. and Roels, J. (2010). Intervention analysis with state-space model to estimate discontinuities due to a survey redesign, Annals of Applied Statistics, 4, pp. 1105-1138.

# Comparison of intercensal updating techniques for local level poverty statistics 

Marissa Isidro, Stephen Haslett, and Geoffrey Jones ${ }^{1}$


#### Abstract

The World Bank has implemented poverty mapping projects in collaboration with national statistical agencies in various developing countries employing a small area estimation method known as ELL (Elbers et al., 2003). This method and its variants require a survey, plus census and/or administrative data and assume that the data sets are gathered at the same time period. In most developing countries, a census is only conducted once in every decade. This poses a problem in the generation of updated small area estimates during intercensal years. It is therefore important to develop an updating method to provide policymakers and/or stakeholders in developing countries with updated estimates of local level poverty statistics. We propose an updating method called Extended Structure PREserving Estimation (ESPREE), an extension of the Structural PREserving Estimation (SPREE) method. In this paper, we compare the ESPREE and the alternative ELL-based updating method used by the World Bank in generating updated small area estimates in the Philippines. The ESPREE method has the advantage over the ELL-based methods that it generates unbiased estimates. An in-country validation exercise was conducted for assessing acceptability and consistency of the ESPREE and ELL-based estimates compared with the available municipal level poverty indicators and expert opinion of key informants.


Keywords: Small area estimation; SPREE; ELL; Intercensal updating; Poverty mapping

## 1. Introduction

A poverty map is an important tool in targeting anti-poverty resources which provides detailed information of the spatial distribution of various poverty indicators within a particular country. Recently, the World Bank (WB) has implemented various poverty mapping projects in collaboration with national statistical agencies in developing countries. These projects usually employ a small area estimation technique proposed by Elbers et al. (2003) widely known as ELL (Elbers, Lanjouw and Lanjouw) method and implemented in PovMap (Zhao, 2006).

The poverty measures or indicators that are generally included in poverty maps are economic-based, such as poverty incidence, gap and severity. These measures are generated based on the premise that a household is poor if its income or expenditure falls below a specified monetary standard, known as the poverty line. These measures can be put into a framework proposed by Foster et al. (1984), the so called FGT measures: $P_{\alpha}=1 / N \sum_{h=1}^{N}\left(\ell-Y_{h} / \ell\right)^{\alpha} I\left(Y_{h}<\ell\right)$, where $N$ is the size of the population in an area, $Y_{h}$ is the income or consumption of individual or household $h, \ell$ is the poverty line and $I\left(Y_{h}<\ell\right)$ is an indicator function (1 when income or expenditure is below the poverty line, and 0 otherwise). Poverty incidence, gap and severity correspond to $\alpha=0,1$ and 2 , respectively. Unlike other small area estimation procedures that fit a model directly to the variable of interest (i.e., poverty status), the ELL method fits a regression model to (log-transformed) income or expenditure first, after which the FGT equation is applied to generate the required estimates of poverty measures.

The ELL method combines the sample survey, census and/or administrative data to come up with small area estimates of poverty. In using these data sets for generating small area estimates, the ELL method assumes that the

[^99]data sets are gathered at the same time period. This assumption is particularly important since the variable of interest (income/consumption or poverty status) is not measured in the census; hence, the model for income is formulated using the survey data and is then applied to the census data. In most developing countries, a census is only conducted once in every ten years. This poses a problem in the generation of updated small area estimates during non-census years or intercensal years. It is therefore important to develop an updating method for small area estimates of poverty measures to provide policymakers and/or stakeholders with an updated estimate of poverty measures. Recently, intercensal updating methods have been proposed based on the ELL method and have been implemented by the WB in developing countries such as Uganda, Thailand, Vietnam and Philippines. As presented in Section 2.1, the implementation of the ELL-based updating methods is basically similar to the 'standard' ELL method except for the data requirements. These updating methods require either a panel survey data (village/household level) or 'time-invariant' variables for cross-sectional survey data. In this paper, we compare the ELL-based updating method which uses time-invariant variables with our proposed updating method - an extension of the structure preserving estimation (SPREE) method by Purcell and Kish (1980), hence called Extended SPREE.

The SPREE is a generalization of the synthetic estimation proposed by Gonzalez (1973). It makes better use of direct estimates and uses the method of iterative proportional fitting (IPF) of margins. IPF is used to adjust the census cell counts of a multi-way table called the association structure (census data) such that the adjusted counts satisfy specified margins, called the allocation structure (survey margins). The cell counts are obtained from the last census while the specified margins represent reliable direct survey estimates of current margins. In this way, SPREE provides intercensal estimates of small area totals of characteristics also measured in the census (Rao, 2003). However, the variable of interest under small area estimation for poverty measures in developing countries is not usually measured in the census. The ESPREE method as discussed in Section 2.2 extends the SPREE method so that updated estimates can be generated by using "pseudo-census data" instead of the census data. The ESPREE method also relaxes the assumption of fixed census data under SPREE, allowing more flexibility in the estimation. In Section 3, the updated poverty estimates generated from the most recent intercensal updating project of the WB conducted in collaboration with the National Statistical Coordination Board in the Philippines are compared with the ESPREE estimates. Substantial differences were observed between the two sets of estimates and the ELL-based estimates appear to be biased. To further assess the quality of the updated estimates, a validation study was conducted in one region in the Philippines, comparing the ESPREE and ELL-based updated estimates with the real poverty situation on the ground. Results showed that the ESPREE estimates were closer to the key informants' assessment than the ELL-based estimates in the municipalities visited.

## 2. Intercensal Updating Methods

### 2.1 The ELL-based updating methods

There are two types of ELL-based updating technique that have been implemented recently. One uses panel survey data and the other one uses "time-invariant" variables. The panel data approach has been used in Uganda (Hoogeveen et al., 2003) and in Thailand (Jitsuchon and Lanjouw, 2005), while the time-invariant variables approach has been implemented in Vietnam and Philippines (Lanjouw and van der Wiede, 2006). The panel data method requires collection of the most recent $\left(t_{1}\right)$ per capita income/expenditure for (a subset of) households included in the sample survey conducted at the same time period as the census $\left(t_{0}\right)$. When household-level characteristics are available as covariates, updated poverty estimates are derived by combining the per capita income/expenditure information at time $\left(t_{1}\right)$ which we denote here as $y_{b h, t_{1}}$ with the household level characteristics that are common to the survey and the census collected in $t_{0}$ denoted by $x_{b h, t_{0}}$ (Hoogeveen et al., 2003). The model is as follows, $y_{b h, t_{1}}=x_{b h, t_{0}} \beta+u_{b h, t_{1}}$ where $\beta$ is the regression parameter and $u_{b h, t_{1}}$ is the random error term and $b$ denotes the $b^{\text {th }}$ primary sampling unit (e.g., barangay) while $h$ denotes the $h^{\text {th }}$ household. In this model, new information on household characteristics are not needed to update poverty estimates, only the most recent information on income/expenditure is required as it uses the auxiliary variables from the census year. This method therefore regresses the most recent $\left(t_{1}\right)$ per capita income/expenditure data on the household characteristics at the earlier period $t_{0}$ rather than $t_{1}$ and assumes that there is no net migration among small areas under consideration. When household level characteristics are not available, Jitsuchon and Lanjouw (2005) suggested the use of village-level
characteristics at the census period, $\left(x_{a t_{0}}\right)$, as covariates, as the case in Thailand. The implementation is similar to the method described above; however, the explanatory variables used are village-level characteristics from the census period.

For countries with two cross-sectional surveys instead of panel surveys, Lanjouw and van der Wiede (2006) proposed an ELL-based updating method that uses "time-invariant" variables from the census period ( $t_{0}$ ). This method has been used in a collaborative project of the WB and National Statistical Coordination Board (NSCB) on intercensal updating of small area estimates of poverty measures in the Philippines. The implementation of this method is quite similar to the panel data technique (i.e. synthetic panel). The income/expenditure data taken from the most recent survey is combined with the time-invariant variables, common to the survey and census data, collected in the census year to generate poverty estimates. The model is as follows, $y_{b h, t_{1}}=\tilde{x}_{b h, t_{0}} \beta+u_{b h, t_{1}}$ where $\tilde{x}_{b h, t_{0}}$ refers to the characteristics that are time-invariant (i.e. $\tilde{x}_{b h, t_{0}}=\tilde{x}_{b h, t_{1}}$, in practice $\tilde{x}_{b h, t_{1}}$ is used for fitting the survey regression); $\quad \beta$ and $u_{b h, t_{1}}$ are as defined above. This method assumes that the covariates in the model are time invariant, i.e., household characteristics and municipality/village means do not change from the census period to the most recent survey; and that migration (at least among small areas) between the census period and the most recent survey is negligible. The implementation in the Philippines did not consider any test for time invariance and only selected those variables that were deemed "logically" time invariant by the user of the method.

### 2.2 The ESPREE Method

The SPREE method can be implemented by fitting a generalized linear model (GLM), $g(\mu)=\mathrm{X} \beta$, to both the census and survey type data, where $g($.$) is the \log$ function, $\mu$ is the expected value of the vector of the dependent variable, which for poverty estimation could be the number of households (which we denote here by Y for the survey period and $\widetilde{Z}$ for the census period) cross-classified by poverty status, province and other related variables. X , is the design or model matrix corresponding to the explanatory variables. Specifically, the model for the census data is as follows: $g\left(\mu^{\tilde{z}}\right)=\mathrm{X}_{1, t_{0}} \beta_{1, t_{0}}+\mathrm{X}_{2, t_{0}} \beta_{2, t_{0}}$ with the subscript $t_{0}$ indicating that the data comes from an earlier period, while the survey model is $g\left(\mu^{\mathrm{Y}}\right)=\mathrm{X}_{1, t 1} \beta_{1, t_{1}}+\mathrm{X}_{2, t_{1}} \beta_{2, t_{1}}$ with the subscript $t_{1}$ indicating that the data comes from a more recent period. We note that $X_{1, t_{0}}=X_{1, t_{1}}=X_{1}$ and $X_{2, t_{0}}=X_{2, t_{1}}=X_{2}$ are the partition of the design matrix corresponding to the partition of the parameter vector $\beta$ in the GLM model given above, which is $\beta_{1, t_{0}}$ and $\beta_{2, t_{0}}$ for the census model, $\beta_{1, t_{1}}$ and $\beta_{2, t_{1}}$ for the survey model. SPREE is equivalent to fitting the census model and then some of the lower order parameters in this model (first partition) are adjusted to align with the most recent information from the survey data, while the higher order parameters (second partition) remain the same, i.e., $\beta_{2, t_{0}}=\beta_{2, t 1}=\beta_{2}$ by assumption. In this way SPREE is used to generate updated small area estimates.

The SPREE model is extended by allowing for a misspecification error $\gamma_{t_{1}}$, i.e. $X_{2} \beta_{2}=X_{2} \beta_{2, t_{1}}=X_{2} \beta_{2, t_{0}}+\gamma_{t_{1}}$. Using the pseudo-census data, the estimation problem is considered in the context of a superpopulation, i.e. we assume that the pseudo-census data has a superpopulation that produces $\widetilde{Z}$. Under the assumption that the pseudocensus data forms a superpopulation, the coefficients $\beta_{1, t_{0}}$ and $\beta_{2, t_{0}}$ of the census model are now considered random with respect to the superpopulation and with expectation $\xi\left[\beta_{1, t_{0}}\right]$ and $\xi\left[\beta_{2, t_{0}}\right]$, respectively. The census model can now be written as: $g\left(\mu^{\tilde{z}}\right)=\mathrm{X}_{1, t_{0}} \xi\left[\beta_{1, t_{0}}\right]+\mathrm{X}_{2, t_{0}} \xi\left[\beta_{2, t_{0}}\right]+u_{t_{0}}$ where $u_{t_{0}}$ is a random variable with respect to the superpopulation assumed for the census data. For the survey data, we also assume that a superpopulation exists for the census from which the survey data is drawn. As in the census model, the parameters $\beta_{1, t 1}$ and $\beta_{1, t_{0}}$ of the survey model are now random with respect to the superpopulation and with expectation $\xi E\left[\beta_{1, t_{1}}\right]$ and $\xi E\left[\beta_{2, t 1}\right]$, respectively. $E$ is the expectation related to the sampling design of the survey data, while $\xi$ is the expectation related to the superpopulation assumed for the census. However, we could assume that $\xi E\left[\beta_{1, t_{1}}\right]=\xi\left[\beta_{1, t_{1}}\right]$ and
$\xi E\left[\beta_{2, t_{1}}\right]=\xi\left[\beta_{2, t_{1}}\right]$ provided that the sample is unbiased and the sampling design is not informative. Hence, the survey model will now be: $g\left(\mu^{\mathrm{Y}}\right)=\mathrm{X}_{1} \xi\left[\beta_{1, t_{1}}\right]+\mathrm{X}_{2} \xi\left[\beta_{2, t_{0}}\right]+u_{t_{1}}+\gamma_{t_{1}}$ where $u_{t_{1}}$ is a random error term for the survey model and $\gamma_{t_{1}}$ is a misspecification error. Hence, our proposed updating model which we call extended SPREE (ESPREE) is as follows:

$$
\begin{equation*}
g\left(\mu^{\mathrm{Y}}\right)=\mathrm{X}_{1} \xi\left[\beta_{1, t_{1}}\right]+\mathrm{X}_{2} \xi\left[\beta_{2, t_{0}}\right]+\varepsilon_{t_{1}} \tag{1}
\end{equation*}
$$

here $\varepsilon_{t_{1}}=u_{t_{1}}+\gamma_{t_{1}}$ and we assume that there is no net migration in between the census and survey periods. The fitting algorithm of the proposed updating method is adapted from the model fitting procedure for classical SPREE.

## 3. Application to Real Data and Validation Study

The two intercensal updating techniques are compared using the Philippine national survey (2000 and 2003) and census (2000) data. The survey data is a combination of the Family Income and Expenditure Survey (FIES) and Labor Force Survey (LFS) which are both conducted by the National Statistics Office (NSO). The FIES is conducted every 3 years and collects information on family income, expenditures and related information. The LFS on the other hand, collects data quarterly on employment and related demographic and socio-economic characteristics of the population over 15 years old. The Census of Population and Housing (CPH) used in this study is also conducted by the NSO and the CPH in the Philippines is done once every ten years. We note that for the ESPREE method, the characteristics of the superpopulation are inferred from a set of 100 bootstrap estimates of poverty status classification of the population, generated by the poverty mapping project conducted by the WB in collaboration with the NSCB in the Philippines employing the ELL method, see Haslett and Jones (2005) and Haslett et al. (2010) for details. We call this set of replicates the pseudo-census.

Space precludes the detailed presentation of the ELL-based updating technique as applied to the Philippine data. We note that the intercensal estimates generated by this method, which will be compared with the ESPREE estimates in this Section, are taken from the publication released by the NSCB (see NSCB, 2009 for more details). The set of variables used to illustrate the ESPREE method is a subset of the variables used in the joint WB/NSCB poverty mapping project mentioned above, see Haslett and Jones (2005) for a complete list and definition of variables. These variables are available in both the 2000 census and the 2000 and 2003 survey data sets. They are strongly correlated with household per capita income and hence the poverty status of the household members, see Isidro et al. (2010) and Isidro (2010) for more details of the models fitted, the auxiliary variables included in the ESPREE model and the variance estimation procedure used.

A summary of the municipal level (small area) estimates of poverty incidence for ESPREE and ELL-based updating methods, their standard errors and CVs are presented in Table 3-1. The mean of the poverty incidence computed from the ESPREE method is higher than the one generated from the ELL-based updating method. This could mean that the ELL-based updating method tends to generate lower values of poverty incidence estimates than the ESPREE method in most of the municipalities or small areas.

Table 3-1
Municipal level estimates via ESPREE and ELL

|  | ESPREE |  | ELL-based |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :---: |
|  | Incidence | SE | CV | Incidence | SE | CV |  |
| Mean | 0.4200 | 0.0418 | 0.1177 | 0.3755 | 0.0413 | 0.1371 |  |
| SD | 0.1713 | 0.0144 | 0.0620 | 0.1843 | 0.0194 | 0.0892 |  |
| Min | 0.0204 | 0.0038 | 0.0358 | 0.0114 | 0.0044 | 0.0140 |  |
| Max | 0.8937 | 0.1725 | 0.5787 | 0.9746 | 0.1812 | 0.8600 |  |

At the provincial level, the ESPREE and ELL-based estimates are compared with the survey-based or direct estimates from the FIES. It appears that the average estimated poverty incidence for the ESPREE and direct estimates are close to each other while the ELL-based estimate is lower. This is seen clearly in the Quantile-Quantile (QQ) plot of ESPREE (Figure 3-1) and ELL (Figure 3-2) estimates versus direct estimates for all provinces.

Figure 3-1
ESPREE and FIES provincial level estimates


Figure 3-2
ELL and FIES provincial level estimates


The updated estimates were also compared at the regional level; it can be observed from the QQ plots (Figures 3-3 and 3-4) that the ELL-based estimates tend to be lower than the survey-based estimates - almost all the points were above the line. The ESPREE estimates on the other hand give points both above and below the line.

Figure 3-3
ESPREE and FIES regional level estimates


Figure 4.4
ELL and FIES regional level estimates


The ESPREE updated estimates of poverty incidence at the provincial and regional levels are evidently closer to the direct estimates that are known to be unbiased. Using these more aggregated direct estimates from $t_{1}$ as the 'gold standard' we note that the ESPREE method is not biased and performs better than the ELL-based updating method.

Acceptability and consistency of the estimates were assessed by comparing them with the expert opinion of key informants and their perception of available poverty related indicators at the small area (municipal) level. These validation activities are adopted from the validation study conducted for the results (small area estimates of poverty measures) of the collaborative WB/NSCB poverty mapping project in the Philippines (NSCB, 2005). The validation exercise was carried out by having a one-on-one interview with each of the identified participants or key informants (representatives from local government units such as the Municipal and Provincial Planning and Development Office, Provincial Social Welfare and Development Office, Provincial Health Office, City Planning and Development Office, National Statistics Office and National Police) using a questionnaire, an adaptation of the validation form used in the WB/NSCB poverty mapping project (NSCB, 2005). This validation form contains the poverty-related indicators that were included in the set of auxiliary variables used in formulating the ESPREE model, other correlates of poverty and indicators of the Millennium Development Goals (MDGs). A total of thirty participants from the four provinces of Region 1 were interviewed. There were two sets of municipal rankings gathered from the participants - the indicator-based and the overall level of poverty assessment. The indicator-based rank is computed as an average of the participants' ranking of municipalities in a particular province based on the indicators included in the validation form. The overall rank is the participants ranking of the municipalities in their province based on their perception of the poverty situation of the said areas. The two sets of ranks were then compared with the ranking of the updated poverty estimates generated from the ESPREE and the ELL-based updating methods. The rank correlations (Rs) are presented in the table below:

Table 3-2
Rank correlation between participants' assessment and the small area estimates (ESPREE and ELL)

|  | Ilocos Norte |  | Ilocos Sur |  | La Union |  | Pangasinan |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
|  | Rs | pvalue | Rs | pvalue | Rs | pvalue | Rs | pvalue |
| Indicator-based vs ESPREE | 0.629 | 0.001 | 0.1614 | 0.3619 | 0.744 | 0.000 | 0.587 | 0.000 |
| Overall rank vs ESPREE | 0.610 | 0.002 | 0.3702 | 0.0312 | 0.837 | 0.000 | 0.062 | 0.677 |
| Indicator-based vs ELL | 0.476 | 0.022 | 0.2046 | 0.2457 | 0.599 | 0.005 | 0.491 | 0.000 |
| Overall rank vs ELL | 0.320 | 0.136 | 0.3106 | 0.0738 | 0.711 | 0.000 | 0.073 | 0.624 |

Table 3-2 shows that the participants' general assessments of poverty agree better with the estimates generated from the ESPREE method than with the estimates generated from the ELL-based updating method.

In conclusion, the comparison of the updated small area estimates of poverty measures in the Philippines generated from the ELL-based updating method using time invariant variables and the ESPREE method showed that the ELLbased updated estimates appear to be biased and less precise. The validation study conducted in one of the regions in the Philippines showed that the key informants' assessment of the poverty situation in their area is generally in agreement with or closer to the ESPREE-based than to the ELL-based updated estimates. Hence, when comparing the two methods, the ESPREE method appears to generate estimates of better quality (unbiased and more precise) that are better able to reflect the real poverty situation on the ground.

## References

Gonzalez, M.E. (1973). Use and Evaluation of Synthetic Estimators, Proceedings of the Social Statistics Section, American Statistical Association, pp 33-36.

Haslett, S., Isidro, M. and Jones, G. (2010). Comparison of Regression Techniques in the Context of Small Area Estimation of Poverty, Survey Methodology, in press.

Haslett, S. and Jones, G. (2005). Local Estimation of Poverty in the Philippines, NSCB /WB Report.

Hoogeveen, J., Emwanu, T. and Okwi, P. (2003). Updating Small Area Welfare Indicators in the Absence of a New Census, preprint.

Isidro, M., Haslett, S. and Jones, G. (2010). Extended Structure Preserving Estimation Method for UpdatingSmall Area Estimates of Poverty, Proceedings of the Section on Government Statistics, Joint Statistical Meetings, pp.1795-1807.

Isidro, M. (2010). Intercensal Updating of Small Area Estimates, unpublished PhD Thesis, Massey University, New Zealand.

Jitsuchon, S. and Lanjouw, P. (2005). Updating Poverty Maps: Emerging Experience with Available Methods and Data from Thailand, preprint.

Lanjouw, P. and van der Wiede, R. (2006). Determining Changes in Welfare Distributions at the Micro-level: Updating Poverty Maps, PowerPoint slides, NSCB/WB Workshop for Intercensal Updating Project.

NSCB (2005). Estimation of Local Poverty in the Philippines, NSCB Philippines.
NSCB (2009). 2003 City and Municipal Level Poverty Estimates, NSCB, Philippines.
Rao, J. N.K. (2003). Small Area Estimation, Wiley Series in Methodology, New York.
Purcell, N.J. and Kish, L. (1980). Postcensal Estimates for Local Areas, International Statistical Review, 48, pp. 318.

Zhao, Q. (2006). User Manual for Povmap, http://siteresources.worldbank.org/INTPGI/Resources.

## RECONCILIATION

# Use of mental health services in Quebec based on data from a population survey and administrative data: Study of underreporting and overreporting of services 

Aline Drapeau, Richard Boyer and Fatoumata Binta Diallo ${ }^{1}$


#### Abstract

This study compared the use of mental health services based on data from Cycle 1.2 of the Canadian Community Health Survey (CCHS-1.2) and data from the Régie de l'assurance maladie du Québec (RAMQ). According to the CCHS-1.2 data, $5.8 \%$ of the 4,235 respondents aged 18 and over received mental health services the previous year, compared to $15.0 \%$ according to the RAMQ data. This underreporting varied by respondents' profile and the type and number of services received according to the RAMQ. The results of this study highlight the importance of taking into consideration both survey and administrative data when estimating the population's needs for mental health services.


Key Words: Mental health services; matching of survey and administrative data; underreporting.

## Introduction

Managers of the health care system have two main sources of data for estimating the population's needs for health care services: population surveys and administrative records. However, these two data sources can yield different profiles of the types of service used and users. Thus, the use of mental health services is lower when documented on the basis of a population sample than on official records (Killeen, 2004; Marshall, 2003; Rhodes, 2002; Rhodes, 2004). This apparent underreporting of mental health services in population surveys is generally attributed to memory bias and social desirability (Beebe, 2006; Killeen, 2004; Rhodes, 2002; Rhodes, 2004).

The fact is that data collection in a population survey makes heavy demands on respondents' memory, since the reference period for using health care services is relatively long, usually consisting of the twelve months preceding the survey. Also, Rhodes and Fung (2004) argue that persons with a mental health problem have symptoms or take medication that can affect their ability to remember. However, recall bias cannot explain why the discrepancy between survey data and official records is greater for the use of services for mental health problems than for physical health problems (Killeen, 2004; Marshall, 2003). The hypothesis of a social desirability bias was put forward by researchers who have shown that the underreporting of mental health services was not random. It was found to be higher among seniors (Rhodes, 2002), less educated persons (Taube, 1986) and the unemployed (Killeen, 2004). The profile of persons at risk of underreporting appears to correspond to the profile of persons who have more prejudices against mental illness (e.g. seniors (Dorvil, 1995)) or those who are the most often stigmatized because of their mental health problems (e.g. unemployed persons (Roeloffs, 2003)).

This study sought to identify the factors associated with the underreporting of mental health services in a population survey compared to data from an administrative register. Two categories of factors were examined: factors related to individuals' sociodemographic profile and factors related to the characteristics of the mental health services provided to these individuals according to the administrative register.

[^100]
## Methodology

This study is based on the matching of data from Cycle 1.2 (Mental Health and Well-being) of the Canadian Community Health Survey (CCHS-1.2) and data from the Régie de l'assurance maladie du Québec (RAMQ). This matching was authorized by the Commission d'accès à l'information du Québec, and the study was approved by the research ethics committee of the Centre de recherche Fernand-Seguin at Hôpital Louis-H. Lafontaine (Montréal). The CCHS-1.2 is a population survey that was conducted by Statistics Canada in 2001-2002, targeting the population aged 15 and over living in private residences (Gravel, 2005). In Canada, the entire population has national health insurance that covers the cost of medical appointments made within the public health system. The RAMQ is the Quebec government agency responsible for paying for the services provided to the public by general practitioners and medical specialists in the public health system.

The sample is composed of Quebec respondents to the CCHS-1.2 aged 18 and over who agreed to have the information that they revealed in the CCHS-1.2 matched with data from government administrative files for research purposes ( $\mathrm{n}=4,554$ ). The data matching was carried out by the Institut de la statistique du Québec in cooperation with the RAMQ (Baulne, 2009), and $93 \%(n=4,235)$ of adult respondents were successfully matched with RAMQ data. Respondents who refused the matching of CCHS-1.2 and RAMQ data were statistically more likely to be unemployed (unemployed, $6 \%$; employed, $4 \%$ ), to not have a legal or common-law spouse (without spouse, $6 \%$; with spouse, $4 \%$ ) or to have children living at home (without children, $4 \%$; with children, $6 \%$ ), but they did not differ by age, sex or the presence of a psychiatric problem. Respondents who could not be matched with the RAMQ data were proportionally more likely to be 45 years of age or over and to have a negative perception of their mental health (Baulne, 2009). The analyses are based on weighted data that correct the effect of this "non-response"; the weights were calculated by Statistics Canada.

Respondents who, in the CCHS-1.2, reported having consulted a doctor (i.e. general practitioner, psychiatrist or other medical specialist) for problems related to their emotions or a drug or alcohol problem in the twelve months preceding their interview, or for suicidal thoughts or a suicide attempt during the same period, were coded as having received a mental health service according to the survey data. The RAMQ's definition of a mental health service was based on three variables in that file: a service provided by a psychiatrist (medical specialist); a medical act in psychiatry (codes 360 to 394 of the RAMQ); and a psychiatric diagnosis (codes 290 to 319 and 2900 to 3199 of version 9 of the International Classification of Diseases - ICD-9). Respondents who, during the year preceding their CCHS-1.2 interview, had received a medical service meeting one of these criteria were coded as having received a mental health service according to the RAMQ. The interval between the interview in the CCHS-1.2 and each mental health service received according to the RAMQ was calculated individually for each respondent. The combination of these two measures (i.e. services received according to the CCHS-1.2; services received according to the RAMQ) produced four categories of respondents: most respondents $(82.8 \%)$ did not receive mental health services according to both the CCHS-1.2 and the RAMQ in the twelve months preceding their interview; $2.1 \%$ are overreporters, meaning that they reported receiving at least one mental health service in the CCHS -1.2 but received none according to the RAMQ; $3.7 \%$ received a mental health service according to both the CCHS -1.2 and the RAMQ; $11.4 \%$ are underreporters, meaning that they received no mental health services according to the CCHS-1.2 but did receive this type of service according to the RAMQ. Overall, the interrater agreement (i.e. CCHS-1.2 vs. RAMQ) is equal to kappa $=0.29$. The analyses largely concern the respondents $(\mathrm{n}=637)$ who received mental health services according to the RAMQ and who did or did not report receiving this type of service in the CCHS-1.2. The characteristics of overreporters are only briefly described.

The sociodemographic profile of the 637 respondents who received mental health services according to the RAMQ is described by gender (women: 62\%; men: $38 \%$ ); age group (i.e. 18-39: $28 \%$; 40-64: $53 \%$; 65 and over: $19 \%$ ); education level (i.e. no diploma: $28 \%$; high school diploma: $20 \%$; post-secondary diploma: $31 \%$; university diploma: $20 \%$ ); employment status (i.e. employed: $63 \%$ vs. unemployed: $37 \%$ ); conjugal status (i.e. with legal or common-law spouse: $56 \%$ vs. without: $44 \%$ ); parental status (i.e. with children in household: $29 \%$ vs. without: $71 \%$ ); psychiatric profile as measured by the Composite International Diagnostic Interview (CIDI) for mood disorders (i.e. major depression and mania: $16 \%$ ) and anxiety disorders (i.e. agoraphobia, panic disorder, social phobia: $8 \%$ ) and by the CIDI Short Form for alcohol or drug dependencies (3\%). The characteristics of the mental health services provided by physicians according to the RAMQ are described by three variables: maximum number of criteria identifying a mental health service (i.e. speciality: psychiatrist, medical act in psychiatry, psychiatric diagnosis) that are met
during a given medical appointment, the number of mental health services and speciality of the physician who provided each mental health service.

Logistic regressions were carried out to estimate the odds ratio of underreporting the use of mental health services in the CCHS-1.2 (compared to the RAMQ), based on the profile of respondents and the characteristics of the services provided by physicians, according to the RAMQ.

## Results

Of the persons who, according to the RAMQ, received mental health services in the twelve months preceding their interview in the CCHS-1.2, the majority ( $481 / 637=75.5 \%$ ) did not report receiving this type of service in that survey. The underreporting of mental health services in the CCHS-1.2 (compared to RAMQ data) increases with age and declines with education level; it is statistically lower for persons who report symptoms meeting the criteria for diagnoses of a mood disorder, an anxiety disorder, a dependency or one of these disorders (Table 3-1). The interactions between gender and the other independent variables were tested. These analyses showed a significant interaction between gender and parental status. The logistic regression model including this interaction suggests that the probability of underreporting mental health services in the CCHS -1.2 is higher for mothers ( $87 \%$ ) than for fathers ( $63 \%$ ) but does not vary according to gender among persons with no children.

Also, underreporting decreases with the maximum number of criteria met during a given medical consultation and with the number of mental health services according to RAMQ data. It is also 4.3 times higher among respondents who received at least one mental health service from a general practitioner, according to the RAMQ, than among those who received this type of service from a medical specialist (Table 3-2).

Lastly, $37 \%$ of respondents who reported having seen a doctor for a mental health problem in the CCHS-1.2 did not receive this type of service according to the RAMQ (data not shown in a table). This apparent overreporting is higher for seniors than for younger adults (ages 18-39: $36 \%$; 40-64: $32 \% ; 65$ and over: $67 \% ; \chi^{2}=9.4, \mathrm{p}$ value $=0.009$ ). The majority of cases of overreporting are explained by telescoping: according to the RAMQ, $74 \%$ and $10 \%$ of overreporters respectively received mental health services between 12 and 18 or between 18 and 24 months before their interview in the CCHS-1.2 rather than in the twelve months preceding the survey as they reported.

## Discussion

The results of this study show that the overall agreement between the CCHS-1.2 and the RAMQ on mental health services is fairly weak (kappa $=0.29$ ) and that the underreporting of this type of service in the CCHS-1.2 is sizable, namely $75 \%$ of users of mental health services identified from RAMQ data. This percentage is higher than the $50 \%$ recorded in the study of Rhodes et al. (2002), who compared the use of mental health services in Cycle 1 (19941995) of the National Population Health Survey with the health insurance register of the Ontario Ministry of Health. In interpreting the findings of this study, two caveats are in order. First, the definition of a mental health service based on RAMQ data is relatively broad, since it rests on three criteria (service provided by a psychiatrist, medical act in psychiatry and psychiatric diagnosis) and since meeting any one of these criteria was sufficient to classify respondents in the category of persons who had received a mental health service. This strategy was adopted to minimize the number of apparent overreporters; it seemed incongruous that researchers' definition of a mental health service was stricter than respondents' definition. However, this strategy tended to increase the number of underreporters. Half of respondents (i.e. 321/637, or $50 \%$ ) met a single criterion, and for 159 of these respondents, the criterion that was met was a psychiatric diagnosis. If the "psychiatric diagnosis" criterion had been excluded from the definition of a mental health service received according to the RAMQ, this would have reduced the percentage of underreporting from $75 \%$ to $55 \%$, a percentage similar to the one observed by Rhodes et al. (2002). Second, RAMQ data are an imperfect gold standard for medical services received by the population, since the recording of these data is constrained by administrative rules that were put in place to manage the payment of medical services. RAMQ data describe health services provided by physicians, and accordingly they must be used in estimating the medical service needs of the population. They partially reflect the public's point of view on the services received, since the underreporting of mental health services in the CCHS-1.2 decreases as a function of the number of services provided according to the RAMQ.

Table 3-1
Odds ratios of underreporting of a mental health service in the CCHS-1.2 compared to RAMQ data, according to respondents' characteristics

|  | n (weighted) | Underreporting ${ }^{\mathrm{a}}\left(\%^{\mathrm{b}}\right)$ | RAMQ-CCHS reporting ${ }^{\text {c }}\left({ }^{\text {b }}\right.$ ) | Odds ratio ${ }^{\text {d }}$ |
| :---: | :---: | :---: | :---: | :---: |
| Total sample | 637 | 75.5 | 24.5 | N/A |
| Gender |  |  |  |  |
| Women | 398 | 73.1 | 26.9 | 1.40 |
| Men | 240 | 79.2 | 20.8 | (0.96 to 2.06) |
| Age group |  |  |  |  |
| 18 to 39 | 181 | 64.1 | 35.9 | 2.31 |
| 40 to 64 | 339 | 74.9 | 25.1 | (1.73 to 3.09 ) |
| 65 and over | 118 | 94.1 | 5.9 | (1.73 to 3.09) |
| Education |  |  |  |  |
| Pre-secondary | 174 | 82.8 | 17.2 |  |
| Secondary | 125 | 74.4 | 25.6 | 0.78 |
| Post-secondary | 196 | 72.4 | 27.6 | (0.66 to 0.93 ) |
| University | 125 | 68.8 | 31.2 |  |
| Employment status |  |  |  |  |
| Unemployed | 234 | 77.4 | 22.6 | 0.85 |
| Employed | 402 | 74.4 | 25.6 | (0.58 to 1.24) |
| Marital status |  |  |  |  |
| No legal or common- | 281 | 66.2 | 33.8 | 2.48 |
| law spouse |  |  |  | (1.72 to 3.60) |
| Legal or common-law spouse | 356 | 82.9 | 17.1 |  |
| Parental status |  |  |  |  |
| No children | 454 | 76.9 | 23.1 | 0.77 |
| Children | 183 | 72.1 | 27.9 | (0.52 to 1.14) |
| Psychiatric disorder ${ }^{\mathrm{e}}$ ( ${ }^{\text {e }}$ |  |  |  |  |
| Mood disorder | 100 | 27.0 | 73.0 | $\begin{array}{r} 0.08 \\ (0.04 \text { to } 0.11) \end{array}$ |
| Anxiety disorder | 54 | 35.2 | 64.8 | $\begin{array}{r} 0.14 \\ (0.08 \text { to } 0.26) \end{array}$ |
| Dependency (drug/alcohol) | 20 | 30.0 | 70.0 | $\begin{array}{r} 0.14 \\ (0.05 \text { to } 0.36) \end{array}$ |
| One of the above | 140 | 32.1 | 67.9 | $\begin{array}{r} 0.07 \\ (0.04 \text { to } 0.10) \end{array}$ |

${ }^{a}$ No mental health service received in the year preceding the interview according to CCHS-1.2 data, but at least one mental health service provided according to RAMQ data.
${ }^{\mathrm{b}}$ Percentages are based on totals (number of respondents) per row.
${ }^{c}$ At least one mental health service received in the year preceding the interview according to CCHS-1.2 data and at least one mental health service provided according to RAMQ data.
${ }^{\mathrm{d}}$ The figures in parentheses represent confidence intervals at the 0.95 level for odds ratios.
${ }^{\mathrm{e}}$ Presence of a psychiatric disorder according to symptoms reported by respondents in the CCHS-1.2. The reference category for estimating each odds ratio is "other disorder or no disorder."

Table 3-2
Odds ratios of underreporting of a mental health service in the CCHS-1.2 compared to RAMQ data according to characteristics of medical consultations ( $n=637$ )

|  | n (weighted) | Underreportin $\mathrm{g}^{\mathrm{a}}\left(\%^{\mathrm{b}}\right)$ | RAMQ-CCHS reporting ${ }^{\text {c }}\left(\%^{b}\right)$ | Odds ratio ${ }^{\text {d }}$ |
| :---: | :---: | :---: | :---: | :---: |
| Maximum number of criteria met during same medical consultation |  |  |  |  |
| One | 321 | 86.3 | 13.7 |  |
| Two | 259 | 71.0 | 29.0 | (0.24 to 0.42) |
| Three | 57 | 35.1 | 64.9 | (0.24 to 0.42) |
| Number of mental health services received according to the RAMQ |  |  |  |  |
| One | 247 | 87.0 | 13.0 | 0.44 |
| Two | 138 | 84.8 | 15.2 | (0.35 to 0.55) |
| Three or more | 253 | 58.9 | 41.1 | (0.35 to 0.55) |
| Physician's specialty according to RAMQ |  |  |  |  |
| General medicine ${ }^{\text {e }}$ | 603 | 77.3 | 22.7 | $\begin{array}{r} 4.31 \\ \text { (2.14 to } 8.68 \text { ) } \end{array}$ |
| Psychiatry ${ }^{\text {e }}$ | 63 | 36.5 | 63.5 | $\begin{array}{r} 0.15 \\ (0.09 \text { to } 0.26) \end{array}$ |
| Other specialty ${ }^{\text {e }}$ | 16 | 75.0 | 25.5 | $\begin{array}{r} 0.88 \\ (0.29 \text { to } 2.68) \end{array}$ |

${ }^{a}$ No mental health service received in the year preceding the interview according to CCHS-1.2 data, but at least one mental health service provided according to RAMQ data.
${ }^{\mathrm{b}}$ Percentages are based on totals (number of respondents) per row.
${ }^{c}$ At least one mental health service received in the year preceding the interview according to CCHS-1.2 data and at least one mental health service provided according to RAMQ data.
${ }^{\mathrm{d}}$ The figures in parentheses represent confidence intervals at the 0.95 level for odds ratios.
${ }^{\mathrm{e}}$ The reference categories for estimating odds ratios are: General medicine (i.e. general practitioner) vs. Other specialty; Psychiatry vs. General medicine or other speciality; Other specialty vs. General medicine or psychiatry. When a respondent received more than one service from the same category of physician, he/she is counted only once in that category.

The results of this study show that the underreporting of mental health services in the CCHS-1.2 (compared to RAMQ data) is not random, since it is higher for certain categories of respondents, such as seniors, less educated persons, persons with a legal or common-law spouse and women with children. Underreporting of mental health services in population surveys is generally attributed to recall bias and social desirability (Beebe, 2006; Killeen, 2004; Rhodes, 2002; Rhodes, 2004). Theoretically, two categories of people could have faulty recall: seniors, because memory tends to decline with age and persons with a psychiatric disorder, because they have symptoms or are taking medication that may affect their memory. These two categories of people would also appear to be more sensitive to social desirability issues related to prejudices against mental illness (Dorvil, 1995; Roeloffs, 2003). The hypothesis that recall bias affects the underreporting of mental health services is supported by the fact that both the underreporting and the overreporting of mental health services in the CCHS-1.2 are higher among seniors. Accordingly, it would be appropriate in population surveys to apply strategies that make it easier to remember past events (for example, use a calendar and reduce the reference period). The finding that persons who meet the diagnostic criteria of a mood disorder, an anxiety disorder or alcohol or drug dependency in the CCHS-1.2 are less likely to underreport mental health services contradicts the hypotheses of a recall bias or a social desirability bias affecting a particular population. On the other hand, the association observed between underreporting and marital status for women and men and parental status for women could be linked to social desirability. However, in the absence of a measure of prejudices against mental illness in the CCHS-1.2, that hypothesis cannot be rejected or confirmed in this study.

Finally, underreporting was 4.3 times higher among persons who had received a mental health service from a general practitioner according to the RAMQ data than among those who had received this type of service from a psychiatrist or other medical specialist. General practitioners constitute the vast majority of the physicians, and for this reason, they are the category of physicians most often consulted for both mental and physical health problems. The high percentage of underreporters among persons who received a mental health service from a general practitioner according to the RAMQ data could be an indication that these persons consulted a general practitioner for a physical health problem that was interpreted as a mental health problem by the physician; this hypothesis is supported by the fact that the percentage of underreporters is also high for non-psychiatric medical specialists.

In conclusion, the results of this study highlight the importance of taking into consideration both survey data and administrative data when estimating the population's needs for mental health services, since these two data sources provide sometimes divergent but complementary information on the use of mental health services in the general population. This study is ongoing and could add further to the advancement of knowledge on the underreporting of mental health services in population surveys.

## Acknowledgements

This study was conducted under a grant from the Canadian Institutes of Health Research (\#167330), and the main author (A. Drapeau) had a research grant from the Fonds de recherche en santé du Québec (\#12230). We wish to thank the Institut de la statistique du Québec and the Régie de l'assurance maladie du Québec for collaborating in this project. We also wish to thank Anne-Marie Parent, who conducted the preliminary analyses.

## Bibliography

Baulne, J. (2009). Influence de l'ancrage social sur le comportement et l'attitude des femmes et des hommes par rapport à la maladie mentale : L'exemple de la sous-déclaration du recours aux services psychiatriques dans une enquête populationnelle. Projet d'Appariement EPSEBE - Méthodologie et résultats. Direction de la méthodologie et de la qualité. Institut de la statistique du Québec, Québec.

Beebe, T.J., McRae, J.A., and Barnes, S.A. (2006). A comparison of self-reported use of behavioral health services with Medicaid agency records in Minnesota, Psychiatric Services, 57, pp. 1652-1654.

Dorvil, H., Bisson, J., and Sévigny, O. (1995). Attitudes et croyances des Montréalais à l'égard des personnes malades mentales et des personnes déficientes intellectuelles. Groupe de recherche sur les aspects sociaux de la santé et de la prévention : Université de Montréal.

Gravel, R., and Béland, Y. (2005). The Canadian Community Health Survey: Mental Health and Well-Being, Canadian Journal of Psychiatry, 50, pp. 573-579.

Killeen, T.K., Brady, K.T., Gold, P.B., Tyson, C., and Simpson, K.N. (2004). Comparison of self-report versus agency records of service utilization in a community sample of individuals with alcohol use disorders, Drug \& Alcohol Dependence, 73, pp. 141-147.

Marshall, S.F., Deapen, D., Allen, M., Anton-Culver, H., Bernstein, L, Horn-Ross, P.L., Peel, D., Pinder, R., Reinolds, P., Ross, R.K., West, D., and Ziogas, A.L. (2003). Validating California Teachers Study selfreports of recent hospitalization: Comparison with California Hospital Discharge data, American Journal of Epidemiology, 158, pp. 1012-1020.

Rhodes, A.E., Lin, E., and Mustard, C.A. (2002). Self-reported use of mental health services versus administrative records: should we care?, International Journal of Methods in Psychiatric Research, 11, pp. 125-133.

Rhodes, A.E., and Fung, K. (2004). Self-reported use of mental health services versus administrative records: care to recall?, International Journal of Methods in Psychiatric Research, 13, pp. 165-175.

Roeloffs, C., Sherbourne, C., Unűtzer, J., Fink, A., Tang, L., and Wells, K.B. (2003). Stigma and depression among primary care patients, General Hospital Psychiatry, 25, pp. 311-315.

Taube, C.A, Schlenger, W.E., Rupp, A., and Whitmore, R.W. (1986). Validity of Medicaid household respondent reporting of ambulatory visits for mental disorders, Journal of Economic and Social Measurement, 14, pp. 243-256.

# Understanding victimization statistics from two national crime data sources: The Uniform Crime Reporting Survey and the General Social Survey on Victimization 

Amanda Halladay, Maire Sinha, Kathy AuCoin, Sarah Franklin ${ }^{1}$


#### Abstract

Should self-reported victimization rates and police-reported crime rates follow the same trends? This is a question that many Canadians may ask as they compare trends from the General Social Survey (GSS) on Victimization with policereported crime trends from the Uniform Crime Reporting (UCR) Survey. These surveys are two complementary instruments that measure levels of crime in different ways and which have their own strengths and weaknesses. The GSS on Victimization is a voluntary sample survey that is conducted every five years, which collects self-reported victimization data from approximately 25,000 Canadians. The UCR is a mandatory administrative census that collects police-reported data on criminal activity from police departments across Canada. The purpose of this paper is to reduce the level of confusion arising from the use of crime data from these two very different sources and to highlight the major differences between the two surveys. Topics covered include the scope of each survey, the target populations, sources of error and possible bias, and survey definitions.


Key Words: Non-sampling errors; survey design; crime rate; victimization rate.

## 1. Introduction

### 1.1 Background

Statistics Canada produces two main indicators on the prevalence of crime using two distinct data sources: the Uniform Crime Reporting (UCR) Survey is used for police-reported crime and the General Social Survey (GSS) on Victimization is used for victim-reported crime. Each of these surveys produces valuable information about crime in Canada. Together they provide a more comprehensive view than either could on its own. While they are complementary data sources, their differences in terms of survey design, content, population coverage, counts of criminal incidents and rate calculations mean that analysts should approach each with an awareness of their analytical benefits and limitations. Because of their many differences, the two surveys' measurements of the level of crime in Canada should not be expected to mirror each other.

### 1.2 Measures of Criminal Activity

In co-operation with the policing community, the Canadian Centre for Justice Statistics has been systematically collecting police data through the UCR Survey since 1962. The UCR Survey is a census of all Canadian Criminal Code offences that are known to, and substantiated by, police services. It also includes federal and provincial statutes. The current version of the UCR (called UCR2) collects microdata on a number of factors, including: victim and accused characteristics; the time and location of an incident; the presence of a weapon; and any resulting injuries to the victim. Police-reported crime data are used to measure the volume of crime and trends over time in Canada. The data from the UCR Survey are used to calculate two main statistics: the traditional police-reported crime rate and the police-reported Crime Severity Index (CSI).

[^101]The traditional crime rate measures the volume of crime reported to the police and has been used since the inception of the UCR. This measure represents the sum of all police-reported criminal incidents falling under the Criminal Code (excluding traffic), divided by the population count. (The crime rate does not include Federal Statute offences, such as drug offences). In this calculation, all offences contribute equally and the crime rate is expressed as a rate per 100,000 Canadians. By contrast, the newly developed CSI was created in 2009 to address the issue of the overall crime rate being driven by high volume, less serious offences such as minor theft, mischief and minor assault. In the calculation of the CSI, each offence is assigned a weight that is derived from court sentences: the more serious the offence, the larger the weight. Consequently, more serious offences contribute more to the CSI (Wallace et al, 2009).

The objective of Statistics Canada's General Social Survey (GSS) program is to gather data on social trends on a variety of topics. The annual GSS focuses on one of five regular themes: social networks, time use, families, caregiving and victimization (Statistics Canada, 2009). The victimization cycle collects information on the extent and nature of self-reported criminal victimization in Canada and has been collected approximately every five years since 1988. Questions are asked to individuals 15 years of age and over on their personal accounts of criminal victimization for eight types of crime: sexual assault, robbery, physical assault, break and enter, motor vehicle/parts theft, theft of household property, vandalism and theft of personal property. The survey produces two key statistics: the twelve-month victimization rates (violent and household) and the five-year spousal victimization rate. Respondents are also asked about the impact and consequence of their victimization and all respondents - regardless of whether they were victimized or not - are asked about their fear and perceptions of crime, and the justice system.

## 2. Differences between the UCR Survey and the GSS on Victimization

### 2.1 Survey design and population

A major advantage of the UCR Survey is its virtual full coverage of police services in Canada. The UCR Survey is a mandatory administrative census, which means that all crimes that come to the attention of the police must be reported. By contrast, the GSS is a voluntary sample survey targeting approximately 25,000 respondents, aged 15 years and over, from the 10 provinces. A sample of dwellings is selected using Random Digit Dialling (RDD) of landline telephone numbers. Once the dwelling has been contacted one member of the household is randomly selected. The target population for GSS is all non-institutionalized people aged 15 and over (i.e., individuals living in households). Households without a telephone or with only cellular service are excluded from the survey. According to the Residential Telephone Services Survey (RTSS), these two groups represented 4\% of the GSS target population in 2004 and $9 \%$ in 2008.The implicit assumption made by GSS analysts is that the victimization of people without a phone or with only a cell phone does not differ substantially from those surveyed by GSS.

In 2009 , the GSS response rate was $61.6 \%$. After accounting for non-response and out-of-scope units, there were approximately 19,500 respondents in the 10 provinces. By contrast, the UCR respondents include more than 1,000 separate police detachments from 202 police services across the country.

Both surveys produce national and provincial estimates. In the case of the GSS, estimates are calculated using survey weights. In 2009, each GSS respondent represented roughly 1,400 Canadians aged 15 years and over. Measures of sampling error (e.g., coefficients of variation) indicate the reliability of the GSS estimates.

### 2.2 Survey content

The UCR Survey collects data on more than 200 types of offence. This information is captured electronically, typically by the police officer who is present at the incident. Prior to the introduction of the incident-based microdata survey (UCR2), crime data were only available in aggregate form from the police. The UCR2 survey and the GSS on Victimization began in the same year: 1988. The UCR2 Survey captures data regarding the location of the incident, weapons involved, types of property stolen, level of injury sustained, relationship between the victim and accused, and the sex and age of both the victim and the accused. Since 2008, the UCR2 Survey has essentially replaced the previous aggregate survey (UCR1) and covers $99 \%$ of the Canadian population.

The GSS captures information surrounding criminal incidents directly from victims. The GSS asks respondents if they have been a victim of eight crime categories, including three violent crimes (sexual assault, robbery and physical assault), theft of personal property and four household crimes (theft of household property, break and enter, motor vehicle/parts theft and vandalism). One advantage of the GSS is that it captures information on crimes that are not reported to the police. In 2009, for the offences covered by the GSS, roughly two-thirds of incidents were not reported to the police (Perreault and Brennan, 2010). Another advantage of the GSS is that in addition to the eight offences covered, the survey adds content each cycle based on input from justice stakeholders. The GSS is therefore able to gather information on new or emerging areas of crime where data have not previously been gathered, or where the offences are not thought to be well represented in crime statistics. For example, the 2009 GSS collected information on hate crime, spousal violence, stalking, cyber-bullying and non-Criminal Code acts, namely discrimination. With the exception of spousal violence that occurred in the past 12 months, these latter crime categories do not feed into the calculation of victimization rates.

Unlike the UCR Survey, the GSS does not cover victimless crimes, crimes against businesses, corporate crimes, homicides or crimes against individuals under the age of 15 . The GSS does, however, capture victim demographic characteristics that are not available on the UCR Survey, including: Aboriginal identity, visible minority status, sexual orientation, marital status, income, employment, education, religion, and health background. Victims are also asked about the physical and emotional impact of their victimization, the associated financial costs, the extent of their reporting to police, and their use of informal and formal victim services. In addition, the GSS asks all respondents - both victims and non-victims - about their perceptions of crime in their neighbourhood, their views of the justice system (police, courts, corrections) and precautionary measures that they take in their day-to-day lives. As a result, analysis of these factors and how they differ among victims and non-victims is available with GSS.

### 2.3 Initiation date, survey frequency and reference period

While the GSS on Victimization has been conducted roughly every five years since 1988, the UCR Survey has been conducted annually since 1962 . The GSS relies on the accurate memory of respondents since it asks them to recall any criminal incident that occurred in the past 12 months and, in instances of spousal violence, the past five years (covering both current and ex-spouses). Collection of the GSS starts in January and continues over a 12-month period. As such, the overall reference period spans two to six years since a respondent contacted in January reports for a different 12 -month and five-year period than a respondent who is contacted in December. By contrast, the UCR survey spans one calendar year. Data from the UCR survey are generally received on a monthly basis, generally within 45 days of month's end. The final aggregate counts are published and available on an annual basis. Each year, the published UCR data from the previous year are updated to reflect changes that have occurred. The GSS data, on the other hand, are collected and released once every five years.

The UCR Survey also captures information on crimes that occurred in the past. For example, a victim of sexual assault may wait several months or years before reporting the event, and a homicide victim may be discovered years after the homicide.

### 2.4 Volume of offences

The UCR Survey captures data on nearly 2.5 million criminal police-reported incidents per year for over 200 crime categories, while the GSS captures information on eight specific offences. Included within the set of over 200 standard crime categories collected by the UCR Survey are violent crimes, property crimes, other Criminal Code violations, traffic violations, drug offences, offences against the Youth Criminal Justice Act (YCJA), and other federal and provincial statutes. The eight GSS crime categories correspond to approximately $50 \%$ of all police-reported crime.

The primary benefit of the GSS is that it collects information on offences that do not come to the attention of the police. There are a number of reasons why a criminal incident may not be reported to the police. The procedures enforced by local police departments and the ease of reporting can affect how the public reports crime. For example, new legislation or policies may change the enforcement practices of the police. In addition, research has shown that victims of spousal violence, sexual assault and teen victims are reluctant to report their victimization to the police and as a result relying on police data alone to monitor these incidents would lead to an undercount (Gannon and Mihorean, 2005).

### 2.5 Sources of error

### 2.5.1 Sampling error

Since the UCR survey is a mandatory census it is not subject to sampling error. The GSS, however, is subject to sampling error. For example, for estimates of the number of victimizations in the past 12 months, the coefficients of variation (CV) at the national level in 2009 were less than $5 \%$. The CVs are generally higher when broken down by province, or for more detailed categories. The GSS considers an estimate to be unreliable if its CV is greater than $33 \%$.

### 2.5.2 Non-sampling error

The UCR Survey and the GSS are both subject to non-sampling errors, which may arise throughout all stages of the survey process. Some of the more common non-sampling errors that may arise are detailed below.

Coding errors: The data from both surveys are heavily processed though a set of automated internal edit routines. However, it is still possible for some processing errors to go undetected.
Respondent error: Each survey has a certain amount of respondent error. For example, a GSS respondent may have trouble recalling past events, placing the event in the correct time period, or may even intentionally misreport an experience. A GSS respondent may also think a crime occurred when it did not (for example, a wallet reported as stolen may have actually been lost). With the UCR, a police officer may make a mistake capturing the data in the Record Management System (RMS).
Varying collection practices: Data from the UCR Survey are submitted electronically by police respondents using an RMS. While there are a limited number of approved extraction programs, reporting practices tend to vary by system as well as by police service. For example, some police services use the latest version with updated violation codes, while others use older versions which have not yet incorporated the newest codes.
Comparability over time: With the UCR Survey criminal violations are updated whenever changes are made to the Criminal Code. This may affect comparability over time. Similarly, changes to the GSS content and concepts over time may affect the comparability of statistics over time. Comparisons between the 1999, 2004 and 2009 cycles are readily available. However, comparisons prior to 1999 should be done with caution due to changes in the sample as well as the definitions of some of the eight offence types.
Coverage error: The target population for the GSS includes all persons aged 15 years and over who do not live in an institution. However, only households with a landline telephone have the opportunity of being sampled by the survey. The proportion of households without a landline phone varies depending on the characteristics of the household. For example, according to the 2008 RTSS, $32.7 \%$ of men between the ages of 25 and 29 live in a household with a telephone, but not a landline (Franklin and Wysocki, 2010). Undercoverage of the target population may cause bias if people living in households with a landline have different victimization profiles than those in households without a landline.
Interviewer error and questionnaire design: Survey interviewers can be a source of error since they have the potential to misinterpret the response and/or bias a respondent's answers through their prompts, wording or tone of voice. However, great care is taken when training the interviewers to minimize this type of error. Another possible source of response error is the design of the questionnaire. For example, the questionnaire wording, length and layout may affect a respondent's answers. Prior to each cycle of the GSS, the questionnaire is reviewed by subject matter experts and rigorously tested by Statistics Canada's Questionnaire Design Resource Centre, as well as by the survey team. There is also a pilot survey to ensure that questions are understood correctly and that the flow of the questions does not impact respondents.
Non-response error: The GSS is a voluntary survey; sampled individuals are not obliged to respond. In 2009, the response rate for the GSS was $61.6 \%$. Non-response bias may arise if GSS respondents do not represent those who did not respond to the survey. The UCR, by contrast, is a mandatory survey with virtually a $100 \%$ response rate.

### 2.6 Rate calculation and units of count

There is an important distinction between the denominators for the crime indicators for each survey. The GSS victimization rates are expressed in terms of the number of incidents per 1,000 households or per 1,000 noninstitutionalized persons aged 15 years and over within the 10 provinces, whereas the traditional UCR crime rate is
expressed per 100,000 people. In calculating the victimization rates, the GSS also has a total cap on the number of incidents recorded for each victim, as well as a series cap on each crime category. The latter limits the number of times a respondent may report being a victim of the same crime. For example, in 2009, there was a cap of five incidents per crime type, and the cap on the total number of personal incidents of crime was not allowed to exceed 20. Refer to table 2.6-1 for specific details of the rate calculations.

Table 2.6-1
Comparing rate and index calculations for the UCR Survey with rate calculations for the GSS on Victimization

| UCR | Traditional Crime Rate (violent and non-violent): The sum of reported criminal <br> incidents per 100,000 people in the population. <br> Crime Severity Index (violent and non-violent): The weighted sum of reported <br> criminal incidents divided by the population, and standardized to a base year of <br> 100 in 2006. |
| :--- | :--- |
| GSS on <br> Victimization | Overall Victimization Rate: The percentage of non-institutionalized Canadians <br> aged 15 and over who were a victim of a criminal incident (covered by the GSS) in <br> the previous 12 months. Victims of multiple incidents are only counted once. <br> Household Victimization Rate: The sum of GSS criminal incidents per 1,000 <br> households in the 10 provinces. <br> Violent Victimization Rate: The sum of violent GSS criminal incidents per 1,000 <br> non-institutionalized people aged 15 and over in the 10 provinces population. <br> Spousal Victimization Rate: The percentage of people who have been a victim of <br> at least one spousal violence incident in the previous year or five years (current or <br> past union). |

In addition, the manner in which crimes are counted also differs between the two surveys. The UCR Survey counts violent crimes based on the number of victims, while property and other Criminal Code offences are counted based on the number of incidents. With the GSS, incidents are counted once regardless of the number of victims involved. The GSS count of crimes is related to the individual victim for personal crimes (violent crimes and theft of personal property) and the household victim for household crimes (vandalism, break and enter, theft of household property).

## 3. Analytical Guidelines: Choosing the UCR Survey and/or the GSS on Victimization for crime analysis

From a subject matter perspective, analysts must determine which data source is most appropriate when approaching a particular research topic. The UCR Survey must be used if analysts want to include the following in their analyses: crimes not covered by the GSS (e.g., homicides); commercial and victimless crimes; crimes against the homeless, people living in institutions or households without landlines; or crimes against children under 15 years of age. The GSS on Victimization must be used if analysts wish to examine unreported crimes, expanded information on the context of the crime, details on the extent and nature of spousal violence, or the general population's perceptions of crime, the justice system and their use of available services.

### 3.1 Advantages of the GSS on Victimization

The main advantage of the GSS on Victimization is that it captures both crimes that are reported and unreported to the police. Information is obtained directly from victims along with more detailed information about the sociodemographic characteristics of the victim. This allows analysts to look at patterns of reporting, and non-reporting, to the police based on crime types and victim characteristics for the eight crime categories covered by the GSS. The GSS is thus able to capture more contextual information about an incident than the UCR. The GSS on Victimization is updated at each cycle and now contains a number of special modules such as: level of social engagement, sense of belonging and trust, social disorganization, fear of victimization, methods of crime prevention and the criminal justice system. The GSS is currently the best tool to provide an assessment of the physical, emotional and financial impacts of crime in Canada, as well as perceptions of crime and the criminal justice system by all Canadians. By
capturing information on both victims and non-victims of crime, the GSS allows for a greater understanding of how victims and non-victims differ. (Information on non-victims is not captured by the UCR.) Elucidating differences between the victim and non-victim populations can assist researchers and policy-makers in the development of such initiatives as crime prevention programs, particularly those based on social development and situational crime prevention models.

### 3.2 Advantages of the UCR

One major advantage of the UCR Survey is that it captures all Criminal Code and Federal Statute offences grouped into more than 200 crime categories. The UCR Survey captures all crimes that come to the attention of the police, including commercial crimes, crimes against children, drug offences and a number of other crimes that are not captured with the GSS. The UCR is also mandatory and as such, disaggregation at low levels of geography is available, unlike with the GSS where reliable estimates may not be possible for small domains. With the incidentbased UCR2 survey, geocoding of incident locations is available to plot crime rates along with socio-demographic measures from other data sources, such as the census of population. This allows for neighbourhood-level analysis of crime. Lastly, the aggregate UCR has been operational on an annual basis since 1962 and maintains a historical record of crime that allows for trend analysis over time, unlike the GSS, which is only conducted every five years. When performing trend analysis using UCR microdata, it should be noted that UCR2 coverage has been steadily increasing over the years.

## 4. Conclusion

Neither the UCR Survey nor the GSS on Victimization is able to capture a complete picture of all criminal activity in Canada. The GSS complements the UCR and helps explain the large amount of crime in Canada that goes unreported to the police as well as providing a better understanding of victims' experiences with the justice system and their socio-demographic characteristics. Rates and trends may differ for the two surveys for a number of reasons: different survey designs, different coverage of the population and different coverage of crimes. Each survey has advantages over the other in terms of analytical potential. As such, crime analysts must consider the benefits and limitations of each with respect to their research topic before determining which data set is most appropriate. Together, the two surveys provide a more complete picture of crime in Canada than either can provide on its own. While in some respects they are similar surveys, they have fundamental differences that affect the interpretation of the resulting crime statistics. As such, attempting to reconcile trends between the two survey instruments is not appropriate.

## References

Franklin, S. and Wysocki, M. (2010). Evaluating RDD Coverage using the 2008 Residential Telephone Services Survey, paper presented at Statistics Canada's RDD working group, Ottawa, Canada.

Gannon, M. and Mihorean, K. (2005). Criminal Victimization in Canada, 2004 , Statistics Canada catalogue no. 85-002-XPE, Ottawa, Canada: Statistics Canada.

Mihorean, K., Besserer, S., Hendrick, D., Brzozowski, J.A., Trainor, C. and Ogg, S. (2001). A Profile of Criminal Victimization: Results of the 1999 General Social Survey. Statistics Canada Catalogue no. 85-553-XIE, Ottawa, Canada: Statistics Canada.

Perreault, S. and Brennan, S. (2010). Criminal Victimization in Canada, 2009, Statistics Canada Catalogue no. 85-002-X, Ottawa, Canada: Statistics Canada.

Statistics Canada (2009). The General Social Survey: an Overview, Statistics Canada Catalogue no. 89F0115XWE, Ottawa, Canada.

Wallace, M., Turner, J., Matarazzo, A. and Babyak, C. (2009). Measuring Crime in Canada: Introducing the Crime Severity Index and Improvement to the Uniform Crime Reporting Survey, Statistics Canada Catalogue no. 85-004-X , Ottawa, Canada: Statistics Canada.

# School education data in India: Emerging issues in increasing multiplication of sources of data 

Anugula N. Reddy ${ }^{1}$


#### Abstract

In India, data on school education are being collected and collated by multiple agencies. The Ministry of Education responsible for the administration of education at both the federal and provincial levels publishes annual data on various dimensions. An autonomous government organization, the National Council of Educational Research (NCERT), was also entrusted the responsibility of providing somewhat detailed data on school education recurrently for every five years. Many household surveys like the Census, National Sample Surveys (NSSs) and National Family and Health Surveys (NFHSs) also provide periodic data on education, particularly on attendance status. In the recent past, owing to several factors, many agencies such as civil society/NGOs and autonomous research/advisory organizations of government including the National University of Educational Planning and Administration (NUEPA) are also beginning to collect and collate data on school education through household and school surveys. The emergence of multiple agencies supplying data on school education raises several questions. Are the gaps in data filled and time lag reduced by multiplying the sources of data? Are the data compatible? How do we reconcile the wide divergence between administrative data on the one hand and household data on the other, particularly with respect to enrolment and attendance? How can we supplement the administrative and household data on school education in monitoring the progress towards goals like Education for All, Education MDGs and the Right to Education in the Indian context? The present paper gives a panoramic view of educational statistics in India and examines the context in which the divergence of sources of data is taking place.


[^102]
# Validating diabetes disease status in claims and survey data 

Joseph Sakshaug, David Weir and Lauren Hersch Nicholas ${ }^{1}$


#### Abstract

The purpose of this study was to validate measures of diabetes status in Medicare claims and survey data. Data from the 2006 Health and Retirement Study were linked to the Beneficiary Annual Summary File for 2,028 Medicare enrollees. Medicare claims and self-reported measures of diabetes status were validated against measured Hemoglobin A1c levels. The Medicare claims measure yielded a higher rate of diabetes ( 27.0 percent) than the self-reported measure ( 20.4 percent). Subjects classified as diabetic by the claims measure, but not the self-report measure, had significantly lower levels of Hemoglobin A1c compared to those classified as diabetic by both measures. We conclude that the claims-based measure tends to slightly overestimate diabetics relative to self-reports.


Key Words: Diabetes classification; Medicare claims; Self-report; Hemoglobin A1c.

## 1. Introduction

### 1.1 Background

Diabetes has become one of the most common and expensive medical conditions amongst older adults. Nearly $25 \%$ of all adults age 60 and older have diabetes in the United States, and this group accounts for over half of the more than $\$ 100$ billion in health care expenditures attributed to this disease (CDC, 2008; ADA, 2008). Obtaining the most precise estimates of diabetes prevalence among various population subgroups will be crucial to inform future health reforms aimed at better managing the financial burden forced upon payers and patients. The most ideal measure of diabetes status is a clinical diagnosis obtained from measured Hemoglobin A1c or, previously, glucose testing (ADA, 2010). Given the paucity of clinical diagnoses in health services research, researchers must rely on selfreports or claims algorithms to assess the prevalence and characteristics of diabetics.

Both measures have strengths and limitations for identifying the diabetic population. Self-reported diabetes status is a common measure collected in population-based health surveys. However, self-reports are susceptible to recall and social desirability effects (Tourangeau, Rips, Rasinski, 2000), which may lead to underestimation of chronic health conditions. In contrast, claims-based measures are applied to patient billing records, which are collected systematically and avoid problems related to patient misreporting. However, claims-based measures may not match actual diagnoses found in medical records (Losina et al., 2003). Furthermore, coding definitions used in the claims may be vague and indicate utilization without knowing the outcome of diagnostic tests, which can lead to inconsistent designations of disease (Iezzoni, 1997). All of these limitations can cause claims-based measures to be biased (Keating et al., 2003; Gruber and Rudnitsky, 2002).

In light of this evidence, an increasing number of studies are linking Medicare claims to population-based survey data (Lillard and Farmer, 1997; NCHS, 2010). Linking these two data sources affords researchers the opportunity to answer important policy-oriented research questions that cannot be answered using any one data source (Sloan et al., 2004; Ellerbeck, 2004). A consequence of data linkage is that it can produce discordant measures of health status

[^103](Nicholas, 2010; McAlpine et al., 2007; Davern et al., 2008). Although researchers commonly treat claims-based measures the "gold standard" for which self-reports can be validated, there is evidence that such measures may misclassify diabetics (Keating et al., 2003; Gruber and Rudnitsky, 2002). In light of the conflicting evidence, there is no strong consensus on which diabetes measure is most valid for identifying and studying the diabetic population (Fowles, Fowler and Craft, 1998; Miller, Safford and Pogach, 2004).

This paper uses measured Hemoglobin A1c levels, collected in the Health and Retirement Study, to validate measures of diabetes obtained from self-reports and claims.

## 2. Methods

### 2.1 Measures of Diabetes Status

Measures of diabetes status for subjects in the analysis sample were gathered from: (1) respondent self-reports from the 2006 Health and Retirement Study (HRS), (2) Medicare Chronic Condition Warehouse algorithm, and (3) measured Hemoglobin A1c (HbA1c) levels obtained from blood spots collected in the 2006 HRS.

Our primary data set was the HRS (Juster and Suzman 1995). HRS is a nationally-representative longitudinal study of Americans over the age of 50. The study began in 1992 and includes a biennial core interview. The study collects information on health behaviors, disease and disability, medical care usage, and other topics. In each wave, respondents are asked whether they have ever been diagnosed with diabetes ("Has a doctor ever told you that you have diabetes or high blood sugar?"). We use data from the 2006 wave. Our analysis sample was limited to persons aged 65 years and older who were Medicare-eligible at the time of interview and consented to have their Medicare claims linked for research purposes. Proxy interviews were excluded.

Eligible HRS participants were asked to release their Medicare records from the Centers for Medicare and Medicaid Services. For respondents who consented (86 percent), Medicare Beneficiary Annual Summary Files from 2006 were obtained and linked. We restrict the analysis sample to beneficiaries continuously enrolled in fee-for-service Medicare during the study period. Potential Veterans Affairs users were excluded to ensure that all health care utilization was reflected in the claims.

The Medicare Chronic Condition Warehouse (CCW) algorithm used to flag diabetics is defined as the presence of at least one inpatient, skilled nursing or home health visit or two outpatient visits with a diabetes-related ICD-9 code (Buccaneer, 2009). This algorithm was found to be adequately sensitive ( $\geq 70 \%$ ), highly specific ( $\geq 97.5 \%$ ), and reliable (kappa $\geq 0.80$ ) by Hebert et al. (1999). In the cited study, the authors obtained self-reports of diabetes status from participants in the Medicare Current Beneficiary Survey who were matched to their Medicare claim record. They used self-reported diabetes status as the "gold standard" to validate claims-based measures of diabetes status.

In 2006, a random-half of the HRS sample was asked for consent to participate in a blood draw. (In 2008, the other random half-sample were asked to participate in the blood draw.) For respondents who consented ( 83 percent), field interviewers collected blood spots and sent them to a laboratory where Hemoglobin A1c levels and other measurements were assayed. We used the clinical threshold of having a HA1c level of 6.5 or higher as a criterion for the presence of diabetes (ADA, 2010).

We classified respondents as diabetic if they had at least two indications of the disease. Because the blood spot collection took place at a single point in time and the self-report measure covers ever told of having diabetes, we cannot use measured HA1c levels solely to validate measures of diabetes status. Furthermore, our primary interest lies in validating measures of diagnosed diabetes, rather than all diabetes (diagnosed and undiagnosed), which the HA1c measure alone cannot disentangle. Therefore, we define our verified measure of diagnosed diabetes as either having an indicator of diabetes in both claims and self-reports (concordance), or having at least one diabetes indicator from either data source and having a HA1c level of 6.5 or greater.

### 2.2 Statistical Analyses

The Medicare claims, HRS, and biomarker data were matched in the HRS secure data enclave at the Institute for Social Research. The analysis sample is composed of 2,028 respondents who had measures of diabetes status from all three sources. We assess the prevalence of having a diabetes diagnosis based on the Medicare CCW and HRS definitions. Mean HA1c levels are reported for discordant cases, those flagged as having a diabetes diagnosis by only one measure, are calculated to assess the validity of the measures. All analyses were performed using SAS version 9.1.3. Adjustment weights accounting for socio-demographic and health status differences between biomarker participants and non-participants were also applied to the estimates.

## 3. Results

### 3.1 Percent Distribution of Diabetics

Table 1 shows the distribution of diagnosed diabetics based on each measure. Out of 2,028 respondents, 440 (20.4 percent) were classified as diabetic based on the HRS self-report measure, 568 subjects ( 27.0 percent) were classified as diabetic based on the Medicare claims measure; the difference between the two measures was statistically significant $(p<0.05)$. The validated measure of diabetes yielded a prevalence rate of 20.4 percent, which was statistically significant relative to the Medicare claims measure, but not to the self-report measure. No statistically significant were found between self-report and claims measures across socio-demographic characteristics, with the exception of age: respondents flagged as diabetic based on the claims measure were significantly older than those flagged as diabetic based on the self-report measure. Self-reported health rating and physical activity levels were not statistically discernible across the two diabetes measures.

### 3.2 Mean Hemoglobin A1c Levels

A total of 196 subjects were flagged as diabetic based on only one of the two measures. $29 \%$ of subjects flagged as diabetic using the claims measure were not flagged as diabetic using the self-report measure. In contrast, only $8 \%$ of self-reported diabetics were not flagged as diabetic by the claims measure. Table 2 shows the average A1c level for each discordant group. Discordant subjects classified as diabetic based solely on the claims measure had a mean A1c of 5.86, whereas those classified as diabetic based solely on the self-report measure only had a mean A1c of 6.32 . The percentage of subjects meeting the clinical definition of diabetes (A1c $\geq 6.5$ ) was 30.5 for the discordant selfreport group and only 12.4 for the discordant claims group. All differences were statistically significant ( $p<0.05$ ). It is unlikely that patients who do not self-report diabetes could unknowingly maintain the tight glycemic control suggested; thus, we suspect that the claims-based algorithm over-classifies diabetics.

Table 1. Percent Distribution of Diabetics among Medicare Beneficiaries Aged 65 and Older, by Self-Report and Claims Measures and Characteristics

|  | HRS Self-Report | Medicare Claims | Validated <br> Diabetes |
| :--- | :--- | :--- | :--- |
| \% Diabetic (overall) $20.4(1.0 ; 2028)^{*}$ $27.0(1.0 ; 2028)^{*} \dagger$ <br> Age  $20.4(1.0 ; 2028)$ <br> $65-74$ $53.9(2.5 ; 261)^{*}$ $45.7(2.1 ; 295)^{*}$ | $51.8(2.6 ; 254)$ |  |  |
| $75-84$ | $37.0(2.4 ; 150)$ | $42.2(1.9 ; 220)$ | $38.6(2.4 ; 156)$ |
| $\quad 85+$ | $9.1(1.6 ; 29)$ | $12.1(1.6 ; 53)$ | $9.6(1.7 ; 30)$ |
| Gender <br> $\quad$ Male | $45.3(2.0 ; 200)$ | $43.4(2.1 ; 246)$ | $45.8(1.9 ; 201)$ |
| Ethnicity <br> $\quad$ Hispanic | $7.0(1.9 ; 37)$ | $6.2(1.6 ; 43)$ | $7.3(1.9 ; 39)$ |
| Race <br> $\quad$ Non-White | $10.0(1.5 ; 62)$ | $11.3(1.3 ; 89)$ | $10.9(1.7 ; 67)$ |
| Total number of flagged diabetics | 440 | 568 | 440 |

Notes: Standard errors and sample sizes are in parentheses.
Column percentages may not always add to 100 due to rounding error.

* Difference between self-report estimate and claims estimate is statistically significant $p<0.05$.
$\dagger$ Difference between estimate and validated diabetes is statistically significant $p<0.05$.
Sources. 2006 Health and Retirement Study; 2006 Medicare Beneficiary Annual Summary File.
Table 2. Mean Hemoglobin A1c Levels among Medicare Beneficiaries Aged 65 and Older, by Discordant and Concordant Diabetes Classification Group

|  | Self-Report Only <br> (Discordant) | Claims Only <br> (Discordant) | Concordant <br> (Self-Reports and <br> Claims) |
| :--- | :--- | :--- | :--- |
| A1c (mean) | $6.32(0.03)^{* \dagger}$ | $5.86(0.04)^{*} \dagger$ | $6.60(0.05)$ |
| Ha1c $>=6.5(\%)$ | $30.5(0.5)^{* \dagger}$ | $12.4(2.2)^{*} \dagger$ | $45.7(2.4)$ |
| Total number of flagged diabetics | 34 | 162 | 406 |

Notes: Standard errors are in parentheses.

* Difference between self-reported estimate and claims estimate is statistically significant $p<0.05$.
$\dagger$ Difference between discordant estimate and concordant estimate is statistically significant $p<0.05$.
Sources. 2006 Health and Retirement Study; 2006 Medicare Beneficiary Annual Summary File.


## 4. Conclusions

We used measured Hemoglobin A1c levels collected from nationally representative survey respondents to validate self-report and claims-data assessed measures of diabetes. We found that claims data algorithms led to false positives relative to self-report and physical measures. As researchers increasingly work with linked self-report and claims data, it is important for researchers to know whether self-report or claims measures should be treated as the gold standard when the two measures yield conflicting results. Although researchers commonly assume that measures based on administrative data are the 'gold standard', the results in this paper suggest otherwise. A disproportionate number of subjects were classified as diabetic according to the claims-based measure, but not the self-report measure. These discordant cases had significantly lower levels of mean HA1c levels compared to the concordant cases. This suggests that the Medicare Chronic Condition Warehouse algorithm is flagging a substantial portion of subjects as diabetic when they are unlikely to have the disease. The overestimate of diabetes status based on the claims measure could be attributed to older, heavier users of care who have more opportunities for screening and diagnostic testing, resulting in an increased likelihood of triggering the Chronic Condition Warehouse algorithm.

Future work should consider the impact of using multiple measures of diabetes on substantive health outcomes. Determining whether different measures produce dramatically different conclusions on health care utilization could have serious implications for policy reform. This work is currently underway.

## References

American Diabetes Association (ADA). (2008). Economic costs of diabetes in the U.S. in 2007. Diabetes Care 31 (3): 596-615.

American Diabetes Association (ADA). (2010). Standards of Medical Care in Diabetes-2010. Diabetes Care 33 (1): 511-561.

Brandt, J., Spencer, M. and Folstein, M. (1988). The Telephone Interview for Cognitive Status. Neuropsychiatry, Neuropsychology, and Behavioral Neurology 1: 111-7.

Buccaneer. (2009). Chronic Conditions Warehouse User Manual Version 1.5.
Centers for Disease Control and Prevention (CDC). (2008). National Diabetes Fact Sheet: General Information and National Estimates on Diabetes in the United States, 2007. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Davern, M., Call, K.T., Ziegenfuss, J., Davidson, G., Beebe, T.J. and Blewett L. (2008). Validating Health Insurance Coverage Survey Estimates: A Comparison of Self-Reported Coverage and Administrative Data Records. Public Opinion Quarterly 72 (2): 241-259.

Ellerbeck, E.F., Engelman, K.K., Williams, N.J., Nazir, N. and Markello, S.J.. (2004). Variations in Diabetes Care and the Influence of Office Systems. American Journal of Medical Quality 19 (1): 12-8.

Fowles, J.B., Fowler, E.J. and Craft, C. (1988). Validation of Claims Diagnoses and Self-Reported Conditions Compared with Medical Records for Selected Chronic Diseases. Journal of Ambulatory Care Management 21 (1): 24-34.

Gruber, T. and Rudnitsky, B. (2002). Can We Be Certain of PRO Accuracy and Accountability? Inaccurate'Report Cards' Generated by Medicare Claims Data as a Marker for Laboratory Testing in Diabetic Patients. American Journal of Medical Quality 17 (5): 171-4.

Hebert, P.L., Geiss, L.S., Tierney, E.F., Engelgau, M.M., Yawn, B.P. and McBean, A.M. (1999). Identifying Persons with Diabetes Using Medicare Claims Data. American Journal of Medical Quality 14 (6): 270-7.

Iezzoni, L.I. (1997). Assessing Quality Using Administrative Data. Annals of Internal Medicine 127: 666-74.
Juster, F. and Suzman, R. (1995). An Overview of the Health and Retirement Study. Journal of Human Resources 30 (5): S7-56.

Keating, N.L., Landrum, M.B., Landon, B.E., Ayanian, J.Z., Borbas, D. and Guadagnoli, E. (2003). Measuring the Quality of Diabetes Care Using Administrative Data: Is There Bias? Health Services Research 38 (6, Part I): 1529-46.

Lillard, L.A. and Farmer, M.M. (1997). Linking Medicare and National Survey Data. Annals of Internal Medicine 127 (8, Part 2): 691-5.

Losina, E., Barrett, J., Baron, J.A. and Katz, J.N. (2003). Accuracy of Medicare Claims Data for Rheumatologic Diagnoses in Total Hip Replacement Recipients. Journal of Clinical Epidemiology 56 (6): 515-9.

McAlpine, D.D., Beebe, T.J., Davern, M. and Call, K.T. (2007). Agreement Between Self-Reported and Administrative Race and Ethnicity Data Among Medicaid Enrollees in Minnesota. Health Services Research 42 (6, Part 2): 2373-88.

Miller, D.R., Safford, M.M. and Pogach, L.M. (2004). Who has Diabetes? Best Estimates of Diabetes Prevalence in the Department of Veterans Affairs Based on Computerized Patient Data. Diabetes Care 27 (Suppl. 2): B10-21.

National Center for Health Statistics (NCHS). (2010). Linkages between Survey Data from the National Center for Health Statistics and Program Data from the Centers for Medicare and Medicaid Services. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention.

Nicholas, L.H. (In Press). Can Food Stamps Help to Reduce Medicare Spending on Diabetes? Economics and Human Biology.

Sloan, F.A., Brown, D.S., Carlisle, E.M., Picone, G.A. and Lee, P.P. (2004). Monitoring Visual Status: Why Patients Do or Do Not Comply with Practice Guidelines. Health Services Research 39 (5): 1429-48.

Tourangeau, R., Rips, L.J. and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge University Press.

Welsh, K., Breitner, J. and Magruder-Habib, K. (1993). Detection of Dementia in the Elderly using Telephone Screening of Cognitive Status. Neuropsychiatry, Neuropsychology, and Behavioral Neurology 6: 103-10.

## SESSION 11A

## COMPLEMENTARY SURVEYS

# Use of the Census as a supplementary frame for the Canadian Survey of Household Spending 

Denis Malo ${ }^{1}$


#### Abstract

The Survey of Household Spending (SHS) is designed primarily to provide provincial estimates of household expenditures. The stratified two-stage sample is selected from an area frame covering the entire population, and the data are collected annually by means of personal interviews. Improving the quality of estimates of expenses on home renovations and repairs has long been a major issue for Canada's System of National Accounts. The Renovation and Home Repair Survey was developed to try to remedy this situation; it was conducted in February and March 2010. To complete the area frame used for the SHS, a supplementary sample of high-income households was selected from the 2006 Census database, and administrative data were used to update the Census frame. We describe the different options considered, the survey design used and the collection results. We also briefly present findings on the improvement of data quality.


Key Words: Dual sampling frames; integration of samples; skewed household expenditures; census.

## 1. Introduction

### 1.1 Background

The Survey of Household Spending (SHS) is conducted annually by Statistics Canada. It collects data on the expenses of households during the reference year preceding collection. The SHS sample is selected from the area frame of the Labour Force Survey using a stratified two-stage design. Geographic areas are selected in the first stage and dwellings are selected in the second stage. The sample consists of approximately 20,000 households, and collection is accomplished by means of computer-assisted personal interviews (CAPI). The System of National Accounts (SNA) is a major user of the survey data, which are also used to periodically update the weights of the basket of consumer items used to derive the Consumer Price Index.

### 1.2 Problem with data quality

One of the sections in the SHS questionnaire concerns improvements and repairs made to owner-occupied dwellings. It measures expenses related to both outside work (landscaping, roofs, exterior siding, etc.) and inside work (renovation of kitchens/bathrooms, painting, plumbing, etc.) performed by owner-occupants. Secondary residences and rented dwellings are excluded from the scope of these questions. Each question is divided into two parts designed to measure complementary concepts. The first part covers the concept of "improvements" through expenses related to the addition or complete replacement of items. The amounts reported are considered investments made by the household. The second part covers the concept of "repairs," using expenditures related to the repair or maintenance of items. These amounts are considered to be expenses incurred by the household. According to SHS data, Canadian households spent an average of $\$ 5,000$ on improvements and repairs in 2008, for a total of $\$ 43.9$ billion. Most of this spending was devoted to improvements ( $\$ 39.4$ billion) and the rest to repairs ( $\$ 4.5$ billion).

These data constitute an important input for the SNA, which for several years has been hoping to see an improvement in the quality of provincial estimates. This has contributed to the search for a method that would yield better quality data for the 2009 reference year. That year was exceptional, given the financial crisis that was taking

[^104]place at the time.
Expenses for improvements and repairs have a highly skewed distribution. Furthermore, there is a strong relationship between the level of these expenses and household income. Figure 1.2-1 shows average expenditures by income decile according to the 2008 SHS. In this context, it was clear that household income had to have a dominant place in any plan to reduce the variance of the estimates. The use of a supplementary sample of high-income households was seen as a means to achieve this objective. It was also hoped that this sample would improve the representativeness of these households in the SHS and thereby have an effect on the level of expenditures.

Figure 1.2-1
Average expenditures on improvements and repairs by income decile - 2008 SHS, Canada


## 2. Development of a sampling plan for a supplementary sample

### 2.1 Feasibility study

A feasibility study was conducted. Its objective was to explore different options for increasing the number of highincome households from what the SHS sample provided, either by selecting a supplementary sample (two-phase design, list frames) or by modifying the SHS plan by increasing the oversampling of high-income households. The common denominator of these options was that the SHS would remain the main vehicle for measuring these expenditures. The study consisted in identifying the possible options and retaining a few for more detailed analysis. The options were classified into three groups, namely surveys based on a two-phase design, changes to the current SHS sample design and use of list frames. The options were compared rather qualitatively according to the following three criteria: design efficiency (ability to reduce the variability of estimates), the cumulative response burden, and lastly the cost of development and implementation. Surveys using a two-phase design proved to be of particular interest with respect to efficiency, provided that the household income variable in the first-phase survey was of good quality. However, response burden might be greater under this approach. The development and implementation costs of the options based on list frames were higher, but the response burden was lower in these cases. Modifying the current sample design of the SHS offered very little potential for improvements in efficiency, and the response burden was substantial. For details on the analysis of the options and the criteria used, see Malo and Tremblay (2008).

In light of these preliminary analyses, it was proposed that two options be chosen and analyzed in greater detail using a simulation. The options chosen were to create a sampling frame of high-income households using either the Census of Population or a tax file based on income tax returns. These options offered a very good potential for improving quality (the main objective of the supplementary sample) while maintaining the response burden at a quite acceptable level. The simulation study showed that either of these options could generate a potential reduction of $23 \%$ in the
coefficient of variation (CV) of the total expenditure on improvements and repairs, with a supplementary sample of 10,000 high-income households and assuming a conservative response rate of $50 \%$.

### 2.2 Choice of sampling frame

Since the deadlines for development and implementation were quite tight, the choice of sampling frame had to be made very quickly, and the 2006 Census was chosen. In that census, a detailed questionnaire was given to one household in five according to a systematic design. This questionnaire contained some fifty questions, including one on the income of members of the household. It was possible to constitute a sampling frame of high-income households using this rich source of information. The main factors that favoured the census are as follows. This database is used as the sampling frame for a growing number of surveys. Its strengths and weaknesses and the technical details relating to its use are increasingly well-known. This frame also had greater potential than the tax file to reduce the variance of the estimates, because in addition to income, it provided several dwelling-related variables associated with the level of improvement and repair expenditures. For example, the SHS data showed that the year of construction of the dwelling, the number of rooms and the estimated resale value of the dwelling were good predictors of the level of these expenditures. These variables are available in the census frame and cannot be obtained from the tax file. The detailed questionnaire can also be used to obtain a large number of auxiliary variables that are useful in the modelling to correct for non-response or to determine time slices for contacting respondents at an optimal time of day.

Although the census was chosen as the sampling frame, one of its main drawbacks when used to target specific subpopulations is that the information defining these populations deteriorates over time. Income was a key variable for the population of interest. The reference year for the income variable in the 2006 Census is 2005 , whereas the targeted reference year was 2009. To try to minimize the transitions between these periods, matching was carried out between the census frame and the T1FF tax file produced by Statistics Canada. The T1FF is at the census family level and is mainly created from the income tax returns of Canadian households. The tax file available at this point is for the 2007 reference year, and matching by address was carried out. Since the unit targeted was the dwelling, the matching was intended to reduce the impact of moves as well as the effect of possible changes in earnings. Analyses showed that our target population was much less mobile than the population as a whole. Approximately $75 \%$ of households with an income above the $80^{\text {th }}$ percentile in each province were living at the same address five years before the census. In contrast, the corresponding proportion for the population as a whole was $59 \%$. A single direct matching was attempted, and the success rate was $55 \%$. Addresses in rural areas are more difficult to standardize, since a post office box or a description of the dwelling may have been provided in the census. An income growth model was developed on the basis of the matched households (changes observed between 2005 and 2007) and applied to unmatched households to update their income. Note that the deterioration in the coverage of the census frame (dwellings constructed since the census) was not critical, given the dual frame context and the fact that the main survey frame (LFS area frame) provided coverage of these dwellings.

## 3. Renovation and Home Repair Survey

### 3.1 Sample design

The target population of the Renovation and Home Repair Survey (RHRS) includes private households that owned their dwelling at the time of the census. Collective dwellings, Indian reserves and other Aboriginal areas as well as Canada's three northern territories were excluded from the scope of the survey. The target population was also limited to households whose "updated" income was above the $80^{\text {th }}$ percentile of the income distribution of households in their province. This threshold had been used during the simulation study, and the associated reductions of variance were considered quite promising. Also, because high-income households generally have a lower rate of participation in surveys, this threshold was considered a compromise between improving quality and obtaining an acceptable level of participation. Figure 3.1-1 shows the broader context in which the survey was developed. The SHS sample, selected from the area frame, already ensures complete coverage of the target population. That sample was selected using a stratified two-stage design, and personal interviews were conducted. The 2006 Census frame served to ensure coverage of high-income households, and it was the frame from which the RHRS sample was selected using a two-phase design (selection of households receiving the detailed questionnaire in the census,
followed by selection of the RHRS sample), which is more efficient than the SHS design. Collection for the RHRS was conducted via computer-assisted telephone interviews (CATI), a less costly method than CAPI. Of course, a portion of the SHS sample falls within the overlapping domain of high-income households, and the estimation methods will take account of the multiple probabilities of selection of these households.

Figure 3.1-1
Dual sampling frame context for RHRS


The size of the RHRS sample was set at 10,000 households, primarily on the basis of budget considerations. However, this size had been used for the simulation study, and the results were promising, even considering a conservative response rate. Three stratification variables were used: province, year of construction and estimated resale value of dwellings (deciles). These variables were considered because they are stable over time. Given the frame used, we wanted to avoid excessively large transitions. Also, these variables are related to the level of expenditures on improvements and repairs. For example, the older the dwelling, the larger such expenditures tend to be. Similarly, the higher the estimated resale value, the greater the expenditures tend to be. The sample was first allocated to each province using a power allocation technique. The sample size for each province was proportional to the size of the province (to the power $\mathrm{p}=0.4$ ) multiplied by the CV of total improvements and repairs as measured by the SHS over a few years (weighted mean). For two provinces of similar size, this process therefore amounted to allocating more sample to the province for which, in the SHS, the quality of the estimate of improvement and repair expenditures was lower. The sub-provincial allocation used is an allocation proportional to the square root of $x$, where $x$ is the estimated resale value of dwellings.

### 3.2 Collection

### 3.2.1 Pre-collection activities

Clearly, the information used to contact households had to be of the best possible quality. A letter was sent to the selected households shortly before collection. The addresses used were primarily drawn from the Address Register, via the link that is maintained between dwellings in the register and those in the census. In some cases, the address used came instead from the information entered by the respondent on the census questionnaire. In general, these two sources yielded addresses of excellent quality. As mentioned above, since the selection unit was the dwelling, if the household at the time of collection was different from the one living at that address during the census, the interview still took place with the new household. In such cases, the telephone number obtained from the census was no longer linked to the address selected. For this reason, the telephone numbers used came from different sources. The primary source was the census, but the Address Register was also used, since it is linked to administrative sources (telephone companies' billing files) that serve to update the Register's information, including telephone numbers, on a quarterly basis. Lastly, when the RHRS was being developed, Statistics Canada's Operations and Integration Division (OID) was creating a centralized unit responsible for pre-tracing. The sample was therefore sent to this group at the very end of the process, and it yielded new telephone numbers (in relation to the Census and the AR) for approximately $5 \%$ of cases. All numbers obtained from the different sources were included for collection activities, with emphasis on the most recent numbers. It is important to note that for our target population, the proportion of households reporting a telephone number in the census is very high ( $95 \%$ ).

Given the low mobility of our population, census data were also used to determine time slices, a procedure designed to optimize the chances of contact with respondents according to their characteristics. Two groups were created, based on the age of the reference person, and calls were allocated differently for the two groups.

### 3.2.2 Main results

The RHRS collection was carried out from six regional offices in February and March 2010, and the interview time was approximately 20 minutes. The questionnaire included a short section on household composition, questions on dwelling characteristics, the module on improvements and repairs (the same as was used for the 2009 SHS) and lastly, questions on household income. The final response rate was $68 \%$, which is comparable to the response rates obtained in the SHS in recent years. The regional offices identified three factors that worked in favour of the survey: the short duration of the interview, the non-intrusive nature of the survey and pride that respondents felt when reporting the work they had undertaken. The percentage of units outside the scope of the survey was very low at $2 \%$; these were mainly households that were renting at the time of the survey. Using the variable on the year in which the respondent moved in, we deduced that nearly $88 \%$ of the households surveyed were the same as those that comprised the sampling frame. This result is very positive, in the sense that it ensures that the population covered is stable. Furthermore, preliminary analyses show that "new" households have characteristics similar to those that were initially targeted. Average expenditures on improvements and repairs amounted to nearly $\$ 9,000$, which is much higher than the $\$ 5,000$ obtained in 2008 for the population as a whole. Once again, this result is encouraging, considering the objective of the supplementary sample.

### 3.3 Estimation

There were two RHRS estimation methods used: first, the survey's usual weighting and variance estimation methods, and secondly, methods applied in integrating (combining) the sample with the 2009 SHS sample. Since the RHRS has a two-phase design, the initial weights were obtained by multiplying the base census weights (sample of one-fifth of households, namely those that received the long form questionnaire) by the second-phase weights. With respect to non-response, the score method was used for reweighting purposes. Thus, two logistic regression models were developed to estimate response probabilities. One model was developed for non-contacts ( $11 \%$ of the sample) and another for the rest of non-response ( $21 \%$ of the sample). The score method was used because it made it fairly easy to exploit the wealth of the census database with respect to auxiliary variables. Two models were used because the level of non-response was sufficient in each case, but especially because the independent variables were different. Although most of the auxiliary variables chosen came from the census, a paradata item was also considered, namely the number of attempts to contact each household. This variable proved to be the one most strongly associated with response status, which is not surprising. In light of this strong association, an analysis was conducted and revealed that this variable was also associated with household income and, by extension, expenditures. Thus, it was hoped that by including this variable in the models, the non-response bias could be reduced, and the impact on the variance was considered reasonable. Once response probabilities were obtained, a classification algorithm was used to form reweighting classes, and the weight of respondent households was multiplied by the inverse of the weighted response rate in each class.

For variance estimation, the bootstrap method was selected for the RHRS. This choice was based on practical considerations, the main one being that the sample was combined with the SHS sample, and this method was the one preferred for that survey. Initial bootstrap weights were derived, and each series of weights was corrected for nonresponse in the same way as the survey weights, although fixed non-response classes were maintained.

The next step consisted of integrating or combining the RHRS and SHS samples. When the samples were combined, a correction factor had to be applied to the weights of high-income households falling within the overlapping portion of the sampling frames in order to consider these households' multiple probabilities of selection. In our case, identifying SHS households belonging to this domain was no small matter. This is usually done by matching the sample from the area frame with the list sampling frame. Since the RHRS sampling frame is itself a sample, the matching rate would have been very low (a maximum of $20 \%$, given the census sampling rate). For this reason, the identification of SHS households falling within the overlapping portion was instead done conceptually. Since the RHRS sampling frame covered households whose income was above the $80^{\text {th }}$ percentile in each province, the same definition was applied to SHS households. Since this threshold could not be derived from any administrative source at the time the samples were combined, it was determined using current SHS data. Also, SHS households belonging
to the overlapping portion of the sampling frame had to live in dwellings built before the census. Figure 3.3-1 returns to the dual sampling frame situation. The corresponding sample sizes (respondent households) have been added.

Figure 3.3-1
Integration of the RHRS and 2009 SHS samples


The estimator of total expenditures takes the following form: $\hat{Y}=\hat{Y}_{a}^{A}+\left[\alpha \hat{Y}_{a r}^{A}+(1-\alpha) \hat{Y}_{a r}^{R}\right] ; \hat{Y}_{a}^{A}$ represents total expenditures from SHS households outside the overlapping portion; $\hat{Y}_{a r}^{A}$ is the estimator of the total for SHS households falling within the overlapping portion; and $\hat{Y}_{a r}^{R}$ is the estimator of the total for RHRS households. The choice of an alpha between 0 and 1 ensures that the estimator is unbiased. Alpha is determined based on an objective of minimizing the variance. Several solutions exist, as noted in Hartley (1962) and Merkouris (2001), among others. Some are based on the sample sizes of each component, while others are based on the design effects of the estimators to be integrated. Since the bootstrap was the recommended method for variance estimation, it became relatively simple to have alpha vary (by intervals of 0.05 ) and to measure the impact of the variance on the estimator. For each new alpha, the survey weights and the bootstrap weights were multiplied by the corresponding factor ( $\alpha$ for SHS cases falling within the overlapping domain and $(1-\alpha)$ for all RHRS cases). Figure 3.3-2 shows the relationship between $\alpha$ and the CV of the estimator of total expenditures on improvements and repairs at the national level.

Figure 3.3-2
CV of the estimator of the total as a function of the integration factor $\alpha$, at the time of integration


Using an alpha of 1 corresponds to considering only the SHS sample for estimating total expenditures; the CV corresponding to this alpha is $4.6 \%$. At the national level, the optimum alpha is 0.10 . It yields a CV of $3.2 \%$, which represents a CV reduction on the order of $30 \%$. The same analysis was conducted for each province, and the alpha yielding the largest CV reduction ranged from 0.10 to 0.50 . The final choice of the alpha(s) to use had to take a possible mode effect into account, since the SHS interviews were conducted by CAPI, whereas CATI was the preferred approach for the RHRS. If, for example, the expenditures obtained from personal interviews were consistently higher than those obtained from telephone interviews, the use of different alphas for different provinces could generate differences in the estimates that would be entirely due to the relative importance assigned to each survey. For this reason, even though at the provincial level the optimal alphas were different, a single alpha of 0.30 was chosen, without overly sacrificing optimality. At the national level, an alpha of 0.30 yields a CV of $3.3 \%$, which is close to the optimum CV.

Following integration, the methods developed for the SHS to detect influential data were applied to the data of the combined surveys. Finally, the survey weights were calibrated using projections of the population counts for individuals and households at various geographic levels. A comparison of data from the SHS and the combined samples showed a reduction in the CV of the final estimates (including calibration) in the order of $25 \%$.

## 4. Census, surveys, administrative data: a summary

This project involved use of the census, survey data and administrative data, as summarized in Table 4.1-1. The SHS was the main survey and the RHRS was the supplementary survey. The Census of Population served as the sampling frame for the RHRS, and it provided excellent coverage of high-income households. Also, the contact information available in the census was of very good quality. The census also made it possible to use several auxiliary variables that were quite useful for stratifying and allocating the sample. With respect to data collection, the census variables were useful for determining time slices. In weighting, the census variables served as auxiliary variables for modelling non-response, and census counts were used in calibration. As for administrative data, tax data were used to update the sampling frame, the Address Register was used to improve contact information, and billing files were used in pre-tracing efforts prior to data collection.

Table 4.1-1
Use of survey data, the census and administrative data

| Surveys | Stage/Item | Use |
| :--- | :--- | :--- |
|  | SHS | Main survey |
|  | RHRS | Supplementary survey |
|  | Sampling frame | Excellent coverage |
|  |  | Contact information |
|  |  | Time slices |
|  | Weighting | Auxiliary variables (non-response/calibration) |
| Administrative <br> data | Tax data | Address Register |
|  | Billing files | Contact information |
|  |  | Pre-tracing |

## 5. Conclusion

The summary findings presented in Section 3.3 are consistent with what was obtained in the simulations, and they served to achieve the variance reduction objectives that had been set. The option selected could be considered for inclusion in the survey program on household expenditures, provided that the budget required for this supplementary survey is available. The survey was implemented very quickly, and it was quite well received by respondents and interviewers alike. The RHRS was developed under very tight time constraints, with a limited budget. One of the desired improvements includes the matching between the sampling frame and the tax files. One possibility is a probabilistic matching. This article focused on the results obtained with respect to variance reduction, but other analyses will seek to compare the level and type of expenditures obtained by the two surveys. Among other things,
these analyses will aim at determining whether there is a collection mode effect and at assessing how representative the SHS sample is for high-income households.

Clearly, the option chosen may be seen as a fairly costly approach. A survey was developed as a means to improve the quality of data in a single module of the SHS questionnaire. That module is indeed an important one, but it represents only a small portion of all household spending. A possible compromise would be to expand the content of the supplementary survey by adding variables that are also strongly associated with household income. In doing so, the response burden should be maintained at an acceptable level in hopes of maintaining the response rate obtained for the RHRS.

## Bibliography

Hartley, H.O. (1962). Multiple frame surveys, Proceedings of the Social Statistics Section, American Statistical Association, pp. 203-206.

Malo, D. and Tremblay, J. (2008). Preliminary analysis of options for improving the quality of repair and renovation expenditure data as part of the Survey of Household Spending redesign, unpublished report, Ottawa, Canada: Statistics Canada.

Merkouris, T. (2001). Cross-Sectional Estimation in Multiple-Panel Household Surveys, Survey Methodology, December 2001, Vol. 27, No. 2, pp 189-200.

# Surveys and administrative files in France, The Impact of hashing 

Benoît Riandey ${ }^{2}$


#### Abstract

In the 1970s, there were hopes that with the computerization of administrative files, registers could be matched using a unique identifier, resulting in a major step forward in statistics. The new technology also offered the possibility of reducing response burden and improving data quality. While these advances actually materialized in northern Europe, they were stymied in France by the Information Technology and Civil Liberties Act. Fear of "Big Brother" made the idea of matching unthinkable-until the current technological revolution. Today, the law is less draconian, minds have changed, and files can be matched anonymously, in part using identifier hashing techniques. Epidemiologists have used them to count AIDS patients anonymously without double counting, for inter-hospital statistics and for enhancing patient cohorts. France's public statistics system is now taking an interest in those techniques, but has not yet taken the first step. We will provide examples of their potential uses in job search surveys, health surveys and surveys of discrimination in the private and public sectors: matching administrative files, complementing surveys with anonymous administrative data, or adding sensitive data to an anonymous statistical file derived from administrative or management data. It is one way of resolving the dilemma that statisticians often face: produce the desired results without collecting the required data, due to privacy rules.


## 1. Hopes dashed

France has long had a wealth of administrative files, and considerable progress has been made in recent years. One can imagine gathering various kinds of administrative data about an individual, like sections in a survey questionnaire, by matching files using a common identifier: the Vital Statistics registration number (NIR).

Such an approach would work for employment, health, education and taxation data. Employers are required to submit an annual report of social data (DADS) and salaries of all paid employees at a particular point in the year (Lagarde, 1998). Independent occupations will soon be added to that file, known as the DADS file. There is also an historical file of unemployment episodes covering all individuals who have ever registered for unemployment, as well as a file of the complete history of students' university studies, and there may soon be one for pre-university studies as well (Goy, 2005). These files are an exceptional resource. Other sources are the comprehensive annual file of household taxable income (Goutard, 2008), soon to include household social income, and the comprehensive record kept by the National health insurance inter-plan information system (SNIIR-AM) of all personal medical goods and services covered by the plans (Lenormand, 2005). The combined (and controlled) use of these files would produce an extraordinary social statistics system, like the one our counterparts in northern Europe have already established.

Unfortunately, that is not the way it turned out in France. Public mistrust of the government, born of the dark periods of the Second World War, has impeded the development of the public statistics system. The 1978 Information Technology and Civil Liberties Act failed to recognize the guarantees provided by statistical confidentiality and crippled the statistical function by subjecting it to the same constraints as administrative processes. All that changed in 2004 thanks to the beneficial impact of European law.

In concrete terms, the 1978 law included a cumbersome procedure requiring publication of a cabinet order before the Vital Statistics identifier, the NIR, could be used. That administrative red tape resulted in self- censorship by statisticians, despite a few brilliant successes. The statistical system was the law's primary target, due to a heated, awkward, unfair controversy when the Vital Statistics inventory was digitized. The rulings of the National

[^105]Commission on Information Technology and Civil Liberties (CNIL), which is responsible for enforcing the Act, resulted in the institutional compartmentalization of identifiers, without regard for validation or coordination. Use of the NIR was restricted to employment, social assistance and health insurance. A special identifier, the INE, had to be introduced for education, and there have been serious difficulties in ensuring its universal use in the education system and in eliminating duplication. A distinct identifier was also required for taxation services, which made it impossible to correctly identify taxpayers. This led to another heated dispute, which was resolved by a Constitutional Council decision allowing taxation services to use the NIR to validate the identifier and communicate with social services. Unfortunately, the Council's judicious ruling did not extend to the national education system, but we have a similar solution to suggest.

## 2. Innovations by epidemiologists and the health insurance system

Driven by the imperative to heal and served by the historical recognition of medical confidentiality, epidemiologists could not accept self-censorship. The undisputed public interest of their mission was strengthened by legal support of the highest order, since the protection of human life is enshrined in the preamble of the French constitution, on the same footing as the Information Technology and Civil Liberties Act, and well above the authority of the Statistics Act of June 1951.

It was their duty to match confidential files from different medical institutions, for such purposes as anonymously counting HIV-positive patients or from inter-hospital birth clinic files (Quantin, 2005). To that end, they introduced, in France, irreversible encryption or hashing techniques using Shamir's Secure Hash Algorithm (SHA) in simple (or deterministic) matching based on one identifier or using the probabilistic methods proposed by Fellegi and Sunter (1969), which Jaro (1995) combined with the SHA hashing. On the recommendation of Professor C. Quantin, these techniques are used by the Health Monitoring Institute to record reportable diseases and to find out from Vital Statistics whether former cancer patients are alive or dead.

At that time, the health insurance system had to assemble individual records of stays in private or public hospitals by matching successive stays by the same person (the information systems medicalization program (PMSI)); medical confidentiality required an anonymous procedure. Hashed NIRs were used for that purpose (Trouessin, 1997) and more recently for the gigantic new health insurance database (SNIIR-AM). The plan is to record all benefit payouts made to 60 million health insurance plan members over the last three years. If we assume that on average, each person has 10 prescriptions or medical appointments a year, the SNIIR-AM will contain 1.8 billion records.

Using tested techniques, epidemiologists and health economists carried out multiple matches of sensitive administrative files, registers of diseases such as cancer, and survey data. Below are some examples of various uses.

The last two household health surveys conducted by the National Institute of Statistics and Economic Studies (INSEE) and the Ministry of Health (Bouvier, 2010) used the census frame and were administered in the field by interviewers in order to collect data. Thanks to a cabinet order, the NIRs were collected unencrypted and, after hashing, were used to retrieve data from the SNIIR-AM on the medical goods and services for which each household was reimbursed by the health insurance plan.

In contrast, the Institute for research and information in health economics (IRDES) administers the health, health care and insurance survey (ESPS) (Allonier, 2010) to a sample drawn from health insurance files. Collected by telephone or personal interview, the data complement the information in the SNIIR-AM, which relieves the IRDES of the need to access NIRs. However, the addresses, which are often erroneous, are not from an optimized sampling scheme. The respondents are selected from a permanent sample of persons covered by social insurance. They form a panel, which is surveyed over four years (half every two years).

The IRDES also runs matches on the HYGIE file, which deals with absenteeism due to illness (Debrand, 2010). The sample is taken from employers' annual reports (DADSs). Since that administrative file contains NIRs, it can be matched, after hashing, with health insurance data while maintaining medical confidentiality. To the same end, the COSMOP project: Monitoring of mortality by occupation cohort (Cohidon, 2010) matches causes of death with the occupations file derived from past DADSs.

The EDISC file tracks occupational mortality (Saurel, 2009). It is based on INSEE's permanent demographic sample (see section 3 below). That longitudinal file is a compilation of successive census questionnaires of a $1 / 100$ sample based on day of birth. Matched with causes of death, it forms the basis of studies of occupational mortality. Census data on occupations are much more informative than Vital Statistics data on occupational status at the time of death. Similarly, the Health Monitoring Institute's COSMOP project looks at mortality by cause of death for a sample of paid employees whose data are provided in the DADSs (Thuret, 2009).

The National institute for demographic studies's (INED) end-of-life survey, for which collection has just been completed, deals with a subject that is even more sensitive, because it relates to the Hippocratic oath: whether doctors should intervene at the very end of a person's life. The subject is more sensitive in the Catholic countries of southern Europe than in the Protestant countries of northern Europe. Deaths are pulled from the Vital Statistics inventory without using NIRs, then the confidential questionnaire is completed by the pronouncing physician. In the European survey (Bilsen, 2007), anonymity was maintained with the assistance of a trusted third-party lawyer. In France, the INED's procedure was fully automated using the Internet and triple hashing.

Epidemiologists conduct completely different types of surveys through their cohorts. For example, the INED and the National institute of health and medical research (INSERM) will take the cohort of children from the French Longitudinal Study of Children (ELFE): 20,000 births that take place on 24 days scattered evenly through the year at a sample of clinics. A perinatal survey at the birth clinic will be followed by a survey at home, retrieval of data from the SNIIR-AM, medical examinations, a school survey, and so on for 20 years. In this case as well, the team is not eager to undertake the process of obtaining a cabinet order and is therefore considering the procedure for accessing the SNIIR-AM. The CNIL's view is that a hashed NIR is not an NIR and therefore does not require a cabinet order. This means that a respondent could enter his or her NIR, which would immediately be hashed. The data anonymization and segregation module is particularly sophisticated. Law and survey methodology are tightly interwoven.

The Constances cohort (Goldberg, 2008) takes the process to a whole new level: the records of 200,000 current or former paid workers will be extracted from the longitudinal pension contributions file, including their occupational data, a summary of their health insurance data and medical examinations, and ultimately their vital status and causes of death. How is that for industrial-scale statistics! The INSERM and the health insurance system plan to establish a secure matching centre (Plastico platform) to carry out all these operations.

These innovations are not without their limitations. One relates to the statistical unit, the other to the nature of the data. We have already discussed the legal nature of the identifier.

The public health insurance system was founded on a Bismarckian concept of worker-based social rights. The statistical unit is therefore the person covered by social insurance along with his or her dependent spouse and children. The information system has evolved toward a truly demographic frame in which the beneficiary is the statistical unit, but in this transitional phase, the identifier is still based on the worker's NIR along with the beneficiary's date of birth. This means that when the beneficiary changes status - by taking a job - his or her identifier changes. This in turn leads to a series of statistical problems, which will ultimately be resolved when the information system reaches maturity and the beneficiary's NIR becomes the identifier. The demographic identifier consistency is currently maintained on a $1 / 100$ sample that will be used to select the new waves for the IRDES's health, health care and insurance survey (ESPS).

The other limitation relates to the nature of the data. In contrast to the hospital PMSI and except for long-term conditions (cancer, diabetes, etc.), SNIIR-AM data do not include medical diagnoses. As encompassing as those economic data on health are, they are insufficient for the purpose of steering health policy, and they need to be anonymously enhanced with data direct from medical records. In the debate about the introduction of the personal medical file (DMP) - sometimes considered shared and sometimes condidered personal, the CNIL decided that medical data would have a distinct identifier from reimbursement data. So the national health identifier (INS) was born. Epidemiologists levied two requirements on the INS: it had to be anonymized to keep medical records secure, and there had to be coordination between identifiers, particularly between the INS and the NIR, so that data on expenses or hospital stays could be matched with diagnoses and mortality data (vital status, date of death and causes of death).

Professor Quantin, an expert in secure matching of medical data, had pointed out a limitation of hashing or irreversible encryption: if one team establishes a diabetes register and another team sets up a register of visually impaired persons, with each register protected by a different INS hashing, the teams will not be able to use the INS to match the two databases for the purpose of analyzing the risks of blindness due to diabetes. They can, of course, use probabilistic matching, but careful management of the INS offers another solution. The following procedure for anonymizing epidemiological files was proposed: The INSs in all the files would be hashed with a common key and then encrypted with a specific key for each project; the project keys would be stored at an identifier coordination office, a sort of institutional safe under the CNIL's authority. Matching processes approved later could be performed at a secure matching centre such as Plastico. The French Statistical Society plans to use the same method with NIRs for public statistics. It is worth noting that the introduction of the INS requires the creation of a correspondence table between INSs and NIRs or between hashed INSs and hashed NIRs, and that in any case, epidemiologists are still planning, with the CNIL's approval, to have a trusted third party break a respondent's anonymity if a diagnosis made during an epidemiological study reveals that the respondent has a serious illness requiring treatment.

## 3. Social statistics: Accomplishments and opportunities

The statistical hopes of the 1970s were put in the deep freeze by the Information Technology and Civil Liberties Act and the need for a cabinet order before NIRs could be used. Nevertheless, INSEE and the public statistical system had performed a small number of brilliant matches, including the taxable income survey and the permanent demographic sample.

For many years, INSEE produced the best income estimates for France by asking the income tax directorate to extract the income tax returns of a sample of households drawn from the census. That survey of taxable income, when matched with the household consumption survey, the French component of the European socio-economic panels and employment surveys, resulted in a substantial advance in the quality of the economic data from those surveys. The amendment of the Act in 2004 allowed INSEE to access the comprehensive taxable income file, soon to include social income (Goutard, 2008).

The permanent demographic sample (EDP) is based on the compilation of census questionnaires of a $1 / 100$ sample of the population selected by day of birth, together with Vital Statistics data for the sample persons or the same dates (Couet, 2006). INSEE carefully coordinated a sample taken from employers' salary reports (then DASs, now DADSs) (Lagarde, 1998), but the CNIL's reluctance about matching and the EDP had dissuaded INSEE from performing that match until recently, despite the exceptional wealth of information it was expected to provide. That gold mine of information is especially rich since the DADS panel was recently expanded to include public servants and in the future will include self-employed workers.

In large towns, the revised census has replaced full coverage with an annual $8 \%$ survey. That change, which was ideal for national or regional cross-sectional analyses, seemed to put an end to the longitudinal operation of the EDP. However, household income tax returns constitute an annual population census that demographers have long ignored. The returns will provide the demographic information for the EDP that the new "census" does not collect, along with economic data.

Such matches used only names and addresses, except for social data readily identified by NIR. In contrast, ministry statistical offices tended to use the NIR when dealing with retirement pensions or employment. The old age security system is so fragmented in France that many retirees receive basic pensions from a number of institutions. As a result, for many years, the administrative statistics system reported the number of pensions paid out but overestimated the number of pensioners (because of pensioners receiving multiple payments), understood the distribution of the pensions being paid out, but did not know the distribution of pension income. In the late 1980s, data were collected from a permanent sample of retirees selected from a sample with unencrypted NIRs, through extraction of information thus identified in the files of all those pension plans. More recently, an identical process was carried out for pension plan contributors, yielding very useful economic projections for the pension system. Data were also collected from a panel of social assistance beneficiaries with the NIRs unencrypted, as the files of employees and the unemployed are matched. While the letter of the law is followed when a cabinet order is obtained, one cannot help but wonder whether, ethically, it would be better to use anonymous hashed NIRs, which serve the purpose and are simpler from an administrative perspective. That would make the administrative statistics system
more dynamic, since it tends to focus on one administrative file at a time and is unable to take advantage of complementary information from other files, because the need-to-know principle forbids one administrative office from knowing what the next office is doing.

That is the position of the French Statistical Society, and its position is strongly supported by the CNIL, witness the fact that the CNIL's director of legal affairs concluded her presentation in a Society-sponsored course on secure matching with the injunction "Statisticians, be daring!" (Gensbittel, 2007). She stressed the fact that since a hashed NIR is not an NIR, its use does not require a cabinet order. That message was heard by the national education system's statistical office (Goy, 2005), which hashes the INE identifier to match statistical files about students and uses probabilistic matching when the INE identifier is unreliable, for example, in statistical files of examination results. The statistical office also has a more ambitious project (FAERE) to anonymously record students' entire educational history using hashed INE.

The INE has the disadvantage of not being validated, whereas the NIR is, thanks to the Vital Statistics inventory. As a result, many students have more than one INE, which makes matching ineffectual for them. We propose the following remedy: Every student should enter his or her INE and NIR on a computer; after the NIR is hashed, the INE/NIR pair would be transmitted to a national office, which would determine whether two INEs correspond to the same hashed NIR and therefore refer to the same person.

The correspondence table between INEs and hashed NIRs would not only be useful for unduplicating INEs but also serve a second, more basic purpose. It would allow cross-tabulation of occupations and education histories. That is the goal of the Centre of studies and research on qualifications (Céreq) generation survey, which tracks the transition to work over five years of a cohort of people exiting the educational system (Olivier, 2008). At the time of the survey, a large portion of the sample could not be reached. Quality information about the initial employment in France of both non-respondents and respondents can be obtained directly through anonymous annual extraction from the DADS employment file.

This is a case where Professor Quantin's suggestion regarding the use of hashed INSs would truly benefit public statistics if it were applied to the hashed NIR (Picard, 2010). If every NIR-managed administrative file is retained for statistical purposes and uses identically hashed NIR together with specific encryption, all matching validated by the National Statistical Information Council and approved by the CNIL can be carried out.

While a conventional panel of persons leaving unemployment is not without interest, the employment re-entry of a sample of unemployed people could be observed continuously with the DADS file. It represents a potential revolution for French public statistics, a revolution that nevertheless complies with the unavoidable requirements of data security. It is worth noting that the Ministry of Labour has just performed an experimental match between the DADS file and the historical unemployment file.

And yet, as we all know, methodology isn't enough. Without public trust, it cannot succeed. A case in point is the social controversy surrounding the pupils' database. The latter is an elementary school (maternelle and primaire) management file, and it is the first step in the FAERE project. Together with the secondary school (collège and lycée) file and the postsecondary file (SISE), it will form a longitudinal file covering students' entire educational history. However, a cabinet order gave people from outside the national education system access to the database. Parents' associations and teachers' unions were afraid that the file would be used to identify and deport parents who were illegal immigrants, and they managed to have all information about nationality, country of birth and mother tongue removed from the file. The statistical file that would be derived from the administrative file in its current state would lack the variables required to assess how well the French educational system meets the needs of people of non-francophone origin. Will we succeed in reintroducing those variables into an anonymous statistical file using secure matching?

Similarly, we can imagine a secure, anonymous statistical system (Quantin, 2009) for observing businesses from outside and analyzing any discrimination that might be occurring within them, without giving the businesses permission to collect such sensitive data themselves, which the CNIL opposes. However, such systems can work only if there is public trust, which does not seem consistent with the current climate in France.

## 4. Conclusion

At this time of very tight budgets for statistics, there is renewed interest in the efficiency of administrative statistics in France, led in part by epidemiologists and health economists. Secure matching techniques are likely to play a significant role in that movement. Will statisticians start using them? Will they be able to enter into confidencebuilding communications with the public? That will be a major issue for statistics in the coming years.

## References

Allonier, C., Dourgnon, P. and Rochereau, T. (2010). Enquête sur la santé et la protection sociale 2008, IRDES rapport 1800. (www.irdes.fr/Publicatons/Rapports 2010/rap1800.pdf)

Bilsen, J., Cohen, J. and Deliens, L. (2007). La fin de vie ne Europe: le point sur les pratiques médicales,
Population et sociétés $\mathrm{n}^{\circ} 430$ (www.ined.fr).
Bouvier, G. and Niel, X. (2010). Les discriminations liées au handicap et à la santé, INSEE Premières $\mathrm{n}^{\circ}$ 1308. (www.insee.fr).

Cohidon, C., Geoffroy-Perez, B. and al. Suicide et activité professionnelle en France: Premières exploitations de données ((COSMOP)) disponibles, avril 2010, www.invs.sante.fr

Couet, Ch. (2006). L'échantillon démographique permanent de l'INSEE, Le Courrier des statistiques ${ }^{\circ}$ 117-119, (www.insee.fr).

Debrand, Th. (2010). Le projet Hygie, ( www.irdes.fr/hygie).
Felligi, I. and Sunter A.B. (1969). A theory for record linkage, JASA n ${ }^{\circ}$ 4, pp 1183-1210.
Gensbittel, M.H., Riandey, B. and Quantin, C. (2007). Appariements sécurisés : statisticiens, ayez de l'audace! Le Courrier des statistiques $\mathrm{n}^{\circ} 124$ (www.insee.fr).

Goldberg, M., Quantin, C., Guéguen, A. and Zins M. (2008). Bases de données médico-administratives et épidémiologie : intérêts et limites, Le Courrier des statistiques $\mathrm{n}^{\circ} 124$, (www.insee.fr).

Goutard, L. and Pujol, J. (2008). Les niveaux de vie en 2006, INSEE Premières n ${ }^{\circ}$ 1203, (www.insee.fr).
Goy, A. L'appariement sécurisé de fichiers d'etudiants grâce au «hachage » des identifiants, Le Courrier des statistiques $\mathrm{n}^{\circ}$ 113-114, (www.insee.fr).

Jaro, M.A. (1995). Probabilistic-linkage of large public health data files, Statistics in medicine, 14, pp 491-498.
Lagarde, S. (1998). La nouvelle exploitation exhaustive des DADS, Le Courrier des statistiques $\mathrm{n}^{\circ}$ 85-86, www.insee.fr (publications, Collections nationales).

Lenormand, F., (2005). Le système d'information de l'assurance maladie, le SNIIR-AM et les échantillons de bénéficiaires, Le Courrier des statistiques $\mathrm{n}^{\circ}$ 113-114, (www.insee.fr).

Olivier, J., Lopez, A., and Ryk, F. (2008). Génération 2004 : Des jeunes pénalisés par la conjuncture, Bref n ${ }^{\circ} 248$ (www.cereq.fr).

Picard, N., Riandey, B., Solaz, A., and Quantin, C. (2010). Les besoins des démo-économistes en matière d'appariements sécurisés, Le Courrier des statistiques n ${ }^{\circ} 129$, (www.insee.fr).

Quantin, C., Gouyon, B., Allaert, F.-A., and Cohen, O. (2005). Méthodologie pour le chaînage de données sensibles tout en respectant l'anonymat: application au suivi des informations médicales, Le Courrier des statistiques $n^{\circ}$ 113-114, (www.insee.fr).

Quantin, C., Riandey, B., Benoit, A.M., and Cohen, O. (2009). Proposition à l'intention du COMEDD d'un dispositif sécurisé de suivi statistique de la diversité dans les entreprises à l'attention des entreprises signataires de la Charte de la diversité, annexe du rapport du COMMED.

Saurel-Cubizolles, M.J., Chastang, J.F., Menvielle, G., Leclerc, A. and Luce, D. (2009). Social inequalities in mortality by cause among men and women in France, J Epidemiol Community Health 63(3):197-202.

Thuret, A., Geoffroy-Perez, B., Luce, D., Goldberg, M. and Imbernon, E. (2009). Analyse de la mortalité prématurée dans la construction, $\mathrm{BEH}^{\circ}$ 30, 21 juillet 2009 pp 325-328.

Trouessin, G. and Allaërt, F.A. (1997). FOIN: a nominative information occultation function, MIE, 3, pp196-200.

# American Community Survey design and statistical methodology 

Alfredo Navarro ${ }^{1}$


#### Abstract

The American Community Survey is the Census Bureau's alternative for replacing the decennial long form in the 2010 Census. New ventures are usually accompanied by challenges. Census long form data are heavily used for model development by transportation planners and grants allocation, as well as in a variety of other economic and social applications. Data users will now have to start developing these applications based on ACS estimates. The ACS design, data collection, and statistical methodology are somewhat different to methods used by the traditional census long form. The paper will touch on key design decisions, such as residence rules, data collection methodology, and the construction of the sampling frame. It also includes a description of key aspects of the ACS statistical methodology for the production of multi-year estimates. The switch to continuous measurement and the production of multiple sets of estimates result in a whole set of new issues regarding the use and interpretation of these estimates. The paper addresses some of these issues from a data user perspective.


[^106]
## SESSION 11C

VALIDATION

# Linking the National Health and Nutrition Examination Survey to Supplemental Nutrition Assistance Program administrative data: A one state pilot study 

Lisa B. Mirel, Vicki Burt, Donna Miller and Christine Cox ${ }^{1}$


#### Abstract

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative survey designed to assess the health and nutritional status of non-institutionalized adults and children in the United States. The NHANES questionnaire includes food security questions about receipt of Supplemental Nutrition Assistance Program (SNAP) benefits, formerly called food stamps, and can be used to estimate the prevalence of food stamp receipt. However, estimates of receiving food stamps benefits based upon reported NHANES information are lower than estimates from the United States Department of Agriculture. To examine potential reasons for differences in the estimates, a pilot study was initiated to link NHANES data with the SNAP administrative database in the state of Texas. Analysis of the linked data showed that of the NHANES participants who reported household authorization of food stamps in the past 12 months, $77 \%$ were identified as SNAP recipients in the Texas administrative database; of NHANES participants who reported no household authorization in the past 12 months, $8 \%$ were identified as SNAP recipients in the Texas administrative database. We compare the demographic characteristics of individuals by concordance on SNAP benefit status between NHANES and the administrative record database. The results of this pilot study provide a basis for evaluating the utility of NHANES food security data and offer preliminary insights into the issues of accurately collecting this information in a self reported survey questionnaire.


Key Words: National Health and Nutrition Examination Survey (NHANES); Supplemental Nutrition Assistance Program (SNAP); Food Stamp; Linkage.

## 1. Introduction

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative survey that collects information on health and nutritional status of adults and children in the United States, including detailed information on food intake and nutrition. A recent National Academies' panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs recommended linking the NHANES data with food assistance records (National Research Council, 2005). Acting on this recommendation, the National Center for Health Statistics and the United States Department of Agriculture's Economic Research Service conducted a pilot study to link the 2005-08 NHANES data to one state's Supplemental Nutrition Assistance Program (SNAP) administrative records, which were formerly called food stamp records. A main goal of this project is to develop a model for conducting linkages between NHANES and state SNAP databases. This pilot study provides the opportunity to assess methodologies for linking NHANES to a state database, to help define the logistics involved in conducting such linkages, and to learn from results, including learning possible ways to improve the questionnaire design in NHANES.

Linking a nationally representative survey to SNAP administrative records has notable challenges. There is no national database for recipients of SNAP. Instead SNAP data are collected at the state level, and linkages of national surveys to SNAP administrative data must be done state by state.

[^107]Previous studies have linked population survey data, such as the American Community Survey (ACS), to state SNAP administrative data (Taeuber et al, 2004). The results of the ACS linkage to food stamp records found underreporting of food stamp receipt in the survey compared to the administrative records. The study found that underreporting was a function of the person in the household receiving the benefit, the length of time for the benefit and the interval between the survey and the receipt of benefit. In addition, the ACS study found that the family definition and the survey coverage contributed to the underestimate.

Linking NHANES to SNAP administrative records can provide some information that was not available from the previous ACS study. For example, ACS does not collect detailed information on food intake and nutrition, and, therefore, the linkage results of ACS cannot address issues related to food consumption and SNAP benefit receipt.

To date, there has never been a linkage between a nationally representative survey that collects detailed information on food intake and nutrition to state SNAP administrative databases. The NHANES linkage in this pilot study to Texas administrative SNAP records begins to fill that gap. The results of this pilot study will provide a basis for evaluating the utility of NHANES food security data and will offer preliminary insights into opportunities to improve the collection of program participation information in a self-reported survey.

### 1.1 NHANES Background

Currently NHANES is a continuous survey (1999-present) that includes an interview in the household followed by an examination in a Mobile Examination Center (MEC). NHANES is a representative cross-sectional sample of the U.S. civilian, non- institutionalized population that is selected using a complex, multistage probability design. Detailed descriptions of the NHANES sample design have been described elsewhere and only will be summarized here (Mohadjer and Curtin, 2008). The first stage selects the primary sampling units (PSUs). These units are mostly single counties or, in a few cases, groups of contiguous counties which are selected with probability proportional to a measure of size. At the second stage of sampling, the PSUs are divided into segments, which are generally city blocks or their equivalent, and selected with probability proportional to size. At the third stage households within each segment are listed, and a sample is randomly drawn. At the fourth stage individuals are chosen to participate in NHANES from a list of all persons residing in selected households. Individuals are drawn at random within designated age, sex, race/ethnicity, and low income screening sub-domains. There may be more than one survey participant per household. Participants can be children, adults or both. An adult proxy responds to the majority of the questionnaire for children who are under the age of 16 .

There are some key aspects about the NHANES design that should be noted as they relate to this pilot project. For example, NHANES is not designed to provide state estimates. The NHANES annual sample consists of 5,000 survey participants drawn from approximately 15 counties. The results from this pilot project are not generalizable to Texas or appropriate for national estimates. Additionally, the sample design allows multiple members of the same household to be selected as survey participants.

The NHANES household interview includes food security questions about receipt of SNAP benefits. This article focuses on one of these NHANES food stamp questions asked at the household level for survey years 1999-2006: "[In the last 12 months], were $\{y o u /$ you or any members of you household $\}$ authorized to receive Food Stamps [which includes a food stamp card or voucher, or cash grants from the state for food]?" In NHANES 2007-08 this question was changed to "In the last 12 months, did \{you/ you or any member of your household\} receive food stamp benefits?"

This article examines the 2005-06 NHANES data that were linked to the administrative database and focuses on the 2005-06 NHANES question on household benefits in the past 12 months. This analysis will be relevant for future assessments of the 2007-08 NHANES food stamp related questions which included a questionnaire change to ask about household receipt of benefits, rather than authorization to receive benefits.

## 2. Methods

### 2.1 Texas

Texas was chosen for this pilot study because it was visited several times by NHANES and it has a centralized administrative food stamp database. There are 254 counties in the state of Texas, and the 2005-06 NHANES sample includes several Texas counties. The Ray Marshall Center (RMC) at the University of Texas, in Austin, performed the linkage of NHANES data to the Texas administrative database under contract. Records were considered potential matches if the NHANES participant matched to the Texas database at any time 48 months before or after the NHANES interview. However, in this article we limit the match results to only those that matched at some point in the 12 months prior to the NHANES interview to correspond to the NHANES food stamp question about household authorization in the past 12 months.

### 2.2 Linkage eligibility criteria

NHANES participants were eligible for linkage if they were part of the Texas NHANES sample, if they did not refuse to provide a social security number (SSN) at the NHANES interview and they provided sufficient personal identification information (PII) for linkage. Eligibility for linkage for an NHANES participant was independent of the questionnaire responses to the food stamp questions and independent of a participant's SNAP program eligibility.

Linkage was performed at the survey participant level and not at the household level. NHANES does not collect PII on members of the household who are not survey participants and, therefore non-surveyed household members could not be linked. However, the NHANES question was asked at the household level.

### 2.3 Scoring algorithm

To determine potential matches between the NHANES records and the Texas administrative database we developed a scoring algorithm; based on the algorithm, survey participants could have none, one, or more potential matches, each with a different match score. Different weights were assigned to agreement on each personal identifier and these weights were summed to calculate a final match score. Potential matches were required to match on gender. Birth day, birth month and a three category race variable were required to match exactly to receive a weight. Weights for year of birth and zip code differed for exact and near matches. If there was a match on year of birth, a higher weight was assigned than if the year of birth was within 3 years. Zip code was weighted based on the date of NHANES interview and the date of food stamp benefit. If the zip code matched and if the food stamp receipt was in the same month as the NHANES interview, the weight for that match was greater than if the zip code matched but the food stamp receipt was within 2 years of the survey. We also used a match tolerance for name and social security number. The spelling distance function available in $S A S ®$, which determines the likelihood of two words matching using an asymmetric spelling distance between the words, was used to match names (SAS Institute Inc., 2010). Weights for names with a spelling distance value less than 0.8 were given a weight of zero. For SSN the weight was calculated as the proportion of the 9 digits numbers that matched.

The maximum match score was 10 . We treated survey participants who matched to a SNAP record with a score of 7.8 or higher as matches. We considered high quality matches to be those with a score of 8.2 or higher. Most potentially matched survey participants generated at least one high quality match. About $6 \%$ of the match results were below 8.2 which were considered to be a lower quality match. The assumption was that a "true" cut-off lies somewhere around 8.0 and potential matches near the cut-off have a moderate probability of being misclassified.

### 2.4 Evaluation

Concordance and discordance were defined conditional on NHANES responses and overall. Discordant and concordant groups were compared using gender, age, Hispanic origin, and matching scores. As this project is a case study, not a full NHANES analysis, survey weights and design variables were not used in the calculations. Further, statistical tests were not done; statements such as "lower" or "higher" are based on inspection of the percents, not statistical significance.

## 3. Results

There were about 1,000 NHANES participants who were part of the Texas sample in 2005-06. About $12 \%$ of these participants were ineligible for linkage. Of the 882 eligible for linkage, 188 participants matched to the Texas administrative database for having received food stamp benefits at some time point in the 12 months prior to the NHANES interview (Figure 1). Of the 882 NHANES participants who were eligible for linkage, 858 answered the question about authorization for food stamp benefits in the past 12 months. Of the 858 , 159 indicated during the NHANES interview that their household had been authorized for food stamp benefits in the past 12 months and 699 did not. Similarly, of the 858 who answered the question about authorization for food stamp benefits in the past 12 months, 174 matched at some time point to the Texas administrative records in the 12 months prior to the NHANES interview and 684 did not.

Figure 1. Flow chart for Texas sample from the 2005-06 NHANES linkage to Texas food stamp administrative records.

NHANES Texas sample:
$\mathrm{N}=1,004$

| Ineligible for linkage: | Eligible for linkage: |
| :--- | :--- |
| $\mathrm{N}=122$ | $\mathrm{~N}=882$ |

Did not answer the NHANES food stamp question: $\mathrm{N}=24$

Answered the NHANES food stamp
question: $\mathrm{N}=858$

Did not match at anytime to the Texas administrative database (+/-48 months of interview): $\mathrm{N}=532$

Matched at some time to the Texas administrative database (+/- 48 months of interview): $\mathrm{N}=326$

Did not match at any time 12 months prior
to the NHANES interview: $\mathrm{N}=152$

Matched at some time 12 months prior to the interview $\mathrm{N}=174$

If the reporting of food stamp benefits from NHANES participants were perfect we would not expect to find a Texas administrative record for those NHANES respondents who reported that no one in their household was authorized to receive food stamp benefits in the past 12 months. Similarly, we would generally expect to find an administrative record for those NHANES respondents who reported that someone in their household was authorized to receive food stamp benefits. However, an administrative record would not be expected if the NHANES respondent was not one of the household members authorized to receive food stamp benefits.

### 3.1 Concordance

Of the respondents who said "No" to the question in NHANES, $8 \%$ were found in the Texas administrative database (making the agreement level $92 \%$ ). Of the respondents who said "Yes" to the question in NHANES about being authorized to receive food stamps, $23 \%$ were not found in the Texas administrative database ( $77 \%$ agreement). Combining these numbers leads to an overall agreement of $89 \%$.

The match scores for the $8 \%$ who said "No" to the food stamp question in NHANES but were found in the Texas administrative database had a mean value of 9.09 and median value of 9.35 , indicating on average, a high quality match. The mean income to poverty threshold ratio for this group was 1.32 and the median value was 0.89 . Among those who reported no household authorization for food stamp benefits, the percentage who were Hispanic was higher among those with a Texas administrative record ( $82 \%$ of 55 ) than those with no Texas administrative record ( $53 \%$ of 644) (Table 3.1-1). In addition, among those who reported no household authorization for food stamp benefits, the percentage who were less than 20 years of age was higher among those with a Texas administrative
record ( $62 \%$ of 55) than those with no Texas administrative record (53\% of 644) (Table 3.1-1). Gender did not differ by concordance with the Texas administrative database among those who reported no food stamp benefit authorization.

Table 3.1-1
Demographic characteristics of 2005-06 Texas NHANES participants who reported that no one in the household was authorized to receive food stamps in the past 12 months by concordance with the Texas administrative database

|  | Concordant <br> No administrative record <br> No NHANES <br> N=644 | Discordant <br> Yes administrative record <br> No NHANES <br> N=55 |
| :--- | :--- | :--- |
| Characteristic | Percent | Percent |
| Gender: Male | 48.8 | 52.7 |
| Age (years): <20 | 52.8 | 61.8 |
| Race/ Ethnicity: Hispanic | 53.0 | 81.8 |

Among respondents with positive responses in NHANES, the income to poverty threshold ratio for those who were not found in the Texas administrative database had a mean value of 1.51 and median value of 1.41 . Among those who reported household authorization for food stamp benefits, the percentage who were less than 20 years of age was higher among those with a Texas administrative record ( $73 \%$ of 123 ) than those with no Texas administrative record ( $39 \%$ of 36 ) (Table 3.1-2). In addition, among those who reported household authorization for food stamps benefits, the percentage who were Hispanic was slightly lower among those with a Texas administrative record (45\% of 123 ) than those with no Texas administrative record ( $53 \%$ of 36 ) (Table 3.1-2). Gender did not differ by concordance with the Texas administrative database among those who reported food stamp benefit authorization. More research is needed to better understand the differences in concordance observed by age and Hispanic origin.

Table 3.1-2
Demographic characteristics of 2005-06 NHANES Texas participants who reported that someone in the household was authorized to receive food stamps in the past 12 months by concordance with the Texas administrative database

|  | Concordant <br> Yes administrative record <br> Yes NHANES <br> $\mathrm{N}=123$ | Discordant <br> No administrative record <br> Yes NHANES |
| :--- | :--- | :--- |
|  | Percent | $\mathrm{N}=36$ |
| Characteristic | 45.5 | Percent |
| Gender: Male | 73.2 | 41.7 |
| Age (years): <20 | 44.7 | 38.9 |
| Race/ Ethnicity: Hispanic | 52.8 |  |

The group of 36 participants who reported household authorization for food stamps in NHANES but were not found in the Texas administrative database is of particular interest. In these cases, the respondent reported that the household was authorized to receive food stamp benefits in the past 12 months but no record was found for the survey participant in the Texas administrative database in the 12 month time frame. Some of these participants did match in the 48 months prior to the interview or in the 3 months after the interview (about $30 \%$ ). The other $70 \%$ did not match at any time point which may be because someone in the household, other than the NHANES participant, was authorized to receive the food stamp benefit.

## 4.Conclusion

This pilot project was successful in linking the NHANES data to a state level SNAP administrative database. Most of the NHANES Texas participants who reported household authorization for food stamp benefits in the last 12 months could be matched to the Texas database. For those who said "Yes" that someone in the household had authorization to receive food stamps, there was a $77 \%$ agreement with the Texas administrative database; for those who reported
no household authorization in the past 12 months, there was $92 \%$ agreement with the Texas administrative database. Given that most people in our sample did not report food stamp benefits the combined agreement was $89 \%$, close to that for the "No" responses. Possible explanations for discordances concern incorrect recall of when benefits were authorized and a household member other than the survey participant being authorized, and highlight areas that may benefit from improvements in questionnaire design.

There were some limitations to the pilot study. Relative to the overall sample design in NHANES, about 10,000 survey participants in each 2 year cycle, the sample size for the linkage evaluation was small, only 882 survey participants were eligible from the 2005-06 Texas NHANES sample. The linkage occurred at the survey participant level but the questionnaire asked about benefits at the household level. The results from this pilot are not generalizable for national or Texas level estimates.

In future work, we plan to assess the 2005-06 linkage, taking into account additional 2005-06 NHANES questions on food stamps. In addition, we plan to assess the 2007-08 questionnaire design change and its impact on the linkage with administrative data. This pilot study showed that linking NHANES participants to a state level administrative database is possible. However, as we plan for future linkages, we know that each state is unique and will present new challenges.

## Acknowledgements

We would like to thank our colleagues at the Ray Marshall Center at the University of Texas, in Austin, Daniel Schroeder, Patty Rodriguez and Deanna Schexnayder, for performing the linkage. We would also like to thank our colleagues at the United States Department of Agriculture, Economic Research Service, Michael Wiseman, John Kirlin, David Smallwood and Mark Denbaly, who were integral in the success of this linkage project.

## References

Mohadjer, L. and Curtin, L.R. (2008). Balancing Sample Design Goals for the National Health and Nutrition Examination Surve, Survey Methodology, Vol. 34, No. 1, pp. 119-126.

National Research Council. (2005). Improving Data to Analyze Food and Nutrition Policies. Panel on Enhancing the Data Infrastructure in Support of Food and Nutrition Programs, Research, and Decision Making. Committee on National Statistics, Division of Behavioural and Social Sciences and Education. Washington, DC: The National Academies Press.

SAS Institute Inc. 2010. SAS® 9.2 Language Reference: Dictionary, Third Edition. Cary, NC: SAS Institute Inc.
Taeuber, C., Resnick, D., Love, S., Staveley, J., Wilde, P. and Larson, R. (2004). Differences in Estimates of Food Stamp Program Participation between Surveys and Administrative Records, (Unpublished Paper). Washington, DC/Baltimore, MD: Bureau of the Census, Jacob France Institute, U.S. Department of Agriculture, and the Maryland Department of Human Resources. (http://www.ubalt.edu/jfi/jfi/reports/FdStmps_070903.pdf, accessed November, 2010

# Quality and quantity: Using administrative data for scientific purposes in labour market research 

Patrycja Scioch ${ }^{1}$


#### Abstract

Administrative data are gaining increasing attention in labour market research, in combination or as sole base data. Despite their advantages, such as large sample sizes, long time periods and extreme wealth of information, they suffer from such shortcomings as inconsistencies and missing values. Unfortunately, the research in quality and the improvement of the data is scarce. I focus on the education variable in widely used German administrative data, the Integrated Employment Biographies (IEB), which are very important for analyses of wage inequalities and opportunities for employment. The quality of this variable has decreased significantly over the last 10 years, with $35 \%$ individuals having no information on educational attainment. Some studies have tried to improve quality by developing imputation rules to fill the gaps in dependence on existing information without using additional sources. I use an additional data source: German patent data. The educational information in these data is very reliable and detailed. Using record linkage methods (via name, address) these data are linked to the IEB to verify the information. Therefore, the highly-educated are identifiable via their name or titles (professor, doctor or diploma). The educational information is compared to the original information and to the outcomes of the replicated imputation rules.


Multivariate models explain the mechanism of measurement errors of the original and the various imputed education variables. For the first time, the quality of an administrative data set will be quantified by highly reliable external data. To conclude the presentation, practical recommendations for a better quality in administrative data will be given.

[^108]SESSION 12 A
CLOSING SPEECH

# A Strategic vision for the use of administrative data at Statistics Canada 

Michel Cloutier ${ }^{1}$


#### Abstract

Over the past ten years and in particular in the last five years, Statistics Canada has made important progress in the greater use of Administrative data. These data are currently used in a very large number of economic and social programs. Administrative data currently in use includes tax records, customs data, health files and various registries, license databases of all kinds, education records, employment insurance records, address information and billing files, justice records and public accounts files, etc. More sources are being explored by various programs (for example credit card data). This paper presents a proposed vision to guide the future use of administrative data. The Vision paper starts with some background and definitions. It then outlines a possible strategic direction for Statistics Canada, including increased coordination, further research and possible efficiencies. Short-term recommendations are presented including the development of a policy for the use of administrative data, an inventory of administrative data, a research program and governance. Finally, some long-term options are discussed.


## Background

The Statistics Canada Quality Guidelines released in 2009 state the advantages of using administrative data and the business case to move further in that direction: "Administrative records present a number of advantages to a statistical agency and to analysts. Demands for statistics on all aspects of our lives, our society and our economy continue to grow. These demands often occur in a climate of tight budgetary constraints. Statistical agencies also share with many respondents a growing concern over the mounting burden of response to surveys. Respondents may also react negatively if they feel they have already provided similar information (e.g., revenue) to administrative programs and surveys. Administrative records, because they already exist, do not require the cost of direct data collection nor do they impose a further burden on respondents. It is important to note that the explosion of technology has also permitted statistical agencies to overcome the limitations caused by the processing of large datasets. For all these reasons, administrative records are becoming increasingly usable and are being used for statistical purposes. "

Over the past ten years and in particular in the last five years, Statistics Canada has made important progress in this direction. Administrative data are currently used in a very large number of economic and social programs. Administrative data currently in use include tax records, customs data, health files and various registries, license databases of all kinds, education records, employment insurance records, address information and billing files, justice records and public accounts files, etc. More sources are being explored by various programs (for example credit card data).

Statistical uses of administrative records include (i) use for survey frames, directly as the frame or to supplement an existing frame, (ii) replacement of data collection (e.g., use of taxation data for small businesses in lieu of seeking survey data for them), (iii) use in editing and imputation, (iv) direct tabulation, (v) indirect use in estimation (e.g., as auxiliary information in calibration estimation, benchmarking or calendarisation), and (vi) survey evaluation, including data confrontation (e.g., comparison of survey estimates with estimates from a related administrative program).

As the Chief Statistician notes in his paper A long-term Vision for Statistics Canada (2009), it is crucial that the Agency continue making progress towards increasing use of administrative data: "In the context of the kind of data

[^109]we collect, we need to keep making progress on using administrative data as a substitute for survey data where possible. The last ten years have seen a very successful effort to increase the utilisation of administrative records for statistical purposes, especially the use of taxation data for the production of business statistics. However, there is still much more potential that we need to exploit. The use of administrative data as a replacement for direct data collection reduces respondent burden and saves money. These savings create budget room that can be used to fill unmet high priority information needs and help enhance the relevance of the Agency's products and services."

As a follow-up to this Bureau-oriented vision statement, a group of senior managers was asked to take stock and develop a vision to guide future use of administrative data. The vision will be articulated around the three specific areas noted by the Chief Statistician in his Long Term Vision for Statistics Canada: quantity, quality, and efficiency.

## Scope of the Vision

Administrative data are defined as all data arising from the administration of an activity or program, covering a full population or a subset of that population. It includes individual transactions or aggregations that were originally collected for non statistical purposes.

This vision covers all types of administrative data and sources as well as all Statistics Canada programs including the economic, social and census programs. The vision establishes a framework to move the Agency to a more comprehensive and efficient use of administrative data within five years. The objective of the vision is to present a broad based corporate perspective on the use of administrative data, while still allowing specific business decisions to be made at the program level. It is also acknowledged that while using administrative data will lower collection costs, processing and analysis costs are likely to be higher especially in the development and implementation phases as well as in the initial years of production.

## Vision Statement

Over the next five years, Statistics Canada will evolve towards a statistical model where the strong preference by default will be given to using administrative data rather than surveys as the source of our statistical information. Two elements are part of the vision: a strategy organized around quality, efficiency and increased output of statistical information and an organization that will support this vision.

As part of this vision, programs would be required to identify, as part of their Quadrennial Program Review (QPR) or when proposing a redesign or a new development, the business case for not using administrative data.

## More specifically, under this vision, in five years we will have:

- increased coordination
- to identify and obtain new administrative data sources to replace or supplement existing data collection;
- to work with holders of administrative data to ensure cooperation and support in this undertaking;
- in the area of research
- put in place coordinated research programs to continually assess and address the quality and further use of administrative data;
- carried out research to address areas where data replacement by administrative data is more difficult (e.g., complex firms; programs with rigid conceptual requirements);
- realized some efficiencies by
- replacing some existing survey data collection (in part or in whole) in social and economic programs with administrative data;
- integrating into our statistical programs more administrative data that have not yet been tapped;
- increasing standardized approaches to the use of administrative data through clear responsibility for each data source to generate coherent statistical program data.
- used administrative data as a priority data source to fill data gaps via new programs or expansion in existing programs, thus avoiding additional costs to develop and collect new survey data
- improved key elements of quality assurance processes such as coverage, coherence analysis, editing, imputation and allocation through the greater use of administrative data and directly replaced poor respondent data with better quality administrative data where survey data are still used.
- increased the usefulness of administrative data by linking more files to survey data and other administrative files to create new information and provide new insights.


## Short-term recommendations

The agency will have to organize itself to facilitate and optimize the use of administrative data by addressing the following issues within the short-term:

## - Policy

- Currently Statistics Canada does not have an explicit policy for the use of administrative data. However there is an implicit practice which has been documented in the Statistics Canada Quality Guidelines as follows:
"It is Statistics Canada's policy to use administrative records whenever they present a costeffective alternative to direct data collection. As with any data acquisition program, consideration of the use of administrative records for statistical purposes is a matter of balancing the costs and benefits. Administrative records start with a huge advantage they avoid further data collection costs and respondent burden, provided the coverage and the conceptual framework of the administrative data are compatible with the target population. Depending on the use, it is often valuable to combine an administrative source with another source of information."
- It is proposed that a policy be fully developed, which formalizes this vision and the preference for using administrative data whenever it is cost effective and appropriate to meet user's needs.


## - Inventory and management responsibility

- Create an inventory and identify for each source the area responsible for that administrative data.
- Provide a single entry point and point of contact for each group of administrative data sets based on the source organization (as appropriate).
- Assign clear responsibility for the management, analysis, storage, distribution and promotion of each administrative data source
- Provide a communication mechanism to ensure that various programs are aware of how administrative data are being used in other areas and to promote a coherent approach to its use across programs (web site, discussion groups, meetings, etc.).


## - Research program

- Mobilize already existing research resources and assign them to priority research projects related to administrative data to ensure that existing sources are fully understood and mined for any potential benefits to the statistical system and to maintain quality.
- Allocate resources to research new administrative sources to
- fill priority data gaps once financing is available;
- to work with admin data collectors to increase usefulness of currently used admin data;
- to work with data collectors to eliminate duplicate collection effort.


## - Governance

- Given the large and varied nature of the administrative data files that the Bureau has and will continue to have under this initiative, there is a need for coordinated governance of the management, use and priority setting associated with administrative data. There are currently two models in use in the Bureau: that of Tax Data Division, which manages, promotes, analyzes and
distributes business tax administrative data, and that of other areas in which subject matter divisions manage the acquisition and use of administrative data. The suggested governance model would ensure that there is a centre of coordination and expertise identified for each administrative data source.


## Long-term Options

In the long-term, Statistics Canada could consider a range of options that are presented starting with the most aggressive strategies for increasing the use of administrative data and ending with the status quo.

- Statistics Canada could request changes in legislation both at a Government wide level and within the Statistics Act to make the use of administrative data mandatory wherever feasible or to increase the authority of Statistics Canada to influence content and concepts in government administrative data which would increase the usability of these sources for statistical purposes. As implemented in other countries, it could include provisions such as mandatory use of common identifiers in all government programs, mandatory use of standard statistical classifications in administrative data sets, and mandatory consultation with Statistics Canada before modifying administrative forms or addition of Statistics Canada's questions on regulatory questionnaires.
- Statistics Canada could reorganize itself around the use of administrative data as the primary source of data for most programs. This would include an important redevelopment of business processes and programs to maximize the use of administrative data. Examples of such models can be found in Portugal, New Zealand, Netherlands, Slovenia, Finland, etc.
- Statistics Canada could mandate that certain programs must use administrative data instead of or in addition to traditional surveys, assuming a positive cost benefit analysis.
- Statistics Canada could promote and facilitate the greater use of administrative data by funding research and communication and ensuring that all program redesigns fully consider this option.
- Statistics Canada could create a full inventory of all administrative data currently being used, as well as other known sources that could potentially be used to better inform program managers of the possibilities of these data sources.
- Statistics Canada could maintain its current model of gradual evolution without prescription.


## Conclusion

Under this vision, program managers would need to make decisions regarding the use of survey or administrative data based on a clear business case. Such a business case, including a cost benefit analysis, would be required to justify business decisions for not using administrative data. It is felt that having a framework and mechanisms to coordinate the use of administrative data and set priorities (possibility: Administrative Data Review Board) would help managers make the transition to the proposed model.

This vision will have a significant impact on the way Statistics Canada does business. The Corporate Business Architecture project should be used to facilitate the changes suggested in this document. In particular the creation of Data Service Centers and common and standardized repositories for micro data are key components required to implement this vision. This will also facilitate access and sharing of administrative data files as well as better information management practices.

In addition the idea that only one area should be responsible for any particular service is central to the Corporate Business Architecture project. This principle needs to be applied to the provision of all administrative data for the agency as it has already been done for tax data. It is likely that given the amount of data and the large number of subject matter areas involved we will need quite a few groups to fulfill this role, but there should be no overlap in their roles and responsibilities.

## References

Statistics Canada (2009). Statistics Canada Quality Guidelines, Statistics Canada Catalogue no.12539XWE, Ottawa, Canada.

Statistics Canada (2009). Long Term Vision Paper for Statistics Canada. Ottawa.


[^0]:    ${ }^{1}$ François Maranda, Statistics Canada; 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6.

[^1]:    ${ }^{1}$ Jelke Bethlehem, Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands (jbtm@cbs.nl)

[^2]:    ${ }^{1}$ Marc Christine, Institut National de la Statistique et des Études Économiques (INSEE), France ; Sébastien Faivre, Institut National de la Statistique et des Études Économiques (INSEE), France.

[^3]:    ${ }^{1}$ Roberta Vivio, Administrative archives data and statistical registers - Istat, Via Adolfo Ravà n.150, 00142 RomeItaly, (vivio@istat.it); par.1, 2, 4. Carla Runci, Administrative archives data and statistical registers - Istat, Via Adolfo Ravà n.150, 00142 Rome- Italy, (runci@istat.it); par. 3

[^4]:    ${ }^{2}$ A person not legally resident in Italy needs a tax code in several cases: to rent an apartment, to have temporary medical care, etc.

[^5]:    ${ }^{1}$ France Labrecque (France.Labrecque@statcan.gc.ca); Suzelle Giroux (Suzelle.Giroux @ statcan.gc.ca); Andrew Quigley (Andrew.Quigley@statcan.gc.ca), Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6.
    ${ }^{2}$ The age groups for estimation are different from the age groups assigned to the dwelling strata, which are " $3-11$ ", "12-19", "20-39", "40-59" and "60-79".

[^6]:    ${ }^{1}$ John L. Eltinge, Bureau of Labor Statistics, U.S.A., (Eltinge.John@bls.gov)

[^7]:    ${ }^{1}$ Stephen Jivraj, Centre for Census and Survey Research, School of Social Sciences, University of Manchester, Humanities Bridgeford Street, Manchester, M13 9PL, UK stephen.jivraj@postgrad.manchester.ac.uk

[^8]:    ${ }^{2}$ There were 104 health authorities in England and Wales in 2001.
    ${ }^{3}$ Output Areas were created using the result of the 2001 Census and contain, on average, 300 people
    ${ }^{4}$ There were 354 local authority districts in England in 2001.

[^9]:    ${ }^{1}$ Louise Morris, Office for National Statistics, Government Buildings, Cardiff Road, Newport NP10 8XG (louise.morris@ ons.gov.uk); Minda Phillips, Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hants PO15 5RR (minda.phillips@ons.gov.uk)

[^10]:    ${ }^{2}$ House of Commons Treasury Committee (2008), Counting the Population, Eleventh Report of Session 2007-08. London: The Stationery Office.
    ${ }^{3}$ Report by the Social Exclusion Unit (2001), A New Commitment to Neighbourhood Renewal, National Strategy Action Plan. London: Cabinet Office
    ${ }^{4}$ National Strategy for Neighbourhood Renewal (2000), Report of Policy Action Team 18: Better Information. London: The Stationery Office.

[^11]:    ${ }^{5}$ UK Statistics Authority (2009), Code of Practice for Official Statistics. UK Statistics Authority
    ${ }^{6} \mathrm{http}: / / w w w . n e i g h b o u r h o o d . s t a t i s t i c s . g o v . u k / d i s s e m i n a t i o n / ~$

[^12]:    ${ }^{2}$ These results are not reproduced here.

[^13]:    ${ }^{3}$ One Swiss franc (CHF) is approximately equal to one Canadian dollar.

[^14]:    ${ }^{1}$ Jannes Hartkamp, DESAN Research Solutions, The Netherlands, (hartkamp@desan.nl); Hans Rutjes, DESAN Research Solutions, The Netherlands, (rutjes@desan.nl)

[^15]:    ${ }^{1}$ Y. Celia Huang, Mary E. Thompson, and Christian Boudreau, Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1; Geoffrey T. Fong, Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1

[^16]:    ${ }^{1}$ Jacques Légaré, Université de Montréal, (jacques.legare@umontreal.ca); Yann Décarie, Institut national de la recherche scientifique ; Patrick Charbonneau, Université de Montréal ; Samuel Vézina, Mount Saint Vincent University ; Janice Keefe, Mount Saint Vincent University.

[^17]:    ${ }^{1}$ Jonathan A. Schwabish, Congressional Budget Office, Ford House Office Building, $2{ }^{\text {nd }}$ and D Streets, SW, Washington DC 20515, (jonathan.Schwabish@cbo.gov)
    ${ }^{2}$ For a more thorough overview of the CBOLT model, see Congressional Budget Office (2009). For the interested reader, the CBO website (www.cbo.gov) also contains a large number of background and technical documents about the CBOLT model.

[^18]:    ${ }^{3}$ For a more thorough description of these approaches, see CBO (2006c).

[^19]:    ${ }^{4}$ Individuals are grouped into birth cohort by decade of birth-for example, people born between 1940 and 1949 constitute one birth cohort.
    ${ }^{5}$ Those inputs include the fertility rate, morality improvements, net immigration, disability incidence and termination, total factor productivity growth, inflation, unemployment, gap between the marginal product of capital and real (inflation-adjusted)10-year bond rates, changes in the earnings share of compensation, and the gap between the consumer price index and the core price index.

[^20]:    ${ }^{1}$ Michael Wolfson, Statistics Canada, (michael.wolfson@statcan.gc.ca)

[^21]:    ${ }^{1}$ Mahlape Mohale, Statistics South Africa, (mahlapem@statssa.gov.za); Harebatho Moletsane, Statistics South Africa.

[^22]:    ${ }^{1}$ Leah B. Marshall and Amy B. O’Hara, Center for Administrative Records Research and Applications, U.S. Census Bureau, 4600 Silver Hill Rd, Washington DC 20233, USA (leah.b.marshall@census.gov, amy.b.ohara@census.gov).
    ${ }^{2}$ This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[^23]:    ${ }^{1}$ Martin Pantel, Statistics Canada, (Martin.Pantel@statcan.gc.ca)

[^24]:    ${ }^{1}$ Lori Stratychuk, Statistics Canada, Canada, KiA 0T6 (Lori.Stratychuk@statcan.gc.ca; Jean Dumais Statistics Canada, Canada, K1A 0T6 (Jean.Dumais@statcan.gc.ca)
    ${ }^{2}$ Detailed information about the OSPH is contained in Leenana et al. (2008).
    ${ }^{3}$ The titles given to each of the ethnic groups of interest are intended to be concise but informative; they do not completely reflect the true diversity of each of the ethnic groups

[^25]:    ${ }^{4}$ CSDs completely partition the province and are roughly equivalent to municipalities.

[^26]:    ${ }^{5}$ See Statistics Canada, 2003, p. 187 for details on organiziang interviews.

[^27]:    ${ }^{1}$ Chris Dibben, University of St Andrews, U.K. (cjld@st-andrews.ac.uk); Gemma Wright, University of Oxford, U,K.; Michael Noble, University of Oxford, U.K.

[^28]:    ${ }^{1}$ Rainer Schnell and Tobias Gramlich, Methodology Research Group, Faculty of Social Sciences, University of Duisburg-Essen, Lotharstrasse 65, D-47057 Duisburg, Germany (rainer.schnell@uni-due.de) and (tobias.gramlich@uni-due.de). Alexander Mosthaf and Stefan Bender, Research Data Centre of the German Federal Employment Agency at theInstitute for Employment Researh (IAB), Regensburger Strasse 104, D-90478 Nurember, Germany (alexander.mosthaf@iab.de) and (stefan.bender @iab.de)
    ${ }^{2}$ Due to legal constrainsts only administrative data and response status could be used for the analysis. With the exception of para data, no survey data has been used.

[^29]:    ${ }^{3}$ www.iab.de/en
    ${ }_{5}^{4}$ http://fez.iab.de/en.aspx
    ${ }^{5}$ See Jocabebbinghaues and Seth (2007 for a description of administrative data prepared and provided by the research data centre

[^30]:    ${ }^{6}$ Using the stat 'holdeck'ado by Adrian Mander and David Clayton, and the 'survwgt' ado by Nick Winter.

[^31]:    ${ }^{1}$ Transcendra McDonald, M.A., OIC National Operational Information Services (C/M), Operational Systems Services Section, Contract and Aboriginal Policing Directorate, Royal Canadian Mouned Police, 73 Leidin Drive, Ottawa ON K1A 0R2, (transcendra.mcdonald@rcmp-grc.gc.ca)

[^32]:    ${ }^{1}$ Anthony Matarazzo, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6
    (anthony.matarazzo@statcan.gc.ca)

[^33]:    ${ }^{1}$ Department of Sociology and Legal Studies, University of Waterloo, Waterloo, ON N2L 2B9 (pjc@uwaterloo.ca)

[^34]:    ${ }^{1}$ Stephanie Eckman, Institute for Employment Research, Germany, (steph.eckman@gmail.com); Frauke Kreuter, Institute for Employment Research, Germany.

[^35]:    ${ }^{1}$ Elisabeth Neusy, Statistics Canada, Ottawa, Canada, K1A 0T6 (Elisabeth.neusy@statcan.gc.ca); Yves Bélanger, Statistics Canada, Ottawa, Canada, K1A 0T6 (yves.belanger@statcan.gc.ca)

[^36]:    ${ }^{1}$ Paul Smith (paul.smith@ons.gov.uk), Karl Ashworth, Charles Lound, Office for National Statistics, Cardiff Road, Newport NP10 8XG, UK.

[^37]:    ${ }^{1}$ Jenna Fulton, University of Maryland, U.S.A., (jfulton@ survey.umd.ed)

[^38]:    ${ }^{1}$ Denis Hamel, statistician, 945, rue Wolfe, Québec, Canada, G1V 5B3; Robert Pampaloon, géographe, 945, rue Wolfe, Québec, Canada, G1V 5B3; Philippe Gamache, statistician, 945, rue Wolfe, Québec, Canada, G1V 5B3

[^39]:    ${ }^{1}$ Richard Madden, University of Sydney, Australia, (richard.madden@sydney.edu.au); Leonie Tickle, Macquarie University, (Australia); Lisa Jackson Pulver, University of NSW, Australia; Ian Ring, University of Wollongong, (Australia); Lee Taylor, NSW Department of Health, Australia

[^40]:    ${ }^{1}$ Paul A. Peters, Health Analysis Division, RHC-24M, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 (paul.a.peters@statcan.gc.ca); Michael Tjepkema, Health Analysis Division, RHC-24Q, 100 Tunney’s Pasture Driveway, Ottawa, ON, Canada, K1A 0T6 (michael.tjepkema@statcan.gc.ca

[^41]:    ${ }^{1}$ Ivan Fellegi, Statistics Canada, (Ivan.P.Fellegi@statcan.gc.ca)

[^42]:    ${ }^{1}$ Antoine Chevrette, Statistics Canada, (Antoine.Chevrette@statcan.gc.ca); Michael Wenzowski, Statistics Canada (Michael.Wenzowski@statcan.gc.ca
    ${ }^{2}$ Fellegi, Ivan; Sunter, Alan (December 1969). "A Theory for Record Linkage". Journal of the American Statistical Association 64 (328); pp. 1183-1210.

[^43]:    ${ }^{1}$ Istat, via Cesare Balbo 16, 00184 Roma, Italy. Nicolette Cibella (cibella@istat.it); Marco Fortini (fortini@istat.it); Daniela Ichim (ichim@istat.it); Tiziana Tuoto (tuoto@istat.it)

[^44]:    ${ }^{1}$ William E.Yancey, US Census Bureau, Washington, DC USA, (William.E.Yancey@census.gov)

[^45]:    ${ }^{1}$ Manuela Maia, Catholic University of Portugal, (mmaia@porto.ucp.pt); Paula Vicente, Lisbon University Institute, (Portugal), (paula.vicente@iscte.pt)

[^46]:    $\overline{{ }^{1} \text { Kirsten West, U.S. Census Bureau, Washington, DC 20233, U.S.A. (Kirsten.K.West@census.gov); }}$
    J. Gregory Robinson (J.Gregory.Robinson@census.gov); Jason Devine (Jason.E.Devine@census.gov); Renuka Bhaskar (Renuka.Bhaskar@census.gov).
    ${ }^{2}$ The component approach cannot be used for cohorts born before 1935 because the birth registration system in the United States did not include all states until 1933.

[^47]:    ${ }^{3}$ The term, "ROYA method" is used in this paper to represent the overall methodology currently used to estimate all components of net international migration, though some components are not estimated based on place of prior residence data.

[^48]:    ${ }^{4}$ For DA 2010, projections for 2008 to April 2010 are based on preliminary and provisional NCHS totals.

[^49]:    ${ }^{5}$ The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

[^50]:    ${ }^{1}$ Kate Wilder, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (Kate.Wilder@statcan.gc.ca); Steven Thomas, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (Steven.Thomas@ statcan.gc.ca)

[^51]:    ${ }^{1}$ Stefan Bender, Institute for Employment Research, Germany, (Stefan.bender@email.de); Jörg Heining, Institute for Employment Research, Germany, (Joerg.heining @iab.de)

[^52]:    ${ }^{1}$ Karen Daley, DROOD 4, Directorate Research - Organizational and Operational Dynamics, Director General Miliary Personnel Research and Analysis, Department of National Defence, 101 Colonel By Drive, Ottawa, Ontario, K1A 0K2, Canada.

[^53]:    ${ }^{1}$ Alan M. Zaslavsky, Harvard Medical School (USA), (zaslavsk@hcp.med.harvard.edu); Yulei He, Harvard Medical School (USA), (he@hcp.med.harvard.ed)

[^54]:    ${ }^{1}$ Division of Biostatistics, The Ohio State University, Columbus, Ohio, 43210, (randridge@cph.osu.edu)
    ${ }^{2}$ U.S. Bureau of Census and Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, 48109, (rlittle@umich.edu)

[^55]:    ${ }^{1}$ David Haziza, Université de Montréal (Canada), (haziza@DMS.UMontreal.ca); Pierre Duchesne, Université de Montréal (Canada), (duchesne@DMS.UMontreal.ca)

[^56]:    ${ }^{1}$ Martin Axelson, Statistics Sweden, 70189 Örebro, Sweden, (martin.axelson@scb.se); Dan Hedlin, Statistics Sweden, Box 24300, Stockholm, Sweden, (dan.hedlin@scb.se); Anders Holmberg, Statistics Sweden, 70189 Örebro, Sweden, (anders.holmberg@scb.se); Ingegerd Jansson, Statistics Sweden, Box 24300, Stockholm Sweden, (ingegerd.jansson@scb.se)

[^57]:    ${ }^{1}$ Christine Bycroft, Statistics New Zealand, PO Box 4741 Christchurch, New Zealand.
    (Christine.bycroft@stats.govt.nz)

[^58]:    ${ }^{2}$ Information about LEED, Statistics NZ, www.stats.govt.nz

[^59]:    ${ }^{1}$ Danilo Dolenc, Statistical Office of the Republic of Slovenia, Vozarski pot 12, Slovenia, 1000 Ljubljana (danilo.dolenc@gov.si)

[^60]:    ${ }^{2}$ Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July on Community Statistics on Migration and International Protection and Repealing Council Regulation (EEC) No 311/76 on the Compilation of Statistics on Foreign Workers, L 199/23, 31 July 2007

[^61]:    ${ }^{1}$ Andrew Heisz, Income Statistics Division, 170 Tunney's Pasture Driveway, K1A 0T6, Ottawa, Canada (Andrew.Heisz@statcan.gc.ca); Manon Langevin, Income Statistics Division, 170 Tunney's Pasture Driveway, K1A 0T6, Ottawa, Canada (Manon.Langevin@statcan.gc.ca); Jeffrey Randle, Income Statistics Division, 170 Tunney's Pasture Driveway, K1A 0T6, Ottawa, Canada (Jeff.Randle@statcan.gc.ca).

[^62]:    ${ }^{2}$ The Pension Plans in Canada Survey is a complete annual survey of registered pension plans in Canada. It includes information on the various terms and conditions of these plans, membership and contributions.

[^63]:    ${ }^{3}$ It is difficult to assess the longitudinal bias introduced by restrospective linkage of immigrants because there are several possible reasons for the failure of their linkage. Because they are likely to be assisted by a family member upon their arrival in Canada or to leave the country several times after their immigration, it is difficult to know if their data are missing because they did not file an income tax return during these years (no introduction of bias) or because of achange in SIN over time.

[^64]:    ${ }^{4}$ At the time of this analysis, information from T4 files was not available for years prior to 2001.

[^65]:    ${ }^{5}$ Employment earnings expressed in 2007 constant dollars.
    ${ }^{6}$ In the literature, this growth is often associated with the growth in returns to education.

[^66]:    ${ }^{1}$ Melissa Riley Pfeiffer, NYC Dept of Health and Mental Hygiene, (mpfeiffer@health.nyc.gov);
    M. Mavinkurve, M.E. Slopen. S.Sedlar, A.E. Curry, J. Ho, K.H. McVeigh

[^67]:    ${ }^{1}$ Jannes Hartkamp, DESAN Research Solutions, PO Box 10288, 1001 EG Amsterdam, The Netherlands (hartkamp@desan.nl).
    ${ }^{2}$ For a discussion of the concept, definitions and measurement issues in an international context, see European Commission Directorate General responsible for Education and Culture (2005).

[^68]:    ${ }^{3}$ Source: http://statline.cbs.nl.
    ${ }^{4}$ Excluding company courses.
    ${ }^{5}$ ISCED-3 in the Dutch system includes the types of education known as havo (5-year general upper secondary education, as a rule starting at age 12 and ending at age 17), vwo (6-year pre-academic general upper secondary education, age 12-18) and the medium and higher levels of $m b o$ (vocational upper-secondary education and training, varying from 2 to 4 years, starting around age 16 or 17, after vmbo (4-year lower secondary education, age 12-16) or havo.

[^69]:    ${ }^{6}$ See table 'Early leavers from education and training' published at http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table\&language=en\&pcode=tsisc060.
    ${ }^{7}$ As of 2004 some changes have been made in operational definitions, which is why two percentages are shown for that year, according to the old $(23,9 \%)$ and the new definitions $(24,9 \%)$. Those who read Dutch can find the details in Hartkamp (2009), available via www.jos.rotterdam.nl.

[^70]:    ${ }^{1}$ Changbao Wu, University of Waterloo, Canada, (cbu@waterloo.ca); Jae-kwang Kim, Iowa State University, USA, (jkim@iastate.edu)

[^71]:    ${ }^{1}$ Yves G. Berger, University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, UK (E-mail: y.g.berger@soton.ac.uk)
    ${ }^{2}$ Rodolphe Priam, University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, UK (E-mail: R.Priam@soton.ac.uk)

[^72]:    ${ }^{1}$ Takis Merkouris, Athens University of Economics and Business, Patision 76, Athens, Greece, 10434

[^73]:    ${ }^{1}$ Tom Haymes, Statistics Canada, Ottawa, Canada, K1A 0T6 (tom.haymes@statcan.gc.ca); Jean-Luc Bernier, Statistics Canada, Ottawa, Canada, K1A 0T6

[^74]:    ${ }^{1}$ Éric Langlet, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, Canada, K1A 0T6, (eric.langlet@statcan.gc.ca)

[^75]:    ${ }^{1}$ Pamela Tallon, Statistics Canada, (patricia.tallon@statcan.gc.ca); Linda Ramsey, Statistics Canada, (linda.ramsey@statcan.gc.ca)

[^76]:    ${ }^{1}$ Jennifer Taylor, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6
    ${ }^{2}$ The remaining dwellings received a questionnaire hand delivered by a census enumerator or were enumerated by personal interview.

[^77]:    ${ }^{4}$ There were a few returns through the telephone help line, amounting to an additional $0.24 \%$ return rate at the most.

[^78]:    ${ }^{1}$ Tobias Bachteler (tobias.bachteler@uni-due.de); Rainer Schnell (rainer.schnell@uni-due.de); Jörg Reiher (joerg.reiher@uni-due.de), University of Duisburg-Essen, Lotharstr, 65, D-47057 Duisburg, Germany

[^79]:    ${ }^{1}$ Antoine Chevrette, Statistics Canada (Antoine.chevrett@statcan.gc.ca)
    ${ }^{2}$ Fellegi, I., and Sunter, A. (December 1969). A Theory for Record Linkage, Journal of the American Statistical Association 64 (328): pp. 1183-1210.

[^80]:    ${ }^{3}$ Paper 011-2008, SAS® Integration Technologies, UNIX and Visual Basic .Net Integration Procedure
    Antoine Chevrette, Statistics Canada, Ottawa, ON

[^81]:    ${ }^{1}$ Rainer Schnell (rainer.schnell@uni-due.de); Tobias Bachteler (tobias.bachteler@uni-due.de); Jörg Reiher (joerg.reiher@uni-due.de), University of Duisburg-Essen, Lotharstr. 65, D-47057 Duisburg, Germany

[^82]:    ${ }^{2}$ For academic use, the record linkage program Merge ToolBox (MTB) is freely available from http://www.recordlinkage.de/. For further information see Schnell et al (2004)

[^83]:    ${ }^{1}$ J. N.K. Rao, Carleton University, Ottawa (Canada), jrao@math.carleton.ca

[^84]:    ${ }^{1}$ William R. Bell, U.S. Census Bureau, Washington D.C. USA, (William.R.Bell@census.gov); Elizabeth T. Huang, U.S. Census Bureau, Washington D.C. USA, (Elizabeth.T.Huang@census.gov)

[^85]:    ${ }^{1}$ Stefan Bender, Institute for Employment Research, Germany, (Stefan.bender@email.de); Jörg Heining, Institute for Employment Research, Germany, (Joerg.Heining@iab.de); Tanja Hethey, Institute for Employment Research, Germany; Patrycja Scioch, Institute for Employment Research, Germany

[^86]:    ${ }^{1}$ Karen Guerts, HIVA-K.U.Leuven, Parkstraat 47 Leuven, Belgium, 3000 (karen.geurts@hiva.kuleuven.be); Monique Ramioul, HIVA-K.U.Leuven, Parkstraat 47 Leuven, Belgium, 3000 (monique.ramioul@hiva.kuleuven.be); Peter Vets, National Social Security Office, Victor Hortaplein 11 Brussels, Belgium, 1060
    (peter.vets@onssrszlss.fgov.be)

[^87]:    ${ }^{1}$ Author \#1 Daniela Hochfellner, Institute for Employment Research, Regensburgerstr.104, Germany, 90478 Nuremberg (daniela.hochfellner@iab.de); Author \#2 Axel Voigt, Institute for Employment Research, Regensburgerstr.104, Gemany, 90478 (axel.voigt@iab.de)

[^88]:    ${ }^{1}$ Martina Huber, Institute for Employment Research, Germany; Alexandra Schmucker, Institute for Employment Research, Germany; Stefan Bender, Institute for Employment Research, Germany, (Stefan.Bender@iab.de)

[^89]:    ${ }^{1}$ Asaph Young Chun, University of Chicago, USA, (Chun-Young @ norc.org); Fritz Scheuren, University of Chicago, USA, (Scheuren@norc.org

[^90]:    ${ }^{1}$ Jae-kwang Kim, Iowa State University, U.S.A., (jkim@iastat.edu); Wayne Fuller, Iowa State University, U.S.A.

[^91]:    ${ }^{1}$ Department of Mathematics and Statistics, College of Science and Mahematics
    ${ }^{2}$ Coles College of Business, Kennesaw State University, Kennesaw, GA, USA, (mlawso16@kennesaw.edu)

[^92]:    ${ }^{1}$ Consultant analyst with the Centre d'accès aux données de recherche de l'Institut de la statistique du Québec (CADRISQ) and pilot with the Services d'accès aux données à des fins de recherche, Institut de la Statistique du Québec, Direction des Services informationnels et technologiques, 3535 Queen Mary, Room 420-4, Montréal, QC, Canada, H3V 1H8, (annie.belanger@stat.gouv.qc.ca
    ${ }^{2}$ Expert advisor on research processes, Services d'accès aux données à fins de recherche Institut de la Statistique du Québec, Direction des services informationnels et technologiques, 200 Chemin Ste-Foy, 9th floor, Québec City, QC, Canada, G1R 5T4, (carol.gilbert@stat.gouv.qc.ca)

[^93]:    ${ }^{3}$ Ethnicity includes language and culture.

[^94]:    ${ }^{1}$ David Price, Statistics Canada, (David.Price@statcan.gc.ca); Martin Lessard, Statistics Canada, (Martin.Lessard@statcan.gc.ca)

[^95]:    ${ }^{1}$ Peter Meyer, National Center for Health Statistics, Hyattsville MD, USA (http://www.cdc.gov/nchs/r\&d/rdc.htm)

[^96]:    ${ }^{1}$ Bradley Brooks, 150 Tunney’s Pasture Driveway, Ottawa, ON K1A 0T6 (Bradley.Brooks@statcan.gc.ca); Michael Haan, University of New Brunswick, 135A Carleton Hall, Fredericton, New Brunswick, E3B 5A6 (mhaan@unb.ca); July Lee, Office of Statistics and Information, Government of Alberta. (Judy.lee@gov.ab.ca)

[^97]:    ${ }^{1}$ Institute of Fundamental Sciences - Statistics, Massey University, New Zealand
    ${ }^{2}$ Department of Geography, University of Sheffield, England

[^98]:    ${ }^{1}$ Jan van den Brakel, Statistics Netherlands, department of Statistical Methods and Maastricht University School of Business and Economics, department of Quantitative Economics, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands (e-mail: JBRL@CBS.NL); Sabine Krieg, Statistics Netherlands, department of Statistical Methods, P.O. Box 4481, 6401 CZ Heerlen, The Netherlands (e-mail: SKRG@CBS.NL)

[^99]:    ${ }^{1}$ Marissa Isidro, PhD Student, Institute of Fundamental Sciences, Massey University, Palmeston North, New Zealand (maris_isidro@yahoo.com); Stephen Haslett, Professor (Statistics), Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealsnd (s.j.haslett@massey.ac.nz); Geoffrey Jones, Associate Professor (Statistics), Institute of Fundamental Sciences, Massey University, Palmeston North, New Zealand (g.jones@ massey.ac.nz)

[^100]:    ${ }^{1}$ Aline Drapeau - Département de psychiatrie de l'université de Montrèal et Centre de recherche Fernand-Seguin, 7331 rue Hochelaga, Montréal (Québec) H1N 3V2, (adrapeau.hlhl@ ssss.gouv.qc.ca);
    Richard Boyer - Département de psychiatrie de l'université de Montréal et Centre de recherche Fernand-Seguin, 7331 rue Hochelaga, Montréal (Québec) H1N 3V2, (richard.boyer@umontreal.ca);
    Fatouma Binta Diallo - Centre de recherche Fernand-Seguin et Département de médecine sociale et préventive de l'université de Montréal, C.P. 6138 succ. Centre-Ville, Montréal (Québec), (fatoumata.binta.diallo.2@umontreal.ca)

[^101]:    ${ }^{1}$ Amanda Halladay, Statistics Canada (Amanda.Halladay@statcan.gc.ca); Maire Sinha, Statistics Canada (Maire.Sinha@statcan.gc.ca); Kathy AuCoin, Statistics Canada (Kathy.AuCoin@statcan.gc.ca); Sarah Franklin, Statistics Canada (Sarah.Franklin@statcan.gc.ca); 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6.

[^102]:    ${ }^{1}$ Anugula N. Reddy, National University of Educational Planning and Administration, India, anreddy@nuepa.org

[^103]:    ${ }^{1}$ Joseph Sakshaug, Institute for Social Research, University of Michigan, 426 Thompson Street, Room 4050, Ann Arbor, MI, USA, 48104 (joesaks@umich.edu); David Weir, Institute for Social Research, University of Michigan, 426 Thompson Street, Room 3053, Ann Arbor, MI, USA, 48104 (dweir@umich.edu); Lauren Hersch Nicholas, Institute for Social Research, Universityof Michigan, 426 Thompson Street, Room 3005, Ann Arbor, MI, USA, 48104 (lnichola@umich.edu)

[^104]:    ${ }^{1}$ Denis Malo, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, (Denis.Malo@statcan.gc.ca)

[^105]:    ${ }^{2}$ National Institute of Demographic Studies, Paris, France (riandey@ined.fr)

[^106]:    ${ }^{1}$ Alfredo Navarro, U.S. Census Bureau, Washington, DC, USA, (alfredo.navarro@census.gov)

[^107]:    ${ }^{1}$ Lisa B. Mirel, Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyattsville Maryland, USA, 20782 (L.Mirel@cdc.gov); Vicki Burt, Centers For Disease Control and Prevention, National Center for Health Statitics, 3311 Toledo Road, Hyattsville Maryland, USA, 20782; Donna Miller, Centers for Disease Control and Prevention, National Center for Health Statistics, 3311 Toledo Road, Hyatsville Maryland, USA, 20782; Christine Cox, Centers for Medicare and Medicaid Services, 7500 Security Boulevard, Baltimore Maryland, USA, 21244. Christine Cox conducted this research while employed at the National Center for Health Statistics prior to becoming an employee at the Centers for Medicare and Medicaid Services.

[^108]:    ${ }^{1}$ Patrycja Scioch, IAB - Institute for Employment Research, Germany, (patrycja.scioch @iab.de)

[^109]:    ${ }^{1}$ Michel Cloutier, Statistics Canada (Michel.Cloutier@statcan.gc.ca); 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6.

