



Strategies for Standardization of Methods and Tools - How to get there

Proceedings

**Statistics Canada's International
Methodological Symposium**

November 1-4, 2011



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 2011 — Catalogue no. 11-522-XCB

End-use licence agreement

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2012

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

STRATEGIES FOR STANDARDIZATION OF METHODS AND TOOLS – HOW TO GET THERE

TABLE OF CONTENTS

PREFACE	6
KEYNOTE ADDRESS	
01A-1 Standardising methods and tools: A look back over 30 plus years of experience	8
S. Linacre, Australian Bureau of Statistics, Australia	
SESSION 2A: CORPORATE BUSINESS ARCHITECTURE: HOUSEHOLD SURVEY FRAMES	
02A-1 Redesign at Statistics Netherlands	22
F. Hofman, Statistics Netherlands, The Netherlands	
02A-2 Progress towards standardisation at Statistics New Zealand	30
J. Lopdell and G. Dunnet, Statistics New Zealand, New Zealand	
02A-3 Development of a common frame for household surveys at Statistics Canada	31
L. MacNabb, M. St-Pierre and M. Grenier, Statistics Canada	
SESSION 02B: SAMPLING AND ESTIMATION	
02B-1 Differential design effects in school-based samples sources	39
C. Dahmen and M. Fuchs, Darmstadt University of Technology, Germany	
02B-2 Canadian Forces Personnel Surveys: A consideration of weighting and non-response	46
F. Larochelle, T. Gou and I. Goldenberg, Dept. of National Defence and the Canadian Military, Canada	
02B-3 Building an integrated sampling framework for Business Surveys: Simulation studies to evaluate the efficiency of a two-phase sample design	52
Y. Li and F. Picard, Statistics Canada	
02B-4 Sparse and efficient replication variance estimation for complex surveys	58
J. K. Kim, Iowa State University, USA and C. Wu, University of Waterloo, Canada	
SESSION 03A: HARMONIZATION OF METHODS AS PART OF LARGE-SCALE STANDARDIZATION PROJECTS FOR BUSINESS SURVEYS	
03A-1 Harmonizing methodologies through a system integration project: Challenges and lessons learned	60
J. Andrews, F. Brisebois, I. Delahousse, C. Dochitoui, M. Lachance, R. Philips and S. Pursey, Statistics Canada	
03A-2 The multiple facets of the redesign of IT tools for producing short-term statistics at INSEE: The Premice program	66
F. Guggemos, National Institute of Statistics and Economics Studies, France	
03A-3 Standardizing UK sub-annual Business Surveys	73
S. Merad and P. Brodie, Office for National Statistics, UK	

SESSION 03B: DISSEMINATION AND DATA ACCESS

03B-1 Eurostat project SICON: Secure Infrastructure for confidential data access and sharing81
D. Buono, Eurostat, Luxembourg

03B-2 Enhanced table production system overview and technical description: Standardizing the
production of custom tables for data quality and efficiency82
P. Timusk, M. Mansour, É. Pelletier and E. Turgeon, Statistics Canada

SESSION 04A: SELECTIVE EDITING

04A-1 Selective data editing and its implementation at Statistics Sweden90
P. Brundell, Statistics Sweden, Sweden

04A-2 SeleMix: An R package for selective editing via contamination models97
M. Di Zio and U. Guarnera, Istat, Italy

04A-3 Selective editing methods and tools: An Australian Bureau of Statistics perspective105
E. Brinkley, K. Farwell and F. Yu, Australian Bureau of Statistics, Australia

SESSION 04B: CONFIDENTIALITY

04B-1 G-Confid: Statistics Canada’s confidentiality software114
C. Rondeau and J.-M. Fillion, Statistics Canada

04B-2 Assessing disclosure risk in perturbed microdata120
N Shlomo, University of Southampton, UK

04B-3 Privacy preserving probabilistic record linkage (P3RL) from A to Z: a GRLS example
from SwissLinkage128
Spoerri, K. Schmidlin, University of Bern, Switzerland, R. Schnell, University Duisburg-Essen, Germany
and K. Clough-Gorr, University of Bern and National Institute of Cancer Epidemiology and Registration,
Switzerland

04B-4 Standardised outputs and regionally-based classifications for minority populations as part
of the England and Wales 2011 Census129
J. Traynor and E. White, Office for National Statistics, England

SESSION 05A: WAKSBERG AWARD WINNER ADDRESS

05A-1 Modelling of complex survey data: Why model? Why is it a problem?
How can we approach it?131
D. Pfeffermann, Hebrew University of Jerusalem, Israel, University of Southampton, UK

SESSION 06A: STANDARDS AND GUIDELINES FOR THE DESIGN AND TESTING OF INTERNET QUESTIONNAIRES

06A-1 Developing electronic questionnaire guidelines: Issues and challenges in a changing environment ...133
A.-M. Côté, D. Lawrence and P. Kelly, Statistics Canada

06A-2 GINO++: A generalized system for web surveys139
R. Torelli, National Institute of Statistics (ISTAT), Italy

06A-3 Embedded experiment for non-response follow-up methods of electronic questionnaire
collection147
M. Karaganis, K. Fox, J. Claveau, J. Leung and W. Lin, Statistics Canada

SESSION 06B: OUTLIERS AND IMPUTATION

06B-1 Retrofitting a simpler outlier detection procedure into a complex generalized system156
L.T Bechtel, U.S. Census Bureau, USA

06B-2 Outlier detection tool at Statistics Canada162
N. Émond, Statistics Canada

06B-3 An assessment of methods to impute risk exposure into model actor’s risk profile
for microsimulation169
D. Hennessy, Ottawa Hospital Research Institute and Statistics Canada, C. Bennett, M. Tuna, Hospital Research
Institute, C. Nadeau, W. Flanagan, Statistics Canada and D. Manuel, Ottawa Hospital Research Institute

SESSION 07A: CONFIDENTIALITY METHODS AND TOOLS FOR ACCESSING DATA WHILE PRESERVING CONFIDENTIALITY

07A-1 De-identification methods for public use health files177
K.E. Emam, University of Ottawa, Canada

07A-2 The U.S. Census Bureau’s microdata analysis system178
M. Freiman, U.S. Census Bureau, J. Lucero, Freddie Mac, L. Singh, Georgetown, J. You, University of
California, M. DePersio, University of Delaware and L. Zayatz, U.S. Census Bureau, USA

07A-3 Providing access to microdata for statistical purposes: Experiences of the Australian Bureau
of Statistics with remote analysis servers187
J.O. Chipperfield, F. Yu and M. Gare, Australian Bureau of Statistics, Australia

SESSION 07B: CONTENT AND COLLECTION

07B-1 Some implications of standardizing methods for quality monitoring of survey interviewing196
D. Currvan, D. Stone, K. Fuller, S. Kinsey and H. Speizer, RTI International, USA

07B-2 European Health Examination Survey: From a sampling and recruitment perspective203
J. Heldal , S. Jentoft, Statistics Norway, (Norway), K. Kuulasmaa, P. Koponen and S. Ahonen, National
Institute for Health and Welfare, Finland

07B-3 Preliminary collection planning Collection Front Door210
A. Marcil, Statistics Canada

07B-4 Implementing quality control procedures at NASS’s National Operations Center215
J.M. Boone, J.L. Parsons, S.R. Feld, J.N. Levy and K.L. Flaherty, USDA National Agricultural Statistics Service,
USA

SESSION 08A: USING STANDARDIZED METHODS AND TOOLS FOR POST-COLLECTION PROCESSING

08A-1 Standardization of post-collection processing in Business Surveys at Statistics Canada.....222
S. Godbout, Statistics Canada

08A-2 Standardization of processes230
F. Hofman, A. Camstra and R.Renssen, Statistics Netherlands, The Netherlands

08A-3 Coding of survey responses – quality assurance efforts and IT tools at Statistics Sweden239
J. Svensson, Statistics Sweden, Sweden

SESSION 08B: QUESTIONNAIRE DESIGN AND COLLECTION MODE EFFECTS

08B-1 Survey data quality provisions in Statistics Canada E-Questionnaire solution: Retrospective and perspectives	244
Y. Abiza, Statistics Canada	
08B-2 Designing a questionnaire to examine the Canadian Forces sports program	245
K.K. Hachey, Department of National Defence and the Canadian Military, Canada	
08B-3 Harmonized content: The new paradigm in developing surveys at Statistics Canada	250
R. Nadwodny and P. Best, Statistics Canada	
08B-4 Calibrating mode effects in the Dutch crime survey	256
B. Buelens and J. van den Brakel, Statistics Netherlands, The Netherlands	

SESSION 09A: STANDARDIZATION IN INTERNATIONAL COMPARATIVE STUDIES: BENEFITS AND CHALLENGES

09A-1 Designing, standardizing and monitoring survey operations in international large-scale educational research	258
R. Carstens, IEA Data Processing and Research Center, Germany	
09A-2 Summarizing item responses in large scale assessment	267
E. Gonzalez and M. Von Davier, Educational Testing Service, USA	
09A-3 Standardization of sampling plans and quality assurance in comparative surveys	268
M. Joncas and S. LaRoche, Statistics Canada	

SESSION 09B: STANDARDIZED SOFTWARE

09B-1 Harmonisation of seasonal adjustment practices through the development of the DEMETRA+ software	275
J. Palate, National Bank of Belgium and P. Jacques, Eurostat, Luxembourg	
09B-2 How does CANCEIS work & can it benefit more users?	276
C. W. Liu, S. Crowe and A. Alavi, Statistics Canada	
09B-3 Development of the social survey processing environment	282
L. MacNabb, Statistics Canada	
09B-4 Implementing methodology changes or enhancements in a standard processing system or How can an Advisory Group help facilitate change	287
K.J. Thompson, U.S. Census Bureau, USA	

SESSION 10A: FRAMEWORK

10A-1 Developing a methodology for the Canadian Framework for Culture Statistics	295
M.K. Allen, Statistics Canada	
10A-2 How many Canadians live in a city? Conceptualization, definition and proposed dissemination for alternative standards	296
R.D. Bollman and P. Murphy, Statistics Canada	
10A-3 The role of data quality standards in the standardization of survey methods and tools	303
J.L. Eltinge, Bureau of Labor Statistics, U.S.A.	

10A-4 Enterprise architecture work at Statistics Sweden	304
M. Axelson, J. Engdahl, Y. Fossan, E. Holm, I. Jansson, B. Lorenc and L.G. Lundell, Statistics Sweden, Sweden	

SESSION 10B: CALENDAR EFFECTS AND TEMPORAL COHERENCE

10B-1 Benchmarking and forecasting: A top-down approach for combining forecasts at multiple frequencies	312
M.A. Trovero, E. Blair and M.J. Leonard, SAS Institute, USA	
10B-2 Improving calendarization using X-12-ARIMA: Application to GST data	320
R. Manríquez, Statistics Canada	
10B-3 The error in business cycle estimates obtained from seasonally adjusted data	328
T. McElroy, U.S. Census Bureau, USA	

SESSION 11A: BUILDING AND USING GENERALIZED SYSTEMS

11A-1 Generalized systems: The Statistics Canada experience	330
Y. Deguire, L. Reedman and M. Wenzowski, Statistics Canada	
11A-2 Triton: A general tool for data collection and micro editing	337
J. Erikson, Statistics Sweden, Sweden	
11A-3 Statistics New Zealand's standard methodology toolbox	343
J. Lopdell and G. Dunnet, Statistics New Zealand, New Zealand	

SESSION 11B: MODELING AND ESTIMATION

11B-1 Model-based and semi-parametric estimation of time series components and mean square error of estimators	345
M. Sverchkov, R. Tiller, Bureau of Labor Statistics, USA and D. Pfeffermann, Hebrew University of Jerusalem, Israel and University of Southampton, UK	
11B-2 Challenges and issues in weighting the Travel Survey of Residents of Canada.....	346
F. Labrecque-Synnott, Statistics Canada	
11B-3 Statistical methods for evaluating secular trends using estimates from annual independent cross-sectional complex probability sample surveys	351
P.J. Smith and Z. Zhao, Centers for Disease Control and Prevention, USA	

Preface

Symposium 2011 was the twenty-seventh in Statistics Canada's series of international symposia on methodological issues. Each year the symposium focuses on a particular theme. In 2011, the theme was: “**Strategies for Standardization of Methods and Tools – How to get there**”.

Symposium 2011 was held from November 1-4, 2011, at the Ottawa Convention Centre in Ottawa, Ontario, and it attracted more than 400 people from countries around the world. Three workshops and 61 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 2011 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 thank all the participants that went to the different presentations.

The Symposium 2011 Organizing Committee

Colin Babyak, President

Logistics and operations

Lyne Guertin, Chair

Jack Singleton

Lori Stratychuk

Program

Sarah Franklin, Chair

José Gaudet

Richard Laroche

KEYNOTE ADDRESS

Standardising methods and tools: A look back over 30 plus years of experience

Susan Linacre¹

Abstract

National Statistical Agencies operate in an industry that has been continually expanding over recent decades. The demand for information continues to grow in an increasingly competitive and cost conscious market. At the same time technological developments have been moving rapidly to open up a continuing range of new opportunities to revolutionise the way we work.

One has been to seek efficiency and quality gains, doing more, and more complex things with less, through standardised methods and tools. The search for standardisation has been underway for as long as I have been working in statistics, and with continuing technological innovation, and associated improvements in information resources, new cycles of standardisation can be expected to continue into the future. But what do we learn from each cycle about how to do it better in the next?

This presentation gives my view of why we work so hard to standardise our methods and tools, and where and why the search has proved relatively successful, and where less so.

It builds on my experience within the Australian Bureau of Statistics, starting with a generalised system developed for business surveys in the early 1970s, and passing through several subsequent cycles of standardisation and modernisation. It also builds on some personal experience at the Office for National Statistics in the UK.

I am not a professional expert in systems development. The presentation provides my view of what works and doesn't, both as a seasoned methodologist, seeking to develop standardised methods, and to implement standardised methods and tools, and also as a subject statistician, seeking to gain the many benefits of the standardised approach.

1. Introduction

Statistical Infrastructure is core to the functioning of an official statistical agency. Such infrastructure includes reusable conceptual components such as classifications, and standards for defining items, question modules with associated editing modules and derivation sets, as well as sample design and estimation methods. It also includes the various tools and systems used to support these, including metadata systems to describe them. It includes, for example, coders, selection and estimation modules, editing tools and so on.

I have worked as a methodologist for over three decades in two countries (albeit, in one for only a couple of years) and as a senior subject matter statistician for a further six years. This paper presents some of my personal views, based on this experience. It covers my thoughts on the things it is important to consider in standardising infrastructure for use across a statistical organisation.

I start by giving a brief and selective description of the building of statistical infrastructure in the Australian Bureau of Statistics (ABS) over a number of decades. It is selective for reasons of space, and I have focussed on the business statistics side, as the area where I have the longest methodological history to draw on. Also I focus on the ABS, as I was not at the Office for National Statistics (ONS) for a long enough period to obtain a proper perspective on the issues and outcomes. It is from this sort of history that we need to draw lessons in embarking on new standardisation projects with the aim of improving the efficiency and effectiveness of our statistical production.

Following this history, I go on to provide my thoughts on why we want to standardise methods and tools, and when we might not want to. I also discuss some of the key decisions we face when we are doing it. I conclude by giving

¹Susan Linacre, Australian Bureau of Statistics, Australia, susan.linacre@abs.gov.au.

some thoughts on what seems to work well in undertaking major programs of standardisation, as well as some pitfalls to avoid, based on past experience.

But first it is useful to think about what a well performing statistical agency might look like.

Some characteristics of such an organisation would be:

- the reliable production of a strong backbone of robust, high integrity statistics, flowing from a well designed, well documented and well supported set of processes;
- resources and high end capability of the organisation focussed on developing new sets of statistics to bring into this backbone, or to use in conjunction with it; or on strengthening and amending its components on a controlled basis;
- The capacity to flexibly and efficiently meet new ad hoc requirements, either through new collections or the combination of data from available sources.

To achieve the above outcomes, an organisation would need to achieve a high level of operational excellence in its core work, freeing capacity to improve, innovate and remain relevant. Such operational excellence would have the additional benefits of providing a strong sense of job satisfaction for staff who could use their creativity to develop and build meaningful outputs rather than recovering from mis-specified derivations and corrupted files.

The main rationale for appropriate standardisation is that it is a key enabler of operational excellence, and hence of the rigour, agility and relevance that is required of national statistical offices. It enables operational excellence, because it supports repeating processes with well designed, efficient methods and tools. Done well, it delivers a high quality low cost approach for multiple users.

2. Building statistical infrastructure at the ABS over the decades

In the late 1960s and early 1970s, the ABS, was establishing a single approach, based on an area frame, for a program of household surveys. At the same time it recognised that common approaches also applied across business surveys. A system to support this common approach, called the General Survey System (GSS) was developed. This aimed to provide an end to end system for business surveys in the ABS, from design through selection, despatch, collection, editing, estimation (including standard errors) and table generation. Those surveys that used the GSS were known collectively as Business Surveys. They consisted of 3 quarterly surveys: stocks, capital expenditure and profits. Numerous other business surveys run in the ABS, chose not to use the GSS.

The concept of the GSS was powerful and it provided an excellent learning platform, however GSS did not provide the flexibility that survey owners felt necessary to allow them to meet their diverse needs in terms of estimation, or treatment of business births and deaths etc. Once you had opted to use GSS, you entered a tunnel with no alternative strategies available til you got out with tables at the other end. The experience demonstrated the need to provide the ability to adapt to new methods and approaches, as well as uses of new sources of auxiliary data, that might become efficient for surveys over time. It also demonstrated the need for parameter driven approaches to provide adequate flexibility.

During the 1980s, the ABS needed to move all its applications onto a new platform, and as part of this move, a second generation of generalised survey tools, known as Survey Facilities, was developed, as a modular suite. The design and estimation tools were expanded to cover a wider array of options, and, having learnt some lessons from the inflexible GSS, the functionality in the new tools was broken down into very small component parts, which could be put together to build an application. There was a concern that it was impossible to predict what actual estimation techniques might be being used in 5-10 years. The design was therefore to ensure that whatever techniques were being applied, these small components (eg sums of squares and cross products) could be used as building blocks.

However the pendulum had swung too far. Each component part solved such a small part of the overall problem, that there was excessive work in putting together an actual application, particularly as the interfaces were difficult. Experts were required to build functioning systems. This was well short of a plug and play sort of flexibility in the hands of users. However on a positive note, as part of a major program of work moving to the new computing

environment, the development of the survey facilities was very tightly managed, with strong senior level input. Standards were developed and used across all relevant surveys, and documentation for the Survey Facilities was excellent. This latter feature meant that it was possible to go on adapting the facilities to new environments well past what would normally be considered a reasonable use by date. Furthermore the general use of the facilities across all applications, made the change to new facilities more straightforward than would otherwise have been the case.

During the 1980s, apart from those 3 surveys still using the old GSS, business surveys used application specific systems, built around common design and estimation tools. However as endeavours to improve efficiency by making greater use of available technology increased, for example in input and editing, there was demand for further shared use of methods and tools to facilitate this. The ABS developed the Survey Processing Environment for the Economic Division (SPEED), in the late 80s, early 90s. The main aim of this was to provide business survey statisticians with a common look and feel across the different surveys, and an ability to build tailored systems using standard components from design and selection, despatch and collection, data input, and editing and estimation. The new development provided an environment with user friendly interfaces from which the various required components could be called in a menu driven system.

The business case for SPEED promised to provide a standardised but flexible approach to business surveys, implementing the latest technology in an efficient way, and delivering cost savings through improved processes. However the focus, and most of the funding, was directed on providing an environment rather than tools within it, although there were some new imputation and estimation modules included in the development. The result was that while SPEED was successful to a degree, it did not boost new efficient approaches to using technology in processing, as follow through to outcomes in terms of tools provided and used in areas such as processing and editing was limited. Different business areas were free to apply their own approaches to using the environment and applications developed their own content around collection and processing.

As a result the cost of applications development and maintenance continued to grow, with several slightly varying approaches coexisting inefficiently within the “standard” environment. While the intention had been to modernise the full set of Survey Facilities as part of building this environment, a shortage of funds and reduced scoping of the project with cost overruns, meant this was not possible. GENINT and GENEST were developed to provide a variety of estimation and imputation options that were more usable than the very fine components that made up the earlier survey facilities.

Alongside the development of generalised systems for business surveys, ran the development of a business register. A basic register was used in the ambitious program of economic censuses run by the ABS in the late 1960s. This register was substantially upgraded in the 1980s, with a very ambitious program of 'bells and whistles' such as date stamping, ostensibly included. Severe performance problems meant that many of these enhancements could not be turned on. The register system was also very inflexible in coping with business unit model changes, and with storing any additional information, such as auxiliary administratively based information, at the unit level. In the second half of the 1980s, the complexity of the register also reduced the ability of the ABS to adapt to inputting electronic files from tax data, rather than paper based inputs. Complexity meant that changes made to accommodate the electronic input led to system problems that developed substantial lags and leakages in business updating. These impact of these lags and leakages on key economic indicator series was a significant concern, and appropriate adjustments needed to be managed manually over many months.

With the experience of the very complex earlier register, a very tightly managed redevelopment was implemented in the 1990s. This redevelopment deliberately avoided trying to add clever functionality to that already working well in the previous register, and focussed on building a higher level of flexibility, particularly in regard to the units model through an object oriented approach. While successful in its aims, unfortunately, this development used a leading edge technology that failed to become an industry standard and a further conservative redevelopment followed not long after.

At the time that particular problems arose in the late 1980s, with very slow update times, particularly for complex units, and the lags and leakages of business units that accompanied a move to electronic transfer from the taxation office, methodologists had begun work on developing common business rules across all our business surveys for how survey operators dealt with deaths, births, industry changes, mergers etc of selected sample units, including the

use or not of new business provisions. A variety of practices were being implemented across surveys, generally for historical rather than rational reasons.

In an effort to overcome the problems arising with the use of the register, business areas engaged actively in this work, with a senior level committee drawn from across economic and labour areas of the ABS, overseeing the development and implementation of new common approaches as part of a program known as Survey Integration.

While this program started in the late 1980s, it continued through the early 1990s, with a revised focus on bringing coherence to different surveys for example quarterly and annual measures of the manufacturing industry. Up until this time the approach had been to explain differences between surveys in terms of different practices used. But with a greater focus of the ABS on its new mission statement of providing information for decision making, it was more clearly recognised that our job was not simply to put out a number of publications from different surveys, but rather to inform. Rather than simply explaining the differences, we moved on to removing those differences that were within our control.

Even so it was slow progress to achieve the cultural change required across the organisation. To achieve standard methods and processes, diverse areas needed to give up control, and implement change not invented by themselves, while maintaining business continuity. The issues involved in the sample and frame maintenance procedures developed were quite complex, and trust that it would all work was difficult to achieve. Areas had their own priorities. The issue came to a head however in the early 1990s, when the ABS data failed to pick up a turning point in the economy as quickly as it should have. This was due in part to an administrative problem with tax updates to the register, and in part to the particular treatments of births and deaths still being applied to some key surveys.

Senior management moved strongly behind the survey integration project. Standard definitions of survey reference period, standard times for drawing frames from the register and a standard approach to new business provisions, consistent with the definition of survey reference period were applied. A new, single Common Frames unit that worked with the register area was set up. This unit developed quality measure to monitor updating of new businesses to the register, and to produce the modelled New Business Provision, to account for businesses operating during the reference period, but still in the updating pipeline at the time of selection. The unit drew quarterly frames from the register, undertook quality assurance on the frames, and selections and despatch for all surveys. Survey areas no longer had discretion over their frame selection time, the ability to adjust their list of businesses according to 'local knowledge', or to adjust their selections. The unit also produced and monitored business demographic statistics, both as a quality check and as statistical output.

Sample and frame maintenance procedures were standardised across all surveys, and needed to be implemented during collection. In the 1990s collection was decentralised, and the procedures needed to be rolled out across survey areas. An understanding of these procedures needed to be developed and maintained in all the survey areas, and initial attempts at this, based on a large look up manual were not very successful. Following a review and discussions with user areas, an alternative approach based a questionnaire/ flowchart approach was much more successful, as it spoke the language of the survey area much better, and could be used directly during the collection process. Subsequent centralisation of collection for business surveys has facilitated the appropriate application of the standard rules.

Through the second half of the 1990s, recognition of the value of being able to readily retrieve data on a given topic across surveys and other sources led to a strong senior focus and corporate approach to data management. This included the development of an output data warehouse of publishable data, and a standardised publication system to draw and publish from this warehouse. The data management development program ocussed initially on large, relatively simple cross sectional datasets. Time was built into the data model as just another dimension, analogous to industry, and for these datasets this approach worked well. However when the data model was extended to include time series, where functionality in manipulation of the time series was required, the inefficiency of this approach was substantial. Significant debate ensued, and it was only after some heated debate and independent review that it was agreed that time series data was far better treated using a model that recognised the intrinsic properties of time. A time series specific data warehouse with functionality for time series manipulation was acquired. Subsequently when tools for managing and disseminating large and complex, household survey datasets were required, it was readily agreed that an alternative set of facilities based on Blaise and Supercross would be added as corporate infrastructure. Unfortunately these elements never interfaced properly with the rest of the ABS data management infrastructure.

In the early 2000s, for business surveys, the integration that was achieved at the front end through the register and common frame unit, and the back end through the output data warehouse and publication system, was extended further, initially through a common input warehouse, and next through a centralised approach to data collection for business surveys. This latter move was in response to continuing requirements to cut costs, including the costs of maintaining a large number of disparate collection and input processing systems that had flourished in the SPEED environment, and which all required maintenance and enhancement, for example to take advantage of web technology.

Senior management recognised that a move from a survey based approach to a functional organisation would require a major organisational change, and committed to an extensive and well resourced change management program, known as the Business Statistics Innovation Program (BSIP). This was managed from a very senior level, and addressed people issues, changes in business processes, organisational structures, governance processes and technical infrastructure. In general the program of change, involving the movement of functions across all 8 offices, was very successful. Benefits in terms of a substantially reduced number of systems were relatively immediate, however some other efficiency outcomes hoped for from the change are some time in coming, and governance processes and senior focus have moved on to other priorities, putting their eventual achievement at risk. A small unit to follow through in support of these outcomes has been maintained, but resources to pursue gains are limited.

A new set of survey facilities was also developed as part of BSIP. In terms of estimation, a variety of options had already been updated from the initial survey facilities under GENEST, but here another lesson had been learnt. The number of options provided, when applied to the number of possible survey designs possible, multiplied out the number of different variance estimation options required, if direct variance estimation was used. Recognising this, and wanting to yield an adequate range of options for design and estimation while keeping variance estimation manageable, a conscious decision to use a replicated weighting approach to variance estimation was taken in the new survey facilities. In other words, generalisability of method had become a selection criterion for choosing a standard method. This approach has worked very well in practice.

3. Why standardise?

The above very partial view of ABS statistical infrastructure developments, show that over the past 5 decades the ABS has invested significantly in developing general methods and tools to support its work. This development effort continues today with new generations of statistical infrastructure under development. The benefits of standardisation of methods and tools are many. Different drivers may be key in different situations, but there are at least five good and related reasons for such standardisation.

The first is efficiency. With multiple users, development and maintenance cost per user are kept down, and maintenance is particularly important here. The cost of keeping a myriad of different systems across an organisation maintained, and enhanced as new technological or scientific opportunities arise, is often prohibitive and the most visible driver to standard approaches. Training costs are also lower with standard approaches, as fewer training tools need to be developed, and more importantly, staff who gain experience in one part of the organisation, can reuse skills relating to the methods and tools in another part, reducing the cost of staff mobility. As development effort is focussed on a standard tool or method, that development can afford to be more rigorous, better managed and documented and thoroughly tested, reducing the scope for errors and the need for rework. Possibly the greatest benefit in terms of efficiency though, is the increased ability to upgrade with newer methods and tools across the organisation as the opportunity arises. Upgrades to standard components can be readily distributed across the organisation, if implemented in a common way.

A second area of benefit is accuracy. A better method or tool may be able to be afforded if the development costs are shared many ways, and a well tested standard is less likely to create error. As well, with a reduced number of ways of doing things, standard methods and tools can be well documented and understood, and managed over time. They are less likely to become poorly understood and potentially misused black boxes.

Standardisation also supports greater coherence and comparability of data across the organisation. By eliminating the variation inherent in a way of doing things, it is easier to focus on variation in the underlying statistic of interest, or patterns in the relationships of the underlying statistics. This is clear where standards relate to such things as item

definitions and classifications. However it is also true where standards relate to methods and tools used, for example to account for births and deaths of businesses, the treatment of outliers or seasonal adjustment.

Fourthly, there is an improved level of control of process and hence outputs that comes with standardisation. With fewer methods and tools in play, it is easier to establish corporate policy, for example in relation to security, data management and dissemination. It is also easier to monitor and audit the implementation of such policy.

A fifth good reason for standardisation of methods and tools is that they position an organisation to be responsive to new requirements, and maintain relevance. They support rapid development of new applications. Instead of starting from scratch, they provide the potential for relevant methods and tools to be chosen and integrated, with appropriate parameter settings for a new use.

3.1 Is there a downside?

There can be downsides to standardisation. If it were easy to implement standard methods and tools, we would not still be working so hard at it, so many decades on from the first efforts at general survey systems.

Standards generally require compromise. The global optimum is not the local optimum. While it may be possible to provide ways of fitting a standard method or tool to a particular circumstance, through particular parameter settings, it will often be the case that providing sufficient flexibility to closely fit all known types of situations will substantially increase the complexity of the standard. This may affect performance to the detriment of all users. For this reason the bells and whistles provided with a standard may be minimised, reducing the fit for some applications.

Sometimes the cost of generalising is simply too great to be worthwhile. A common process is not always a good enough reason for a common method or tool, if the circumstances differ too greatly. For example, in many ways the process of traffic control is the much the same whether the traffic is in the air, on road, rail or the water, but the circumstances differ considerably, and while there is some language and logic in common, there are few tools that are common across these different types of traffic. Or for a more statistical example, the ABS experience in trying to generalise elements of data management across large cross sectional and relatively small time series data, demonstrated that while time could indeed be considered as just one dimension along which data could be described, the particular attributes of time made this a very inefficient approach for managing time series as output data. Performance of the generalised ABS data warehouse system for the management and manipulation of publishable time series of statistics was prohibitively costly, and the cost of generalisation across these different data types too large.

As well as the global optimum not being the local optimum, another downside is the loss of local control on the maintenance and improvement of the tools central to their effectiveness. If funds for new developments are centralised, there will be no resources to update locally as the need is perceived to arise. Furthermore, with shared use of corporate tools there can be issues of performance depending on load placed on them. Local areas lose control of parameters affecting their ability to deliver to required timetables.

This loss of local control to adjust methods and tools, may also have a negative impact on creativity and innovation, reducing motivation to look for better ways of doing things. This might worsen as business areas lose familiarity with their corporatised methods and tools, with the result that they become less aware of opportunities for enhancement of processes and tools as technology develops. Corporate owners, if not sufficiently aware of business issues may similarly miss such opportunities for improved business outcomes. This points to a need for a close working relationship between those with responsibility for the methods and tools, and the business areas.

There is also a need to find ways to continue to encourage local innovation and the sharing of successful ideas across the organisation. The Apple strategy of encouraging the development and spread of applications that build on the standards, and allow those that build a following to survive while others come and go, works well in the commercial context with a profit motive to drive it. The ABS experience with the SPEED environment where business areas could bid for funds to develop applications within the standard environment, was well received by user areas in terms of local building within a corporate infrastructure, however it did not reward shared use or cross ABS promotion of new local developments.

Another drawback with implementing a new standard is the change it requires. An area which already has a local application that works adequately from their perspective, may be reluctant to carry the cost (both in terms of dollars and energy) to make a change to a new corporate standard. If change is to be implemented across many areas, this cost can be very large, and a funding source, such as a capital injection may be needed. While statistical agencies can often attract user funding to undertake important new statistical work, it can be much harder to arrange for new sources of funding to develop and implement statistical infrastructure for more generalised use.

Furthermore the change required is rarely simply a technical one such as the removal of one system to plug in another. Instead the change may require different governance mechanisms with central control of things previously controlled locally. It may require different functions to be performed by staff, different skills to be needed, new organisational structures, new ways of working across the organisation and so on. In other words developing and implementing a standard may require significant change management across all affected areas, and if a standard is to be used across a large part of the organisation, the change will need to be managed as a significant organisational change, with all the complexity of cultural and organisational as well as technological change.

If a major program of standardising methods and tools is to be embarked on, the magnitude of the task needs to be planned for and managed through to full implementation and realisation of outcomes, as a change program. The magnitude of the program embarked on must be commensurate with the benefits to be gained and the resources available to achieve the change.

4. Planning a significant development and implementation of standard tools and methods across an organisation: six key decisions that need to be made

Undertaking a program to standardise tools and methods across a statistical organisation, is a major change project. It will have significant associated costs, both in terms of dollars spent, and in the use of scarce capabilities. Standard methods and tools will seek to incorporate the best opportunities that new technology and statistical science have to offer, and will place demands on key staff for development and implementation, while also requiring the maintenance of 'business as usual' based on the old methods and technology.

Some key decisions for management are:

1. What are the key drivers for the standardisation. For example, is it to achieve efficiency, or to achieve control, or for other quality related benefits.
2. The scope of such standardisation: how much of the end to end processes of the organisation will be covered: over which subject areas/ survey areas, *etc.*, will the tools extend; and for what time frame are the standards being built (for example do they need to be able to link with technical opportunities that can be seen coming to maturity in 5 years say)
3. Who owns and directs the development and implementation. In particular who makes key go/no go decisions, prioritises use of resources, determines changes to scope if needed, *etc.* Related to this is the governance structures that are needed over the life time of the program through to benefits realisation.
4. The degree of standardisation to be achieved, and the degree of complexity in solutions that will be tolerated (how much flexibility for alternatives will be offered, possibly at the expense of cost, performance and ease of use, and maintenance requirements)
5. Whether the methods and tools being developed will be compulsory or optional for business areas.
6. The degree of granularity to use in designing modules (the tension between flexibility and usability)

Some discussion of each of these issues follows.

4.1. Drivers for Standardisation

The benefits of standardisation are discussed above. Which combination of these benefits are key in the implementation of a proposed program will determine the approach taken. If there is an immediate cost imperative, then those elements of standardisation that will readily yield most savings will be prioritised, whereas if the key motivation is quality and control, those areas of greatest risk will be the key initial focus. Where trade-offs need to be made, for example between efficiency (*e.g.*, in number of tools, bells and whistles) and quality (*e.g.*, in levels of functionality to meet various local needs), they can be made consistently based on an agreed overall direction.

As a program of standardisation is a change management program, and is likely to require cultural change and personal commitment of staff across a number of areas, it is valuable to provide clear messages of purpose. If a key driver is to improve efficiency, it is tempting to also sell this as a push for greater quality, particularly when greater quality is a likely outcome. The danger here is that realisation of the intended benefit, efficiency will be muted by a tendency for staff to make trade-off decisions in favour of quality, not necessarily in line with the key management requirement.

4.2. Scope

Scope should be determined according to the key outcomes required, the resources available (both dollars and capability), and any impediments, *e.g.*, cultural impediments that need to be overcome.

This is stating the obvious, but we frequently see examples of a mismatch between ambition and capability resulting in work commencing on a grand plan that later needs to be substantially scaled back, for example to provide 'at least the functionality of the previous system', with a substantial reduction in benefits realised over the initial business plan promises. A major disadvantage of this 'over ambition', is that resources tend to be over allocated at the front end in exploring and developing wide ranging options, and under invested at the back end in realising outcomes through effective implementation and monitoring.

Some potential causes of over ambition regarding scope?

- Simple greed: we see opportunities and we don't like to say no to ourselves or others, so we bite off more than we can chew, from both a capability and budget perspective.
- Lack of honest appraisal of what our real level of capability is, and how quickly the level can be built to that required.
- A general level of over optimism, for example in relation to things working first time, or low levels of staff turnover.
- Lack of trust that the cost estimates supplied are reasonable. (Have they been intentionally padded as a risk mitigation strategy, or because there is an expectation that only partial funding will be provided? The latter concern leads to a cycle of over quoting and under funding that can be difficult to break). This distrust often leads management to reduce budgets while seeking full scope.
- A related problem can arise where management, bidding for funds from an external source, may bid for a larger scope than they have the capability to achieve, in the expectation that they will receive only partial funding and will have to cut the scope. If the full funding is received, the agency can find itself with a capability shortfall.
- A failure to appreciate that the program must be planned, managed and resourced from initial stages, right through to benefit realisation. This leads to a tendency to only consider development costs in determining scope, on the assumption that business implementation costs will be relatively minor, and can be "absorbed".
- Related to the previous point, can be a failure to appreciate the energy and cost associated with moving an organisation's culture and understanding to a position where major corporate developments are understood, valued and contributed to by the business areas.

The first four points above are the key reasons why the outputs achieved from a program of standardisation, in terms of methods and tools, will often fall short of the initial plan, in terms of coverage or functionality. However a business plan justifies costs against realisable benefits, not outputs, and it is the last two points that is most likely to be the key reason, that even with the outputs delivered, the benefits set out in the business case (*e.g.*, reduced processing costs, faster outputs) fall well below the promised, and theoretically achievable level.

The last point has different levels of relevance across organisations. Some are well structured to support corporate approaches, with a strong corporate culture already in place. The ABS falls fairly well into this category. Other organisations, possibly formed from recent amalgamations of different organisations, or funded on the basis of distinct stove pipes of work, may find it more difficult to build a culture where people understand and value the corporate benefits and are motivated to help drive to the outcomes sought.

In looking at required resources to achieve benefits, it is important to cost all elements of the change process, included those needed to make intended changes stick in the business areas across the organisation, and stay stuck. It may be better to work initially with a limited scope, but ensure resources are available, and governance is maintained, right through to realising the benefits.

A common concern regarding major developments in the 1970s and early 80s, was the investment trap, where large projects promised substantial benefits, but actually incurred significant overruns in time and money. Management faced the problem that the only way to achieve any value from the spent investment was to keep on spending, with promises that benefits were imminent.

One strategy the ABS implemented effectively through the 1980s, was to institute a rule that each new development must have a stand alone deliverable within 6 months and have no more than 10 people working on it at one time. Larger projects were broken into smaller pieces capable of yielding value on their own.

However this strategy has some pitfalls that need to be avoided. In particular it encourages the incremental development of standards from particular “prototype” applications, or “proof of concept” approaches that may not fit well all the applications intended to use the standards. If the first deliverable in 6 months, is a set of methods and tools for editing in a particular collection, and the intention is to further develop this set as a standard for use across many collections, there needs to be care that the development path taken will actually prove effective for the full intended suite.

In general it will be necessary, in determining the scope of the project to fit resource availability, to undertake some form of staged program, and the question arises as to how to define the scope for each phase, in terms of which parts of the statistical system to cover, not only in terms of the subject fields and collections to be covered, but also in terms of parts of the statistical value chain to be included in the development. Issues to consider here include:

- Where do logical breaks in the statistical process occur? How will connections be built between redeveloped processes and old processes, and how can the cost of these connections be minimised?
- Where do dependencies arise: is it possible to realise the benefits in the business case with the scope chosen; are all components that are needed to deliver the benefit included in scope (including training and implementation components)?
- Where are the opportunities for achieving outcomes greatest (*e.g.*, currently the most inefficient areas, or the areas of highest risk)?; Where are the opportunities easiest to achieve (*e.g.*, less complex processes, or more engaged business areas)?
- In which areas are opportunities for improvement still developing, for example through emerging technology?; it may make sense to postpone work in such areas til some maturity is achieved in these developments.

However the scope is determined to fit available resources, the full range of the program should be mapped out up front, to ensure for example that if low hanging fruit are to be the focus of early work, the development will none the less be capable of effective expansion to fit all the fruit in the intended fruit bowl. Trying to fit more complex and less willing applications into a mould that does not suit, is likely to lead to significant performance problems and failure to realise benefits.

4.3. Who owns and directs the development and implementation?

If you want to make significant changes within an organisation, it is important to understand how these changes will affect the organisation's people, and what the associated motivators for different groups will be. Statistical agencies are often driven by strong values, which may either help or hinder change processes. These values centre on integrity, quality and service. Costs and efficiency tend to come in at a lower order. Cost becomes a driver when there is a threat to one of these core values. For example a significant budget cut may become a key motivator for

implementing more cost effective approaches, if this will retain integrity, quality and service levels which might otherwise be threatened. Whatever the driver, in order for change to be implemented effectively, it must have strong ownership, from development to implementation, by the business areas, from senior management levels down. This does not mean that development must be done in business areas, but that business areas must have a strong sense that the development is being done for them, and with their engagement.

Business areas are readily motivated by their senior management, and by clients. They are not so easily motivated by a bright idea or tool coming from another part of the organisation such as Methodology or Technology Services areas. If there is not business area ownership of the proposed project, with strong business area belief that the intended outcomes are both achievable and valuable, support for development will be low, and implementation will be extremely difficult. Regardless of who undertakes development, senior management of business areas must own the project, support it with appropriate expert input, ask hard questions of progress and prioritisation, and join the decision making over any needed rescoping. They must stay engaged, at the senior level, right through development, implementation and benefit realisation.

The skills and knowledge of the business areas are critical to developing the new approaches, and solving problems as they arise. In particular the practical issues in relation to what might work and what might not, in terms of processes, as well as the statistical issues and client needs, all need to be taken into account in the development and implementation. On the other hand, the development should not be constrained by past ways of doing things, so ideally those with business knowledge will be able to think outside the box of past practice, to what might be possible with new technology and methods. Standardised infrastructure must take into account the varying needs across the organisation. For this reason it is generally effective to separate development work from the day to day pressures of maintaining business as usual.

Methodologists and technology areas are important parts of the team for building corporate infrastructure, along with the business areas. Methodologists, in particular, if they have been exposed in their careers to a variety of business areas and practical developments, will have a cross cutting view of how things might be done, and a cultural propensity for embracing change and opportunity. On the other hand, methodologists who have been kept separate from business areas, in highly specialised roles, or who see themselves solely as the guardians of a narrow view of quality, may resist changes to past methods and processes as resolutely as any others.

In general, the ideal is to have a well motivated cross disciplinary team, “owned” by senior management of the business areas, co-located and working as one, to one integrated work program, and with one set of governance mechanisms determining priorities.

Issues that arise in relation to such a team, include:

- How to free the important business capability, and methodological and technical capability that is also needed to maintain business as usual (*i.e.*, how to redirect scarce resources from the day to day operational requirements, to the strategic changes)?
- How much to separate the development teams from the 'business as usual' team, especially if resources are scarce?

These two points are interrelated, as significant separation of development work from 'business as usual', can increase the demand for scarce business experts compared to combining ongoing and development work under one management structure. The best approach depends on the scale of project undertaken, but for very large projects there is a logic in creating the 'new world' version of infrastructure under a separate management structure from that used for business as usual, to allow a dedicated focus, and high level input, on the development work. Related projects can be brought together in a program, and help support each other under this model, with the energy of the project being maintained. Moving old business into the new world can then be a struggle if the engagement of 'business as usual' has not been sufficient during the development. The ABS suffered from this separation in some major projects, for example the Data Management project of the 1990s. Other projects such as the Business Statistics Innovation Program worked hard to maintain that engagement from the most senior level, to good effect.

An alternative approach, being taken within ABS in its current modernisation program, is to create a separate management structure for the development program, but to include within the span of senior management control, a

mix of business areas most involved in the key aspects of the development program at a particular time, with the expectation that business areas might move in and out of this management structure as the development progresses. Management is then responsible for both the business as usual and development work of those areas where modernisation is most focussed, but with a changing mix over time. Engagement with other business areas will need to be maintained, as generalised tools must meet broader requirements, and a key learning of past experience is that a focus on management right through to benefit realisation will be needed, particularly in deciding when, and under what governance arrangements, businesses areas move back out of the modernisation group.

4.4 The degree of standardisation to be achieved, and the degree of complexity in solutions that will be tolerated

Two related issues management must consider are: the degree to which one method and/or tool will be used across all possible applications; and the degree of complexity that management will tolerate, manage and resource in a standardisation to allow it to be tailored to different applications.

There is a strong logic to the mantra of “keep it simple”. Simpler systems tend to perform better, and be easier to use, maintain, and upgrade. However if the tool or method is to fit a variety of applications, without losing too much local optimality, it is likely to need to be possible to set a number of parameters to achieve an adequate fit. This will increase complexity. At heart this is a cost benefit issue, where decisions should be made on a good analysis of the costs of complexity and the benefits of more effective local implementations. The important issue is to ensure both the costs and benefits have been well tested. The situation of the ABS register development of the 1980s where the substantial functionality built into the register could never be turned on because of the associated drop off in performance, provides a lesson in the need for thorough performance testing throughout the development cycle.

Where complexity is the result of providing greater functionality to handle what are in fact complex circumstances that arise in reality, it may be necessary to accept the complexity and manage it through good design. The problems we are seeking to solve as statisticians are becoming more complex each decade, with greater data demands for integrated and cohesive data from a variety of sources, at a variety of levels, viewed over time, minimising load on providers through choices of collection modes, and maximising likely use by analysts through multiple and changing dissemination methods, while maintaining confidentiality in the face of increasing sources of information that impact identifiability. If these increasingly complex problems are not solved in our sets of standard methods and tools, the problem of solving them will fall back on the local areas, at significantly higher cost to the organisation.

Hence, rather than the mantra of “keep it simple”, it is more relevant to think in terms of managing complexity, through good design and implementation (including good training), and through adequately testing impacts on performance, and resourcing resolution of issues. We need to remind ourselves that the complexity of landing a man on the moon was managed. It was just expensive. Complexity should not be avoided per se, but the degree of complexity should be based on costs and benefits of managing it, and complexity should not be visible to the user. In a plug and play approach, experts should be able to take off the covers and understand the complexity, but regular users should not need to. Management of complexity requires considerable relevant expertise, and complexity should certainly be avoided in the absence of this expertise.

4.5 Whether the methods and tools being developed will be compulsory or optional for business areas

Another management decision that must be made relates to whether the standard methods and tools are to be compulsory for all relevant business areas to pick up and use, or whether they are to be optional, and available to the business areas to use. This decision will depend in part on the purpose of the standardisation. If the aim is to achieve control, for example to ensure that all published and publishable outputs are appropriately disseminated and discoverable and satisfy organisational standards for quality and management of confidentiality, then the standards will be compulsory.

If the purpose on the other hand is to achieve efficiency gains, and there is a case that business areas are in the best position to judge their applicability to that business, it may be considered appropriate to leave the business area to judge whether to pick up the standard. This will also remove the monopoly situation developers might otherwise enjoy, and encourage them to make sure their development fits the need of the business, and is maintained and upgraded as required by the business. However this logic does rely on the business area having strong motivators to

choose the option that best fits corporate needs rather than local needs. If a local area already has adequate resources, and will not itself benefit from an increased efficiency whether local or corporate, there may be no incentive to take on a change process which involves a loss of control, as discussed above.

If compulsion is used, the onus is on senior management who have made this decision, to make it work. This places greater accountability on corporate management and is less comfortable for senior management.

4.6 The degree of granularity in tools

Standard methods and tools are often built as modules, so that those developing an application can pick and choose from a suite of options, joining them through friendly interfaces, giving rise to the plug and play analogy. This provides flexibility in tailoring to particular applications, and to meeting as yet unseen needs that may arise in the future.

The ABS General Survey System of the 1970s predated a recognition of this modular approach, and the user was delivered a full, but largely inflexible system from beginning to end (though there were some more flexible mechanisms for interrogation and amendment of databases that were also provided). The ABS Survey Facilities approach of the 1980s recognised the need for flexibility to meet a variety of needs, including some future needs, that could be achieved with modules, but the pendulum swung too far... each module solved too little of the problem. It required an expert to put the modules together. With high staff mobility, tools need to be able to be picked up and used in an intuitive way, to do something useful, with minimum previous experience.

5. Managing a major standardisation projects, some do's and don't based on past experience

When management is embarking on a significant program of standardising methods and tools in an agency, setting the scope set and making the various decisions that will broadly define the approach to the development, past experience points to some strategies that seem to be effective in achieving successful outcomes. In my view the following are key:

- Dream big and map the possibilities broadly, over a larger budget than might be available, and over a long time frame of possibilities.
- Step back to reality and plot the steps forward, trying as far as possible to sequence components on a small project, quick delivery basis, not precluding longer term intentions, and ensuring full senior management understanding and buy in.
- Identify the parts of the program with big interdependencies and plan these carefully... the initial application in these areas may provide some quick deliverables, but it is important to ensure that the shape of the solution will fit with all of the interdependencies, not just the first application.
- Plan and manage for outcomes, not outputs; understand the changes in individual and organisational behaviours and roles needed to realise the benefits.
- Use good project management approaches to ensure clear objectives, good decision making processes, clear accountability and effective monitoring systems right through to securing outcomes; don't drop project management off once outputs are developed.
- Ensure senior management continues to drive the project til delivery of the outcomes.
- Ensure business buy in; ideally they should be identifying the problems that are being addressed, and discussing solutions for example in group discussions.
- Be realistic about resources and capability even if it is uncomfortable; be clear and honest in communication; if some components such as implementation, need to be financed from ongoing current budgets, make explicit what will drop off the work program.
- Ensure the project team is brought together as one unit, with a common set of goals, priorities, timetables and language whether they are from technical methodological or business streams.
- Test early, test often and be honest (engage the business owner, build trust and seek input to problem solve).
- Fess up when things go wrong and deal with them early (don't wimp).
- Reprioritise when needed to deliver maximum value, and be clear on any changes and their implications for the realisation of benefits. Don't allow a trust gap to grow with business clients.

- Communicate regularly, clearly, honestly.
- At the outset, avoid the perfect being the enemy of the good.
- Towards the finish, avoid “near enough is good enough”, keep the energy up til the outcome is secured; this means providing follow through support in the language of the business areas.
- Ensure the business area is accountable for outcomes not just the developers; business areas must focus on problem solving during implementation, and not adopt “I told you so” behaviour.
- Monitor (say 12 months, 3 years after implementation) and take action to ensure the benefits are realised and the new methods and tools stick and stay stuck.

On the flip side, history tells us that there are certain danger signals to look out for:

- the solution looking for a problem (someone’s bright idea often is a bright idea, but it needs to be engaged with and owned as solving a real world problem by the business areas);
- something ill defined... hazy and out there which resists clear definition and management;
- advice from a vested interest;
- the cutting edge (plan for a much bigger budget in terms of dollars and capability, if for some reason you want to experiment at the cutting edge);
- a divided team;
- an under resourced project, especially if it is under resourced for the implementation stage, and especially if shortcuts are being taken with components like documentation and testing;
- performance problems;
- turning off governance once the output is achieved, but before the outcome is delivered.

6. Conclusion

Standardisation has many benefits, and because of this, it is not surprising that statistical agencies have been working at it, with considerable success, but some learnings, for a number of decades. This experience helps guide new efforts as technology, as well as other opportunities and needs arise. We know up front, some of the issues that need to be addressed and decisions that will need to be made, and can set ourselves up to make the decisions wisely. We know some of the pitfalls to avoid, and some of the things that do seem to lead to successful outcomes.

There will, of course, be new issues and new answers to be found as the environmental context changes, but probably the most significant learning from experience is likely to remain fully relevant. Clever technology or methodology, is never enough. To achieve enduring outcomes, a significant project of standardising underlying statistical infrastructure must be recognised as a change process. Success will depend on senior managers having a clear understanding of, and belief in, the outcomes required, and providing strong leadership through to ensuring their enduring realisation. Success is more about people than it is about technical solutions.

SESSION 2A

**CORPORATE BUSINESS ARCHITECTURE:
HOUSEHOLD SURVEY FRAMES**

Redesign at Statistics Netherlands

Frank Hofman^{1,2}

Abstract

Statistics Netherlands (SN) faces a number of major challenges: improving efficiency and the quality of key statistics, while lowering the administrative burden at the same time. To meet those challenges, SN has developed an enterprise architecture, generic business services and a standard toolbox. Subsequently, the statistical processes have to be redesigned one by one according to the architectural principles and using the generic services and the toolbox. A redesign approach has been developed to facilitate the individual redesign projects.

This redesign approach is based on the statistical design process as it is seen by the enterprise architecture and combined with the RUP process for software development. Two additional principles complement the main process of statistical design and software development. These are: 1) a preliminary investigation and a project initiation precede the main process and 2) the most important artefacts of a redesign project are centrally reviewed. In general, the redesign approach has proven to be useful, although some comments are appropriate.

Key Words: Approach; Architecture; Design process; Document.

1. Introduction

Statistics Netherlands is in the middle of radical changes. A number of very heterogeneous driving forces pose major challenges that can only be met when the way the institute operates is thoroughly reconsidered. Van der Veen (2007) gives an overview of the present situation and the challenges ahead in a broad context. In particular, efficiency and quality of key statistics must be improved, while at the same time the administrative burden has to be lowered considerably.

In order to stay in control of these challenges, an ambitious modernisation programme, the Master plan ‘Counting on Statistics’ (Ypma and Zeelenberg, 2007) has been started in 2005. As part of this Master plan, the enterprise architecture, some generic business services and a standard toolbox have been developed (see Braaksma, 2009). The ultimate goal of the Master plan is to redesign all statistical processes according to the architectural principles, which implies wide adoption of the toolbox and the generic services. Redesign of all statistical processes, which amount to several hundreds, is a huge task and can only, be accomplished step by step. SN has developed a redesign approach to help individual redesign projects on their way.

In this paper, we will explore this redesign approach, which was developed by the redesign teams and is now generally adopted within SN. We start by describing the redesign process, before focussing on the most important products to be delivered: the Business Analysis Document (BAD), the Methodology Advisory Document (MAD) and the Software Architecture Document (SAD). In the final section, we will reflect upon experiences with the redesign approach in redesign projects and we will look at current and future developments.

This paper is a shortened and slightly updated version of an earlier paper: Hofman and Leerintveld (2010).

¹Frank Hofman, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands, f.hofman@cbs.nl.

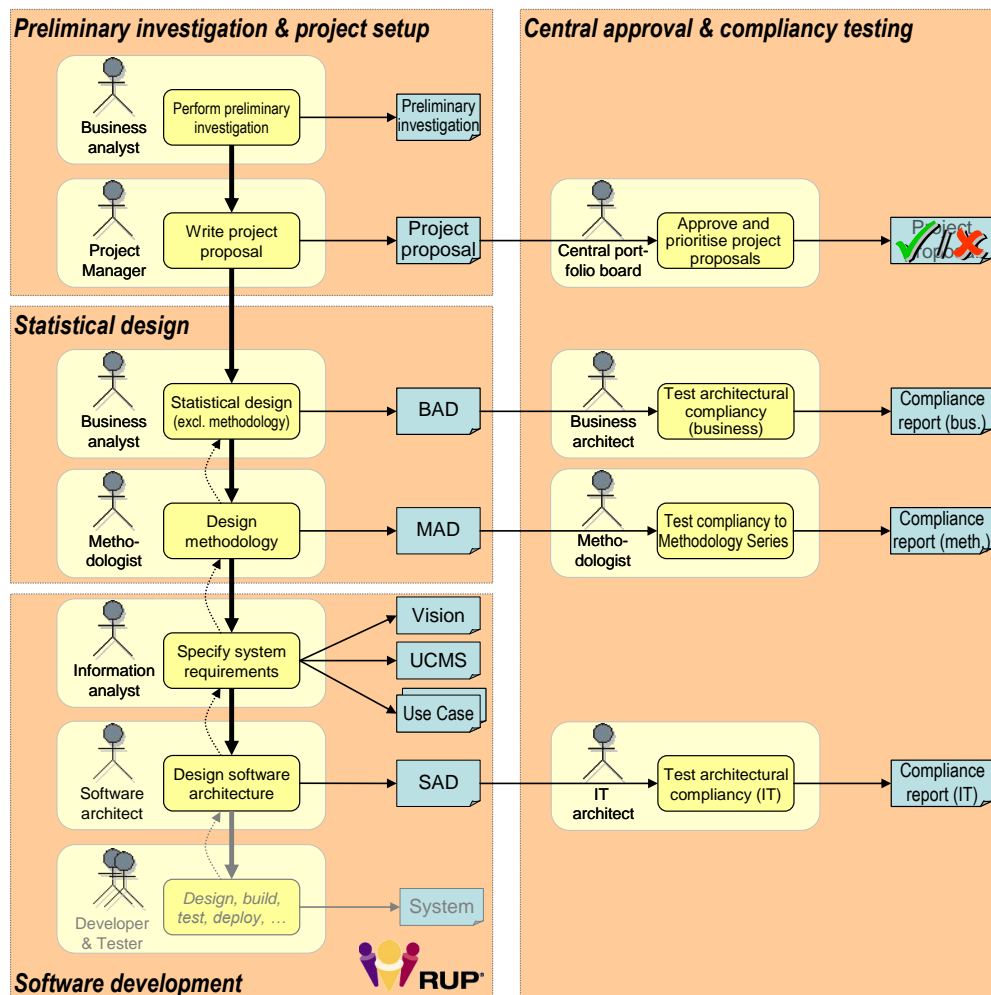
²The view expressed in this paper are those of the authors and do not necessarily reflect the politics of Statistics Netherlands.

2. Redesign process

The redesign process has been developed to support redesign projects with a standard way of working. The process gives a number of related activities that need to be carried out during an average redesign project. Each activity is carried out by a specific role (indicating a set of competences) and produces (a part of) a document (or other artefact). Although the sequence of activities may suggest a strict sequential order, they may also occur in parallel or iterative.

The scope of the redesign process is broad, starting even before a redesign project has been set up with a preliminary investigation, and extending to the delivery of the new information system. The focus in this paper is on the first part of the process up to the 'statistical design' and its connection to the system development. SN has adopted some well known standards for parts of the redesign process: RUP (Rational Unified Process³) for system development, TMap⁴ for testing and Prince2⁵ for project management. This paper focuses on the non-standard parts of the redesign process.

Figure 2-1
Redesign process



³http://en.wikipedia.org/wiki/IBM_Rational_Unified_Process

⁴<http://eng.tmap.net/Home>

⁵<http://en.wikipedia.org/wiki/PRINCE2>

Figure 2-1 shows the redesign process. A quick overview shows that before starting an actual redesign project, the project proposal needs to be written and approved. In order to properly fill out a project proposal, it is often necessary to carry out a preliminary investigation into methodology, statistical products, tools, *etc.*, focussing on needed changes and setting the course for the actual redesign project. During the statistical design the statistical products are designed as well as the process and methodology to transform the input data into these products. The statistical processes generally need supporting systems (software) that have to be developed.

Some documents that are produced during a redesign project, need central approval or need to be checked for architectural compliance.

In this section we will examine each sub process of the redesign process.

2.1 Preliminary investigation & project setup

For each (redesign) project the project manager has to draw up a project proposal, describing, among other things, the business case and goals. Since SN uses Prince2 for project management, the project proposal consists of the project mandate, project brief and/or the project initiation document (PID).

If it is deemed necessary by the project manager or (members of) the designated steering committee, a project proposal can be preceded by a preliminary investigation. The investigation usually starts with the as-is situation, briefly describing the current process to indicate problems and/or desired changes. It may also include an analysis of external changes that have to be dealt with. During the investigation alternatives for the future situation are examined. Balancing the pros and cons of each alternative results in a choice by the steering committee setting the course for the actual redesign project.

The investigation is usually carried out by a business analyst, but other specialists like a methodologist, an information analyst and a software architect may also be involved, if appropriate. A fixed format for reporting a preliminary investigation is under development.

2.2 Statistical design

Since this part of the process is most specific for a national statistical institute (NSI), we discuss the statistical design in more detail than the other steps. The enterprise architecture of SN identifies five steps in (re)designing a statistic, of which the first four are compressed into the first step in figure 2-1:

1. Determine statistical information needs
2. Design statistical product
3. Design data sources
4. Design process model
5. Design methodology

As for the entire redesign process, these five steps do not necessarily have to be carried out sequentially. They can also occur in parallel or iterative. For example, it may appear to be difficult or expensive to obtain the exact input during the design of data sources, while a register with slightly different information is readily available. In such a case, a business analyst needs to go back to the design of the statistical product and even to the customers to evaluate whether a slightly different output also covers their needs.

A business analyst is responsible for writing the Business Analysis Document (BAD). Writing the Methodological Advisory Document (MAD) is the responsibility of a methodologist.

Determine statistical information needs

As a basic principle, the design process is output-driven. So we start by determining the statistical information needs. During this step, we identify our customers and their needs.

This step adds the internal and external customers for the statistic in development to the context diagram. This is described in the BAD.

Design of statistical product

Knowing the needs of our customers, we can now design the actual statistical product. We determine the table(s) to be produced, the population, variables and aggregation levels as well as the quality metadata (indicators and standards) and the publication frequency of the statistic.

To minimise (near) duplicate publications and maximise the coherence between the statistical products, it is important to be well-informed on existing output and intermediary products within the same statistical theme.

Typically, a statistic has external customers; however, many have internal users as well. In that case, we also design the intermediary product(s) which will be reused by those internal users in this step.

The result of this step is a description of the steady states for the output and possibly for the intermediary products in the BAD. Steady states are a key element in SN's architecture consisting of data sets with guaranteed quality which are made available for re-use through one of the generic business services, the Data Service Centre.

Design of data sources

Knowing what the statistical process has to deliver, we can now turn to the means of achieving the desired output. This means we have to design the input data sources needed to produce the output. The description of the input data sources is comparable to that of the output: conceptual metadata and the quality-metadata.

Even more so than during the design of the statistical product, it is important to have an overview of available input and intermediary products when designing data sources. To reduce the administrative burden on citizens and businesses, new surveys or questionnaires may only be considered, when none of the available sources within SN or other governmental organisations is suitable for the desired statistical product.

The result of this step is a description of the steady states for the input in the BAD.

Design of process model

In the process model, we describe the steps (activities) to be completed to produce the statistical product (output) from the data sources (input), incorporating the generic business services. From each step we describe goal, input and output. The input may consist of auxiliary information, for example calculated the weights needed for "grossing up" sample survey results to make them representative of the target population.

The flow through all steps is as important as the steps themselves, especially when the flow is complex, containing branches and loops. The criteria used for decisions, like the stop criteria of loops, are often derived from the quality needs of the statistical product.

For more complex processes, we do not only describe the production process, but also the management process, especially when the statistic is part of a chain of statistical processes. We state the process indicators needed from the production process to accurately manage the process.

The results of this step are a process model as well as (additional) intermediary products in the BAD.

Design of methodology

The final step in designing a statistic is choosing the suitable methodology to produce the output at the required quality from the selected input. In practice, the main methodology is often explored before the process model is designed. Then, in iterations, the methodology and process model are elaborated upon.

Another basic principle is to use only validated methods. Therefore, SN has developed the Methodology Series, a series of scientifically based, well documented and proven methods, which are the preferred methods to use in a specific statistical process.

The result of this step is a description of all methodology described in the MAD.

2.3 Software development

To go from statistical design to information system, SN combines the statistical design process with the RUP as system development process. In this paper, we will focus on the requirements discipline and the software architecture part of the analysis and design discipline.

The requirements discipline is crucial for a smooth transition from statistical design to system development. The information analyst captures the system requirements in the Vision document and elaborates the systems usage in the Use Case Model Survey. The detailed description of the system's function is captured in several Use Cases.

The software architectures of the individual projects play a crucial role in managing the overall IT landscape. An important guiding principle is the reuse of existing software within SN, before buying standard software, before developing bespoke software.

The software architect describes the projects software architecture in a Software Architecture Document (SAD).

2.4 Central approval & compliance testing

All documents are reviewed within the redesign project. Additionally, some documents have to be reviewed by a central authority. Currently these are: the project proposal, the Business Analysis Document (BAD), the Methodological Advisory Document (MAD) and the Software Architecture Document (SAD).

Project proposals for redesigns have to be approved by a central body, the 'Central Portfolio Board', consisting of the deputy director-general, the directors of the three statistical divisions, the director of IT and the director of methodology. The board checks the legitimacy of the business case and prioritises the projects. Only after a redesign project proposal has been approved and the necessary resources have been allocated, a redesign project may actually start.

After the BAD has reached a certain degree of completeness, it is submitted to the business architects for review. The business architects determine whether the proposed statistical process complies with the enterprise architecture of SN and document their findings in a compliance report to the steering committee of the redesign project. If the BAD is found to comply with the enterprise architecture, the redesign project will continue. If not, the steering committee may decide to either fix the BAD or ask the Central Architecture Board for permission to deviate from the enterprise architecture. The Central Architecture Board consists of the vice-director of the methodology division and the managers of the development departments.

Similarly to the BAD, the MAD and SAD are checked by the (corporate) methodology division and the IT architects for compliance on the Methodology Series and IT part of the enterprise architecture, respectively.

3. Main design documents: BAD, MAD and SAD

In the previous section, we have shown which design steps contribute to which documents. In this section, we will focus on the main documents themselves: the BAD, the MAD and the SAD.

3.1 The Business Analysis Document (BAD)

The BAD captures the statistical design except for the methodology (which is captured in the MAD). The main topics in the BAD are:

- **Context**

The context describes the surroundings of the statistical department. It contains the external and internal customers of the statistic as well as the suppliers of the data sources. Furthermore, it contains the generic

business services which are used within the process, like the Data Collection Service and the Data Service Centre.

- **Products**

This section describes all steady states involved in the process. Not only the output, but also the data source and the intermediary products (at micro or macro level) are given. For each steady state, its conceptual and quality metadata are explored, like its population, variables, frequency and quality.

For more complex statistics, we also develop a business object model. This model provides an overview of the 'real world' objects (units) involved in this statistic, together with their relationships. For example, it shows how a job is related to both a business and a person. It may also be used to relate the input units to the statistical units, such as relating a legal unit to a business unit. Conforming to the RUP, we use an UML class diagram covering the statistically relevant classes and their main attributes.

- **Process model**

The process model provides the overview of the entire process and its relation to the steady states. Generally, the overall process is broken down into sub processes going from one steady state to the next. The process model also shows how the generic business services are used within the process. For the exact functioning of each step within the process references are made to the MAD.

In more complex processes, the production process and the management process and/or (parts of) the design process are captured in the model. It states the (process) information needed to monitor and to adjust the production process.

- **Architectural conformance**

In this appendix, the business analyst may account for deviations from the enterprise architecture. When these deviations are well-founded, the BAD may still be approved.

A BAD of an average redesign project takes one up to several month to draw up and counts 30 to 80 pages. If a preliminary investigation has been conducted, the BAD will generally take less time and effort.

The Central Architecture Board has recently approved a new template for the BAD. Aside from the topics mentioned above, this template covers the requirements (or prior conditions) for the process, the process planning and the initial migration strategy. In an appendix we ask to give the architectural conformance of the design. Does the design comply with the business architecture? And if not, what are the reasons to deviate?

3.2 The Methodological Advisory Document (MAD)

The methodology is captured in the MAD. The methodology in the MAD covers the entire statistical process from data collection to the dissemination of the final product and defines the actual operation of each step of the process. Examples of methodological themes are:

- Sample surveys
- Questionnaire design
- Non-response correction
- Imputation
- Correction for seasonal variation
- Disclosure prevention

The MAD not only states the methods chosen, but also why they have been selected and how they are applied: the parameterisation.

Since the MAD has been introduced more recently, there is as yet no template or guideline concerning its content. However, further development of the MAD has been put into motion.

3.3 Software Architecture Document (SAD)

In agreement with the RUP, we use Kruchten's⁶ 4+1 view model for the SAD. Special attention is paid to the reuse of existing tools and either the use or the development of new but reusable software services. Cost efficiency is

⁶<http://www.ibm.com/developerworks/wireless/library/wi-arch11/>

considered to be very important and the development of custom-made software is to be avoided, wherever COTS⁷ tools or reusable services are available. A typical, finished SAD has around 40 to 70 pages.

Like the BAD, the SAD also contains an appendix named ‘Architectural conformance’. A template for the SAD is available.

4. Reflections and (future) developments

In this section, we evaluate the current redesign process and mention new developments aiming to improve or expand the redesign process.

4.1 Connection between statistical design and IT development

The general redesign process in section 2 suggests a smooth connection from statistical design to IT development. However, the RUP does not explicitly relate the Vision document (or other requirement documents) to process designs. Instead, the Vision document generally elaborates the stakeholders and their needs before identifying the features of the IT system. We explicitly trace the features and use cases to the steps in the process model, indicating the IT support required (or wanted) to perform each step in the process. Doing so, we can trace all functional requirements back to the process.

4.2 Iterative development versus the waterfall model

It is easy to see there is a certain tension between the iterative⁸ nature of the RUP and the entire redesign process on the one hand and the sequential approval of documents (project proposals, BADs, MADs and SADs) by central authorities on the other. For example: if experiments with software in the software development phase show the proposed process model leads to poor performance on large datasets, there will be a tendency to increase performance by using more powerful hardware instead of re-thinking the process model itself, since the process model has been formally approved.

Understandably, project managers and steering committees are weary of re-submitting previously approved documents and business analysts, information analysts and software architects are therefore hesitant to advise them to do so. Whether or not this fear is justified, is largely irrelevant in practice. The simple fact that a formal approval process is necessarily a hurdle and is perceived as such, leads to potentially sub-optimal solutions.

If these hurdles are such that backtracking is made impossible (either on purpose or psychologically), the development process will resemble the so-called waterfall model⁹, which has been criticised extensively. The multi-disciplinary redesign team is an attempt to alleviate the problems associated with this model, as is the preliminary investigation. Still, backtracking and re-thinking decisions is hard and requires insight in the statistical process as well as in software development.

4.3 Standard process steps

A research project concerning the possibilities of Standard Process Steps (SPS) has been started. The basic idea of SPS is to design and develop prefabricated building blocks that can be used when designing a statistic. These building blocks cover one or more steps within the statistical process and provide a standard methodological solution. The business analyst redesigning a statistic ideally combines several of these building blocks and configures them to make them suitable for the particular statistic. The aim of SPS is to reduce the time and costs for (re)design of statistical processes as well as their maintenance. For further information, we refer to Renssen *et al.*, (2009) and Hofman, Camstra and Renssen (2011).

⁷Common of the Shelf: shrink-wrap software readily available; requires only configuration to specific implementation.

⁸http://en.wikipedia.org/wiki/Iterative_and_incremental_development

⁹http://en.wikipedia.org/wiki/Waterfall_model

4.4 Differentiation of the redesign approach

Within Statistics Netherlands, the differentiation of the redesign approach for different kinds of projects is a major issue. Here are two questions: 1) ‘How to distinguish different types of redesign projects?’ and 2) ‘How to tailor the general approach to individual redesign projects?’

The Central Portfolio Board is developing a classification of the redesign projects. The idea is to distinguish three kinds of projects, based on their size and the importance of the statistic. Nevertheless, the exact criteria for the classification are still under discussion. Another discussion in this respect is whether the decentralised development, projects staffed locally using standard tools, should use this more formal approach.

The second question has not been answered yet. Some people plead for reducing the number of documents to be made, whilst others would rather reduce the level of detail while holding on to all documents.

4.5 Process assurance

As part of the quality assurance system which is mandatory for (some) governmental agencies, the information security of each statistical process needs to be described and a risk assessment indicating possible countermeasures has to be made.

Although the process description for the process assurance is quite similar to that of the process design, these used to be two different process descriptions. This was partly due to the fact that different tools were used for process design and assurance. SN has now switched to one tool (Mavim) for both uses and has started a project to further align the two process descriptions. The aim is to design a process only once, to expand this design with the information needed for the risk management and to keep the process description up-to-date when minor revisions occur. Doing so, we realise up-to-date, detailed process descriptions of all statistical processes during their life cycle.

References

- Braaksma, B. (2009), “Redesigning a Statistical Institute: The Dutch case”, *Proceedings of MSP2009, workshop on Modernisation of Statistics Production 2009*.
- Hofman, F. and B. Leerintveld (2010), “Redesign at Statistics Netherlands”, *Proceedings of Statistics Canada Symposium 2011*, Ottawa, Canada.
- Hofman, F., Camstra, A. and R. Renssen (2011), “Standardisation of Processes”, *Proceedings of World Statistics Congress (ISI2011)*, Dublin, Ireland.
- Renssen, R., Morren, M., Camstra, A. and T. Gelsema (2009), “Standard processes”, unpublished report, The Hague, The Netherlands: Statistics Netherlands.
- Van der Veen, G. (2007), “Changing Statistics Netherlands: driving forces for changing Dutch statistics”, paper presented at the Seminar on the Evolution of National Statistical Systems, New York, USA.
- Ypma, W.F.H. and C. Zeelenberg (2007), “Counting on Statistics; Statistics Netherlands’ Modernization Program”, paper presented at the Seminar on Increasing the Efficiency and Productivity of Statistical Offices at the plenary session of the Conference of European Statisticians, Geneva, Switzerland.

Progress towards standardisation at Statistics New Zealand

John Lopdell and Gary Dunnet¹

Abstract

Statistics New Zealand employs standard concepts in its high level survey designs, and has a number of common, but not standard, processes and methods. However the infrastructure to support these is unique to each output. Our 'standardisation roadmap' provides a pathway towards increased standardisation of our methods, processes, data management and technology.

Several initiatives are underway to progress standardisation within Statistics New Zealand. One key initiative is the development of 'platforms' based on common infrastructure and survey 'clusters'. Five platforms are being developed to support the various stages of the generic business process model (gBPM). There is a Household Survey platform for cross-sectional social statistics, a Business and Economic Statistics (BES_t) platform for micro-economic statistics, a CRM-based platform for data collection, a platform for macro-economic statistics, and an OECD stats based platform for data dissemination. In addition to gains in efficiency, these platforms will increase the coherence of our statistical outputs, and facilitate increased use and analysis of the data.

This paper describes the 'platform' approach being developed in Statistics New Zealand, progress to date in integrating collections on to the platforms, and future work, with the primary focus being the Household Survey platform and BES_t platform. The paper presents (1) the principles applied in developing the platforms, (2) the approach used in the development of the platforms, (3) the requirements and functionalities for methods/tools in these platforms (*e.g.*, statistical data editing and imputation), and (4) the actual development phases involved.

¹John Lopdell and Gary Dunnet, Statistics New Zealand, New Zealand.

Development of a common frame for household surveys at Statistics Canada

Larry MacNabb, Martin St-Pierre and Marco Grenier¹

Abstract

In order to meet the high quality standards established by its statistical programs, the Corporate Business Architecture (CBA) initiative, implemented in 2009, is addressing challenges that Statistics Canada must overcome. The goal of the CBA is to improve organizational efficiency and the robustness of systems and processes, while also accelerating the execution of new projects and programs. One of the CBA's recommendations is to create a common frame for household surveys and to standardize the processes involved in its creation and maintenance. This common frame will be based on the Address Register (AR), a dwelling frame already used for the census and some household surveys. We will link both contact and socio-demographic information from the census and various administrative sources to the AR. To do this, standardization and validation tools will be developed to process the various data files. The use of this common frame will result in efficiency gains by, among other things, eliminating the duplication of tasks between different surveys, reducing collection costs, and developing more efficient sample designs. In addition, the excellent coverage provided by this frame, combined with a multi-modal collection strategy, will offer alternatives to random digit dialling surveys. This presentation begins with the history and justification of the project. Then, we will examine the development of various functions of the common frame, particularly the standardization tools and the steps involved in processing the sources of data integrated into the frame, and we will provide an overview of the potential uses of this frame for household surveys.

Key Words: Corporate Business Architecture; Survey frame; Address register; Household surveys.

1. Statistics Canada Corporate Business Architecture (CBA)

In April 2009, a Statistics Canada task force produced a report outlining the potential of achieving a 5% savings in Statistics Canada's (STC) operational budget through an improved corporate business architecture. In response to this report, STC formally launched the Corporate Business Architecture Initiative. Led by a task force of senior managers and reporting directly to STC's Policy Committee, the initiative had three main objectives:

1. Generate a harvestable efficiency on ongoing operational costs of 5% within 5 years;
2. Develop a set of reduced and unduplicated systems and processes that are properly maintained and documented;
3. Generate improved responsiveness in terms of the delivery of new statistical programs.

In order to achieve these goals, several key CBA principles were adopted across the agency. The main principles brought forward were as follows:

- Decisions regarding processes to be optimized globally rather than locally;
- Use of corporate services were to be optimized, and in many instances, become mandatory;
- Existing business processes and computer systems were to be reused when possible;
- Use of corporate business applications and tools was to be encouraged and;
- Processes should be developed with the intent of generating and using metadata to document and drive (control) the process.

¹Larry MacNabb, Larry.MacNabb@statcan.gc.ca; Martin St-Pierre, Martin.St-Pierre@statcan.gc.ca; Marco Grenier, Marco.Grenier@statcan.gc.ca, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada K1A 0T6.

2. Common Frame Initiative

2.1 Principles

In the current collection environment it is becoming increasingly difficult to design sampling strategies which will not only ensure adequate representation in household surveys but also provide efficient means of contacting survey respondents. To that end a set of methodologies is required to address the need for both sophisticated sample designs and multiple vectors for contacting respondents. Link et al. (2009) believe that Address Based Sampling (ABS) is the sampling base upon which such a set of methodologies can be built, providing a stable sampling base, a rich source of characteristic and geographic data for facilitating sophisticated designs, and an opportunity to utilize multiple modes for contacting and conducting surveys with households.

Statistics Canada launched the Common Frame project with the overall objectives of providing a means of addressing the challenges associated with the design of unbiased sampling strategies and, at the same time, providing a foundation for supporting multiple modes of collection. The overarching vision of the project was to develop one address based frame to be utilized by the census and most major household surveys within Statistics Canada. Much of the infrastructure to achieve this already existed to support the Address Register (AR), which was targeted as the foundation upon which the remaining common frame infrastructure would be based.

Another objective associated with the project involved the centralization of all frame activities within one functional unit. This included the standardization of administrative files, storage of files, frame development and maintenance and coordination of field listing operations for the purposes of managing the quality of the common frame. Wherever possible the project was to ensure the harmonization of processes and use of common tools in relation to the standardization of addresses and the treatment of telephone numbers as extracted from administrative sources.

2.2 Description and Roles of the Address Register

The Address Register (AR) is a list of addresses covering the majority of private and collective dwellings in Canada. Prior to the 2011 Census, the AR contained approximately 15 million addresses. The AR stores addresses based on two concepts: location addresses and mailing addresses. The location address is used to physically locate the dwelling in the field. Location addresses can be broadly categorized into civic style or non-civic style, which refers to presence or absence of a unique house number on the dwelling. The mailing address is the address used by the Canada Post Corporation (CPC) to deliver mail to that address and can be different from the location address of the same dwelling in Canada. The AR must maintain a mailing address for each location address of a dwelling in order to carry out the mailing out of census questionnaires. However, when the mailing address is non-civic (*e.g.*, a post office box), it is not available in the AR. For this type of dwelling, only the location address is available. Additional information on the various address concepts (location, civic or mailing) is available in an internal report by McClean and Charland (2011). Following the 2006 Census, the AR was able to determine a civic style address for more than 95% of dwellings in Canada and a mailing address for almost 87% of dwellings in Canada. The proportion of civic and mailing addresses is much higher in urban areas than rural areas. In fact, in urban areas, the addresses are virtually all civic style with an associated mailing address.

Each dwelling covered by the AR is also linked, through its address, to the statistical or administrative geographic hierarchy. Statistics Canada jointly manages a national street network file with Elections Canada called the National Geographic Database (NGD). Statistics Canada has developed address linkage software that links or ‘geocodes’ addresses to the block-faces found in the NGD. Dwellings that are geocoded to a block-face are by extension coded to a specific block. Blocks themselves are rolled up to higher level geographies such as dissemination areas (DA), census tracts (CT) and census subdivisions (CSD). These entities are part of the statistical geographic hierarchy used by Statistics Canada for disseminating census information.

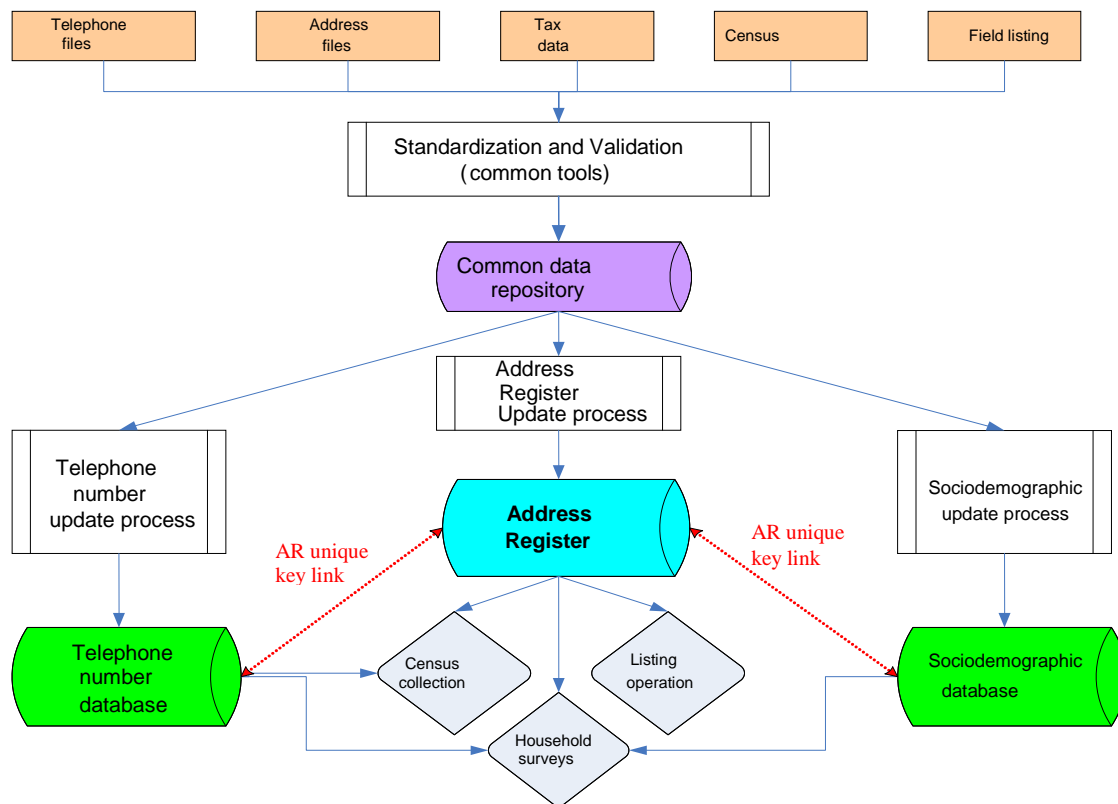
The AR is updated quarterly using data processing from administrative sources and by field verification. The purpose of updating the AR from administrative sources is to identify new residential addresses. Addresses from these sources are matched quarterly to the existing AR. Currently, the administrative sources used are:

- Telephone company billing files
- Telephone directory files
- Data files from the Canada Revenue Agency
- Canada Mortgage and Housing Corporation Starts and Completions Survey
- Canada Post Corporation Points of Call
- Various municipal sources

A field verification activity known as “listing” is also used to update the Address Register. Field staff is provided with lists of addresses for a given small area (listing unit) and a map identifying the listing blocks. Field staff then verifies the list of addresses, adding and deleting addresses, if necessary, as well as verifying the block numbers. Listing is an ongoing activity and a set of listing units is provided to field staff every quarter.

Until now, the primary role of the AR has been to provide an address database for Canada’s Census of Population, which is conducted every five years. Since 2006, the census collection method has been a mail-out of questionnaires to a high percentage of dwellings. The AR was used to create the mail-out list. For the 2011 Census, a mail-out method was used in areas covering approximately 80% of the dwellings in Canada. The AR is also used as input to the Labour Force Survey (LFS) area frame. That frame is also used for other household surveys and consists of a list of clusters representing small geographic areas. For collection in the LFS and other surveys, the AR provides the initial dwelling list for most sampled clusters. A complete dwelling listing is needed for other clusters a few months before the collection. This listing activity is also used to update the AR. Lastly, the AR is used by the Canadian Health Measures Survey (CHMS) to complete their dwelling sampling frame.

Figure 2.2-1
Common frame quarterly updating process



With the new common frame initiative for household surveys, the role of the AR will be considerably expanded in the next few years. Although its primary role will remain that of providing an address base for the census questionnaire mail-out, the AR will become the common frame for most household surveys. Household surveys, and particularly random telephone number surveys, may be able to use an address-based sampling methodology to improve their coverage. If certain surveys continue to use the telephone as their primary collection mode, the AR will have to be permanently linked to the telephone numbers drawn from other administrative sources. In addition, to enhance the efficiency of sample designs of household surveys using the AR, socio-demographic information from census and tax data files will be linked to the AR dwellings. Under the new initiative, the AR will continue to be updated quarterly; and will provide, at the same frequency, a list representing the dwelling universe from which the methodologists responsible for household surveys can select samples. Lastly, the AR will become the initial and final dwelling source for most clusters for the LFS area frame.

Figure 2.2-1 provides a general overview of the AR quarterly updating process, including the new functionalities of the common frame. The different data sources will pass through a common standardization and validation module and then standardized files will be retained in a common data repository (CDR). The CDR will be the source of three distinct processes. First, in the centre of the figure is the AR updating process, which will be similar to the current process. Then, two new processes will be developed to create a telephone number database and a socio-demographic information base. These two bases will be linked to the AR using the unique AR key.

3. Common Frame Functions

3.1 Common Processes and Tools

As the previous figure shows, all data files used to update the common frame, namely the AR, will have to be integrated in a common standardization and validation module. The module will, among other operations, standardize addresses and names, and validate telephone numbers. Some of these processes, or common tools, have been developed from existing Statistics Canada processes, while others require the development of new ones. In addition, a common address linking process will have to be developed to allow matching between files or with the AR. This subsection describes four common tools that will be used to build the common frame.

3.1.1 Address Standardization

Addresses from different data sources arrive in the format of that source. They need to be standardized, using a common process, in order to find a common format for all sources and to improve address matching between those sources. A standardized address process was already in place to handle data from administrative sources used to update the AR. All that was necessary was to review the processing steps in that process and make changes and improvements to meet the new requirements of the common frame. The process also had to be optimized so that it would be generic and simpler to use.

The address standardization tool parses the address from the source files into several fields such as municipality, street name, civic number, apartment number, street type, *etc.* Each address field is then standardized to a set of standard codes. This tool also links addresses to the Canada Post Corporation (CPC) base to obtain the corresponding mailing addresses (including postal code), the format of which may be different from that of the original address. Given that the CPC base is updated monthly, it is important that addresses from the different administrative files are linked to the CPC base each quarter, even if the administrative file has not been updated since the last quarter. As mentioned in Section 2.2, the last step is to link the addresses to the NGD so that they can be associated with a block-face, and by extension, with a specific block. Once this has been done, the addresses can be linked to higher level geographies such as DAs, CTs and CSDs.

3.1.2 Address Matching

A linkage process is absolutely essential when connecting the addresses in administrative data files with the addresses in the AR. This type of process already exists and was developed to enable the identification of addresses in administrative data sources that are not in the AR (growth identification process). Consequently, a common

address linkage module can be developed from the current process with only a few minor improvements. The objective is to make this process generic and accessible to users interested in linking addresses from any set of files containing standardized addresses.

3.1.3 Telephone Number Validation

A new tool is being developed to validate telephone numbers from administrative telephone number sources. The tool's main goal is to identify telephone numbers that are valid and should be retained in the new telephone number database. A number is deemed valid if the area code and prefix (ACP = first six digits of the number) are valid in Canada. The source used to determine the list of valid ACPs is the NPA-NXX file, which is maintained by the Canadian Numbering Administrator (CAN). This tool is also used to link an area code to a telephone number without an area code or with an invalid code. The address (province and postal code) is used to identify the area code associated with the number. However, this association is not always possible since there is sometimes more than one possible value. Lastly, the tool links complementary information to telephone numbers. For example, using various external files, it may be possible to identify whether a telephone number is connected to a land line or a cell phone.

3.1.4 Name Standardization

Several of the administrative data files used to create the common frame contain the names of people. To increase the data linkage success rate between these various administrative sources, there are benefits to standardizing the names from these sources using the same process for all sources. To this end, a name standardization module will be developed and used when processing data sources. As a number of name standardization processes are already in place at Statistics Canada, the task will involve choosing and implementing one of them for the common frame. A few modifications and improvements will be required to effectively handle all of the administrative data sources.

3.2 Telephone Number Database

One very important component of the new common frame is the creation and updating of a telephone number database that will combine the information from all administrative data files containing telephone numbers and link it to the AR. The creation of this database will address several objectives. First, the telephone numbers from all sources must be linked to the AR, so that surveys selecting their samples through the AR will have access to telephone numbers with which to contact the households living in the selected units. Not all telephone number sources were linked to the AR in the past. The objective is not merely to create a linkage between telephone numbers and addresses; but also to try to link the best numbers possible and to allow for more than one number per address, if possible, to generate efficiencies at collection. Another objective of this database is to create a separate, exhaustive list frame of telephone numbers, which could be used directly as a sampling frame, notably for dual-frame methodology. In certain rural areas, it is difficult to associate telephone numbers with AR addresses given that non-civic style mailing addresses (*e.g.*, post office box or rural route) are often used in telephone number sources. Creating this telephone number list frame would improve dwelling coverage and provide telephone surveys with an alternative sampling and collection strategy in these areas. Lastly, this frame will link quality and other indicators to each telephone number, such as the updating date, the list of sources that include the number or the type of service associated with the number (land line or cellular).

Several sources will be used to create and update the telephone number database. One of those sources is the monthly Info-Direct file, which contains the telephone numbers of the directories of the principal telephone service providers in Canada (land lines only). We will also be using the numbers from service provider billing files normally available each quarter. Also included will be the telephone numbers provided on the census short form questionnaires and on different tax data files. In the future, if we obtain access to other telephone number sources, we will include them when the database is updated. We also plan to include in the database the telephone numbers provided by household survey respondents in part so that we can avoid contacting these respondents again during subsequent surveys, thereby reducing their response burden.

The telephone number database will contain only one entry per telephone number and will retain, as much as possible, the best address for each number. Here is a summary of the steps to build that database. First, as indicated on the diagram in Section 2.2, all telephone number sources will have to have been integrated into the common

standardization and validation module and then added to the CDR. The first step involves reading the files in the CDR and excluding invalid or non-residential telephone numbers. Next, duplicates in each source will be eliminated in order to retain only one entry per telephone number. If a number is associated with more than one address in a single source, the best address is retained based on address quality criteria. If it is impossible to choose the best address, only the telephone number is retained at this stage (the address is deleted). All sources are then combined by matching telephone numbers. When a telephone number appears in more than one source, a second duplication elimination process is carried out to select the best address for each number. Besides address quality criteria, additional criteria, such as the reliability of the source and source reference date, are used at this stage to select the best address associated with the given number. After this step, the list frame is complete and contains one entry per telephone number. In the process of selecting the best address, it is important to point out that we may retain the best civic-style and non-civic style address (= mailing address) for the same number. Lastly, the telephone numbers in the database are linked to the AR. This linkage is done automatically since the telephone number sources will have already been linked to the AR during the quarterly update. Once the links have been established, the next step is to determine, for each AR address, the “best” telephone number or numbers to use at collection. This involves establishing a priority order for the numbers (*e.g.*, first number to call, second number to call, *etc.*). This order takes into account the reliability and reference date of the sources, as well as the type of telephone line, giving priority to land lines over cellular numbers. The entire telephone database building and AR linkage process will take place each quarter. This means that, each quarter, the database will contain only the telephone numbers that were found in the most recent versions of the telephone number sources. Telephone numbers no longer appearing in any of the sources will not be retained.

3.3 Socio-demographic Information Database

The common frame will be composed of another database containing household socio-demographic information. To this end, a new process must be developed to build this base and associate the socio-demographic information from it to the dwellings in the AR. This database will address two primary needs. First, the database’s information will be used as auxiliary information for household surveys, perhaps in the sampling design phase to create dwelling clusters, to create dwelling strata or even to determine the sample allocation. The information might also be used in the survey estimation phase to improve non-response adjustments or for the imputation of certain variables (*e.g.*, household income). The second need to be addressed by building this database is to centralize the processing and linkage of files containing socio-demographic information by taking advantage of the common tools described in Section 3.1. In recent years, several surveys or programs have used and processed these files separately, which has resulted in duplication of efforts and a corporate-wide loss of efficiency. Including a socio-demographic database in the common frame will address the objectives of the CBA.

Statistics Canada has several sources of socio-demographic data. The main source is the census, which is conducted every five years. The short census questionnaire contains information on household members such as address of residence, name, birth date, sex, language or languages spoken, and the relationships among the people in the household. The last census took place in May 2011 and the data have been available to users since fall 2011. In 2011, the detailed content of the census long questionnaire was collected for the first time using the National Household Survey (NHS) and was administered to approximately one-third of households on a voluntary basis. The detailed socio-demographic information from the NHS is also a useful source of information for building the common frame. Another source of socio-demographic data is Canadian tax data; notably, the annual income tax returns known as T1 files. These files contain information about the tax filers such as place of residence, name, birth date, sex, and marital status, as well as income information. The T1 files and other administrative data sources are used to create the internal T1 Family File (T1FF), which consists of an enhanced and more complete version of the T1. The T1FF contains information not only on the tax filers but also on their spouses and non-filing children; and links individuals to each other based on the census family concept. The T1FF provides the composition of households in a more complete and accurate manner than the T1. However, the T1FF is only available five to six months later.

The socio-demographic database will be created using these data and will include mainly two-level information: at the dwelling level and at the census dissemination area (DA) level. Using data from the short questionnaire and the T1FF, dwelling level information (address) on each dwelling will be retained in a file in the AR. The dwelling level file will include household information (size, type, income, *etc.*) and information on the characteristics of the people living in the dwelling (birth date, sex, language, income, *etc.*). The first version of this file will be built using data from the 2011 Census along with income data from the 2010 T1FF. There are two options for linking the 2010 T1FF data to the 2011 Census. Linkage could be at the level of individuals, offering a higher matching rate and better

quality but requiring much more effort. It could also be only by address matching, requiring much less effort but reducing the quality of the matching results. The 2010 T1FF data could also be used to complete the missing information on non-responding dwellings or those presumed vacant at the time of the census. The dwelling-level file will then be updated annually using the last available T1FF. The T1FF will be linked to the socio-demographic database by address only. If a link is established in the T1FF for any given address, then all the socio-demographic information will be replaced by the information available in the T1FF. However, if no T1FF link is established, the previous information will be retained in the database. This process ensures that, for each AR address, the socio-demographic database will contain the most recent information on the household and the people residing there.

The second file in the socio-demographic database will contain information at the DA level. The DAs represent a small area composed of one or more neighbouring dissemination blocks and approximately 400 to 700 residents. All of Canada is divided into DAs. The data from the 2011 Census and NHS will be used to aggregate the household and individual information at the DA level. It is more prudent to aggregate the detailed NHS information at the DA level since its information is only available for a subset of households (about 20%). This file will contain variables such as the number of people and households in the DA, median income, median level of education, proportion of immigrants and others. This data will be useful to household surveys for building dwelling clusters or strata and for non-response modelling.

4. Conclusion

In recent years, the quality of the AR maintained by Statistics Canada for the Census of Population has improved considerably and several processing and administrative data linkage procedures have been developed. The need for another frame for household surveys based on random telephone numbers, combined with the objectives of the CBA initiative, led to the decision to use the AR as a common dwelling frame for the household surveys program. These surveys will take advantage of the existing infrastructure to update the AR and all its administrative data processing procedures. This will allow these surveys to migrate to an address-based sampling methodology and to improve their coverage of the population. Using a common frame will also facilitate the coordination of samples among household surveys. The process of building a telephone number list frame and its linkage to the AR will allow for a more efficient use of telephone number information from various data sources. It should result in better telephone contact rates for targeted households. In addition, this telephone number database will represent an enhanced list frame that can be used to select samples directly, notably for dual-frame sampling methodology. Building a socio-demographic database linked to the AR will enrich the common frame and help develop more efficient sampling designs. Lastly, introducing a common frame meets the objectives of the CBA for standardization and centralization of processing, storage and data access processes.

Acknowledgements

The authors would like to thank Windie Gagné, Jocelyne Marion and Jean-Louis Tambay for their contribution to this project and Yves Lafortune and Denis Poulin for their comments.

References

- Link, M., Daily, G., Shuttles, C., Yancey, T., Burks, A. and C. Bourquin (2009), "Building a New Foundation: Transitioning to Address Based Sampling After Nearly 30 Years of RDD", *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 5654-5665.
- McClellan, K. and K. Charland (2011), "Address Register: Future State after the 2011 Census of Population", unpublished report, Ottawa, Canada: Statistics Canada.

SESSION 2B
SAMPLING AND ESTIMATION

Differential design effects in school-based samples

Caroline Dahmen and Marek Fuchs¹

Abstract

Several large-scale school-based surveys make use of cluster samples and can therefore suffer from design effects, having an impact on the effective sample size. In this paper, we will demonstrate that design effects differ considerably across various subgroups of the sample like different school forms and grades. In addition, the school-level and class-level design effects differ in magnitude across various types of schools within the same sample. Thus, no uniform strategy is suitable for most school-based surveys and one has to consider differential design effects when planning a study.

Key Words: Complex samples; Design effect; School-based surveys; Effective net sample size.

1. Introduction

Clustered samples are often used in social sciences, but while they are more cost-effective than simple random samples the resulting samples may suffer from design effects (*deff*). The composition of clusters is often based on natural groups like classes or dwellings. Thus, elements from the same group or cluster tend to be more similar to each other than elements from a simple random sample (SRS) drawn from the same population. For example elements may resemble each other with respect to socio-demographic characteristics or elements may be exposed to the same environmental effects. These similarities result in a certain level of homogeneity characterising the group or cluster. Since this affects variance estimation, the sample size for clustered samples has to be increased in order to yield the same precisions like a SRS. This is necessary because we obtain less new information by surveying more elements from the same group or cluster (Cornfield, 1951; Groves *et al.*, 2009; Kish, 1965; Lohr, 1999).

For example school achievement surveys like PISA (OECD, 2009) or TIMSS & PIRLS (2007) apply two-stage cluster samples, by sampling multiple students from randomly selected schools or multiple classes within each school. To compensate for the complex design, the design effect is included in calculation of the necessary sample size. This adjustment is required because the statistical value of a clustered sample is lower than the value of a simple random sample of the same size.

In principle, the design effect indicates by what factor the variance of the estimate in a clustered sample is underestimated in comparison to simple random sample. Thus, “the design effect provides a measure of the precision gained or lost by use of the more complicated design instead of an SRS” (Lohr, 1999, page 309). Hence, the design effect indicates by how much the sample must be increased to reach the desired precision.

When planning a clustered survey, the actual design effects are often unknown and must be estimated. The *deffs* for clustered samples depend on the individual characteristics of the survey, like homogeneity of the clusters or size of the clusters and can only be estimated based on the data produced in a certain survey. In addition, design effects are not identical for all estimates derived from the same sample (Kish, 1995) and they can also differ for subgroups within one sample (Verma and Lê, 1996). Despite this fact, in most surveys a constant overall design effect is applied when calculating the necessary sample size. In certain cases this might not be appropriate. It might lead to a bias in the estimation of the variance or to higher costs (Kish, 1965).

To develop a sample that is both cost-efficient and precise careful estimation of expected *deffs* in a sample is advisable. However, with the actual data not yet available, design effects cannot be determined during the planning phase of a survey. To come up with a reliable estimator of *deff*, researchers planning for a survey have two options to

¹Caroline Dahmen, Darmstadt University of Technology - Institute of Sociology, Residenzschloss, 64283 Darmstadt, Germany; Marek Fuchs, Darmstadt University of Technology - Institute of Sociology, Residenzschloss, 64283 Darmstadt, Germany, fuchs@ifs.tu-darmstadt.de.

approximate the values of *deff*: either conduct a pilot study to retrieve the necessary data to estimate *deffs* or search for an already existing study similar in design, target group and variables.

In our case we were faced with the problem of designing a sample survey in Germany for a large scale comparative study among vocational students in seven European countries. Since no complete list of all German vocational students existed the survey was planned as a school-based cluster sample from the outset (students in classes in schools). In order to come up with proper estimates of *deff* we used an existing school-based survey in the same population. During the analysis two main research questions arose: “To what extent do *deffs* differ across subgroups within the target population?” and “If differences within these subgroups exist, what are the driving factors for differences in *deff*?”

2. Differential Design Effects

2.1 Data & Method

The data used for the analysis derived from the Violence at Schools study 2010 (Fuchs, 2009). This large scale school-based trend study is conducted every five years in Bavaria, Germany. The sample size in 2010 was approximately 6,000 students clustered in a total of 173 schools. Typically two classes per school were included in the sample. Within this sample, students from four different school types existing in Germany were included.

Germany features a tracked secondary school system with students separated at age ten based on their school achievements in primary school. Lower secondary school, intermediate secondary school and the upper secondary school are the major school forms existing, catering each for a specific group of students (*e.g.*, students with very good grades at primary school are admitted into upper secondary school). Further, the duration of the education differs between the school forms. While students at lower secondary school finish their education after grade nine or ten, students at upper secondary schools finish their secondary education after twelve years (afore 13 years) of schooling. Next to the three forms of general secondary schools, vocational schools were part of the sample. After having finished one of the three general schools, students can enter vocational education and training, or start studying at a university as long as they hold the necessary entrance diploma. Young people striving for a vocational education and training will enrol in a vocational school either full-time or part-time depending on the training program.

Kish suggested estimating the design effect by dividing the variance of an estimator of a complex design by the variance of the estimator of an SRS of the same size (Kish, 1965). Using the formula given below we estimated *deff* on school level and on class level.

$$Deff = 1 + \rho(B-1) \quad (2.1-1)$$

In formula 2.1-1 (Kish, 1965) the intraclass correlation coefficient (ρ), the measure of homogeneity of a cluster, is included as well as the size of the cluster (B). Therefore, effects that occur due to the clustered design are reflected.

While the cluster size can easily be determined, estimating the intraclass correlation is more complex. We applied a three-level regression model to decompose the variance that can be explained on each level present in our design. In formula 2.1-2, the intraclass correlation is calculated by taking the between and within school level into account. Here, between-school variance is divided by the sum of the between and within school variance. While σ^2_B indicates the between-school variance σ^2_w represents the within-school variance (Foy, 2004).

$$\rho = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_w} \quad (2.1-2)$$

The range of values of ρ is zero to one, with zero stating total heterogeneity of the sample, *e.g.*, a simple random sample, while one indicates that the sample is completely homogeneous.

In our analyses the formula was slightly modified because we included the variance that could be explained on the individual level (σ^2_i) as well (see 2.1-3 and 2.1-4). This is necessary due to the three levels (*e.g.*, school, class, student) present in our sample. Here σ^2_i indicates the variance that can be explained on the student level.

$$\rho_{school} = \frac{\sigma^2_B}{\sigma^2_B + \sigma^2_W + \sigma^2_I} \quad (2.1-3)$$

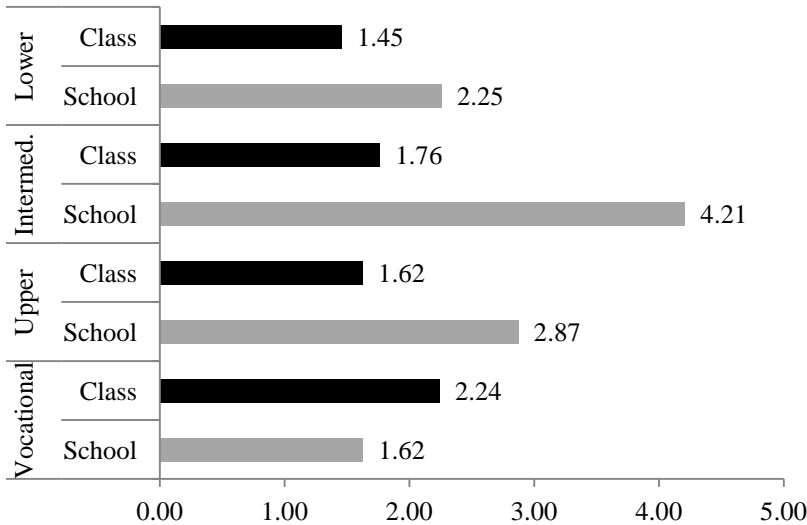
$$\rho_{class} = \frac{\sigma^2_W}{\sigma^2_B + \sigma^2_W + \sigma^2_I} \quad (2.1-4)$$

Following the equations 2.1-3 and 2.1-4 recommended by Hox (2002) we were able to calculate ρ and hence *deff* separately on school and class level. Calculating values for 51 variables separately for school and class level, for all grades and all school forms resulted in a total of 816 estimates. These estimates delivered the base material for our further analysis.

2.2 Results

The value given next to the bars in Figure 2.2-1 indicates the design effect for each specific school form and level. Here, the mean *deff* of all 51 variables was calculated.

Figure 2.2-1
Mean *deffs* separately for school forms on school and class level



The grey bars in Figure 2.2-1 represent the design effects on school level. When comparing these values we could see large differences across the school types. Here, the effect on the school level at the intermediate schools was more than twice the size than at vocational schools. Even when comparing only the general secondary schools, design effects differed considerably. In contrast, differences on class level, the black bars, were smaller in size. Also, class level design effects were generally smaller compared to school level design effects.

Vocational schools differed from the general secondary schools, which all showed the same pattern, namely a similarity in size of the class *deffs* and varying school *deffs* which were substantially larger than their counterparts on the class level. At vocational schools though, the class level *deff* was larger than the school level *deff*. Further the vocational schools showed the smallest mean *deff* of all school forms. We assumed that this can be explained by the fact that in Germany students in each class at vocational schools mostly train for the same occupation, and that we find classes of different occupations within one school (*e.g.*, one school can offer programs for cooks, chemical production technician and hairdressers). So the vocational classes themselves are more homogeneous than the vocational schools.

Because the Violence in Schools study also provided the data from different grades, we were able to determine differences in the magnitude of the design effects for different age groups within one school type and therefore obtained even more precise estimates for the design effects. We therefore categorized the students by junior, middle and senior grade.

Table 2.2-2
Average school level *deff* of different school forms by grade

		Grades			Total
		Junior (1)	Middle (2)	Senior (3)	
School form	Lower (a)	1.82 ^{(a2)(b1)}	2.46 ^{(a1)(b2)}		2.25
	Intermediate (b)	3.48 ^(a1)	4.48 ^{(a2)(c2)}		4.21
	Upper (c)	2.82	2.34 ^(b2)	2.71 ^(d3)	2.87
	Vocational (d)			1.62 ^(c3)	1.62

Note: Superscripted combinations of letters and digits denote significant differences ($p < .05$) to other values in the same row or column using the Scheffé post-hoc test.

As shown in Table 2.2-2 we determined differences in the design effects within one school type between two age groups. However, only the *deffs* at lower secondary schools differed statistically significantly between the age groups indicated by the superscripts. While *deff* increased from junior to middle grade at the lower and the intermediate school, the upper secondary school showed smaller effects for the oldest age group compared to the junior group. Thus the design effects at the school level and at the class level did not only differ significantly between school types but also across age groups, even though it became statistically significant only for lower secondary schools. Accordingly, we have to consider multiple subgroups within the sample that vary considerably regarding the magnitude of the design effect.

To identify driving factors for differences in *deff* on school level we analysed the components that determine *deff*.

Table 2.2-3
Components determining *deff* of different school types and levels

			Cluster size	Intraclass correlation (ρ)	Mean <i>deff</i>
School form	Lower (a)	Class	16	.029	1.45 ^{(b)(d)}
		School	29	.053	2.25 ^(b)
	Intermediate (b)	Class	23	.034	1.76 ^{(a)(d)}
		School	46	.085	4.21 ^{(a)(c)(d)}
	Upper (c)	Class	20	.034	1.62 ^(d)
		School	39	.068	2.87 ^{(b)(d)}
	Vocational (d)	Class	20	.069	2.24 ^{(a)(b)(c)}
		School	39	.017	1.62 ^{(b)(c)}

Note: with () = $p < .05$ Post-hoc test Scheffé

In Table 2.2-3 the main components, the cluster size and the intraclass correlation (ICC) coefficient, that determine *deff* following the equation 2.1-1 are listed separately for class and school level for four different school forms. It is apparent that schools were very dissimilar with regard to these components. While the class size at lower secondary schools was on average 16 students, this number was substantially higher at intermediate schools amounting to 23. More interesting though, were the differences of the ICC. All in all, the values seemed rather small, with .085 being the largest value. Still, they differed drastically not only between schools on the same level but also within schools on different levels. Especially vocational schools provided a rather small ρ of only 0.017. In the last column, the mean *deffs* for each level and school form are stated again, and we could show that the differences in *deff* between the schools were in most cases significant, indicated by the superscripts. By applying a post-hoc test, we demonstrated that *deffs* on school level at intermediate schools differed highly significantly from all other *deffs* on the same level. Further, values of *deff* differed significantly at least from one other value on the same level.

In a subsequent analysis we assessed the question regarding the driving factors responsible for the differential design effects across subgroups. From our point of view two factors can be held responsible for the differential design effects: the composition of the student body in each class or school with respect to the socio-economic background of the families, and the fact that the students in each class or school spend a lot of time together and may influence each other and thus may develop similar attitudes and behaviours. Since we could not analyse the effect of social interaction in the classroom, we focused on the composition of the classes and schools instead. As shown above homogeneity of clusters differed drastically across sociodemographic characteristics. We assumed that composition of a cluster provides information on the inherent homogeneity.

We assessed six sociodemographic characteristics: The education of the father, the education of the mother, the income status of the family, the community size of the place where the students live. Since we assumed that the composition of a class or school is the key factor, and not individual characteristics of students, we computed the standard deviation of the sociodemographic variables within each class or school. In addition we included percentage of girls (percent deviation from 50%) and the percentage of students with an immigration background in the analyses.

Analyses indicated that school types differ with respect to the composition of their student body; however there were also substantial differences within schools. When comparing the composition properties for the different school types, for example the percentage of students with immigration background, considerable differences occurred. While 33 percent of the students at lower secondary schools (junior grade) possessed an immigration background this number was much smaller at upper secondary schools (only around 20% at junior grade). For the education of the mother it can be determined that all school forms showed a similar standard deviation at junior grade of approximately 0.83. At lower secondary schools though, that value declined to 0.77 (at intermediate secondary school the value declined to 0.71). This means that the student body was slightly more homogeneous in the middle grade than in the junior grade. At upper secondary schools the standard deviation increased from junior to middle grade. Here the student body showed a more pronounced heterogeneity in the older age group.

To determine which characteristics of the composition had a significant effect on the magnitude of the design effects on school level, we calculated hierarchical linear regression models (see Table 2.2-4) using the socio-demographic standard deviation as predictors. We used the general intermediate school form as the reference category, and grade as a control variable.

Table 2.2-4
Hierarchical linear regression for *deffs* on school level (standardized regression coefficients)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Level 1								
Lower school†	-1.84***	-1.84***	-2.15**	-1.72***	-2.28**	-1.12	-1.82***	-0.84
Upper school†	-1.36**	-1.48**	-0.85	-0.84*	-0.58	-1.25	-0.35	-0.67
Vocational school†	-2.36***	-2.71***	-2.74**	-2.58***	-3.11***	-3.19 ⁺	-2.78***	-1.68
Grade		0.24	0.15	0.13	0.04	-0.33	0.12	-0.08
Level 2								
Education father _{SD}			-4.91					
Education mother _{SD}				-7.38*	-8.06*	-6.86 ⁺	-11.12*	-6.48
Income status _{SD}					5.90			
Community size _{SD}						3.81		
Girls % ₁							-0.03	
Immigration % ₂								-5.89
Log likelihood	-931.94	-931.95	-928.84	-927.79	-924.59	-924.51	-929.99	-924.08
n	408	408	408	408	408	408	408	408
Pseudo R ² _{McFadden}	0.00703	0.00702	0.01033	0.01145	0.01486	0.01495	0.00911	0.01541

Note: + p < .1; * p < .05; ** p < .01; *** p < .001, SD = standard deviation, %1 = percentage of deviation from 50 percent of girls in the cluster, %2 = percentage of immigrants in the cluster
† reference category = intermediate school

Our results indicated significant negative effects for all school types relative to the reference category. For most models these effects remained significant, at least for the vocational and the lower secondary schools. When including the sociodemographic variables, the standard deviation of the education of the mother was the only variable that yielded a significant effect (Model 4). When introducing the other variables the strength of this effect declined, except when introducing the percentage of girls (Model 7).

The variable we controlled for in Model 6, the standard deviation size of the community where the students live, led to non-significant results, both the school types and the education of the mother were not significant anymore. The variable “percentage of immigrants” (Model 8) removed the significance of the education completely and in this model we could also see that the school type was not significant any more.

These findings suggest that the sociodemographic homogeneity of the composition of the clusters cannot explain the differences in the *deffs* with one exception, the education of the mother. At this point further research with other variables is necessary to determine whether composition has an influence at all.

2.3 Conclusion

Results from our analyses have affirmed that *deffs* differ extensively for subgroups of a sample. Because the design effect is an estimator of a specific variable instead of a whole sample values for different subgroups can vary distinctively. We showed that in school surveys different school types and grades can be determined as relevant subgroups characterized by differential *deffs*. For school-based samples, which consist of different school forms and grades, it is recommended to include customized *deffs* in the sample size estimation. Further, more research is necessary to determine how these differences come about.

Our assumption, that homogeneity of the subgroups is responsible for different *deffs* could not be confirmed by our analysis. Hence, we were able to determine relevant subgroups of school-based samples, yet the driving factors responsible for differential *deffs* could not be identified.

The sociodemographic variables applied in this paper could not explain the differences with regard to homogeneity. Still, we cannot abandon the assumption that the composition of the student body is responsible for differences in *deff*. Instead, the analysis should continue with other sociodemographic variables. In addition, next to the composition of classes, other sources could be responsible for the varying effects. One could think of the different competences of schools to integrate their students into a class or school.

All in all, it is recommended to abandon the approach of using one overall *deff* when calculating the sample size. Especially in complex designs consisting of several subgroups which differ considerably with regard to the homogeneity of the groups, a ‘one fits all approach’ is not advisable. Here, more flexibility is needed to be able to adjust the sample size in the best way possible, providing both a cost-effective and precise sample. Therefore, the specifics of a survey should be reflected and customized *deffs* should be incorporated in a sampling calculation.

References

- Cornfield, J. (1951), “Modern Methods in the Sampling of Human Populations”, *American Journal of Public Health*, Vol. 41, No. 6, pp. 654-661.
- Foy, P. (2004), *Intraclass Correlation and Variance Components as Population Attributes and Measures of Sampling Efficiency in PIRLS 2001*, Hamburg: IEA Data Processing Center.
- Fuchs, M. (2009), “Impact of school context on violence at schools - A multi-level analysis”, *International Journal on Violence and Schools*, 7(2), pp. 20-42.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J., Singer, E. and Tourangeau, R. (2009), *Survey Methodology*, 2nd ed., New York: Wiley.

- Hox, J. (2002), *Multilevel Analysis: Techniques and applications*, Quantitative Methodology Series, New York: Psychology Press.
- Kish, L. (1965), *Survey Sampling*, New York: Wiley Series.
- Kish, L. (1995), "Methods for Design Effects", *Journal of Official Statistics*, Vol. 11, No.1, pp. 55-77.
- Lohr, S. L. (1999), *Sampling: Design and Analysis*, 2nd ed., Pacific Grove: Duxbury Press, 592 pages.
- OECD (2009), *PISA 2006 Technical Report*, OECD Publishing.
- TIMSS & PIRLS International Study Center (2007), *TIMSS 2007 Technical Report: Chapter 5. Sample Design*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Verma, V. and T. Lê (1996), "An Analysis of Sampling Errors for the Demographic and Health Surveys", *International Statistical Review*, 64(3), pp. 265-294.

Canadian Forces Personnel Surveys: A consideration of weighting and non-response

François Larochelle, Tingting Gou and Irina Goldenberg¹

Abstract

The Director General Military Personnel Research and Analysis (DGMPRA) conducts research to support the development of policies and programs related to the management of Canadian Forces personnel. Much of this research is based on data collected through personnel surveys. This paper reviews DGMPRA's journey in the search for methods of calculating accurate and reliable survey estimates that will be used to provide evidence-based information to enable personnel-related program and policy development. While there are benefits associated with the use of population-based weighting for estimation in an operational research context, this paper discusses the practical challenges associated with producing accurate estimates. Chief among these challenges is survey non-response. This paper discusses the results of a case study undertaken to identify demographic variables associated with non-response and the way ahead for addressing this important challenge.

Key Words: Canadian Forces Your-Say Survey; Non-response; Population-based weighting.

1. Introduction

Effective personnel management is essential to the Canadian Forces (CF) in terms of recruiting, training, preparing, supporting, and honouring military personnel for service to Canada. The Director General Military Personnel Research and Analysis (DGMPRA) within the Department of National Defence (DND) supports personnel management by providing research to enable evidence-based policy and program development. In order to achieve this, personnel surveys of military members are an essential research tool, and every year DGMPRA designs, administers, analyzes, and reports on a variety of surveys.

DGMPRA is a relatively small organization working in an environment where the time available to clean and analyze data is generally limited because the client needs to be informed of top line results within a short time period after the data is collected. The majority of scientists within DGMPRA have a social science background, usually psychology or sociology. This provides them with the skills and knowledge required to work on a broad range of personnel-related research problems, both quantitative and qualitative. While subject-matter expertise in survey development and analysis is available, DGMPRA does not currently have scientists specialized specifically in sampling-related survey methodology, whose official role would be to provide this type of guidance and expertise.

The response rates to personnel surveys administered to CF members are generally below 40% which makes the analysis of data more complex and raises issues regarding the bias of survey estimates. Given the operational research context in which DGMPRA operates, addressing these challenges is not simple since it requires advanced knowledge and expertise in survey methodology, as well as resources. DGMPRA is aware of this reality, and also recognizes the critical role that survey research plays in the organization. Therefore, in recent years greater attention has been paid to the issue of non-response, to the methods used to produce survey results, and to the validity and ability of these results to be generalized. This examination is helping DGMPRA to move forward by developing more effective and scientifically rigorous methods for calculating estimates, and measuring their reliability.

This paper reviews DGMPRA's journey in the search for efficient methods to calculate accurate and reliable survey estimates. While there are benefits associated with the use of population-based weighting over other weighting

¹François Larochelle, Director General Military Personnel Research and Analysis, Department of National Defence, 285 Coventry Road, Ottawa (Ontario), Canada, K1A 1V0; Tingting Gou, PhD in Statistics, Ottawa (Ontario), Canada; Irina Goldenberg PhD, Director General Military Personnel Research and Analysis, Department of National Defence, 285 Coventry Road, Ottawa (Ontario), Canada, K1A 1V0.

methods for estimation in an operational research context, this paper also discusses the practical challenges related to the production of accurate estimates. Finally, the paper presents the results of a case study that was undertaken to identify demographic variables associated with non-response.

2. Basic Notations

N denotes the total size of the target population from which a survey sample is selected, and for each index i in the set $\{1, 2, \dots, N\}$ Y_i denotes the value of a study variable of interest for the i^{th} member in this population. This variable is measured from the survey for a sample of n units, and S_r denotes the set of indexes for the n population units who responded to the survey.

3. Estimation within DGMPRA

3.1. A Consideration of Sample Descriptive Statistics

A simple category of statistics that can be considered for reporting survey results is the ‘sample descriptive statistics.’ These statistics are used to describe the characteristics of the survey respondents, and they include statistics such as the sample median, the sample mode and the sample mean

$$\bar{y} = \frac{\sum_{i \in S_r} Y_i}{n}. \quad (1)$$

Sample descriptive statistics provide information on the sample of respondents and, as long as they are not generalized to a larger population from which the sample was selected, there is no sampling error associated with them. In this context, these statistics are very advantageous for delivering timely and accurate results to the client because they can be calculated without particular attention to or analysis of non-response and coverage errors. In addition, they can be reported without any measure of sampling error, such as the standard deviation or the coefficient of variation.

Although there are advantages to reporting sample descriptive statistics to describe survey respondents, in an operational research context, clients are typically interested in obtaining survey results which accurately describe the target population from which the sample was selected. For this reason, additional efforts must be invested in producing reliable and relevant estimates that can be generalized to the overall population of interest.

3.2. A Consideration of Estimators for the Population Mean

For the surveys conducted within DGMPRA, the unknown population parameter of interest to the clients is most often the population mean

$$\bar{Y} = \frac{\sum_{i=1}^N Y_i}{N}. \quad (2)$$

The sample mean \bar{y} can be used to estimate \bar{Y} but, unlike when it is used as a sample descriptive statistics, there is a sampling error associated with this and the sample mean is also subject to bias in this context. The bias and variance of \bar{y} must therefore be investigated in order to determine whether \bar{y} is an appropriate estimator, or whether another estimator with a smaller bias and/or variance can be calculated within an acceptable budget.

When analyzing the bias of an estimator, three important mechanisms for consideration include the sample selection probabilities, as well as non-response and under-coverage. For example, Cochran (1977) and Biemer & Christ (2008) use simple frameworks for illustrating and analyzing the potential bias of \bar{y} resulting from non-response and

coverage errors. In order to reduce bias resulting from unequal selection probabilities, non-response and under-coverage, survey statisticians have proposed various techniques for calculating weighted estimators of the form,

$\bar{y}_w = \left(\frac{1}{N}\right) \sum_{i \in S_r} w_i Y_i$, where the sum of the weights $\sum_{i \in S_r} w_i$ equals the total population size N , and the weight

$w_i = \pi_i^{-1} a_{ri} a_{pi}$ is the product of the sampling weight π_i^{-1} with two weight adjustment factors, a_{ri} and a_{pi} as defined below:

- 1) π_i^{-1} is the inverse of the sample selection probability for the i^{th} population unit;
- 2) a_{ri} is a non-response weight adjustment factor. The inverse a_{ri}^{-1} is an estimate of the probability that the i^{th} population unit responds to the survey given that it is selected in the sample; and
- 3) a_{pi} is a post-stratification adjustment factor. The inverse a_{pi}^{-1} is an estimate of the probability that the i^{th} population unit is included in the sampling frame from which the sample was selected.

Statistics Canada (2003) and Biemer & Christ (2008) provide a more detailed discussion on the construction of weighted estimators of the form

$$\bar{y}_w = \left(\frac{1}{N}\right) \sum_{i \in S_r} \pi_i^{-1} a_{ri} a_{pi} Y_i . \quad (3)$$

While the value of π_i is known and determined by the sample design, the calculation of the adjustment factors a_{ri} and a_{pi} requires an analysis of the mechanisms underlying under-coverage and non-response. Weighting is an important topic of survey methodology and, as mentioned earlier, various methods have been proposed by survey statisticians for calculating a_{ri} and a_{pi} depending on the type of auxiliary information that is available on the target population, the respondents, and the non-respondents (for example, see Holt & Elliot, 1991; Deville & Särndal; 1992, and Lynn 1996). In addition to data requirement, the complexity, robustness, impact on variance and computational intensity of these methods must be considered when deciding on an appropriate weighting strategy to be used.

Population-based weighting is presented in Lynn (1996) as a non-response weight adjustment method that can be used when there is no available auxiliary information about non-respondents. This is carried out by creating weighting classes which satisfy the following conditions:

- the weighting classes are mutually exclusive and jointly exhaustive;
- the population size N_k of each weighting class k is known from an external data source such as an administrative list;
- the weighting class to which a survey respondent belongs can be identified using the auxiliary information collected from the survey; and
- the distribution of the variable of interest Y among the respondents who belong to the same weighting class is similar to the distribution of Y in the population for this class.

The population-based weighted estimate has the general form

$$\bar{y}_{pb} = \left(\frac{1}{N}\right) \sum_{i \in S_r} \pi_i^{-1} a_i Y_i , \quad (4)$$

where a_i is the weight adjustment factor for the i^{th} population unit and satisfies

$$a_i = \frac{\sum_{k=1}^L \delta_i^k N_k}{\sum_{j \in S_r} \sum_{k=1}^L \delta_j^k \pi_j^{-1}} . \quad (5)$$

In the above equation, L denotes the number of weighting classes and δ_i^k is a binary variable which is equal to 1 if the i^{th} population unit belongs to class k in $\{1, \dots, L\}$ and 0 otherwise. Based on the definition of a_i in equation (5), it follows that a_i is the same for all units belonging to the same weighting class, and that the sum of the weights $\sum_{i \in S_r} \delta_i^k \pi_i^{-1} a_{pi}$ for the respondents in class k is equal to the population size N_k of the class. When compared with the weighted estimate \bar{y}_w defined in equation (3), the population-based estimate \bar{y}_{pb} only requires the calculation of one adjustment factor a_i instead of two (a_{ri} and a_{pi}) and, as noted in Lynn (1996, p. 210), “as well as correcting for non-response, weighting in this way simultaneously incorporates an element of post-stratification.”

This makes population-based weighting an attractive option for DGMPPRA, given that population-based weighting classes can be constructed using the demographic information available on the CF population already collected by the Director Human Resources Information Management (DHRIM) within DND. The administrative database maintained by DHRIM contains information such as the gender, first official language, marital status, age, years of service, rank, military occupation and CF unit of CF members. When designing a survey, the database is used to extract a snapshot of the target population from which the sampling frame will be created. This extraction can also be used to create weighting classes based on selected variables that are both available through DRHRIM and that are collected through the demographic survey questions.

3.2. Considerations to Using Population-based Weighting in Practice

While population-based weighting appears to meet the needs of DGMPPRA as an efficient method for calculating accurate and reliable estimates in an operational research environment, in practice, this method still poses some challenges, including:

- The population counts for the weighting classes may not be accurate due to the reliability of the information on the weighting variables within the DHRIM database. In this case, it may be preferable to use an estimator which adjusts sampling weights for non-response only (no post-stratification adjustment) based on auxiliary information collected from respondents and non-respondents (sample-based weighting).
- In order to effectively reduce non-response bias, the demographic variables selected for defining the weighting classes must be associated with non-response and with the survey variable Y of interest. Making a good selection in this regard requires the collaboration of subject matter-experts, and if there is more than one survey variable of interest, it may be challenging to identify a reasonable set of variables for defining the weighting classes.
- As highlighted by Lynn (1996), the demographic variables available from DHRIM for creating weighting classes may not be good predictors of non-response. If this is the case, population-based weighting will not effectively reduce non-response bias.
- As discussed in Biemer & Christ (2008), Kim *et al.*, (2007), and Elliot (1999), some weighting classes may contain a small number of respondents and/or result in extreme weight adjustments. In this case, subject-matter experts should be consulted to effectively collapse these classes with other classes in order to avoid an increase in the variance of \bar{y}_{pb} which is not worth the bias reduction resulting from population-based weighting.
- Due to the low response rate of surveys conducted by DGMPPRA, assessing the effect of non-response on the reliability of \bar{y}_{pb} is very important. However, it is also challenging and costly because it requires an analysis of the non-response mechanism and of the remaining bias within the weighting classes if non-response does not occur at random.

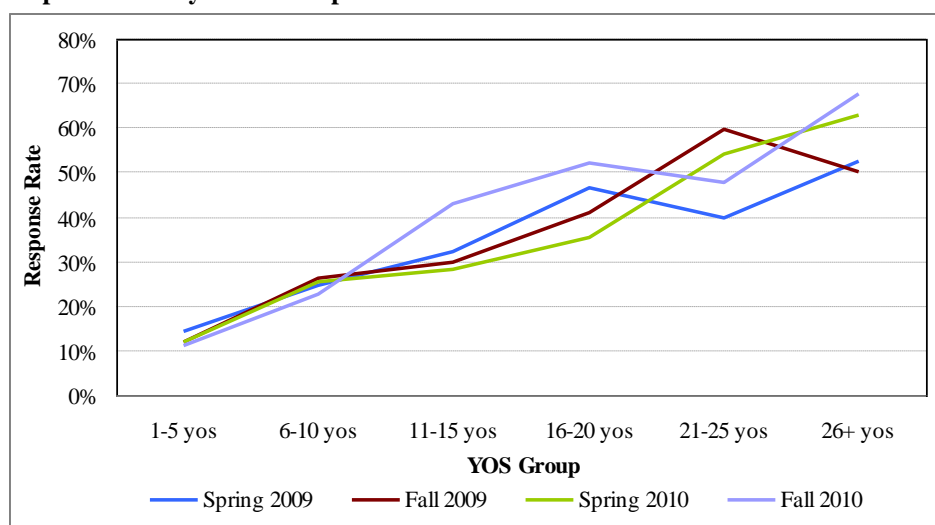
Despite the relative simplicity of population-based weighting, the challenges listed above highlight the importance of planning and allocating sufficient resources for the calculation of survey weights after the data collection period. They also highlight the importance of analyzing non-response and coverage errors for assessing the reliability of \bar{y}_{pb} and reducing its bias, as well as the benefits of having access to survey methodology expertise (internally and/or externally) for the effective calculation of accurate estimates.

4. Consideration of Non-response for the CF Your-Say Survey

As discussed in the previous section, the low response rates (generally below 40%) for CF personnel surveys raise fundamental issues regarding non-response bias and the reliability of survey estimates for generalization to the target population. Assessing this bias requires the investment of resources in the analysis of the non-response mechanism of CF members in order to measure the residual bias within the weighting classes that are used to adjust the weights of survey estimators for non-response. As an initial step in this direction, DGMPPRA analyzed patterns of non-response using a personnel survey that is administered at regular intervals on a repeated basis in order to identify demographic variables associated with non-response to web-based surveys.

The CF Your-Say Survey (YSS) is a bi-annual web-based survey which is used to gather information on the attitudes and opinions of Regular Force members on a variety of personnel-related topics (Urban, 2007). Non-response from the four YSSs administered between 2009 and 2010 (Spring 2009, Fall 2009, Spring 2010 and Fall 2010) was modelled using logistic regression. Five explanatory variables for which information exists for both the respondents (from the YSS) and for the units in the sampling frame (from DHRIM) were considered for the analysis. These included gender; element (Canadian Army, Royal Canadian Navy, Royal Canadian Air Force); first official language (French or English); military rank; and years of service (YOS). Among these variables, YOS was found to be the best predictor for non-response and, as can be seen from Figure 4-1, the probability of response is the lowest for CF members who have been serving for less than five years. Also, the probability of response generally increases as the YOS of a CF member increase. Due to the positive association between rank seniority and YOS, military rank was also found to be an important variable associated with non-response.

Figure 4-1
Response Rate by YOS Group for the Four YSSs Administered in 2009 and 2010



The results of the analysis of the YSS patterns of non-response have allowed for the more accurate prediction of non-response to web-based CF personnel surveys, enabling the integration of non-response into the calculation of sample size requirements during the sample design stage. For example, the following approach describes how a non-response model (based on one or more demographic variables such as rank and YOS) can be used to estimate the sample size requirement per stratum under stratified random sampling when one or more predictor variable is not among the stratification variables.

- 1) Determine the number of respondents desired in each stratum.
- 2) Use the non-response model to calculate the predicted response rate in each stratum based on the distribution of the stratum population with respect to the predictor variables.
- 3) Calculate the sample size required to produce the desired number of respondents in each stratum based on the predicted response rate.

With the integration of non-response models into sampling design, DGMPPRA can more accurately predict survey non-response and calculate the sample sizes required to produce population-based weighted estimates with sampling

variances that meet precision criteria. However, further analysis of non-response is still required to measure the remaining bias within the population-based weighting classes.

5. Conclusion

The recent initiatives taken to examine current estimation methods are leading DGMPPRA to refine these methods in order to improve the quality of survey results. While population-based weighting is an effective method of estimation in an operational research context, it still poses challenges, many of which can be addressed through consultation with survey methodology experts in order to facilitate the delivery of timely and reliable estimates to clients. In this regard, mechanisms are being sought and developed to efficiently integrate such expertise into survey research within DGMPPRA.

Non-response of CF members to personnel surveys remains the principal challenge to the collection of quality survey data. The identification of key demographic variables associated with non-response is certainly a good starting point for understanding non-response. As a next step, the analysis of 'paradata' (data about the data collection process) (Laflamme, 2008) from future surveys can be used to help answer the following questions related to non-response:

- How does the response rate increase with time throughout the data collection period?
- How does the response rate increase with email reminders during the data collection period?
- Do the characteristics of survey respondents vary as the number of reminders increases?

Answers to such questions can help to better understand non-response and find cost-effective solutions for reducing non-response. Indeed, as highlighted in Holt & Elliot (1991, page 333), "the best approach to the problem of unit non-response in surveys is to make strenuous efforts to minimise it in the first place."

References

- Biemer, P., and S. Christ (2008), "Weighting survey data", in *International Handbook of Survey Methodology*, J. Hox, E. de Leeuw, & D.A. Dillman (Eds.), European Association of Methodology, Taylor & Francis Group, pp.317-341.
- Cochran, W. (1977), *Sampling Techniques: Third edition*, Wiley series in Probability and Mathematical Statistics, 1977.
- Elliot, D. (1999), "Report of the Task Force on Weighting and Estimation", Government Statistical Series Methodology Series No 16, United Kingdom: Office for National Statistics.
- Holt, D. and D. Elliot (1991), "Methods of weighting for unit non-response", *The Statistician*, 40, p. 333-342.
- Kim, J., Li, J. and R. Valliant (2007), "Cell collapsing in poststratification", *Survey Methodology*, Vol. 33, No. 2, pp. 139-150.
- Laflamme, F. (2008), "Understanding survey data collection through the analysis of paradata at Statistics Canada", *American Association for Public Opinion Research 63rd Annual Conference, 2008 American Statistical Association, Proceedings of the Section on Survey Research Methods*.
- Lynn, P. (1996), "Weighting for non-response", *Survey and Statistical Computing 1996*, Chesham: Association for Statistical Computing, pp. 205-214.
- Statistics Canada (2003), "Survey Methods and Practice", catalogue no 12-587-X, Statistics Canada.
- Urban, S. (2007), "Your-Say: A Review of Current Administration Procedures and Survey Content", DRDC CORA TN 2007-28, Center for Operational Research and Analysis, Department of National Defence, Canada.

Building an integrated sampling framework for Business Surveys: Simulation studies to evaluate the efficiency of a two-phase sample design

Yi Li and Frédéric Picard¹

Abstract

In 2010, Statistics Canada launched the development of the Integrated Business Statistics Program (IBSP), which aims to redesign the existing Unified Enterprise Survey platform into a generalized model for producing business statistics. The IBSP will use the Business Register (BR) as the sampling frame. However, due to the different natures of the surveys that will be covered by this program, the information on the BR may not be efficient for the stratification. In order to accommodate the special needs of different surveys, it is proposed to have a two-phase design as an option. In this study we evaluated the efficiency of two-phase design against that of one-phase design.

Key Words: Integrated Business Statistics Program; One Phase Design; Two Phase Design; Simulation.

1. Introduction

The Integrated Business Statistics Program (IBSP) is a new project that will integrate most of Statistics Canada's business surveys to improve the efficiency in production of business statistics. A key element of the IBSP is to build an Integrated Sampling Framework that can be applied to a wide variety of business programs at Statistics Canada. Due to the different natures of the surveys that will be covered by this program, the information on the Business Register (BR), which will be used as the frame, may not be efficient for the stratification of those surveys. In order to meet the special needs of different surveys, an option is to have a two-phase design. That is, the information collected from the first phase sample will be used to update and enrich the frame and a second phase sample will be selected using that information to better target the population either through filtering or by more efficient stratification and allocation. The reader is invited to consult Chapter 9 of Särndal *et al.*, (1992) for a comprehensive review of the two-phase sample design theory.

In order to compare, in terms of cost and data quality, the effectiveness of the two-phase approach with that of the one-phase design, we performed simulations on two selected IBSP surveys, the Annual Survey of Manufactures and Logging (ASML) and the Capital Expenditures Survey (CAPEX). This paper will summarize the results of the study. We will start with an overview of the IBSP (section 2), followed by the description of the methodology for the simulations (section 3). We will present the main results in section 4 and conclude with recommendations and a list of future work (section 5).

2. An overview of the IBSP and its sampling design

We briefly describe the methodology for the IBSP sample design in this section. More details about the IBSP can be found in Godbout (2011).

The IBSP is a redesign of the current Unified Enterprise Survey, an annual program developed in 1997 which now covers around 60 surveys. The coverage of the IBSP will be extended to around 120 surveys, including annual and sub-annual (quarterly and monthly) surveys. It will give priority to global optimums rather than local optimums of the individual surveys, develop flexible methodology and systems for business surveys, and make an optimal use of administrative data.

¹Yi Li, Business Survey Methods Division, Statistics Canada, Ottawa, ON, Canada, K1A 0T6, (Yi.Li@statcan.gc.ca); Frédéric Picard, Business Survey Methods Division, Statistics Canada, Ottawa, ON, Canada, K1A 0T6 (Frederic.Picard@statcan.gc.ca).

While most of the IBSP surveys are defined by industry, some of them, for example the CAPEX, are activity based. Some IBSP surveys target data at the enterprise level, and some others collect data at lower levels such as establishment or location. The populations of different IBSP surveys could overlap. For example, the population for the annual survey of transportation and that for the monthly survey of transportation are almost the same. An activity-based survey could overlap with IBSP surveys driven by industry. It is therefore necessary to have a plan for sample coordination and rotation to control response burden.

The IBSP population, covering almost all the industries of the Canadian economy, will be stratified by industry and province so that reliable provincial and industrial estimates can be produced to meet the needs of the System of National Accounts. The definition of industry will be based on the North American Industry Classification System. Due to the highly skewed nature of business population, the population will also be stratified by size to ensure that the important and large enterprises receive a higher probability of selection and that the small ones have a lower selection probability. The size variable(s) could be revenue, expense, asset, land, *etc.*

Stratified Bernoulli sampling has been proposed as the sampling method as it facilitates sample coordination and rotation. The IBSP samples will be selected using the Generalized Sampling System developed at Statistics Canada. Interested readers are encouraged to consult Demnati and Turmelle (2011) for more information about the IBSP sampling design.

It is also proposed to have a two-phase design as an option so that the sample design can be as simple and flexible as possible on one hand and it can meet the specific challenges and requirements that each survey may have on the other. With this design, a unified sampling plan will be applied at the first phase, and basic or general information will be collected from the first phase sample; then a second phase sub-sample will be selected to collect financial and other detailed information. The information collected at the first phase will be used to help target the respondents better or re-stratify the population more efficiently. The theory of two-phase sampling has been well developed and it is known that it has the potential to reduce survey cost and response burden in some situations. However, because of the complexity of the IBSP and the fact that the surveys involved are multi-purpose, we needed to perform the simulation study to evaluate the effectiveness of the two-phase design in comparison with the one-phase design.

3. Description of the simulation and its methodology

3.1 Synthetic populations

Two synthetic populations, one for each of the selected surveys, were built with a mixture of collected and imputed data in order to mimic the true population. They contained stratification variables (industry, province and revenue for both surveys with the addition of the country of control for the CAPEX) and the variables of interest (shipments of commodities for the ASML and expenditures on Capital Construction (CC) and Capital Machinery and Equipment (CM) for the CAPEX). The imputation was based on the nearest neighbor donor method and was implemented by using BANFF, a generalized edit and imputation system developed at Statistics Canada.

3.2 Sampling unit definition and population stratification

In order to compare the simulation results with those from production, the stratification and sampling unit definitions used in production were applied for both surveys.

First, a sampling unit for the ASML was establishment and that for the CAPEX was all establishments from the same enterprises and within the same cell. There were 34,828 sampling units for the ASML and 799,691 for the CAPEX in the synthetic populations. Second, both surveys derived exclusion thresholds to exclude from the populations the smallest businesses that accounted for up to 5% or 10% of the total revenue of a cell. The portion of the units with revenue below the exclusion thresholds was called Take-None stratum. Third, the populations of the units with revenue above the exclusion thresholds were stratified by industry and province (plus Country of Control for the CAPEX). Then, the Lavallée-Hidiroglou (L-H) algorithm was used to stratify these units by revenue into Take-All (TA), Take-Some-Large (TSL) and Take-Some-Small (TSS) strata. The L-H algorithm is described in Lavallée and Hidiroglou (1988). While the units in the TA stratum were selected with certainty, those in the TSL got higher

probabilities of selection than those in the TSS. In the derivation of size boundaries, the CV requirement was set as the same as those used in the production for the two surveys. The units for which the L-H algorithm could not assign a stratum were all flagged as ‘must-take’ which means they were selected with certainty.

3.3 Sample allocation and selection

Stratified Bernoulli sampling was used for selection of the samples for both the one-phase and two-phase designs.

In the one-phase design, the expected overall sample sizes were the same as those in production: 13,500 for the ASML and 30,000 for the CAPEX. The sample was allocated using power allocation (with the revenue variable) for the ASML and Square-Root-N allocation for the CAPEX. The Square-Root-N method was used for the CAPEX due to the weak correlation between the stratification variable (revenue) and the variables of interest (CC and CM).

In the two-phase design, the expected overall sample sizes for the first phase were 18,000 for the ASML and 30,000 for the CAPEX. They were again allocated by Power Allocation based on Revenue for the ASML and by Square-Root-N for the CAPEX. This study assumes there is no non-response.

In a real two-phase survey, each unit selected in the first phase will be asked to answer short questions that will be used for the filtering of units from the second phase or the allocation of the second phase sample. For example, we would ask the unit to list its top three commodities (no dollar amount) for the ASML, or if it has any expenses on CC or CM (Yes/No) for the CAPEX. In the simulation, for each unit selected in the first phase, we obtained the information from the synthetic populations. The CAPEX units that answered no to both questions were screened out from the population for the second phase.

For the second-phase, the ASML sample was allocated by minimizing the multivariate objective function $F = \sum_j \widehat{CV}(\widehat{Y}_j)$ under constraints on the second-phase sample size and sampling fractions, where \widehat{Y}_j was the estimator for the total of the variable of interest, $\widehat{CV}(\cdot)$ was the estimator of the coefficient of variation due to the second phase sampling variance, and the sum was taken over the domain (commodity*province) of interest. We used the top three commodities information obtained at the first-phase to estimate the coefficients of variation. For the CAPEX, out of the 1,000 replicates selected at the first-phase, an average of 19,286 units were in-scope and screened-in. Then, we simulated on two scenarios, one to take a census and another to select a subsample of 15,000 units. Square-Root-N was still used for allocation of the second phase sample.

3.4 Estimation

Estimates of totals for the variables of interest were produced and for comparison purposes, estimates of totals for the size variable (revenue) were also produced.

3.4.1 Calibration of weights and levels of estimation

The initial weight of an in-sample unit is usually the inverse of its inclusion probability. However, since the sample size of STBS was random and the realized sample sizes varied from sample to sample, the weights were adjusted to make the sum of the weights of the units that were selected in a stratum for phase 1 (or phase 2) equal to its population stratum counts for phase 1 (or phase 2). The calibrated weights were used to calculate the estimates for each replicate.

3.4.2 Estimation of the precision

To evaluate the precision of the one-phase and two-phase sample designs, 1,000 Monte Carlo replicates were selected for each of the designs. A Monte-Carlo replicate of a two-phase design consisted of a selection of a first phase sample followed by a selection of a second phase subsample. For each replicate, we calculated the sampling weights and estimated the totals for the variables of interest.

We calculated the Relative Root Mean Square Error (RRMSE), which measures the total variation from the reference value, as follows:

$$(1) \quad RRMSE = \frac{1}{Y} \sqrt{\frac{1}{1000} \sum_{r=1}^{1000} (\tilde{Y}_r - Y)^2}$$

where \tilde{Y}_r is the estimate from the r^{th} replicate and Y is the value of the total of the variable of interest from the synthetic population.

3.4.3 Estimation of the costs

Based on the operational costs information, the following unit costs were applied for estimation of data collection cost for the one-phase and two-phase designs.

Table 3.4-1
Unit cost of data collection for one- and two-phase designs

One-phase		Two-phase	
Collection	Unit cost	Collection	Unit cost
Pre-contact	1	First Phase (ASML)	1.5
		First Phase (CAPEX)	2
Regular questionnaire	7	Second Phase: Regular questionnaire	7
Electronic questionnaire	5	Second Phase: Electronic questionnaire	5

Note that the first phase of a two-phase sample design can be seen as an extended pre-contact from a view of data collection. The reason for lower cost of the first phase for the ASML than for the CAPEX is that there will be significant overlap of samples from year to year and the information asked at the first phase for the ASML (the top three output commodities) is considered stable enough that it can be assumed that it does not need to be updated for units that were in the sample for the previous year.

4. Simulation results

In this section, we present the simulation results for both the one-phase and two-phase designs for each survey. Only the most interesting results are presented in this section. Detailed results can be found in Pacquelet (2011) for the ASML, and in Xie et al (2011) for the CAPEX.

4.1 Analysis of the ASML simulation

Data quality was improved for the two-phase design as we can see from Table 4.1-1 that more domains (defined by commodity*province) had smaller RRMSE for the two-phase than for the one-phase design with almost the same total collection cost (Tables 4.1- 2 and 4.1-3). Also, if we sum up the RRMSE of those 282 domains, the total with the two- phase design was 1.26, which was a substantial reduction from 3.99 with the one-phase design.

Table 4.1 -1
ASML: Comparison of RRMSE between the one-phase (RRMSE1) and two-phase (RRMSE2) designs.

Category	Number of domains	Percentage
RRMSE1 > RRMSE2	162	58%
RRMSE1 = RRMSE2=0	32	11%
RRMSE1 < RRMSE2	88	31%
Total	282	100%

Table 4.1-2**ASML: Cost of the two-phase design with expected overall sample size of 18,600 for the first-phase and 11,000 for the second-phase**

Collection	Average overall sample size	Electronic questionnaire (1 st phase: 0%; 2 nd phase: 60%)		Regular questionnaire (1 st phase: 100%; 2 nd phase: 40%)		Grand total cost
		Unit cost	Total cost	Unit cost	Total cost	
1 st phase	18,600			1.5	27,900	27,900
2 nd phase	11,000	5	33,000	7	30,800	63,800
Total						91,700

Table 4.1-3**ASML: Cost of the one-phase design with expected overall sample size of 13,500**

Operation	Average overall sample size	Electronic questionnaire (pre-contact: 0%; collection: 60%)		Regular questionnaire (pre-contact: 100%; collection: 40%)		Grand total cost
		Unit cost	Total cost	Unit cost	Total cost	
Pre-contact	13,500			1	13,500	13,500
Collection	13,500	5	40,500	7	37,800	78,300
Total						91,800

We see that with costs similar to that of the one-phase design, the two-phase design can improve the quality of the estimates. This is because we have the possibility at the second-phase to better target the sample using the information collected at first-phase. We performed other simulations in which the number of domains of interest varied and we noticed that as the number of domains of interest increased, the gain in precision using the two-phase design became smaller. The two-phase design would have no advantages over the one-phase if there were too many variables of interest.

4.2 Analysis of the CAPEX simulation

In the case of a census for the second phase, the data quality for the two-phase design is equivalent to that for the one-phase design. And in the case of a sample survey for the second phase, the data quality of the one-phase design is only slightly better than that for the two-phase design by comparing the RRMSE. However, tables 4.2-1 and 4.2-2 showed a cost reduction of 16% even when the second phase was a census. Even more can be saved when a sample is selected for the second phase.

Table 4.2-1**CAPEX: Cost of a two-phase design with expected overall sample size of 30,000 for the first phase and a census for the second phase**

Collection	Average overall sample size	Electronic questionnaire (1 st phase: 0%; 2 nd phase: 60%)		Regular questionnaire (1 st phase: 100%; 2 nd phase: 40%)		Grand total cost
		Unit cost	Total cost	Unit cost	Total cost	
1 st phase	30,000			2	60,000	60,000
2 nd phase	19,286	5	57,858	7	54,001	111,859
Total						171,859

Table 4.2-2**CAPEX: Cost of a one-phase design with expected overall sample size of 30,000**

Operation	Average overall sample size	Electronic questionnaire (pre-contact: 0%; collection: 60%)		Regular questionnaire (pre-contact: 100%; collection: 40%)		Grand total cost
		Unit cost	Total cost	Unit cost	Total cost	
Pre-contact	30,000			1	30,000	30,000
Collection	30,000	5	90,000	7	84,000	174,000
Total						204,000

5. Conclusion/Recommendations

The results of the study have shown that employing a two-phase sample design has the potential of reducing the collection cost and response burden over a one-phase design without negatively impacting the data quality; or it can help improve data quality with the same amount of resources.

In the future, studies could be conducted for other IBSP surveys. When the correlation between the variables of interest and revenue is not high, it may be helpful to have a model-assisted approach to generate auxiliary variables having relatively high correlations with the key variables of interest by making use of all available information rather than just revenue. Problems resulting from random realized sample sizes for both phases due to Bernoulli sampling need to be resolved. The impact caused by non-response or measurement error at the first phase on the sample of the second phase should be evaluated as well.

References

- Demnati, A and C. Turmelle (2011), “Proposed Sampling and Estimation Methodology for the Integrated Business Statistics Program”, *Advisory Committee on Statistical Methods*, Statistics Canada, Ottawa, Canada.
- Godbout, S. (2011), “Standardization of Post-Collection Processing in Business Surveys at Statistics Canada”, *Proceedings of Statistics Canada Symposium 2011*, Statistics Canada, Ottawa, Canada.
- Lavallée, P. and M. Hidioglou (1988), “On the Stratification of Skewed Populations”, *Survey Methodology*, Volume 14, no 1, Statistics Canada, Ottawa, Canada.
- Pacquelet, L (2011), “Simulation de différentes méthodes d’allocation pour un plan de sondage à deux phases dans le contexte du PISE”, Internal Document, Statistics Canada, Ottawa, Canada.
- Särndal, C.-E., Swensson, B. and J. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag, pp. 343-385.
- Xie, H, Li, Y and J. Gaudet (2011), “A Preliminary Sampling Study for the Integrated Business Statistics Program Using the Capital Expenditures Survey Data”, Internal Document, Statistics Canada, Ottawa, Canada.

Sparse and efficient replication variance estimation for complex surveys

Jae Kwang Kim and Changbao Wu¹

Abstract

It is routine practice for survey organizations to provide replication weights as part of survey data files. These replication weights are meant to produce valid and efficient variance estimates for a variety of estimators in a simple and systematic manner. Most existing methods for constructing replication weights, however, are only valid for specific sampling designs and typically require a very large number of replicates. In this paper we first show how to produce replication weights based on the method outlined in Fay (1984) such that the resulting replication variance estimator is algebraically equivalent to the fully efficient linearization variance estimator for any given sampling design. We then propose a novel application of the calibration method for replication weights to simultaneously achieve efficiency and sparsity in the sense that a small number sets of replication weights can produce valid and efficient replication variance estimators for key parameters. Our proposed method can be used in conjunction with any existing resampling techniques for large-scale complex surveys. Some extension to balanced sampling designs is also discussed. Simulation results showed that our proposed methods perform very well. Our proposed strategies will likely have impact on how public-use survey data files are produced and how these data sets are analyzed.

¹Jae Kwang Kim, Iowa State University, USA and Changbao Wu, University of Waterloo, Canada.

SESSION 3A

**HARMONIZATION OF METHODS AS PART OF LARGE-SCALE
STANDARDIZATION PROJECTS FOR BUSINESS SURVEYS**

Harmonizing methodologies through a system integration project: Challenges and lessons learned

J. Andrews, F. Brisebois, I. Delahousse, C. Dochitioiu, M. Lachance, R. Philips and S. Pursey¹

Abstract

In 2007, Statistics Canada's Data Quality Assurance Review project identified the processing systems for some of the Industry Statistics Branch monthly business surveys as being areas of concern. Some of these systems were aging in terms of technology, had complex processing designs, had had multiple customizations incorporated over time, or were lacking important functional components. In response, a processing systems renewal project was launched. The vision for this project was to deliver a streamlined, shared, generalized system that can be used by multiple monthly surveys. Due to the common challenges of a monthly production cycle for the surveys involved, a harmonized processing environment was to benefit all these surveys in gaining efficiency and sharing expertise when facing common issues. This project was also taken as a good opportunity to harmonise methodologies used for various survey steps such as the calendarization of reported data, the edit and imputation, as well as the estimation step. In creating this new environment a number of challenges in terms of methodological differences, communication, and human resources were met and overcome. This paper will discuss these challenges and the lessons learned during the process.

Key Words: Harmonization; Integration; Generalized system; Lessons learned.

1. Introduction

Statistics Canada conducts nearly 350 surveys covering virtually all aspects—both social and economic—of Canadians' lives. Despite the diversity of these surveys, they must all meet common quality requirements and ensure compliance with data privacy. For most of these surveys, the operations and methods used are, broadly speaking, often quite similar. For example, except for a few censuses, most surveys use sampling. The data collected are then checked and validated to ensure that various quality criteria are met. Everything is then disseminated via various products that meet high standards of confidentiality compliance. However, when we look at individual processes more closely, it quickly becomes clear that the similarity ends there. The differences, which are often legitimate, arise from the fact that each survey has specific objectives designed to meet the particular needs of its client and users. Even though these differences are unavoidable, they often result in complexities, thereby reducing the statistical institute's overall effectiveness in delivering its products. Harmonization is therefore a desirable initiative.

The purpose of harmonization is to prevent or eliminate differences in technical processes that all have the same goal. Why develop and maintain different data processing methods if they all yield similar results of equivalent quality? However, in a context of continual production such as exists in statistical institutes, it is difficult to stop and examine the situation clearly and then make major changes toward harmonization. Hence the process of harmonizing surveys, while desirable, is a laborious exercise. Nevertheless, the benefits can be many: less complexity and duplication in the processes, improved efficiency and reduced operating costs, less costly computer technology operations owing to a common infrastructure and simplified maintenance, and finally, more effective introduction of best practices within the various teams involved in the different surveys, as well as co-operation among them.

This article provides an overview of the *Industry Statistics Branch Monthly Surveys Systems Integration Project*, which is designed to integrate three surveys into the same production system and to harmonize the methods and practices used. The challenges faced and the lessons learned during the project are described. First, Section 2 provides some background to the implementation of the project, then Section 3 describes different parameters of the

¹Andrews (jessica.andrews@statcan.gc.ca), F. Brisebois (francois.brisebois@statcan.gc.ca), I. Delahousse (ivelina.delahousse@statcan.gc.ca), C. Dochitioiu (catalin.dochitioiu@statcan.gc.ca), M. Lachance (martin.lachance2@statcan.gc.ca), R. Philips (robert.philips@statcan.gc.ca) and S. Pursey, Statistics Canada, Tunney's Pasture, 120 Parkdale Avenue, Ottawa, Ontario, Canada, K1A 0T6.

project, such as its goal and objectives, the surveys involved and the timetable. The project can be broken down into four phases leading to the final product; Section 4 presents these four phases, highlighting their particular challenges and lessons learned. Section 5 provides a summary of these lessons as well as others encompassing the project as a whole.

2. Background

Statistics Canada conducts quality assurance reviews that assess the soundness of quality assurance practices in its different programs. Reviews target specifically the execution of a program, not its design, and aim at identifying risks that could result in a program not being able to deliver its regular product. In 2007, nine mission critical programs were reviewed, and improvements were proposed and implemented.

The Quality Assurance Review that was conducted for the Monthly Survey of Manufacturing (MSM) identified the processing system as a significant area of risk due to its age, complex design and the multiple customizations that had been incorporated over time. The report noted that the system was overdue for renewal and that the condition of the current processing environment significantly increased the risk of making and releasing errors.

The review performed for another monthly survey, the Monthly Wholesale and Retail Trade Survey (MWRTS), indicated that their production system was a more current system, and was being slowly migrated to corporate architecture standards. However, the MWRTS production system was lacking some very useful and important functional components involving mainly macro data analysis, such as times series and revision tools.

In response to these challenges identified through the reviews, a proposal was submitted by the Industry Statistics Branch to look into a harmonized approach for its monthly business surveys. The project was to start with the inclusion of the two surveys reviewed and mentioned above (MSM and MWRTS), as well as another important monthly survey covering the food industry, that is, the Monthly Survey of Food Services and Drinking Places (MSFS).

3. Project parameters and implementation

3.1 Project goal and objectives

The purpose of the integration project is to deliver a streamlined, shared and generalized system that can be used by several monthly surveys. Given the similarity of the challenges encountered during a monthly production cycle, this new system will enable the surveys to benefit from a harmonized processing environment while gaining in efficiency and sharing expertise when common challenges arise.

This new system will fulfil the main objective, which is to simplify and harmonize the existing procedure for processing surveys, to automate processing operations and to distribute roles and responsibilities across the teams currently supporting the monthly surveys of the Industry Statistics Branch. It is important to note that the scope of the project includes all processing stages following data collection, thus excluding stages such as sampling and collection as such.

Efforts therefore focused on implementing a new, integrated system. A secondary objective, consisting of harmonizing existing methodologies where possible, was added to the project. This objective was not to reinvent the existing methodology for each of the three surveys, but instead to take advantage of the existence of generalized systems and to identify opportunities for harmonization that may arise in the future. This objective therefore reflects a migration toward Statistics Canada's generalized systems where they are not already being used. It should be noted that Statistics Canada has and supports a range of products covering the entire spectrum of a typical survey production cycle, extending from sampling (Generalized Sampling system, GSAM) to estimation (Generalized Estimation System, GES), including edit and imputation (BANFF system) and a number of other operations. Deguire, Reedman and Wenzowski (2011) provide an overview of these systems.

3.2 Description of surveys targeted

The systems integration project targets three monthly surveys of Statistics Canada's Industry Statistics Branch. Although these surveys cover different populations, they have various conceptual, operational and methodological similarities. For example, all three are compulsory, use Statistics Canada's centralized business register as a survey frame and utilize tax files to reduce the response burden on smaller businesses. However, they differ in the condition and complexity of their respective production systems. Their methodology also differs in a few cases, but these differences often lie in the parameters used in the methods applied, and not in the methods as such. The paragraphs that follow give an overview of the three surveys, identifying the systems-related needs of each.

The MSM produces statistical series on the activity of the manufacturing industry, measuring sales of goods manufactured, inventories, unfilled orders and new orders. The MSM data serve as indicators of the economic situation of manufacturing industries and are used in calculating Canada's gross domestic product. The most recent redesign of this survey dates back to 1999. Computer technology has greatly evolved since then, and the methodology has also been refined over the years. For example, a major change in methodology was incorporated into the MSM in 2004 to reduce the response burden on small businesses. Instead of submitting those businesses to the usual data collection procedures, it was decided to use the Goods and Services Tax (GST) files to derive their data (Thomas and Cook, 2005). Thus, since the 1999 redesign, the production system has undergone several modifications and additions. It is also important to note that despite the use of a large number of Statistics Canada's generalized systems, others could be introduced to replace various custom computer programs.

The MWRTS provides information on the performance of the wholesale and retail trade sector and is also an important indicator of the health of the Canadian economy. The survey provides monthly estimates of wholesalers' sales and inventories and the number of business locations for retail trade. The last major redesign of the survey was completed in 2004. As with the MSM, one important change was to incorporate the use of GST tax data in order to reduce the response burden on small businesses (see Trépanier (2004) for further information on that redesign). As regards the review of quality assurance, it should first be noted that the existing system for the MWRTS was much more up to date than that of the MSM, at least with respect to the technology used. However, the survey was much in need of more user-friendly tools, such as for analyzing consistency between the annual and monthly surveys (including the capacity to make the necessary revisions), as well as for being able to examine more closely the time series produced for the survey.

Finally, the MSFS provides estimates of sales for restaurants, caterers and drinking places. At the time of the integration project, the survey had just completed a full overhaul with respect to both methodology and computer technology. The system in place used current technology, and in the great majority of cases, it used the generalized systems of Statistics Canada. Although this survey had not necessarily undergone a quality review, it was one of the Industry Statistics Branch's three monthly surveys, and to ensure complete harmonization, it was included in the integration project. The expertise acquired as a result of its recent redesign has contributed the development of the project.

3.3 Overall course of the project

The project began in 2008 and was initially intended to last two years. However, because of the complexity and scope of the task, the timetable had to be extended by two years, and thus in 2011, the newly developed system was used for a first survey. The project can be broken down into four major phases, each entailing quite separate challenges. The section that follows examines these four phases, highlighting the challenges encountered and the lessons learned.

4. Phases of the integration project with their challenges and lessons learned

4.1 Understanding and documenting the tools and methods in place

Even before undertaking any major change in such a large-scale process as those used in monthly surveys, it is important to first get a full picture of the situation. It is therefore necessary to break the process down into

components and make sure that the importance of each is well understood. It is also crucial to have a good grasp of how these components fit together in order to develop a new system that produces the same high-quality final product, but does so more efficiently.

A review of the process was conducted, and in a sense it marked the launch of the project. The three systems that were in place for the selected surveys were examined closely and a functional model was defined for the new system to be developed. In all, then, there was a complete description of the process, a list of input and output files for each and a flow chart showing the movement of data through the process and how all these files were interconnected. The review exercise was also designed to clearly define the roles and responsibilities of each work team involved, to underline the importance of using a common technical language, and to include a few technical recommendations regarding the computer technology to be used.

At the same time, a complete review was conducted of the statistical methods in place. There was a detailed examination of several stages: calendarization of reported data, edit and imputation, and estimation. The general finding was that the three surveys use fairly similar methodology on the surface but exhibit a number of differences in the parameters used within these methods. Computational tools also differ from one survey to another. Table 4.1-1 provides an excerpt of the conclusions drawn from this review of methodological processes.

Table 4.1-1
Summary of the conclusions drawn from the review of methodological processes for selected survey stages

Stage	Conclusion
Calendarization of reported data	The three surveys have a similar approach, except for the MWRTS, which uses a daily weight. The tools used differ and do not include generalized systems.
Edit	Each survey has rules that determine whether a value is too large or incoherent in relation to the historical or auxiliary information available. The general approaches are similar but use different parameters (<i>e.g.</i> , tolerance rules). The tools used are different; one survey already uses the BANFF system, while the other two plan to adopt it soon. Outlier detection is also performed for all three surveys. The MSM uses a different method (based on the Mahalanobis distance) from the other two (Hidirogrou-Berthelot) and employs different tools for this purpose.
Imputation	Similar data sources are used for imputation: historical data (with trend), data from the survey frame or tax data, or data collected from respondents (imputation by class, for example using the mean). When classes are used, they are constructed in similar fashion using industrial classification variables, geographic information and a measure of size. The tools used are different.
Estimation	The methods are similar for the three surveys, except for estimation for the take none portion, for which the three surveys use tax data via different approaches. The GES is used by all three surveys.

For this first phase of the integration project, the main lesson originates in the finding made in the integration phase (described in Section 4.3), when the different processes had to be linked together to form a complete and functional system. The lesson consists in spending considerable time and effort in this initial, documentation phase, so as to obtain a very complete and detailed picture of the processes in place. The integration project had taken care to scrutinize the systems in place, while concentrating mainly on the main processing stages. The intermediate stages, which are often carried out in the background but serve to interconnect some of these stages, had received less attention. They proved to be a rather complex puzzle and required additional development time. An example of such an intermediate stage is the extraction of tax data from centralized databases and then their temporary storage so that they could subsequently be used and manipulated in the processing of the surveys. It was therefore important to devote a large portion of the planned timetable to this first phase, since oversights here at the outset could lead to turbulence along the way, when efforts would be focused on other tasks.

4.2 Harmonization of tools and methods

The work performed in the first phase served to give direction to the harmonization efforts that were subsequently undertaken. Since the main thrust of the project was harmonization of systems, computer technology experts were able to examine opportunities with respect to the technology and systems to be used. Statistics Canada maintains and supports a wide range of generalized products, and the main idea was obviously to use them, provided, of course, that they could satisfy needs. In short, when a processing stage was accomplished by two surveys using different systems, the approach using a generalized system was favoured.

As noted above, a secondary objective of this project was to harmonize the statistical methods used. The two main changes in methodology were with respect to harmonizing the method of calendarizing the data reported and the method of detecting outliers in the edit stage. For calendarization, a method based on the one used in the processing of tax data was adopted. That method was, in a sense, already in place for the MSFS and therefore served as a starting point. The use of monthly and daily “weights” was made more consistent for the three surveys. As regards outlier detection, the method in place for the MSM was changed. This survey used a method based on the Malahanobis distance, whereas the other two surveys instead used the Hidiroglou-Berthelot (HB) method (see Hidiroglou and Berthelot, 1986). Since the HB method was available in the BANFF generalized edit and imputation system, and since it yielded results very similar to those obtained with the Malahanobis method, it was adopted for the MSM.

The main challenge faced during this phase was to bring about a change in mentality. The existing three surveys are conducted and processed by three different teams. Even though these teams all work more or less within common divisions at Statistics Canada, this is nevertheless an approach based primarily on a traditional “silo” structure. The project sought to introduce a more “global” approach that would better meet the common needs of the surveys and lend itself to horizontal co-operation, integration and solution sharing among teams. During this harmonization period, efforts essentially focused on the use of existing solutions and tools. This might in some cases be perceived as an obstacle to innovation, which is an aspect of work on surveys that is often more attractive and motivating. However, this interruption in innovation was to be only temporary, and from this point on a “global” approach was to be followed.

4.3 Development and integration of components of the new system

The third phase was designed to take the processes or tools identified previously and develop what was to be the new system. Initially, the plan was to set up this system using a totally new mechanism while incorporating the right tool when required. For example, for the data edit and imputation stage, BANFF was the preferred tool. Thus, it was now necessary to make sure to provide the proper specifications for each survey (the “parameters”) but to do so within the same harmonized process, thereby ensuring that the same features are available for the three surveys. For most of the survey stages, these tools were developed in this way. However, with a lack of human resources and time, a few processes already in place in production were simply reused in their existing form and were plugged into the new system that had been developed. For example, the estimation programs specific to each survey were reused, undergoing only the technical changes that would allow connection to the integrated system and ensure good transmission of parameters, linkage with input files, *etc.*

As noted above, some major intermediate processes that had been neglected during the initial, documentation phase had to be developed on the spot during this phase. This meant that there was an increase in the already heavy workload of team members and at the same time an increased the risk of errors.

4.4 Testing and implementation

The last phase of the project consisted in conducting numerous tests and then ultimately moving on to the start of production. An elaborate testing strategy was developed, extending from local testing (one process at a time, independently) to the final testing, consisting of a parallel test (complete production simultaneously on the old system and the new, integrated system). These tests serve to ensure the production is going smoothly and that the same high-quality results are obtained. It had been agreed that the estimates ultimately produced might be different from those obtained using the current production system, since some changes made when methods were harmonized

would have an impact—although not a significant one—on the figures produced. However, potential impacts were well documented, so as to have a better grasp of the differences observed and to be able to explain them.

The official start of production under the new system was planned for late 2011 or early 2012. At the time this article was written, the MSFS was going to be the first survey to make the transition to the new system. The other two were to follow in the next months, with an attempt being made to coordinate this transition as well as possible with other major stages in the survey process (such as annual revision of estimates).

5. Project summary

In addition to the challenges discussed during the four phases, various other important aspects add to the complexity of this project. For example, the new system had to be developed while maintaining the monthly production of the surveys. The production schedules for these surveys are very tight, and this left very little time for testing the new system in real time in order to imitate actual production. Also, most of the personnel involved in developing the new system were people already assigned, either directly or indirectly, to monthly production. Therefore, this involvement, on the one hand, was demanding for these persons, but on the other hand it was necessary in order to ensure that there was a good understanding of the systems in place and to properly determine the requirements of the new system.

In conclusion, the development of an integrated production system addresses several existing risks as well as needs of the production teams. After all, despite a time frame that was longer than expected and a final product that was less generalized than was hoped for, this project is a step in the right direction. Following this integration of monthly surveys, Statistics Canada has undertaken a larger-scale project on integrating its annual business surveys, in what is known as the Integrated Business Statistics Program. This program will actually be a redesigned and expanded version of the existing United Enterprise Survey. The lessons learned in the monthly surveys integration project can thus be put to good use.

References

- Deguire, Y., Reedman, L. and M. Wenzowski (2011), “Generalized systems: the Statistics Canada experience,” *2011 International Methodology Symposium*, Statistics Canada.
- Hidioglou, M.A. and J.-M. Berthelot (1986), “Statistical editing and imputation for periodic business surveys”, *Survey Methodology*, June 1986, Vol. 12, No. 1, pp. 73-83, Statistics Canada.
- Thomas, S. and K. Cook (2005), “Combining administrative and respondent data by the monthly Survey of Manufacturing,” *Symposium 2005: Methodological Challenges for Future Information Needs*, Statistics Canada.
- Trépanier, J. (2004), “The Redesign Canadian Monthly and Retail Trade Survey: A Postmortem of the Implementation,” *Proceedings of the Survey Research Methods Section*, American Statistical Association.

The multiple facets of the redesign of IT tools for producing short-term statistics at INSEE: The Premice program

Fabien Guggemos¹

Abstract

The mission of INSEE's Premice program is to have common implementation and development for projects involving the redesign of IT tools for managing short-term business indicators (indicators of current business activity, price indexes, business climate survey indicators). Firstly by coordinating different projects and secondly by providing or developing shared IT services or statistical models, the Premice program seeks to improve the quality and consistency of short-term indicators, to secure and streamline their production, and to achieve productivity gains in development and then in maintenance.

Key Words: GSBPM; Short-term indicators; Pooling; Statistical processes; Common IT services.

1. Introduction: Background and creation of the Premice program

The idea of pooling the methods of producing short-term indicators at INSEE (National Institute of Statistics and Economic Studies, France) came about in the mid-1990s. At that time, a project called Propice was developed, consisting of a set of statistical tools organized in an IT application and designed to be used simultaneously in applications for producing indexes on industrial production, new orders and turnover and for managing the survey on products, expenses and assets. Propice was based on the finding that there was great similarity between both the objects being processed and the functions to be performed. At the time, a choice was made to use the SAS language and create "macro-program" catalogues. While those tools are still in use today, they are not without their shortcomings. The processing chains are not modular and cannot be interrupted; they call for complete processing from the control of input data through to the formatting of results, with a run time ranging between three hours and overnight. The lack of flexibility in these chains ultimately led to the development of self-serve programs that duplicate entire parts of the application.

More recently, the INSEE project CRPI (Internet collection and return) devised a tool for collection via the Internet, which is widely used in the field of short-term indicators. As a result of the 2001 merger of three projects (Internet-based collection for business climate surveys and for industrial prices, as well as a portal for businesses), the website entreprises.insee.fr was launched in late 2003. The CRPI application is operational and offers businesses the opportunity to respond online to a number of surveys managed by different divisions of INSEE. Maintenance has a sizable upgrading component, and therefore financial resources are very largely devoted to investment in new surveys. However, the cost of incorporating the new surveys is often considered prohibitive for potentially appropriate applications. In particular, it is necessary to rework both the survey and the CRPI tool to construct the mechanism for doing this.

These different projects offered an opportunity to acquire experience with successfully implementing inter-divisional projects, and they also shed light on commonalities among short-term indicators. As regards the first point, the CRPI was the first project of this nature: the collaboration was always satisfactory... although the issue of ownership of the application was never settled. As to the second point, it was confirmed by an urbanization study, conducted in early 2008 at the request of INSEE management, examining the overall system of short-term business statistics. From that study, the following three findings emerged:

¹Fabien Guggemos, Institut National de la Statistique et des Etudes Economiques (INSEE), 18, boulevard Adolphe Pinard, 75675 Paris Cedex 14, FRANCE (fabien.guggemos@insee.fr).

- applications for managing short-term indicators are mostly outmoded, since they were developed with aging or even obsolete technologies and software;
- this obsolescence poses significant risks for the production of indicators and high operating costs, firstly for maintenance teams and secondly for users, who are sometimes obliged to create, on a self-serve basis, “warts” that perform services not provided by the application;
- the procedures for these applications at the macro level are quite similar, as are the statistical processes involved, and therefore it seems appropriate to look comprehensively at the redesign of these various applications.

In this situation, in which redesigns of the tools for producing price indexes (Papaye project) and managing business climate surveys (Conj2 project) were being launched concurrently, all the elements were in place for creating a structure ensuring coordination of the projects and their inclusion in an initiative for sharing or pooling what they had in common. It was thus that the Premice program, whose name is an acronym for Programme de REfonte avec Mutualisation des Indicateurs Conjoncturels d’Entreprises (redesign program for shared short-term business indicators), was officially created in early 2009.

2. Strategic objectives of the Premice program

The mission of the Premice program is to “federate” the implementation and development of projects for the redesign of IT tools for managing short-term business indicators. However, the pooling initiative that underlies this mission is conceived “prudently.” The latter statement is characteristic of the spirit of the Premice program. Mindful of the difficulty of conducting a single project and wiser from experience with the Propice project mentioned in the introduction, Premice opted to position itself toward support and facilitation of projects rather than prescription and standardization. This, then, is not a rigid structure weighing down on projects but rather an approach supporting them. Accordingly, Premice’s success is measured in the quality of the service rendered to users rather than in the number of modules developed in common. In short, the Premice program is both ambitious and pragmatic.

The Premice program is designed to pursue and achieve the following three strategic objectives:

- to improve the quality and consistency of short-term indicators;
- to secure and streamline data production;
- to achieve productivity gains in both development and maintenance.

2.1 Improve the quality and consistency of short-term indicators

Projects generally provide opportunities to explore the processes for development of statistics. In other words, beyond the creation of a new IT tool that is generally the most visible deliverable, a project can advance both statistical methodology and the index production process. Under the Premice program, statistical project teams are in ongoing contact, which promotes communication and experience sharing. At the same time, the program does not seek to impose methods when project leaders do not subscribe to them.

One of the criticisms made of the above-mentioned Propice project was its monolithic nature; the processing chain is launched in one go and takes a long time to deliver its data. Yet, in the analysis of cyclical or short-term data, a value added that index managers provide is that they are able to propose different scenarios (*e.g.*, to produce indexes, raw or seasonally adjusted, for different items and existing classification levels). One of the ambitions of Premice, then, is to maximize the number of scenarios run so as to increase the quality of the indicators produced, which requires that IT processing be sufficiently flexible and fast. The producer must have a number of mechanisms for acting rapidly on the indicator to construct a sufficient number of scenarios. Nevertheless, care must be taken to maintain a reasonable number of possibilities offered and to make the processing secure by providing a high degree of traceability.

The common discussion framework provided by Premice is also intended to improve the consistency between the different current indicators. Such consistency is necessary when comparing different indexes, something that is done,

for example, when developing quarterly accounts. In the absence of such consistency, a comparative analysis of seasonally adjusted current indicators would not, for example, be very meaningful. By clarifying the methods of producing each indicator, each project team can see where it stands in relation to others and can accordingly adopt a production strategy for its own indicator that is compatible with that of the other indexes.

The most tangible proof of the consistency of different processings is the use of the same tools. While it may be difficult to create common modules, there are different degrees of pooling; this can extend from a shared vision of a process (using the same vocabulary, the same description of a given task) to common processing (performing the same action). The purpose of Premice is not to force producers to adopt a given tool but rather to provide them with the technical capacity to implement best practices.

Ultimately, the idea is to go from an organization with parallel processes to a policy of lively exchanges within a community. Beyond the quality of the indexes, there is the prospect that the actors involved will find it easier to move between the different indicators and acquire considerable expertise in these domains.

2.2 Secure and streamline data production

The production of short-term business indicators is quite a common operation in the field of statistical operations; most often, indicators are published on a monthly basis. The challenge is therefore to be able to produce these figures monthly at the scheduled time and date. With the Premice applications, the processing time for the different indicators should decline significantly. Not only will faster processing result in shorter lead times for publication, but it will also improve data quality by leaving more room for analysis of indicators and their revisions.

Moreover, the time allowances for implementing the process are also short; it is therefore necessary to be very efficient and strive to avoid any unnecessary action. In Premice, the objective is to pool experiences in order to derive from them all the elements for producing indicators as quickly as possible with an adequate standard of quality. Once again, the gains anticipated are in terms of quality in the analysis stage.

Finally, to make production thoroughly secure, it is necessary to be able to react quickly to various unforeseen events. One of the objectives of Premice is to provide the quality of service of an application with the flexibility of a self-serve function. Among the elements essential to making application software available to statisticians, traceability is reinforced and facilitated. The automated production of indexes will thus be carried out using updated and easily maintainable tools, while the various production chains will be documented in detail.

2.3 Achieve productivity gains in both development and maintenance

The development of common modules requires a sizable initial investment to ensure that the services rendered will adequately meet the needs of the different projects. However, this potential extra cost is offset by many advantages. Where a number of issues relating to the production of current indicators are tackled in a coordinated fashion and inter-team relations are increased, each project benefits from ideas and comments regarding the various other projects, and this can lead to a convergence toward common or similar solutions. In fact, exchanges between projects help to define a common culture and a common language in the field of current indicators. More concretely, the development of a common module prevents functionalities that are identical between two or more projects from being programmed more than once. A project can therefore benefit from the functionalities of a model developed by another project team or adapt them if needed, based on its own specificities.

The modular set-up that must be adopted insofar as possible by projects under the Premice program also helps to improve the quality of application software products. This serves to reduce the cost of maintaining the products and also offers opportunities for partial redesign. A component that is not fully satisfactory can be altered quickly and reinserted in place of the old one according to a timetable more compatible with the requirements of the project ownership. Not only will there be gains due to the fact that the maintenance of a common module, carried out by a single team, will benefit all the applications, but the modular set-up will make pooling much easier by avoiding linear and non-interruptible processing chains.

Ultimately, both the development of common modules (for which maintenance will be centralized) and the closeness between IT architectures (which favours a common culture within development teams and reduces the cost of

training newcomers) ensure that Premice's applications will be long-lasting and contribute to productivity gains, a major issue for INSEE.

3. Scope of the Premice program

The Premice program was created in late 2009 on the occasion of the redesign of the applications for producing price indexes (Papaye project) and managing business climate surveys (Conj2 project). This program is designed to incorporate the various applications for producing short-term business indicators as these applications come up for redesign. Thus, the systems for producing turnover indexes (Harmonica project) and industrial production indexes (Ocapi project) joined the program when their overhauls were launched in late 2010 and late 2011 respectively. Ultimately, the program's scope encompasses the following:

- **activity indicators**
 - industrial production indexes - (*Ocapi project, from late 2011 to 2015*)
 - turnover indexes - (*Harmonica project, from late 2010 to 2014*)
 - new order indexes
- **indices de prix** - (*Papaye project, from late 2009 to 2013*)
 - producer and import price indexes
 - agricultural price indexes
- **business climate surveys** - (*Conj2 project, from late 2009 to mid-2014*)
- **other indicators**
 - construction, rent benchmark, and commercial rent indexes
 - monthly survey of food superstores

4. Four operational dimensions of the Premice program

The three strategic objectives encompass four operational dimensions, relating to the general organization of the program as well as to its concrete expression in computational and statistical terms:

- define a coordinated strategy for the Premice program as a whole;
- link the design and implementation of the different application software products within the scope of the program;
- set up IT services that can be used jointly by the applications that come under the Premice program
- implement statistical processes that lend themselves to pooling or even standardization.

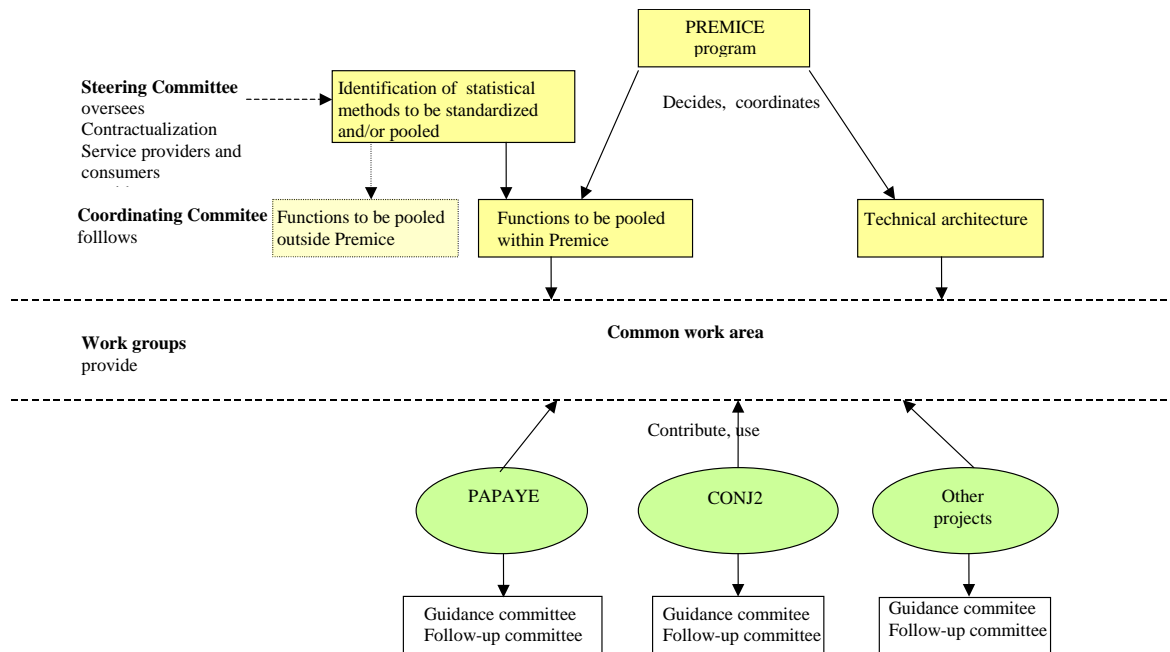
4.1 Functioning the Premice program

Define a coordinated strategy for the Premice program as a whole

The Premice program primarily seeks to set out development principles for the projects concerned. For each system, a prerequisite for setting up a common architecture and technical components is to identify functions that are durable and stable over time and functions that can steadily evolve, and to bring out common characteristics (in particular the need to have tools for making frequent and rapid calculations). This common IT organization must be able to meet future applications' need to react and evolve; it must also secure their functioning and reduce the expenses associated with both development and maintenance. Beyond this organization, it is necessary to identify IT services that can be used jointly as well as statistical modules that lend themselves to pooling. It is up to the steering committee, the program's actual decision-making body, to decide on opportunities for pooled investment. More generally responsible for ensuring that the program operates smoothly, the steering committee decides on the services and modules to be pooled. It also determines how the development work on those modules will be distributed among the different project teams and designates an owner—responsible for maintenance—for each of these common services

and modules. The steering committee undertakes strategic arbitration for the program in the event of disagreement among the various projects' ownership teams. Its decisions are informed and prepared by a coordinating committee. This committee, the preferred forum for communications between the different project ownership teams, guides and coordinates operational activities, creating and monitoring the working groups set up to identify and study in detail the opportunities for pooling between applications.

Figure 4.1-1
Operational diagram of the Premice program



4.2 Linkage of the different projects

Link the design and implementation of the different application software products within the scope of Premice

The design and implementation of services to be pooled are conferred on one or another of the projects, depending on the availability of the teams. The other projects then have direct access to these services and can possibly make minor changes to them in accordance with their specific needs. Since the projects are carried out concurrently, there can be constructive dialogue among the teams. Such exchanges result in significant homogenization of many aspects (microanalysis; work methods, such as adoption of a standardized grid for analyzing processes, derived from international models such as the Generic Statistical Business Process Model (GSBPM)). However, each project participating in Premice retains its identity and independence; it therefore has its own objectives while using the shared architecture and development platform.

Thus it was clearly established that a project participating in Premice did not have to use pooled modules when they did not perform the service expected by the project. While Premice has the scope of a program, it does not necessarily have all the characteristics of one. Moreover, the resources devoted exclusively to Premice are quite small (it has a single statistical coordinator and a half full-time equivalent dealing with technical and IT aspects). It is the projects themselves that have the financial resources, commensurate with their respective complexity; a project participating in Premice can therefore refuse Premice-related requests that may pose a risk to its smooth functioning.

Despite the limited means at its disposal, the Premice program produces a number of deliverables, of two different types:

- statistical deliverables; Premice provides a framework for discussion on the subject of current indicators; successive working groups examine themes selected for their pooling potential;
- IT deliverables, which may be the counterpart of statistical deliverables (since some working groups have IT themes) or may give concrete expression to statistical deliverables in the form of tools or services that are created and then used by projects.

Premice is therefore a fairly new object in the landscape of INSEE applications and projects. It is not intended to address an issue involving the tooling of one of the Institute's statistical operations, but rather it supports the redesign of part of its information system.

4.3 Implementation of common IT services

Set up IT services that can be used jointly by applications that come under the Premice program

The urbanization study cited in the introduction led to a call for the use of common IT services, especially in the “upstream” and “downstream” stages. The services provided can be of different types, such as a module integrated into the application, or into an external application with which the application will exchange information.

Thus the CRPI, the project for managing collection via the Internet, is already implementing a joint service for data collection between surveys on production prices and short-term business surveys, and this service preserves certain boundaries between the processes for those surveys. Other services have been identified, and the steering committee has decided to have them implemented. This entails a single host platform for data, for operations for disseminating results—in particular to INSEE's macroeconomic database—and for dialogue with INSEE's future business statistics directory (draws of samples, business demography signals). Beyond the obvious productivity gains during the development process, the implementation of shared services should result in application software products that are more homogeneous and are accordingly of better quality and easier to maintain.

4.4 Pooling of statistical processes

Implement statistical processes that lend themselves to pooling or even standardization

The Premice program is structured around the identification of modules—sub-units in the statistical process—that lend themselves to pooling. The projects incorporate these modules into their IT development framework or at least provide for the possibility of taking them into account when they are ready. A list of these poolable modules has been established, based on a comparison of the production processes for the various indexes, with more detailed studies subsequently having been carried out on each of them. A number of potentially poolable areas have been identified. These include analysis of indicator revisions, control methods and data adjustment; application of confidentiality rules; and techniques for correcting seasonal and trading day variation. To evaluate precisely the potential for pooling, studies have thus been carried out by working groups comprised of members of the different project teams. The sharing of good statistical practices serves to improve the consistency of the indicators produced and the overall quality of cyclical analysis.

5. Conclusion

Premice, which extends until 2015, is an ambitious program that is energizing for project teams. Involving several divisions of INSEE, it is progressing in an atmosphere of mutual confidence among the many different actors. The very organization of the program, described in Part 4, is a guarantee of good behaviour, since the success of Premice primarily depends on the success of the projects that comprise it. The visibility of the program therefore depends on the projects. In light of this, the pooling initiative requires ongoing investment and support. For statistical aspects, the working groups play this role, while for the IT aspects, pooling takes place at different levels, from the developer's

toolbox to the sharing of responses to specific problems in the field of short-term indicators. The first concrete achievements are beginning to appear, in the form of prototypes for integrating software packages for seasonal adjustment and confidentiality management. While the road ahead should be easier with Premice, it will be necessary to assist projects to overcome various obstacles and then make good on investments. The question always arises as to what the best collective response is; this is indeed the primary concern of those responsible for Premice.

Reference

METIS Steering Group (2009), “Generic Statistical Business Process Model,” version 4.0, www.unece.org/stats/gsbpm.

Standardizing UK sub-annual Business Surveys

Salah Merad and Pete Brodie¹

Abstract

The UK, along with other member states of the EU, was required to adopt an updated version of its standard industrial classification (SIC, equivalent to NACE), called SIC 2007, in its sub-annual business surveys from 2010. The office took this opportunity to review the whole survey process, from customer requirements to publication. In particular, it was decided to merge all sector specific monthly surveys into a single monthly business survey (MBS). In the first stage, only two surveys, covering the production sector and the distribution and services sector, have been harmonised, standardised and processed together. Other surveys, including those covering retail and construction, have adopted the same survey name, MBS, but their harmonisation with the first two surveys will be undertaken at a later date.

In this paper, we describe the different aspects of the harmonisation and standardisation, including questionnaire design, sample design and allocation, response chasing strategies, data validation rules, imputation classes and estimation methods. We also discuss the operational problems encountered during the implementation process and review the lessons learnt from our experience.

Key Words: Standardised methods; Survey design; Integration; Response burden.

1. Introduction

1.1 Background

Up until 2010, the Office for National Statistics was running a set of monthly surveys covering different economic sectors: production distribution and services; retail; construction. The surveys of production and distribution and services collected turnover, the survey of retail collected the value of sales, whereas the survey of construction collected data on the value of output for different types of work. Other variables were also collected: employment was collected monthly in some surveys and quarterly in others; new orders and exports data were collected in the production sector.

The four surveys have different histories and developed independently; as a result, different methods were used, even when the information collected was very similar—as for employment and turnover in the surveys of production and distribution and services. Although there was a degree of integration with respect to the use of a common sampling frame and coordinated sample selection, it was thought that much more integration and standardization could be achieved between the surveys. A project scoping document outlining the objectives and benefits was produced in 2004, but the work was not taken forward because of other priorities at the time.

1.2. Drivers for Standardization

The UK, along with other member states of the EU, was required to adopt an updated version of its standard industrial classification (SIC, equivalent to NACE), called SIC 2007, in its sub-annual business surveys from 2010. Some activities classified under production in the old classification system would move to be classified under services in the new system, and vice versa, some activities would become out of scope and new ones would be introduced. As part of the process of accommodating the new classification, it was decided to review many aspects of the survey operations, and to use this as an opportunity to integrate further the short-term surveys and standardise their methods and processes. This should minimise the operational problems caused by businesses moving between sectors and reduce the perception of increased burden. In particular, it was decided to merge all the sector-specific

¹Salah Merad, Office for National Statistics, Government Buildings, Cardiff Road, Newport, UK, NP10 8XG. salah.merad@ons.gsi.gov.uk; Pete Brodie, same affiliation and address as first author, pete.brodie@ons.gsi.gov.uk.

monthly surveys into a single Monthly Business Survey (MBS). In the first stage, only two surveys, covering the production sector and the distribution and services sector, have been fully integrated, although the standardization of some methods was extended to the surveys of retail and construction; their full harmonization and integration will be considered at a later date. The redevelopment of Workforce Jobs statistics was also taking place at the same time; as employment data collected in the short-term surveys represented the main source in the production of these statistics, the standardization of the methods used in these surveys should lead to more comparable statistics across industries.

In addition to quality considerations, the standardization of methods and processes should lead to efficiencies: running a smaller number of surveys should lead to lower operating and maintenance costs; standardized processes should lead to staff being able to work on different sectors with minimal additional training.

In this paper, we describe the redesign work carried out along the entire survey process, from defining the scope of the survey to publication, which we refer to at ONS as ‘the statistical value chain.’ This paper is based heavily on two papers: James (2010) and Taylor *et al* (2011). In section 2 we describe the redesign work across the statistical value chain, and in Section 3 we discuss the challenges encountered at the implementation stage. In section 4, we list the main lessons we have learnt from our efforts at standardisation and discuss on-going and future work.

2. Redesign Across the Statistical Value Chain

2.1 Redesign Principles and Constraints

An important aim of the redesign work was to present a single survey of monthly economic data to respondents: as such it was decided that a new standardised name be adopted for all monthly surveys. The name ‘Monthly Business Survey’ was chosen; at the same time, the ONS’s structural business survey, Annual Business Inquiry, was renamed ‘Annual Business Survey.’ It was also decided to remove all unnecessary differences in methods and processes.

A separate review and consultation about employment statistics concluded that employment data were only needed quarterly; hence, it was decided to standardize the frequency of collection of employment in all sectors. Furthermore, because employment data are less volatile than turnover, it was decided that only a subsample of businesses should be sent employment questions every quarter, which should reduce the burden on respondents. The implementation of sub-sampling was relatively easy as it was already in use in the services survey. It was also decided to use common editing, imputation and estimation methods where possible and to set the parameter values where needed using the same criteria across all sectors.

We operated under a number of constraints: the total sample size was to remain the same and we could only carry out a joint sample allocation between the production and services sectors. The quality requirements by the survey areas could not be all satisfied with the available sample size and hence compromises were required. Finally, all work had to be completed and changes implemented by January 2010, to coincide with the change in industry classification, which limited the number of methods that could be considered and the time spent to evaluate them to choose between them.

2.2 Brief Descriptions of the Redesign Work

2.2.1 Scope of the survey

As a result of the adoption of the new classification system, SIC 2007, the scope of the survey was expanded. For example, ‘Landscape design’ moved from Agriculture (out of scope of the short-term surveys) to the services sector. Also, it had been a long standing desire of the office to carry out direct collection of turnover in some industries for which data were sourced elsewhere—an example is ‘Veterinary activities.’ Some of these industries were introduced into MBS in 2010, others will follow later to allow more time for questionnaire testing. In industries where only employment is needed, because turnover is obtained from an external source, another survey called ‘Quarterly Business Survey’ was introduced.

One of the lengthy tasks of the redesign was to agree the SIC groupings for which outputs would be published. Output managers for both turnover and employment, who represented external customers, representatives from the main internal customers (e.g., National Accounts) and Methodology were involved in agreeing the publication groups. An important consideration from the methodological perspective was to reduce the level of detail, as the publication groupings formed the basis for stratification: too many strata results in an inefficient sample allocation. In the retail sector, the number of publication groups stayed the same at 27 but in the production and services sectors the number of groups was reduced from just over 300 to 150. In the employment only industries, the number of groups was reduced from about 40 to about 30 in the Quarterly Business Survey.

Another consultation was conducted to identify which variables were no longer needed and which components to include in the definition of turnover. This exercise led to dropping a number of industry-specific questions, and this resulted in the reduction of the number of questionnaire types from 65 to 26.

2.2.2 Questionnaire Design

As was mentioned above, a single name, Monthly Business Survey, was adopted for all monthly ‘turnover’ surveys. However, for the retail sector an additional short description was adopted: Monthly Business Survey - Retail Sales Index. This was thought to be important by the survey area to ensure continuity with the old name (Services - Retail Sales Index). Reaching agreement about the survey name was not easy and necessitated extensive discussions between the survey areas and Methodology about the potential impact on response. In the end, it was agreed to add sector-specific survey notes on page 1 and a description of the purpose of the survey on page 2. As an example, the notes for the production sector are:

The information you supply contributes to the Index of Production (IoP), which shows changes in the manufacturing sector output and is a key measure of manufacturing companies' contribution to the economy. The Bank of England and HM Treasury use GDP and the IoP as key indicators to monitor and forecast economic growth and to inform vital policy decisions. These measures are also used by various Trade Associations when making international comparisons.

A general review of the questionnaires was carried out to ensure that each industry received an appropriate questionnaire. In many cases, there was a corresponding questionnaire under the old classification that could be used, but in others the list of components to include and exclude in turnover was changed or some questions were removed. The latter led to an additional benefit: in questionnaire types where the number of questions was small, Telephone Data Entry (TDE) became the default mode of collection. Also, the small number of questionnaire types has made the extension to the scope of TDE much easier.

Questionnaires for the newly in-scope industries have been tested using cognitive testing methods on a small purposive sample of businesses. An important aspect that has not been considered as part of this work is the harmonisation of the turnover question between the MBS and the Annual Business Survey. This is being considered as part of a wider harmonisation project that is currently under way in the office.

2.2.3 Sample Design and allocation

The samples for each survey were stratified by industry grouping and employment size, where the industry grouping tended to be at the most detailed level in the industry classification system. In MBS, it was agreed to use, broadly, the grouping used in National Accounts, which resulted in a substantial reduction of industry groups in the production and services sectors. In these two sectors and retail, four employment size bands were used in each industry grouping, with the stratum containing the largest businesses being completely enumerated. Different sets of size band boundaries were used in each survey, and in some surveys more than one set was used. Although operationally convenient, using a single set of size band boundaries in all MBS would not be appropriate as, for example, the largest businesses in the production sector tend to be smaller than the largest ones in the services sector.

For practical reasons, it was decided to limit the total number of size band sets to eight; therefore, we had to determine the eight best combinations that could cover all industries in MBS. Different rules, such as ‘cumulative root f’ and ‘cumulative root x,’ were used to obtain optimal boundaries in each industry. After an examination of the different sets, a compromise was reached on the best eight sets to be used. In addition to the four size bands, in the

retail and services sectors there was a completely enumerated stratum for businesses with small and medium employment and high turnover. This band was introduced in the production and construction sectors under MBS.

The sample allocation between strata was done jointly in the production and services sectors, but separately in retail and construction. Previously, Neyman optimal sample allocation was used, where the objective was to minimise the total variance under a fixed sample size (see, for example, Cochran (1977)). In MBS, it was decided to allocate the sample so that the coefficients of variation (CVs) at some specified levels of aggregation do not exceed set targets. This problem can be expressed as a multivariate allocation problem, where the objective is to minimise the total sample size subject to constraints on precision (see, for example, Särndal *et al.*, (1992, page 470)). The problem can be formulated as the minimisation of a convex function subject to a number of linear constraints; obtaining the optimal solution can be computationally very intensive but fast heuristics that give good sub-optimal solutions are available. We used the solution developed at the Australian Bureau of Statistics (see Preston (2004)) and implemented in a SAS macro.

Specifying target CVs was not easy: the survey areas were not able to say which levels would be satisfactory to their users. So, as a guide, we estimated the CVs, at the publication groups, that would result from a Neyman allocation at the overall level, and examined their distribution. It turned out that the majority of the CVs at the lowest level of aggregation in the publication groups were quite low (below 5%); so we used 5% as the target CV at this level. Another consideration was that the overall CV should be close to the CV under the Neyman allocation. Applying the multivariate macro with these targets, and other targets at other levels of aggregation, resulted in a sample size larger than the maximum available; therefore, we increased some of the targets. After a few iterations, we obtained an allocation with a total sample size very close to the available size and resulting in CVs that the business areas found satisfactory. The rebalancing of the sample, compared with that obtained using the Neyman allocation, was small in most strata but notable in a few.

2.2.4 Editing and Imputation

The editing rules for similar variables were harmonised across all industries in the services and production sectors, and any rules that were industry-specific were reviewed and changed to accommodate the new industry classification. Selective editing, which identifies records with potential errors that could have a big impact on the estimates if uncorrected, has been used in these sectors but the parameters for the method had not been updated for a number of years. The redesign of the survey presented us with an opportunity to review the selective editing parameters using a common criterion in all industries: the maximum percentage bias was set at 1%—see Hooper *et al* (2011). Selective editing has also been introduced into the retail sector, and it has seen a 20 percentage point fall in the percentage of records identified for re-contact to correct potential errors. Selective editing in the construction sector will be considered at a later date.

Imputation is used in all sectors to deal with unit and item non-response, but the imputation classes were defined differently in each survey. Because of limitations in our processing system, the definition of imputation classes needed to be the same across the full survey. As part of the integration, the production and services would be processed together, and this meant that we had to choose a single definition of imputation classes. The methods used in production and services sectors, and a few other variations, were tested using past data, and it was found that the method used in the production sector offered the best compromise and was therefore adopted.

2.2.5 Estimation

In all sectors, ratio estimation was used for both turnover and employment: turnover from the business register was used as an auxiliary for financial variables, whereas register employment was used for employment variables (total employees, full-time and part-time employees and gender splits). Other financial variables, such as exports and new orders, use the same method as turnover. In the distribution and services sector, calibration takes place within each sampling stratum, which means that separate ratio estimation was used, whereas in the production sector combined ratio estimation across the sampled strata within each industry grouping of the stratification was used. In retail, the separate ratio estimator was used in some industries and the combined ratio estimator was used in others.

For MBS, it was decided that the default position would be to use the separate ratio estimator—the combined ratio estimator should only be used when a stratum has a small sample size and there is a ‘similar’ stratum to combine it with. Empirical investigations were used to decide where combined ratio estimation could be used.

2.2.6 Results Processing and Publication

As was mentioned above, the production and services sectors are now fully integrated under MBS, including their data being stored on a single database and processed together. Also, the combined results processing for both sectors have moved under the same team, which necessitated a number of changes in operation: this gave us an opportunity to eliminate inconsistencies in timing and methods of delivery of turnover data to National Accounts. There is now a harmonised approach for the generation of data to inform early estimates of GDP (this was previously only carried out in the Services sector) and to revising results across all industries.

The principal use of the survey outputs being inputs to the Index of Production and Index of Services, only a limited set of survey outputs was published but they appeared in two separate publications: one covering ‘Engineering,’ which is part of the production sector, and another covering the services sector. However, since January 2010, a new joint publication with a wider coverage of production, called *Turnover, Orders in Production and Services Industries*, was launched. Also, a new briefing strategy, covering both growth and revisions, has been introduced in MBS, leading to consistency across industries in the production and services sectors.

Because of time pressures, the integration of the retail and construction sectors with the production and services sectors will be considered at a later date. We expect that this will be more challenging than the integration of the production and services sectors: retail data are for different periodicities (five weeks, four weeks, four weeks), as opposed to a calendar month, and require more complex calendar adjustments, whereas for the construction sector we collect data for variables that are very different from those in the other sectors. We need to consider carefully the trade-offs between the benefits of combined processing and the potential costs in terms of loss in flexibility and increased complexity of systems.

3. Implementation of the redesign

An offline, test version, of the processing system was configured, based on the new sample design, questionnaire type, revised editing rules, imputation and estimation methods. This meant that full testing could be carried out in advance of going live in January 2010. Also, a volume test of the combined database was carried out successfully.

The change in classification and design meant that many more businesses were introduced to the sample than would normally be the case under the sample rotation rates in place—the office uses rotational sampling, with the overlap controlled by the use of Permanent Random Numbers (PRNs). Efforts were made to ensure that the sample overlap was adequate by specifying an appropriate PRN start in each stratum. Even with these efforts, the number of businesses with no previous return was higher than usual, which made data validation difficult—the editing rule that compares a current return against the previous return could not be used in many more cases than usual.

There were also some challenges with businesses that moved between sectors as they received a different questionnaire type. For example, a business moving from retail, where ‘retail sales’ are collected, to the services sector, would be requested to return ‘total turnover,’ which led to more time spent on data validation.

Although MBS adopted the new classification from January 2010, National Accounts needed data under the old classification for a period of nearly two years. This necessitated writing a new processing system and testing it, which put more pressure on everyone. Also, running the two systems in parallel was found to be difficult and resource-intensive by the survey areas: training was organised but a lot of support was still needed in the first few months.

MBS was launched in January 2010 and early warning letters were sent to respondents; this ensured that things went smoothly with no major problems reported. The main issue arose when respondents contacted the office to clarify things: there was some confusion when the respondents were referring to MBS whereas our staff were still thinking in terms of the old survey names. Although the changes to the surveys were communicated to staff, the training

provided to some areas was not sufficient. More emphasis was given to the production of outputs and communication with external users about the change in the series as a result of the change in industry classification.

The project to create MBS took place in parallel with the project to implement the change in industry classification and the Workforce jobs redevelopment. Each project was managed by a team, but Methodology was represented by the same person in all three projects. This, and the fact that the MBS project team had representatives from all interested parties (Data Collection, survey managers of all relevant outputs, Methodology and Information Management) made communication between the work streams more effective. Feedback about the effectiveness of the group was sought from participants, and changes in operation were made.

Plans were developed early and stakeholders were consulted; this proved useful, although it was difficult to have full engagement from those not directly affected at the time. As a result, last-minute requests for change in the details of the survey were submitted, which put the project members under pressure to deliver under the already very demanding timetable.

4. Conclusion and Future Work

The redesign of all the processes and methods for the production of short-term statistics should lead to more relevant and better quality estimates. Also, the standardization of processes and methods, and the full integration of two sectors, production and services, should lead to efficiencies. The meticulous checking of specifications and the rigorous quality assurance of the systems, the support given to the results team, along with strong project management and coordination, has resulted in a successful implementation of the changes.

The main lessons learnt are that compromises are often needed, stakeholder engagement is important but difficult to secure, and time for last-minute changes should be built into the planning. Also, an evaluation of the changes after one year was planned, and this has proved useful: we are making some changes to the sample designs now that better information has become available.

The work fit well with other projects in the office, and paves the way for future improvements. One area is the harmonization of response-chasing across MBS: we currently use simple prioritisation rules, but they are not consistent across sectors and may not be optimal. The office has been considering this problem over the last few years: work by Southampton University (see Berger, 2009) investigated several scoring methods and tested them in a simulation based on data from the distribution and services sector. The work found that there was interaction between the imputation classes, imputation method and scoring method. Therefore, the office decided to wait until data under the new designs are available for analysis before a decision on which response chasing strategy to adopt is made. Analysis of response patterns is under way. Another important area of investigation is the use of administrative data to produce short-term statistics. This work is being carried out as part of a European project into the modernisation of European economic and trade statistics (ESSnet on the Use of Administrative and Accounts Data in Business Statistics). The main challenge in this work is in dealing with the mixture of periodicities of the administrative data and their timeliness (see Orchard *et al.*, 2011). A third area concerns the use of electronic data collection, which covers two aspects: the first is the use of internet questionnaires, and the second is to obtain direct data-feed from businesses. The office is investigating the feasibility of data-feed from pay roll data in the first instance.

It remains a challenge to pull together the different strands of work and develop coherent plans to make the required changes to our systems and processes to implement the methods that will come out of ongoing and future work.

References

- Berger, I.G. (2009), "Priority Response Chasing", unpublished technical report, University of Southampton, UK.
- Cochran, W.G. (1977), *Sampling Techniques*, 3rd edition, New York: Wiley.
- Hooper, E., Lewis, D. and C. Dobbins (2011), "The application of Selective Editing in the ONS Monthly Business Survey", *Survey Methodology Bulletin*, Office for National Statistics, UK, No 68, pp. 1-11.

James, G. (2010), “Improving the Design of UK Business Surveys”, paper presented at the European Conference on Quality in Official Statistics, Helsinki, Finland, available at http://q2010.stat.fi/media/presentations/session-14/james_paper.pdf.

Orchard, C., Moore, K. and A. Langford (2011), “Practices for Using VAT Turnover Data Within the UK to Produce Estimates of Growth and Monthly Turnover”, Technical Report, Deliverable 2.4 for Work Package 4 of the ESSnet on the Use of Administrative and Accounts Data in Business Statistics, available at <http://essnet.admindata.eu/Document/GetFile?objectId=5278>.

Preston, J. (2004), “Optimal Sample Allocation in Multivariate Surveys: An Integer Solution”, unpublished document, Australian Bureau of Statistics.

Särndal, C.-E., Swensson, B. and J. Wretman (1992), *Model-Assisted Survey Sampling*, New York:Springer-Verlag.

Taylor, C., James, G. and P. Pring (2011), “The Development of the Monthly Business Survey”, *Economic and Labour Market Review*, Office for National Statistics, UK, Volume 5, No 2, pp. 95-103.

SESSION 3B

DISSEMINATION AND DATA ACCESS

Eurostat project SICON: Secure Infrastructure for confidential data access and sharing

Dario Buono¹

Abstract

The overall aim of the project is to develop and establish a pilot of infrastructure, services and documentation for accessing EU confidential datasets held in Eurostat by external partners, mainly NSIs in view of integrating MSs and Eurostat processes. The infrastructure is aimed to be used by Eurostat production unit (the main users) interested in developing more integrated EU statistics production processes with key partners. In the first phase, only data viewing and remote actions are envisaged; file transfer function is not going to be implemented. Solutions proposed should ensure security of the data, draw as much as possible on solution already developed at Commission level (DIGIT-RACHEL environment for CITRIX and WEBGATE/INTRAGATE) and be compatible with the IT infrastructure of NSIs. Guidelines and protocols for running and managing the system are to be set up and included in the manual for handling confidential data in Eurostat. Pilot tests are foreseen to take in account the needs of the two related projects Euro Group Register and Decentralised Access for Scientific Purposes. At the end of the project, recommendations will be drawn to scale the infrastructure to the actual needs and will provide a business model for running and maintaining such an infrastructure.

¹Dario Buono, Eurostat, Luxembourg.

Enhanced table production system overview and technical description: Standardizing the production of custom tables for data quality and efficiency

Peter Timusk, Mamdouh Mansour, Éric Pelletier and Eric Turgeon¹

Abstract

An overview and technical description of a new enhanced tabulation system that integrates Statistics Canada's Corporate Business Architecture compliant tools and delivers more effective quality compliance, as well as efficiencies, by standardizing the production of custom tables. This paper discusses the goals of the enhanced system, notes the programming efficiencies that were gained and discusses the role of methodologists in the adoption of standardization strategies.

Key Words: Tabulation; Custom Tables; Standardization Strategies; Programming Efficiency; Quality Compliance.

1. Introduction

1.1 Introduction

An enhanced table production system has been implemented that harnesses the benefits of Statistics Canada's Corporate Business Architecture (CBA) compliant generalized systems components into a Statistical Analysis Software (SAS) environment and also engages a SAS Output Delivery System (ODS) output system in a leading-edge new application (that is also described). This system has been developed and also proven for the past year in Statistics Canada's Business Special Surveys and Technology Statistics Division—whose table production specialists are tasked with maintaining the system's ability to produce reliable, relevant descriptive tabular output for a variety of custom cost-recovery surveys. Specialists are tasked with producing custom tables that pertain to constantly changing subjects, measures, concepts and standards. For this reason, a highly-flexible, agile system was needed to respond to table structures, input specifications and output requirements that change with each and every project. The one constant is the use of Generalized Estimation System (GES) and CONFID2/G-CONFID, as well as standardized business processes that the unit uses to support survey managers.

1.2 Overview of the Production Process

To produce custom tables of statistics, we use two Statistics Canada CBA compliant tools. One tool, the GES, allows the calculation of estimates of ratios, means, totals, and counts for survey results. This program also calculates a coefficient of variance or, if needed, a standard error by taking into account the sampling and survey design. The other CBA tool we use, G-CONFID (previously CONFID2), is a Statistics Canada generalized system that offers a methodology designed to prevent the release of confidential data. It allows a suppression pattern of Xs to be computed for a table where the estimate will be replaced with the letter 'X.' This is so that the resulting table cannot be used to determine any individual company or person's responses and ensures only aggregate results will be shown.

In summary, we use our table production system to produce

- custom tables of survey estimates
- reliability indicators (coefficients of variance or standard errors)
- confidentiality (suppression and non-publication of some estimates).

¹Peter Timusk, Mamdouh Mansour, Éric Pelletier and Eric Turgeon, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa ON K1A 0T6.

2. Issues with Past Table Production Systems

There were a number of technical issues with our past table production system. We used a cumbersome graphical user interface for GES that took painstaking editing for large number of variables. There were also system size limits as the programs processed large numbers of estimates at the same time. In tabulation, estimates were placed in Excel tables as character variables instead of as numbers—which is what they were. If an estimate has too high a coefficient of variance or standard error, it receives an ‘F’ for quality and is not published. A letter ‘F’ replaces the estimate in the published table. For estimates that were X’d out for suppression reasons or F’d out for quality reasons, a letter is published instead of the number. For this reason numbers were published as character variables.

After the tables were created, a subject matter expert was tasked with checking the quality of the tables. This was a manual quality control (QC) process and was prone to some error.

The table production section dealt with custom tables of an ad hoc design from ad hoc surveys, no simple standard processing system was designed to accommodate these requests. More often than not, the table designs were complex cross tabulations which required complex programming needed to create the estimates and complete these tabulation outputs which were also complex.

In the past there had been reliance on one senior programmer to complete this complex programming work. This structure and inherent dependency left the section vulnerable.

Additionally, the documentation on these complex tables were either limited to a few lines of comments in the codes or were not in a standard location or readable format. This meant that when a table required data from an older survey a great deal of effort was needed to find out how the older survey had been tabulated and what the exact specifications were in the older survey.

3. Redesign of the Table Production System

It was decided that the table production system would be enhanced. First, a good model of the present table production system was needed. Business system analysts were engaged to document production systems, processes and procedures via formal Business Process Modelling (BPM). In this change to the system, the team making the change worked closely with production staff to document what is done and how in a formal BPM framework

A systems developer was employed to develop a set of SAS Macros to allow a simplified approach to GES and CONFID2.

4. Enhanced Table Production System: Summary of Changes, Overview and Example

4.1 Summary of Changes

In the new enhanced table production system (ETPS), the approach was to use simplified programming (one program per table). The design of the tables was simplified by having the clients design the tables with fewer variables per table. The design of the tables was migrated to the clients rather than being completed by the production staff. Transfer of table design to the survey analyst produced formal written agreements on table specifications. This then lead to written confidentiality and reliability specifications.

Through the design of the tables and the simplification of the programming, a more modular approach is now used. The system is now more flexible for multiple ongoing custom requests of tables. A repeatable process and procedure is used in the new system.

As well, it has been possible to implement a number of standardizations in the new system involving naming conventions. These have been successfully applied to file naming and questionnaire variable naming.

A more up-to-date tabulation system can now be used with SAS ODS, which is also used in other ways to compute frequency datasets for suppression by counts when CONFID2 is not used. Using SAS ODS and SAS formats allowed the use of SAS ‘special missings’ to code quality letters and Xs as numerical variables. This allows the estimates which are numbers to remain in a numerical format when tabulated.

SAS programs are now regularly documented with ‘comments’ that allows other programmers to clearly understand code and allows other programmers to easily locate code by program sections. This has enabled programs to be worked on by multiple programmers with different skill levels; typically this means one junior programmer and one senior programmer working together.

This has allowed more programmers to write and edit programs and for a general rise of the section’s SAS programming skills. A great deal of education and extensive training was applied to the section’s programmers. As well, the section experienced an influx of experienced SAS programmers. Dependency on a limited number of programmers is no longer a risk.

A tool was introduced that allows the final output Excel tables to be quality checked against independent estimates of the same values. With this tool, quality control is now fully automated and a 100% comparison can be made for all estimates for thousands of tables and cells.

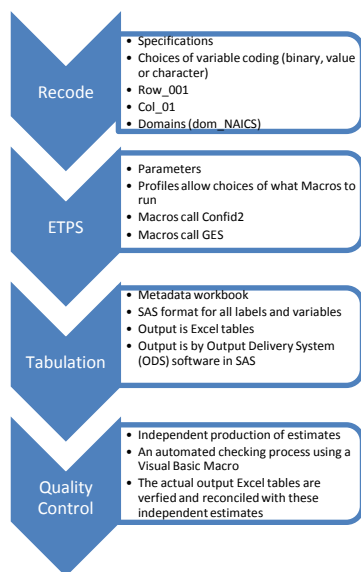
4.2 ETPS Overview

Programmers are now responsible for recoding of survey data into datasets for input into the CBA business processes and this is still completed by the table production team. This can be completed by all levels of programmers in the team.

- Specifications are detailed and provided by the client in ‘spec tables’.
- Parameters are used in both the ETPS core and in the tabulation system.
- There were real gains in efficiency that could be made by having proper documentation.

The following diagram shows the steps now followed:

Figure 4.2-1
ETPS Process Overview



4.3 ETPS Example

For the spec table, the client specifies the table design and identifies the variables in database (often a survey results file). The client further identifies values of variables for all rows and columns and then any populations and subpopulations. The following three diagrams give an idea of how this is completed by the client. Although the tables appear the same in each diagram, each concerns a different specification from a point of view of the programming that will be completed based on this specification.

Figure 4.3-1
Specification Example – Variables

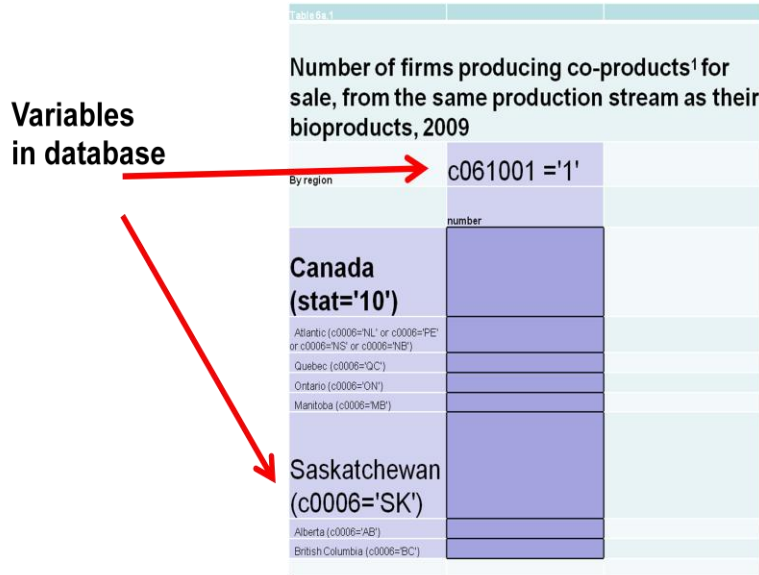


Figure 4.3-2
Specification Example – Rows and Columns

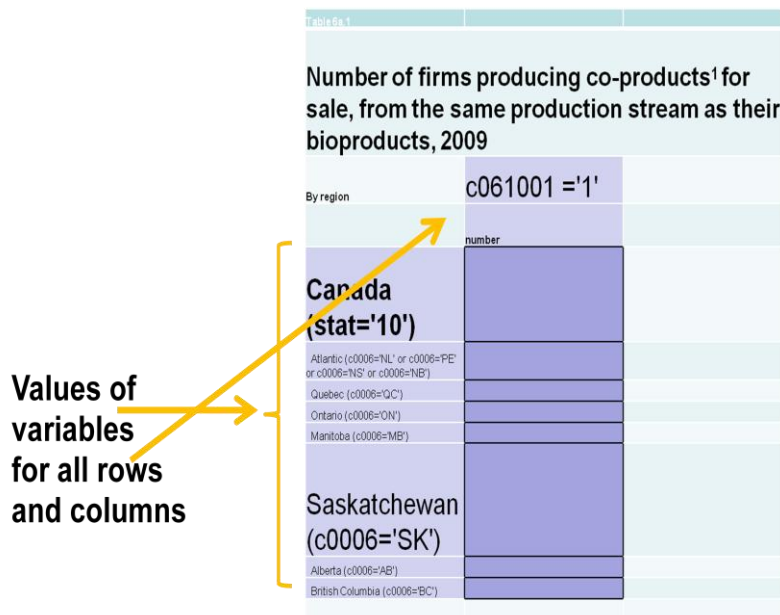


Figure 4.3-3
Specification Example – Populations

Table 4.3-1

Number of firms producing co-products¹ for sale, from the same production stream as their bioproducts, 2009

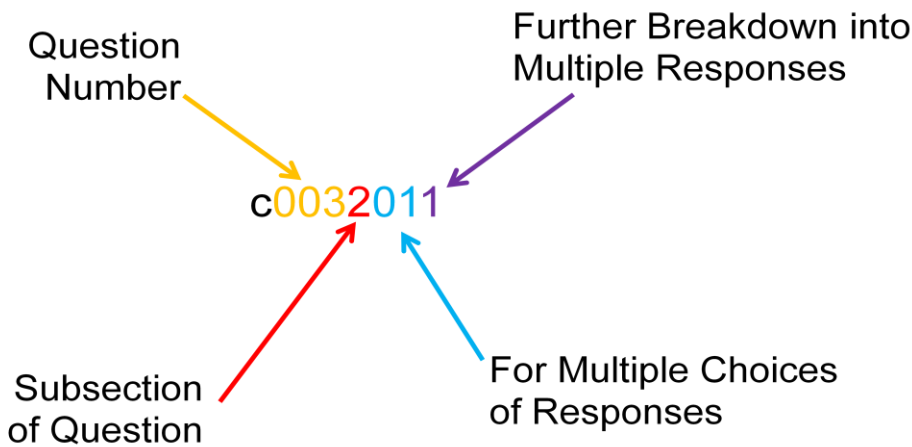
By region	c061001 = '1'	
	number	
Canada (stat='10')		
Atlantic (c0006='NL' or c0006='PE' or c0006='NS' or c0006='NB')		
Quebec (c0006='QC')		
Ontario (c0006='ON')		
Manitoba (c0006='MB')		
Saskatchewan (c0006='SK')		
Alberta (c0006='AB')		
British Columbia (c0006='BC')		

Populations and subpopulations.

5. Standardization

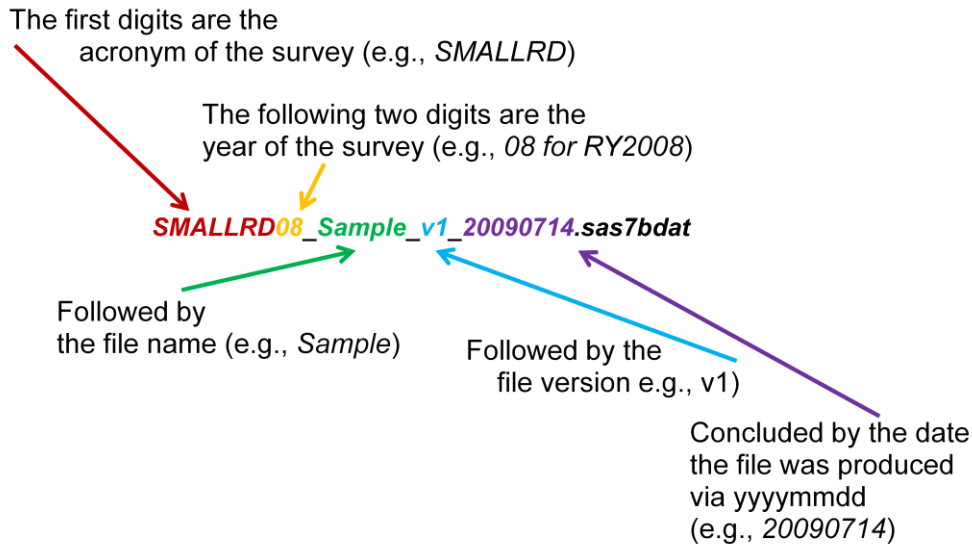
The following two diagrams show examples of standardization of naming conventions that have been adopted in the ETPS. The first shows a standardization of questionnaire cell numbering which is used as the variable name for the variable on the survey results data file that contains the responses to this question.

Figure 5.1-1
Question Numbering Standardization



The second diagram shows the naming convention for data files used in the enhanced table production system. This too was standardized by methodologists and adopted in the table production section. At times when a dataset changes often and the file is referenced elsewhere in many places, the date will not be used in the name, as changing the file name would require too many downstream edits and the date a file is produced will change often in a production project.

**Figure 5.1-2
File Naming Standardization**



6. Data Quality

There was an effort made to automate and document QC. In this context, QC means independent calculation and comparison of estimates. A Visual Basic macro was identified at Statistics Canada that enables cell-by-cell automatic comparison for any two Excel output tables. The macro generates an output log file of differences in estimates and reliability values. A system developer helped transform the macro into a robust application suiting the enhanced table production system needs. Quality control is thus now fully automated and a 100% comparison can be made for all estimates for thousands of tables and cells.

7. Effectiveness, Efficiencies and Work Distribution Advantages

Programming work is now assigned to different programmers determined by the table's complexity.

We now implement a 100% automated check of programmed output by using independent tabulations: One by the programmer in production of the tables; and a second independent tabulation by the subject matter officer. By constructing one program per table we reduce programming complexity and the load on processing systems such as GES.

As well, the use of standardized naming conventions has increased efficiency.

8. Remaining Challenges

Additional business processes are not yet integrated into ETPS. For example, SEVANI, a generalized system to compute variances due to imputation, is not yet integrated into the macros and profiles available in the ETPS. The latest data confidentiality features in G-CONFID are not yet integrated into the ETPS. The system was built for the CONFID2 version of CONFID.

The ETPS QC process is now enhanced by checking estimates and changes to coefficients of variance, if independent GES runs are also performed. We can still go further with more frequent use of automated processes to check suppression patterns by counts. We are starting to use independent comparison of counts from SAS Proc Freq to SAS Proc Tabulate to check counts of contributors for suppression reasons.

In integrating the ETPS system with CBA complaint tools we ensure corporate support for the systems used to make tables. In the long term both base funding and ongoing cost-recovery projects will benefit from this initiative.

9. Conclusions

Our new ETPS simplifies production by enabling multiple programmers to collaborate efficiently. It improves quality control by automating the process and enables us to complete a 100% QC despite volume of programmed output.

We now have restructured the use of GES and CONFID2, so that our work can focus on input dataset creation. We have also used the ETPS while adopting standardized naming conventions that assists with efficient, error-free identification of files, and identification and resolution of issues.

10. Acknowledgements

The authors wish to thank everyone who worked in the Enhanced Table Production Team for all their work on the project. Also Frances Anderson provided the authors with some key guidance for structuring the presentation and telling the story for the enhanced table production project. As well Paula Thompson and Greg Peterson reviewed the final drafts of the presentation slides and George Sciadas review the proceeding's paper. The authors are indebted to these reviewers for their assistance.

SESSION 4A
SELECTIVE EDITING

Selective data editing and its implementation at Statistics Sweden

Pär Brundell¹

Abstract

Selective data editing can reduce the often resource demanding editing process in business surveys. Therefore a generic software for selective (micro) editing, Selekt, has been developed at Statistics Sweden. It now exists in version 1.2 and the current and near future plan is to implement Selekt in half a dozen business surveys and in the process gain useful experiences and knowledge to build on. This paper will give a brief insight of Selekt and some of its underlying theoretical base as well as a mentioning on the implementation work in some business surveys.

Key Words: Selective data editing; SAS macro Selekt.

1. Background

1.1 Errors

Errors in survey data can arise in raw data delivered by respondents to the statistical agency or in data transmissions. The statistics production process is a mixture of many activities with risks of introducing errors. Types of errors could, for example, consist of item non-response, non-valid values, model errors or contradictions. Suspected data values could be divided into two types; suspected deviation errors (outliers) and definition errors (inliers) when many respondents misunderstand a question or when data is fetched from info systems with other definitions than desired. Suspected deviation errors often require manual follow-up which takes time and is expensive. In many cases, the edits used to detect such errors have a low hit-rate and many of the changes made in data have very little impact on the final statistics. Definition errors could be difficult to find and some ways of finding them could be combining the editing for several surveys, deep interviews in focus groups, look for high proportions of item non-response, or use graphical editing.

1.2 Editing

Editing is an activity of detecting, resolving and understanding errors in data and produced statistics. The editing activities generally involve at least some of the following operations:

- A. Respondent editing
- B. Manual editing before data registration
- C. Data registration editing
- D. Production editing / micro editing
 - 1 'Traditional' editing
 - 2 Selective editing
- E. Coherence analysis
- F. Output editing / macro editing
- G. Evaluation
- H. Delivery control

The activities are many but a large effort is often spent on the production editing (D above) and it is here that selective editing can hopefully reduce the amount of work at the statistical agency. Traditional editing as opposed to

¹Pär Brundell, Statistics Sweden, Klostergatan 23, Sweden, SE-70189.

selective editing often consists of finding and setting acceptance limits in edits. Such limits could be derived from previous survey rounds. Selective editing can use such traditional edits and a score function in order to select only suspected errors of the largest expected importance. With the purpose of a large-scale implementation of selective data editing at Statistics Sweden a tool, Selekt, has been developed. It consists of SAS macros and through a ‘control panel’ where the user can set necessary and optional parameters to design the selective data editing for a data set defined by the user.

2. Selective data editing

2.1 Basic concepts

The purpose of selective data editing is to reduce the cost for the statistical agency and for the respondents, without a significant decrease of the quality of the output statistics. A selective editing approach could be to target some of the microdata records (observations). Each record would firstly be classified as either having suspected errors or not being suspected. Then, the potential impact on all statistics of each suspected record would be calculated and only the records with the most potential impact on the final statistics would be flagged for manual follow-up.

The selective data editing approach at Statistics Sweden, however, involves taking what is described in Figure 1 one step further. Instead of a dichotomous suspicion value it is allowed to be continuous, with values between zero and one. The potential impact on output is initially calculated for each record (observation) and for each variable. The expected impact of the record and variable is the product of suspicion and potential impact. The flagged observations would then be those with the highest aggregated values of expected impacts.

The selective data editing approach consists of constructing a score function that prioritizes variables and records. The score is computed at the first stage by multiplying the suspicion with the impact. The suspicion ranges continuously from zero to one and the impact is always an absolute number since the direction of the impact is not of interest in this approach.

Figure 1. A first approach to selective data editing could be to flag suspected observations with the largest potential impact (the upper right square in the figure)

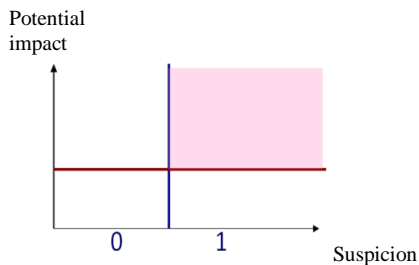
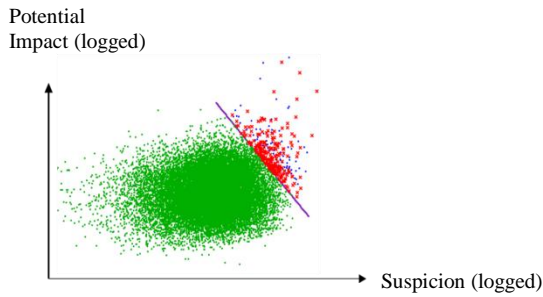


Figure 2. The approach to selective data editing at Statistics Sweden involves continuous impact and suspicion



2.2 Suspicion

The suspicion can be set individually for each record (*i.e.*, observed unit) and each variable. The suspicion extends continuously from zero (no suspicion) to one (fully suspected). The basis for computing the suspicion is the comparison between the unedited value that, for example, is sent in by the respondent and its expected value and variation. The method in Selekt for obtaining the expected value can be chosen by the user through one of two main alternatives. Either the expected value is obtained a) from analysing the time series of the survey variables for each observed unit or b) from computing the mean or median for each survey variable within edit groups defined by the user of Selekt. There is a built-in facility in Selekt that, based on the defined edit groups, obtains the mean or median at the lowest possible level in terms of number of observations. This is done through a tree analysis methodology.

The suspicion in Selekt is defined as $Suspicion = R / (\text{Tau} + R)$ where R is defined as

$$R = \begin{cases} (\tilde{z}_{j,k,l} - KAPPA \cdot (\tilde{z}_{j,k,l} - \tilde{z}_{j,k,l}^L) - z_{j,k,l}) / (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l}^L) & \text{if } z_{j,k,l} < \tilde{z}_{j,k,l} - KAPPA \cdot (\tilde{z}_{j,k,l} - \tilde{z}_{j,k,l}^L) \\ 0 & \text{if } \tilde{z}_{j,k,l} - KAPPA \cdot (\tilde{z}_{j,k,l} - \tilde{z}_{j,k,l}^L) < z_{j,k,l} < \tilde{z}_{j,k,l} + KAPPA \cdot (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l}) \\ (z_{j,k,l} - \tilde{z}_{j,k,l} - KAPPA \cdot (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l})) / (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l}^L) & \text{if } z_{j,k,l} > \tilde{z}_{j,k,l} + KAPPA \cdot (\tilde{z}_{j,k,l}^U - \tilde{z}_{j,k,l}) \end{cases}$$

The above formula can be found in Norberg, A. *et al.* (2010, p 22 under definition 4.1). The parameters Kappa and Tau are set by the user. Kappa could be said to define the 'gap' of the acceptance range. A small Kappa gives a suspicion larger than zero already at a small deviation from the expected value. Tau is used to adjust the suspicion. If Tau for example is 0.001 then the suspicion becomes either zero or one. If Tau is larger for example 10 then the suspicion becomes proportional to the distance from the middle of the distribution. In order to compute the suspicion in Selekt the following quantities are needed:

$\tilde{z}_{j,k,l}^L$ = Lower limit

$\tilde{z}_{j,k,l}^U$ = Upperlimit

$\tilde{z}_{j,k,l}$ = Expected value

$z_{j,k,l}$ = Unedited value

Below are some graphs of the suspicion for different settings of the parameters Kappa and Tau. These examples are based on a set of historical observations but the principle of the suspicion characteristics based on the settings are general.

Figure 3. When Kappa is equal to zero the suspicion is almost always larger than zero

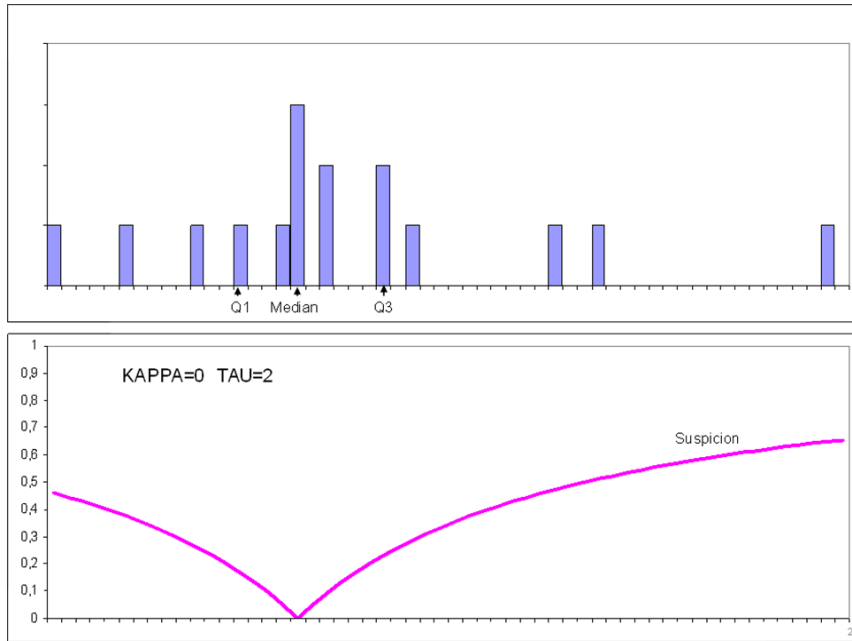


Figure 4. Here, Kappa is equal to one instead of zero as above

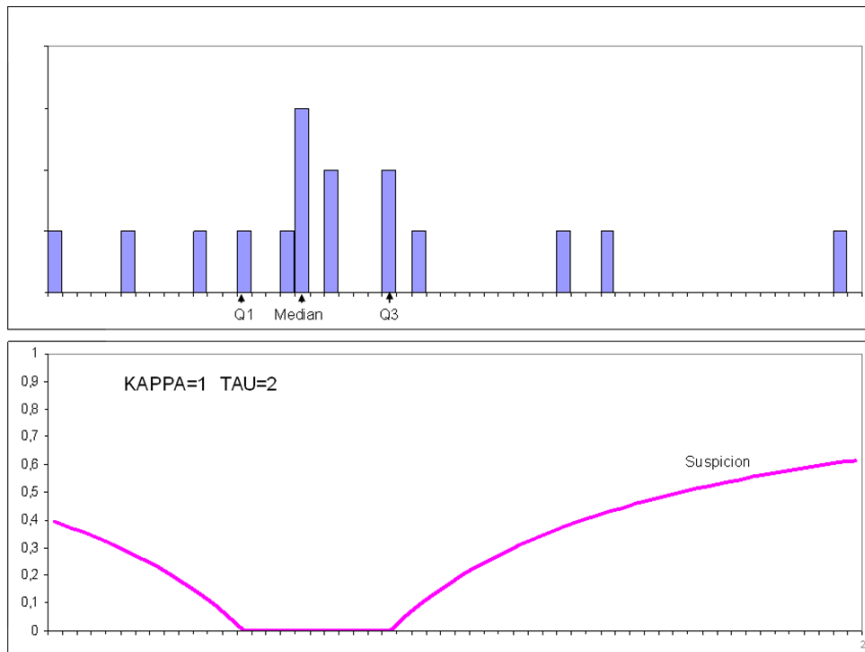


Figure 5. Kappa is equal to one and Tau is equal to five

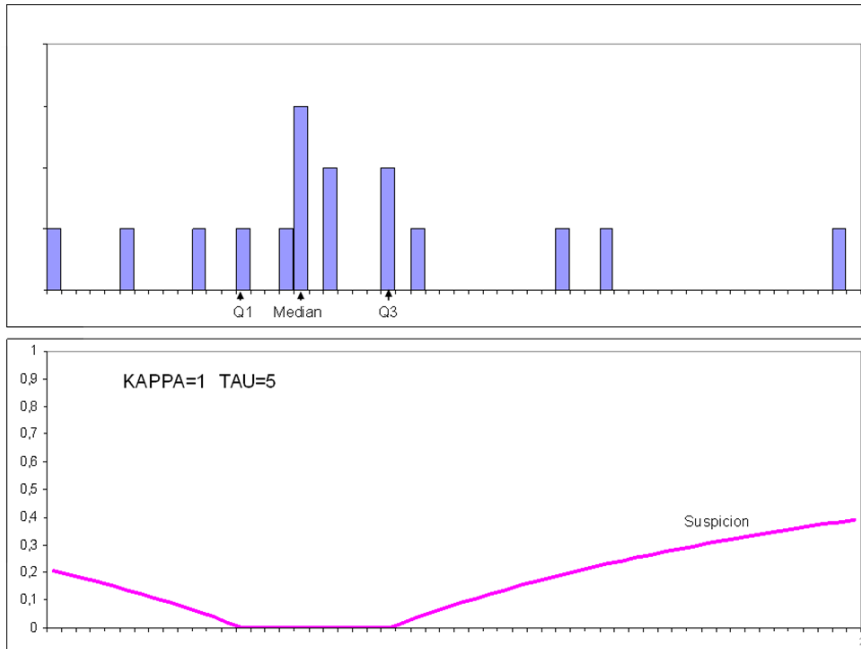
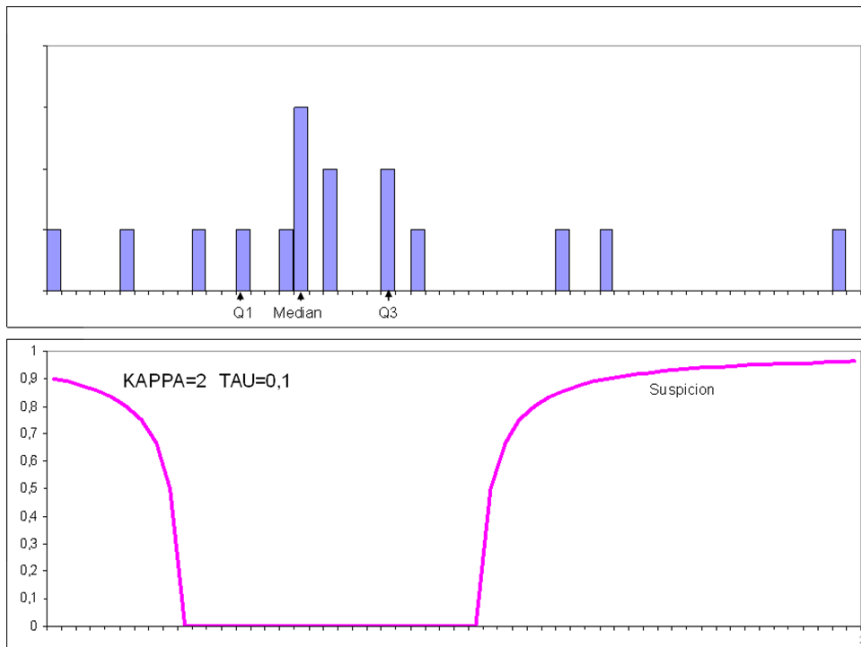


Figure 6. When Tau is small the suspicion rises abruptly



2.3 Impact

The idea of the impact is to measure how much the unedited instead of the ‘corrected’ value (or at least what would be considered to be the correct value) would affect the output statistics. If there is no reason to think that there is an error in the supplied value, the impact would be zero. If, on the contrary, the unedited value would differ from the correct one, the absolute value of the impact on the output statistics would be larger than zero. In our terminology there are in fact several sorts of impact. The subscript *une* denotes the unedited value, *i.e.*, the value obtained from

the survey respondent before any editing. The subscript *edi* denotes the edited value, *i.e.*, the corrected value. In many cases the unedited value is the same as the edited one. The actual impact is defined as $w (y_{une} - y_{edi})$ which for an observation is the impact on the estimated domain total of variable *y* if y_{une} is kept instead of making a review to find y_{edi} . The impact measure contains the survey weight *w* since it is the output statistics or result that is interesting here. The potential impact is defined as $w (y_{une} - y_{pred})$ and is a proxy for the actual impact to be used in practice. The y_{pred} is a prediction, *i.e.*, expected value, for y_{edi} since y_{edi} is not known before a review is done. The expected impact (per domain, variable, observation) is the product of the suspicion and the potential impact.

2.4 The score function

The local score produced by multiplying the suspicion and the potential impact can be further enhanced by adding parameters concerning the importance of variables or domains. The parameter $VIOLIN_j$ (default=1) adjusts the importance of variable *j* by multiplying the score by this factor. The parameter $CLARINET_{c(d)}$ (default=1) adjusts the importance of classification *c*, which defines the domains *d*, by similarly multiplying the score by this factor. The parameter $OBOE_j$ is an adjustment for the size of the estimated total or its standard error for variable *j*. The parameter $CELLO_ALFA$ relates the impact to either the estimate of the total or the standard error of that estimated total. The $VIOLIN$, $CLARINET$, $OBOE$ and $CELLO_ALFA$ are all part of the whole adjustment factor called $CELLO$. The product of the suspicion and the potential impact is hence multiplied by the $CELLO$ factor so that the score is equal to $suspicion_{j,k,l} \times potential\ impact_{d,j,k,l} \times CELLO_{d(c),j}$ (for domain *d*, variable *j*, primary sampling unit *k* and secondary sampling unit *l*). The formula for $CELLO$ is

$$CELLO_{c(d),j} = \frac{VIOLIN_j \times CLARINET_{c(d)}}{\left(maximum \left\{ CELLO_ALFA_j \times \hat{T}_{d,j,t0}, SE(\hat{T}_{d,j,t0}) \right\} \right)^{OBOE_j}} .$$

The local scores by domain, variable, second stage unit (if any) and primary sampling unit are aggregated up to global scores for the primary sampling unit and finally up to respondent unit if needed. *Selekt* offers a choice of the aggregation method. The alternatives are currently to use the sum, the sum of squares or the maximum. At every aggregation level (domain, variable etc.) *Selekt* offers the user to set a threshold so that the score in fact at each aggregation level is the maximum of zero or the score minus the threshold.

3. Implementation of selective data editing at Statistics Sweden

The objective of implementing selective data editing at Statistics Sweden is to reduce the editing work without significantly decreasing the quality of the output statistics. The objective is also in the long run to cut costs even if a certain investment has to be made in implementing selective editing in a survey. The following list gives a brief indication in which surveys selective editing has been implemented so far at Statistics Sweden.

Survey	Implementation stage
Wage & salary structures in the private sector (SLP)	Selective editing using <i>Selekt</i> has been implemented
Business activity indicators (Kortind)	Selective editing using <i>Selekt</i> has been implemented
Short-term statistics, wages and salaries, private sector (KLP)	Selective editing using <i>Selekt</i> has been implemented
Commodity flow survey (VFU) 2009	Selective editing using <i>Selekt</i> has been used in the survey
Foreign trade - exports and imports of services (UHT)	Selective editing using <i>Selekt</i> has been implemented
Rents for dwellings (HIB)	Selective editing using <i>Selekt</i> has been implemented
Short-term employment (KS)	Selective editing using <i>Selekt</i> is planned to be implemented beginning of 2012
Turnover and inventory statistics for the service	Selective editing using <i>Selekt</i> is planned to be implemented beginning of 2012
Revenues and expenditure of multi-dwelling buildings (IKU)	Selective editing using <i>Selekt</i> is planned to be implemented beginning of 2012

References

- Granquist, L. (1995), "Improving the Traditional Editing Process", in Cox *et.al.*, (eds.) *Business Survey Methods*, Wiley.
- Jäder, A. and A. Norberg (2006), "A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics", *Background facts on Economic Statistics 2006:3*, Statistics Sweden.
- Norberg, A. and *et al.* (2010), "A General Methodology for Selective Data Editing", version 1.0 2010-02-04, unpublished report, Statistics Sweden.
- Norberg, A. and *et al.* (2011), "User's Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing", unpublished report, Statistics Sweden.

SeleMix: An R package for selective editing via contamination models

Marco Di Zio and Ugo Guarnera¹

Abstract

The aim of ‘selective editing’ is to identify errors with high impact on the target estimates in order to correct them through an accurate interactive editing procedure. This task is usually accomplished by means of ‘score functions’ that express the importance of errors affecting observations.

Score functions are generally defined on the basis of the analysis of residuals with respect to some predictions of the data. Models used in the traditional methods generally do not take explicitly into account the ‘intermittent’ nature of the error mechanism. As a consequence, it is difficult to distinguish the component of the observed variability associated with the natural deviations from a mean behavior, from the component due to the presence of measurement errors.

In order to overcome this difficulty, an approach based on explicitly modeling both data and error mechanism has been recently proposed. It is based on the use of contamination models, that, is a latent class model where the latent variable is to be interpreted as an indicator variable for the error occurrence. This formalization allows us to relate the values of the score function to the expected errors in data.

In order to ease the use of contamination models, the R package SeleMix has been recently implemented. The paper introduces the model and shows the main functionalities of SeleMix.

Key Words: Influential Errors; Mixture Models; Score Function.

1. Introduction

Selective editing is based on the principle of looking for units affected by significant errors so that accurate editing procedures can be applied only to this subset. The aim is to reduce the cost of the editing and imputation phase maintaining an acceptable level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994). In practice, observations are prioritized according to the values of a score function measuring the importance of the potential errors that they contain (Latouche and Berthelot, 1992; Hedlin, 2003). The units, whose score is above a given threshold, are selected to be accurately edited.

The most commonly used methods to determine the scores are based on the comparison of observed with predicted values (Hedlin, 2008). This residual is composed of the possible error and the natural variability of the analysed quantity. In the usual setting, there is no possibility of distinguishing these two elements, hence the score of an observation is not directly related to the expected error of that unit. As a consequence, the value of the threshold will not be directly interpretable as the level of accuracy of the final estimates and a stopping rule for determining the subset of units to be selected will be available only in a simulative context when edited (considered as true) and raw data of an outdated survey are available (de Waal *et al.*, 2011).

Di Zio *et al.*, (2008) have proposed to use a latent variable model allowing, under certain assumptions, to estimate the expected error associated with each unit. In this setting the threshold value is directly interpreted as the level of accuracy of the edited data allowing to establish a stopping rule related to the amount of error left in data. The method is based on the use of contamination normal models where the erroneous data are assumed to follow the same distribution as the error free data but with inflated variance (see Ghosh-Dastidar and Schafer, 2006). Results of experiments on both simulated and real data show that selective editing based on contamination models is useful in many context (Bellisai *et al.*, 2009; Buglielli *et al.*, 2010).

¹Marco Di Zio, Istat - Istituto Nazionale di Statistica, Roma - Via Cesare Balbo 16, Italy, 00164 (e-mail: dizio@istat.it); Ugo Guarnera, Istat – Istituto Nazionale di Statistica, Roma - Via Cesare Balbo 16, Italy, 00164 (e-mail: guarnera@istat.it).

In order to ease the usage of the method, an R package named SeleMix has been implemented. It is freely available on the R website. The functions included in the package allow to apply the procedure also to data affected by missing values. In this case, for each incomplete record, missing items are replaced with their expected values conditional on the observed ones. Expectations are computed according to the contamination model, so that the predictions take into account the possible presence of errors in observed items. In this sense, the contamination model approach can be used also as a “robust” imputation method.

The paper is structured as follows. In Sections 2 and 3 we shortly describe the contamination model and the selective editing approach respectively, more details can be found in Buglielli *et al.*, 2011. Section 4 is devoted to the illustration of SeleMix with its main functionalities.

2. The Model

The key elements of the approach presented in this paper are: 1) specification of a parametric model for the true (non-contaminated) data, and 2) specification of an error model. This allows us to derive the distribution of the true data conditional on the observed data. This distribution is central in the proposed selective editing method. An important point is that the model specification reflects the intermittent nature of the error mechanism. This means that errors are assumed to affect only a subset of data, or in other words, each unit in the dataset is affected by an error with an (unknown) a priori probability. The assumption of intermittent error, which is very common in the context of survey data treatment, naturally leads to the model specification of the error model in terms of a mixture of probability distributions. As a consequence, the observed data distribution is also a mixture whose components correspond to error-free and contaminated data respectively. Such models are often referred to as contamination models and are commonly applied in the context of outlier identification. In the following, the model is described in some detail.

2.1 True Data Model

We assume that two sets of variables are observed: the variables of the first group, say X -variables, are assumed to be correctly measured while the second set of variables, say Z -variables, corresponds to items possibly affected by measurement errors. In this set-up, which can be useful when some variables are available from administrative sources or are measured with high accuracy, it is quite natural to treat the variables that are observed with error as response variables and the reliable variables as covariates. This framework includes as a special case, the situation where reliable covariates X are not available, so that what is to be modelled is the joint distribution of the Z variables.

In the following we model true data through a lognormal probability distribution. This seems a reasonable assumption in many cases where economic data are to be analysed.

According to the previous assumptions, true data corresponding to possible contaminated items are represented as a $n \times p$ matrix Z^* of n independent realizations from a random p -vector assumed to follow a lognormal distribution whose parameters may depend on some set of q covariates not affected by error. Thus, if $Y^* = \ln Z^*$, we have the regression model:

$$Y^* = XB + U \quad (1)$$

where X is a $n \times q$ matrix whose rows are the measures of the q covariates on the n units, B is the $q \times p$ matrix of the coefficients, and U is the $n \times p$ matrix of normal residuals whose i th row U_i is normally distributed with zero mean and covariance matrix Σ :

$$U_i \sim N(0, \Sigma), \quad i=1, \dots, n. \quad (2)$$

2.2 Error Model

In order to model the intermittent nature of the error mechanism we introduce a Bernoulli r.v. I with parameter π , where $I=1$ if an error occurs and $I=0$ otherwise. In the sequel, Z and Y will denote possible contaminated variable in original and logarithmic scale respectively. Thus, given that $I=0$, it must hold $Z=Z^*$ ($Y=Y^*$). Furthermore, given that $I=1$, errors affect data through an additive mechanism represented by a Gaussian r.v. with zero mean and covariance matrix Σ_ε proportional to Σ , i.e., given $\{I=1\}$:

$$Y = Y^* + \varepsilon, \quad \varepsilon \sim N(0, \Sigma_\varepsilon), \quad \Sigma_\varepsilon = \lambda \Sigma, \quad \lambda > 0.$$

It is convenient to represent the error model through the conditional distribution:

$$f_{Y|Y^*}(y | y^*) = (1 - \pi)\delta(y - y^*) + \pi N(y; y^*, \Sigma_\varepsilon) \quad (3)$$

where π (*mixing weight*) is the "a priori" probability of contamination and $\delta(t'-t)$ is the delta-function with mass at t .

In case that the set of X -variables is empty, the variables Y_i ($i=1, \dots, n$) are normally distributed with common mean vector μ . It is worthwhile noting that, due to the intermittent error assumption, it is conceptually possible to think of data as partitioned into true and erroneous, and to estimate, for each observation, the probability of being true or corrupted. The distribution of the observed data is easily derived multiplying the normal density for the true data implied by (1) and (2) and the error density (3), and integrating over Y^* :

$$f_Y(y) = (1 - \pi)N(y; B'x, \Sigma) + \pi N(y; B'x, (\lambda + 1)\Sigma) \quad (4)$$

The distribution (4) refers to observed data and can be easily estimated by maximizing the likelihood based on n sample units via an ECM algorithm.

3. Selective Editing

In order to use the contamination model for selective editing, we need to derive the distribution of error-free data Y^* conditional on observed data (including covariates X).

Straightforward application of Bayes formula provides:

$$f_{Y^*|X,Y}(y^* | x, y) = \tau_1(x, y)\delta(y^* - y) + \tau_2(x, y)N(y^*; \tilde{\mu}_{x,y}, \tilde{\Sigma}) \quad (5)$$

where τ_1 and τ_2 are the posterior probabilities of belonging to true and erroneous data respectively:

$$\tau_1(x_i, y_i) = \Pr(y_i = y_i^* | x_i, y_i)$$

$$\tau_2(x_i, y_i) = \Pr(y_i \neq y_i^* | x_i, y_i) = 1 - \tau_1(x_i, y_i), \quad i=1, \dots, n$$

and

$$\tilde{\mu}_{x,y} = \frac{(y + \lambda B'x)}{\lambda + 1}; \quad \tilde{\Sigma} = \left(\frac{\lambda}{\lambda + 1}\right)\Sigma.$$

It is immediate to derive the corresponding conditional distribution in the original scale:

$$f_{Z^*|Z}(z^* | z) = \tau_1(\ln(z))\delta(z^* - z) + \tau_2(\ln(z))LN(z^*; \tilde{\mu}_{x,\ln z}, \tilde{\Sigma}) \quad (6)$$

where $LN(\cdot, \mu, \Sigma)$ denotes the lognormal density with parameters (μ, Σ) , and for the sake of simplicity, we have suppressed the X -variables in the notation whenever they appear as conditioning variables. Estimation of the distribution (6) is obtained by replacing the estimates of $(\mu, \Sigma, \pi, \lambda)$ resulting from the ECM algorithm to the corresponding parameters.

Once the target distribution (6) has been estimated, “predictions” of “true” values z_i^* , conditional on observed values z_i , can be obtained for all the observations $i=1, \dots, n$ as:

$$\hat{z}_i = E(z_i^* | z_i) = \int z_i^* f_{Z^*|Z}(z^* | z) dz_i^*$$

Thus, for $i=1, \dots, n$ the *expected error* ε_i can also be defined as:

$$\varepsilon_i = (\hat{z}_i - z_i) = \tau_2(\ln(z_i))(z_i - \tilde{\mu}_{x_i, \ln z_i}). \quad (7)$$

In the context of official statistics, estimates of some quantities (such as totals or means) of a finite population U are typically of interest. The contamination model approach can be combined with randomization inference in order to obtain robust estimates. For concreteness, let us suppose that the target estimate is given by the total T_z of the variable Z , i.e., $T_z = \sum_{i \in U} z_i$ and an estimator $\hat{T}_z = \sum_{i \in S} w_i z_i$ is used where w_i are sampling weights attached to each unit of a sample S of size n . A robust version of \hat{T}_z is given by $\hat{T}_z^* = \sum_{i \in S} w_i \hat{z}_i$, where the last estimator is obtained from the previous one by replacing observed values z_i with the predictions \hat{z}_i .

Using the contamination model directly for estimation purposes is appealing, but the results could be too sensitive with respect to departures from the model assumptions. In the present context, we are interested in selective editing. Thus, we define a score function in terms of the absolute value of the estimated expected error according to the contamination model.

This definition is particularly useful in that it makes it possible to estimate the residual error remaining in data after the editing of the units with the highest expected errors. It follows that the number of units to be reviewed can be chosen such that the residual error is below a prefixed threshold. Specifically, the threshold can be defined in terms of ratio between the expected residual error and a (robust) reference estimate such as that obtained from the contamination model itself. In order to define the score function, let us introduce the relative individual error r_i as the ratio between the (weighted) expected error and the reference estimate T_z^*

$$r_i = \frac{w_i (\hat{z}_i - z_i)}{\hat{T}_z^*}$$

Note that the expected error appearing in the previous formula is the product, according to formula (7), of $\tau_2(\ln(z_i))$ and $(z_i - \tilde{\mu}_{x_i, \ln z_i})$ that can be interpreted as a risk component and influence component respectively (Jäder and Norberg, 2005).

The score function is defined as $SF_i = |r_i|$. Moreover, let R_M be the absolute value of the expected residual percentage of error remaining in data after removing errors in the units belonging to the set M (the absolute pseudo-bias, Latouche and Berthelot, 1992):

$$R_M = \left| \sum_{i \in \bar{M}} r_i \right|, \text{ where } \bar{M} \text{ denotes the complement of } M \text{ in } S.$$

Once an “accuracy” threshold η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of SF_i ;
2. selecting the first \bar{k} units for reviewing, where:

$\bar{k} = \min\{k \in (1, \dots, n) \mid R_{M_j} < \eta, \forall j > k\}$ and M_m is the set composed of the first m units.

These two steps guarantee an upper bound not only for the total amount of error remaining in data (namely the threshold η), but also for the error affecting each single not edited observation. In fact, it is easy to check that SF_i is below 2η for $i > \bar{k}$.

The algorithm so far described is easily extended to the multivariate case by defining a global score function GS_i as $\max_p SF_{i,p}$, where $SF_{i,p}$ is the local score function for the i th unit and the p th variable. This global score ensures that the previous properties still hold for each variable of interest.

4. The R Package SeleMix

In order to implement the selective editing method based on contamination model, R functions have been developed and included in a package. The package can be used also in cases where the multivariate contaminated variable Y contain missing values. In these cases robust predictions for missing values are provided. The software allows to include in the model also a set of “cleaned” variables X to be used as explanatory variables. This characteristic is useful, for instance, when auxiliary information (e.g., administrative or historical data) is available.

The core of the package is composed of three functions `ml.est`, `pred.y`, `sel.edit`. Further graphical tools are available.

The main output of the package is the identification of critical units corresponding to the most influential errors given a prefixed threshold of accuracy. In the following we describe in more detail the functions of the package *SeleMix*.

`ml.est`. This function estimates the parameters $\theta = (B, \Sigma, \pi, \lambda)$ on observed data, using the ECM algorithm. It returns also “anticipated” values (predictions) for the Y variables for all the observations.

The input of the `ml.est` function is the matrix of observed data and optionally the matrix of covariates X .

By default the algorithm starts with $\lambda = 3$ and $\pi = 0.05$ but it is possible to define different starting points. The user must specify if true data are assumed to follow normal or lognormal distribution. In the latter case zeros in data are replaced by a small value (10E-8) and a warning is returned.

The starting values of the regression coefficients B and the covariance matrix Σ are computed on the input data Y and X via OLS (i.e., as if they were error free). The EM algorithm consists of repeatedly applying the expectation and the maximization step until convergence or until some user specified maximum number of iterations is reached.

The function computes for each unit, the posterior probability τ that it belongs to the mixture component corresponding to contaminated data. This probability is used to define a flag of outliyness that is 1 if τ is greater than a specified threshold (by default equal to 0.5), and 0 otherwise.

The function returns the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) scores in order to evaluate the goodness of fit of the mixture model versus the standard normal model.

This information helps the user to assess the validity of the use of a mixture model.

The output of the `ml.est` function is provided as a list whose components are: the model parameters θ , the anticipated values, the BIC and the AIC scores, the outlier flags, and the posterior probabilities τ .

The function `ml.est` includes a call to the function `pred.y` that calculates the predictions for the variables Y .

`pred.y`. This function aims to estimate the true data distribution conditional on observed response variables and covariates. It needs as input the parameters $\theta = (B, \Sigma, \pi, \lambda)$ and a set of observed data. Note that missing values are not allowed for X variables. It returns, for each unit, a "prediction" for both observed and missing items of each Y variable, the outlier flag and the posterior probability τ .

`sel.edit`. This function prioritises observations according to the score function values and flags the units to be edited such that the expected residual error is below a prefixed level of accuracy.

It is worth noting that `sel.edit` can be used independently of the other SeleMix functions. In fact, the identification of influential units can be performed regardless of the particular model used for the prediction.

As input, the function receives: the matrix of observed data and the matrix of corresponding anticipated values, the reference total estimate of each Y variable, the sampling weights and the prefixed level of accuracy.

The reference total of Y is optional, if omitted it is computed as the weighted sum of the predicted values. The weights are assumed to be equal 1 if not differently specified, and the default threshold of the level of accuracy is 0.01.

Influential units are selected according to the values of a global score computed as follows. First a local score for a given variable is defined as the weighted absolute difference between observed and anticipated values standardised with respect to the reference total estimate. Then, the global score is obtained by computing the maximum of the local scores and the observations are ranked according to the descending values of the global score.

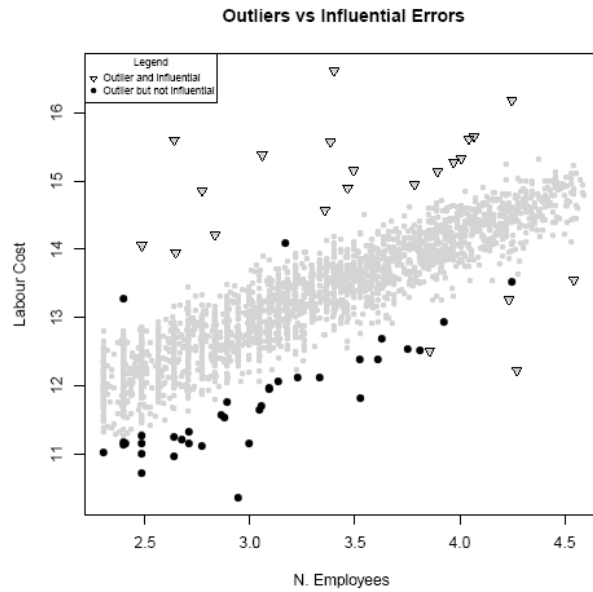
The last step is to find the first k units such that, for all variables, the (expected) total residual error remaining in the other $(n-k)$ units, is below the prefixed threshold.

The output of `sel.edit` is a matrix containing the flag of influential units, the rank according to the global score, the global and local scores, and the residual cumulative error for each variable.

The following two figures obtained by means of a graphical tool available in the SeleMix show outliers versus influential errors, and estimated versus true residual errors with respect to an experiment carried out on an Istat business survey. The details of the experiment are provided in Buglielli *et al.*, 2011.

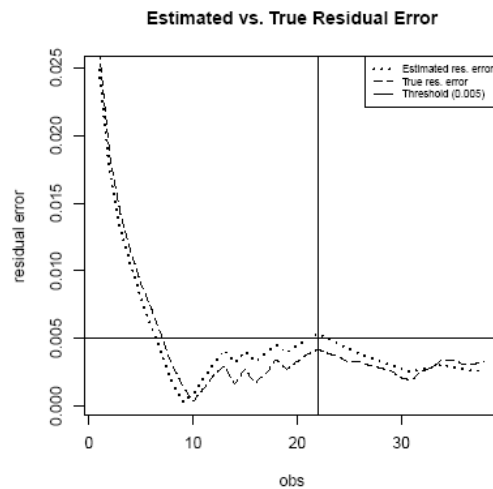
In Figure 1, the observations depicted with grey triangles are those classified as influential errors, while the black dots are observations classified as outliers but not influential. The selection is made with respect to a threshold η equal to 0.005. In this application, all the influential errors are outliers, but we notice also that there are outliers that are not influential errors. This is an important peculiarity of selective editing that allows to save resources for data revision. In fact, even if observations are classified as errors their impact on the estimates is negligible with respect to the chosen level of accuracy.

Figure 1. Outliers and influential errors according to a threshold η equal to 0.005. Logarithmic scale is used.



In Figure 2, the dashed line shows the true residual error, while the dotted line depicts the estimated residual error for the variable *Labour Cost* on the subset of the first 40 observations. All the units on the left of the vertical lines are the influential observations. We note that both true and estimated cumulative residual error curves are below the prefixed threshold for some units before the last observation considered as influential. This is due to the fact that the cumulative error is computed on the difference between observed and anticipated values, and the values can compensate each other. We remind that the stopping criterion ensures that the residual error is below a certain level of accuracy from the last selected unit.

Figure 2. Estimated (dotted line) and true (dashed line) residual error according to a threshold η equal to 0.005.



References

- Bellisai, D., Di Zio, M., Guarnera, U. and O. Luzi (2009), "A Selective Editing approach based on contamination models: An application to an Istat business survey", *UNECE Work Session on Statistical Data Editing*, Neuchatel, 5-7 October 2009.
- Buglielli, M.T., Di Zio, M. and U. Guarnera (2010), "Use of Contamination Models for Selective Editing", *Q2010, European Conference on Quality in Survey Statistics*, 4-6 May 2010, Helsinki.
- Buglielli, M.T., Di Zio, M., Guarnera, U. and F.R. Pogelli (2011), "Selective Editing of Business Survey Data Based on Contamination Models: An Experimental Application", *NTTS 2011 New Techniques and Technologies for Statistics*, Bruxelles, 22-24 February 2011.
- de Waal, T., Pannekoek, J. and S. Scholtus (2011), *Handbook of Statistical Data Editing and Imputation*, New Jersey, Wiley.
- Di Zio, M., Guarnera, U. and O. Luzi (2008), "Contamination Models for the Detection of Outliers and Influential Errors in Continuous Multivariate Data", *UNECE Work Session on Statistical Data Editing*, Vienna.
- Ghosh-Dastidar, B. and J.L. Schafer (2006), "Outlier Detection and Editing Procedures for Continuous Multivariate Data", *Journal of Official Statistics*, Vol. 22, No. 3, 2006, pp. 487-506.
- Hedlin, D. (2003), "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics", *Journal of Official Statistics*, Vol. 19, No. 2, pp. 177-199.
- Hedlin, D. (2008), "Local and global score functions in selective editing", *UNECE Work Session on Statistical Data Editing*, Vienna.
- Jäder, A. and A. Norberg (2005), "A Selective Editing Method considering both suspicion and potential impact, developed and applied to the Swedish Foreign Trade Statistics", *UNECE Work Session on Statistical Data Editing*, Ottawa.
- Latouche, M. and J.M. Berthelot (1992), "Use of a score function to prioritize and limit recontacts in editing business surveys", *Journal of Official Statistics*, 8, n. 3, pp. 389-400.
- Lawrence, D. and C. McDavitt (1994), "Significance Editing in the Australian Survey of Average Weekly Earnings", *Journal of Official Statistics*, Vol. 10, No. 4, pp. 437-447.
- Lawrence, D. and R. McKenzie, (2000), "The General Application of Significance Editing", *Journal of Official Statistics*, 16, n. 3, pp. 243-253.

Selective editing methods and tools: An Australian Bureau of Statistics perspective

Eden Brinkley, Keith Farwell and Frank Yu¹

Abstract

Early last decade, the Australian Bureau of Statistics initiated a major re-engineering program aimed at transforming processes, methodologies and technologies used in the production of economic statistics. One of our processes for early attention was data editing, and one of the new methodologies introduced was selective editing for micro data using a home-grown tool called the Significance Editing Engine. The paper outlines the methodology underpinning our approach to micro selective editing and the Significance Editing Engine, and concludes with a summary of the key messages from our journey so far and how our thinking has evolved.

Key Words: Selective editing; Significance editing; Significance Editing Engine; Re-engineer; Transform; Messages.

1. Introduction

Early last decade, the Australian Bureau of Statistics (ABS) initiated a major re-engineering program known as the Business Statistics Innovation Program. This program aimed to transform processes, methodologies and technologies used in the production of ABS Economic Statistics. Significant work was undertaken to review business processes within the context of a consistent high level ‘end to end’ framework, and to replace the disparate processes with a smaller number of more common and better practice approaches. New methodologies were also introduced to help reduce costs and provider load, and to improve data quality.

One of our processes for early attention was data editing, and one of the new methodologies introduced was significance editing for micro data using a home-grown tool called the Significance Editing Engine (SigEE). The paper outlines the methodology underpinning the ABS approach to micro significance editing and SigEE. It also details the functionality offered by SigEE, and how it deals with a variety of micro editing requirements. A new macro significance editing tool, which uses significance scores to detect anomalous estimates, will also be highlighted. The paper concludes with a summary of the key messages from our journey so far, and how ABS thinking has evolved over the last 10 years.

2. What is significance editing?

Significance editing is a form of selective editing and is based on the principle that if we can predict the impact of our editing actions on the results we are trying to achieve, we will be in the best position regarding what to edit and how much to edit (Farwell and Raine, 2000). This is done by identifying and prioritising significant differences between what is observed and what is expected in the data. The significance of the difference is assessed in terms of the impact of the difference on expected outputs. Scores are created based on measures of the predicted impact of editing. The scores can be used to prioritise anomalous data items, provider records, or estimates. The data is ranked by score size and cut-offs are applied to select the anomalous data. The choice of cut-off value involves a cost/benefit analysis involving a trade-off between the cost and benefit of editing. Once important anomalies in the data have been identified, processes need to be put into place to resolve the nature of the anomaly and implement a method of treatment.

¹Eden Brinkley, Australian Bureau of Statistics, Keith Farwell, Australian Bureau of Statistics, Frank Yu, Australian Bureau of Statistics, Locked Bag 10, Belconnen, ACT, Australia, 2616.

The views expressed in this paper are those of the authors and do not necessarily reflect those of the Australian Bureau of Statistics.

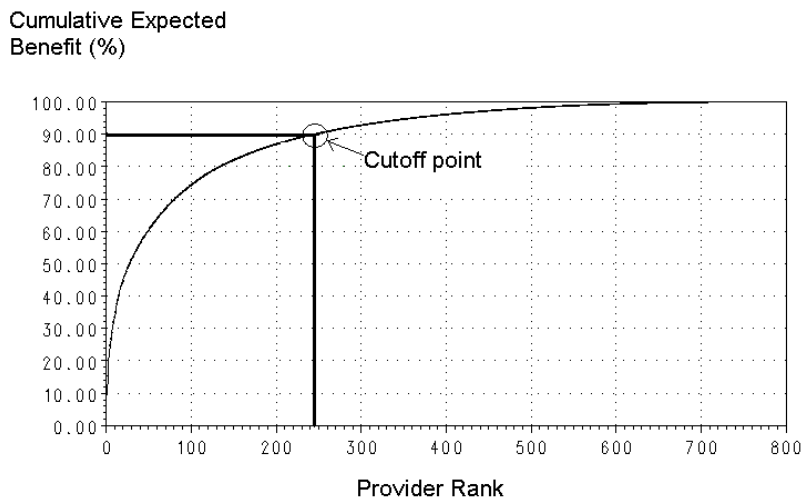
Significance editing was first developed for editing micro data, but the general framework has now been extended to include the selective editing of macro data. The ABS is currently building a complementary tool to SigEE for macro significance editing (Farwell, 2009). It will be used to detect anomalous estimates.

3. Significance editing concepts

For micro significance editing, the idea is to select a set of key variables (*i.e.*, key items) which feed into a set of key estimates at a particular level (referred to as the ‘target’ or ‘significance level’). Each key item value within a unit record is given an ‘item score.’ Item scores are derived using expectations of key items and target estimates (referred to as ‘expected values’ and ‘expected estimates’ respectively). The item response value, expected value, and expected estimate (and other attributes such as design weights) are used to predict the impact of editing. The standardised impact of editing is also referred to as ‘benefit’ and the score is an estimate of ‘expected benefit’ (‘score’ and ‘expected benefit’ are used interchangeably in this paper).

The significance editing framework allows for local scores such as item scores to be combined using a metric to create global scores. For example, it is often desirable to score data providers rather than data items. Each provider record will have several item scores (one for each key item) and these can be combined into a global score for each provider (a ‘provider score’). The local and global score sizes can be used to generate local and global ranks. Provider scores are used to rank data providers rather than data items. For example, if a questionnaire is used to collect the data, the provider score can be used to rank completed questionnaires. Cut-offs are then used to select anomalous records. An example of a basic cut-off is where data is considered anomalous if the associated score is greater than the cut-off value. Also, plots of cumulative expected benefit verses rank can be used to select cut-offs ‘as we go.’ The plots are referred to as ‘cost/benefit curves.’ They are a form of Lorenz curve and can be ordered by a GINI coefficient. They display the trade-off between the cost of editing and the benefit of editing. In Figure 3-1 below, about 90% of the expected benefit is due to about 250 of the 800 providers submitted to SigEE.

Figure 3-1
Example of a Cost/benefit curve



4. Significance scores

4.1 Generic significance score

A generic significance score (Farwell, 2009), which can be used to create micro or macro significance scores, has the following form:

$$\text{Generic score} = 100 * \left| \frac{\text{Predicted impact of editing}}{\text{Standardising value}} \right| \quad (1)$$

$$\text{Impact of editing} = \text{Adjusted expected target estimate} - \text{Expected target estimate} \quad (2)$$

The generic score (1) can be considered to ‘target’ a particular estimate. The measure of impact of editing (2) is defined in terms of differences between the expected and observed data. For micro editing, there is also a need to take the probability that the record is erroneous into account when measuring the predicted impact (not needed for macro editing). Refer to section 4.2 for an example and brief discussion.

The expected target estimate for micro editing is, technically, the sum of weighted expected values. The adjusted expected target estimate is a recalculation of the expected target estimate with a reported value replacing its expected value (all other expected values remain unchanged). It is calculated on a value by value basis to create an adjusted expected target estimate (and a measure of impact of editing) for each reported value. The score creation process is the same for macro editing except we replace item values with estimates (from a set of study estimates). For estimates of total, the expected target estimate is a sum of the expected study estimates, and estimates in the study set aggregate to estimates at the target level. Refer to Farwell (2009) for more details.

Within the significance editing framework, the standardising value in (1) can be either an expected target estimate or a multiple of the expected standard error for the target estimate (options within SigEE), and it enables scores for different variables to be compared and combined.

While the generic score caters for macro scores, SigEE does not include functionality for macro significance editing. The remainder of the paper will therefore concentrate on micro significance editing.

4.2 Basic SigEE item scores

Using (1) and (2), with a Horvitz-Thompson estimate of total as the target estimate, an item score for provider i , item j , is:

$$s_{ij} = 100q_j \left| \frac{Y_{adj,ij}^* - Y_j^*}{Y_j^*} \right| = 100q_j \left| \frac{w_i^*(y_{ij} - y_{ij}^*)}{Y_j^*} \right| \quad (3)$$

where w_i^* is the expected estimation weight for provider i ; y_{ij} is the reported item value for item j , provider i ; y_{ij}^* is the expected item value for item j , provider i ; Y_j^* is the expected target estimate for item j ; q_j is the probability of misreporting for item j (*i.e.*, the probability that y_{ij} is erroneous); and $Y_{adj,ij}^* = Y_j^* - w_i^* y_{ij}^* + w_i^* y_{ij}$.

In a more technical sense, the impact of micro editing a data item can be defined as the absolute reduction in bias in the target estimate due to misreporting (Farwell, Poole, and Carlton, 2002). The impact must be predicted since it can only be known after editing. For (3), this resolves to estimating the expectation of $\left| w_i^*(y_{ij} - y_{ij}^*) \right|$ which we currently approximated with $q_j \left| w_i^*(y_{ij} - y_{ij}^*) \right|$. This can be justified if q_j and $\left| w_i^*(y_{ij} - y_{ij}^*) \right|$ are independent. Unfortunately, this is not strictly true and further research is needed in this area. Within ABS, it is often the case that little is known about q_j and the tendency is to set $q_j = 1$.

So to calculate this score, we require a reported item value, an expected item value, an expected target estimate, an expected estimation weight, and a value for q_j . If we have these terms prior to receiving responses, a score can be calculated as soon as a response is obtained. This score can be compared to a pre-specified cut-off value and a decision can be made immediately on whether to edit the value.

While the expression $w_i^*(y_{ij} - y_{ij}^*)$ can be used to measure the impact of editing for an estimate of total, the generic definition (1) is also useful for deriving more complex scores such as those for estimates of rate and standard errors, and for macro editing. For example, we can derive a micro editing item score for the ratio of two Horvitz-Thompson estimates of total ($R_{jk} = Y_j / Y_{k \neq j}$, for items j and k) as follows. We use the ratio of the expected numerator and denominator target estimates (Y_j^* / Y_k^*) as the expected target estimate (R_{jk}^*). The adjusted expected target estimate becomes the ratio of the adjusted expected numerator and denominator target estimates, resulting in $R_{adj,ijk}^* = Y_{adj,ij}^* / Y_{adj,ik}^*$, where $Y_{adj,ij}^* = Y_j^* - w_i^* y_{ij}^* + w_i^* y_{ij}$ and $Y_{adj,ik}^* = Y_k^* - w_i^* y_{ik}^* + w_i^* y_{ik}$.

5. ABS editing scenarios covered by SigEE

The ABS first implemented significance editing in the Australian Survey of Average Weekly Earnings (AWE) in the early 1990's (Lawrence and McDavitt, 1994). It was based on item scores developed from (1) for estimates of rates. AWE was a collection with a small number of key item scores, a good supply of pre-edited and post-edited data, historical data, and stable estimates. The pre-edited and post-edited values allowed an analysis to determine effective cut-offs. Expected values and estimates were easy to obtain from the available historical values and estimates, and the target estimates were easy to predict.

Various applications of the significance editing framework were implemented for other collections after the initial AWE application, particularly in Agricultural collections. As part of ABS's drive towards more generic processes and a common business processing infrastructure, it was decided to test the versatility of the framework by applying it to several surveys, each presenting a different set of editing challenges. The tests were conducted during 2002 and 2003 using a suite of SAS programs which 'evolved' into SigEE. The tests included collections that did not have pre-edited and post-edited data; new and one-off collections with no historical data; collections with very erratic response values; collections with estimates that were difficult to predict; and collections with many key outputs and many key data items. Refer to Farwell (2004) and Farwell (2005) for further details. These led to the development of interactive cut-offs, cost/benefit curves, alternative item scores, and provider score options.

6. SigEE functionality

SigEE contains functionality that allows significance editing to be performed with or without expected values; with or without expected estimates; and with or without pre-specified cut-offs. A feature of SigEE functionality is the use of four types of item scores. The four item score approaches are referred to as *Paths* within SigEE, and the item scores are called *Path A*, *Path B*, *Path C* and *Path D item scores*.

6.1 Item score choices: Paths A, B, C, and D

The generic score (1) is used to create *Path A* item scores. These scores require expected values, expected estimates, and expected weights (SigEE uses $q_j = 1$ as the default for the probability of misreporting, equivalent to assuming each reported value is equally erroneous).

Path B scores were developed to deal with the situation where expected estimates are not available and only expected values and expected weights are available. They are, essentially, standardised *Path A* scores where we divide each *Path A* score by the sum of the *Path A* scores and express the result as a percentage. This results in there being no need for expected estimates. A *Path B* score for provider i , item j , has the form $s_{ij}^* = 100 * \frac{s_{ij}}{\sum_i s_{ij}}$ where the

sum is over all provider records requiring editing for item j . For example, the Path B score for an estimate of total is:

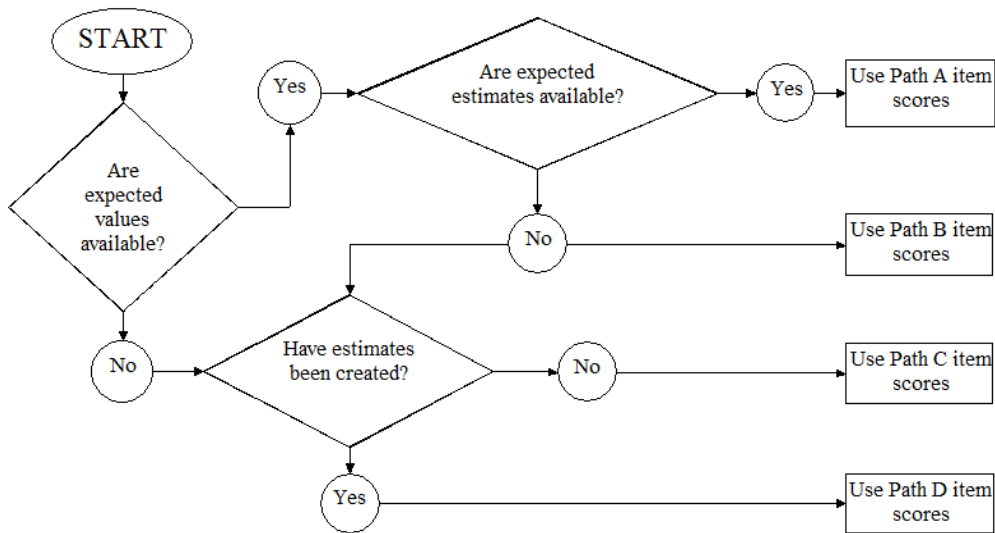
$$s_{ij}^* = 100 \frac{|w_i^*(y_{ij} - y_{ij}^*)|}{\sum_i |w_i^*(y_{ij} - y_{ij}^*)|}$$

Path B scores can be affected by extreme values for s_{ij} and SigEE has extreme score adjustment functionality to manage this. The data items responsible for the extreme scores are removed from the scoring process and placed into the critical stream for editing and new Path B scores are then calculated.

Path C and D item scores are based on combined standardised contributions to the level, movement, and standard error of a target estimate. The *initial* scores for level, movement, and standard error are combined using a weighted Euclidean metric to create an item score. These item scores were created to overcome the situation where no expected values and no expected estimates are available. Path C items scores estimate the contributions to level, movement, and standard error (since there will be insufficient good quality responses to support creation of estimates). Path D scores require the by-product information from the creation of estimates to calculate the contributions. Therefore, Path D scores can only be created once estimates have been produced.

Diagram 6.1-1 below displays how the SigEE path is selected.

Diagram 6.1-1
Choosing the appropriate Path



6.2 Contributor scores

The initial scores used for Path D scores can be used to rank contributors to estimates, movements in estimates, and standard errors of estimates; the estimates of unit contribution are far more accurate than those used for creating the Path C scores. This functionality can be used to generate ordered lists of unit contributors which are often used during macro editing. In this sense, macro editing often involves a micro editing component. In order to separate macro editing tasks from micro editing tasks, we call the initial scores for Path D *contributor scores* when used during macro editing.

6.3 Expected value generation

SigEE has expected value generation capability. SigEE can obtain historical values for use as expected values if they are available. Optional adjustment factors can be applied to the historical values. SigEE can also generate expected

values using weighted means or ratio imputes using a previous or current survey data file. SigEE also allows the user to specify expected values in a file.

6.4 Provider score choices

SigEE offers the following three choices for provider scores using weighted item scores:

- (i) the maximum of the weighted item scores;
- (ii) the Euclidean norm of weighted item scores; and
- (iii) the root mean square (RMS) of weighted item scores.

For example, the weighted RMS provider score is defined as $s_i = \sqrt{\frac{\sum a_j s_{ij}^2}{n}}$ where s_i is the provider score for provider i ; s_{ij} is the item score for item j , provider i ; a_j is a user-defined item weight for item j ; and n is the number of non-missing key items.

The weighted Euclidean score is the RMS score with $n=1$. The item weight (a_j) allows users to make one item more important than another when creating a provider score. The item weight can be used to account for the probability of misreporting when q_j is set to 1 in SigEE. The Euclidean and RMS scores (the SigEE default) work well when there are large numbers of key items or when interactive cut-offs are required. The score generates meaningful cost/benefit curves for interactive cut-off selection.

6.5 Multivariate surprise outlier detection

Contributor scores can also be used for surprise outlier detection. Contributor scores for level can be combined using the RMS norm for the key items. The combined score can be used to select multivariate surprise outliers. These are outliers which have a strong impact on the set of key items.

6.6 Cut-off options

SigEE has the option to use pre-specified cut-offs (created prior to obtaining responses) or to interactively choose cut-offs. It is preferable to use pre-specified cut-offs if possible. However, pre-specified cut-offs can become inefficient over time (leading to more data items or providers being selected for editing) and so they need to be maintained. Cut-offs can be determined interactively by reference to lists and cost/benefit curves.

SigEE provides four cut-off methods which can be used for editing based on item scores or for editing based on provider scores. These are referred to as: a *score* cut-off; a *top-k* cut-off; a *cumulative score (%)* cut-off; and an *iterative cumulative score (%)* cut-off. The score cut-off is a value where scores greater than the cut-off are selected for editing. The top-k cut-off selects records with ranks greater than or equal to k. To apply the cumulative score (%) cut-off, standardised scores are cumulated in rank order (starting at rank 1). All records needed to achieve the cumulative score percentage total are selected. The iterative cumulative score (%) cut-off is a more complicated application of the cumulative score (%) cut-off. The method iteratively selects records over several edit runs in such a way as to achieve the pre-specified cumulative score percentage cut-off. SigEE keeps track of the benefit totals and adjusts the cut-off target for each succeeding edit run. SigEE also has extreme score functionality for use with cut-offs based on standardised scores (*i.e.*, cut-offs involving cumulative score percentages). Records with extreme scores are removed and the standardised scores are recalculated. The cut-off is applied to the revised scores. Records with extreme scores are placed in the critical stream for editing.

6.7 Editing based on provider and item cut-offs

Most economic collections select entire provider records for editing because the data tend to have a balance sheet structure. An error in one data item can be related to possible errors elsewhere on the provider record. Accordingly,

the primary cut-offs used are provider cut-offs, as the aim is to detect anomalous provider records. Editors can use item scores and ranks, and other information, to resolve and treat errors in the entire record.

There are some economic collections which collect data that do not have a balance sheet structure (*e.g.*, certain activity or commodity collections). For these collections, errors in one part of the provider record tend not to be related to errors elsewhere in the record, and so individual data items, rather than provider records, are selected. For these types of provider records, editors should not spend unnecessary time looking through the entire record to edit particular data items. Accordingly, the primary cut-offs used are item cut-offs, as the aim is to detect anomalous item values. There is still a need to prioritise the provider records to manage workflows, but the main focus is on specific data items within the provider records.

The main factors which determine the set of functionality used within SigEE for each collection are: primary cut-off type (*i.e.*, item or provider cut-offs); method for creating cut-offs (*i.e.*, pre-specified or interactive cut-offs); and what kind of item score is used (*i.e.*, Path A, B, C, or D item score).

7. Ten years on ...

7.1 The changing environment

Significance editing and SigEE were introduced 10 years ago. However, like all statistical agencies, the ABS has seen substantial change over this timeframe. Client demand for data that is 'better, faster, cheaper' has increased, and there is now a greater focus on the use of administrative data to help deliver our statistical solutions. The pace of change for technology has also increased, with many new tools and systems in place to deal with the added demands on our statistical infrastructure. Declining budgets have left fewer staff to do the work, and this is compounded by increased staff turnover as the workforce becomes more mobile. This has increased our need for better knowledge management, and to build staff capability much more quickly.

There have been some good successes with significance editing and with SigEE. The methodology is now routinely used by quite a few collections within the ABS and there have also been significant savings made. While our successes are pleasing, the ABS has not been universally successful in introducing significance editing and /or SigEE to economic collections. This has led us to reflect on why this is the case, and what we should do to ensure greater adoption of the methodology going forward.

7.2 Key messages from our journey so far

Perhaps the first message to emerge is for business areas to recognize implementation of selective editing represents a significant change process. It involves the development and adoption of new methods, processes and systems, as well as new roles and skills for staff. All too often the change is considered a relatively minor change to collection operations, and the complexities involved are underestimated. Areas that set aside suitable resources to make the change happen invariably have greater success, as there are many complex changes to think through. They also tend to invest in better procedures and documentation, and this helps in passing knowledge from one generation to the next.

Senior managers need to buy into the change process, and stay in for the duration. Areas where senior managers opt out of the change process along the way, or delegate to lower levels as they become distracted with other priorities, are far less successful in adopting these new methods. Senior managers should also be made accountable for making the change happen.

One of the best ways to ensure success is to set a big savings target, or possibly a significant reduction in time to release (or both). This really helps to 'focus the mind'. Making senior managers responsible for delivering the savings targets also helps to keep them engaged.

Business areas need to recognize that changing editing processes is a long term commitment, and it should be treated as such. Key stakeholders need to be engaged, not just early on, but on a continuing basis. For instance,

methodological and systems support should be available during the development phase, the implementation phase, and on an ongoing basis. If there is a rapid change in staff, skills in an area can be depleted severely, leaving little knowledge to support new processes. Good methodological and systems support will help get these business areas back on track. Good documentation and training materials will also help, and ideally these materials should be available online.

Related to this point, effort should be made to monitor, evaluate and improve selective editing processes over time. This will ensure they remain optimal for the given collections.

For areas promoting or supporting change within an organization, we suggest focusing on business areas where senior managers and staff want to change, at least as your first priority. You will be far more successful with these areas than with areas where there is little interest, or where people are distracted with other priorities, even if the savings to be gained appear significant. It also helps to leverage your success stories as you build them. This will gain interest and support for change, as genuine savings or other improvements cannot be dismissed easily.

Lastly, SigEE was built to handle a great variety of editing situations. With our changed work environment over the last 10 years, some people now say it takes too much time to learn and set up, and that the tool is too difficult to use in practice. In developing new tools we should focus on delivering functionality that gives the greatest gains, and perhaps some of the ‘nice to have’ functionality should be foregone. Investing in making the tools easy to use is also important for getting business area buy in (*e.g.*, good interfaces, well integrated with the processing environment, *etc.*).

References

- Farwell, K. (2004), “The general application of significance editing to economic collections”, *Methodology Advisory Committee Papers*, cat. No. 1352.0.55.066, Australian Bureau of Statistics, Canberra.
- Farwell, K. (2005), “Significance Editing for a Variety of Survey Situations”, *Paper presented at the 55th session of the International Statistical Institute*, Sydney, 5-12 April.
- Farwell, K. (2009), “The use of scores to detect and prioritise anomalous estimates”, *Methodology Advisory Committee Papers*, cat. No. 1352.0.55.104, Australian Bureau of Statistics, Canberra.
- Farwell, K., Poole, R. and S. Carlton (2002), “A technical framework for input significance editing”, *Conference paper for DataClean2002*, Jyvaskyla, Finland.
- Farwell, K. and M. Raine (2000), “Some current approaches to editing in the ABS”, *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, USA.
- Lawrence, D. and C. McDavitt (1994), “Significance editing in the Australian Survey of Average Weekly Earnings”, *Journal of Official Statistics*, 10:4, pp. 437-447.

SESSION 4B
CONFIDENTIALITY

G-Confid: Statistics Canada’s confidentiality software

Caroline Rondeau and Jean-Marc Fillion¹

Abstract

Under the *Statistics Act*, Statistics Canada must protect respondents’ confidential data. Cell suppression is a technique used to protect tabular data. The automated confidentiality software G-CONFID (formerly called CONFID2), developed at Statistics Canada, is used to implement this technique. This software is user-friendly, uses the same structure as Statistics Canada’s other generalized systems and can incorporate new approaches. It can also be used to deal with potentially voluminous multi-dimensional tables. G-CONFID is also part of Statistics Canada’s methods and systems consolidation project. The main objective of G-CONFID is to provide the appropriate level of protection for confidential cells while minimizing the loss of information. To achieve this objective, linear programming is performed to optimize residual suppression. This presentation covers the functionality and characteristics of G CONFID. The emphasis will be on the use of user-specified cost variables.

Key Words: Sensitivity; Confidentiality; Residual suppression; Cost variable.

1. Introduction

1.1 Introduction to G-Confid

Economic data are often presented in the form of tables with different levels of detail. This way of presenting such data can lead to a problem with disclosure in one of the following situations: (i) where there are very few respondents in a cell, or (ii) where there are only one or two respondents who contribute the most in a cell, which is called a dominance situation. Cell suppression is a technique employed to ensure that tabular data are protected. The automated confidentiality software G-Confid is used to apply this technique.

G-Confid is a generalized system that performs cell suppression for economic-type data. Not only is it a standardized tool used as a model for other generalized systems at Statistics Canada, it is also very flexible and user-friendly. Since the input/output files are in SAS, it offers the possibility of modifying information and defining parameters and options according to our needs. This system is used by several types of surveys at Statistics Canada.

1.2 Organization of this paper

Section 2 explains the different components of G-Confid. Section 3 takes a closer look at the use of user-specified cost variables, using an example. The advantages of G-Confid are described in Section 4 by way of a conclusion.

2. Components of G-Confid

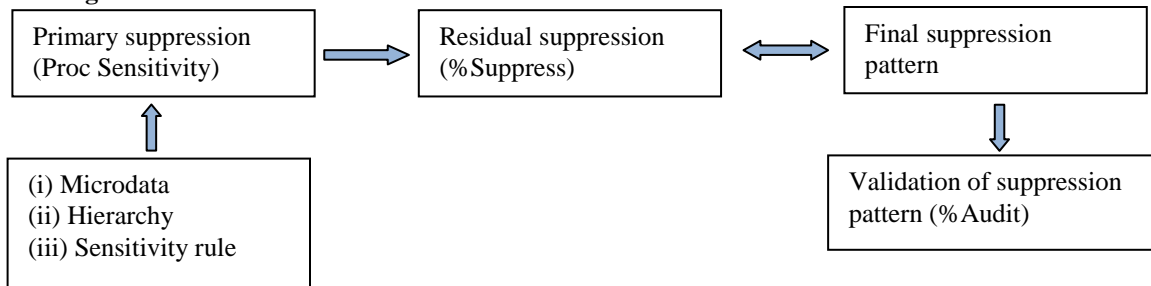
2.1 General description of G-Confid

G-Confid is a suite of three SAS components. The first component (Proc Sensitivity) performs primary suppression. The second component (%Suppress macro) performs residual suppression, meaning that it identifies the additional cells to be suppressed using linear program optimization methods. The third component (%AUDIT macro) verifies

¹Caroline Rondeau, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, K1A 0T6, caroline.rondeau@statcan.gc.ca; Jean-Marc Fillion, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, K1A 0T6, jean-marc.fillion@statcan.gc.ca.

that the residual suppression adequately protects confidential cells. Figure 2.1-1 shows how G-Confid works; its components will be explained in greater detail in sections 2.1.1 to 2.1.3.

Figure 2.1-1
Functioning of G-Confid



2.1.1 Primary suppression

In primary suppression (the Proc Sensitivity component), the total value of each cell and its sensitivity, based on microdata, are calculated. Sensitivity is a value that determines whether or not the cell is considered confidential. Three elements are essential for primary suppression: (i) the microdata file, (ii) the hierarchy, and (iii) the sensitivity rule. The microdata file must include the following information: an ID number (generally the business number), the table dimension variables (*e.g.*, industry and region) and the variable of interest (*e.g.*, income). The hierarchy is used to specify the structure of the data tables to be protected in G-Confid and is related to the dimension variables. The hierarchy may be defined at several levels; for example, one may want to deal with different regions of Canada, such as the East, the West and the Maritimes, and each of these regions is divided into its respective provinces. The sensitivity rule serves to identify confidential cells. Several rules exist, and they are particular forms of the general linear sensitivity rule described in formula (1).

$$S = \sum_{i=1}^r \alpha_i x_i, \text{ où } \alpha_1 \geq \alpha_2 \geq \dots \geq -1 \quad (1)$$

where

S represents the sensitivity of the cell,

α_i are fixed coefficients,

r is the number of contributors to the cell,

x_i represents the values of the different contributors to the cell in descending order ($x_i \geq 0$).

If S is positive, then the cell is confidential. There are several sensitivity rules defined in G-Confid: the nk rule, the pq rule and other Statistics Canada rules.

The following is an example of the Sensitivity procedure.

PROC SENSITIVITY

```

DATA=data      OUTCONSTRAINT=outconstraint  OUTCELL=outcell
HIERARCHY= "Total_naics N912 N913 N92 N931 N932 N933 N934 N941 N942;
           Canada East Quebec Ontario West:
           East Newfoundland and Labrador Maritimes West Prairies British Columbia;"
SRULE="pq 0.15"; ID EnterpriseID;
VAR income; DIMENSION Industry Region;
  
```

As we can see, the general formulation is similar to all SAS procedures. The SAS microdata file is defined in the DATA parameter, the SAS output files in the OUTCONSTRAINT and OUTCELL parameters. The OUTCONSTRAINT file identifies the relationships between the cells using the coefficient representing linear equations. The OUTCELL file includes the dimension variables, the total value of each cell, its sensitivity and its status. The hierarchy, the sensitivity rule (pq 0.15), the identity variable (EnterpriseID), the variable of interest (income) and the dimensions (Industry and Region) must be specified in the procedure.

2.1.2 Residual suppression

The main objective of residual suppression (the %Suppress component) is to provide confidential cells with the appropriate level of protection while minimizing the loss of information. For each confidential cell, complementary cells are identified, by solving the linear programming problem while minimizing suppression costs subject to the condition that each cell be well protected. Confidential cells combined with complementary cells form the final suppression pattern.

There are two stages (called phases) for residual suppression. Residual suppression is performed sequentially. Phase 2 does a little housekeeping, in that it frees up cells from the suppression carried out in Phase 1. Each of these phases uses a cost function. The available costs are: “constant,” “size,” “digit” and “information.”

Below is an example of the Suppress macro call.

```
%SUPPRESS(INCELL=outcell, CONSTRAINT=outconstraint, CFUNCTION1=size,
CFUNCTION2=information, CVAR1 = , CVAR2 = , OUTCELL=outcell_sprs);
```

The output files (OUTCELL and OUTCONSTRAINT) in the Sensitivity procedure are used as an input file for this macro (INCELL and CONSTRAINT). The cost functions for each phase are defined in CFUNCTION1 and CFUNCTION2 and the output file is in OUTCELL. By default, the variable of interest (that is, the cell total) is used as the cost variable for each of the phases (CVAR1 and CVAR2). However, it is possible to define new ones using the parameters CVAR1 and/or CVAR2. This will be discussed further in Section 3.

Table 2.1.2-1 presents an example in tabular form of the output of the Suppress macro where the confidential cells are shown in red and the residual cells in blue. Please note that this is not the G-Confid output; it is a representation in tabular form.

Table 2.1.2-1
Total income by industry and region

	Canada	East	Newfoundla nd and Labrador	Maritimes	Quebec	Ontario	West	Prairies	British Columbia
N912	2016	74	0	74	677	342	923	838	85
N913	22115	3	0	3	20197	382	1533	692	841
N92	3875	355	3	352	245	549	2726	2071	655
N931	3014	88	0	88	1164	637	1125	791	334
N932	3435	35	0	35	2750	548	102	98	4
N933	3947	209	0	209	1393	1266	1079	787	292
N934	231	59	0	59	86	32	54	50	4
N941	7019	113	0	113	784	1221	4901	4695	206
N942	66	0	0	0	0	13	53	52	1
Total	45718	936	3	933	27296	4990	12496	10074	2422

If only the primary suppression were carried out (red cells), there would be no final suppression pattern, since it would still be possible to derive certain suppressed values. For example, in the column representing Canada, only one industry is confidential (N932), and it is therefore easy to find its value. Hence, residual suppression is needed to obtain a final suppression pattern.

2.1.3 Suppression validation

The main objective of suppression validation (the %Audit component) is to verify the quality of the suppression pattern, which was (i) created in G-Confid and modified by the user or (ii) created outside G-Confid and provided by the user.

3. Cost variable

3.1 Advantage

As noted above, the Suppress macro enables users to define their own suppression costs. The use of the cost variable makes it possible to (i) maintain consistency among suppression patterns for the same survey (*e.g.*, old survey versus new), (ii) suppress cells with a high coefficient of variation (CV) first, and (iii) favour publication of areas important to users.

To tell G-Confid how to identify the cells to be suppressed first, this need only be specified in the cost variables (CVAR1 and/or CVAR2) in the Suppress macro call. The specified variable will be used instead of the variable of interest (cell total) in calculating the suppression cost. The lower a cell's cost, the greater the likelihood of its being suppressed. There is also the option of changing the cell's status, that is, imposing a suppression pattern. But we will focus on the use of cost variables, by way of an example.

3.2 Example

Data from the Survey of Employment, Payrolls and Hours (SEPH) will be used to show how cost variables are used. That survey encounters the following problem: cells for the territories (small population in northern Canada) are selected as complementary cells because their employment numbers are low (cost variable: total employment), but these cells are of some importance for the territories in question. Therefore, the SEPH would like to publish more cells for the territories by focusing on their contribution to employment (proportion of employment in the territory) instead of the total employment number.

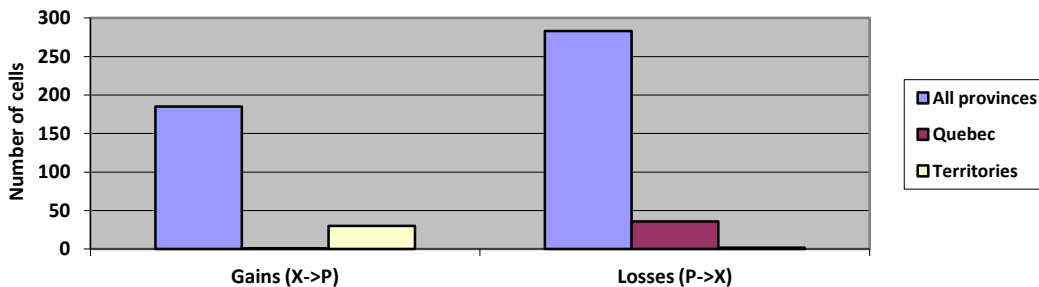
The cost variable for Phase 1 is the proportion of employment for a given industry in its province and is defined according to formula (2):

$$prop_{ij} = total_{ij} / \sum_{i \in j} total_i, \quad \text{where } i = \text{industry and } j = \text{province} \quad (2)$$

This cost variable combined with the 'size' cost function is used to suppress cells with the lowest proportions of employment. For Phase 2, the cost variable is total employment combined with the 'size' cost function, which will have the effect of freeing up the cells with the most employment.

A simulation was carried out using 12 months of data. If a cell is suppressed for a given month, it is considered suppressed for the final pattern. On average, we have 5,300 cells of which 2,000 are suppressed. The purpose of the simulation is to compare gains and losses between the basic method (the cost variable is total employment for both phases) and the new approach described above, namely using the proportion of employment. Gains are the number of cells that went from suppressed status (X) with the basic method to published status (P) with the new approach. Figure 3.2-1 shows the gains (X-P) and losses (P-X) with the new approach for the territories, Quebec and all provinces combined.

Figure 3.2-1
Number of cells showing losses/gains with respect to publication



As may be seen, there is indeed less suppression with the new approach for the territories. However, this gain for the territories entails an increase in losses for the provinces, since we must look elsewhere for residual cells. This leads to a higher suppression rate for large provinces such as Quebec. Note that for this simulation, there are approximately 425 cells and 330 cells respectively for Quebec and the Territories.

Despite the gain for the territories, this new approach was not implemented because of the increase in suppression in the large provinces (e.g., Quebec). A better function, taking account of both employment and the proportion of employment, must be found. One solution proposed would be to use the idea of a proportion, but not only at the provincial level but also for Canada as a whole. We could use a certain percentage (p) of the provincial proportion and 100-p for the proportion for Canada as a whole, as defined with Formula (3).

$$prop_{ij} = \frac{p}{100} \frac{total_{ij}}{\sum_{i \in j} total_i} + \frac{(100-p)}{100} \frac{total_{ij}}{\sum_{ij} total_{ij}}, \quad \text{where } i = \text{industry and } j = \text{province} \quad (3)$$

A simulation was carried out with p = 25, 50, 75 and 100. With p= 100, we are back to Formula (2). Figures 3.2-2 and 3.2-3 show losses and gains respectively between the basic approach and the proposed approach according to different values of p.

Figure 3.2-2

Number of cells with losses

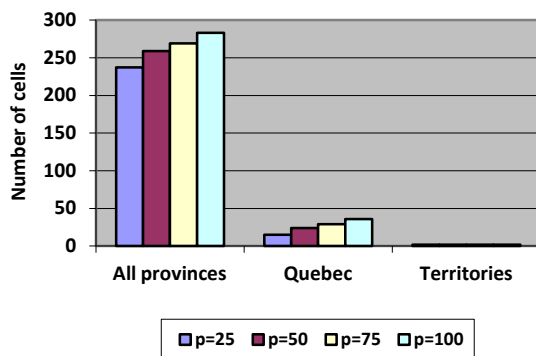
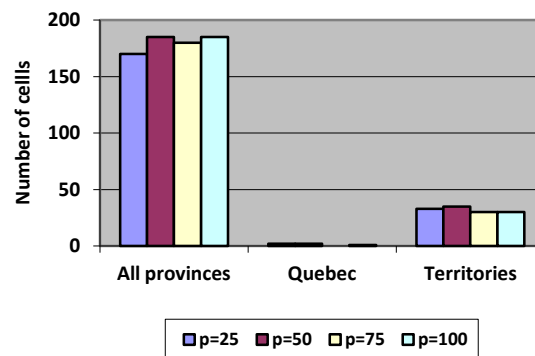


Figure 3.2 -3

Number of cells with gains



For the territories, the amount of loss is unaffected by the value of p. For Quebec, p = 25 or 50 represent the least loss without hindering the gain. For Canada as a whole, p = 25 represents the least loss but slightly less gain than, say, p=50. The proposed solution seems to be a good compromise. There are fewer losses for the large provinces and the gain is still satisfactory for the territories. Note that the suppressed cells could, according to other analyses, be negligible. These analyses should include losses and gains in terms of percentage of employment.

Other simulations will have to be made. As we can see, G-Confid is very flexible when it comes to using cost variables.

4. Conclusion

In addition to being a standardized and very user-friendly tool, G-Confid has several advantages. It can be incorporated into a larger SAS program. It is capable of resolving very large tables, it has good performance for residual suppression (SAS OPTMODEL solver) and it handles multiple decomposition (for example, Canada can be split into provinces, economic regions and census metropolitan areas). Moreover, since G-Confid is developed in-house, new approaches can be added in future years. Finally, G-Confid can be used with SAS Enterprise Guide, meaning that the G-Confid functions can appear in the form of customized tasks. These will generate the SAS code needed for G-Confid via a graphic interface.

Acknowledgements

I would like to thank the people in charge of SEPH, especially Yves Morin and Shou Xiang Chen, who continues to work on finding a better function that takes account of both employment and the proportion of employment (a compromise between the two).

References

Frolova, O., Fillion, J.-M. and J.-L. Tambay (2009), “Confid2: Statistics Canada’s new tabular data confidentiality software”, SSC Annual meeting, *Proceedings of the Survey Methods Section*.

Tambay, J.-L. and J.M. Fillion (2011), “New business survey confidentiality software G-Confid”, paper presented at teh Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spaine, October 26-28 2011.

Statistics Canada (2011), “Guide de l’usager G-Confid”, internal document.

Assessing disclosure risk in perturbed microdata

Natalie Shlomo¹

Abstract

We reconcile two methods for assessing disclosure risk in survey microdata based on the probability of a correct match to an external population file. The methods are the Fellegi and Sunter (1969) probabilistic record linkage framework and the probabilistic modelling framework based on the Poisson log-linear model. Skinner (2008) showed that the two methods are essentially the same. We provide empirical evidence of this result and demonstrate how disclosure risk can be assessed for a highly perturbed dataset containing business data from a 1982 Queensland, Australia Survey of Sugar Farms. We propose to estimate the probability of a correct match using probabilistic record linkage based on distance metrics between original and perturbed values.

Key Words: Probabilistic record linkage; Poisson log-linear model; Additive noise; Matching probability.

1. Introduction

Disclosure risk occurs when there is a high probability that an intruder can identify an individual in released microdata and confidential information may be revealed. For sample microdata arising from social surveys, the disclosure risk scenario is typically based on the assumption that an intruder can link the sample microdata to available public data sources through a set of identifying key variables that are common to both sources. The identification of an individual can then be used to obtain more sensitive information and lead to the disclosure of attributes. In order to limit the risk of identification, the statistical agency will implement disclosure limitation methods on the identifying variables through coarsening, sub-sampling or use perturbative disclosure limitation methods which alter the data by introducing forms of misclassification.

Disclosure risk is typically assessed through the notion of population uniqueness (see: Bethlehem, Keller and Pannekoek, 1990, Skinner and Holmes, 1998, Elamir and Skinner, 2006, Skinner and Shlomo, 2008). The probabilistic modelling framework relies on distributional assumptions to estimate population parameters for model-based disclosure risk measures. One individual disclosure risk measure is defined as the matching probability of a sample unique to the population based on a common set of key variables. The global measure of disclosure risk is obtained by summing over the matching probabilities of the sample uniques to obtain the expected number of correct matches. Shlomo and Skinner (2010) expanded the original probabilistic modelling framework of Elamir and Skinner, 2006 and Skinner and Shlomo, 2008 to take into account measurement errors in the key variables, either arising naturally through data processing stages or purposely introduced into the data as a perturbative disclosure limitation method.

The disclosure risk assessment for business microdata arising from Establishment Surveys should take a different approach. The surveys typically employ large sampling fractions where large businesses are sampled with certainty and the distributions are skewed. Therefore, Business microdata are typically treated as a full-population Census. The disclosure risk scenario is based on inferential disclosure where the intruder is assumed to be one of the competing businesses in a cell of a table with prior knowledge of the distribution of values within the cell and can infer on the values of competing businesses. In this case, sensitive variables are also key variables and need to be taken into account in the risk assessment. In recent years, there has been much research directed to releasing microdata arising from business surveys by applying highly perturbative disclosure limitation techniques. The disclosure limitation method for these types of datasets are based on generating partially synthetic datasets where

¹Natalie Shlomo, Southampton Statistical Sciences Research Institute, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom. E-mail:N.Shlomo@soton.ac.uk.

values are drawn from models derived from both perturbed and unperturbed variables (see: Raghunathan, Reiter, and Rubin, 2003, Reiter, 2005).

Assessing disclosure risk for highly perturbed datasets has typically been carried out through the probabilistic record linkage framework of Fellegi and Sunter, 1969. One of the first examples was carried out in Spruill, 1982 who linked perturbed sample microdata back to the original sample using distance based record linkage. More recent examples use the Fellegi and Sunter (F&S) record linkage theory (Yancy, Winkler and Creecy, 2002, Hawala, Stinson and Abowd, 2005 and Torra, Abowd and Domingo-Ferrer, 2006). In the F&S record linkage framework, each potential pair is assigned a matching weight. The matching weights are sorted and appropriate cut-offs determined according to pre-specified Type I and Type II error bounds. Pairs with high matching weights are considered to be correct matches and pairs with low matching weights are considered to be correct non-matches. The matching weights are used to calculate the probability of a correct match given an agreement.

In this paper we reconcile the two methods for disclosure risk assessment as shown in Skinner (2008). In section 2 we introduce the notation and theory of the two frameworks for disclosure risk assessment: the F&S probabilistic record linkage framework and the probabilistic modelling framework, and provide empirical evidence of the relationship between the two methods when taking into account errors arising from perturbation. In Section 3 we demonstrate the risk assessment on a highly perturbed business dataset containing sugar cane farms from a 1982 survey of the Sugar Cane Industry in Queensland, Australia (Chambers and Dunstan, 1986). We end with a conclusion in Section 4.

2. Notation and Theory

In this section we describe the F&S probabilistic record linkage framework and the probabilistic modelling framework based on the notion of population uniqueness whilst taking into account the perturbation. We reconcile the two frameworks.

2.1 Fellegi and Sunter Probabilistic Record Linkage

Using the notation of Skinner, 2008, let \tilde{X}_a denote the value of the vector of cross-classified identifying key variables for unit a in the microdata ($a \in s_1$) and X_b the corresponding value for unit b in the external database ($b \in s_2$). Note that s_2 can be the population P or any subset $s_2 \subset P$. The different notation of X allows for different values of the two vectors due to natural misclassification in the data or an application of a perturbative disclosure limitation method to the sample microdata file. Denote this misclassification matrix by:

$$P(=\tilde{X}_a k | X_a = j) = \theta_{kj}. \quad (1)$$

Based on the F&S theory of record linkage, a comparison vector $\gamma(\tilde{X}_a, X_b)$ is calculated for pairs of units $(a, b) \in s_1 \times s_2$ where the function $\gamma(.,.)$ takes values in a finite comparison space Γ . For the disclosure risk scenario we assume that the intruder uses the comparison vector to identify pairs of units which contain the same unit $(a, a) \in s_1 \times s_2$. Typically the intruder will use a combination of exact matching and probabilistic matching by considering only pairs that are blocked through an exact match on some subset $\tilde{s} \subset s_1 \times s_2$. The intruder seeks to partition the set of pairs in \tilde{s} into a set of matches: $M = \{(a, b) \in \tilde{s} | a \in s_1, b \in s_2, a = b\}$ and non-matches: $U = \{(a, b) \in \tilde{s} | a \in s_1, b \in s_2, a \neq b\}$. The approach by F&S is to define the likelihood ratio $m(\gamma)/u(\gamma)$ as the matching weight where: $m(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma | (a, b) \in M)$ and $u(\gamma) = P(\gamma(\tilde{X}_a, X_b) = \gamma | (a, b) \in U)$. We denote $m(\gamma)$ as the m -probability and $u(\gamma)$ as the u -probability. The higher values of the likelihood ratio are more likely to belong to M and the lower values of the likelihood ratio are more likely to belong to U . In addition, under the assumption of independence the m -probability and the u -probability can be split into individual components for each separate key variable. Let $p = P((a, b) \in M)$ the probability that the pair is in M . The probability of a correct match $p_{M|\gamma} = P((a, b) \in M | \gamma(\tilde{X}_a, X_b))$ can be calculated using Bayes Theorem:

$$p_{M|Y} = m(\gamma)p / [m(\gamma)p + u(\gamma)(1-p)]. \quad (2)$$

An intruder might estimate the matching parameters $m(\gamma), u(\gamma)$ and p by linking the released microdata to an external file containing all or subsets of the population. The parameters can be estimated using the EM algorithm which is an iterative maximum likelihood estimation procedure for incomplete data. Based on the estimation of the parameters, we calculate the probability of a correct match given an agreement $p_{M|Y}$ by (2). These probabilities can be used as individual record-level measures of disclosure risk and aggregated to obtain global measures of disclosure risk.

2.2 Probabilistic Modelling for Measuring Identification Risk

The probabilistic modelling framework for estimating the risk of identification is based on theory which uses models for categorical key variables. Let $\mathbf{f} = \{f_j\}$ denote a q -way frequency table, which is a sample from a population table $\mathbf{F} = \{F_j\}$, where $j = (j_1, \dots, j_q)$ indicates a cell and f_j and F_j denote the frequency in the sample and in the population cell j , respectively. Denote by n and N the sample and population size, respectively and the number of cells by J . We assume that the q attributes in the table are categorical identifying key variables. Disclosure risk arises from small cells, and in particular when $f_j = F_j = 1$ (sample and population uniques). We focus on a global disclosure risk measure based on sample uniques: $\tau = \sum_j I(f_j = 1)1/F_j$. This measure is the expected number of correct matches if each sample unique is matched to a randomly chosen individual from the same population cell. We consider the case that \mathbf{f} is known, and \mathbf{F} is an unknown parameter and the quantity τ should be estimated. An estimate of τ is:

$$\hat{\tau} = \sum_j I(f_j = 1)\hat{E}[1/F_j | f_j = 1] \quad (3)$$

where \hat{E} denotes an estimate of the expectation. The formula in (3) is naïve in the sense that it ignores the possibility of misclassification. A common assumption in the frequency table literature is $F_j \sim \text{Poisson}(\lambda_j)$, independently, where $\sum_j F_j = N$ is a random parameter. Binomial (or Poisson) sampling from F_j means that $f_j | F_j \sim \text{Bin}(F_j, \pi_j)$ independently, where π_j is the sampling fraction in cell j . By standard calculations we then have:

$$f_j \sim \text{Poisson}(\lambda_j \pi_j) \text{ and } F_j | f_j \sim f_j + \text{Poisson}(\lambda_j(1 - \pi_j)), \quad (4)$$

where $F_j | f_j$ are conditionally independent.

We take the approach as developed in Skinner and Holmes, 1998, Elamir and Skinner, 2006 and Skinner and Shlomo, 2008 and use log linear models to estimate population parameters and estimate identification risk. The sample counts $\{f_j\}$ are used to fit a long-linear model: $\log \mu_j = x'_j \beta$ where $\mu_j = \lambda_j \pi_j$ in order to obtain estimates for the parameters: $\hat{\lambda}_j = \hat{\mu}_j / \pi_j$. Under simple random sample and $\pi_j = \pi$ for all j , the maximum likelihood (MLE) estimator $\hat{\beta}$ may be obtained by solving the score equations: $\sum_j [f_j - \exp(x'_j \beta)] x_j = 0$. Under a complex survey design and differential weights, a pseudo-likelihood approach can be taken for estimating $\hat{\beta}$. Using the second part of (4), the expected individual disclosure risk measures for cell j is defined by:

$$E_{\lambda_j}(1/F_j | f_j = 1) = [1 - e^{-\lambda_j(1-\pi)}] / [\lambda_j(1-\pi)]. \quad (5)$$

Plugging $\hat{\lambda}_j$ for λ_j in (5) leads to the desired estimates $\hat{E}_{\hat{\lambda}_j}[1/F_j | f_j = 1]$ and then to $\hat{\tau}$ of (3).

The probabilistic modelling approach did not consider the case of misclassification naturally arising in surveys or purposely introduced into the data as a disclosure limitation method. Shlomo and Skinner (2010) defined disclosure risk measures that take into account misclassification. In this case, the disclosure risk measure is:

$$[\theta_{jj}/(1-\pi\theta_{jj})]/[\sum_k F_k \theta_{jk}/(1-\pi\theta_{jk})] \quad (6)$$

and it follows that (6) is less than $1/F_j$ with equality holding if there is no misclassification.

If the sampling fraction is small as for many social surveys, we can approximate (6) by:

$$\theta_{jj}/\tilde{F}_j \quad (7)$$

where \tilde{F}_j notes the population arising from the perturbed sample. Note that the approximations in (7) does not depend upon θ_{jk} for $j \neq k$ and so knowledge of these probabilities is not required in the estimation of risk if ‘acceptable’ estimates of θ_{jj} and \tilde{F}_j are available. The definition of risk in (7) applies to a specific record. The aggregated measure across sample unique records, defined from (7) is

$$\tau_\theta = \sum_{j \in SU} \theta_{jj}/\tilde{F}_j \quad (8)$$

where SU is the set of key variable values which are sample uniques in the perturbed sample. Similar to the case with no misclassification, this measure may be interpreted as the expected number of correct matches among sample uniques.

Since the values of F_j or \tilde{F}_j appearing in (7) through (8) are unknown, we need to estimate them. Expression (8) provides a simple way to extend the log-linear modelling approach provided θ_{jj} is known (which is the case if the statistical agency purposely perturbs the data for disclosure limitation). Using the perturbed sample counts \tilde{f}_j , $j=1, \dots, J$ we estimate $1/\tilde{F}_j$ from the \tilde{f}_j , $j=1, \dots, J$ by fitting a log-linear model to the \tilde{f}_j , $j=1, \dots, J$. This results in an estimate $\hat{E}(1/\tilde{F}_j | \tilde{f}_j = 1)$ based on the assumptions of the Poisson distribution for the population and sample counts. These estimates should be multiplied by θ_{jj} values and summed if aggregate measures of the form in (8) are required.

2.3 Reconciling the Frameworks

Skinner 2008 relates the F&S record linkage framework to the probabilistic modelling framework by providing the following examples:

Example 1: Assume no misclassification has occurred, *i.e.*, $\tilde{X}_a = X_a$ in both the population (P) and the sample (s) and that the true match status is known by the agency. Assume that sample (s) was drawn by simple random sampling from the population P . We calculate the decision table in Table 1 for each $X_a = j$ in the realized sample where the rows are a binary agreement/disagreement on the comparison vector: $\gamma(X_a, X_b)$ for pairs $(a, b) \in s \times P$ and the columns the matching status. From Table 1 we calculate directly $p_{M|\gamma} = 1/F_j$. We also obtain: $m(\gamma) = f_j/n$, $u(\gamma) = f_j(F_j - 1)/n(N - 1)$ and the probability of a correct match: $p = 1/N$. Using Bayes formula:

$$P_{M|\gamma} = \frac{1/N \times f_j/n}{1/N \times f_j/n + (1-1/N)f_j(F_j - 1)/n(N - 1)} = \frac{1}{F_j} \quad (9)$$

Small F_j therefore results in a high probability of a correct match given an agreement in the comparison vector.

Table 1: Contingency table of binary agreement status and match status for $X_a = j$ with no misclassification

	Non-match	Match	Total
Disagree	$n(N-1) - f_k(F_k - 1)$	$n - f_k$	$Nn - f_k F_k$
Agree	$f_k(F_k - 1)$	f_k	$f_k F_k$
Total	$n(N-1)$	n	Nn

Example 2: In continuation of Example 1, assume now that the microdata has undergone misclassification (either as a result of errors or purposely perturbed for disclosure limitation). Denote \tilde{f}_j the observed misclassified sample counts with $\tilde{X}_a = j$ derived by $\tilde{f}_j = \theta_{jj}f_j + \sum_{k \neq j} \theta_{jk}f_k$. We calculate the contingency table on the realized misclassified sample in Table 2 for each $\tilde{X}_a = j$ where the rows are a binary agreement/disagreement on the comparison vector: $\gamma(\tilde{X}_a, X_b)$ for pairs $(a, b) \in s \times P$ and the columns the matching status.

Table 2: Contingency table of binary agreement status and match status for $\tilde{X}_a = j$ with misclassification

	Non-match	Match	Total
Disagree	$Nn - n - \tilde{f}_k F_k + M_{kk}f_k$	$n - M_{kk}f_k$	$Nn - \tilde{f}_k F_k$
Agree	$\tilde{f}_k F_k - M_{kk}f_k$	$M_{kk}f_k$	$\tilde{f}_k F_k$
Total	$Nn - n$	n	Nn

From Table 2, we can calculate directly $p_{M\gamma} = \theta_{jj}f_j / \tilde{f}_j F_j \approx \theta_{jj} / \pi \tilde{f}_j \approx \theta_{jj} / \tilde{F}_j$ where \tilde{F}_j is the number of units in the population (P) with $\tilde{X}_a = j$ (imagining that the misclassification takes place before the sampling). We also obtain: $m(\gamma) = \theta_{jj}f_j / n$, $u(\gamma) = (\tilde{f}_j F_j - \theta_{jj}f_j) / n(N-1)$ and the probability of a correct match: $p = 1/N$. Using Bayes formula:

$$p_{M\gamma} = \frac{1/N \times \theta_{jj}f_j / n}{1/N \times \theta_{jj}f_j / n + (1-1/N)(\tilde{f}_j F_j - \theta_{jj}f_j) / n(N-1)} \approx \frac{\theta_{jj}}{\pi \tilde{f}_j} \approx \frac{\theta_{jj}}{\tilde{F}_j}. \quad (10)$$

Expression (10) is similar to expression (8) derived from the probabilistic modelling framework.

3. Assessing Disclosure Risk of Perturbed Business Microdata

We demonstrate how we can assess the disclosure risk of highly perturbed Business microdata. Since Business microdata is treated as Census data, the matching probability in both frameworks depends solely on the probability of not being in error θ_{jj} . In this case, we assume the conditional independence assumption of F&S and analyse the probability of not being perturbed separately for each variable, both the identifying and sensitive variables.

3.1 Perturbing the Dataset

We use the Sugar Farms Dataset corresponding to a sample of 338 Queensland sugar farms. The dataset has one nominal categorical variable (region) and five continuous variables (total area covered, amount of harvest, receipts, costs and profits where profits is the difference between receipts and costs). Traditional statistical disclosure control techniques were applied to perturb the sugar cane farm data with the aim to reduce identity and inferential disclosure risks. First, five records were deleted for having outlying values of receipts (over \$300K). The identifying key variable of area is coarsened by categorizing the variable into nine groups. Coarsening the identifying variable reduces the disclosure risk by making it more difficult to use as a matching variable and generally provides more protection than a perturbative method since it reduces the information that is released. The target survey variables of harvest, receipts, costs and profit were perturbed by additive noise generated from a Multivariate Normal distribution within small groups. The Multivariate Normal Distribution is used in order to preserve the mean and covariance

structure of the perturbed target variables as well as ensure the edit constraint that profit is equal to receipts minus the costs (Shlomo and DeWaal, 2008).

The noise addition was carried out as follows: Consider the four target variables harvest (t), receipts (x), costs (y) and profit (z) where $x - y = z$ and assume that they have a joint mean of $\boldsymbol{\mu}^T = (\boldsymbol{\mu}_t, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{\mu}_z)$ and a covariance matrix $\boldsymbol{\Sigma}$. We generate multivariate random noise: $(\varepsilon_t, \varepsilon_x, \varepsilon_y, \varepsilon_z)^T \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma})$, where the superscript T denotes the transpose. In order to preserve sub-totals and limit the amount of noise, the random noise was generated within small groups of records defined by the quintiles of receipts (note that we drop the index for quintiles). The vector $\boldsymbol{\mu}'$ contains the corrected means of each of the four variables t, x, y and z in the quintile based on a noise parameter δ and calculating: $d_1 = \sqrt{1 - \delta^2}$ and $d_2 = \sqrt{\delta^2}$:

$$\boldsymbol{\mu}'^T = (\boldsymbol{\mu}'_t, \boldsymbol{\mu}'_x, \boldsymbol{\mu}'_y, \boldsymbol{\mu}'_z) = \left(\frac{1 - d_1}{d_2} \boldsymbol{\mu}_t, \frac{1 - d_1}{d_2} \boldsymbol{\mu}_x, \frac{1 - d_1}{d_2} \boldsymbol{\mu}_y, \frac{1 - d_1}{d_2} \boldsymbol{\mu}_z \right).$$

The covariance matrix $\boldsymbol{\Sigma}$ is the original covariance matrix of the four variables in each quintile. For each separate variable, we calculate the linear combination of the original variable and the random noise generated above, for example, for record i : $z'_i = d_1 \times z_i + d_2 \times \varepsilon_{z_i}$. The mean vector and the covariance matrix remain the same as the original data and after the perturbation the additivity constraint is exactly preserved.

3.2 Proposed Method for Disclosure Risk Assessment

The perturbed Sugar Farms microdata is treated as Census data and therefore we have coarsened the identifying variable and perturbed the sensitive variables and all are used in the risk assessment process. We also assume that all records are population uniques since the cross-classification of the identifying and sensitive variables produce unique records. The risk assessment will be carried out by matching the perturbed dataset back into the original dataset and calculating the matching probability in (8) or (10) which depends in this case solely on θ_{ij} , the probability of the records in cell j of the cross-classified variables not being perturbed. In this case, all the cells are uniques.

We define this probability in terms of a metric taking a value between 0 and 1 that measures the (normalized) distance between the original and perturbed values of each variable in each record. The final matching probability for each record is obtained by taking the weighted sum of the metrics across all variables. For the continuous variables, the difference between the original and perturbed values of a variables is the noise that was generated by the multivariate Normal Distribution. For the coarsened variable, we take the distance from the mid-point of the interval. Define the difference for value i of a variable p between the original and perturbed values as ε_i^p .

We consider two options for calculating the distance metric taking on values between 0 and 1:

1. Option 1: Standardize the difference of value i for variable p : $Z_i^p = [\varepsilon_i^p - \bar{\varepsilon}_i^p] / s_\varepsilon^p$ and calculate:
 $Dist_i^p = 1 - |1 - 2\Phi(Z_i^p)|$
2. Option 2: Use the exponential function as follows: $Dist_i^p = \exp[-|\varepsilon_i^p| / med(\varepsilon_i^p)]$ where $med(\varepsilon_i^p)$ is the median values of ε_i^p .

We follow the method of the F&S record linkage to obtain the matching probability for each pair and aggregate them across the true matches to find the expected number of correct matches as shown in (8) and (10). In addition, the F&S record linkage allows the calculation of other types of risk measures which reflect the uncertainty of the intruder in being able to infer a correct match: the proportion of correct matches out of declared links and the odds of a correct match given an agreement (declared links that are true matches divided by declared links that are false matches). The dataset of 333 records was blocked on region (which was not perturbed) resulting in 31,367 possible pairs. We link on all other variables: coarsened area, and the noise adjusted harvest, receipts and costs. For each record, we propose to take a weighted sum of the distance metrics across variables by assuming the conditional independence assumption of the F&S framework. The m -probabilities of the F&S record linkage are represented by

the distance metrics since they reflect the errors induced by the disclosure control methods. The weights should therefore reflect the u -probabilities, *i.e.*, the odds that given an agreement on a value, the pair is a true match. The weights are calculated by a logistic regression where the response variable is 1 for a true match and 0 if not, and the independent variables are the distance metrics. These odds are then normalized to sum to one. We compare the weighted average of distance metrics to the case where a simple average is taken of the distance metrics.

In Table 3, we examine the proposed method for assessing disclosure risk on the Sugar Farms dataset under two levels of perturbation and the two options for weighting the distance metrics. The threshold for the F&S record linkage is determined by a pre-specified Type I error of 1.4%.

Table 3: Results of the Record Linkage Procedure for Assessing Disclosure Risk in the Perturbed Business Dataset (Type I error is 1.4%)

		$\delta = 0.4$		$\delta = 0.7$	
		Option 1	Option 2	Option 1	Option 2
Equal Weights	True Matches / Declared Links	0.297	0.290	0.160	0.151
	True Matches / False Matches	0.423	0.409	0.191	0.178
	Sum of Matching Probabilities	307.5	290.0	289.8	263.9
Weights Based on Normalized Odds	True Matches / Declared Links	0.307	0.313	0.168	0.175
	True Matches / False Matches	0.443	0.455	0.201	0.213
	Sum of Matching Probabilities	309.0	295.6	299.9	292.7

Table 3 shows the higher disclosure risk associated with the lower perturbation rate of $\delta = 0.4$. The weighted average of distance metrics gives more power to the record linkage framework compared to using the simple average. The two options for distance metrics are inconsistent for the two F&S risk measures based on the number of true matches out of declared links and the true matches to false matches. Whilst option 1 gives higher disclosure risk measures under the simple average with equal weights compared to option 2, the opposite occurs when using the weighted average according to the normalized odds, *i.e.*, option 1 gives lower F&S disclosure risk measures compared to option 2. As in all perturbative methods, the expected number of correct matches obtained from summing over the matching probabilities on true matches is high. The perturbation however raises the level of uncertainty in correct matches as seen in the true to false match rate. The results from this study show that the record linkage with the most power would use the distance metric of option 2 and the weighted average of distance metrics based on the normalized odds.

4. Conclusion

In this paper, we have provided evidence of the reconciliation between the F&S record linkage framework to the probabilistic modelling framework for estimating disclosure risk using a matching probability that can be calculated under both frameworks. We show how this matching probability can be calculated under the F&S record linkage framework for the case of a highly perturbed business dataset where it is typically treated as a Census and both sensitive and identifying variables are included in the disclosure risk scenario. In addition the F&S record linkage framework allows other types of risk measures which reflect the uncertainty of the intruder in obtaining a correct match.

Acknowledgements

This work is funded by the Grant Agreement No. 244767 under Theme 8 of the 7th Framework Programme of the European Union, Socio-economic Sciences and Humanities: BLUE-ETS.

References

- Bethlehem, J., Keller, W. and J. Pannekoek (1990), "Disclosure control of microdata", *Journal of the American Statistical Association*, 85, pp. 38-45.
- Chambers, R.L. and R. Dunstan (1986), "Estimating distribution functions from survey data", *Biometrika*, 73, pp. 597-604.
- Elamir, E. and C.J. Skinner (2006), "Record-level measures of disclosure risk for survey micro-data", *Journal of Official Statistics*, 22, pp. 525-539.
- Fellegi, I. and A. Sunter (1969), "A theory for record linkage", *Journal of the American Statistical Association*, 64, pp. 1183-1210.
- Hawala, S., Stinson, M. and J. Abowd (2005), "Disclosure risk assessment through record linkage", in *Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality*, Geneva.
- Raghunathan, T.E., Reiter, J. and D. Rubin (2003), "Multiple imputation for statistical disclosure limitation", *Journal of Official Statistics*, 19, No. 1, pp. 1-16.
- Reiter, J.P. (2005), "Releasing multiply imputed, synthetic public-use microdata: An illustration and empirical study", *Journal of the Royal Statistical Society, A*, Vol.168, No.1, pp. 185-205.
- Shlomo, N. and T. De Waal (2008), "Protection of micro-data subject to edit constraints against statistical disclosure", *Journal of Official Statistics*, 24, No. 2, pp. 1-26.
- Shlomo, N. and C.J. Skinner (2010), "Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata", *Annals of Applied Statistics*, Vol. 4, No. 3, pp. 1291-1310.
- Skinner, C.J. (2008), "Assessing disclosure risk for record linkage", in J. Domingo-Ferrer and Y. Saygin (Eds.) *Privacy in Statistical Databases*, Lecture Notes in Computer Science 5262, Berlin: Springer, pp. 166-176.
- Skinner, C. and D. Holmes (1998), "Estimating the re-identification risk per record in microdata", *Journal of Official Statistics*, 14, pp. 361-372.
- Skinner, C.J. and N. Shlomo (2008), "Assessing identification risk in survey microdata using log-linear models", *Journal of the American Statistical Association*, Vol. 103, No. 483, pp. 989-1001.
- Spruill, N.L. (1982), "Measures of confidentiality", *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 260-265.
- Torra, V., Abowd, J.M. and J. Domingo-Ferrer (2006), "Using Mahalanobis distance-based record linkage for disclosure risk assessment" in J. Domingo-Ferrer and L. Franconi (Eds) *Privacy in Statistical Databases*, LectureNotes in Compute Science, 4302, Berlin: Springer, pp. 233-242.
- Yancey, W.E., Winkler, W.E. and R.H. Creecy (2002), "Disclosure Risk Assessment in Perturbation Micro-Data Protection", in J. Domingo-Ferrer (Ed.) *Inference Control in Statistical Databases*, New York: Springer, pp. 135-151.

Privacy preserving probabilistic record linkage (P3RL) from A to Z: a GRLS example from SwissLinkage*

Adrian Spoerri, Kurt Schmidlin, Rainer Schnell and Kerri Clough-Gorr¹

Abstract

Probabilistic record linkage is at the top of the agenda of current health research. Decreasing research resources limit the production of new data. Therefore, existing data – often routine data – should be analysed whenever possible. Data protection issues prevent using the best existing data for many record linkage projects. Newly developed methods in privacy preserving record linkage like Bloom filters allow encryption of names, addresses, and other restricted personal data. Until recently it was not possible to apply probabilistic record linkage methods to encrypted data. This study aims to demonstrate the entire process of a privacy preserving probabilistic record linkage combining data Swiss Childhood Cancer Registry (SCCR) data with National Institute for Cancer Epidemiology and Registration (NICER) data using Bloom filters and GRLS linkage software. As the linkage team is not allowed to look at restricted personal data (ensures no breach of confidentiality) pre-processing of linkage variables has to be automated at the registry sites. We will demonstrate a tool for pre-processing as well for encryption and exportation of data to be used by the responsible persons at SCCR and NICER. The linkage process will then be done by SwissLinkage (independent third party). The linkage team will never have access to restricted data (*e.g.*, names) and will perform linkage using GRLS. GRLS of Statistics Canada was not designed to perform linkages with encrypted variables. We demonstrate the application of Bloom filter hash codes and calculation of a similarity measure (Dice coefficient) using GRLS. The ability to link data sources using important discriminating information such as patient name without breach of confidentiality – Privacy Protected Probabilistic Record Linkage (P3RL) – has the potential to ethically transform epidemiology research by making available cancer-related data anonymously and thus heretofore not previously accessible.

¹Adrian Spoerri and Kurt Schmidlin, University of Bern, Switzerland; Rainer Schnell, University Duisburg-Essen, Germany; Kerri Clough-Gorr, University of Bern and National Institute of Cancer Epidemiology and Registration, Switzerland.

*SwissLinkage is a centre of record linkage excellence at the University of Bern.

Standardised outputs and regionally-based classifications for minority populations as part of the England and Wales 2011 Census

Joe Traynor and Emma White¹

Abstract

The advantage of a census over survey samples is the ability to provide a broad range of information about the population in a small geographical area.

Disclosure control constraints are designed for the general population to protect personal information, including that pertaining to minority groups, at the smallest geographical level. However there are areas where minority populations are large enough to enable detailed outputs to be produced specific to them.

To address this issue, ONS plans to produce an algorithm that can be applied to any minority population e.g. an ethnic or religious group, such that once that population meets a specified threshold size at a given level of geography, predefined outputs are produced to meet the needs of that specific community.

The paper will describe the steps taken to design and test the algorithm, including the related user consultation process.

The successful design of such an algorithm has a number of benefits:

- Value is added to census information that is already collected;
- Identification of areas where minority groups are concentrated;
- Increased outputs relating directly to minority groups will lead to more informed and representative planning and policy decisions;
- Support for the equality agenda.

¹Joe Traynor and Emma White, Office for National Statistics, England.

SESSION 5A

WAKSBERG AWARD WINNER ADDRESS

Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?

Danny Pfeffermann¹

Abstract

Survey data are frequently used for analytic inference on statistical models, which are assumed to hold for the population from which the sample is taken. Survey data typically differ from other data sets in five main aspects.

- 1- The samples are selected at random with known selection probabilities, which allow using the randomization distribution over all possible sample selections as the basis for inference, instead of using the distribution underlying the population model. A combination of the two distributions is in common use.
- 2- The sample selection probabilities in at least some stages of the sample selection are generally unequal; when these probabilities are related to the model outcome variable, the sampling becomes informative and the model holding for the sample is then different from the target population model.
- 3- Survey data are almost inevitably subject to various forms of nonresponse, often of considerable magnitude, which again may distort the population model if the response propensity is correlated with the outcome of interest.
- 4- The data are often clustered due to the use of cluster samples. The clusters are ‘natural units’ and the observations within the same cluster are therefore correlated.
- 5- The data available to the modeler may be masked in order to protect the anonymity of the respondents.

Many approaches have been proposed in the literature for estimating population models from complex survey data possessing these features. The approaches differ in the conditions underlying their use, the data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. This heterogeneity means that there exists no single approach that can be considered as best in all situations. That being the case, a fundamental question arising is which approach or approaches should be used for a given practical application.

In the present paper I review the various approaches proposed in the literature for dealing with these features, discussing their merits and limitations in light of the properties mentioned above. I also present simulation results, which compare the approaches when estimating regression models from a stratified sample in terms of bias, variance, and coverage rates.

Dr. Pfeffermann’s complete paper can be found in *Survey Methodology* December 2011.

¹Danny Pfeffermann, Hebrew University of Jerusalem, Israel, and University of Southampton, UK.

SESSION 6A

**STANDARDS AND GUIDELINES FOR THE DESIGN AND TESTING OF
INTERNET QUESTIONNAIRES**

Developing electronic questionnaire guidelines: Issues and challenges in a changing environment

Anne-Marie Côté, David Lawrence and Paul Kelly¹

Abstract

Multi-mode data collection, in particular the use of self-administered Internet questionnaires, presents challenges for Statistics Canada surveys. To ensure that the collected information meets our quality requirements, special care must be taken during the development of the conceptual framework for the Internet questionnaires. The design, and testing procedures as well as the survey content must be taken into account. A corporate Internet questionnaire solution that includes standard processes for the rendering of electronic questionnaires is currently underway. This paper addresses some of the issues and challenges concerning development of guidelines regarding the design and development of Internet questionnaires at Statistics Canada.

Key Words: Electronic questionnaire; Guidelines; Standards; Mode of collection.

1. Introduction

Statistics Canada has been offering various Electronic Data Reporting (EDR) solutions to respondents for some time. In 2010, a new corporate initiative to use electronic questionnaires (EQ) as the primary mode of collection was approved. The initiative is the result of continuous attempts to develop an EDR option that complies with Statistics Canada's strict confidentiality and security requirements, is technically sustainable, and meets respondent expectations. The option must also be compliant with the Common Look and Feel guidelines for Canadian Government websites and meet accessibility requirements to ensure equitable access to all content on Government of Canada websites.

This paper provides a brief overview of the corporate initiative especially as it relates to data collection with EQ. The paper will also cover the establishment and implementation of EQ guidelines and standards at Statistics Canada; how the existing guidelines were developed, as well as how they continue to evolve.

2. New EQ Project

2.1 Objectives

The focus on the EQ as the primary mode of collection is motivated by changing expectations of the Canadian population as well as by the current economic conditions. Statistics Canada's vision is to design surveys to make EQ the primary mode of collection and the first mode in a sequential multi-mode environment. This focus is motivated by the need to look for more efficient ways to conduct data collection as well as capitalize on generic functionalities, reducing the need for customized collection tools.

¹Anne-Marie Côté, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (anne-marie.cote@statcan.gc.ca); David Lawrence, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (dave.lawrence@statcan.gc.ca); Paul Kelly, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (paul.kelly@statcan.gc.ca).

By offering EQ, Statistics Canada will achieve the following:

- Provide a collection mode that is convenient for respondents and available 24 hours a day, seven days a week.
- Reduce costs associated with interviewer administered and postal surveys, assuming adequate Internet take-up rates.
- Improve timeliness of responses via Internet compared to paper.
- Reduce development time as Statistics Canada anticipates a shorter development period for EQ compared to Computer-Assisted Interview (CAI) applications.
- Maintain or improve data quality through the use of online edits, automated skips and additional help functions to ensure that survey questions and concepts are understood by respondents, which will lead them to provide higher quality answers.
- Meet respondents' expectations as more and more respondents stated that they would prefer completing an EQ instead of spending time with an interviewer or filling out a paper questionnaire.

2.2 Development and Deployment

There are six main steps that comprise the EQ development process: initial assessment, input specifications, electronic questionnaire generation, testing, data collection, and retrospective evaluation (post-mortem).

Since each survey is unique, the first step is an initial assessment. Each survey is reviewed by the collection experts in collaboration with the subject matter area to assess how the electronic questionnaire strategy would fit with existing collection and processing requirements.

The EQ input specifications are then developed where generic functionalities are used and existing EQ guidelines and standards are respected. If required, additional functionalities may also be developed to meet the needs of a specific survey. However, these functionalities must be generic in order for Statistics Canada to be able to reuse them in other surveys.

EQ is automatically generated from a generic dynamic input specifications template. In other words, once questionnaire specifications (*e.g.*, questions, flows, validation messages, *etc.*) are prepared, the specifications file is inputted into the Electronic Questionnaire Generation System (EQGS), which automatically processes the inputs and delivers a complete EQ as an output. The EQGS was developed in-house by Statistics Canada. It is expected to reduce the time required to develop and test each EQ application. Moreover, applications that are generated by the EQGS are also compliant with Statistics Canada confidentiality and security requirements, with the Common Look and Feel Standards for the Internet (CLF) guidelines of Canadian Government and meet accessibility requirements. The EQGS is the cornerstone of the new EQ project.

Following the generation of the EQ application, thorough testing is conducted to ensure the application is in accordance with Statistics Canada standards of quality. The testing phase includes quality assurance review, user acceptance testing, end-user usability testing and compliance with accessibility and CLF requirements.

Next is data collection. An invitation email is sent to each respondent. The email contains a brief introduction to what the survey is and provides a link and a secure access code to log into the EQ. Once the link is activated, the respondent can access the secure electronic portal and is able to fill out his/her questionnaire online. When the respondent submits the questionnaire, the collected information is integrated with the information gathered from other modes of collection in the existing data processing system.

Finally, there is a retrospective evaluation where the observations and lessons learned are used to improve the process for following cycles or surveys.

2.3 Collection Partners

A standardized development and deployment process was created to serve all collection partners. Special care was taken to consider all aspects, such as questionnaire content, collection tools, processing systems, etc. The process

schedule is generally short and involves different service areas, which, in turn, requires continuous coordination of activities between all involved partners and rigorous use of project management tools.

At Statistics Canada, different divisions and resource centers offer specialised services to survey program areas. In the case of the development and deployment of the EQ, specialised services are offered by the Operation and Integration Division (OID), the Collection System and Infrastructure Division (CSID), the Collection Planning and Management Division (CPMD) as well as the Questionnaire Design Resource Centre (QDRC).

The EQ development project is coordinated and overseen by CPMD. They also ensure that the new mode of collection integrates well with the existing collection process.

In a typical survey, subject matter areas provide specifications to OID who are responsible for EQ design and preparation of an Excel template, which feeds into the EQGS. OID and CPMD are also responsible for ensuring that the specifications respect all of the EQ guidelines and standards. Once the EQ Excel template is ready, OID renders the EQ application and initiates the testing process.

CSID is responsible for developing and maintaining the entire EQ platform, which includes the EQGS. CSID also ensures that the required functionalities for a given survey are ready and available in the EQGS.

Questionnaire design experts from QDRC assist in the development of the EQ design and specifications. The QDRC also coordinates and carries out EQ testing with end-users.

3. Design Guidelines and Standards

3.1 EQ Standards Committee

At Statistics Canada, the EQ Standards Committee was established in 2010 to develop and compile a set of design guidelines for the EQs to ensure a common interface for respondents, consistency in screen display and layout, the use of Canadian Government requirements for a common look and feel, as well as standardized functionalities and approaches.

The committee is a multidisciplinary team. This group includes system designers, usability experts, questionnaire design specialists, accessibility experts, business and social collection coordinators and operations managers.

3.2 Guidelines

The goal of the EQ guidelines and standards is to provide a common interface for users while reducing development, testing and training times as well as costs.

The first version of Statistics Canada's design guidelines and standards has been available internally since May 2011. The standards were formulated based on external documentation and literature, existing practices at Statistics Canada, and findings and observations from end-user testing. To date, a single set of guidelines and standards has been established for both business and social surveys.

The standards describe how EQ should be rendered and displayed. For the moment, they cover design features of web-based surveys, such as navigation, usability, accessibility and screen layouts. Traditional components of questionnaire design, such as content wording and ordering, are not yet addressed. The standards will continue to evolve in response to EQ testing, research findings, and results from ongoing data collection. They will also be expanded to include new survey requirements. The standards committee is currently looking at more complex issues, such as harmonization of common content, to be included in the guidelines.

3.3 In-scope Elements

The committee has examined and established standards for the following items:

- Government of Canada CLF elements and compliance;
- survey window interface;
- consistency in tool bars;
- navigation keys and buttons functionality and placement;
- question categories and placement;
- instructions;
- response categories, placement and selection;
- movement and skip functions;
- fonts and styles;
- consistent use of colors and/or black and white;
- character limitations;
- functionalities for negative values, decimals and thousands;
- generic content for the introductory page and contact information;
- message display (edits, warnings);
- session time outs;
- browsing and printing.

4. Project Status

To date, collection using an EQ has either been conducted or is currently underway for 16 business surveys, two agricultural surveys and two institutional social surveys. Approximately 10 additional surveys should be in production by March 31, 2012.

There is also ongoing testing for social surveys. Some selected modules of the Canadian Labour Force Survey (LFS) have been tested with former LFS respondents. An integrated electronic version of the General Social Survey's questionnaire was also qualitatively tested with potential respondents.

Statistics Canada's objective is to have close to 200 business and household surveys using EQ as a mode of collection by March 31, 2015.

5. Challenges

5.1 Implementation of a New Methodology

The implementation of a new methodology, such as EQ data collection, brought many challenges and all collection partners have had to adapt. For example, many business surveys have been traditionally administered using a paper questionnaire. Subject matter areas have had to understand that an EQ cannot always be an electronic replica of the paper instrument. In these situations, steps were taken to ensure that the data collected via EQ can be integrated into existing collection and processing systems and are analytically comparable with what is collected through other modes. Special attention was paid to try to mitigate possible mode effects.

5.2 Communication

Having a constant and continuous line of communication was a challenge as there were many collaborators and the design guidelines were new and evolving. The business process associated with the development and deployment was new and there were many subject matter areas that were going through this process for the first time.

5.3 Questionnaire Content

Another challenge introduced by the new collection method is the need for some standardized elements for the questionnaire content. Although a policy for reviewing and testing questionnaires has existed at Statistics Canada for almost 20 years, to date there is no documented design standards for business paper instruments. Over time, each paper questionnaire has been ‘customized’. During the EQ specification process, there is more to consider than just what is presented on the paper questionnaire. A variety of elements can affect the visual design of the EQ such as questions and instructions wording and placement, question numbering, section headings and edit messages. These features need to be well-defined. To improve the EQ development process, the content review should be thorough and questionnaire design experts should be involved early in the process, ideally, during the input specification step.

5.4 Planning and Implementing Qualitative Testing

There were challenges in planning and implementing the qualitative testing of the EQ. At Statistics Canada, qualitative testing for paper questionnaires only involved the subject matter area and the questionnaire design experts (QDRC). For EQ, the qualitative testing takes more time to implement, involves more collection partners and requires a careful coordination of activities.

To date, time constraints have limited the ability to fully test the EQ in a live environment on the respondent’s computer. Qualitative testing has primarily been conducted by deploying a test version of the EQ on Statistics Canada laptops. So far, this has restricted the ability to explore all aspects of the EQ experience such as receiving an email invitation and logging into the EQ portal while using user-specific settings established by respondents on their computers.

5.5 Technical and Technological Changes

With an increasing number of surveys that are planning to use EQ as a mode of collection and the fact that each survey tackles different subject matter content, new functionalities must be developed to accommodate different survey needs. Also, given the vast number of applications and Web-based surveys that can be found on the Internet, respondents are becoming more aware of what is available in terms of functionalities and visual design. In order to preserve the response rates and reduce response burden, Statistics Canada must keep up with these innovations when offering an online response option.

The speed at which new technologies appear and their constant evolution is also a challenge for Statistics Canada. For example, the release of a new version of a common browser leads to more testing and possible adjustments to the EQ platform as the agency needs to be compatible with the new version to meet the Government of Canada Web sites accessibility requirements.

6. The Future

The EQ design and development process continues to evolve and all collection partners continue to learn and adapt. Findings from qualitative testing and data collection must still be incorporated back into the EQ guidelines and standards and, eventually, traditional aspects of questionnaire design, will also be incorporated.

In order to gain efficiencies, questionnaire design experts will be involved in the early stages of the EQ content, specification and rendering processes. Questionnaire design experts are also looking into ways to improve pretesting strategies and skills. Different approaches using cognitive interviews coupled with alternate lines of information gathering such as paradata or new technological tools (*e.g.*, eye tracking system) are being considered.

The members of Statistics Canada’s EQ development team keep themselves up to date with the new and emerging technologies. This includes ensuring the EQGS is compatible with different browsers and operating systems. The team is also evaluating the feasibility and potential gains that could result if Statistics Canada had a collection application for mobile devices.

The use of EQ collection for social surveys is still in development. Questionnaire experts, collection partners and various subject matter teams are exploring different methods to transition interviewer-assisted questionnaires into self-complete instruments. The introduction of the EQ collection mode to social surveys also brought the need to study the feasibility of using prefilled information collected in previous interviews. Issues regarding confidentiality and respondent reactions will be evaluated.

Finally, all Statistics Canada collection partners are keeping apprised of ongoing research and developments by international colleagues.

References

Karaganis, M. (2011), “Development and Implementation of E-questionnaire as a Primary Collection Mode at Statistics Canada”, internal unpublished document, Ottawa, Canada, Statistics Canada.

Lawrence, D. (2011), “Developing Electronic Questionnaires at Statistics Canada: Experiences and Challenges in a Changing Environment”, paper presented at the Internet Survey Methodology Workshop, Den Haag, Netherlands.

Statistics Canada (2011), “E-Questionnaire Design Guidelines and Standards”, internal unpublished report, Ottawa, Canada.

GINO++: A generalized system for web surveys

Renato Torelli¹

Abstract

The generalized system, GINO++, allows the researcher himself to: design and implement web questionnaires independently and quickly through graphical interface; accompany the questions by hints and tool tip; include checks on the data entered; monitor the progress of the survey in real-time.

The final user (survey respondent) is able to: complete the questionnaire in multiple sessions by saving each time and by sending only at the end of compiling; make local prints (html or excel) of the questionnaire; insert notes for each individual variable (if provided by the researcher during design); include the sources of each question (if provided by the researcher during design).

Whenever the respondent saves the entered data, it is directly inserted on tables of a DB Oracle generated by the application with a user-defined name.

GINO++ features two main areas: Management and design of questionnaire and metadata; Monitoring of the survey.

Key Words: Web surveys; Generalized software; Questionnaires design; Monitoring of surveys.

1. Introduction

1.1 Description

GINO++ [*much more than Gathering Information Online*] allows the statistician himself (which means without software developers or computer scientists!) to perform three stages of a survey: designing, capturing and monitoring.

In the “designing stage” the statistician can design a questionnaire, but also modify questionnaire just before the survey (and improve the layout thereafter) and insert rules to check the entered values.

In the “capturing phase” he can get data online and put them directly into a database. Thereafter it is possible to view each questionnaire, even if partially filled, exactly as it appears to the compiler (after a “Save” or “Send” action). Moreover, he can export just entered data in excel format, for example, and visualize the series by means of diagrams (provided that time series are planned).

In the “monitoring phase”, the researcher can constantly monitor, step by step, the activities of respondents and supervisors such as the number of accesses and questionnaires status. Furthermore, he can urge the laggards immediately, in every step of the survey: initial user registration, temporary submission, final submission. Lastly, he can make quality analysis using detailed reports about errors.

1.2 Phases of a survey

In the survey life cycle, GINO++ stands after the design of survey and before the phases of edit & imputation and data broadcasting. So, it deals with the middle phases of a survey: design of questionnaire and data capturing.

It is important to note that, while using this system, we obtain two results: the first is that we remove a burden on developers to implement ad-hoc questionnaires and the second is that we give autonomy to researchers.

¹Renato Torelli, Istat, via Cesare Balbo 16, Rome, Italy, 00184, torelli@istat.it.

1.3 Evolution

In Istat, we started the development of the first version of the system in 2008 and it was firstly adopted by the survey “*urban environment in Italy*”. Year by year we have been able to upgrade the system, with new functions implemented and new surveys as test cases.

Up to now we have covered three main areas of interest: environmental statistics, social statistics and economic statistics.

2. Architecture

2.1 Technological Components

The system consists of two main technological components: PHP is the web programming language server side and Oracle is the Relational DataBase Management System.

The DB is composed of two set of tables: Data tables and Metadata tables. As usual, data tables are typically few and large, while metadata tables are many and small.

Furthermore, the Data tables are divided in:

- *Source Data*: it is a static table and it may contain initial data for the personalization of the questionnaire and
- *Incoming Data*: it is a dynamic table and contains the data that come hand by hand by respondents

2.2 Configurations

The simplest management structure of a survey supported by GINO++ consists of two levels: the first is the owner level, *i.e.*, the Responsible of the Survey and the second is the respondent level, where respondents may be individuals, households, organizations, institutions, companies, and so on.

A more complex structure of a survey supported by GINO++ still consists of two levels (owner and respondent), but each respondent must answer to different themes (*i.e.*, many questionnaires for the same survey). In this configuration there is the coexistence of two roles at the respondent level: only one *thematic responsible* and several *thematic respondents*.

A third possible structure consists of three levels and differs from the previous due to the presence of an “intermediary body”.

The last configuration allows up to four monitoring levels.

2.3 Users management

Permissions to read and write are dynamically set from the responsible of the survey to each role, depending on each *status* of the questionnaire (*unprocessed, in process, submitted, validated and completed*).

The concurrency between users is managed. So, when the first user with write permission logs on to the questionnaire, all remaining users (as well as another session that was opened by the same user) are not allowed to write.

2.4 Basic components

The basic building blocks of the survey are: *Theme*, *Replication*, *Questionnaire*, *Section*, *Question*, *Variable*, *Rule* and *Classification*.

As shown in Figure 2.4-1, a survey may investigate several *Themes*, in such a case it results composed by several thematic questionnaires.

A *Theme* may recur in several *replications* (i.e., in different editions of the same survey).

A *Questionnaire* may be used from several surveys, themes and replications. Moreover, a questionnaire is composed of several *Sections*.

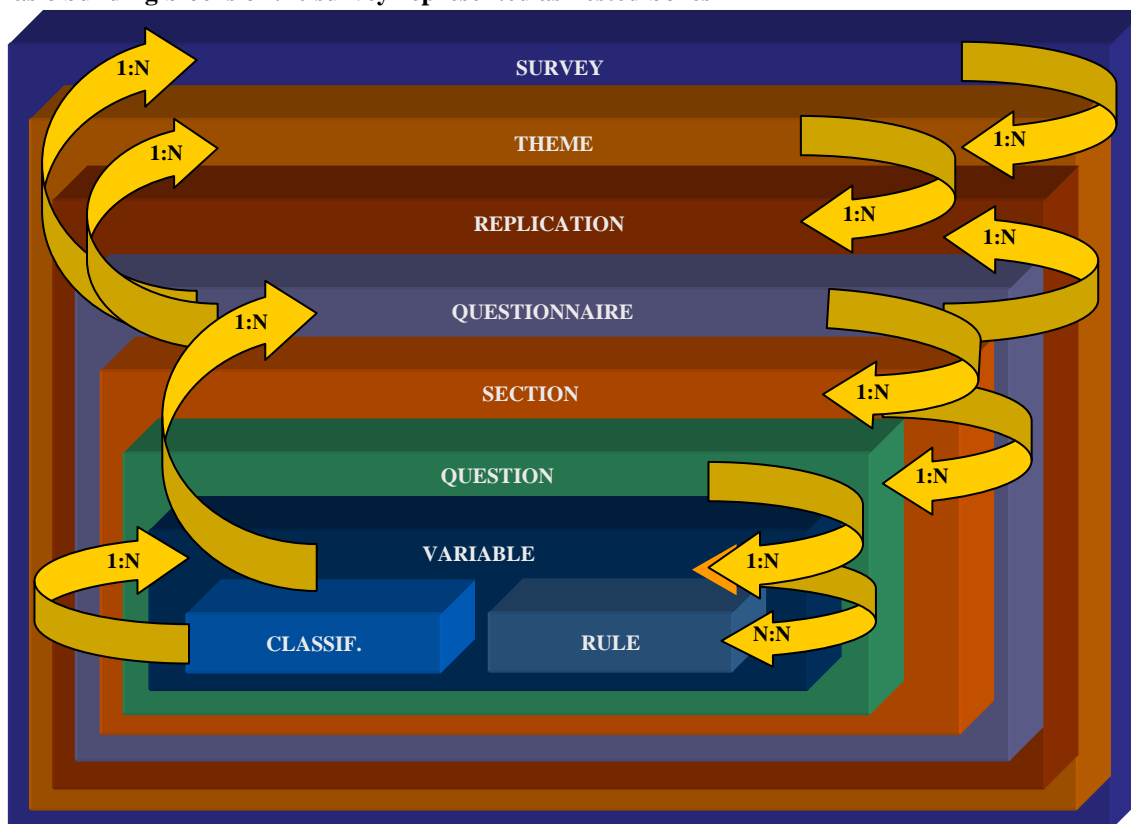
A *Section* is composed of several *Questions*, unless it is only an header.

A *Question* is composed of several *Variables*.

The same *Variable* may appear in multiple questionnaires, in multiple *Rules* and a *Rule* can be used from several *Variables*.

Finally, a *Classification* can be used for multiple *Variables*.

Figure 2.4-1
Basic building blocks of the survey represented as nested boxes



2.5 Features

The statistician will find in the system two main areas: one for the Management of metadata and one for the Monitoring of the survey.

The first area deals with the definition of the survey and with the building of the questionnaire.

The second area is concerned with Monitoring the registration of users, the status of questionnaires and, in general, the progress of the investigation.

3. Metadata management

3.1 Description

The metadata management allows to create, update and delete:

- The features of a survey including Themes, Replications, Status and Users;
- The overall world of Variables and Classifications;
- The Questionnaire organized in Sections, Questions and Rules.

3.2 Sections

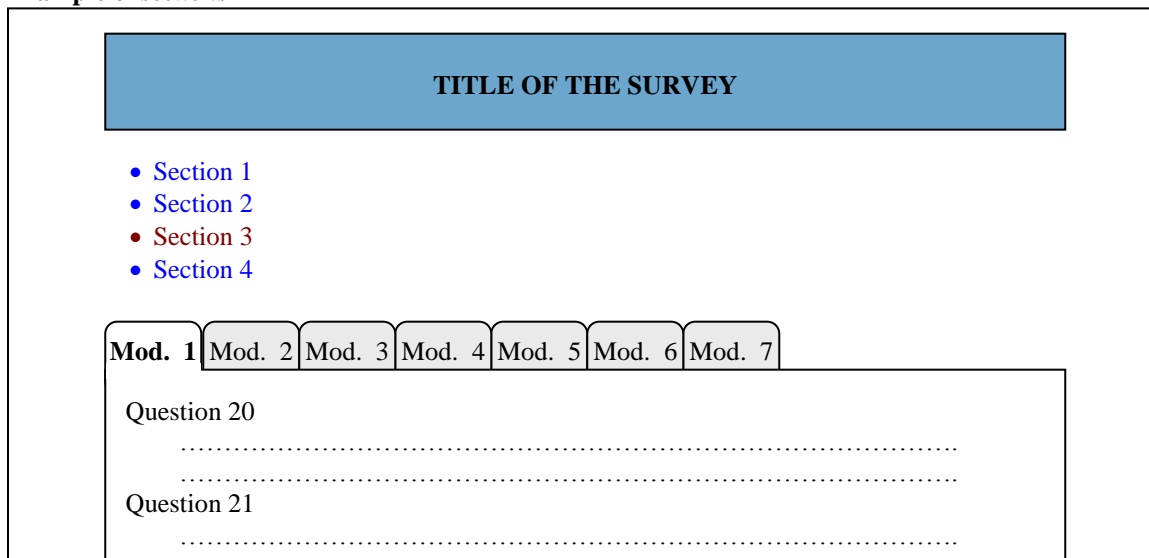
These are the possible types of Sections:

- Header /not header;
- Paginated / not paginated;
- Modular / not modular.

In the Figure 3.2-1 a full example of the three types of sections:

- The title of the survey with the sky-blue background, has been obtained through a section of type header;
- The bulleted index, in blue characters, is the access to one section per page;
- At the bottom the section three made up of seven modules. A module is a block of questions that is repeated, just like for the members of a household or the branches of a company.

Figure 3.2-1
Example of sections



3.3 Questions

A variety of formats is available for questions:

A first type is "row" question (Figure 3.3-1), *i.e.*, all the variables associated with the question shall be placed horizontally.

Figure 3.3-1
Example of row question

Q8.Text of the question.....

No Yes Nr.

A second type is “column” question (Figure 3.3-2), *i.e.*, all the variables associated with the question shall be placed vertically.

Figure 3.3-2
Example of column question

Q14.Text of the question.....

a. Males

b. Females

c. Tot

The “table” question (Figure 3.3-3) is a line of variables which is repeated a certain number of times, even not fixed *a priori*.

Figure 3.3-3
Example of table question

Q31.Text of the question.....

	Var1	Var2
1	<input type="text"/>	<input type="text"/>
2	<input type="text"/>	<input type="text"/>
3	<input type="text"/>	<input type="text"/>

The fourth type of question is the “matrix” (Figure 3.3-4) that is a way to arrange prefixed m times n variables.

Figure 3.3-4
Example of matrix question

Q4.Text of the question.....

	Set1	Set2	Set3
Case1	<input type="text"/>	<input type="text"/>	<input type="text"/>
Case2	<input type="text"/>	<input type="text"/>	<input type="text"/>
Case3	<input type="text"/>	<input type="text"/>	<input type="text"/>
Case4	<input type="text"/>	<input type="text"/>	<input type="text"/>

The questions have some options available:

- Hints: additional text, in order to explode/complete some concepts;
- Tool tip: when the cursor is over the question a message appears;
- Sources: to allow the compiler to describe the source of data for that question;
- Box notes: a square area with colored background to alert the compiler about some notes;
- The possibility of automatic numbering of questions with numbers or letters.

3.4 Variables

Variables may have the following formats:

- Check box for multiple-choice answer;
- Radio button and Drop-down menu for single-choice answer;
- Text, Numeric and Date field.

Moreover, variables can have these options:

- a field notes associated with it, to allow the user to explain his answer with a free text. If there is something written in this field, the icon changes and a pencil appears on the image of a sheet;
- opportunity to see the variable time series;
- read-only mode;
- automatic or manual numbering.

3.5 Rules

The rules associated with each variable can be of the following types:

- Enable/Disable a Field;
- Jump Sections/Questions/Variables;
- Consistency check (for example $A+B$ must be less than C);
- Calculated Variables (for example A must be equal to $B+C$);
- Filter Variable on the value of another variable (for example “provinces list” controls “cities list”).

Rules also can have some options:

- a rule can be enabled or disabled during the survey;
- a rule may be blocking or non-blocking for sending the questionnaire.

4. Monitoring

4.1 Description
















To perform an effective monitoring activity, we have several reporting forms, five of them following.

4.2 First form

In this form (Figure 4.2-1) for each respondent are shown:

- some details in order to better identify and select / filter it;
- current status of the questionnaire (*i.e.*, in process, sent, and so on);
- the icon of a cogwheel to change the working status and carry it forward or backward;
- a padlock icon to indicate whether the questionnaire is currently on compiling (closed padlock). This icon is also used to unlock the sessions have been blocked due to the abrupt closing of the browser without properly exiting from the application;
- a semaphore to indicate the level of violation of rules.

Figure 4.2-1
First monitoring form

Respondent	Territorial info	Status	Last modified	Last User	Change status	Involved users	Questionnaire	On compiling	Level of violation
Respondent1	Rome	Sent	20/07/11	User3					
Respondent2	Milan	In process	14/10/11	User8					
Respondent3	Florence	In process	03/09/11	User5					

4.3 Second form

A second form shows, for each type of user, identifying data and how to contact him. In particular are displayed:

- respondent;
- type of user (Thematic Respondent, Thematic Responsible, Intermediary Body);
- name of user;
- date of registration;
- telephone;
- email;
- any affiliation.

4.4 Third form

A third form shows the status of the survey providing the absolute value and the percentage of questionnaires in each status of processing (*unprocessed, in process, submitted, validated and completed*) and at different territorial levels.

4.5 Fourth form

A fourth form gives an indication on quality of the questionnaires, providing absolute value and percentage of fields filled in any status of processing. Of course this information has full meaning only in the case of not customized questionnaires (*i.e.*, where some of the questions are not already partially precompiled).

4.6 Fifth form

The fifth and last form provides two summary information at different territorial levels and for each theme:

- a report on how many Thematic Respondents are registered and how many are not;
- a report on how many questionnaires are *unprocessed, in process, submitted, validated and completed*.

5. Other features

5.1 Upload/Download

An useful feature for respondents (or intermediary body) who already have the required data in electronic format, is the possibility of uploading data at different levels:

- the whole Questionnaire;
- individual Sections;
- individual Questions.

Quite remarkable is that you can do the same controls activated by online compilation.

Downloading is allowed as well.

5.2 Quality analysis

Another feature is the Quality analysis of the questionnaire both “*respondent side*” and “*designing side*”:

- Quality “respondent side” means quality of answers, *i.e.*, the error count at the last save x error type (Severe, Intermediate, Slight – user defined partitioning the set of rules);
- Quality “designing side” means analysis on the quality of questions, paths, and so on and is obtained by error count over all saves.

5.3 Miscellaneous

Finally, the latest version of the application allows:

- the management of multiple languages;
- the so-called Multi-Questionnaires, *i.e.*, dynamic creation of more questionnaires for each user whenever a certain event occurs;
- management reminders by generating lists of addresses or directly by sending e-mails.

5.4 Strengths

Strengths of the system are:

- A high level of stability (four versions and several surveys);
- A wide set of features;
- Uniformity of presentation for external users and the same application interface for internal users;
- Cost reduction.

Acknowledgements

I am lucky to have met so many smart people who have worked on this project and unfortunately I can not mention them all. I would just like to thank Corrado Carmelo Abbate who allowed me to begin this work. We would not have come so far if Linda Laura Sabbadini and Saverio Gazzelloni did not strongly believe in this work. Fortunately, Silvia Montagna had the courage to experiment with a still young software, suggesting improvements and proposing it to others. A special thanks to Teresa Di Sarro for the constant effort and the great ability to think in general terms and for her accuracy in the test phase. Finally, many thanks to Angela Ciocci, who has contributed significantly to the project with her technological expertise.

Embedded experiment for non-response follow-up methods of electronic questionnaire collection

Milana Karaganis, Karla Fox, Jeannine Claveau, Joanne Leung and Wei Lin¹

Abstract

Currently, Statistics Canada is undertaking a general restructuring of its business statistics programs. One of its goals is to let electronic data collection become the principal mode of collection for business surveys. Until now, follow-up methods used for electronic questionnaire were based on strategies used for paper collection methods (fax and/or telephone reminders). In order to establish a standard collection follow-up strategy for annual business surveys using an electronic questionnaire as a main collection mode, Statistics Canada built an experimental design to compare different non-response follow-up methods. This paper summarizes initial results of this experiment.

Key Words: Collection; Non-response; Follow-up; Paradata; Experimental design.

1. Introduction

At Statistics Canada, business survey collection consists of many steps and uses more than one collection mode. Many business surveys continue to use paper questionnaires for data collection. Recent developments in Internet technologies have greatly impacted survey data collection as use of electronic questionnaires (EQ) for data collection have exploded over the last ten years. The EQ surveys could take on a variety of forms from simple email surveys to sophisticated web survey systems. Statistics Canada is currently undertaking a general restructuring of its business statistics programs. One of the goals is to let electronic data collection become the principal mode of collection for business surveys.

Until now, follow-up methods used for EQ surveys at Statistics Canada were based on methods used for paper collection, which is a combination of fax and telephone attempts. International and Statistics Canada experiences showed that electronic respondents require different follow-up patterns. In order to establish a standard collection follow-up strategy for annual business surveys using an electronic questionnaire as a main collection mode, Statistics Canada built an experimental design to compare different non-response follow-up methods (NRFU), combining telephone and email reminders at different time points throughout the collection period. Seven surveys belonging to the Unified Enterprise Statistics (UES) program in the 2011 collection cycle were used for this experiment. An embedded balanced factorial design was used for this experiment. By doing an embedded experiment, we aimed at finding a follow-up strategy that produces the best response rates and is the most efficient in terms of cost. We would also like to know how far we can get with sending email reminders only, and to see if there is any importance of having the first follow-up attempt via telephone versus email.

This paper summarizes initial results of this experiment. Section 2 gives an overview of UES collection. Section 3 gives the methodology of the experimental design. Section 4 presents the results of the experiment, including results of analysis of variance tests. Section 5 summarizes conclusions and section 6 gives recommendations for future EQ surveys.

¹Milana Karaganis, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (milana.karaganis@statcan.gc.ca); Karla Fox, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (karla.fox@statcan.gc.ca); Jeannine Claveau, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (jeannine.claveau@statcan.gc.ca); Joanne Leung, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (joanne.leung@statcan.gc.ca); Wei Lin, University of Toronto, Ontario, Canada (wei.lin@utoronto.ca).

2. Overview of UES Collection

The UES program consists of close to 60 annual business surveys which are integrated in terms of content, collection and data processing. In 2011, web-based electronic questionnaires (EQ) were built and used for collection of seven UES surveys. Before 2011, only one survey in the UES program was using electronic questionnaire as a collection mode.

The UES collection process is done in two steps. The first phase of the annual collection cycle for UES surveys consisted of a telephone pre-contact made for new enterprises selected in the sample to confirm their contact information as well as the activity codes based on the North American Industry Classification System (NAICS). In 2011, for seven surveys with EQ, telephone pre-contact was conducted not only to confirm existing information, but also to obtain email addresses of the respondents. Respondents were advised that collection would be done via the electronic questionnaires and were asked to supply email addresses. The only respondents assigned to paper questionnaires were the ones that adamantly refused EQ or the ones that we did not reach during pre-contact. For units that previously experienced EQ collection, a heads-up email was sent instead to advise that their responses would once again be collected through EQ. Overall, the 2011 sample for the seven UES surveys with EQ collection mode was 9,324 units and 6,457 units (approximately 70%) were assigned to the EQ collection mode after completion of the pre-contact phase.

The next phase of the collection consisted of mailing out either paper questionnaires or email invitations to the sampled units. The final phase of collection for the UES surveys consisted of receiving completed questionnaires, imaging paper questionnaires, uploading both imaged and electronic questionnaires into the central collection system (Blaise) for edit verifications, follow up with outstanding respondents (NRFU) and respondents whose questionnaire failed editing in Blaise (FEFU), finalization of the cases in Blaise and output to the subject-matter division for further processing.

For the seven surveys using EQ in 2011 collection, follow-up on the units responding via EQ who did not submit completed questionnaires was subject to the experiment. Units responding via a paper questionnaire followed the standard NRFU strategy (telephone follow-up and/or fax reminders). For all telephone NRFU regardless of the collection mode, a maximum of five attempts was allowed before a unit was considered as a final non-response unit.

3. Methodology of the Experimental Design

3.1 Embedded Design

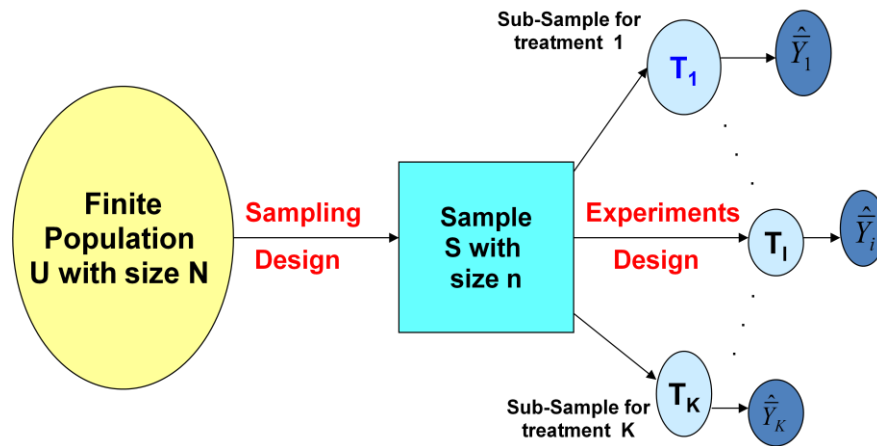
Researchers agree that it is important for survey methodologists to study how different survey methodologies and implementation strategies affect non-response, quality and efficiency (Van den Brackel and Renssen (2005), Jackle et al. (2010), Groves (2010)). Randomized controlled trials are the most common way of attempting to determine if a cause-effect relation exists between a specific intervention and an outcome. Other study designs, can detect associations between an intervention and an outcome. But they cannot rule out the possibility that the association was caused by a third factor linked to both the intervention and the outcome. Although randomized trials are powerful tools, their use is often limited by ethical and practical concerns. In the case of the UES it would not be feasible, due to the cost, to set up an experiment separate from the survey to study different non-response follow-up strategies. However, recent research literature demonstrates that experiments embedded in ongoing sample surveys are particularly appropriate to investigate the effects of alternative survey methodologies on response behaviour or estimates of finite target population (Van den Brackel and Renssen (2005), Van den Brackel and Berkel (2002)). Even if it turns out that the alternative approach has significant effects, we still could use the data collected from the current survey approach for publication (*i.e.*, estimation) and use all the data for testing treatment differences (*i.e.*, inference).

An experiment within a survey can be thought of as a variation of a two-phase survey design as illustrated in Figure 3.1-1 below. As such, to test if the treatments are significantly different we need to use methods that take into account both the experiment phase and the sampling phase (Van den Brackel and Renssen, (1998)).

For our experiment, as there was no previous estimate of effect size, we used an embedded balanced factorial design. Treatments were assigned in a balanced fashion randomly within survey strata. This randomization was done at the time of the survey design. This was done for operational reasons as randomization of non-responding units was not feasible in the current Blaise system framework. Randomization at this time point does allow for comparisons between treatments, but we need to study the effect of differential response from time of randomization on the treatment (Jackle *et al.* (2010)).

The embedded experiment was to be conducted on live surveys, with the collected data to be used to produce usual estimates that have been typically produced every cycle. This placed a significant constraint on the experiment. We had to ensure that the end collection results would not be affected. In other words, we could not jeopardize the ultimate goal of producing estimates from this collection. Therefore, the experiment was designed to take place during the first four months of collection. Once the experiment was completed, a final blitz of all outstanding units was conducted to improve response rates.

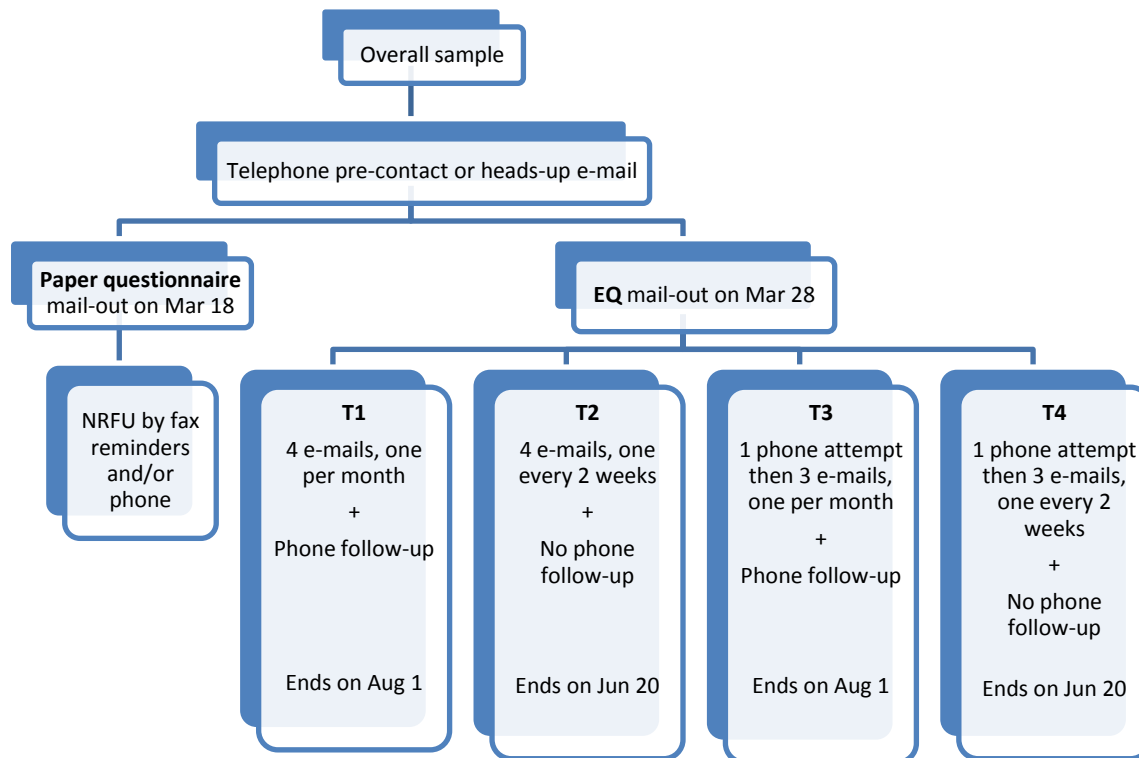
Figure 3.1-1: Illustration of an experiment embedded in a sample survey



3.2 Treatments Assignment

The set up of the experiment is illustrated in Figure 3.2-1. Note that the split between units receiving paper questionnaire and electronic questionnaire was not performed by randomization. Instead, it is determined by the respondents. Therefore, we cannot make comparison on the results between paper and EQ collection. In fact, paper collection was composed of businesses that were not contacted during pre-contact and those who refused to use EQ. Thus, these businesses are more prone to be a non-response and are suspected to be qualitatively different from the units who went with the EQ collection mode.

Figure 3.2-1: The set up of the experiment



Units with an EQ collection mode were randomly assigned to one of the four treatments, where each treatment contained approximately 25% of the EQ sample. The randomization was done within each combination of survey, strata group and type of questionnaire. The three strata groups are: (1) small take-some stratum, (2) large take-some stratum and (3) take-all and must-take stratum. The type of questionnaire was identified by its length, *i.e.*, long form or short form. It has to be noted that stratification was applied to the six surveys while the Head Office survey was a census. As a result, we ended up with a randomized block design for service industries (surveys and strata used as blocks) and a completely randomized design for the Head Office survey. Thus, results had to be analyzed separately for service surveys and the Head Office survey to account for differences in the design.

On March 28, 2011, each unit in Treatments 1 to 4 (T1 to T4) received an email invitation that contained a hyperlink and an access code for completing the survey online. Non-response follow-up started approximately one month into the collection, on April 26, 2011. On that date, all outstanding units were sent to NRFU by a specified approach indicated for T1 to T4. Each treatment had a different NRFU approach using a combination of email reminders and phone attempts. NRFU stopped for a unit as soon as the questionnaire was returned to Statistics Canada.

The four treatments were designed as follows. Treatment 1 ('standard') was designed to mirror the usual NRFU strategies for paper questionnaires. In other words, we allowed phone follow-up throughout the entire collection and once a month, (April 26, May 26, June 23 and July 25) we emailed a reminder to complete outstanding questionnaires (thus, substituting fax reminders with email reminders). Treatment 1 was to be used as a baseline for comparison with other strategies. This treatment was scheduled to finish on August 1, to allow for the final blitz by interviewers.

Treatment 2 was designed to test how far we could get with email reminders only, *i.e.*, what kind of response rates we could achieve without any phone follow-ups. We did not allow telephone follow-up, but interviewers were allowed to answer phone calls from respondents and make firm appointments to collect data, if insisted by respondents. Because there was no phone follow-up at all, we decided to send email reminders more often, once

every two weeks (April 26, May 9, May 26 and June 7). Thus, this treatment was scheduled to close on June 20, to allow for the same number of email reminders as the other treatments.

Treatment 3 was designed to measure the impact of doing the first follow-up attempt via phone rather than by email. In this case, the first follow up attempt was conducted over the phone and the rest of follow-up reminders were done via email, once per month (April 26-29 (phone attempt), May 26 (email), June 23 (email) and July 25 (email)). Phone follow up calls were conducted similar to Treatment 1. For Treatment 3, experiment ended on August 1.

Treatment 4 was designed to combine both having first follow-up attempt done over the phone and then switching to email reminders only (April 26-29 (phone attempt), May 9 (email), May 26 (email) and June 7 (email)). As such, we did not allow telephone follow-up, but interviewers were allowed to answer phone calls from respondents and make firm appointments to collect data. For Treatment 4, experiment ended on June 20, again, similar to Treatment 2 as we chose to compress the follow-up schedule with email reminders every second week.

At the end of the experiment (June 20 or August 1), all outstanding units were sent to telephone NRFU. Although for Treatments 2 and 4 the end of the experiment was on June 20, telephone NRFU just began on July 8 for these treatments. Note also that these two treatments received another email reminder on July 7. After August 7, experiment was completed for all treatments and telephone NRFU blitz for all four treatments was in operation. Active collection continued until October 14, 2011 where all follow-up actions stopped. This marked the end of the collection cycle for 2011.

4. Results

4.1 Descriptive Statistics

There were a total of 9,324 units sampled for these seven surveys. Among these units, 6,457 units were assigned to the EQ collection and were then divided randomly into four treatments of approximately the same size. The four treatments had 1,615, 1,613, 1,615 and 1,614 units respectively.

Through follow-up of non-response, respondents could request a switch of a collection mode. Thus during collection, we observed 338 units switching from paper to EQ and 521 units switching from EQ to paper collection. Also, 1,098 turned out to be out of scope of the survey or out of business, as confirmed during collection. At the end of collection, the numbers of in scope units in the four treatments were 1,375, 1,350, 1,394 and 1,376 respectively.

4.2 Return Rate

Return rate, which indicates the percentage of questionnaires completed and returned, is often used as a key measure of survey progress. It is determined when the questionnaire is submitted to Statistics Canada or if the unit is declared a respondent via another collection mode (CATI, fax, ...). Business surveys have highly skewed populations, meaning a relatively small number of units can account for a large portion of the economic activity. Therefore, return rates ought to be calculated both on a weighted and an unweighted basis. The unweighted return rate indicates the percentage of questionnaires received or completed, among all in scope units, whereas the weighted return rate is a percentage of the revenue contribution of the received or completed units, among the overall revenue contribution of the in scope units.

The progression of the total weighted return rate for each treatment from the EQ Mail-out (March 28) to the end of collection (October 14) is shown in Graph 4.2-1. The return rates were computed based on all in scope units at the end of collection. All the events happened on the key dates are indicated by the dotted lines labelled from "E1" to "E8".

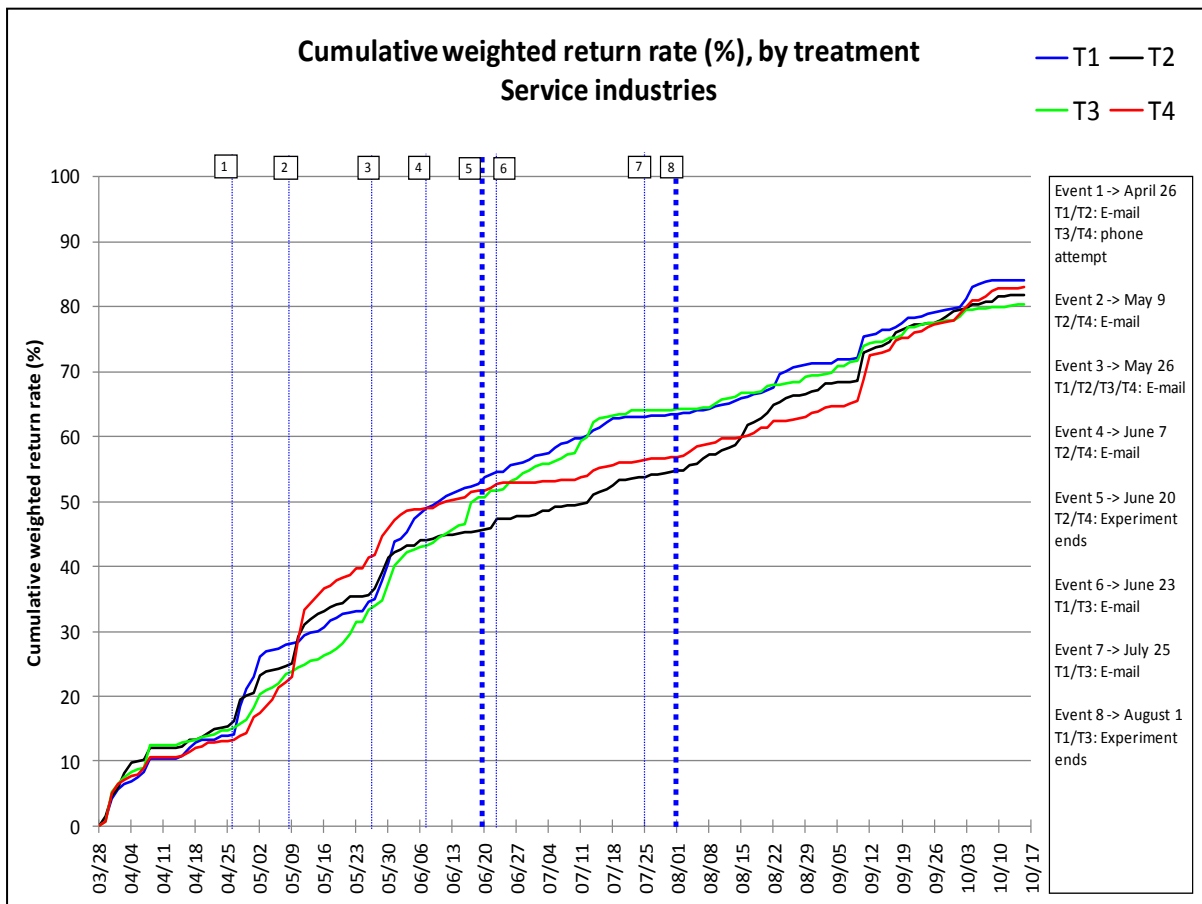
At the beginning of the non-response follow-up (April 26, E1), the weighted return rate of the four treatments were similar (14% to 16%). On April 26, Treatments 1 and 2 received an email reminder and Treatments 3 and 4 received a phone attempt in that week. Two weeks later (May 9, E2), all the treatments and paper collection had similar weighted return rates (23% to 25%). On May 9, Treatments 2 and 4 received an email reminder. It appears that this

helped and after May 9, Treatment 2 and especially Treatment 4 were in the lead, in terms of return rate, over the other two treatments. The advance was kept until the end of May for Treatment 2 and until mid-June for Treatment 4. It seems that having a reminder action every two weeks instead of every month gave Treatments 2 and 4 the opportunity to advance at the beginning. On May 26 (E3), all four treatments received an email reminder. On June 7 (E4), Treatment 2 received its fourth email reminder and Treatment 4 its third one. However, this fourth email reminder for Treatment 2 did not seem to have a similar positive impact as the three previous ones.

By June 20 (E5, the end of the experiment for Treatments 2 and 4), return rates were around the 50% mark for all four treatments. We were able to get more than 40% of questionnaires returned by just sending email reminders and around 50% by just doing one phone attempt plus email reminders. It is even more of interest to note that Treatment 4, where only one phone attempt was made between April 26 and April 29, attained almost the same results as Treatment 1 (standard treatment) and as Treatment 3, where telephone follow-up as per score function was applied continuously since April 26.

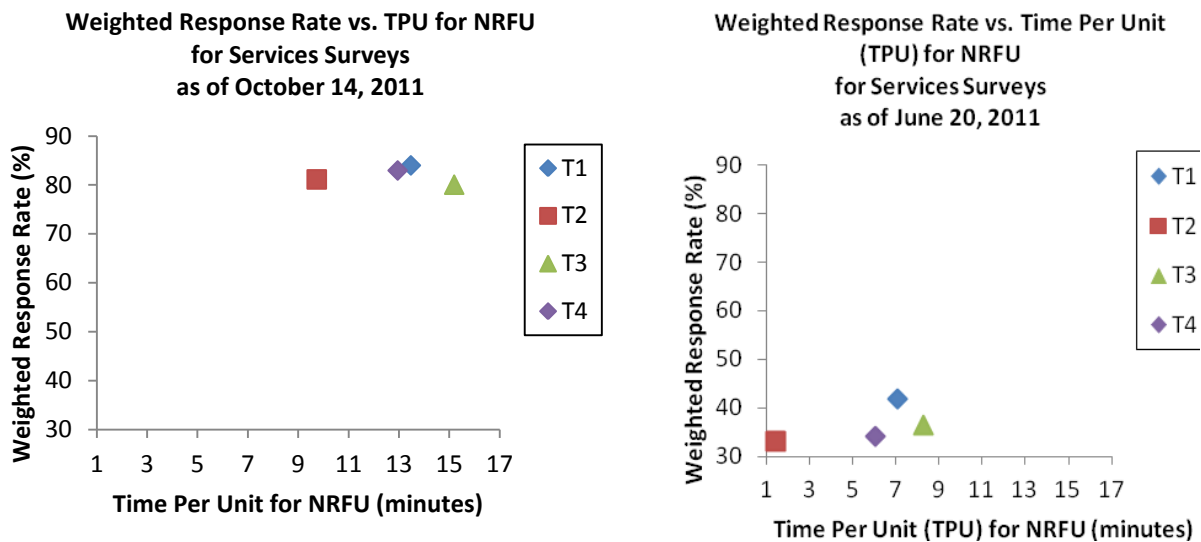
After June, Treatments 1 and 3 began to overtake Treatments 2 and 4. On August 1 (E8), these treatments exceeded Treatments 2 and 4 by more than 10%. Since regular telephone follow-up for Treatments 2 and 4 began after the end of the experiment (in fact on July 8 due to operational issues), the impact of the telephone follow-up was seen only at the end of summer. By the end of September, the four NRFU treatments show very similar return rates. This steady pattern continued until the end of collection with all treatments finishing with over 80% return rate.

Graph 4.2-1: Cumulative weighted return rate observed on each day during collection for Service Industries



Next, we looked at collection costs and how they compare across different treatments. The start and end times for each phone attempt are recorded in the Blaise System. We are then able to compute the duration of each attempt, also known as Time Per Unit (TPU). The scatter plots of the weighted response rate versus the TPU for non-response follow-up on June 20 and October 14 are shown in Graph 5. Treatment 2 (red squares) had similar response rates as the other three treatments, but with a lower TPU spent on NRFU. Since it is more costly to spend time on telephone follow-up than to send email reminders (which is an automated process), a treatment was considered more costly if the TPU on NRFU is higher. Therefore, from the scatter plots, we observed that Treatment 2 was able to give similar response rates with lower cost on telephone follow-up, compared to Treatments 1, 3 and 4.

Graph 4.2-2: Scatter plots of weighted response rate versus the average duration of Time per Unit (TPU) for NRFU (on June 20 and on October 14, 2011)



4.3 Analysis of Variance

Analysis of variance (ANOVA) was performed on weighted and unweighted return rates, number of days elapsed to receive questionnaire, the number of NRFU attempts and time spent per unit (TPU) on NRFU.

Recall that the Head Office Survey is a census where each unit has the same sampling weight, while for Services surveys a stratified sample is selected. Because of the difference in the sampling designs, the analysis of variance tests were run separately for Head Office and for Services surveys. For tests on Head Office units, the F-Test in the SAS procedure PROC GLM was used. For Services surveys, the Wald Test adjusted for design (Van den Brakel and Renssen (2005)) was used.

In Table 4.3-1 below, the results of ANOVA tests showed that on both June 20 and October 14 (end of collection), the unweighted and weighted return rates were not significantly different among the four treatments for all surveys. The NRFU number of attempts and NRFU TPU of the four treatments were significantly different. The number of days elapsed to receive questionnaire was significantly different among the four treatments for the Head Office Survey as of June 20, but is not significantly different on October 14. It was not significantly different for Services surveys on both June 20 and October 14.

Table 4.3-1: Testing treatment differences on June 20 and on October 14, 2011

Main Effects: T1 = T2 = T3 = T4	Head Office		Services	
	Jun 20 p-value	Oct 14 p-value	Jun 20 p-value	Oct 14 p-value
Weighted return rate	—	—	0.9996	0.9999
Unweighted return rate	0.0621	0.1590	0.6512	0.9999
Number of days elapsed to receive questionnaire	<0.0001	0.7637	0.0978	0.5231
NRFU attempts	<0.0001	0.0256	<0.0001	0.0066
NRFU TPU	<0.0001	0.0363	<0.0001	0.0104

5. Conclusion

We successfully implemented an embedded experiment to test four strategies for the non-response follow-up of electronic questionnaires for seven UES surveys in the 2011 collection cycle. The strategy consisting of email reminders every two weeks at the beginning of collection succeeded in obtaining the first 40% of response at lower cost. Results showed that even if telephone follow-up starts three months later than in the current follow-up strategy, very similar final response rates can be achieved, with fewer phone attempts and less effort.

The initial tests run to compare implemented treatments seemed to support these conclusions obtained from the descriptive analysis. However, more detailed analysis is yet to be performed to finalize the overall conclusions from this experiment. In particular, we would like to look at the implications of using different modes for the collection and assess if there is any impact on the estimates.

6. Acknowledgement

This experiment was made possible due to efforts of many areas within Statistics Canada. Authors would like to acknowledge all the hard work and contribution by Service Industries Division, Enterprise Statistics Division, Business Survey Methods Division, Social Survey Methods Division, Collection Systems and Infrastructure Division, Operations and Integration Division, Collection Planning and Management Division, and Sturgeon Falls Regional Office that made this study possible.

References

- Groves, R.M. and L. Lyberg (2010), “Total survey error past, present, and future”, *Public Opinion Quarterly*, 74(5), 849-879.
- Jackle, A., Roberts, C. and P. Lynn (2010), “Assessing the effect of data collection mode on measurement”, *International Statistical Review*, 78(1), 3-20.
- Van den Brakel, J.A. and R.H. Renssen (1998), “Design and Analysis of Experiments Embedded in Sample Surveys”, *Journal of Official Statistics*, 14(3), 277-295.
- Van den Brakel, J.A. and R.H. Renssen (2005), “Analysis of experiments embedded in complex sampling designs”, *Survey Methodology*, 31(1), 23-40.
- Van den Brakel, J.A. and C.A.M. Van Berkel (2002), “A design-based analysis procedure for two-treatment experiments embedded in sample surveys. An application in the Dutch labor force survey”, *Journal of Official Statistics*, 18(2), 217-231.

SESSION 6B

OUTLIERS AND IMPUTATION

Retrofitting a simpler outlier detection procedure into a complex generalized system

Laura T Bechtel¹

Abstract

Distributions of economic data tend to be highly skewed, making outlier detection difficult because some values flagged as outliers are legitimate. Often, outlier detection procedures applied to economic data employ ratio comparisons. Economic Programs at the U.S. Census Bureau commonly employ the Hidiroglou-Berthelot (HB) statistical edit. For a long time, the HB edit was the only outlier detection method available in the Standard Economic Processing System (StEPS). However, the HB edit has two limitations: input data must be positive, and the HB edit cannot be used to detect single-item outliers. These two limitations were prohibitive for the Quarterly Financial Report (QFR), which collects income (which can be negative) as one of its key items. Consequently, the StEPS outlier detection module was enhanced by adding resistant fences, which has excellent demonstrated results for other U.S. Census Bureau applications. In this paper, we present how we retrofitted resistant fences into the existing outlier detection software, describing both the challenges encountered during the process and their resolution.

1. Introduction

The Economic Programs at the U.S. Census Bureau collect and publish data for over 100 business surveys. Many of these surveys perform outlier detection using the statistical ratio edit proposed by Hidiroglou and Berthelot (HB) edit in 1986. The HB edit is popular among business surveys because it accounts for the size of the unit when identifying ratio outliers. Consequently, for over 10 years, the Standard Economic Processing System (StEPS) only had an HB edit module instead of an outlier detection module. While this was useful for most of the surveys that use StEPS, some surveys do not have data that lends itself to using the HB edit.

In this paper, we discuss the procedures taken to add the resistant fences outlier detection method to StEPS. We first give a brief overview of StEPS and associated terminology in section 2. In section 3, we present the HB edit and resistant fences outlier detection methodology. Key characteristics of StEPS implementation of HB edit are explained in section 4. Section 5 outlines the stages we employed to retrofit resistant fences into the existing StEPS outlier detection module. We conclude in section 6 by reflecting on lessons learned and listing the benefits of the retrofitting process.

2. StEPS Background

StEPS is generalized software used for implementing many different economic surveys. The modules that make up StEPS include data collection, data editing, outlier detection, imputation, and estimation.

As with any software, StEPS has some software-specific terminology. The term ‘item’ is used to describe a variable containing numeric data stored for publication or data processing, usually on the survey instrument. For example, if a survey collected the value of sales, then the variable sales would be an item. Additionally, a reporting or tabulation unit is referred to as an ‘ID.’ When discussing survey data processed in StEPS, ‘ID-item level’ is used to mean the

¹Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau, Washington, DC 20233 (Laura.Becht@ census.gov). This report is released to inform interested parties of research and to encourage discussion. Any views expressed on methodological or operational issues are those of the author and not necessarily those of the U.S. Census Bureau. I thank Anne Russell, James Hunt, Xijian Liu, and Katherine Thompson for their careful review and thoughtful comments on earlier versions of this manuscript, and the members of the SMAG for their comments and suggestions on the presentation.

specific variable being reported for a specific reporting or tabulation unit. For instance, the outlier detection module allows a user to define specifications for an outlier edit at the ID-item level. This means that each ID will be subjected to the outlier edit test for the item specified in that test.

In addition to having its own terminology, StEPS also has its own change control process. In order to make a change to the system, a change request (CR) must be submitted to the StEPS Change Control Board (CCB). The membership of the CCB is comprised so that all StEPS users are represented, and CRs must be approved by the CCB for implementation. If the CR is approved, there are three stages that it must go through before the change is migrated to StEPS: 1) gathering and documenting requirements; 2) aligning requirements with the existing software; and 3) testing and implementation. When the CR involves methodological issues such as outlier detection, the CR is overseen by the StEPS Methodology Advisory Group (SMAG), a permanent committee comprised of methodologists representing the StEPS user areas.

3. Outlier Detection Methodology

3.1 HB Edit

The HB edit is a selective, ratio editing procedure that flags less data as ‘suspicious’ than traditional ratio tests. The ratios are of the form, $R_i = x_i/y_i$ where x_i and y_i are variables reported for ID i that are positively correlated. Typically, y is a previously reported value of x , but can also be a different item reported in the same statistical period.

The HB edit is used to generate tolerances that identify the ratios as outlying or not. Before the tolerances are developed, there are several transformations to the data. The first transformation is the centering transformation: where the transformed observations (S_i) are centered around R_m , the median ratio. The centered observations are then subjected to the size importance transformation, $E_i = S_i \times \{\max(x_i, y_i)\}^u$ where E_i is the HB statistic and u is the size parameter that ranges from [0, 1], with the default set to 0.5.

With the HB statistics calculated, the next step is to generate tolerances (upper and lower bounds). Generally, this is done using quartiles and their differences. However, the HB edit uses a second term to avoid zero-valued differences between the quartiles:

- $D_{q1} = \max\{(E_m - E_{q1}), |A \times E_m|\}$ where E_m is the median HB statistic, E_{q1} is the first quartile of the HB statistics and A is a multiplier (usually defaults to 0.05) used when $E_m - E_{q1}$ is near zero.
- $D_{q3} = \max\{(E_{q3} - E_m), |A \times E_m|\}$ where E_{q3} is the third quartile of the HB statistics.

Once D_{q1} and D_{q3} are calculated, observations with HB statistics that fall outside of $[E_m - c \times D_{q1}, E_m + c \times D_{q3}]$ are flagged as outliers, where c is a predetermined constant.

For a survey to use the HB edit, its data must satisfy some underlying assumptions. The two key assumptions are that compared data items are positively correlated, and studied data items must be **strictly non-negative**. These conditions are satisfied by many, but not all, surveys in StEPS. For example, the Quarterly Financial Report (QFR) cannot apply the HB edit to their income variable because income can be negative. A macro-review of the income variable is, however, necessary prior to publication. Therefore, it was proposed that resistant fences be added to StEPS for surveys like QFR.

3.2 Resistant Fences

When compared to HB edit, resistant fences is a much simpler form of outlier detection because no transformations of the data are necessary. In fact, HB edit is a special form of resistant fences. Resistant fences methods use tolerances that are computed from the distribution’s quartiles to flag outlying observations. Within each specified analysis cell, a distribution of analysis variables is specified as either a single item, a ratio of an item to its previously reported value, or a ratio of an item to another highly correlated item from the same collection period. After defining the analysis variable, the method obtains the following statistics: q_1 , the first quartile; m , the median; q_3 , the third quartile; H , the interquartile range ($q_3 - q_1$).

These statistics are then used to generate bounds that will be used for flagging outliers in the micro data. Note that these quartiles can use either weighted or unweighted data, depending on user preference. There are two general methods of calculating these bounds (Thompson, 1999):

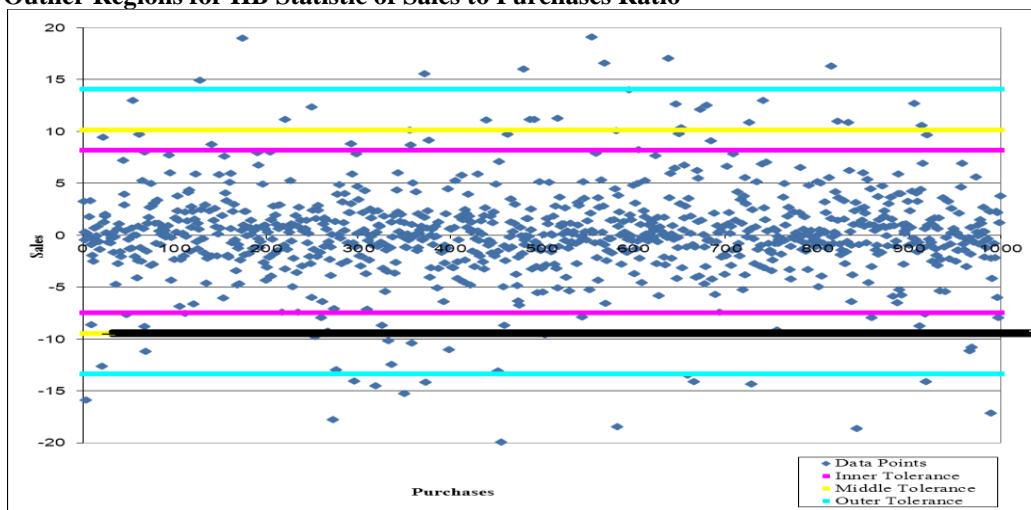
- **Traditional Resistant Fences:** Flag if the analysis variable is less than $q_1 - k \times H$ or greater than $q_3 + k \times H$
- **Asymmetric Resistant Fences:** Flag as an outlier if the analysis variable is less than $q_1 - k \times (m - q_1)$ or greater than $q_3 - k \times (q_3 - m)$

where k is a pre-specific constant that determines the width of the ‘fences.’ The parameter k is analogous to the c value used to determine the HB edit tolerances

4. HB Edit in StEPS

The HB edit in StEPS can be defined at the ID-item level; each HB edit test is specified to flag one particular item. StEPS provides the user with the flexibility to specify up to three outlier regions for each test by allowing the user to provide up to three c values. Below, Figure 4.1 demonstrates how these three different regions (tolerances) work in StEPS

Figure 4.1
Outlier Regions for HB Statistic of Sales to Purchases Ratio



The values of the observations that fall between the inner tolerances (the pink lines) in Figure 4.1 are considered legitimate values according to the HB edit test and are not flagged as outliers. The observation values that fall between the inner (pink) and middle (yellow) tolerances are values that vary only slightly from the distribution and are flagged for review by an analyst. The values between the middle (yellow) and outer (blue) tolerances are a bit more suspicious and are flagged to be suppressed from imputation base in addition to being flagged for review. Finally, the values lying outside the outer tolerances (blue) are so egregious that they are flagged for imputation in addition to being flagged for suppression and review.

The following set of parameters can be specified for the HB edit at the ID-item level: the numerator, the denominator, three different c values, the u value, the A value, and the class variable being used to define the outlier cell.

There are several points that need to be made about the HB edit parameters. HB edit parameters are specified at the ID-item level, which means that the user cannot specify different HB edit parameters for each outlier cell. Only the numerator item is flagged as an outlier. One item can be used repeatedly as a numerator; one item can have several different HB edit tests associated with it. However, only one final outlier flag is assigned at the ID-item level. The StEPS program is designed for outlier cells to be defined by at most one classification variable, so to define an

outlier cell using more than one class variable, the user must create a recoded variable that combines all the desired variables into one class variable.

5. The Retrofitting Procedure

5.1 The Change Request

As outlined in section 2, a CR needs to be submitted to the StEPS CCB for approval before a change can be implemented. Originally, the resistant fences CR was submitted from the requesting survey for a very specific version of resistant fences: asymmetric fences. However, the CCB immediately asked the SMAG to review the CR. The SMAG determined that the CR needed to be generalized so both versions of resistant fences would be implemented. In cooperation with the original CR requestor, the SMAG resubmitted a more generalized CR, which the CCB approved.

5.2 Gathering Requirements

Before we began gathering the requirements, some precedents set forth by the way HB edit was implemented in StEPS needed to be met:

- The parameters needed to be specified at the ID-item level.
- Up to three sets of tolerances could be specified for each outlier test.
- The outlier cell level for the outlier test could be specified.
- Only one final outlier flag would be assigned at the ID-item level.
- The (positive) ratio tests should be analogous to HB edit.

In addition to the precedents set forth by HB edit, the SMAG needed to ensure that the requirements allowed the requesting survey (QFR) to implement their outlier detection procedures. Simultaneously, the requirements needed to be general enough for other program areas. Clearly, the SMAG had many perspectives to consider. Not surprisingly, many sticking points were encountered, resulting in much discussion and very little resolution.

Finally, after feeling like we were spinning our wheels, the SMAG looked at methods employed in the past to help resolve disagreements. An efficient but fair solution was an open issues document. It was simple – the unresolved requirements were listed and each division representative proposed a resolution. If we could not come to a unanimous decision, we let the majority rule, giving a little more weight to the requesting division. This helped the requirements gathering process gain a lot of momentum and the requirements were completed within a couple of meetings.

One question arose that could not be resolved using an open issues document: What should the final outlier detection flag be if an item is subject to more than one outlier detection test and the resulting flags do not agree? Below, Table 5.2.1 illustrates this phenomenon. The ID and item are listed in first and second columns, respectively. The remaining three columns present the resulting flags from three different outlier detection tests for the ID-item combinations presented in the first two columns. As previously mentioned, there is only one final outlier flag at the ID-item level. So, the question is for IDs 0001-0003 is: what is the final outlier flag for quarterly revenue (QREV)? For ID 0001, it is a simple answer – flag it for imputation. For IDs 0002 and 0003, the answer is not as simple and a set of rules for assigning the final outlier flag is needed.

Table 5.2.1.
Outlier Detection Flags for One Item

ID	Item	Resistant Fences Test 1	Resistant Fences Test 2	HB Edit Test 1
0001	QREV	Impute (I)	Impute (I)	Impute (I)
0002	QREV		Refer to Analyst (R)	Suppress (S)
0003	QREV	Suppress (S)		Impute (I)

When HB edit was designed, no choice was given; if more than one HB edit test was run on the same ID-item combination and the tests resulted in different outlier flags, the ‘most’ outlying flag was selected as the final outlier flag. The program area that requested resistant fences wanted to do exactly the opposite and select the ‘least’ outlying flag. In more technical terms, the users wanted to address this problem by either choosing to minimize Type I or Type II error. Type I error is the probability of flagging a legitimate value as an outlier and Type II error is the probability of not flagging a legitimate outlier. Looking at Table 5.2.2, the last two columns present how IDs 0002 and 0003 would be flagged when minimizing either Type I or Type II error, respectively.

Table 5.2.2
Outlier Detection Flags Minimizing Type I or Type II Error

ID	Resistant Fences Test 1	Resistant Fences Test 2	HB Edit Test 1	Minimize Type I Error	Minimize Type II Error
0001	I	I	I	I	I
0002		R	S		S
0003	S		I		I

The resulting requirement was to give the user the choice of minimizing Type I or Type II error for each item being subjected to outlier detection, with the default being to minimize Type II error.

5.3 Design and Implementation

While requirements seemed to be very challenging for us methodologists to gather, the programmers had an equally, if not more difficult, job. First, they had to dissect the existing code for HB edit and then figure out how they were going to program our requirements into the module. Communication and compromise were essential for this phase of the retrofitting process.

The programmers walked us through the design once it was complete. For the most part, the design matched with our requirements. There were a few requirements that programmers had not incorporated into the design. For example, we had specified that the resistant fences module should allow the user to specify up to six variables to define an outlier cell. The programmer came back and told us that was not feasible due to limitations of the existing HB edit code. Specifically, they would have to rewrite HB edit in order to accommodate this request, and that was too risky. Instead, like HB edit, resistant fences would only have the capability to specify one variable. While we would have liked the flexibility to specify six class variables instead of one, we did not push the issue because 1) we did not want to compromise the functionality of the existing HB edit module and 2) we could use six variables as long as we combined them into one variable. There were more necessary requirements that could not be compromised.

Sometimes it was not compromising with the programmers, but compromising among SMAG members. During the design walkthrough, SMAG members had trouble agreeing on issues like how things should appear on the screen and naming conventions for the new variables. Finally, after much negotiation, the design was complete and ready for testing.

5.4 Testing and Implementation

It was essential that testing of this new StEPS functionality be thorough and complete before being officially released to the production environment. However, because it was new functionality, many users did not know how to test it. As a result, the SMAG developed a generalized testing plan that told the users exactly what parameters they should specify and what output they should expect. Select SMAG members also developed and implemented training for testing. Once the testers were trained, they used the test plan to complete testing. Of course, some small problems were found and corrected. Finally, resistant fences was implemented in StEPS.

6. Conclusion

In the end, it took just about one year for the entire retrofitting process. A great deal of the time was spent gathering requirements and designing the software. This may not have been the longest retrofitting process ever encountered, but it most certainly could have been improved upon. We have a few lessons learned that might make similar endeavors more efficient.

First, when retrofitting any software module, it is important that current users of the software are involved. These users should be very familiar with the pros and cons of the existing software. In our case, our software experts provided invaluable input. Second, choose your battles wisely. This seems obvious, but can be quickly forgotten when trying to develop the “best possible” software. As you go through the requirements process, there are inevitable disagreements; it is important to differentiate the “must haves” from the “nice to haves.” Third, develop a specialized group to gather requirements. This will help keep focus and momentum and prevent the group from becoming sidetracked on other issues. Fourth, use an open issues document to track and resolve issues. Having the open issues and the proposed solutions in writing lessened the discussion and sped up the decision making process. Finally, establish separate groups for functional (how you implement it) and nonfunctional (methodology) requirements. Having methodologists deciding how a program should be implemented (*e.g.*, how the screens should look) could hinder the process of making a decision if they are not the end users.

Pleasantly, there are some unexpected benefits from going through the retrofitting process. The new functionality for resistant fences was implemented in StEPS in such a way that would lend itself to improving the existing HB edit functionality. During the requirements gathering process, known shortcomings of the way HB edit was implemented were improved upon in the implementation of resistant fences. By doing this, our goal was to further justify making these improvements to HB edit so it would be consistent with its outlier detection counterpart. Finally, the entire retrofitting process facilitated discussion and understanding about the different philosophies for implementing the same (or similar) outlier detection methods. In the end, not only did we provide our users with a more flexible outlier detection system, we provided them with a broader insight into existing methodologies and philosophies.

References

- Hidioglou, M.A. and Berthelot, J.-M. (1986), “Statistical Editing and Imputation for Periodic Business Surveys”, *Survey Methodology*, 12, 73-83.
- Thompson, K.J. (1999), Ratio Edit Tolerance Development Using Variations of Exploratory Data Analysis (EDA) Resistant Fences Methods. Statistical Policy Working Paper 29, available from the Federal Committee on Statistical Methodology (<http://www.fcsm.gov/99papers/thompson.pdf>).

Outlier detection tool at Statistics Canada

Nelson Émond¹

Abstract

Many outlier detection methods are available, and the subject is amply covered in the literature. The final choice of a method will be based on the experience of the subject matter expert and the expertise of the methodologist, since they will take the structure of the data into account. However, survey managers facing time constraints and limited resources and budgets are forced to consider a number of alternatives, and they often opt for an existing method that is more accessible instead of choosing one that might be more suitable.

To resolve this problem, a tool has been developed in SAS. It combines the methods most commonly used at Statistics Canada. This makes it possible to preserve expertise for the benefit of other surveys with the same profile. The originality of this tool is that it gives an overview of the data, enabling users to view the data using interactive graphs in order to compare the methods. Consequently, there are no development costs for surveys and it is possible to choose the most appropriate method while optimizing the parameters. The consequences of an inappropriate choice of methods and/or parameters should not be neglected, since this can make imputation more onerous and affect the quality of the estimates.

Key Words: Detection; Outlier; Influential; Atypical; Robust.

1. Introduction

1.1 Definition

To date, there is no consensus on how to define an outlier. A good attempt was made by Hawkins (1983). According to his definition, the majority of observations follow an assumed model and the observations that deviate sufficiently from it are called outliers. This model concept will run throughout this article. A number of detection methods will be presented and the choice will depend on the underlying model.

1.2 Description

The structure of the observations and the detection objective influence the choice of one outlier detection method over another. With this in mind, a tool was designed to offer a number of detection methods and a means to compare them. The tool provides a graphic representation of the data and displays the impact of the choice of parameters. It is essentially a data development and analysis tool. Unlike the Banff generalized system, which among other things is designed for outlier detection, it is not oriented toward production. On the other hand, the Banff system has no graphical interface for displaying data. The Hidioglou-Berthelot and interquartile methods are already in Banff. A third, the sigma-gap method, is currently being integrated. The tool presented here includes these three methods as well as five others.

2. Methods

2.1 Background

In an internal survey at Statistics Canada asking survey managers what method they used, the Hidioglou-Berthelot and sigma-gap methods proved to be the most popular. The interquartile and top contributor methods are often considered, since they are quite simple. The other methods are often more complex, and they will be described only

¹Nelson Émond, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa ON, Canada, K1A 0T6, Nelson.Emond@statcan.gc.ca.

briefly. The tool offers the choice of eight methods: Hidiroglou-Berthelot, sigma-gap, interquartile, Sequential Bayesian Outlier, M estimation, Mahalanobis' distance, classical methods and top contributors. The tool was developed using SAS and processes only continuous data. It is fast and easy to use. It includes a number of functions to assist the user. These will be described in greater detail in a later section.

2.2 Methods

2.2.1 Hidiroglou-Berthelot

This is the method most often used at Statistics Canada (see Hidiroglou and Berthelot, 1986). This is a bivariate method which was originally developed for historical data, but which can also be used for correlated variables, such as revenue (x) versus profit (z). Its main significance lies in the fact that it considers size a factor in identifying outliers. There are three restrictions to bear in mind when applying this method:

- The observations must be strictly positive and continuous;
- There must be a linear relationship based on the ratio between the two variables;
- The line goes through the origin.

The steps in the method for a unit i in a specific class h are as follows:

1. Calculation of the ratio: $r_{hi} = \frac{x_{hi}}{z_{hi}}$;

2. Transformation of the ratio: $s_{hi} = \begin{cases} 1 - \frac{r_{hM}}{r_{hi}} & \text{if } 0 < r_{hi} < r_{hM} \\ \frac{r_{hi}}{r_{hM}} - 1 & \text{if } r_{hi} \geq r_{hM} \end{cases}$ where r_{hM} is the median of r_{hi} ;

3. Calculation of effects: $e_{hi} = s_{hi} [\text{Max}(x_{hi}, z_{hi})]^U$ where $0 \leq U \leq 1$, with U being the curvature factor;

4. Calculation of the first quartile (e_{hq1}), the median (e_{hM}) and the third quartile (e_{hq3}) of the effects (e_{hi});

5. Distance: $d_{hq1} = \text{Max}(e_{hM} - e_{hq1}, |A \cdot e_{hM}|)$
 $d_{hq3} = \text{Max}(e_{hq3} - e_{hM}, |A \cdot e_{hM}|)$

where $A (=0.05$ by default) is determined by the user and ensures a minimum value for the distance;

6. A value will be considered an outlier if $\begin{cases} e_{hi} < e_{hM} - C_{crit} \cdot d_{hq1} \\ e_{hi} > e_{hM} + C_{crit} \cdot d_{hq3} \end{cases}$

where the variables A , C_{crit} , and U are parameters determined by the user.

Figure 2.2.1-1 shows the main significance of the method, which takes account of the size variable on the horizontal axis. The larger the size, the lower the acceptance ratio. Figure 2.2.1-2 is a different representation of the same phenomenon.

The graphical representation is established according to a standard that includes the curves that delimit the acceptance zone beyond which values will be outliers. These curves are available when this is possible. Additional information is provided in the title. It is also possible to put the list of identifiers of the outlier observations in the chart (see Figure 2.2.1-2) or in the legend. All the other methods use the same graphical standard.

Figure 2.2.1-1
Representation of rate of variation vs. size

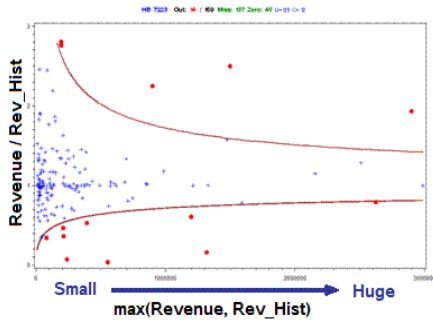
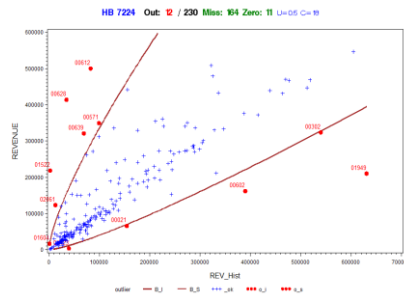


Figure 2.2.1-2
Representation of revenue vs. historical revenue



2.2.2 Sigma-gap

This is a univariate or bivariate method. In the event that it is bivariate, it can be compared to the Hidroglou-Berthelot method, since it is ratio-based. The data do not have to be exclusively positive. This method is based on the distance between two consecutive observations according to the value of the standard deviation. The steps in the method in a class h are as follows:

1. Assume that $r_{hi} = x_{hi}$ in the univariate case and $r_{hi} = \frac{x_{hi}}{z_{hi}}$ in the bivariate case;
2. Calculate the standard deviation (σ_h) according to one of the four methods available in the tool;
3. Put the values of r_{hi} in ascending order;
4. The user will provide the percentile (K) starting at which the distance criterion will be applied. Thus, all the values below K are not eligible to be outliers. For the others, when the distance between two consecutive observations satisfies the following condition:

$$r_{hi} - r_{hj} > C * \sigma_h \text{ where } j = i - 1 \text{ and } C > 0,$$

all the observations larger than r_{hj} will be considered to be outliers.

The user will have to determine the parameters K , C and the method of calculating the standard deviation. By default, the system uses the *Mad* method to calculate σ_h .

2.2.3 Interquartile

This is a ratio-based univariate or bivariate method. It does not take the size of the enterprise into account. It is a special case of the Hidroglou-Berthelot method in the bivariate case with $U=0$. It is often used, since it gives the user the option of having good control over the number of outliers. Here are the steps to follow:

1. Assume that $r_{hi} = x_{hi}$ in the univariate case and $r_{hi} = \frac{x_{hi}}{z_{hi}}$ in the bivariate case;
2. Put the values of r_{hi} in ascending order
3. Calculate the values of the 25th (Q_1), the median (Q_2) and 75th (Q_3) percentile;
4. An observation is an outlier if it satisfies one of the following criteria:

$$r_{hi} \begin{cases} < Q_2 - K_{inf} * [Q_2 - Q_1] \\ > Q_2 + K_{sup} * [Q_3 - Q_2] \end{cases}$$

The user will have to provide the parameters K_{inf} and K_{sup} that are a multiple of the interquartile distance.

2.2.4 Sequential Bayesian procedure

This method assumes *a priori* that there is a linear relationship between the independent variables and the variable of interest. This is a complex method, and you can consult Philips and Gutman (2006) for detailed explanations. In brief, it is an iterative process that removes one observation in each iteration. During an iteration, each observation receives a probability of being an outlier according to the influence that it has on the regression line. The observation that will be removed is the one with the highest probability of being an outlier. That probability depends on the RStudent/Student ratio.

The search for outliers stops when an acceptable correlation is reached or when there is no longer an observation with a sufficiently high probability of being an outlier or the maximum percentage of outlier observations, determined by the user, is reached. The user will have to provide the maximum percentage that is acceptable.

2.2.5 M Estimation

One of the goals of outlier detection is to find a robust estimator. Several techniques use non-parametric methods that do not assume a distribution *a priori*. But new parametric methods were created in the early 1960s, and one of these was M Estimation, pioneered by Huber (1964). There has been considerable development since that time. Briefly, the Estimation M technique can be summarized as follows:

$$\text{Minimize}_{\theta} \sum_{i=1}^n \rho(r_i)$$

Where ρ is a symmetrical function, *i.e.*, $\rho(-t) = \rho(t)$, θ is the vector of the regression line parameters and $r_i = y_i - \mathbf{X}_i^T \theta$.

By deriving in relation to the coefficients θ_j , we obtain the following function:

$$\sum_{i=1}^n \psi(r_i) x_i = 0$$

where x_i is the vector of the independent variables.

This yields p equations (p =number of independent variables + 1) which are often difficult to solve. In practice, an iterative process is used to find a solution. There are several choices of the ψ function, but the one offered by the system is Huber's. The originality of the proposed method is that it applies the interquartile outlier detection method to the robust residuals $\hat{r}_i = y_i - \mathbf{X}_i^T \hat{\theta}$, where $\hat{\theta}$ is the robust estimator of θ , in order to determine the outliers.

2.2.6 Mahalanobis' distance

This method assumes that most of the data are clustered around a central point. It is well summarized in the article of Franklin, Thomas and Brodeur (2000). The main steps are as follows:

1. Centre the data using the estimator L_1 (see Rousseeuw and Leroy, 1984);
2. Initialize the initial weights at $\delta_i^n = 1$ for $i = 1, \dots, n$;
3. Determine a desired number of iterations (10 iterations are usually sufficient);
 - a. Randomly generate a normalized vector " Y_1 ";
 - b. Calculate the other vectors Y_2, \dots, Y_p where p is the number of variables such that they form a set of orthogonal-orthonormal vectors;
 - c. Calculate the weights δ_i for each record;
 - d. If $\delta_i^k < \delta_i$ then $\delta_i = \delta_i^k$;
4. Calculate the new weighted vectors ($\hat{\mathbf{u}}$) and the variance-covariance matrix ($\hat{\mathbf{V}}$);
5. Calculate the robust Mahalanobis' distance: $D_i = (\mathbf{x}_i - \hat{\mathbf{u}})^T \hat{\mathbf{V}}^{(-1)} (\mathbf{x}_i - \hat{\mathbf{u}})$;

6. Perform the detection test: if $\frac{(n-p)n}{(n^2-1)p} D_i > F_{\alpha,p,n-p}$ then unit i is considered to be an outlier.

This is a robust iterative multivariate method initially developed by Patak (1990), which does not assume that the observations fit in with a regression model.

2.2.7 Classical methods

Under this heading are a number of statistical tests which emphasize observations that have a major influence on the least squares estimator. The diagnoses consist in a combination of numerical and graphical statistics. Several diagnoses are based on residuals, while others are based on the impact of suppressing an observation.

These tests include: projection matrix, Cook's distance, DfBetas, DfFits, CovRatio and RStudent. They are well documented in Weisberg (1985). SAS procedures are used to calculate these tests.

Ladiray and Ramsay (2003) applied these tests in the bivariate case. They emphasized the impact of atypical observations on the different correlation coefficients such as Kendall's, Spearman's, Pearson's and the R-square in addition to calculating the slope and the y-axis at the origin with or without the presence of atypical data. The user can choose one or more tests and an observation is considered to be atypical as soon as it fails at least one test. Also, atypical values can come from either of the two variables.

2.2.8 Top contributors

This is a highly intuitive method based on the idea that outliers are those values that are the most likely to influence estimates. Consequently, this method is more interested in large values. It leaves the user considerable latitude, and hence its popularity. The user can choose to determine the following:

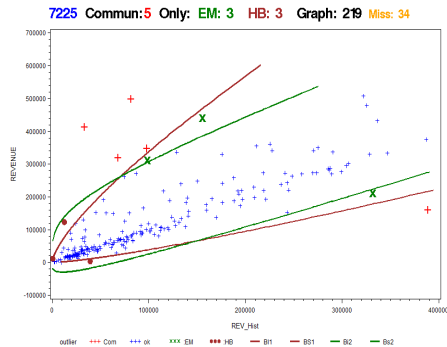
- Those with the largest values per stratum;
- Those that contribute to more than a certain percentage of the total for the stratum;
- Those that exceed certain threshold value per stratum;
- Those where the difference between two consecutive values is greater than a certain percentage;
- Those that are the largest observations that contribute to more than a certain percentage of the total (*e.g.*, 3.85, do the three largest values contribute to more than 85% of the total? If so, identify them as influential values).

An attractive feature of this method is that it generally yields few influential values.

2.3 Comparison of methods using the tool

This module was developed to show graphically the impact of one method compared to another. Various pieces of information related to the methods are shown on the graph. In the title, we see the stratum, the number of outlier observations that are common to the two methods, the number of outlier observations specific to each method, the number of observations on the graph and how many are missing. It is also possible to compare a single method with two sets of different parameters. Also, the curves that delimit each method will be represented when possible.

Figure 2.3-1
Estimation M (EM) vs. Hidioglou-Berthelot (HB)



2.4 Tool options

The user has several options:

- Declare the weight variable in the calculations for some methods;
- Declare the variable for variance structure of the regression model for some methods;
- Exclude values that are equal to zero, missing and/or negative;
- Export a graph in an HTML file;
- Put the number of the identifier on the graph or in the legend;
- Enlarge a particular area of the graph;
- Validate the number of observations per stratum, the type de variables, the name, *etc.*
- Export the list of outliers in a file.

3. Future developments

The emphasis will be on writing a Reference Guide and a Quick Reference Guide. Testing is under way to verify whether bivariate methods can be applicable with multivariate observations. There is the possibility of creating a more user-friendly interface if time permits. The priority will not be on developing new methods in the near future. It will be possible to add new functionalities to the existing methods, depending on the needs and requests of users.

4. Conclusion

Outlier detection is often the neglected part in surveys, owing to a lack of time or resources or a budget constraint. It should not be forgotten that poor detection of outliers has a negative impact on imputation and estimation. We often turn to an acceptable existing method, lacking the time to develop an alternative method. A tool such as the one described in this article makes available a large number of methods assembled in a single environment while making it possible to do graphic analyses. It should be noted that that no one method is better than another. One method will be preferable to another depending on the structure of the data, and on whether it meets the needs of the users.

References

- Hawkins, D.M. (1983), Outliers in “Encyclopedia of Statistical Science”, Eds. S. Kotz and N.L. Johnson, New York: John Wiley & Sons.
- Hidioglou, M.A. and J.-M. Berthelot (1986), “Statistical editing and imputation for periodic business survey”, *Survey Methodology*, vol.12, p.73-83.
- Huber, P.J. (1964), “Robust Estimation of a Location Parameter”, *Annals of Mathematical Statistics*, 35, 73-101.

- Ladiray, D. and L. Ramsay (2003), "Statistical Evaluation of the CoA-based Comparison between Tax and Survey Data", working paper, Ottawa, Canada: Statistics Canada.
- Philips, R. and I. Guttman (2006), "Towards the Robust Estimation of Parameters in the Univariate Linear Model", internal publication, Ottawa, Canada: Statistics Canada.
- Rousseeuw, P.J. and A.M. Leroy (1984), *Robust Regression and Outlier Detection*, New York: John Wiley.& Sons.
- Weisberg, S. (1985), *Applied Linear Regression*, 2nd edition, University of Minnesota St-Paul, Minnesota, New York: John Wiley& Sons, Chapter 5.
- Franklin, S., Thomas, S. and M. Brodeur (2000), "Robust Multivariate Outlier Detection Using Mahalanobis' Distance and Modified Stahel-Donoho estimators", ICES II.
- Patak, Z. (1990), "Robust principal Component Analysis Via Projection Unit", Master's Thesis, University British Columbia, Canada.

An assessment of methods to impute risk exposure into model actor's risk profile for microsimulation

Deirdre Hennessy, Carol Bennett, Meltem Tuna, Claude Nadeau, William Flanagan and Douglas Manuel¹

Abstract

Imputation of missing data exposure risk is a common and important issue in microsimulation models. Data for microsimulation are often assembled from multiple sources and requires imputation of both missing survey responses and missing variables (*i.e.*, variables required for modeling that are not available in the main source data used to initialize the model). While imputation for survey non-response can be achieved using validated methods, imputation of missing variables presents a greater challenge because it involves using external 'donor' data that may or may not be comparable to the initial data source. Microsimulation modellers use a variety of approaches to impute missing variables; however it is unclear which approach produces the most valid results. A variety of approaches to imputation for microsimulation present advantages depending on the purpose of the variable in the eventual microsimulation model. In preparation for constructing a Population Health Microsimulation (POHEM) model of cardiovascular disease and salt intake, we investigated various techniques to create imputed variables for blood pressure and cholesterol, which are core risks for cardiovascular disease. We used the Canadian Community Health Survey 2.2 as the initial data source for most cardiovascular risks including sodium intake. Blood pressure and cholesterol information was obtained from donor data (the Canadian Health Measures Survey). Two approaches to imputation were evaluated. First, regression imputation which involves the use of one or more variables common to both data files. Second we examined 'hot-deck' imputation methods which assign an actual value of blood pressure or cholesterol from the donor data for each model actor. Our approach was to compare the methods based on accuracy, discrimination and validity of the imputed values; however the statistical properties of each method and the implications of using the resulting data in microsimulation also had to be carefully considered.

Key Words: Imputation; Survey-data; Microsimulation; Blood pressure; Cholesterol.

1. Introduction

1.1 Imputation for Microsimulation

Imputation of missing exposure risk is a common, complex and important issue in almost all microsimulation models for health. Multiple sources of data are often required to generate actor health profiles, including socioeconomic characteristics, exposure to health risks (risk exposure) and disease status. Typically, a main source of data is used to initialize each actors profile and then missing characteristics are assembled and imputed from multiple sources of donor data. A variety of approaches are used to impute missing characteristics, however it is unclear which approach produces the best results. That is, a complete dataset containing imputed variables that will behave similarly in comparison with the donor dataset and that will associate in an expected way (according to other published results) with variables in the initial data source. In the context of developing a Population Health Microsimulation (POHEM: CVD) model to illustrate and project the relationship between sodium intake and cardiovascular disease (CVD), imputation was essential because not all important exposures are captured in a single data source. We used the Canadian Community Health Survey (CCHS) 2.2 to obtain data for most characteristics of actors' health profile, including sodium intake and other nutritional information; however this data did not include blood pressure and cholesterol measurements which are core risks that are used in all CVD risk algorithms. Therefore it was necessary

¹Deirdre Hennessy, Ottawa Hospital Research Institute and Statistics Canada, 1053 Carling Ave., Ottawa, ON, K1Y 4E9 (deirdre.hennessy@statcan.gc.ca); Carol Bennett, Ottawa Hospital Research Institute, 1053 Carling Ave., Ottawa, ON, K1Y 4E9 (cbennett@ohri.ca); Meltem Tuna, Ottawa Hospital Research Institute, 1053 Carling Ave. Ottawa, ON, K1Y 4E9, (mtuna@ohri.ca); Claude Nadeau, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6 (claudio.nadeau@statcan.gc.ca); William Flanagan, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON, K1A 0T6, (william.flanagan@statcan.gc.ca); Douglas Manuel, Ottawa Hospital Research Institute and Statistics Canada, 1053 Carling Ave., Ottawa, ON, K1Y 4E9 (dmanuel@ohri.ca).

to impute these measures from the Canadian Health Measures Survey (CHMS) and then validate the results of the imputation before proceeding with the microsimulation model specification.

Two general approaches to imputation were evaluated. First, regression imputation, which involves constructing a model using one or more common variables (variables common to both data files, for instance age, sex and body mass index (BMI)), was examined. An advantage of this method is that the variable to be imputed can be modelled on a potentially large set of other variables that explain their causal relationship (Durrant, 2005). In addition, regression models have the ability to describe the relationship between variables using model coefficients that can be compared to causal relationships established in other studies. The coefficients produced by a regression model can also be used by other researchers without the need for donor data. A potential disadvantage of this of this method is that it may distort the distribution of the predictor variable and inflate the association between the predictor and other variables in the model. In addition, the imputed values are predicted, not actually observed, in another data source and this is a parametric approach and may be sensitive to misspecification of the regression model (Durrant, 2005).

Second, we examined ‘hot-deck’ imputation methods which assign an actual value of blood pressure or cholesterol from donor data randomly within imputation classes, which are constructed based on cross-classifications of fully observed common variables (for instance age, sex and BMI). An advantage of this method is that it is non/semi parametric, and aims to avoid making distributional assumptions, which is an important property if the data to be imputed are skewed. Using hot-deck methods the model actors’ imputed values should have the same distributional shape as the similar survey respondents in the donor data. This is important because if extreme values of the imputed variable represent those individuals at highest risk of disease/death in the population, which is the case with cholesterol and blood pressure, it is important to recreate this distribution in order to accurately predict risk of disease/death. Moreover, because the imputed values are used in a start-up population for simulation, inaccuracies generated at this stage will become additive as the simulation model projects forward in time. Another advantage of this method is that a ‘donor’ can contribute many missing values to a ‘donee’ record. In this study all the physical measures data came from the same database (CHMS) therefore using the same donor to contribute all the missing variables may help to preserve interrelationships between measurements performed on the same individual. A disadvantage of this model is that it depends completely on donor data, which may or may not include variables collected/measured in a similar compared to the initial data. Moreover, important variables which may be causally related to the imputed variable and which could be used in the hot-deck imputation may not be collected at all in the donor data. Another disadvantage of this method is that the sample size of the donor data must be reasonably large in order for it to work well (Durrant, 2005).

We tested regression and hot-deck imputation, assessing each method using three criteria. First we assessed the accuracy of imputation by comparing the imputed values overall, and for different subpopulations, to the CHMS. Next the discrimination of imputation or the ability to discern exposure between individuals or groups was examined. Discrimination was assessed by comparing the distribution and range of values (1, 5, 10 *etc.* percentiles) between the imputed and CHMS data. Finally, the imputed data was examined to determine if known relationships between variables within the data are preserved, for instance the association of older age with higher blood pressure. The relationship between these variables in the imputed data should correspond to the relationship described from systematic reviews and meta-analyses of other studies (Khaw and Barrett-Connor, 1988; Strazzullo *et al.* 2009).

The objective of this study was to investigate various techniques to create imputed variables for blood pressure and cholesterol, using the CHMS 2007-2009 as the donor data and the CCHS 2.2 as the recipient data.

2. Methods

2.1 Data Sources

2.1.1 Initial Source: Canadian Community Health Survey 2.2

The CCHS 2.2 was used as the main or initial data source. This was a nationally representative survey of 35,107 (21,106 adults \geq 18 years) Canadians conducted in 2004. This survey collected data on self-reported health, chronic disease status and physical activity as well as detailed food intake data including a 24-hour dietary recall survey, which is the gold-standard of dietary information available in Canada.

2.1.2 Donor Source: Canadian Health Measures Survey

The CHMS 2007-2009 survey was used as the donor data. This survey collected data on self-reported health, chronic disease status and physical activity, in a similar manner to the CCHS surveys. In addition, physical measures of health were collected, including BMI and blood pressure and cholesterol levels. Wave 1 conducted in 2007-2009 collected data on 5,604 (3,719 adults ≥ 18 years) Canadians at 15 sites across the country.

2.2 Statistical Analysis

2.2.1 Regression Imputation

In regression imputation, a regression model is fitted that relates Y to the auxiliary variables (fully observed data common between the datasets) X . The predicted values obtained through modelling are used for imputation of the missing values in Y . More simply, let $Y_{(imputed)} = E\{Y/X\}$ (Durrant, 2005).

In this study we modelled measured levels of systolic (SBP) and diastolic blood pressure (DBP) as well as total cholesterol and high density lipoprotein (HDL) (all available in the CHMS) on a set of common auxiliary variables available in both the CHMS and CCHS 2.2. Using simple linear regression techniques with survey and bootstrap weights, models were constructed separately for males and females and the predicted values obtained were used to impute the missing values. A summary of the predictor variables in the final models for blood pressure and cholesterol are presented in Table 2.3.1-1.

Table 2.3.1-1
Summary of predictor variables in imputation models of SBP, DBP, total cholesterol and HDL

	Age (c)	BMI (c)	HTN (2) awareness	HTN drugs (2)	Heart disease (2)	Diabetes (2)	Education (2)	Marital status (2)	Ethnicity (2)	Owens house (2)	Physical act. (2)	General health (2)	Daily (2) smoker
SBP													
Males	√	√	√		√								
Females	√	√	√				√						
DBP													
Males	√	√	√	√	√	√		√			√		
Females	√	√	√	√	√								
Total cholesterol													
Males	√	√	√		√	√		√					
Females	√	√	√			√							√
HDL													
Males	√	√						√		√			
Females	√	√			√	√			√		√	√	√

Notes: (c)= continuous variable, (2)= number of categories in the variable, act.= activity, SBP= systolic blood pressure, DBP= Diastolic blood pressure, HDL= high density lipoprotein, BMI= body mass index, HTN= hypertension.

2.3.2 Hot-deck imputation

'Hot-deck' imputation assigns the value from a record with observed data to a record with missing data. Using this method, missing values are imputed from other records in the database that share attributes related to the incomplete variable. To achieve this imputation, classes are constructed based on common auxiliary variables available in both datasets and donor values are selected from within the imputation classes (Kalton and Kasprzyk, 1982; Durrant, 2005).

In this study we used the a subset of the most important predictor variables, obtained in the regression modelling described above, as imputation classes in the hot-deck imputation, see Table 2.3.1-1. We also performed the

imputation by sex. We used a SAS macro for iterative hot-deck imputation that allowed us to impute missing values of SBP, DBP, total cholesterol and HDL simultaneously on our imputation classes (Ellis, 2007).

3. Results

3.1 Regression imputation - Accuracy

The accuracy of the imputation was assessed by comparing estimates of SBP between the CHMS and imputed data overall and by important subgroups. The SBP imputation was chosen as an illustrative case due to space constraints in this paper. Table 3.1-1 shows that regression imputation estimates the mean and the median values of SBP accurately overall and for age, sex and education level. Notably, the range of values estimated using regression imputation overall and across the subgroups is considerably truncated, which is evident from the maximum and minimum values. This was further explored in the next section.

Table 3.1-1
Accuracy of imputed values of SBP overall and by subgroup

	Don.	Imp.	Don.	Imp.	Don.	Imp.	Don.	Imp.
	Mean	Mean	Median	Median	Max.	Max.	Min.	Min.
Overall	116.91*	116.55	114.63	115.90	194.61	150.08	80.22	94.72
Sex								
Males	118.73	116.04	117.42	115.79	194.61	140.073	83.01	99.82
Females	115.28	116.96	111.84	116.07	189.96	150.08	80.22	94.72
Age group								
18-44 years	109.72	106.04	109.05	106.04	153.69	124.74	80.22	94.72
45-64 years	120.31	118.34	118.35	117.54	185.31	134.02	86.73	108.20
65+ years	129.04	129.70	127.65	129.01	194.61	150.08	84.87	114.24
Education								
<Secondary education	119.86	122.55	117.42	123.08	189.96	148.47	83.94	94.72
Secondary or higher	115.55	114.30	113.70	113.65	194.61	150.081	80.22	96.62

Notes: SBP= systolic blood pressure, Don.=donor, Imp.= imputed, Max.= maximum, Min.= Minimum,
*SBP is measured in millimetres of mercury (mm/Hg).

3.2 Regression Imputation - Discrimination

The discrimination of the regression imputation was assessed by comparing the distribution and range of values estimated. Histograms comparing the measured and imputed values of SBP from the CHMS and CCHS demonstrated that the range of imputed values was truncated compared to those in the CHMS (results not shown). Table 3.2-1 compares the percentiles of measured and imputed data and presents a difference between the two values. Especially at the high end of the distribution (90th percentile and above), the imputed data significantly underestimates the CHMS values.

Table 3.2-1
Comparing the distribution of measured to imputed values for SBP

	Don.	Imp.	Difference
Percentiles			
1%	91.38	95.69	-4.31
5%	96.96	98.18	-1.22
10%	99.75	101.56	-1.81
25%	106.26	108.32	-2.06
50%	114.63	115.90	-1.27
75%	124.86	124.30	0.56
90%	136.95	131.98	4.97
95%	143.46	136.71	6.75
99%	163.92	142.05	21.87

Notes: SBP= systolic blood pressure, Don.=donor, Imp.= imputed

3.3 Hot-deck imputation: Accuracy

Table 3.3-1 shows that hot-deck imputation also estimates the mean and the median values of SBP accurately overall and for sex and education level. Within the age groups there was some discrepancy (never more than 4.05 units) between the donor and imputed data. The range of values estimated using hot-deck imputation overall and across the subgroups is very similar compared to the donor data, with the imputed data producing slightly lower minimums than those observed in the CHMS. This was further explored in the next section.

Table 3.3-1
Accuracy of imputed values of SBP overall and by subgroup

	Don.	Imp.	Don.	Imp.	Don.	Imp.	Don.	Imp.
	Mean	Mean	Median	Median	Max.	Max.	Min.	Min.
Overall	116.91	115.01	114.63	113.00	194.61	197.00	80.22	74.00
Sex								
Males	118.73	116.25	117.42	115.00	194.61	197.00	83.01	77.00
Females	115.28	114.00	111.84	111.00	189.96	192.00	80.22	74.00
Age group								
18-44 years	109.72	106.10	109.05	105.00	153.69	153.00	80.22	74.00
45-64 years	120.31	116.72	118.35	115.00	185.31	187.00	86.73	81.00
65+ years	129.04	126.83	127.65	125.00	194.61	197.00	84.87	79.00
Education								
<Secondary education	119.86	119.73	117.42	118.00	189.96	197.00	83.94	74.00
Secondary or higher	115.55	113.41	113.70	112.00	194.61	194.00	80.22	77.00

Notes: SBP= systolic blood pressure, Don.=donor, Imp.= imputed, Max.= maximum, Min.= Minimum

3.4 Hot-deck imputation - Discrimination

Histograms comparing the measured and imputed values of SBP from the CHMS and CCHS demonstrated that the range of imputed values was very similar compared to those measured (results not shown). Table 3.4-1 compares the percentiles of measured and imputed data and presents a difference between the two values. In comparison to regression imputation, the hot-deck method more accurately reconstructs the distribution of imputed values across the whole distribution. In contrast to regression imputation, hot-deck imputation underestimates values at the lower end of the distribution.

Table 3.4-1
Comparing the distribution of measured to imputed values for SBP

	Don.	Imp.	Difference
Percentiles			
1%	91.38	86.00	5.38
5%	96.96	93.00	3.96
10%	99.75	96.00	3.75
25%	106.26	104.00	2.26
50%	114.63	113.00	1.63
75%	124.86	124.00	0.86
90%	136.95	136.00	0.95
95%	143.46	145.00	-1.54
99%	163.92	165.00	-1.08

Notes: SBP= systolic blood pressure, Don.=donor, Imp.= imputed

3.5 Validity

In this study the validity of the imputed values was difficult to assess as we did not have a comparison dataset that incorporated both **measured** sodium intake and **measured** blood pressure. One approach is to examine whether known relationships, for instance increasing levels of high blood pressure with age, body mass index (BMI) and by self-reported hypertensive status, are preserved in the imputed dataset (Khaw and Barrett-Connor, 1988; Strazzullo et al. 2009). As can be seen from Tables 3.1-1 and 3.2-1 above, values imputed via regression and hot-deck methods both increase with age. Moreover, the expected relationships of mean SBP with increasing BMI and self-reported hypertensive status were also preserved in the imputed dataset (results not shown). In addition, we constructed regression models of imputed SBP, by sex, adjusted for age, BMI and self-reported/awareness of high blood pressure diagnosis. The results comparing both the regression and hot-deck methods to the CHMS are presented in Table 3.5-1. The significant association between age, BMI and SBP is preserved in both imputed datasets, with the regression coefficients being similar. Notably the confidence intervals (CIs) around the coefficients produced via regression imputation are very narrow. In contrast, the significant association between hypertension status and SBP is not preserved using the hot-deck method, while it is using the regression method. While there was a statistically significant relationship between the mean SBP when compared across hypertensive status (results not shown), stratifying by sex and adjusting for age and BMI seems to remove the association. These results, particularly in relation to the narrow confidence intervals produced by regression imputation and the non-significant association between hypertensive status and SBP demonstrated by the hot-deck imputation, warrant further investigation and may result in some adjustment to the way in which the hot-deck imputation is implemented, for instance by including additional imputation classes.

Table 3.5-1
Relationship between imputed SBP and other variables in donor and imputed data

Variable	Coefficient		CIs		p-value	
	Male	Female	Male	Female	Male	Female
CHMS- Age (years)	0.26	0.45	(0.22,0.29)	(0.41,0.49)	<0.00	<0.00
Reg.	0.29	0.56	(0.28,0.29)	(0.55,0.56)	<0.00	<0.00
HD.	0.26	0.50	(0.23,0.28)	(0.47,0.53)	<0.00	<0.00
CHMS- BMI (kg/m ²)	0.23	0.41	(0.08,0.38)	(0.23,0.58)	<0.00	<0.00
Reg.	0.25	0.13	(0.23,0.25)	(0.12,0.13)	<0.00	<0.00
HD.	0.33	0.49	(0.39,0.58)	(0.23,0.42)	<0.00	<0.00
CHMS- Hypertension status (no/yes)	5.05	9.07	(2.69,7.42)	(6.62,11.52)	<0.00	<0.00
Reg.	5.09	8.45	(4.92,5.25)	(8.38,8.52)	<0.00	<0.00
HD.	0.27	0.65	(-0.94,1.48)	(-0.69,1.99)	0.34	0.67

Notes: SBP= systolic blood pressure, CIs= confidence intervals, Reg.= regression, HD.= hot-deck, BMI= body mass index.

4. Conclusions

In this study, regression and hot-deck imputation techniques were assessed as methods to impute important risk factors into model actors' profiles for microsimulation. While both methods were successful in reproducing mean and median values of the imputed variable, hot-deck methods were more successful at recreating the distribution of values. In addition, the validity of the imputed values was investigated, albeit in a limited way. These results demonstrated that the expected associations of SBP with age and BMI were preserved in both imputed datasets, while the association of SBP with hypertensive status was not preserved by hot-deck imputation. Further investigation of these results is warranted and future analysis will investigate how relationships between imputed values and variables **not** used in the imputation process are preserved. Moreover, alternative imputation methods, such as regression imputation from a conditional distribution and multiple imputation, will be assessed. Finally, the ultimate association between sodium intake and imputed SBP will also have to be investigated to establish the validity of the imputed values.

Overall, these results suggest that regression and hot-deck imputation methods have different advantages and the choice of method may ultimately depend on the purpose imputed variable in the microsimulation model. In conclusion, accurate imputation that produces valid results is important for microsimulation because the eventual purpose of the simulated data is to provide projections of future trends of many aspects of health status in order to guide decisions about policy. This study aimed to report the performance of two different imputation techniques, which could be used in preparation of datasets for microsimulation, in a transparent and relevant manner.

Acknowledgments

Deirdre Hennessy was employed through a postdoctoral fellowship from the Canadian Institutes for Health Research (CIHR) funded Simulation Technology in Advanced Research (STAR) team. Douglas Manuel holds a Chair in Applied Public Health from CIHR and the Public Health Agency of Canada.

References

- Durrant, G.B. (2005), "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review", Working Paper, United Kingdom: National Centre for Research Methods, Southampton Statistical Sciences Research Institute, University of Southampton.
- Ellis, B. (2007), "A consolidated macro for iterative hot-deck imputation", poster presented at NorthEast SAS Users Group - 2007, Baltimore, MD.
- Kalton, G. and D. Kasprzyk (1982), "Imputing for Missing Survey Responses", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, p. 22-31.
- Khaw, K.T. and E. Barrett-Connor (1988), "The association between blood pressure, age, and dietary sodium and potassium: A population study", *Circulation*, Vol. 77, No. 1, pp. 53-61.
- Strazzullo, P., D'Elia, L., Kandala, N.B. and F.B. Cappuccio (2009), "Salt intake, stroke, and cardiovascular disease: Meta-analysis of prospective studies", *BMJ*, 339, b4567.

SESSION 7A

**CONFIDENTIALITY METHODS AND TOOLS FOR ACCESSING DATA
WHILE PRESERVING CONFIDENTIALITY**

De-identification methods for public use health files

Khaled El Emam¹

Abstract

Increasing amounts of de-identified health data are being made publicly available. For example, CMS in the US is now posting claims data online, the Heritage Provider Network in California has launched a competition to entrants worldwide to develop a model to predict hospitalization using health data on more than 150,000 patients, and CIHR in Canada now requires the public disclosure of individual-level clinical trials data for trials that they fund. This presentation will describe a quantitative methodology for the creation of public use files, under varying access restrictions, for the management of identity and attribute disclosure risks. The methodology has been used to disclose large health data sets over the last 3 years, and examples of the issues that arise during these disclosures will be discussed.

¹Khaled El Emam, University of Ottawa, Canada.

The U.S. Census Bureau's microdata analysis system

Michael Freiman, Jason Lucero, Lisa Singh, Jiashen You, Michael DePersio and Laura Zayatz¹

Abstract

This paper describes a Microdata Analysis System (MAS), currently under development, that allows users to generate analyses of confidential Census Bureau data, such as cross-tabulations and regressions, without having access to the underlying microdata. Users may perform analyses on a universe (subset) of their choosing, subject to restrictions. For allowable universes, a random subsample of observations in the universe is dropped from all further analyses. This 'Drop q Rule' is crucial in protecting tabular data. For regression, there are some further rules regarding how categorical variables are handled and what interactions and transformations are allowed. Once a regression is submitted, it is checked to ensure that the goodness of fit is not so high as to create a disclosure risk. Confidentialized diagnostics are also given, and we are adding additional capabilities to the system, such as summary statistics, histograms and scatterplots—all suitably protected.

Key Words: Disclosure; Remote access system; Tabulation; Regression; Synthetic data.

The U.S. Census Bureau collects its survey and census data under Title 13 of the U.S. Code, which prohibits the Census Bureau from releasing any data “whereby the data furnished by any particular establishment or individual under this title can be identified.” The Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) also requires the protection of information collected or acquired for statistical purposes under a pledge of confidentiality. However, the agency also has the responsibility of releasing data for the purpose of statistical analysis. Like other national statistical agencies, our goal is to release as much high quality data as possible without violating the pledge of confidentiality.

This paper discusses a Microdata Analysis System (MAS) that is under development at the U.S. Census Bureau. Much of the framework for the system was described in Steel and Reznick (2005) and Steel (2006). The system is designed to allow data users to perform various statistical analyses (regressions, cross-tabulations, univariate or bivariate summary statistics, *etc.*) on confidential survey and census microdata without seeing or downloading the underlying microdata.

Section 2 gives some background on the MAS and the motivation for its development. In section 3, we discuss the current state of the system, including its capabilities and the rules that protect confidentiality, with a focus on regression and tables. In section 4, we consider some of the other features available in the system. In section 5, we examine another approach to the problem of creating a remote access system. In section 6, we conclude with remarks on future research and the further development of the system.

2. Background on the MAS

For the U.S. Census Bureau and other statistical agencies throughout the world, the problem of data confidentiality, as well as the increasing desire for a wide variety of customizable data products, has motivated the creation of online remote access systems that allow the user to request a statistical analysis and receive the results without having direct

¹Michael Freiman, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233 (michael.freiman@census.gov); Jason Lucero, Freddie Mac, 8200 Jones Branch Drive, McLean, VA 22102; Lisa Singh, Georgetown University Department of Computer Science, 329A St. Mary's Hall, Washington, DC 20057; Jiashen You, University of California-Los Angeles Department of Statistics, 8125 Math Sciences Bldg., Box 951554, Los Angeles, CA 90095; Michael DePersio, University of Delaware Department of Mathematical Sciences, 501 Ewing Hall, Newark, DE 19716; Laura Zayatz, U.S. Census Bureau. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

access to the underlying microdata. Depending on the specifics of the system, the result might be based on perturbed data, and some queries may be rejected to preserve confidentiality. The concept of a remote access system is not new, and indeed the idea of a system allowing customized queries was proposed at least as early as the Geographically Referenced Data Storage and Retrieval System described by Fellegi and Goldberg (1969).

The Microdata Analysis System will allow the U.S. Census Bureau to provide access to information that may be more complete and detailed than what would otherwise be available to most users. Furthermore, the system will be open to anyone who wants to use it, without a fee or application process, thus extending access to those for whom a visit to one of our Research Data Centers is unfeasible. The MAS will initially allow access to data from demographic surveys and decennial censuses, and we hope that when the system is further developed, it will also be able to handle economic data. We will initially make available regression and cross-tabulations, with other types of analyses to be added in the future. We intend to keep a record of all of the queries entered into the system, but not the identities of the users making the queries. Although this record will not directly affect the output that the system provides to the recorded queries, it will allow us to see how the system is being used, so that we can make tweaks for usability and disclosure avoidance as necessary.

Our current plan—as described in Chaudhry (2007)—is to offer the MAS through a Java interface within the Census Bureau’s free online DataFerrett service. The MAS has a graphical interface that allows users to select variables of interest from a list. In the case of regression, variables can be dragged into models and, with a few clicks, users may create variable interactions and transformations of selected variables.

One suggested use of the MAS has been as a preliminary way of examining data. Although a researcher might wish to perform an analysis more complex than the MAS allows, a related query to the MAS could be used to get some basic information about the variables of interest. This may inform the researcher’s decision of whether to go ahead with the more costly and time-consuming process of gaining access to a Research Data Center. Furthermore, since access to these centers is by application only, use of the MAS may allow a more informed and thoughtful application.

3. Overview of the MAS Confidentiality Rules

The MAS software is programmed with several confidentiality rules and procedures that uphold disclosure avoidance standards, many of which were devised inside the Census Bureau, and some of which were developed by Prof. Jerome Reiter at Duke University. The purpose of these rules and procedures is to prevent data intruders from using the results of one or multiple queries to reconstruct the individual microdata records, either completely or partially.

3.1 Confidentiality Rules for Universe Formation

MAS users may choose a universe, or sub-population, on which to run their analyses. In this paper, we will use $U(n)$ to denote a universe with n observations. This notation is ambiguous, but the particular universe of interest will generally be apparent from the context. The system gives a set of variables and category levels from which a user can define a universe using condition statements on the variables. For example, the user may select a universe consisting of the sub-population of all females in a geographic region. A more complicated universe could consist of all people who are male or unemployed, or of all people whose income in dollars falls into the union of $[9180,20155]$ and $[31662,43468]$, although admittedly the last of these may be of dubious utility (and, as described below, is probably inadmissible as presented here, because of the role of cutpoints in the system). One of the confidentiality rules requires that all variables used to define universes must be categorical.

Since a user may wish to define a universe based on variables that are inherently continuous rather than categorical, each continuous variable is accompanied by a list of allowable interval cutoffs known as *cutpoints*, as outlined in Lucero *et al.* (2009). When a universe is defined using a numerical variable, the only ranges of that variable that may be used are those based on this predetermined list. For example, if the variable of interest is adjusted gross income and the cutpoints are \$0, \$10,000, \$23,000, \$35,000, \$52,500 and \$100,000, then the universe of individuals with income in the interval $(\$23000, \$52500]$ would be allowable, since both of its endpoints are cutpoints, while the

universe of individuals with income in the interval (\$35000,\$38000] would not be allowable, as one of its endpoints is not a cutpoint. This restriction helps to prevent a differencing attack disclosure, a concern further described in Section 3.1.1. Such a disclosure would occur, for example, if a user requested a table for the universe of individuals with an income of more than \$11,313 and the corresponding table for the universe of individuals with an income of more than \$11,314, and then manually compared the two tables. If only one person in the dataset had an income of exactly \$11,314, then this person’s other attributes could easily be deduced.

Cutpoints are assigned when the data are loaded into the MAS, and they remain unchanged thereafter. In some cases, a variable will be given multiple possible sets of cutpoints, with the relevant cutpoints in a given case determined by the other characteristics of the universe. In particular, if the geographical region that defines the universe is widened, then the list of cutpoints may be expanded, so that the user has access to more refined intervals, as larger geographies decrease disclosure risk of individual records.

There are also further requirements on universes that apply both with inherently categorical variables and with cutpoint variables, such as a minimum universe size and a requirement that cells of certain tables associated with the universe definition not be too sparse.

3.1.1 Differencing Attack Disclosures and Random Record Removal

A major concern with tables in a remote access system is that a *data intruder* will be able to identify attributes of an individual respondent (the *target observation*) by subtracting the statistical analysis results obtained through two queries on similar universes, a method called a *differencing attack*. The potential problem is that a data intruder may create two universes in the MAS, $U(n)$ and $U(n-1)$, where both contain the same n observations with the exception of the one target observation missing from the second universe. In this case, $U(n) \setminus U(n-1) = U(1)$ is the universe consisting solely of the target observation. For example, suppose that the intruder knows that there is only one non-citizen among the n residents of the area of interest. Then the intruder may create $U(n)$ and $U(n-1)$, where $U(n)$ is the full universe of people in the area and $U(n-1)$ is the universe consisting of citizens who live in the area. Suppose the intruder then requests two separate cross-tabulations for the same underlying table variables; we call these two tables T_n and T_{n-1} , as shown in Figure 3.1.1-1. The tables below show a differencing attack based on a tabulation of age (a binary classification of whether the person is at least 45 years old) versus income (a binary classification of whether income is at least \$50,000).

Figure 3.1.1-1
An Example of Performing a Differencing Attack by Matrix Subtraction

All People		
T_n	<\$50,000	\geq \$50,000
Age<45	323	170
Age \geq 45	45	58

-

Citizens Only		
T_{n-1}	<\$50,000	\geq \$50,000
Age<45	323	169
Age \geq 45	45	58

=

Non-Citizens Only		
T_1	<\$50,000	\geq \$50,000
Age<45	0	1
Age \geq 45	0	0

The intruder may perform the matrix subtraction $T_n - T_{n-1} = T_1$, where T_1 is a two-way table of gender by employment status for non-citizens, of whom there is only one. As shown in Figure 3.1.1-1, T_1 contains a cell count of 1 in the cell of people under age 45 and with income of at least \$50,000, which tells the data intruder that the non-citizen contained in $U(1)$ has these two characteristics. The intruder has now created a disclosure about the non-citizen, even though the universe creation restrictions in the MAS would not make the universe of non-citizens

available directly. By performing differencing attacks similar to the one just described, a data intruder can successfully rebuild the entire confidential microdata record for the one observation contained in $U(1)$.

A differencing attack may also be a concern when there are exactly two observations that have a certain characteristic—such as being non-citizens—and the intruder is himself one of these two. Such an intruder could then construct the full universe $U(n)$ and the portion of the universe consisting solely of citizens $U(n - 2)$. He may then manually remove himself from $U(n)$ to get $U(n - 1)$ —the universe of all people other than the intruder. Put another way, $U(n - 1)$ is the universe consisting of all citizens and the one non-citizen other than the intruder. The intruder may then perform a differencing attack as above by comparing and to obtain information on the other non-citizen.

Checks made when a universe is selected help to prevent successful differencing attacks, but the main tool to thwart such attacks is the ‘Drop q Rule.’ A user-defined universe that passes all of the checks has q records removed at random. To do this, the MAS will first draw a random integer value of q such that $2 < q < k$ and such that when the universe is modified by omitting q records, the number of remaining records is a multiple of 3. Here k is some predetermined number, which depends on the size of the universe. Then, given q , the MAS will subsample the universe $U(n)$ by removing q records at random from $U(n)$ to yield a subsampled universe $U(n - q)$, which will be used for all subsequent analyses. The MAS will produce only one subsampled universe $U(n - q)$ for each unique universe $U(n)$, and this universe will be used for the lifetime of the system, so that extra information about the original universe cannot be obtained by repeatedly examining different subsamples.

The differencing attacks of most concern require, among other things, that two universes are available that differ in size by 1 or 2. However, under the Drop q Rule described above, all subsampled universes have sizes that are multiples of 3, and no pair of multiples of 3 (including pairs where both numbers are the same) can have a difference of 1 or 2. Hence the Drop q Rule eliminates the possibility of this sort of disclosure, or even of an apparent disclosure of this sort. (By an “apparent disclosure,” we mean a matrix subtraction where the resulting difference has no cells with negative value and where the sum of the values across all cells is equal to the number of observations—in this case either 1 or 2—that the intruder wishes to isolate. In this case, the intruder may believe he has obtained a disclosure, even if the disclosure is inaccurate.)

A less likely scenario involves some group of $j > 2$ people sharing a characteristic (or combination of characteristics) otherwise unseen in the dataset. In this case, $j - 1$ of these people could conspire against the one other person to create two tables $U(n)$ and $U(n - j)$, then manually remove themselves from $U(n)$ to get $U(n - (j - 1))$, which, in combination with $U(n - j)$, can be used to obtain a disclosure against the non-conspiring individual. The Drop q Rule makes such an endeavour unlikely to succeed. However, the possibility of success, or even apparent success, can be eliminated by replacing the requirement that the subsampled universe size be a multiple of 3 with a requirement that the subsampled universe size be a multiple of $m + 1$, where m is the largest value of j for which such collusion is a concern. This is because the collusion scenario requires two universes differing in size by exactly j , but no two multiples of $U(n - j)$, can differ by exactly this amount. However, the conditions necessary for this type of attack to work seem so specific that we are inclined to keep the requirement that the subsampled universe size is multiple of 3 now, rather than use a multiple of some other number.

In addition to the two main types of differencing disclosure described above, another concern is the possibility that there is a relatively small group, all of whose members share some characteristic other than the one(s) used to define the group. For example, suppose that a differencing attack indicates that there are five female Korean War veterans in some region, and their marital status is examined. If we find that two are married, two divorced and one widowed, then there is no apparent disclosure about any particular individual. However, if we find that all five are divorced, then we have made an apparent disclosure about all five. The Drop q Rule helps here, for much the same reason that sample data are inherently less susceptible to disclosure than census data. In this instance, it is not necessarily the case that all female Korean War veterans are divorced, as there may be one or more female Korean War veterans who are not divorced but who were removed from the dataset by the Drop q Rule. Intruders should recognize that they cannot get a disclosure that is known to be correct using this method, even in instances when the MAS output indicates an apparent disclosure.

3.2 Confidentiality Protection for Regression Models

Regressions in the MAS must be performed on an allowable universe, and the system includes several additional rules specifically applicable to regression. There is a limit of 20 independent variables for a regression equation. Numerical predictor and response variables may be transformed, but only transformations from a fixed list are allowable, so as to prevent the user from performing transformations that deliberately overemphasize outliers or other particular observations. Currently, only square, square root and natural logarithm are allowed, although this list will probably be expanded.

Reznek (2003) and Reznek and Riggs (2004) describe a major disclosure risk that can arise from fully interacted regression models with only dummy variables as predictors. Hence we have limitations on interaction terms: no more than three variables may be interacted with each other, and we do not allow fully interacted models. Users create interactions by clicking on predictor variables already in the model, so two-way interactions are only possible if each interacted variable appears uninteracted in the model. A similar situation exists with three-way interactions, and when the system creates a three-way interaction, it also creates all of the corresponding two-way interactions between pairs of variables involved. Categorical predictor variables are incorporated into the model using dummy variables for all categories except one reference category; the variable's most common category is used as the reference category. For a category of any categorical variable to have any dummy variables associated with it or with its interactions, it must have at least some number m of observations; if this minimum is not met, all dummy variables for that category are omitted from the model. In effect, this means that sparse categories are absorbed into the reference category. We initially set $m = 3$, but this can be modified as described below.

Regressions with high values of R^2 pose another potential disclosure risk, as such regressions allow estimation of one variable from a microdata record with a high degree of accuracy if the other variables are known for that record. Hence such regressions are not output by the system. This is somewhat different from the usual regression context, as a more familiar situation is one in which a high R^2 is desirable, whereas here it is seen as problematic. It may also be the case that there exists a dummy variable such that whenever that dummy variable equals 1, the corresponding observation has its response very accurately predicted by the regression. This dummy variable may arise either from a categorical predictor or from the intersection of categories from categorical predictors that are interacted with each other. Regressions with this feature will also not be provided to the user; one may think of this as a check on the local goodness of fit to complement the R^2 check on the global goodness of fit. Furthermore, output will not be given if there exists a dummy variable that assumes a value of 1 very few times in the dataset.

When categorical variables are used as predictors, the rules above can be very restrictive, especially on relatively small datasets or when categorical variables are interacted, making it potentially unlikely that the system will give the desired output. Since our goal is to provide output whenever possible, we make a slight modification to the regression in this case. This is done by increasing the lower bound m on the number of observations that a category must contain to avoid being absorbed into the reference category. By absorbing more categories into the reference category, we hope to alleviate the conditions that prevented the regression from being output. The MAS continues to increase m until either a regression is found that can safely be output—in which case that regression is fit—or m is large enough that one of the categorical predictors is reduced to having just a single level, with all other levels being absorbed into the reference level, leading the system to refuse output.

A shortcoming of our current approach is that it will sometimes combine categories in undesirable ways. Most notably, the method we have described above does not consider any ordinal structure that may be present. For example, if one predictor is a categorical variable describing highest level of education attained, it is possible that the reference category will contain those whose highest degree is a high school diploma, associate's degree or master's degree, while those whose highest degree is a bachelor's degree will be in a separate category, and this does not make intuitive sense. We hope to improve this aspect of the system in the future.

If all of the requirements for a regression are satisfied, either before or after adjusting the parameter m , then the MAS will pass output to the user. Currently, the output includes regression coefficients; their standard errors, t-statistics and P-values; the F-statistic for the regression and its P-value; the R^2 for the regression; and an ANOVA table. All of these are rounded, to thwart any attack based on exact values of regression coefficients from large numbers of regressions. For ordinary least squares (OLS) regression, the coefficients, standard errors, t-statistics and P-values are currently given to four significant digits.

The MAS also has the capability of performing logistic (either binary or multinomial) regression when the response variable is categorical, or Poisson regression when the response variable is a count, and the rules for OLS regression are adapted to the new context. Limits on interactions and the approach to categorical predictors are the same. To measure whether a regression needs to be withheld (or have m increased), we use pseudo- R^2 measures to determine whether the global goodness of fit is too good; if this is the case, then the regression will not be given or will have m increased. Our local goodness of fit measures are somewhat different. For logistic regression, the regression may be withheld (or m increased) based on observations in the dataset whose predicted probabilities of being in a particular response class (based on the model) are close to 1. A similar rule applies in Poisson regression, and we are still examining whether this provides adequate protection. As before, the rounded estimated coefficients, along with their standard errors, test statistics and P-values, are provided, as is the Analysis of Deviance table in the Poisson and binary multinomial cases.

Although regression diagnostics are useful in determining whether an OLS regression adequately describes the data, such diagnostics—especially unperturbed residual plots—pose a disclosure risk, as they allow the intruder to determine information about individual points. Therefore, the system creates residual plots based on synthetic residuals and synthetic predictor and fitted values. Our approach for OLS regression follows closely the method described in Reiter (2003). For logistic regression, whether binary or multinomial, diagnostic plots are also given, following the method in Reiter and Kohnen (2005). We also provide Q-Q plots and test statistics to evaluate the normality of residuals. The Q-Q plots are based on the synthetic data, as otherwise individual points in the plot could lead to disclosures about the data, but the nature of the data synthesis method should lead to the synthetic residuals appearing more nearly normal than the actual data. However, the test statistics are based on the actual data, so our recommendation is that users evaluate normality based on the test statistics and then use the Q-Q plots to evaluate the nature of any deviations from normality. A shortcoming of our diagnostic plots is that given the limitations on analyses that can be run in the system, there may not be a way to rectify any problems that the plots reveal.

4. Additional Features

Among the additional features being developed are histograms and scatterplots of numerical data. Each of these poses a disclosure risk if unperturbed.

4.1 Histograms

Histograms do not seem to be a major disclosure risk, except when outliers are present. The Drop q Rule already gives some protection against disclosure; we adapt the method used to create the histogram so that there is further protection. The main concern with a histogram is that it may be used to find outlier values of the variable being plotted.

We begin by removing from the distribution any extreme outliers. When this has been done, we use a kernel density function to find a smoothed estimate of the distribution of the variable. We then draw a sample from the smoothed distribution equal in size to the original dataset. Note that because of the smoothing, the bounds of the estimated density will fall beyond the bounds of the observed data, so it is possible for the histogram to extend further than the original data. However, if there is one observation that is somewhat more extreme than the others, but perhaps not so extreme as to be excluded as an outlier, it is possible that when drawing from the smoothed distribution, none of the sample will come from the area around that observation, so the histogram may also extend less far in that direction than the real data. The discreteness of the bins of the histogram also acts as a de facto perturbation of the data.

To further protect unusual values, we require that any bin in the histogram must have a minimum of three observations. Bins with fewer than three observations have more observations added to augment them to three. Bins with three observations (after the augmentation) are colored red when the histogram is plotted, whereas the other bins are colored gray.

We are still testing and modifying the method of creating histograms to ensure that it does not create a disclosure risk. We have also tested the histogram procedure to determine how well the synthetic histograms mimic the real histograms. One concern is what constitutes an “extreme outlier” above. We are omitting all points more than four

standard deviations from the mean. However, in skewed distributions, this may eliminate some points that are not outliers.

4.2 Scatterplots and Side-by-side Boxplots

We are considering a variety of approaches to the problem of making a disclosure-proof scatterplot of two numerical variables.

One approach is to use the same method as for synthetic residuals. A possible downside to this is that this method treats the two variables in an asymmetric fashion, so that a synthetic plot of y versus x need not look like a synthetic plot of x versus y . In the case of a residual plot, this asymmetry between the variables is natural, but in a more general scatterplot, we may want both variables treated similarly.

Another approach is to use a method that starts with the true scatterplot—or, if this includes too many points, a subset of the points of the scatterplot—and then moves each point a random distance in a random direction. Points that are in need of relatively little protection are moved only a small amount, whereas an outlier, even a modest one, will stand to be moved more. We are considering two variants of this approach, one of which is described in You (2010).

One potential concern in this method is with dramatic outliers, as these can be moved substantially in any direction. In a simulation of one of the methods being considered, we used a dataset of 1,001 points. The first 1,000 had X -iid $N(0,1)$ and $Y=X^2$, while the 1,001st point was an outlier, with $X=4$ and $Y=16$. This point fell in the extreme upper right corner of the plot, and was rather far from any other point, particularly in the vertical direction. As a result, it had the potential to be moved substantially. This led to perhaps unexpected results: sometimes the point remained an outlier in the upper right, with its extremity increased or decreased, but sometimes it moved to a completely different part of the plot, becoming, for example, an outlier in the upper middle area or even the upper left. Hence, if no changes are made to this method, there may be the necessity to warn the users that the presence of a point far from the others in a plot is evidence of an outlier, but may not give any useful information about where the outlier is.

Sparks *et al.* (2008) use a method of side-by-side boxplots to replace both residual plots and ordinary scatterplots, and we are also considering this approach for the MAS. When this method is used, the x variable is split into bins and a scatterplot of the y variable for each x bin is made, then the scatterplots are plotted side by side. If certain precautions are taken, such as winsorizing the data to protect outliers, disclosure risk can be minimized. Sparks *et al.* argue that in many cases, side-by-side boxplots not only have less disclosure risk than scatterplots, but also have more utility to the user.

4.3 Descriptive Statistics and Tests

The MAS computes a few basic descriptive statistics, and we expect that more will come as the system is developed. We are moving somewhat slowly on this because we want to make sure that no one can manipulate the descriptive statistics to arrive at a disclosure. However, this does not seem likely, as descriptive statistics (with the exception of quantiles) tell little about individual observations in most cases when the dataset is not very small. The statistics are suitably rounded. The MAS also runs t -tests on the mean of a variable, and provides 95% confidence intervals for the mean. We are still considering to what extent the user should be able to determine the confidence level.

5. Another Approach: The Luxembourg Income Study

Remote access systems have been developed that can sometimes provide more versatility than the MAS, but at the cost of being more difficult to maintain and protect. A notable example has been implemented by the Luxembourg Income Study (LIS), a research institute collecting data on income, wealth and various other measurements, founded in 1983 (see Luxembourg Income Study, 2009a). The LIS data are an aggregation of household surveys taken by various contributing countries. LIS's remote access system—called LISSY—allows registered users to submit their own code via email or an online form, which may be written in SAS, SPSS or Stata. Output, when deemed allowable, is returned by email and is viewable on the form. The system does not allow certain commands that could

be used to obtain a disclosure relating to an individual or household. Also prohibited are “sequences of commands and/or variables that would end up breaching the rules on data confidentiality;” these, as well as requests that give overly long output, are flagged for manual analysis or are denied outright. Further specifics are given in Luxembourg Income Study (2009b). Schouten and Cigrang (2003) also note that LIS contains an archive of jobs submitted, which can be further evaluated to make sure the data are being used properly.

6. Future Work

We will soon be testing the confidentiality rules as implemented in the MAS beta prototype to ensure that they are sufficiently strong to provide the necessary confidentiality protections.

The current system does not have any mechanisms in place to deal with missing values, which are often present in a survey such as the American Community Survey, where respondents may be willing to participate in some of the survey but may not wish to divulge the answer to every question.

Another concern is the difference between surveys and censuses. Since most of the datasets on which the MAS is run are based on surveys, extra methods must be put in place to take this into account. An advantage of using survey data is that the mere fact of sampling the population provides a substantial amount of protection to the data, since a unit that is unique in the survey in some regard may have several matches in the population as a whole. However, our methodology up until now has not considered the survey weights that are necessary to draw proper inference from an analysis, and further work will be needed to integrate this into the system and to ensure that analyses, especially tabulations, are done in such a way that the weights cannot be determined and then used to reveal sensitive information about the entities being weighted. With surveys of establishments, there is also the risk that a dominant establishment will stand out in such a way that it could lead to a disclosure. This might happen if one establishment in a table cell is much larger than the others, or to a lesser—but still possibly problematic—extent, in a regression, where such a point could be an influential point.

In addition, we plan to create a set of confidentiality rules for cross-tabulations, and to add different types of statistical analyses within the system, such as an expanded set of descriptive statistics and significance tests. We would also like to add variants on regression, such as forward and backward stepwise regression. However, this will be somewhat challenging, as the choice of variables in each step will have to be integrated with the rules on categorical predictors.

References

- Chaudhry, M. (2007), “Overview of the Microdata Analysis System”, Statistical Research Division internal report, Washington DC: U.S. Census Bureau.
- Fellegi, I.P. and S.A. Goldberg (1969), *Some Aspects of the Impact of the Computer on Official Statistics*, Ottawa: Dominion Bureau of Statistics.
- Keller-McNulty, S. and E. Unger. (1998), “A Database Prototype System for Remote Access to Information Based on Confidential Data”, *Journal of Official Statistics*, 14, pp. 347-360.
- Luxembourg Income Study (2009a), *LIS Micro-data Access*, <http://www.lisproject.org/data-access/lissy.htm>. Accessed December 24, 2011.
- Luxembourg Income Study (2009b), *LIS Micro-data Access – Job Syntax*, <http://www.lisproject.org/data-access/lissy-syntax.htm>. Accessed December 24, 2011.
- Lucero, J., Zayatz, L. and L. Singh (2009), “The Current State of the Microdata Analysis System at the Census Bureau”, *Proceedings of the American Statistical Association, Government Statistics Section*.

- Reiter, J. (2003), "Model Diagnostics for Remote Access Regression Servers", *Statistics and Computing*, 13, pp. 371-380.
- Reiter, J. and C. Kohnen (2005), "Categorical Data Regression Diagnostics for Remote Access Servers", *Journal of Statistical Computation and Simulation*, 75, pp. 889-903.
- Reznek, A. "Disclosure Risks in Cross-Section Regression Models" (2003), *Proceedings of the Section on Government Statistics, JSM*.
- Reznek, A. and T. Riggs (2004), "Disclosure Risks in Regression Models: Some Further Results", *Proceedings of the Section on Government Statistics, JSM*.
- Schouten, B. and M. Cigrang (2003), "Remote Access Systems for Statistical Analysis of Microdata", *Statistics and Computing*, 13, pp. 381-389.
- Sparks, R., Carter, C., Donnelly, J., O'Keefe, C., Duncan, J., Keighley, T. and D. McAullay (2008), "Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics®", *Computer Methods and Programs in Biomedicine*, 91, pp. 208-222.
- Steel, P. and A. Reznek. (2005), "Issues in Designing a Confidentiality Preserving Model Server", *Monographs of Official Statistics*, 9, pp. 29-36.
- Steel, P. (2006), "Design and Development of the Census Bureau's Microdata Analysis System: Work in Progress on a Constrained Regression Server", presentation at Federal Committee on Statistical Methodology Policy Seminar, Washington DC.
- You, J. (2010), "Data-Driven Quality-Preserving Methods for Synthesizing Microdata on a Remote-Access Regression Server", unpublished report, Washington DC: U.S. Census Bureau.

Providing access to microdata for statistical purposes: Experiences of the Australian Bureau of Statistics with remote analysis servers

James O. Chipperfield, Frank Yu and Melissa Gare¹

Abstract

Many national statistical offices are looking to improve their output dissemination strategy by enhancing access to microdata through the use of remote access analysis servers. In this approach, results of statistical analyses or tabulation of the data are released in a form that will not enable any microdata to be linked to individuals. The Australian Bureau of Statistics (ABS) is developing a remote access service which will enable users to submit requests for tabulation of count data and analysis outputs from statistical models, while ensuring that the confidentiality of individuals' information on the microdata is strictly maintained. This paper gives an overview of a methodology currently being considered to ensure the confidentiality of individuals' information on the microdata. This is achieved by query control and perturbation of the outputs

Key Words: Confidentiality; Remote access analysis server; Perturbation.

1. Introduction

Vast amounts of microdata are collected by agencies from Censuses, surveys and administrative sources. Such microdata can be used in the development and evaluation of policy for the benefit, or utility, of society. For this reason, there is very strong demand from analysts, within government and universities, to access such microdata. When allowing analysts access to its microdata, the agency is often legally obliged to ensure that the risk of disclosing information about a particular person or organisation is acceptably low. Managing the risk of disclosure is commonly referred to as Statistical Disclosure Control (SDC). Even after removing personal identifying information, such as name and address, from the microdata the risk of disclosure remains (see for example Willenborg and De Waal, 2001).

Methods of SDC for microdata include reducing the level of detail, replacing real values with synthetic values (see for example Reiter, 2002), sub-sampling, micro-aggregation, swapping attributes between records, and perturbing categorical values. Mathews and Harel (2011), Duncan and Pearson (1991) and give good summaries of many of these, as well as a few more.

One way of potentially improving the trade-off between utility and disclosure risk is a remote analysis server. A simple model for a remote analysis server is as follows:

1. An analyst submits a query, via the internet, to the agency's analysis server.
2. The analysis server processes the analyst's query on the sensitive microdata. The statistical output (*e.g.*, regression coefficients) from the query is modified for the purpose of SDC. Some output may be restricted on the basis that it could allow an analyst to reconstruct the attributes of an arbitrary record.
3. The analysis server sends the modified output, via the internet, to the analyst.

Remote analysis servers provide users with control over the particular outputs they want to extract from a dataset. This is a fundamental shift in the process, from the traditional paradigm where the national statistical offices would decide all the outputs that will be released to one where users can specify what they require when and as they need it. The challenge for the statistical offices is to provide SDC for the different possible outputs.

¹James, O. Chipperfield, Frank Yu and Melissa Gare, Australian Bureau of Statistics, ABS House, Belconnen, ACT 2614, Australia, james.chipperfield@abs.gov.au, frank.yu@abs.gov.au and m.gare@abs.gov.au.

Some advantages of a remote analysis server are as follows:

- Although the statistical output is modified, it is based on the real microdata. This means complex relationships in the microdata are essentially retained.
- The degree to which a particular output is modified can depend upon the output itself. For example, estimates at a broad level may require proportionally less modification than estimates at a fine, or small area, level. Since an analyst is restricted from viewing the attributes of any record, less modification is needed than would otherwise be the case.
- The impact of the modifications on the output can be broadly indicated to the analyst. If the impact is large the analyst may decide to ignore the results altogether.
- Once the server is set up, it can process multiple analyses in real time.
- All submitted programs can be logged and audited. If an audit concludes an attempt at disclosure was made, the agency can revoke the analyst's access to the server and take legal action.

There are some disadvantages of a remote analysis server:

- Some statistical outputs may be aggregated (*e.g.*, record-level residual plots may be replaced with box plots) or perturbed (*e.g.*, regression coefficients), and others may be restricted altogether.
- The analyst may be restricted to use only analysis techniques supported by the server
- Analysis through a remote server may take longer than if the microdata were available on the analyst's personal computer.

There has been some work on managing the disclosure risk of analysis and tabular output, (*i.e.*, on point (B) above). In respect to analysis output see Gomatam *et al* (2008), Lucero and Zayatz (2010), Bleninger *et al* (2010) and Sparks *et al* (2008) and in respect to tabular output see Shlomo (2007). The goal of this literature is to protect against data attacks, which involves an analyst using output from an analysis server to reconstruct attributes for one or more records which, if successful, could be used to attempt disclosure by linking to other microdata.

Section 2 outlines ABS experience and plans with respect to remote servers. Sections 3 outline the ABS' method of managing disclosure risks for count tables, and Section 4 describes the approach for protecting analysis output. Method for protecting tables of continuous measurements is being implemented and will not be discussed in this paper.

2. ABS Experience in Remotes Servers

In 2002 the ABS released the Remote Access Data laboratory (RADL). RADL is a secure online data query service that approved clients access via the ABS website. Within RADL users submit queries in the SAS, STATA or SPSS Statistical languages. These queries are run against a pre-confidentialised microdata file, referred to as a Confidentialised Unit Record File (CURF). The results of the queries are automatically checked and cleared outputs made available to users via their desktops. The underlying microdata are kept securely within the ABS environment and are not accessible for viewing. In contrast to Basic CURFs, which are disseminated on CD-ROM to approved users for use in their own computing environment, the introduction of RADL enabled the ABS to make more detail microdata available for researchers to query.

Census Table Builder (CTB), a remote server that releases frequency tables, was released in 2009 and principally designed for the 2006 Population Census. CTB utilizes the Space-Time Research (STR) SuperSTAR suite of products and incorporates the ABS dynamic perturbation confidentiality routine. This routine is run on the de-identified microdata during the generation of the requested tables and is outlined in Section 3.1 of this paper and in Fraser and Wooton, 2005.

The ABS has commenced development of a new remote server referred to as the Remote Execution Environment for Microdata (REEM). Core components of REEM will be TableBuilders, building upon the perturbation method developed for CTB and incorporating further enhancements to the SuperSTAR suite, and an Analysis Server. There are a number of drivers for this new development. The ABS faces increased demand from users for flexible access to

rich microdata about households and businesses, including a growing need for the analysis of administrative and linked datasets. This cannot be accommodated in the existing remote access service, RADL. RADL relies on a pre-confidentialised microdata file, referred to as a Confidentialised Unit Record File (CURF), and does not always provide information of sufficient detail. Users also have concerns about the impact that SDC has on analytical outputs obtained from CURFs. The viability of the existing RADL based approach is also under threat from the increasing risk of identification due to increased computing power (both hardware and software) and the proliferation of detailed external datasets.

There are other good reasons for developing a remote server as a replacement for RADL. First, it supports the ABS' objective to build capability and continually improve its effectiveness. Specifically, there are a number of inefficiencies in the assessment and production of CURFs for RADL. It is envisaged that replacing RADL with REEM will significantly reduce or eliminate the staff time required for either of these activities. In particular, REEM will incorporate confidentiality routines that are tailored to each type of analysis supported by REEM. These routines will be designed to run dynamically on the de-identified microdata, removing the need to assess and produce CURFs to underlie the future remote server.

Second, a driver for the project is to increase accessibility of ABS outputs. REEM will adopt internationally recognised metadata standards, including the use of DDI/SDMX and machine to machine interfaces (APIs), to support easy discovery of data items and dissemination of output through web services.

The initial version of TableBuilder for weighted social survey microdata, Survey TableBuilder, was released in December 2011 for a restricted set of microdata. A trial version of the Analysis Server is planned for restricted release in mid 2012.

3. Count Data

Cells in a table are either 'internal' or 'marginal.' The count for a marginal cell is a sum of two or more other counts appearing in the table. If a cell is not a marginal cell it must be an internal cell. We now describe the ABS' method for perturbing unweighted (section 3.1) and weighted counts (section 3.2) for internal and marginal cells of a table.

3.1 Census Table Builder

Here we describe the method of perturbing unweighted counts as implemented in Census Table Builder (CTB), an ABS remote server which allows analysts to remotely request contingency tables to be calculated from the Australian Census' microdata. The perturbed tables are automatically returned to the analyst, with generally no intervention from ABS staff. The analyst can define the dimensions of the table and the attributes of the records contributing to the table with only limited restriction (only tables with a high percentage of cells with counts of 0 or 1 are not released).

Denote the i^{th} unweighted sample count for an *internal* cell in a contingency table by $n_i = \sum_{j=1}^n \delta_{ij}$, where $i=1, \dots, C$, $\delta_{ij}=1$ if the j^{th} record on the microdata belongs to the i^{th} cell and $\delta_{ij}=0$ otherwise, $j=1, 2, \dots, n$ and $n = \sum_{i=1}^n n_i$. CTB releases n_i^* to the analyst instead of n_i , where

$$n_i^* = n_i + e_i^* + a_i^*,$$

$n_i^* \geq 0$, $|e_i^*| \leq L_e$, $|a_i^*| \leq L_a$, and L_e and L_a are positive integers specified by the agency. Clearly, the difference between n_i and n_i^* is restricted to be less than $L = L_a + L_e$. The e_i^* s represent the random integer perturbation of the i^{th} cell count. The a_i^* s are derived so that the internal and marginal counts are consistent and so that the changes to the marginal counts are bound (for details see Appendix).

Define $Var_*(\cdot)$ and $E_*(\cdot)$ to be the variance and expectation with respect to the perturbation distribution of e_i^* , which meets the following criteria:

- a) $E_*(e_i^*) = 0$
- b) $Var_*(e_i^*) = \sigma^2$
- c) $Cov_*(e_i^*, e_j^*) = 0$ if $i \neq j$
- d) whenever the same set of records contribute to a cell count, the value for e_i will always be the same (see Fraser and Wooton, 2005).
- e) e_i^* is an integer

Criterion a) ensures the count data are unbiased over the perturbation distribution. Criterion b) means that any cell count has a fixed perturbation variance. Criterion c) ensures that differencing two cells counts does not remove the effect of perturbation. Criterion d) ensures the effect of perturbation is not removed by repeatedly requesting the same cell count.

Table 1 gives an illustrative example of tabular counts before and after perturbation. Perturbed counts are asterisked while original counts are not. For example, a true count of 1 is perturbed to 3.

Table 1
Example of tabular counts before and after perturbation

	Treatment A				Treatment B			
	<i>Success</i>	<i>Trials</i>	<i>Success*</i>	<i>Trials*</i>	<i>Success</i>	<i>Trials</i>	<i>Success*</i>	<i>Trials*</i>
Clinic 1	1	5	3	6	10	20	9	17
Clinic 2	9	10	9	11	5	20	4	18
Totals	10	15	12	17	15	40	13	35

* Perturbed counts

3.2 Survey Table Builder

Survey Table Builder (STB) applies SDC to survey-weighted count data. Denote the i^{th} weighted count in a contingency table by $N_i = \sum_j d_j \delta_{ij}$, where d_j is the survey weight for the j^{th} record. The corresponding perturbed count is $N_i^* = \lceil \tilde{d}_i n_i^* \rceil + A_i^*$, where $\tilde{d}_i = n_i^{-1} N_i$ is the average weight for records belonging to the i^{th} cell, n_i^* is the perturbed sample count described previously, $\lceil x \rceil$ rounds x to the nearest integer, and A_i^* performs an analogous function to a_i^* but for weighted counts (for details see Appendix). STB will not release any information about \tilde{d}_i , e_i^* , n_i^* or N_i to the analyst. If $\tilde{d}_i = 1$ for all i , then the CTB and STB methods of SDC are equivalent. Marley and Leaver (2011) studied the measures of risk and utility associated with STB.

4. Analysis Server

4.1 Without Statistical Disclosure Control (Standard Case)

First we consider the standard case for estimating regression coefficients in a regression model. Consider microdata from which an analyst specifies an outcome variable y and K covariates \mathbf{x} , where the data are $\mathbf{d} = \{(y_j, \mathbf{x}_j) : j = 1, \dots, n\}$. Consider fitting a regression model with parameter $\boldsymbol{\beta}$ using an unbiased estimating function $H(\boldsymbol{\beta})$ (see Chambers and Skinner, 2003). In particular we consider the estimating equation

$$H(\boldsymbol{\beta}) = \sum_{i=1}^n G_j(\boldsymbol{\beta}) \{y_j - f_j(\boldsymbol{\beta})\},$$

where $f_j(\boldsymbol{\beta}) = E(y_j | x_j)$ and $G_j(\boldsymbol{\beta})$ is a vector of order K with k^{th} element $G_{jk}(\hat{\boldsymbol{\beta}})$ which is a function of $\boldsymbol{\beta}$ and x_j but not of y_j . The solution to $H(\boldsymbol{\beta}) = \mathbf{0}$ gives the standard estimate, $\hat{\boldsymbol{\beta}}$, of the regression coefficients.

Data attacks involve obtaining $\hat{\theta}$ from one or more queries in order to reconstruct attributes for an individual record. These attacks can involve differencing, leveraging a single record, isolating a record with a covariate, and by making inferences from a highly accurate model. These are well discussed for example by Gomatam (2008). Data attacks can of course use other outputs, such as plots, diagnostic statistics, p-values in a data attack.

When designing a set of perturbations and restrictions to apply to a set of analysis output, it quickly becomes clear that a series of regressions designed to find the optimal model could be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter.

4.2 With Statistical Disclosure Control

Below we discuss the approach ABS is considering to implement in its remote analysis server.

4.2.1 Estimation of Parameters

Instead of solving $H(\boldsymbol{\beta}) = \mathbf{0}$ and releasing $\hat{\boldsymbol{\beta}}$, the server solves

$$H(\boldsymbol{\beta}) = \mathbf{E}^* \tag{1}$$

and releases the resulting estimator $\hat{\boldsymbol{\beta}}^*$, where $\mathbf{E}^* = (E_1^*, E_k^*, \dots, E_k^*)'$ are perturbations introduced for the purpose of SDC, $E_k^* = u_k^* e_k$, u_k^* is the uniform distribution on the range $(-1,1)$, and $e_k = \max_j \{G_{jk}(\hat{\boldsymbol{\beta}})(y_j - f_j(\hat{\boldsymbol{\beta}}))\}$ is the maximum influence a record may have on the k^{th} estimating equation. For example, for the case of binary variables and the logistic model $e_k = 1$. The distribution of the perturbations, E_k^* , are independent and if the same model is fitted the same value of \mathbf{E}^* is used- this stops an analyst estimating $\hat{\boldsymbol{\beta}}$ by fitting the same model a number of times and averaging over the regression parameters obtained from solving (1).

The size of the perturbation is designed to be of sufficient size to mask the contribution of any record to the estimating equation. Applying the perturbation to the score function is important, since this is where $\hat{\boldsymbol{\beta}}$ imposes a constraint on the data values.

4.2.2 Inference

To make valid inference with $\hat{\boldsymbol{\beta}}^*$ an analyst will need to account for the variance from both the model and the perturbation of the estimating equation. The variance of $\hat{\boldsymbol{\beta}}^*$ is

$$\mathbf{V}_{m^*}(\hat{\boldsymbol{\beta}}^*) = V_m(\hat{\boldsymbol{\beta}}) + V_*(\hat{\boldsymbol{\beta}}^*)$$

where $V_m(\hat{\boldsymbol{\beta}})$ is the variance of $\hat{\boldsymbol{\beta}}$ due to the model (i.e. the absence of any perturbation) and $V_*(\hat{\boldsymbol{\beta}}^*)$ is the variance of $\hat{\boldsymbol{\beta}}^*$ due to the perturbation. We propose estimating $V_m(\hat{\boldsymbol{\beta}})$ using the delete-a-group Jackknife (Rao and Wu, 1988). A benefit of the Jackknife is that it is simple to calculate and is unbiased when the microdata have been collected from a sample with a complex design (e.g., clustered sampling), as is the case for many ABS surveys. The Jackknife method involves allocating all selection units to one and only one replicate group in the same way that the

sample was selected from the population. Using a similar approach to the sandwich variance estimator (see Chambers and Skinner, 2003 pp.105), we derive $V_*(\hat{\beta}^*) = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{D}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$, where $\mathbf{D} = \text{Var}_*(\mathbf{E}^*)$.

We argue that the uncertainty in the Jackknife variance estimator (see pp.196 Shao and Tu, 1996), due to the allocation of selection units to replicate groups, is such that the total variance, $\mathbf{V}_{m^*}(\hat{\beta}^*)$, cannot be used in a data attack.

4.2.3 General Restrictions

Several authors have noted that fixed-distribution perturbation (as used above) alone is not sufficient to protect analysis outputs in the context of multiple queries. Approaches to managing the additional risks have included imposing restrictions into the analysis server (see Gomatam *et al.* 2005; Sparks *et al.* 2008) On the other hand, when designing a set of restrictions to manage disclosure risk, it quickly becomes clear that a series of regressions designed to find the optimal model could be indistinguishable from a sophisticated data attack. Therein lies the challenge: not restricting the former while thwarting the latter. In this subsection, we mention a set of restrictions that do not defend against a particular data attack, but are designed to significantly hinder a data attacker while only making a minor reduction in utility. These general restrictions include:

- $n > 50$.
- $n/K > 10$
- $K > 5$
- models can be fitted to a subset of records, where the subset is defined by at most 4 (always less than K)
- binary variables originally on the microdata
- new binary variables can only be created from two other binary variables that are originally on the microdata.
- new continuous variables can be only be created by using certain transformations
- variables must be non-zero for at least 15 records.
- for models with only binary covariates, the number of covariate patterns in \mathbf{x} must be greater than 50
- $\mathbf{X}'\mathbf{X}$ must be full rank, where $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_n)'$ and \mathbf{x}_j is the K column vector of covariates for the j^{th} record.

The values (*e.g.*,50) used in the above restrictions are used for illustration and can of course be changed.

4.2.4 Additional Attack-Specific Restrictions

As mentioned above, there are some well documented data-attacks (see for example, Gomatam 2008). It makes sense to impose restrictions, in addition to those mentioned in section 3.2.3, to explicitly defend against them. These restrictions are not discussed here, for space, but can be found in Chipperfield and O'Keefe (2011). We do, however, briefly describe three attacks for which explicit defences are constructed.

One such attack is called a *differencing attack*. A differencing data attack involves fitting the same model to two sets of records that are identical except that one record is dropped from one of the sets. Differences in the regression coefficients from the two models could be used in an attempt to reconstruct attributes of the dropped record. For example, if the covariates of the dropped record are known to the attacker, the change in the regression coefficients would allow a binary outcome variable for the dropped record to be derived.

Another such attack involves *fitting different models to the same set of records* and their attributes (*i.e.*, the same data set) by:

1. Swapping the choice for the outcome variable
2. Using a different link function (*e.g.*, linear, logistic and probit)
3. Using variables that are different transformations of the same attributes

Each model imposes K constraints on a set of records' attributes, which are unknown to the analyst. The aim of this attack is to impose enough constraints so that it is possible to solve for the values in the underlying data set.

4.2.5 Diagnostics

A range of test statistics (see Hosmer and Lemeshow, 2000) are available to assess the model assumptions (*e.g.*, normality of residuals) and model fit (*e.g.*, AIC, R-squared). Again, when releasing such statistics the agency needs to balance the disclosure risk against the utility. Ideally, an analyst's model selection should not be influenced by statistical disclosure control.

The approach to SDC for the estimate of the dispersion parameter or diagnostic statistics closely follows that for regression coefficients. Denote such a parameter or statistic by $t^* = t(\hat{\beta}^*, \mathbf{d})$. Instead of releasing $t^* = t(\hat{\beta}^*, \mathbf{d})$ we release,

$$t^{**} = t^* + u^* s(\hat{\beta}^*, \mathbf{d})$$

where u^* is a random variable on the range (-1,1) and $s(\hat{\beta}^*, \mathbf{d})$ bounds the maximum influence that a single record in \mathbf{d} can have on the statistic t^* given $\hat{\beta}^*$.

Diagnostics that involve plotting individual record values (*e.g.*, residual plots) will be aggregated in some way, following Sparks *et al.* (2008). For example, Q-Q plots will be replaced by a smoothed non-parametric regression line and residual plots will be replaced by parallel box plots. Residual plots will be replaced by parallel bar charts.

Appendix

Denote the internal and marginal cells of a table by $t=1, 2, \dots, C, C+1, \dots, T$, where $t=1, 2, \dots, C$ denotes the internal cells of the table. Denote the t^{th} cell count by n_t . Instead of releasing n_t , TB releases $n_t^* = n_t + e_t^* + a_t^*$ which is obtained in two steps. The first step involves calculating the preliminary counts $m_t^* = n_t + e_t^*$, where e_t^* has properties a)-e) from section 2.2. The table's preliminary counts are not consistent: sums of preliminary counts for internal cells are not guaranteed to equal corresponding preliminary marginal counts. The second step involves finding the value for a_t^* that so that The table with counts n_t^* is consistent and $|a_t^*| \leq L_a$ for all $t=1, \dots, L$. This means no preliminary count, for a marginal or internal cell, is changed by more than L_a .

References

- Bleninger, P., Drechsler, J. and G. Ronning (2010), "Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study", *Privacy in Statistical Databases*, Springer.
- Chambers, R.L. and C.J. Skinner (2003), *Analysis of Survey Data*, John Wiley & Sons.
- Chipperfield, J.O. and M.C. O'Keefe (2011), "Disclosure-Protected Inference using Generalised Linear Models", unpublished report, Canberra, Australia: Australian Bureau of Statistics.
- Fraser, B. and J. Wooton (2005), "A proposed method for confidentialising tabular output to protect against differencing", UNECE work session on Statistical Data Confidentiality.
- Gomatam, S., Karr, A.F., Reiter, J.P. and A.P. Sanil (2008), "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk –Utility Framework for Remote Access Analysis Servers", *Statistical Science*, 20, pp.163-177.
- Hosmer, D.W. and S. Lemeshow (2000), *Applied Regression Analysis*, John Wiley and Sons.
- Little, R.J.A. (1993), "Statistical Analysis of Masked Data", *Journal of Official Statistics*, 2, 407-426.

- Marley, J.K and V.L. Leaver (2011), “A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis”, Proceedings of the International Statistics Institute.
- Mathews, G.J. and O. Harel (2011), “Data Confidentiality: A Review of methods for statistical disclosure limitation and methods for assessing privacy”, *Statistical Surveys*, 5, pp. 1-29.
- Rao, J.N.K. and C.F.J. Wu (1988), “Resampling Inference with Complex Survey Data”, *Journal of the American Statistical Association*, 83, pp. 231–241.
- Reiter, J.P. (2002), “Satisfying Disclosure Restrictions with Synthetic Data Sets”, *Journal of Official Statistics*, 18, 531-543.
- Shao, J. and D. Tu (1996), *The Jackknife and Bootstrap*, Springer.
- Shlomo, N. and C. Skinner (2010), “Assessing the Protections provided by Missclassification-based Disclosure Limitation”, *The Annals of Applied Statistics*, 1291-1310.
- Sparks, R., Carter, C. Donnelly, J., O’Keefe, C.M., Duncan, J., Keighley, T. and D. McAullay (2008), “Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics™”, *Computer Methods and Programs in Biomedicine* 91, pp. 208-222.
- Willenburg, L. and T. de Waal (2000), *Elements of Disclosure Control*, Springer.

SESSION 7B
CONTENT AND COLLECTION

Some implications of standardizing methods for quality monitoring of survey interviewing

Doug Currivan, Derek Stone, Kristin Fuller, Susan Kinsey and Howard Speizer¹

Abstract

Standardizing methods and tools for evaluating the quality of survey interviewing across modes and studies has increasingly been an important goal for many survey organizations. RTI has developed a standardized, mode-independent interview quality monitoring evaluation system, QUEST. This system allows in-person and telephone interviewing behaviors to be evaluated using a common set of quality metrics that are stored in a single shared database. The system supports evaluation of interviewing quality for both live monitoring in real time and review of computer audio-recorded interview (CARI) files. QUEST replaces a varied set of RTI quality monitoring processes and tools used across a wide range of in-person and telephone interviewing projects and allows for interviewer quality data to be evaluated over time across surveys and modes. This paper addresses some important methodological issues resulting from implementing standardized quality monitoring processes via QUEST. Specifically, we examine (1) demands to tailor QUEST protocols to address alternative interviewing techniques and specialized interviewing tasks, (2) the impact of multiple protocol changes on monitoring outcomes such as performance blocks used and errors detected, and (3) initial findings from evaluating monitor variability in detecting errors across QUEST monitoring sessions. The paper discusses the next steps planned to enhance these procedures and tools for continued improvements to this standardized quality monitoring effort.

Key Words: Survey Interviewing; Quality Monitoring; Monitor Variability.

1. Background and Introduction

Nearly twenty years ago, Couper, Holland, and Groves (1992) noted that monitoring protocols often (1) followed unsystematic and subjective procedures and (2) included only general impressions of telephone interactions, rather than objective measures of behavior. In recent years, standardizing methods and tools for evaluating the quality of survey interviewing across modes and studies has increasingly been an important goal for survey organizations. RTI has developed a standardized, mode-independent interview quality monitoring evaluation system, QUEST (Speizer, *et al.* 2009; Speizer, *et al.* 2010). This system allows in-person and telephone interviewing behaviors to be evaluated using a common set of quality metrics that are stored in a single shared database. The system supports evaluation of interviewing quality for both live monitoring in real time and review of computer audio-recorded interview (CARI) files.

QUEST provides a standardized system for monitoring interviewer performance with respect to the authenticity of completed interviews and appropriateness of data collection and interview administration protocols. Specifically, this system supports interviewing quality through (1) a standard set of interviewer skills/behaviors for evaluating field and telephone interviewing; (2) a common evaluation form, scoring rubric, performance tracking, and feedback process; (3) increased use of CARI to improve interviewer feedback and track survey item performance (Biemer, Herget, Morton, and Willis, 2000; Thissen, *et al.* 2008); (4) organized interview and interviewer performance data across surveys to monitor and improve quality, and (5) increased efficiency to control quality monitoring costs.

Monitoring interviews using QUEST involves use of a standardized evaluation form to record observed instances of incorrect or inappropriate interviewer behaviors. The form also allows monitors to note exceptional behaviors observed that contributed positively to data quality. The evaluation form combines interviewer behaviors into 12 “performance blocks” reflecting specific skill sets or interviewing outcomes. For each monitoring session, QUEST generates an overall score and a block score for each block considered by the monitor. Scores are based on the number and the “criticality” of observed errors. Criticality refers to the severity of an error and, therefore has implications for the overall score for each session. All QUEST items are defined as either non critical, critical, or

¹RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709.

extremely critical. The system then uses the block scores and total number of observed errors to generate one of three overall session scores, “exceeded expectations,” “met expectations,” or “did not meet expectations.”

This paper addresses three important methodological issues resulting from RTI’s efforts to implement standardized quality monitoring processes: (1) the need to tailor QUEST to address alternative interviewing techniques and specialized interviewing tasks, (2) the impact of multiple protocol changes on monitoring outcomes such as skill blocks use and errors detected, and (3) initial findings from evaluating monitor variability in detecting errors across QUEST monitoring sessions. Although the QUEST system is used for both field (in-person) and telephone data collection efforts, this paper focuses only on phone data collection conducted in RTI’s call center. The following sections describe the rationale and outcomes for each of these three issues, as well as future work to enhance procedures and tools for continued improvements to this standardized quality monitoring effort.

2. Tailoring the System for Alternative Techniques and Specialized Tasks

This section describes recent efforts to tailor QUEST to address alternative interviewing techniques and specialized interviewing tasks that were not previously accommodated by the system. The QUEST forms and scoring algorithm were designed to accommodate the creation and implementation of additional or alternative performance blocks as deemed necessary across different data collection projects. Core blocks and items are used as much as possible to standardize quality monitoring, but the system is flexible enough to accommodate special project needs.

2.1 Addressing Non-standardized Interviewing Techniques

The majority of interviewer-administered surveys at RTI are designed to follow conventional standardized interviewing procedures (Fowler and Mangione, 1990). Initial development of QUEST focused on evaluation of interviewing quality for standardized protocols. Adjustments were required to accommodate a more “conversational” form of interviewing implemented for selected studies (Conrad and Schober, 2000). These adjustments involved the creation of an alternate *Reading Skills* block and implementation of a requirement for projects to specify whether the “conventional” or “conversational” *Reading Skills* block should be used. Items for proper articulation and pronunciation remain in the conversational block, but items focused on standardized administration were replaced with items for conversational techniques. *Figure 2.1.1* shows these two blocks.

Figure 2.1.1
QUEST Reading Skills Blocks for Conventional Standardized Interviewing versus Conversational Interviewing

READING SKILLS – CONVENTIONAL INTERVIEWING	READING SKILLS – CONVERSATIONAL INTERVIEWING
Articulation unclear Pronunciation incorrect Major unscripted word/phrase added Minor unscripted word/phrase added Major word/phrase omitted Minor word/phrase omitted Response categories not read when required Entire question/instruction omitted	Articulation unclear Pronunciation incorrect Conversational interviewing not used or used incorrectly Key question element omitted Inappropriate/over-use of paraphrasing Entire question/instruction omitted Did not use proper grammar

2.2 Addressing Specialized Interviewing Tasks

Although the primary purpose of QUEST was to facilitate quality monitoring of standardized interviewer-administered surveys, for some studies interviewers engage in additional tasks that require quality monitoring. One such task is recontacting sample members in an effort to retrieve data missing from their interviews. The resulting adjustment involved the creation of an additional block, *Recontact and Follow-up/Missing Data Retrieval Items*. This optional block is shown in *Figure 2.2.1*.

A second specialized interviewing task in RTI's call center is providing help desk support for data collection efforts. Help desk activities are unique in that the interviewer's primary role is one of providing assistance to sampled individuals or institutions instead of administering an interviewer. Help desk tasks present unique issues that the original QUEST evaluation form and criteria did not previously address. For this reason, additional evaluation items were created to be added to multiple QUEST blocks for projects involving help desk tasks. *Figure 2.2.2* shows the items added to facilitate evaluation of help desk tasks.

Figure 2.2.1
QUEST Block for Recontact and Missing Data Retrieval

RECONTACT AND FOLLOW-UP/MISSING DATA RETRIEVAL ITEMS
Coded POC information incorrectly
Reason for follow-up incorrect
Did not collect missing data items accurately
Missing data items explanation incorrect
Fax/mail procedures administered incorrectly
Action Due prompt followed incorrectly

Figure 2.2.2
Additional QUEST Items for Help Desk Tasks

READING SKILLS - CONVERSATIONAL INTERVIEWING
Did not use proper grammar
FEEDBACK SKILLS
Gave inappropriate solution
PRESENTATION SKILLS
Wait time not communicated
PROFESSIONAL BEHAVIOR
Did not empathize with caller's concerns
INTERVIEW PROTOCOL
Expertise in study content not shown/did not demonstrate knowledge of study
Call not documented in thorough/timely manner
Additional assistance not offered

3. The Impact of Multiple Protocol Changes on Monitoring Outcomes

On April 1, 2011, two protocol changes were made to QUEST. First, the sampling interval was reduced from 15 to 12 minutes for monitoring sessions where an interview is being observed. Second, new items and scoring criteria were added to the QUEST evaluation form, including redefining the criticality of specific interviewer behaviors. In addition, debriefing sessions were held with monitoring staff to address issues such as when specific QUEST performance blocks should be considered in monitoring sessions. The goal of the protocol changes and monitor debriefing sessions was to improve the amount and quality of monitoring data being collected in RTI's call center. To assess the potential impact of the two protocol changes and the monitor debriefing sessions on QUEST outcomes, we examined (1) the average proportion of performance blocks considered in monitoring sessions and (2) average error rates observed for three specific performance blocks.

Table 3.1
Frequency of Performance Blocks Not Considered Before and After April 1, 2011

Performance Block	Sessions Before April 1		Sessions April 1 and After	
	Live (n=54,528)	Recorded (n=264)	Live (n=13,191)	Recorded (n=112)
Case management	13.8%	68.6%	9.3%	50.0%
Initial contact	55.1%	37.9%	48.4%	30.4%
Keying skills	56.6%	73.5%	52.2%	50.9%
Reading skills	60.1%	42.1%	51.0%	13.4%
Probing skills	67.3%	48.1%	57.9%	17.9%
Feedback skills	70.6%	58.3%	60.0%	24.1%
Presentation skills	57.5%	39.0%	49.4%	9.8%
Professional behavior	55.8%	39.0%	46.3%	9.8%
Interview protocol	73.2%	65.2%	58.1%	22.3%

Table 3.1 presents the frequencies of sessions in which specific QUEST blocks were not considered prior to April 1, 2011 and April 1 and beyond, for both live and recorded monitoring sessions. Overall, these data indicate monitors scored an increased proportion of blocks across both live and recorded monitoring sessions following the protocol changes on April 1, 2011. Although this finding generally appeared to hold for both live and recorded sessions, the number of recorded sessions for the two time frames was quite limited (n=376).

Table 3.2 presents the average rates of error detection for three selected performance blocks prior to April 1, 2011 and April 1 and beyond, for both live and recorded monitoring sessions. For both live and recorded sessions, the error detection rates increased slightly for all three performance blocks. The one exception was that no case management errors were observed in the 112 recorded sessions conducted on April 1 and after. In comparing pre- and post-April 1, 2011 results, we were not able to isolate exactly how the two protocol changes or monitoring debriefing sessions might have affected the rates at which monitors detected specific errors.

Table 3.2
Error Rates for Three Performance Blocks Before and After April 1, 2011

Performance Block	Sessions Before April 1		Sessions April 1 and After	
	Live (n=54,528)	Recorded (n=264)	Live (n=13,191)	Recorded (n=112)
Case Management	0.013	0.023	0.015	0.000
Reading skills	0.009	0.030	0.011	0.089
Probing skills	0.005	0.042	0.009	0.169

4. Initial Findings from Evaluation of Variability among Monitors in Detecting Errors

In theory, the highly standardized QUEST methods and tools should promote high consistency among monitors evaluating live or recorded sessions. The focus of QUEST on collecting objective indicators of interviewer behavior, as opposed to more subjective impressions of interviewing quality, would seem to support this hypothesis. Various approaches can be taken to examining inter-rater agreement, depending on the purpose of the evaluation (Hicks, *et al.* 2010). Our initial assessment of inter-rater reliability in QUEST focused on overall consistency among monitors in detecting errors and error detection in specific performance blocks. These data allowed us to examine (1) overall variability across monitors in terms of their likelihood of observing errors and (2) whether specific monitors appear to detect errors significantly more or less frequently than average rates. The analysis of monitors' variability assumes call center monitors received a random assignment of monitoring sessions within a data collection project, which appears to be a generally sound assumption based on current RTI call center procedures.

To assess monitor variability we organized session data on errors observed for all monitors and by individual monitors for three performance blocks on a single project across three months. Organizing the data by months

supported the assumption that the set of monitoring sessions was generally randomized among monitors, as opposed to shorter time frames where randomization of sessions could be limited for at least some monitors. These data allowed us to determine the mean proportion of sessions where any errors were observed and the mean proportion of sessions where any errors were observed for each of the three blocks. We then examined how the proportions for each monitor compared to the mean and standard deviation for all sessions conducted by these monitors.

Table 4.1
Variability across Monitors in Sessions where One or More Errors were Observed (May 2011)

Category in Relation to Average Proportion	Number of Monitors	Proportion of Monitors
Two standard deviation or more above mean proportion	2	15.4%
Within two standard deviations (above or below) mean proportion	11	85.6%
Two standard deviations or more below mean proportion	0	0.0%
TOTALS	13	100%

Tables 4.1 through *4.3* show the number and proportion of monitors grouped in relation to the mean proportion of all monitoring sessions in which any errors were observed for the months of May, June, and July 2011. Overall, variability among monitors in detecting any errors for this project and period appeared to be low. For all three months, only 1 or 2 monitors were two standard deviations or more above the mean proportion for all monitors. No monitors were two standard deviations or more above the mean proportion in any of the three months. For these three months, monitors appeared to be consistent in the overall likelihood of detecting errors across sessions.

Table 4.2
Variability across Monitors in Sessions where One or More Errors were Observed (June 2011)

Category in Relation to Average Proportion	Number of Monitors	Proportion of Monitors
Two standard deviation or more above mean proportion	1	10.0%
Within two standard deviations (above or below) mean proportion	9	90.0%
Two standard deviations or more below mean proportion	0	0.0%
TOTALS	10	100%

Table 4.3
Variability across Monitors in Sessions where One or More Errors were Observed (July 2011)

Category in Relation to Average Proportion	Number of Monitors	Proportion of Monitors
Two standard deviation or more above mean proportion	1	8.3%
Within two standard deviations (above or below) mean proportion	11	91.7%
Two standard deviations or more below mean proportion	0	0.0%
TOTALS	12	100.0%

Tables 4.4 through **4.6** provide similar data on error detection for three specific QUEST performance blocks, Reading Skills, Probing Skills, and Case Management for May, June, and July 2011. These tables show the mean proportion of sessions in which any errors were detected for these three blocks in the second column. The third and fourth columns indicate the number of monitors whose error rate were either two or more standard deviations the mean or two or more standard deviations below the mean. Similar to the overall error detection rates for these three months shown in **Tables 4.1** through **4.3**, one or two monitors appeared to be more likely than the norm to detect interviewer errors within each of these performance blocks. Beyond these instances, monitors appeared to be rather consistent in detecting errors within these three blocks.

Table 4.4
Variability across Monitors in Sessions in which One or More Errors were Observed for Selected Blocks (May 2011)

Performance Block	Mean Proportion of Sessions with Any Errors	Number of Monitors Two or More SDs Above Mean	Number of Monitors Two or More SDs Below Mean
Reading skills	0.017	1	0
Probing skills	0.007	2	0
Case Management	0.006	1	0

Table 4.5
Variability across Monitors in Sessions in which One or More Errors were Observed for Selected Blocks (June 2011)

Performance Block	Mean Proportion of Sessions with Any Errors	Number of Monitors Two or More SDs Above Mean	Number of Monitors Two or More SDs Below Mean
Reading skills	0.009	2	0
Probing skills	0.006	1	0
Case Management	0.011	1	0

Table 4.6
Variability across Monitors in Sessions in which One or More Errors were Observed for Selected Blocks (July 2011)

Performance Block	Mean Proportion of Sessions with Any Errors	Number of Monitors Two or More SDs Above Mean	Number of Monitors Two or More SDs Below Mean
Reading skills	0.015	1	0
Probing skills	0.012	2	0
Case Management	0.013	1	0

5. Conclusions and Next Steps

QUEST was designed to provide standardization of quality monitoring. Over time, the needs of various data collection efforts have required enhancements to the system to accommodate alternative procedures and special tasks. In addition, multiple protocol changes have been periodically required to improve the amount and quality of monitoring data being collected. Without being able to isolate the impact of specific protocol changes, QUEST monitoring results indicated a higher proportion of performance blocks being considered across sessions and slightly increased error detection rates for three performance blocks following implementation of these changes. Initial data on monitor variability in evaluating sessions showed fairly high consistency across monitors in overall errors detected and errors detected for three specific blocks. Moving forward, RTI plans to maintain standardized protocols for quality monitoring of field and telephone interviewing, while accommodating special techniques and tasks as needed. We plan to continue to analyze potential effects on QUEST outcomes when significant protocol changes are made. Further evaluation of monitor variability in detecting errors will include analyzing monitor variability over longer time periods and assessing error-level agreement among monitors for specific recorded scenarios.

Acknowledgements

The authors thank additional current members of the QUEST team: Richard Heman-Ackah, Sridevi Sattaluri, Curry Spain, Dave Foster, Melissa Cominole, and Nicole Tate. We also gratefully acknowledge the contributions of former team members Rita Thissen, Orin Day, Mai Nguyen, Mary Allen, and Courtney Gainey.

References

- Biemer, P., Herget, D., Morton, J. and W.G. Willis (2000), "The feasibility of monitoring field interview performance using computer audio recorded interviewing (CARI)", in *Proceedings of the American Statistical Association's Section on Survey Research Methods*, pp. 1068-1073.
- Conrad, F. and M. Schober (2000), "Clarifying question meaning in a household telephone survey", *Public Opinion Quarterly*, 64, 1-28.
- Couper, M., Holland, L. and R. Groves (1992), "Developing systematic procedures for monitoring in a centralized telephone facility", *Journal of Official Statistics*, 8, 63-76.
- Fowler, F.J. and T. Mangione. (1990), *Standardized Survey Interviewing: Minimizing Interviewer-related Error*, Sage: Newbury Park, CA.
- Hicks, W., Edwards, B., Tourangeau, K., McBride, B., Harris-Kotejin, L. and A. Moss (2010), "Using CARI Tools to Understand Measurement Error", *Public Opinion Quarterly*, 74, 985-1003.
- Speizer, H., Kinsey, S., Heman-Ackah, R. and R. Thissen (2009), "Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance", presented at *Federal Committee on Statistical Methodology Research Conference*, Washington, D.C.
- Speizer, H., Currivan, D., Heman-Ackah, R. and S. Kinsey (2010), "Developing a common, mode-independent, approach for evaluating interview quality and interviewer performance: lessons learned", presented at the *American Association for Public Opinion Research Annual Conference*, Chicago, IL.
- Thissen, M.R., Sattaluri, S., McFarlane, E. and P. Biemer (2008), "The Evolution of Audio Recording in Field Surveys", *Survey Practice*. <http://surveypractice.org/2008/12/19/audio-recording/> (accessed 2 February 2010).

European Health Examination Survey: From a sampling and recruitment perspective

Johan Heldal¹, Susie Jentoft¹,
Kari Kuulasmaa² Päivikki Koponen² and Sanna Ahonen²

Abstract

In 2009, the European Health Examination Survey (EHES) project was launched with support from the European Commission to collect comparable, high quality data on health and health risks of the European adult population. The survey includes an interview and core physical measurements. Individual countries can include additional content. The purpose of the EHES is to provide data for national and Europe wide planning and evaluation of health policies, health promotion and research.

The EHES Reference Centre coordinates the activity and establishes common standards for all aspects of data collection. Twelve countries have participated in a pilot phase. Five of them have carried out full-size surveys and the others are prepared to start their national surveys.

Establishing common standards that can be applied to create comparable surveys in 30 countries with varying availability of sampling frames, cultures and legislation is challenging. The paper focuses on the variation in sampling, recruitment and survey response among European countries and the approaches to deal with them.

Key Words: Sampling designs; Examination Survey; Participation rates; Pilot surveys; Full size surveys.

1. Introduction

The history of health examination surveys in Europe goes back to the 1960s when Finland and Sweden carried out their first national surveys. Later, several countries have carried out such surveys. However, all these surveys have been done with a national or regional scope with their own protocols and without standardization or coordination across national borders. Many of them were nationally representative sample based surveys (EHRM 2002, FEHES 2008a). The World Health Organization MONICA study where several European countries took part used statistical sampling but the samples were not nationally representative (MONICA 2003).

Section 2 describes the organization of the EHES project. Section 3 gives a brief overview of the recommendations from EHES Reference Center (RC). Section 4 describes the experiences from the pilot surveys and from full size surveys carried out so far. Section 5 presents conclusions.

2. Organization

The European Health Examination Survey (EHES) Pilot Project was launched in 2009 for a two year period as part of European Union's Health Programme. The goal for the project was to *develop and plan a pilot European Health Examination Survey in the European Union and the European Free Trade Association/Agreement of the European Economic Area member states in preparation to test examination modules and field procedures for this survey*. The EHES Pilot Project is coordinated by the EHES RC, established jointly by the National Institute for Health and Welfare of Finland, Statistics Norway (statistical methodology) and Istituto Superiore di Sanità in Italy (legal and ethical issues). The EHES RC is also responsible for establishing common standards for all aspects of data

¹Statistics Norway, Kongens Gate. 6, N-0153 Oslo, Norway (johan.heldal@ssb.no and susie.jentoft@ssb.no).

²National Institute for Health and Welfare (THL), Mannerheimintie 166, FIN-00300 Helsinki, Finland (kari.kuulasmaa@thl.fi).

collection. The EHES RC was funded by the European Commission through a Service Contract³. A Joint Action, with co-funding from the European Union (EU), was established to prepare for and pilot the surveys in twelve countries (Czech Republic, Germany, Greece, Finland, Italy, Malta, The Netherlands, Norway, Poland, Portugal, Slovakia and UK/England). The recommended size for the pilot surveys was about 200 participants.

By December 2011 five of these countries have carried out full size surveys (Germany, Italy, Netherland, Slovakia and UK/England). Three others are planning to start their national surveys in 2012 (Greece, Finland and Portugal). Although there is no European level funding for the full size surveys yet, the countries share a common interest to make their results comparable and to share data with the EHES RC for quality assessment and joint reporting.

The long term goal is to establish a system of national sampling based HES' (Health Examination Surveys) in Europe to collect comparable, high quality data on the health and health risks of the European adult population. This data will be used for Europe wide planning and evaluation of health policies, health promotion and research. The decision to launch the EHES project was based on the recommendations from a feasibility study (FEHES) which in 2008 concluded that the goal was feasible. The EHES Pilot Project is due to be finalized by the end of April 2012. European level funding for the next phase of EHES is pending.

3. The Recommendations

A European manual for EHES is being developed. The manual is published in three parts:

- A. Planning and preparation of the surveys (16 chapters, 194 pages)
- B. Fieldwork procedures (7 chapters, 124 pages)
- C. European level collaboration and co-ordination (7 chapters).

Parts A and B have been published (EHES, 2011). They were based on the FEHES recommendations (FEHES 2008b). Part C will be finalised by February 2012.

3.1. Target Population and Sampling

The core target population for EHES, to be covered by all countries, is all residents aged 25 to 64 years. Some countries extended the age range. However, there was some variation in the coverage of the available sampling frames (Table 3.1-1).

The manual (part A Ch. 3) suggests a geographically stratified, two-stage sampling design in most countries. The Primary Sampling Units (PSUs) are used as data collection areas, each containing one examination site. The PSUs should be selected with Probability Proportional to Size (PPS) sampling. If feasible, individuals can be stratified by sex and age when sampling at Stage 2. Procedures for doing this are described and implemented in an R-application that is offered (RcmdrPlugin.EHESsampling, See CRAN). The recommended sample size for each country is minimum 500 individuals in each of four ten year age groups (25-34, 35-44, 45-54 and 55-64) for each sex. This is based on an assumed design effect of 1.5. The PSUs should be small enough so that invitees can easily travel to the examination site and participate in the survey. Details on the sampling recommendations can be found in the EHES manual: part A (EHES 2011). The full size surveys should cover all seasons and the PSUs should be visited in a random order to avoid confounding effects between season and geography.

For sampling Stage 2, a frame of individuals with high coverage was recommended. In the Health Survey of England a postal address frame was used and all eligible residents at the address were invited to participate. In the pilot survey in Greece, households were selected from the Census 2001 and one individual was selected from each household.

The EHES manual suggests repeating the surveys with the core measurements about every five years, while some additional measurements may be repeated less frequently. As an alternative, a system of continuous data collection in

³Disclaimer: The EHES Pilot project has received funding from the European Commission/DG SANCO. The views expressed here are those of the authors and they do not represent Commission's official position.

which survey data from several years can be aggregated to provide quality estimates may be implemented. The Health Survey of England, the US National Health and Nutrition Survey (NHANES) and the Canadian Health Measure Survey (CHMS) are continuous collection surveys.

Table 3.1-1
Main sampling frames

Country	Sampling frame	Frame coverage
Czech Republic	National register of permanent residents	All residents
Finland	Central register of permanent residents	All residents
Germany	Registries of local populations	All residents
Greece	Pilot: Census 2001, Full size: Census 2011	Private households
Italy	Registry of local residents	All residents
Malta	Central population register	All residents
Netherlands	Population registers	Only citizens
Norway	Central population register	All residents
Poland	National population register	All residents
Portugal	National health service list	All registered under the national health system
Slovakia	Population register	All residents
UK/England	Postal address list	All private households

3.2 Recruitment

The recruitment process should be planned according to what is the most feasible way in each country. The individuals or households selected for the survey should first be contacted with an invitation letter including also an information leaflet. The leaflet should contain the most important information in a concise and interesting form telling about the objectives of the survey, the questionnaire and the measurements and encourage people to participate. It should stress the importance of the survey and of participation; briefly describe the selection process, the strict confidentiality of the survey data, the benefits to public health and to the participants that receive the results and a free health check and other relevant information. A toll free telephone number should be provided for questions. At least one to three re-contacts are recommended for invitees that do not show-up or call. In most countries, personal contacts by phone calls or home visits seem to be more effective than just mailed invitations. Substitution of non-participants is not allowed. The manual discusses factors that can affect participation rates and suggests measures that can be used to increase participation including long opening hours, appeals to employers, profiling in local media and by local community leaders, incentives and coverage of travel expenses. The use of incentives is dependent on the culture and legislation and survey organisation and the resources available in the country. It may be monetary or small gifts.

3.3 Interview and Measurements

The survey includes a questionnaire, core physical measurements including weight, height, waist circumference, blood pressure and blood samples for the measurement of blood lipids and fasting glucose or HbA_{1c}. Some countries included various additional measurements in the survey, such as lung function tests, electrocardiograms, dental health examinations or functional capacity tests. The core questionnaire items are based on the EHIS questionnaire. Translation of the questionnaire already exists for most of European languages. Using the same questions offers comparability to the EHIS surveys. The clinical measurements have been selected based on a number of criteria, the key focus being to address public health problems. For a full discussion, see the EHES manual, Part A, Chapter 5.

3.4 Data Management, Estimation and Access

All data on core measurements from the pilots and full size surveys carried out within EHES will be transferred to the EHES RC in anonymized form and a data sharing protocol is being developed to make it possible to give researchers access. The EHES RC will estimate a number of core indicators for each country. These indicators will mainly be age and sex standardized estimates for ten-year age groups based on flat one-year standardizations. This procedure controls for differential age distributions among the countries and is necessary to make the estimates

internationally comparable over time. Details on the indicators as well as procedures for imputation, weighting and estimation will be included in the manual part C.

4. Challenges

4.1 Experiences from the Pilots

Many aspects of the recommendations were tested in the pilot surveys. In most countries only one or two PSUs/-examination sites were selected. However, the selection of PSUs was often done purposively and not as a test of fully developed design for sampling Stage 1. In some of the countries carrying out full size surveys, the pilots were integrated as a part of the main survey. People were selected randomly within the selected PSU(s), usually stratified by sex and age groups. All countries were visited by representatives of the EHES RC to check how the recommendations had been implemented and to discuss any problems. All national pilots complied with the core survey content and age range. Although minor deviations were observed, there were few problems with the standardization of interviews or physical and clinical measurements. The most significant challenge was to obtain the desired participation rates. The FEHES recommended a target participation rate of at least 70%, a goal that none of the pilots achieved. Table 4.1-1 presents participation rates for nine of the pilot countries, without disclosing their names. The figures are based on preliminary analysis of individual level data on participation status, and may partly reflect possible errors in the coding of the data. The calculation of the participation rates is described in the EHES Manual, Part A, Chapter 13.

Table 4.1-1
Participation rates in the pilot surveys

Country	Sample sizes		Participation Rates (%)	
	Men	Women	Men	Women
1	200	200	40	54
2	370	391	38	44
3	125	125	54	71
4	198	202	57	54
5	1600	1600	42	49
6	1311		23	
7	245	245	41	43
8	300	300	34	47
9	124	126	44	67

In most pilots the individuals were selected directly from population registers (Table 3.1-1). The quality of the sampling frames showed variation. Non-contact rates were high in some pilot studies due to the sampling frame being out-dated and/or low reporting of moving to the main register. In some countries negative attitudes to either surveys or authorities may have contributed to low response rates. Personal contacts by telephone were difficult in some countries as few people had home telephones and mobile telephone numbers were not always available. In one country the ethical committee denied the survey organizers re-contacting the invitees in any way after the first invitation letter, which had a disastrous effect on participation. Few countries did recruitment and/or health examinations by home visits and some reported that home-visits were not appreciated in their country. There seems to be large cultural variations on this. We feel that survey organizers and ethical committees should not be afraid of motivating people in a positive manner to participate, also with personal contacts.

4.2 Sampling for the Full Size Surveys

As pointed out in the beginning of section 4.1, the population sampling could be only partially addressed in the small pilot surveys. Most countries used the same basic frame for selecting individuals in their pilots as they consider for their full-size survey. The role of these frames is to provide sizes for the PSUs in Stage 1 of the designs as well as providing the sampling frame for Stage 2 in the selected PSUs. Most of the national frames cover, at least in principle, all residents of the country, but as table 3.1-1 shows there are deviations to this.

PSUs need divisions that can be clearly defined on a map and for which data on their sizes in terms of people or households exist. This is usually possible for administrative divisions such as municipalities, districts or regions (the meaning of these terms may differ among the countries). Postcode areas are possible as well. Unfortunately, administrative divisions may be too large or too small either in population or extent to be suitable as PSUs. In this case, it is necessary to split or combine divisions to obtain units with suitable sizes. The new divisions should also be clearly identifiable in the sampling frame. What is feasible also depends on whether the individual frames are organized centrally or if they exist only at local level.

Many countries want to use existing facilities like regional health centres or hospitals as examination sites. However, in some countries we see that the density of such facilities is too low and many of them cover a district with too large travel distances to be suitable as PSUs. Cooperation with such facilities can also be a challenge. Other countries have avoided this problem by not using existing facilities. Germany has small vans criss-crossing the country with equipment, rigging temporary sites at rented localities for one week at a time before moving to the next site. The Health Survey of England carries out the examinations in people's homes, but this solution may create other problems for cross-national comparability. The German solution may be considered by other countries, either as a general approach or as a supplement to existing facilities where they are too far away.

NHANES (USA) and CHMS (Canada) uses mobile examination units. Such vehicles will be too expensive for many European countries. They will have to be in continuous operation over many years to be economically defensible. An investment in mobile examination units for the EHES surveys could eventually be done at European level with vehicles that can operate in many countries.

Table 4.2-1
Design and period for some European full size surveys

Country	Design	Survey period
Germany (DEGS)	180 PSUs with PPS sampling → 42 individuals per PSU → 7560 individuals	November 2008 to November 2011
Greece (planned)	PSUs with PPS → 5500 households → one person per household	When Census 2011 is available as frame
Finland (FINRISK, planned)	87 strata → 10 000 individuals in one stage. Only regional coverage.	January to April 2012
Italy (OED)	One centre per region (20) with non-probability → $m \times 220$ examined per centre → 9020 examined. Only people from a vicinity of the centre are invited because of too large travel distances.	September 2008 to March 2012
Malta (planned)	Stratified one stage → 3 600 individuals	2014
The Netherlands (NL de Maat)	7 PSUs (15 planned) → 15 000 invited, 4 000 examined	Phase 1: May to December 2009 Phase 2: October to December 2010
Poland (WOBASZ) (Not EHES survey*)	104 PSUs in 48 strata. Equal probability selection of PSUs within strata. 100 men + 100 women selected in each PSU → 20 800 invited	2003 to 2005
Slovakia	Sample of 4000 individuals from 36 districts with regional public health institutes. 43 districts without health institutes are not covered because of too large travel distances.	October to December 2011
UK/England (HSE 2011) (Not EHES survey*)	576 PSUs with PPS → 16 postal addresses per PSU → 9 216 dwellings → all eligible residents	January to December 2011

* A survey is defined as an EHES Survey if the data from the survey will be shared with the EHES RC. For UK/England this will only be the case for the pilot data.

The full-size surveys carried out so far were planned before the EHES recommendations were published and participate in EHES because they share a common interest to make their results comparable. These surveys adopted the EHES recommendations to the extent that was possible late at the planning stage, and they have not all been able

to comply completely. This concerns coverage and sampling design in particular. Table 4.2-1 presents two aspects of variations among them, the sampling designs and the survey periods.

In one of the surveys, the PSUs were not selected with probability sampling. Otherwise the number of PSUs shows large variation. From a pure sampling perspective it is desirable to have many PSUs with few invitees in each rather than few PSUs with many invitees in order to minimize sampling variation. However, involving a large number of health centres or clinics may be operationally difficult and costly. But this is also a question of how the survey is organized. The two countries with the largest number of PSUs are the two countries that operate their surveys independently of existing stationary facilities.

Table 4.2-1 also shows a large variation in survey periods from three months to almost three and a half years. Some countries plan to use shorter survey periods because of comparability with their earlier surveys. However, for comparability across countries all future full size EHES surveys should cover the same season. Preferably all seasons should be covered equally. NHANES and CHMS take their samples for two years at a time covering all seasons. In EHES there are competing interests that need to be reconciled.

5. Conclusions

Is it possible to establish a high quality system of comparable standardized health examination surveys in Europe? We think it is, but it requires a full commitment from the countries and the EU to target the methodological challenges mentioned in this paper. Anything half way will be waste of money. It also requires improved strategies to increase participation rates. We also anticipate that a full size survey receiving national rather than just local attention will attract more interest than the local pilot.

When this is written only a few of the surveys mentioned in table 4.2-1 have been completed and reported. When this is published in the proceedings we will know more of the results, their successes and failures, and more will be known about the future of EHES. The results based on further analysis of the data will be published later.

Acknowledgements

Lists of key personnel contributing to the EHES Pilot Project are available at <http://www.ehes.info/contact.htm>. EHES Joint Action has received funding from the European Commission (Grant agreement number 2009-23-01). The EHES Reference Centre is funded by the European Commission through a Service Contract (SANCO/2008/C2/02-SI2.538318 EHES).

References

EHES (2011), The EHES manual, part A and B. http://www.ehes.info/manuals/EHES_manual/EHES_manual.htm.

European Health Interview & Health Examination Surveys Database, <https://hishes.iph.fgov.be/index.php?hishes=home>.

European Health Risk Monitoring (EHRM) Project (2002), Review of surveys for risk factors of major chronic diseases and comparability of the results, <http://www.ktl.fi/publications/ehrm/product1/title.htm>.

FEHES (2008a), Review of Health Examination Surveys in Europe. *B18/2008, Publications of the National Public Health Institute*, Helsinki, available from http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2008/2008b18.pdf

FEHES (2008b), Recommendations for the Health Examination Surveys in Europe *B21/2008, Publications of the National Public Health Institute*, Helsinki, available from http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2008/2008b21.pdf.

MONICA (2003), Monograph and Multimedia Sourcebook, World Health Organization, Geneva.

RcmdrPlugin.EHESsampling (2011), Software. *The Comprehensive R Archive Network*. <http://www.r-project.org/>,
User manual available from authors.

Preliminary collection planning Collection Front Door

Anie Marcil¹

Abstract

Statistics Canada's collection program now offers all client divisions a one-stop organizational contact point for collection activities, a "front door" for collection. The Collection Front Door is responsible for the initial stage of assessing the feasibility of collection. This involves examining the survey's specifications, determining how the data collection process will proceed, estimating the collection capacity required and preparing budget estimates in conjunction with the various collection partners.

The Collection Front Door is designed to provide client divisions with consultation services on data collection; coordinate data collection services between the collection partners and the client divisions during the initial data collection feasibility assessment; play an integrative role; and clarify the roles and responsibilities of the collection partners and the client divisions.

This new service helps standardize and centralize collection activities, ensure consistency across all collection activities and use paradata from previous surveys to guide the preliminary planning of Statistics Canada surveys.

Key Words: Collection Front Door; Planning; Budget estimates; Collection activities.

1. Introduction

Over the years, collection services and the procedures associated with delivering those services have continually evolved and increased in complexity. Previously, survey managers were responsible for determining what collection services their survey needed and, for each service, finding a contact who could provide preliminary cost estimates. In some cases, they could spend a good portion of their time looking for that person to get advice or a referral to the services they needed. The information that the managers received about each collection service may have varied and been inconsistent. As a result, budget estimates were not always comparable. In addition, the managers may not have been aware of all the new collection services available to them, and consequently, the procedures followed for their survey were not necessarily the most effective.

In 2007, the new vision for Statistics Canada's collection services called for the creation of a single organizational contact point for client divisions looking for collection services. The Collection Front Door was launched two years later, in 2009. It then became the single central point for the preliminary planning of collection activities for all Statistics Canada surveys.

2. Clients and Collection Partners

2.1 Clients

The Collection Front Door's clients are survey managers responsible for business, household, agricultural and institutional surveys.

For new surveys, the Collection Front Door is instrumental in determining certain survey specifications, providing more information about the available collection services, and identifying the collection process and partners that will

¹Anie Marcil, Statistics Canada, 170 Tunney's Pasture Drive, Ottawa, Canada K1A 0T6. (anie.marcil@statcan.gc.ca).

be involved. For existing surveys that are undergoing a number of changes, it is a way of reviewing the survey specifications and determining what changes are required.

The Collection Front Door's assessments can be adapted to any collection method (computer-assisted telephone interviewing (CATI), computer-assisted personal interviewing (CAPI), paper or electronic collection).

2.2 Collection partners

This section contains a brief description of the main collection partners at Statistics Canada and their responsibilities.

The Collection Planning and Management Division (CPMD) is responsible for

- survey collection procedures;
- training interviewers;
- tracking the progress of collection;
- liaising with clients and the regional offices;
- coordinating capacity with the regional offices; *etc.*

The regional offices (ROs) are responsible for

- hiring interviewers;
- the production plan;
- the interviewers' schedules;
- collecting data from respondents;
- post-collection reports; *etc.*

The Collection Systems and Infrastructure Division (CSID) is responsible for

- the planning, development, support and maintenance of Statistics Canada's collection systems for all collection methods.

The Operations and Integration Division (OID) is responsible for

- survey product design and printing services;
- Distribution Centre services;
- data entry;
- imaging;
- coding;
- processing of administrative data; and so on.

The Communications Division is responsible for

- designing the brochures;
- testing electronic surveys;
- the assistance centre for electronic survey respondents; *etc.*

The Dissemination Division and System Engineering Division (SED) are responsible for

- testing electronic surveys; *etc.*

3. Scope and benefits

3.1 Scope

For new or redesigned surveys, the Collection Front Door is responsible for the preliminary assessment stage of collection. It is the point of contact between the client divisions and the collection partners. This preliminary stage involves examining the survey's specifications, determining how the data collection process will proceed, assessing the capacity requirements, preparing the budget estimates, and sharing the risks and recommendations with the client.

The Collection Front Door is only involved in the preliminary planning of collection. Once it is confirmed that the survey will go ahead, the collection partners and the client hold a hand-off meeting to summarize the preliminary assessment and get the survey development process under way. That is where the Collection Front Door's involvement ends.

3.2 Benefits

The benefits of this service are as follows:

- A single point of contact for all client divisions that want collection services. Survey managers have just one place to go to obtain information about collection activities.
- A more standardized approach that provides a comprehensive, harmonized view of all collection activities. This service ensures that the main assumptions used are the same for all collection partners and that that information is incorporated uniformly and consistently into the general feasibility studies for all Statistics Canada surveys.
- Better communication and better planning of collection activities reduce the risks, the negative effects and the possibilities of error or surprises in developing and conducting the survey.
- The service creates and maintains a centre of expertise on collection activities at Statistics Canada.
- In addition, various scenarios or options can be assessed so that survey managers can make better-informed decisions about the best collection strategy, without losing sight of the survey's main objectives.

4. Process and Collection Feasibility Assessment Report

4.1 Collection Front Door process

The Collection Front Door process is as follows:

- Clients are required to submit a request via a central corporate request management system along with a survey specifications form.
- Then the Collection Front Door meets with the client to review the survey's objectives and specifications. Roles and responsibilities are clarified, and one or more collection strategies are defined. At this initial planning stage, the survey's specifications are usually vague and may change quickly. It is standard procedure, therefore, for the Collection Front Door and the client divisions to hold a number of discussions to clarify outstanding points or issues.
- The specifications gathered at those meetings are then passed on to the collection partners. The latter provide their budget estimates, resource availability and recommendations or risks associated with the collection strategies on the basis of the specifications provided.
- The detailed Collection Feasibility Assessment Report (CFAR) is prepared in conjunction with the collection partners and submitted to the client.
- A hand-off meeting is held when the Collection Front Door receives confirmation that the survey can go ahead. The Collection Front Door's involvement ends at that point.

As mentioned, the Collection Front Door works with constantly changing parameters and potentially confusing terminologies. It is therefore important to allow sufficient time for the Collection Front Door to carry out its assessments and establish realistic, reliable working assumptions. In addition, it is not uncommon for the Collection Front Door to have to redo reports or assessments with new survey specifications.

With the experience it has acquired in the last two years, the Collection Front Door has shown flexibility in addressing clients' needs quickly while following the proper steps in the process.

4.2 Collection Feasibility Assessment Report

The Collection Feasibility Assessment Report contains the following sections:

- a summary of all collection activities for the survey in question;
- a list of the specifications and assumptions used in the assessment;

- an overview of each collection partner's activities;
- an assessment of each partner's capacity;
- a section explaining the risks and recommendations associated with certain aspects of collection for the survey;
- a detailed description of each partner's preliminary costs based on preliminary specifications;
- a few important deadlines, including the survey go-ahead confirmation date.

5. Success of the Collection Front Door

5.1 General success

Since the Collection Front Door was established, it has chalked up a number of significant achievements:

- It is now recognized as a centre of expertise on collection activities. The experience that its members have acquired over the last two years (and continue to acquire) is making consultations with client divisions faster and more productive. It can answer client divisions' questions directly, without checking with the collection partners, which allows them to concentrate on the Agency's other priorities.
- To date, the Collection Front Door has assessed more than 140 surveys using various collection methods and strategies (*e.g.*, multimode surveys (CATI/CAPI), electronic surveys, unimode surveys, postcensal surveys, longitudinal surveys).
- Roles and responsibilities are clearly established at the time of the preliminary collection assessment, which facilitates the management of the survey development activities. The client knows what activities will be carried out by the partner and is therefore able to concentrate on other aspects of the survey's development and implementation. This has resulted in better relations between the collection partners and the client divisions.
- Collection Feasibility Assessment Reports identify the effects of certain collection strategies, which helps survey managers make better decisions.
- The Collection Front Door contributes to the Corporate Business Architecture by assigning collection activities to the right centres of expertise, which helps eliminate duplication of collection activities at Statistics Canada.
- The standardized specifications form gathers basic information about the survey in question. In addition, the form has proven to be a very useful checklist to remind client divisions of all the steps in the collection process and the resources that have to be included in the project.
- Having a preliminary planning tool helps Statistics Canada's survey managers in preliminary discussions with external clients based on predefined scenarios.

5.2 Success for methodology

The benefits for methodology are as follows:

- information about the results of similar surveys;
- information about the available collection services;
- a list of the files or information that has to be provided to partners to begin survey development;
- clarification of each party's roles and responsibilities;
- a comprehensive, harmonized vision of collection activities and the views of collection experts;
- preliminary assessments based on a number of scenarios. Methodologists offer survey managers a few collection strategies, which are then assessed and compared on the basis of their effects on collection activities (capacity, costs, deadlines, *etc.*).

6. Conclusion

In conclusion, the Collection Front Door was designed for the following purposes:

- provide client divisions with data collection consulting services;
- coordinate data collection services between client divisions and collection partners during the initial stage of assessing collection feasibility;

- prepare Collection Feasibility Assessment Reports containing preliminary cost estimates, capacity analyses and a description of the risks associated with the proposed collection strategies;
- play an integrative role by
 - o ensuring that all collection partners concerned take part in the process;
 - o attempting to establish high-level integrated specifications for data collection projects;
 - o standardizing cost estimates and capacity planning and ensuring consistency in the services provided to clients;
 - o serving as a single point of contact for client divisions looking for collection services.
- reduce duplication wherever possible and eliminate gaps in current collection roles and responsibilities.

The Collection Front Door plays a pivotal role in the preliminary planning of Statistics Canada's surveys.

Acknowledgements

The author is grateful to Milana Karaganis for her guidance and encouragement in establishing this new service, and to Edward Joseph and Stuart McFarlane for their continuing efforts to improve the process and meet clients' needs.

Implementing quality control procedures at NASS's National Operations Center

Jeffrey M. Boone, Joseph L. Parsons, Shari R. Feld, Jenna N. Levy and Kristie L. Flaherty¹

Abstract

The National Agricultural Statistics Service (NASS) centralized much of its survey data collection activities at its new National Operations Center (NOC) in August 2011. The NOC Division is responsible for telephone data collection, the processing of paper questionnaires, the maintenance of NASS's list frame, and other survey related activities. The motivation for the creation of the NOC was to reduce the source of error inherent in data collection activities, improve data quality, and reduce operational costs. The business case for the NOC articulated the need for a comprehensive quality control program at the new center to drive performance excellence. This paper examines the current state of the quality control program at the NOC, including metrics and monitoring methods. The process of implementing quality control measures into various functions is discussed as well as future opportunities for enhancement and lessons learned.

Key Words: Quality Control; Quality Assurance; Call Center; Metrics; Monitoring.

1. Introduction

1.1 Overview

Quality control/assurance is an important aspect of any agency. The International Organization for Standardization defines quality as “the degree to which a set of inherent characteristics fulfill requirements” (2005). Quality assurance involves ensuring that objectives or targets are met. Quality control involves the monitoring or evaluating a product or process to ensure that the desired standards are met. For the purpose of this article, the term ‘quality control’ will be used, although both quality assurance and quality control are being examined.

In this paper, the implementation of quality control procedures at the National Agricultural Statistics Service's National Operations Center will be discussed. Current progress and future objectives will be presented, along with a literature review of articles relating to quality control and its application to survey data collection.

An effective quality control program will ensure that a process will provide quality results over the entire span of the process. The *Guide to the Project Management Body of Knowledge* explains that meeting quality requirements has certain benefits, such as less rework, higher productivity, lower costs, and increased customer and stakeholder satisfaction (2008). In the context of survey methods, this corresponds to less call backs, less time spent editing surveys and researching unusual data, more record touches (more interviews), lower overall costs, and increased assurance of high-quality data.

Apart from discovering cost savings, having a quality control program may determine the parts of the process that work well and need to be emphasized. This could lead to suggestions in future training and employee incentives and rewards. The system will also detect problems in the process, also leading to suggestions in more training as well as possible disciplinary actions or other fixes. Maintaining a quality control program can also lead to the recognition of process improvement opportunities, such as Lean Six Sigma (*e.g.*, see George (2003)). This can lead to continuous improvement possibilities throughout the life of the program. Although process improvement can be applied at any stage of a process, the existence of metrics already being measured will enable a smooth and quick application of process improvement.

¹Jeffrey M. Boone, Joseph L. Parsons, Shari R. Feld, Jenna Levy, and Kristie Flaherty, National Agricultural Statistics Service, United States Department of Agriculture, 3251 Old Lee Hwy., Rm. 305, Fairfax, VA 22030, jeff.boone@nass.usda.gov.

1.2 Literature Review

Quality control has been discussed in many works in survey methodology. LaFlamme, Mayden, and Miller describe active management (quality control) as a set of plans and tools used to effectively and seamlessly collect information on data collection processes to better improve survey administration practices. Active management refers to “monitoring progress, conducting timely analysis of indicators, identifying problems, implementing and communication corrective actions, and evaluating successes” (2008).

‘Paradata’ is often used to describe information about the data collection process. This paradata is equivalent to quality control metrics when they are applied to the survey data collection process. Bates et al, writes that types of paradata include call records, observations of interviewers and respondents, audio recordings for interviewer and respondent interactions, items generated by computer-assisted instruments, such as response times and key strokes, *etc.* Paradata is used to measure survey quality in a production environment and to manage production with the goal of optimizing quality and minimizing costs. Paradata can be used for fieldwork monitoring, non-response analysis, responsive designs, aid in assessing measurement error, non-response adjustment, and to improve editing and coding (Bates, N., Dalhammer, J., Phipps, P., Safir, A., and Tan, L., 2010).

Lepkowski *et al.* describe four paradata paradigms: effort, active sample, productivity, and data set balance. For the National Survey of Family Growth, in reference to effort, several sources of data are measured: interviewers working, hours, percent productive, calls per day, calls per hour, percent peak calls, and screener versus main call. The authors also looked at percent occupied, percent eligible, percent nonworking, noncontacts, mean number of calls, percent of 8 or more calls, percent locked buildings, percent resistant, percent hard appointment, and propensity. In productivity, the number of interviews, cumulative interviews, hours per interview, and calls per interview were measured. In data set balance, the response rate, percent with children, percent sexually active, and group rates were measured (2010).

Lyberg explains that paradata can provide continuous updates of progress and stability checks (monitoring, input to long-run process improvement), of product quality (analysis for special and common cause variation), and input to methodological changes (finding and eliminating root cause problems). Paradata is also critical to responsive designs and for providing input to organizational change. It can be used to understand variation and measure cost of poor quality and waste. Paradata are multivariate in nature and may need to be combined to be relevant. Creating paradata archives allow reanalysis so that understanding of what is key can grow or change (2009).

2. NASS and the National Operations Center

2.1 About NASS and the National Operations Center

The National Agricultural Statistics Service (NASS) provides timely, accurate, and useful statistics in service to U.S. agriculture. NASS conducts hundreds of surveys a year, conducts the Census of Agriculture every five years, provides data relating to America’s agricultural products, and provides data used to determine commodity prices. NASS currently has a relatively decentralized approach to data collection, with 46 field offices throughout the United States. NASS began centralizing its data collection procedures with the opening of a National Operations Center (NOC) in St. Louis, MO, in August 2011. One objective of this initiative is to reduce the source of error inherent to data collection activities, improve data quality, and reduce operational costs. The center will have various functions, including telephone interviewing, the processing of paper questionnaires, maintenance of NASS’s list frame, training, survey development, Blaise programming, and Web survey system programming.

NASS currently has some quality control procedures in place. For telephone enumeration, these include the daily reports of response rates, post-survey reports of enumerator performance, post-survey cost reports, incentives for contract interviewers, enumerator monitoring, call backs, and enumerator performance evaluations. The enumerator monitoring and call backs, performed by supervisors, involve a paper form that is filled out by hand and filed. For forms processing, many forms are edited by hand. NASS does not currently track the number of edits or the time it takes for forms to be edited. Some forms, such as the Census of Agriculture and the Cash Rents Survey, are processed at the Bureau of the Census’ National Processing Center (NPC) in Jeffersonville, IN. The NPC utilizes an

electronic tracking system that provides numerous reports, including document tracking information such as location of a form and the number of forms at different steps of the process.

2.3 Quality Control at the National Operations Center

A central point of the creation of the NOC is for the improvement of data quality. With the successful implementation of the quality control program, NASS can continue to evolve from information learned in the data collection processes at the NOC. Although NASS has some quality control procedures, as mentioned above, these may be ineffective in the NOC's environment. First, monitoring evaluations and call backs are currently done on paper. With the large number of enumerators at the NOC, many monitoring reports will need to be created, yielding to an inefficient monitoring process (i.e., these informal methods NASS currently uses will not simply scale up to the amount of work at the NOC). Also, a systematic and standardized system will ensure data quality. Forms processing also uses a quality control program that will detect problems with the system. To ensure the standardization of NASS's frames, the frames maintenance procedure will also need to be monitored for quality to ensure that the frames are consistent and of high quality.

3. Progress

3.1 Steps

The steps to create a quality control system are straightforward. First, the metrics used to monitor the quality of the data collection process must be identified. Both productivity and data quality must be considered. Next, a method for capturing these metrics must be developed. This includes both developing an interface or system to record the necessary information for the calculation of these metrics and identifying a location for this information to be recorded. Finally, a method for displaying these metrics must be developed. This sort of 'dashboard' will provide the users of the system with the information required to make data-driven decisions in a timely manner. This process seems simple, however, in some business environments, many difficulties arise. These issues will be discussed later in this article.

3.2 Metrics

The first step towards creating a quality control system is to identify the metrics that will be used to monitor the quality of the data collection process. Metrics should have the ability to answer a question which results in an objective, data-driven decision. The metrics should measure both the productivity of staff, such as telephone interviewers, as well as the quality of the data. Metrics relating to productivity, such as number of calls, length of calls, and number of refusals, as well as cost, are relatively simple to determine; however, those relating to data quality are more difficult to ascertain. These measures will provide a way to detect issues in the data collection process and with the quality of the collected data. If issues are detected, actions may then be taken to correct or minimize the occurrence of these problems.

When selecting productivity indicators, it is important to consider the following characteristics. The indicators must be 1) easily understood, 2) measureable and comparable at any point in data collection, 3) consistently updated throughout data collection, 4) relevant, interpretable, and comparable at different levels of aggregation. These can be divided into three different types based on times, number of attempts/calls, and time per unit or time per complete (Laflamme, 2009).

Numerous metrics have been identified for use at the NOC. Some metrics include usable completes per hour worked, percent of refusals, average length of call, number of paper forms processed in a given time period, as well as many others. Some metrics can be considered both measures of productivity and data quality, such as the average length of call. For example, the length of the call gives information on how much work is being done and is thus a productivity measure. However, if the average length of calls yielding completes is unusually low, the quality of the data collected may be compromised.

A common issue with the determination of metrics involves the number of metrics to monitor, as well as their importance relative to the question posed. Monitoring too many metrics may not only consume too much time and effort but may signal a problem when there actually is no problem. The more areas one looks for problems, the more false alarms one will find. A common method for alleviating this problem is to create an index combining information from multiple variables. This would enable the monitor to observe fewer variables without losing any information.

Another problem in the case of call center metrics is inherent in the difficulty of some cases. For example, some geographic regions have a naturally lower response rate than others. In this case, it would be erroneous to consider only the response rates of enumerators across all regions, since those with the more difficult cases will probably have lower response rates. Thus, studies have been done on creating a 'difficulty' or 'response propensity' index (see *e.g.*, Laflamme and St-Jean (2011)). This would enable the coach or supervisor to note the difficulty of the cases when examining enumerator productivity. NASS is currently performing research to develop a response propensity index to classify cases that may have a high probability of non-response.

3.3 Storage

An important aspect of a quality control system is data storage. The physical space may sometimes be a problem, but beyond just the concept of size is the business problem of database maintenance. Database maintenance includes ensuring that items allowed to be stored in the database are chosen in a manner such that no information is replicated and every variable may be referenced in a standardized manner. NASS is currently developing a new centralized database for the storage of work-in-progress data. Statistics Canada also uses a centralized database as a paradata warehouse to store all necessary information. Although it is a very good practice to create a database such as the one described above, it creates a bottleneck for system development that relies on the database for storage and data access. This has been a very difficult obstacle experienced in the creation of the quality control system.

Another important aspect in using a newly-deployed centralized database is ensuring the data originating in different systems is in the same standard form when it is placed in the database. If different systems are not using the same identifier for the interviewer, for example, the data from these systems may not be merged properly, and the resulting metrics may not be properly computed. Note that it may not be the case, due to the architecture of the database, for all necessary information to move to that database. Thus, some data may need to be pulled from other databases by the quality control application.

3.4 Monitoring

Once the necessary or desired metrics have been determined, a method must be selected to obtain these metrics. For telephone enumeration, this includes recording the production measures, such as the length of a phone call or refusal rate, in a database. Many of these measures can be obtained from CATI systems, such as the Blaise system. Measures relating to enumerator performance that are not automatically captured, such as the ratings given by the monitoring staff, will need to be recorded. Other information, such as the skills, availability, attendance record, and time in job of each enumerator, will be tracked. For the other functions of the NOC, metrics may be tracked by the computer software, such as the Tracking and Control System and Key From Image/Paper System being developed for forms processing and in-house frames maintenance systems, the Enhanced List Maintenance Assistant and the Enhanced List Maintenance Operation. Data from these systems will also need to be made available in a database.

Once the data is accessible, a computer software application will be used to illustrate the data to the staff that need it. This dashboard will present the data as descriptive tables as well as graphs and charts. It is important to ensure that field managers have the ability to "drill down in trace files to examine details of interviewers" (O'Reilly, 2010). Many software applications have the capability to create dashboards. Something as simple as an interactive webpage may also be sufficient. The choice of the software will be determined by the availability of resources, including current licenses, cost of new software, and employee knowledge.

4. Issues

NASS is currently developing multiple new systems. Many of these systems are high in demand and are necessary in the development of the NOC. The quality control system is very important, but not often considered a requirement to begin operations. Thus, the specialized personnel required for the development of the system are utilized for these other 'higher priority' systems; due to the demand of these staff, it is often difficult to acquire them as resources. Unfortunately, some steps must be completed before subsequent steps can begin (*i.e.*, some steps to complete the project lie on the critical path).

Data originating in different systems must be standardized before being put into the centralized database. Since many of NASS's systems were created by different groups (inside and outside the agency) and at different times, the unique interviewer identifiers are not the same format across the different systems. Thus, when the data is pulled into the centralized database, it must be ensured that the data is being merged properly through some sort of transformation or script to match the interviewers' data.

Not only is it often difficult to prepare the database for the storage of the data, the acquisition of the software to create the dashboard is a difficult process. The process is not as simple as choosing the software, purchasing it, and installing it; deployment of the software on a company- or agency-wide network is an immense task. Security and enterprise architecture personnel approval must be obtained in most cases, which may take a large amount of time. Even if the software is installed and available, it may not be configured properly for this specific use.

5. Conclusion

Creating a quality control system can be a daunting task. The steps to create the system, namely identify the metrics to be monitored, choose the monitoring method, and develop the user interface to use the system, are straightforward. However, the environment in which the system is being developed may cause these simple steps to be far more complicated than first realized. It is important to note that a key resource necessary for the creation of a quality control system is patience (along with strong negotiation skills). Many of these steps in the creation of the system involve many other people and systems that are out of your control. Being patient while still pushing to accomplish the task is an important attitude that one must have in this endeavor, as well as many other cross-functional projects.

Once all the hurdles are leapt and the system is in place (not necessarily at 100%)—many benefits may be realized. The system may pave the way for continuous process improvement opportunities not recognized without the tracking of metrics. If a quality control system is in place and capturing the necessary information for the desired metrics, the step of measuring process metrics is already complete. Another benefit is the opportunity for global optimization of the entire data collection process across all data collection locations as opposed to local optimization that would only be performed at individual data collection sites. Capturing these metrics also provides opportunities for cost savings, such as performing optimal management of data collection and a reduced need for editing. Other benefits include identifying areas for focused training, improving hiring procedures, survey development, and a potential application of responsive or adaptive survey design.

References

- Bates, N., Dalhammer, J., Phipps, P., Safir, A. and L. Tan (2010), "Assessing contact history paradata quality across several Federal surveys", *Proceedings of the American Statistical Association 2010 Joint Statistical Meetings*, American Statistical Association, retrieved from https://www.amstat.org/sections/srms/Proceedings/y2010/Files/306005_55654.pdf.
- George, M. (2003), *Lean Six Sigma for Service: How to Use Lean Speed and Six Sigma Quality to Improve Services and Transactions*, New York: McGraw-Hill.
- International Organization for Standardization (2005), "ISO 9000", *Quality Management Systems-Fundamentals and Vocabulary*, Geneva: ISO Press.

- Laflamme, F., Mayden, M. and A. Miller (2008), "Using paradata to actively manage data collection process", *Proceedings of the American Statistical Association 2008 Joint Statistical Meetings*, American Statistical Association, retrieved from <http://www.amstat.org/sections/srms/proceedings/y2008/Files/300608.pdf>.
- Laflamme, F. (2009), "Experiences in assessing, monitoring and controlling survey productivity and costs at Statistics Canada", *Proceedings of the 57th Session of the International Statistical Institute*, retrieved from <http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/0049.pdf>.
- Laflamme, F and H. St-Jean (2011), "Proposed indicators to assess interviewer performance in CATI surveys", paper presented at the 2011 Joint Statistical Meetings, Miami, FL.
- Lepkowski, J.M., Axinn, W., Kirgis, N., Kruger, S.N., Mosher, W. and R.M. Groves (2010), "Use of paradata in a responsive design framework to manage a field data collection", NSF Paper No. 10-012. Retrieved from <http://www.psc.isr.umich.edu/pubs/pdf/ng10-012.pdf>.
- Lyberg, L. (2009), "The paradata concept in survey research", Presented at NCRM Paradata Network, London, UK, August 24, 2009, retrieved from <http://www.natcen.ac.uk/ncrm-paradata-network/docs/Lyberg-paradata-concept.ppt>.
- O'Reilly, J. (2010), "Paradata and Blaise: a review of recent applications and research", *Proceedings of the 12th International Blaise Users Conference*, retrieved from <http://ibuc2009.blaiseusers.org/papers/7d.pdf>.
- Project Management Institute (2008), *A Guide to the Project Management Body of Knowledge*, 4th ed., Newton Square, PA: Project Management Institute, Inc.

SESSION 8A

**USING STANDARDIZED METHODS AND TOOLS FOR POST-
COLLECTION PROCESSING**

Standardization of post-collection processing in Business Surveys at Statistics Canada

Serge Godbout¹

Abstract

Statistics Canada is undergoing a redesign of its business surveys. This offers a unique opportunity to develop a new processing framework oriented toward selective editing and active collection management. Statistics Canada is proposing to implement a complete set of functionalities to be applied to partially collected data in order to come up with estimates while collection is still going on. With this set of estimates, a carefully chosen set of quality indicators would also be produced. These indicators will be used to make decisions about active collection management, including producing a list of priority units for non-response follow-up and failed edit follow-up. This presentation will give a flavour of the challenges of developing processes to produce estimates and quality indicators based on a portion of the sample being collected and will also describe the active collection management strategy.

Key Words: Standardization; Data processing; Quality indicators; Active collection management.

1. Introduction

In 2010, Statistics Canada launched the Corporate Business Architecture initiative. Increasing financial pressures led to a review of its methods and systems to identify avenues to achieve efficiencies, enhance quality assurance and to improve the responsiveness of new statistical programs. A key element of the initiative is the development and mandatory use of generic corporate services and generalized systems for sampling, collection, processing, publication and archiving of statistical data for household and business survey programs. To meet these objectives, Statistics Canada has launched a major redesign of its business statistics programs. The purpose of the Integrated Business Statistics Program (IBSP) is to develop a common platform for a large number of its business surveys. By 2016, approximately 120 annual, sub-annual or ad hoc surveys from 10 different programs will be integrated in this new harmonized framework.

The objectives of the IBSP are to build a single harmonized platform integrating business surveys, reduce development and maintenance costs, simplify processes, reduce the learning curve for employees and improve the timeliness of surveys, modernize processes, reduce the response burden, and generate efficiencies. To accomplish these objectives, the IBSP strategy is built on six pillars: increased governance, increased use of tax data, a common editing strategy, more efficient management of active collection, multi-mode collection with electronic data collection as the primary collection mode, and use of the Business Register as the sole frame for all business surveys (Statistics Canada, 2010). Standardization will be the focus to ensure efficient production of statistics while meeting the specific requirements of the different surveys integrated in the IBSP.

This article describes the key components for standardization of the IBSP's methods and tools. Section 2 presents the basic processing models. Section 3 describes the means used to ensure standardization of methods and tools, and Section 4 details the business process models and their associated methods and tools.

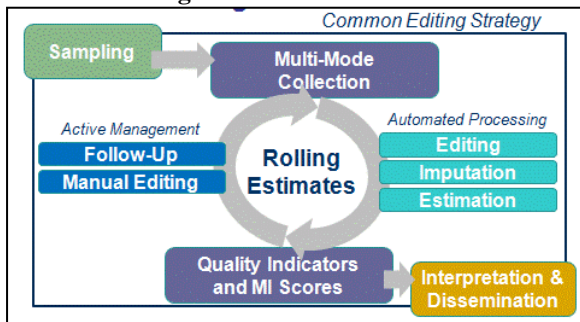
2. IBSP processing models

The IBSP sampling and collection model is built on a two-phase design to better target the population in order to generate financial and commodity estimates and other industry-related characteristics. The sampling and collection model also assumes a top-down approach focused on the enterprise with maximum use of tax data.

¹Serge Godbout, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6.

The IBSP’s post-collection processing is based on a common editing strategy defined using an iterative processing model that combines collection, processing and analysis, periodically produces estimates and quality indicators, and dynamically feeds back to collection and analysis. The common editing strategy will harmonize editing methods and tools, expand automated editing operations, reduce follow-up activities with respondents, and limit manual interventions to influential units. A key element of the common editing strategy is an efficient method to prioritize follow-up and manual editing. This should result in efficiency gains and improved quality in terms of timeliness and accuracy (Saint-Pierre and Bricault, 2011).

Figure 1
IBSP Processing Model



At the centre of this strategy is an iterative processing model referred to as rolling estimates. As Figure 1 shows, the data collected through multi-modal collection from various sources are first integrated and then subject to automated processing for editing, imputation and estimation. When complete, quality indicators are linked to the estimates. If the target quality is achieved, the estimates are set aside for interpretation before dissemination and the collection resources are reallocated. If the quality is not achieved, the quality indicators are disaggregated at the unit level to generate measure of impact scores used to prioritize the units for telephone follow-up and manual editing. During the next iteration, the integrated data are updated using newly collected data and the manual corrections to produce a new set of estimates and quality indicators. The cycle continues until all quality targets are met or the data collection end date is reached.

3. Standardization of methods and tools

Methods and tools are standardized by building exhaustive business process models, using generalized systems and defining effective metadata.

3.1 Role of business process models

The starting point in building business process models is the Generic Statistical Business Process Model (GSBPM). This generic model was developed by the Joint UNECE, Eurostat and OECD work group (METIS) to serve as the basis for standard terminology and statistical metadata systems and processes development (United Nations Economic Commission for Europe Secretariat, 2009). It applies to all activities undertaken by official statistical agencies that result in data outputs. The GSBPM consists of four levels describing the hierarchical relationship of all statistical business processes and sub-processes. There are additional global processes, two of which are closely linked to the model (quality management and metadata management).

Based on the GSBPM, processes and sub-processes are then described in order to build the most detailed level of the business model, which becomes specific to the IBSP. This is a key step in standardizing processes by establishing an exhaustive and ordered list of statistical activities to be carried out for the IBSP, by defining the roles of each partner and a common language. The generic systems in the project are also clearly identified.

3.2 Role of generic integrated services

Statistics Canada has put in place generic integrated services with others still to be developed. The goal of these generic services, including methodology, the business register, the questionnaire design resource centre, the data services centre, collection, publication and centralized post-collection processing, is to enhance operational effectiveness and to standardize their respective services. The Corporate Business Architecture working group recommended that use of these services be mandatory for all programs and projects.

3.3 Role of generalized systems

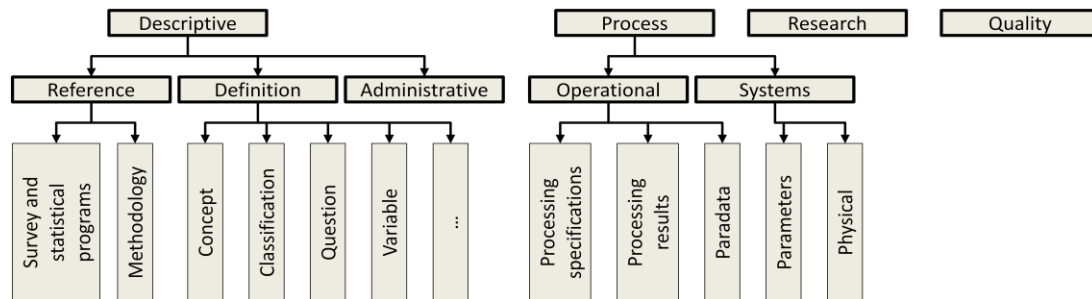
Since the late 1980s, Statistics Canada has developed a set of generalized tools to address survey needs related to sampling, editing, imputation, estimation, confidentiality and time series. Over time, these tools have proven their usefulness in standardizing methods, reducing development costs and facilitating the movement of people between projects (Mohl, 2007). Generalized systems play a very important role in standardization by providing a set of known, accepted and proven statistical methods covering virtually all processing steps. They are therefore applied to all surveys, reducing development and programming requirements. Built-in modules and allowing for flexible parameterization, they adapt to the specific needs of each survey. Development, support and maintenance of these systems are handled by teams of methodologists and programmers, guaranteeing thorough testing, complete documentation, and expert technical support. Once in place, these generalized systems significantly reduce the programming required to construct the specific systems required for each survey and thus become its focal point.

Several of Statistics Canada’s generalized systems are currently being redesigned. In particular, new generalized sampling (G-Sam) and estimation (G-Est) systems will be constructed respectively from the GSAM generalized sampling system and from the CMS and StatMax pair of generalized estimation systems, all developed by Statistics Canada. The Banff system will remain the principal generalized tool for editing and imputation. Other generalized tools, such as the G-Confid and G-Link, were also developed by Statistics Canada (Deguire, Reedman and Wenzowski, 2011).

3.4 Role of metadata

Metadata are generally perceived as being data that define and describe data. More specifically, Gartner Research (Blechar *et al.*, 2010) defines metadata as information that describes various facets of an information asset to improve its usability throughout its life cycle. The different types of metadata included in the IBSP can be organized by the classification set out in Figure 2.

Figure 2
Classification of IBSP Metadata



The IBSP’s fundamental principles include the development of metadata-managed systems on the basis that metadata precede the various processing steps and are used at the outset rather than documented after production is complete. The IBSP vision sees a central metadata repository to eliminate duplication of work, improve effectiveness and reduce the risk of error in all survey phases (Statistics Canada, 2010). Metadata identification and traceability requirements were established on a priority basis.

Metadata play a number of important roles in method and system standardization. Descriptive metadata make it possible to harmonize data and metadata definitions and to ensure ongoing documentation. Process metadata enhance the effectiveness, flexibility and coherence of systems by managing through effective parameterization the complexities generated by the specific requirements of surveys. They improve processes by detailing the operations carried out in the different processing steps and simplify connectivity with generalized systems.

4. Description of business process models and lists of methods

To support the different processing steps of the IBSP model, many statistical methods have to be evaluated and selected, including several that are already available through Statistics Canada's existing generalized systems. In some cases, more than one statistical method may be available to managers to meet specific survey requirements. In other instances, methods may significantly impact the model's complexity, operations or systems development, in which case, the proposed methods must be comprehensively evaluated and compared to make the optimum choice from a global perspective.

Six criteria have been established to aid in this selection, namely, the impacts on (1) collection costs, (2) data quality, (3) complexity, (4) generalized systems, (5) operations and (6) responsiveness and flexibility. If several statistical methods are compared, a weight is assigned to each of the six criteria so that it is possible to determine the relative gain in accuracy (better CV for example) compared to all of the impacts on the other criteria.

Sections 4.1 to 4.3 describe the sampling and collection, post-collection processing, and quality indicators and active management models. Their business processes and list of statistical methods will be presented.

4.1 Sampling and collection

The IBSP sampling and collection model is based on a two-phase design. The first phase is an integrated sample of a larger size chosen from information available on the Business Register. It is used to collect contact information, updated classification variables and basic information on activities, commodities and industry-related characteristics. Note that units for which recent historical information is available are excluded from first phase collection if the information is deemed to be still valid. In the IBSP model, sampling is focused at the enterprise level based on a top-down approach. This means that each collection phase uses a unique base list with the enterprise as the sampling unit, where the part of the enterprise in the survey field, and its contribution to each domain of interest, is taken into account through a multivariate strategy. The IBSP uses rotation to coordinate intra-survey samples and samples between surveys as a means to manage response burden.

The IBSP collection model seeks to make optimum use of tax data to reduce response burden. Electronic collection will be the main data collection mode for surveyed units. The non-response follow-up strategy includes batch actions with low marginal cost (fax and email reminders, voice messages) and follow-ups by computer-assisted telephone calls where the marginal cost is higher.

Table 3 describes the generalized methods and tools available for sampling and collection for the processes identified in the business process model (Figure 6 appended). Note that processes 4.1.2 to 4.1.5 and 4.1.7 to 4.1.10 are missing from the table because they are essentially metadata management and approval steps unrelated to statistical methods. The operations and tools associated with sampling frame and collection management are the responsibility of Generic Services of the Business Register (Statistics Canada, 2009). For the key components of the sampling design – sample stratification, allocation and selection – the G-Sam generalized sampling system (Statistics Canada, 2006) enables use of generally accepted business survey methods. Collection operations are not explicitly part of the IBSP business process model because they are handled in their entirety by Generic Services.

Table 3**Generalized sampling and collection methods and tools**

Business process steps	Statistical methods available and/or recommended	Generalized tools
4.1.1- Create and revise survey frame	Parameterization	Business Register
4.1.6.1- Stratify according to size	Cumulative Root f (Dalenius-Hodges); Geometric (Gunning-Horgan); Lavallée-Hidiroglou	G-Sam
4.1.6.2- Allocate sample	Multivariate power allocation (includes univariate power / Neyman)	G-Sam
4.1.6.3- Select sample	Stratified Bernoulli / Stratified SRS / Poisson; Permanent random numbers; Collocated sampling	G-Sam
4.1.6.4- Produce diagnostic reports	Parameterization	IBSP system
4.1.11- Customize collection entities	Parameterization	Business Register

4.2 Post-collection processing

Once the data has been collected and integrated, editing identifies the records and variables to be imputed or excluded from the donor group and/or imputation models. Under the IBSP common editing strategy, the bulk of automated edits will be moved out of collection to be embedded in post-collection processing in order to effectively prioritize the units that failed editing and to significantly reduce the time between when the data are received and are available for processing.

In addition to using partial non-response, the IBSP strategy is to impute units for total non-response when auxiliary data are available. In addition, the selective editing strategy proposed for the IBSP requires imputation of key variables for respondents to be used as fitted values. The nature of the imputed data will vary depending on the collection phase. In the first phase, imputation should be limited to variables of size and classification for non-responding units or for units excluded from first phase collection when the information is still deemed to be valid. In the second phase, survey data and administrative data will be subject to imputation.

At the estimation stage, IBSP surveys will estimate totals, ratios, proportions and trends for financial, commodity and characteristic data collected in the second phase. Systems will produce domain estimates and their associated variances for the surveyed and non-surveyed portions. In addition, second phase sampling algorithms, and especially the stratification and allocation algorithms, require estimates of auxiliary totals used as input parameters.

Table 4 describes the generalized methods and tools available for post-collection processing for the processes identified in the business process model (Figure 6 appended). Editing and imputation methods will be supported by the Banff generalized system (Statistics Canada, 2011) and the G-Est generalized system (Statistics Canada, 2005) will support estimation.

Table 4
Generalized methods and tools for post-collection processing

Business process steps	Statistical methods available and/or recommended	Generalized tools
5.2.1 and 5.5.2- Detect outliers	Hidiroglou-Berthelot; Sigma-Gap	Banff
5.2.2- Localize errors	Fellegi-Holt principles	Banff
5.3.1- Impute missing data	More than 20 imputation methods available including donor imputation, previous value, trend, etc.	Banff
5.3.2- Impute respondents	Similar imputation methods	Banff
5.3.3- Prorate	Basic or scaling prorating	Banff
5.5.1- Reweight for non-response	Adjustments by homogeneous response groups	G-Est
5.6.1- Calibrate weights	Hajek; Generalized regression with double calibration	G-Est
5.6.2- Calculate estimates and variances	Methods to be determined	G-Est

4.3 Quality indicators and active management

Quality indicators are generally used to evaluate the quality of outputs. In the IBSP, they also play an important role in the reallocation of resources linked to follow-up operations and editing and in cutting off active collection. The IBSP model provides for the use of quality indicators based on total variance, combining sampling variance and non-response variance, or on the mean square error when bias is measured. These indicators are rounded out with others based on response and coverage rates, weighted or unweighted, in order to prevent errors caused by the volatility of quality estimates at the start of collection. Quality indicators are then disaggregated to the unit level, generating measure of impact (MI) scores to predict their respective impact on quality. MI scores are also used as size variables in selecting follow-up samples and to prioritize those samples for manual edit. Since each unit has an MI score for each key variable and quality indicator considered, a global MI score (Hedlin, 2008) has to be derived to combine them into a single size variable.

There are two components of active management in the IBSP model: active management of collection and active management of analysis. Active collection management seeks to make the best use of resources for non-response and failed-edit follow-up (Godbout, Beaucage and Turmelle, 2011). Active analysis management prioritizes units for manual review in a selective editing process (Brundell, 2011). The IBSP model proposes using MI scores to draw a random sub-sample of influential units for non-response follow-up and to prioritize units and domains for failed-edit follow-up and manual revision.

Table 5 describes the generalized methods and tools available for quality indicators and active management for the processes identified in the business process model (Figure 6 appended). Most of the quality indicator and active management methods will be specific to the IBSP system, except for the sub-sample for non-response follow-up, which will use certain methods in the G-Sam generalized system for initial sampling.

Table 5
Generalized methods and tools for quality indicators and active management

Business process steps	Statistical methods available and/or recommended	Generalized tools
6.1.1.1- Calculate quality indicators	Synthesis of quality measures from previous steps	IBSP system
6.1.1.2- Calculate MI scores	Disaggregation of quality indicators by linearization	IBSP system
6.1.2.1- Compare QI to targets	Derived variables calculations	IBSP system
6.1.2.2- Derive global MI scores	Derived variables calculations	IBSP system
6.1.4.1- Stratify and allocate follow-up sub-sample	Multivariate power allocation	G-Sam
6.1.4.2- Select follow-up sub-sample	Stratified Bernoulli / Stratified SRS / Poisson; Permanent random numbers	G-Sam
6.1.3- Generate Diagnostic Data	Table aggregation	IBSP system

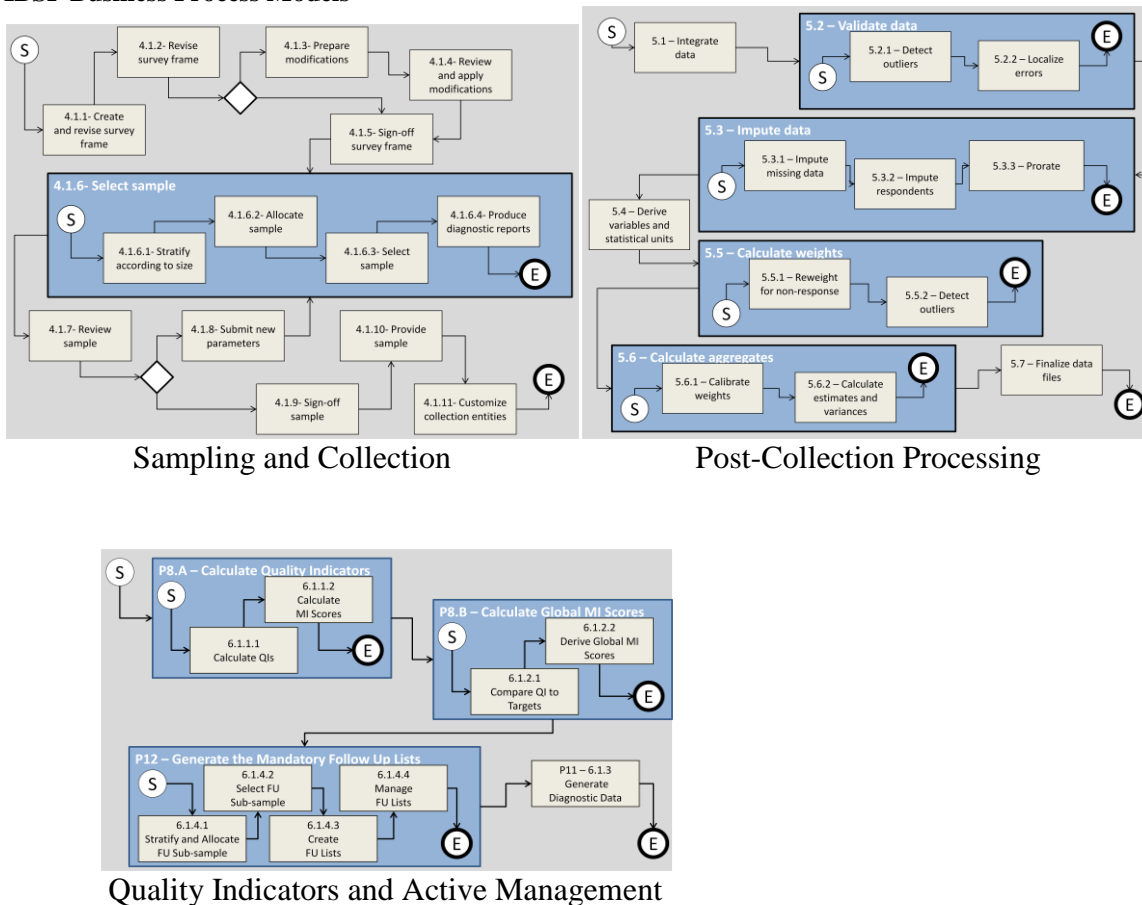
Conclusion

Standardization is a key element of the IBSP given the complexity of processing, the specificities of each integrated survey and the flexibility required for ad hoc surveys.

Business process models establish a complete list of the statistical activities to be accomplished. The roles of each partner are clearly defined along with a common language. Generalized systems and generic integrated services become core elements of the processes. They ensure high quality development, support and maintenance of components often deemed more complex or specialized. Metadata make it possible to effectively manage the complexity and specificities. They harmonize definitions, ensure ongoing documentation and enhance the effectiveness, flexibility and coherence of the data. Metadata also serve as the link between the IBSP's specific systems and the generalized systems. The statistical methods proposed for the different surveys are thoroughly evaluated and selected based on several considerations, not only the accuracy of estimates.

Appendix – Business process Models

Figure 6
IBSP Business Process Models



References

- Blechar, M., Beyer, M.A., Thompson, J., Lapkin, A. and N. Gall (2010), Gartner Clarifies the Definition of Metadata, 2H10-1H11, Gartner Research, August 19th 2010.
- Brundell, P. (2011), “Selective data editing and its implementation at Statistics Sweden”, 2011 International Methodology Symposium, Ottawa, Canada.
- Deguire, Y., Reedman, L. and M. Wenzowski (2011), “Generalized Systems: The Statistics Canada Experience”, 2011 International Methodology Symposium, Ottawa, Canada.
- Godbout, S., Beaucage, Y. and C. Turmelle (2011), “Achieving Quality and Efficiency Using a Top-Down Approach in the Canadian Integrated Business Statistics Program”, Conference of European Statisticians, Ljubljana, Slovenia.
- Hedlin, D. (2008), “Local and Global Score Functions in Selective Editing”, Conference of European Statisticians, Work session on statistical data editing, Vienna, Austria.
- Mohl, C. (2007), “The Continuing Evolution of Generalized Systems at Statistics Canada for Business Survey Processing”, International Conference on Establishment Surveys III, Montréal (Canada).
- Saint-Pierre, É. and M. Bricault (2011), “The Common Editing Strategy and the Data Processing of Business Statistics Surveys”, Conference of European Statisticians, Ljubljana, Slovénia.
- UNECE Secretariat (2009), “Generic Statistical Business Process Model, Version 4.0”, joint UNECE/Eurostat/OECD work session on statistical metadata (METIS).
- Statistics Canada (2005), GES V4.3 – User Guide, Ottawa (Canada).
- Statistics Canada (2006), “Generalized Sampling System version 2.3, Guide de l'utilisateur”, Business Survey Methods Division, Ottawa (Canada).
- Statistics Canada (2009), “Guide sommaire sur le Registre des entreprises”, Business Register Division, Ottawa (Canada).
- Statistics Canada (2010), “Integrated Business Statistics Program Blueprint”, Business Survey Methods Division, Ottawa (Canada), internal document.
- Statistics Canada (2011), “Banff – Functional description of the Banff system for edit and imputation version 2.04”, Banff support team, Ottawa, Canada.

Standardization of processes

Frank Hofman, Astrea Camstra and Robbert Renssen^{1,2,3}

Abstract

Over the past five years, Statistics Netherlands has been working on an ambitious program to redesign the statistical process. The general ideas of this program are represented in a comprehensive enterprise architecture. More recently, the architecture has been complemented by a series of standard methods and standard tools that should facilitate the design of the production process. Before standard methods or tools can be readily applied in the production of statistics, the statistical processes themselves need to be standardized to a certain extent. For this purpose, a conceptual business model for processing statistical data was developed. An important concept in this model is the standard process step. Standard process steps correspond to applications of statistical functions which can be implemented as business services. Statistical functions are usually based on standard statistical methods. By identifying these standard steps and providing guidelines for their use, the model aims to close the gap between the high-level view taken in the business architecture and designing statistical processes in practice. This paper briefly discusses the model and its concepts. The model will be illustrated by applying it to the field of data-editing.

Key Words: Standard process step; Functional component; Building block; Function.

1. Introduction

Statistical institutes are under constant pressure to improve efficiency and reduce reporting burden, in particular for businesses. In addition, they are facing demands to maintain high quality standards, to enhance flexibility, and to focus more on rapidly changing user needs and product innovation. To meet these contrasting objectives, Statistics Netherlands has been working on an ambitious redesign program (see Braaksma 2009, for an overview). The general ideas of this program are embedded in comprehensive enterprise architecture (*e.g.*, Huigen *et al.* 2009).

One of these general ideas is to promote reuse. The reuse of data enables more efficient production of statistics, whereas the reuse of methods, processes and tools contributes to more efficient (re)design of statistics. Recently, the architecture has been complemented by a series of standard methods and a set of standard tools that can be used to further streamline the core production process. The methodology series is a catalogue of approved statistical methods presently used at Statistics Netherlands, and is meant to inform and assist statisticians in applying the correct methods. Eventually, all statistical processes should only (re)use methods included in the series. The need for standard tools originates from the continuing high IT maintenance costs caused by the great diversity of (often tailor made) tools. A study evaluating the current tools in statistics production (Renssen, Wings and Paulussen 2008) has resulted in a preliminary list of 18 preferred tools for the domain of statistical data processing, the two most important criteria for making the shortlist being the possibilities of the tool to deal with metadata and its ability to separate design from implementation.

Before standard methods or tools can be applied in the production of statistics, the statistical processes also need to be (partly) standardized. Renssen *et al.* (2009) give some initial ideas for the standardization of statistical processes. In Renssen (2010) these ideas are extended to a general conceptual business model for data processing. An important concept in this model is the so-called standard process step. Standard process steps correspond to applications of statistical functions implemented using a standard statistical method. There may be several different methods that can achieve the same functionality, for example, hot deck and nearest neighbour are two methods used for

¹Frank Hofman, Astrea Camstra and Robbert Renssen, Statistics Netherlands, Henri Faasdreef 312, 2492 JP The Hague, The Netherlands. f.hofman@cbs.nl, a.camstra@cbs.nl, rh.rensen@cbs.nl.

² The views expressed in this paper are those of the authors and do not necessarily reflect the politics of Statistics Netherlands.

³Link to the methods series (in English): www.cbs.nl/en-GB/menu/methoden/gevalideerde-methoden/default.htm.

imputation. Alternatively, a specific method can be applied for different functions. A regression method for example, can be used for imputing data values or for estimating population totals. The idea is to design processes in terms of standard process steps which then serve as the link between methods and tools. By creating a repository of standard process steps, the transparency and flexibility of the design process will increase and the reuse of methods and processes (and ultimately tools) will be facilitated.

Section 2 describes the theoretical models that have been developed at this stage of the research. These models have been applied to the field of data-editing, which is briefly explained in section 3. Finally, section 4 gives some preliminary conclusions and an overview of future research.

This paper is partly based on two previous papers on this research (Camstra and Renssen, 2011 and Renssen and Camstra, 2011b).

2. The conceptual model for data processing

The concept of a standard process step does not stand alone but only becomes meaningful in the context of our model for designing statistical processes. Section 2.1 briefly introduces this model, describing the first step in the statistical design: the operationalisation of concepts. The second step is to identify the statistical functions needed (section 2.2) and finally the application of these functions in statistical processes (section 2.3). Section 2.4 explains how we aim to standardize the process design by providing generic building blocks. Finally an overview of all concepts and their relations is given in section 2.5.

2.1 Operationalization of concepts

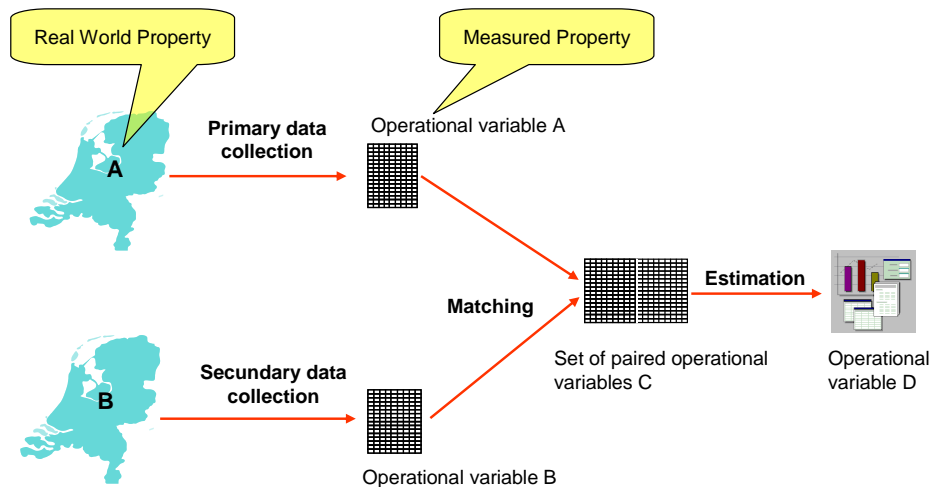
The design of a statistical process begins with determining the statistical information needs and translating them into statistical products. This entails defining the real-world properties to be considered, including the population and time period being reported on. For example, unemployment figures relate to the population of Dutch residents aged 15-65 years, “unemployment” is defined as “working less than 12 hours per week and actively looking for (more) work” and the reference period is a calendar month.

The property “unemployment” is called a conceptual attribute variable. To measure a conceptual attribute variable, it must first be translated into an operational attribute variable. An operational definition describes the actions required to measure the concept, including classification and measurement scales. For example, unemployment can be operationalized by means of one or more questions in a survey.

In the practice of producing statistics, the operationalisation of a statistical concept is generally a multi-step process. In Figure 2.1-1 a very simple example is given. To publish figures on the total number of unemployed by age group (D), the property “unemployment” (A) is measured in a sample using the Labour Force Survey, while person characteristics such as age (B) are obtained from the population register. These measured operational variables are combined into a new measurement C, from which D is finally estimated. Due to data collection errors, sampling errors and/or matching errors the resulting estimate may differ from the conceptual total that had to be estimated.

Frequently, there are several (sets of) operationalisations to achieve the same statistical output (D), although statistical strategies can limit the possible choices. These strategies usually stem from policies for efficiently managing the statistical production processes and often favour a particular solution. The general strategy for data collection at Statistics Netherlands states that every effort must be made to reduce statistical reporting, meaning that primary data collection can only be conducted when there is no alternative (secondary) data source. Hence, in figure 2.1-1 two data sources are shown instead of collecting all data in the survey.

Figure 2.1-1
Operationalization of a variable in several steps



2.2 Statistical functions

The series of operationalisations in figure 2.1-1 already shows the outlines of the design of the statistical process. The next stage in the design is to elaborate these steps further. This means identifying the set of functions needed for each step to realize the end product from one or more input products. In Figure 2.1-1 the second step uses a matching function to obtain C from A and B. In the third step an estimation function is used. Matching and Estimation are examples of statistical functions. These statistical functions are in turn related to statistical methods. The estimation function, for example, is often based on a regression method.

Performing a particular step may require a number of preparatory actions. For matching two data sets it might be necessary to derive and/or encode a set of variables to obtain a unique key variable. This is modeled as applying successively a derivation function, a coding function and a matching function. To obtain a successful matching result several iterations of the matching function may also be needed, using different key variables in each iteration.

Measurements of real world properties may also contain errors or missing data. In addition to measuring the properties themselves, it is therefore important to provide information about the quality of these measurements. Furthermore, we consider quality functions as specific statistical functions that measure quality according to some statistical method or subject matter knowledge.

In Section 3 we will take a closer look at statistical functions, especially those involved in data editing.

2.3 Designing a statistical process

After identifying the required functions, the input, output and method of each function application (the process steps) need to be exactly specified. Traditionally, we described statistical processes in terms of activities (steps) which were sometimes grouped into sub processes. Input and output were mainly described on the level of the sub processes and only occasionally were the activities referring to the methods series for their internal functioning. Although this way of describing processes does provide an overview of a process and serves as the starting point for the development of (IT) tools, it appeared to be quite hard to compare different process designs or to identify possible reusable parts. A study of several recently redesigned statistical processes revealed several causes:

- Similar steps within several processes are quite differently named and described, making it hard to recognize their similarity.

- Applications of statistical methods are not recognized, especially when these applications are trivial. For example, when estimating population totals based on register data, these data are implicitly weighted using weights that are equal to 1.
- Applications of statistical methods may be very complex and need preparatory activities. For example, when using t-1 data in an editing process, these t-1 data should be matched beforehand. In addition, there are procedures in case the t-1 data is not complete.
- Applications of specific tools may need preparatory (non-statistical) activities, like a transformation of format, reshuffling data columns or renaming variables.
- Re-use of data and mixed mode strategies complicate the activities in a process, because several data sources should be combined and processes should be mutually linked.

By standardising the process steps, we aim to speed up the design and ultimately the implementation process. To accomplish this aim, we will establish a library of generic building blocks. By combining and configuring these building blocks one can easily design any specific process.

Figure 2.3-2
Specification of a standard process step

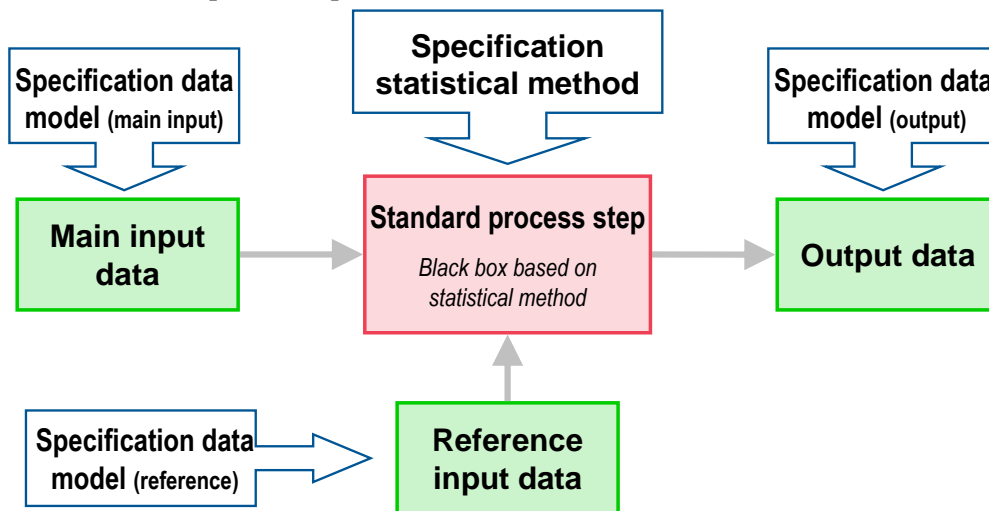


Figure 2.3-2 shows how such a building block is implemented as a standard process step within a specific statistical process. During production it will require main data and possibly some reference data as input to deliver the required output data (all shown in green). To design this standard process step we have to select the right building block and specify the blue elements: the input and output data models as well as the method used. Examples of the latter are the edit rules for an edit function or the matching variables and criteria for a matching function.

We have explicitly distinguished the main input data from the reference (or auxiliary) input data. The main input data is the data that is actually processed or to which value is added during this step, whereas the reference input data is not changed in any way. When imputing, for example, the unit (record) that will be imputed is the main input data, whereas the t-1 data is only used as reference input data. If the reference data is not readily available, we see additional sub processes to gather and prepare this reference data. We have come across additional sub processes that are quite complex, making the entire process less transparent and the design, implementation and/or production more costly. Instead it might be feasible to use a different method, requiring readily (or easily) available reference data, but resulting in similar output. By clearly distinguishing the main data and process flow on the one hand and the reference data and auxiliary sub processes on the other hand, we get a better overview of the entire process. This also enables a better assessment of the quality of the output versus the costs.

The distinction between main and reference input data increases flexibility of the processes. We see the main input and output data to be identical for a function, no matter what method is used. When we replace one method for another one (for the same function), we only need to change the reference input data and method specification. For a simple group mean imputation, for example, we only need one or more auxiliary variables to establish the group and

all valid units (with no missing data) within this group. In contrast to ratio imputations or regression imputations, those require other (and often more) reference data. So we don't have to change the main data flow, only this one process step and possibly auxiliary sub processes for the reference input data.

The idea of designing a statistical process using a library of generic building blocks has been tested in two situations. The first case was a kind of dry run, where we re-engineered the existing, recently redesigned statistical process of the Short Term Business Statistics on paper only (see Renssen and Slootbeek, 2011). The second case was a real-life redesign of the statistical process of some Environmental Statistics, which is currently being implemented. The first case proved it to be possible to apply the concept of the standard process steps. The second case confirmed the aims of more transparency and of the possibilities for reuse even within one process.

2.4 Functional component

In the previous section we have illustrated how we want to use standard process steps to design statistical processes. We aim to establish a repository of building blocks that can be implemented as standard process steps in a specific process. These building blocks are generic, whereas the standard process steps based on them are configured for the use in a specific statistical process. The formal name for a building block is a functional component. A functional component describes its function (added value) and the method used as well as the requirements and restrictions for its specification.

Figure 2.4-3
From a functional component to a standard process step

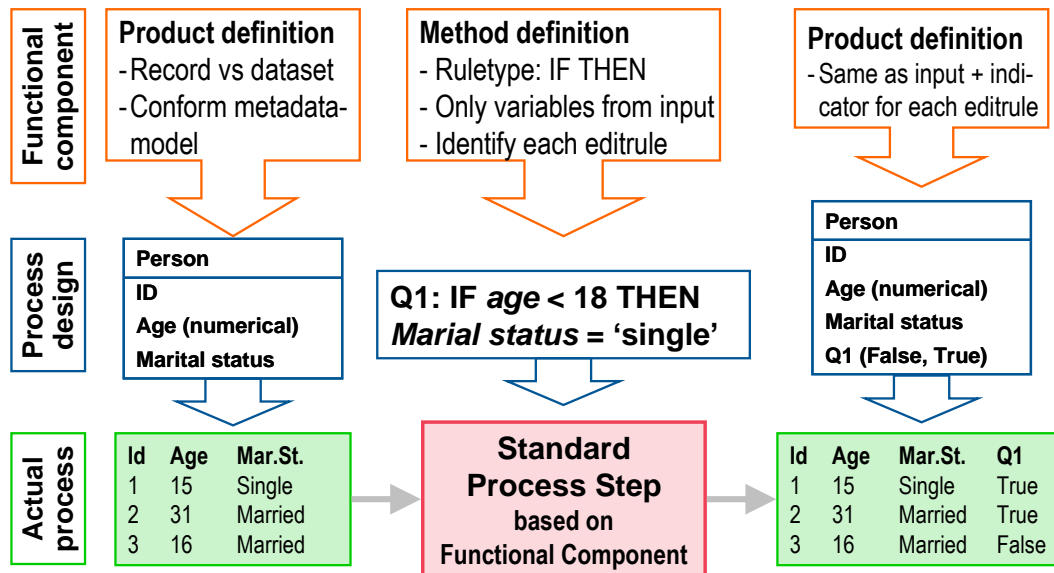


Figure 2.4-3 illustrates how functional components can be used in process design. At the bottom in green, we see the actual production process. Here we have a simple dataset of persons with their age and marital status. We want to validate whether the data in this dataset conforms to the edit rule saying 'people under 18 years cannot be married'. After processing we expect the first two records to be correct and the third to be incorrect.

In the blue middle layer, we design this simple process by picking a functional component for data validation and then specifying the data models of both input and output as well as the edit rule.

The orange top layer shows the requirements and restrictions that come with the functional component and that need to be met during the design. For example this data validation component may process individual units (records), whereas an aggregation component will need a full set of units. Another type of restriction might be the data type of

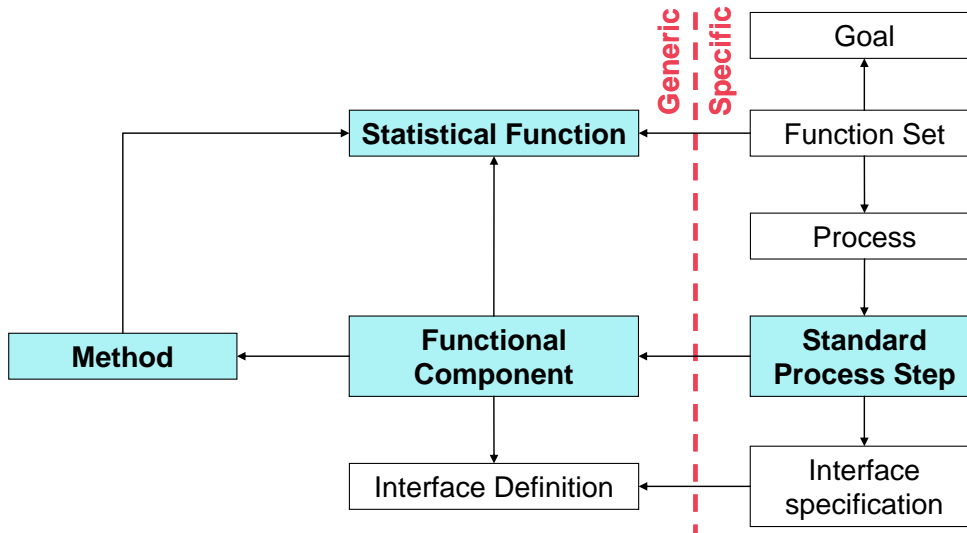
the variables to be imputed (numerical or categorical). For the method specification in the example rules are restricted to the 'IF THEN' type. Further research is needed to elaborate upon these requirements and restrictions.

2.5 Information model

The information model in figure 2.5-4 provides an overview of the main concepts and their relations presented so far. In this section, we briefly summarize the main concepts. For more information we refer to Gelsema and Hofman (2011).

The left side of the figure shows all generic concepts, whereas the concepts on the right side are specific to one survey, for example the Labour Force Survey. The functional component is the central concept. It is the generic building block which may be used to design a specific process. Each functional component implements a statistical function by applying a (statistical) method. Examples of statistical functions are the imputation, derivation or validation functions. Furthermore, for a specific function we may have several functional components corresponding to the different methods that may be used, like the nearest neighbour or regression method for the imputation function.

Figure 2.5-4
Information model



At the right hand side of the figure, we see the concepts used when designing a specific statistical process. At the highest level, each process has one or more goals, generally the output to be produced (or the information needs that need to be satisfied). Knowing the input and output of a statistical process, we can draw up a rough sketch of the statistical functions we will need. Only then we start the actual design of the (statistical) process, choosing the functional components (thus the methods we will use) and then configuring these components for their specific use, turning the components into standard process steps.

The interface specification describes the configuration of each standard process step, like the data models of the input and output and the (further) specification of the method. The generic interface definition of each functional component describes the requirements and restrictions for the specification.

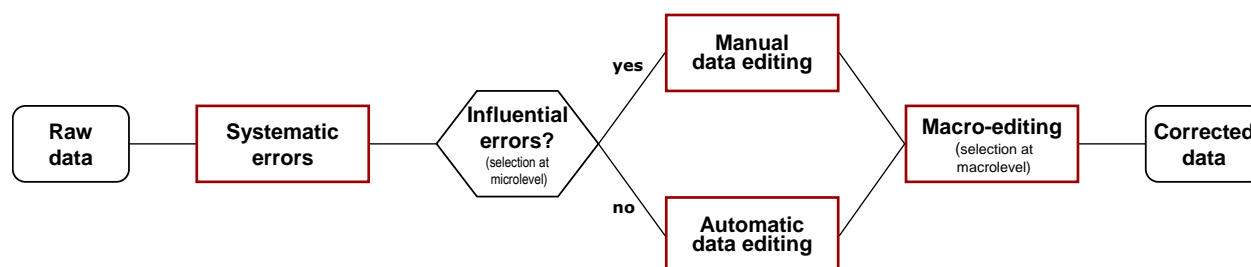
3. The model applied to data editing

In this section we will apply the main concepts of the model described in the previous section to the processing domain of data editing. Section 3.1 discusses the strategy for data editing used at Statistics Netherlands, while in Section 3.2 an example of a very simple data editing process is given.

3.1 Strategy for statistical data editing

The subject “Data Validation and Correction” of the methodology series (Hoogland *et al.* 2010) discusses the data editing techniques most frequently used at Statistics Netherlands. The authors describe an overall strategy for the data validation and correction process as shown in figure 3.1-5. The specific way this strategy is applied for different statistical processes may vary and not all steps have to be completed.

Figure 3.1-5
Data editing strategy



In the first step of the data validation and correction process “obvious” systematic errors are detected and corrected. An example of a systematic error is a ‘thousand-error’, *i.e.*, a value that is wrong by a factor of 1000. If the systematic errors are known they can be easily corrected using deductive methods. One of the next steps is to manually check and correct the data. Given the general data editing instructions, the subject matter specialist determines which data are wrong and how they should be corrected. Corrections are usually made on the basis of expert knowledge, possibly supplemented with reference data. Since manually checking and correcting data is costly and time consuming, this is often restricted to influential errors that cannot be reliably solved automatically. To determine which errors should be processed manually and which automatically, a score is assigned to each unit indicating the expected impact on publication figures if this unit were to be manually corrected (this is also called selective editing) High scores have a high priority to be examined interactively. The remaining less important errors can subsequently be edited automatically. A set of data values of a unit are checked against a predefined set of edit rules and the wrong values(s) are automatically located. At Statistics Netherlands, error localisation methods are frequently based on the Felligi-Holt paradigm. In the last step, provisional publication figures are estimated and compared with historical data or external data sources. If the estimated figures are implausible, the underlying micro-data are analysed further and corrected if necessary. This process of macro-detection and micro-correction is called macro-editing. Macro-editing can be interpreted as a form of selective editing where the selection of influential errors is made through the population estimates.

3.2 Statistical functions involved in data editing

In the different techniques covered by the overall editing strategy described in the previous section, a number of basic data editing operations or functions can be distinguished. We define the following four editing functions:

- Data validation function: checks variables for errors or inconsistencies.
- Error localisation function: determines, in case of inconsistencies, which variable is ‘wrong’ and needs to be corrected.
- Score function: identifies influential observations, *i.e.*, observations that have a substantial impact on publication figures.
- Correction function: corrects the (located) errors or inconsistencies.

The four functions are sufficient to design a process that implements the data editing strategy discussed in the previous section, although one step in the strategy generally requires several functions. For example, the processing of systematic errors requires the validation, localisation and correction functions, whereas the determination of influential errors requires the validation and score functions.

Note that although these four functions cover the entire strategy, different methods can be used for each function, so more than four functional components will be needed. However, the full elaboration of this strategy into functional components takes too far for this paper.

4. Conclusions & future research

So far, the ongoing research on standardization of statistical processes has shown the idea to design statistical processes from a repository of generic building blocks to be feasible. The aim of more transparency in process design appears to be confirmed by two case studies. At this stage of the research, the aims of flexibility and reuse cannot yet be proven, although some reuse even within one statistical process has been accomplished.

After focussing on the functions involved in data editing we will expand our scope. We have already identified about 15 functions that are most frequently used in statistical processes (see Renssen and Camstra, 2011b). These will need further scoping and elaboration to describe them as functional component.

The purpose of the present research is primarily descriptive and we are not directly concerned with the physical implementation of standard process steps and its relation to the standard toolbox. This will be addressed at a later stage in collaboration with the IT-architects. The ultimate goal is to build a repository of functional components (on paper) as well as IT components, which form the basis for both the design and implementation of statistical production processes.

References

- Braaksma, B. (2009), Redesigning a Statistical Institute: The Dutch case, *Proceedings of MSP2009, workshop on Modernisation of Statistics Production 2009*.
- Camstra, A. and R. Renssen (2011), Standard process steps based on standard methods as part of the business architecture, *Proceedings of World Statistics Congress (ISI2011)*, Dublin, Ireland.
- Gelsema, T. and F. Hofman (2011), Conceptual Information Model for Standard Process Steps, unpublished report (in Dutch), The Hague, The Netherlands: Statistics Netherlands.
- Hoogland, J., Van der Loo, M., Pannekoek, J., and S. Scholtus (2010), Methodology series, theme Data validation and correction, unpublished report, The Hague, The Netherlands: Statistics Netherlands.
- Huigen, R., Bredero, R., Dekker, W., and R. Renssen (2009), Statistics Netherlands Architecture; Business and Information model, unpublished report, The Hague, The Netherlands: Statistics Netherlands.
- Renssen, R., Morren, M., Camstra, A., and T. Gelsema (2009), Standard processes, unpublished report, The Hague, The Netherlands: Statistics Netherlands.
- Renssen, R. (2010), Basic principles of a conceptual business model for data processing, unpublished report (in Dutch), Heerlen, The Netherlands: Statistics Netherlands.
- Renssen, R. and A. Camstra (2011a), The data validation function, unpublished report (in Dutch), Heerlen, The Netherlands: Statistics Netherlands.
- Renssen, R. and A. Camstra (2011b), Standard process steps in statistics, *Proceedings of the Meeting on the Management of Statistical Information systems (MSIS 2011)*, Luxembourg.
- Renssen, R. and M. Slootbeek - Van Laar (2011), The conceptual model for data processing applied to the Short Term Business Statistics, unpublished report (in Dutch), Heerlen, The Netherlands: Statistics Netherlands.

Renssen, R. Wings. J. and R. Paulussen (2008), Processes, methods and tools, unpublished report (in Dutch), Heerlen, The Netherlands: Statistics Netherlands.

Vale, S. (2009), Generic Statistical Business Process Model, Version 4.0. Joint UNECE/Eurostat/OECD paper presented at Statistical Metadata (METIS), Luxembourg.

Coding of survey responses – quality assurance efforts and IT tools at Statistics Sweden

Jörgen Svensson¹

Abstract

Coding of survey responses is time-consuming, costly and error-prone. This presentation is about the endeavors of Statistics Sweden to assure the quality of the coding process. Work has been done on a wide front.

Verification of the coding is being implemented 2011-2012 in all relevant surveys. For computer-assisted (manual) coding, quality control comprises of independent verification of the coding, conducted for at least five percent of the coded records in a survey. If the original code and the verification code differ, the 'correct' code is decided by an adjudication (reconciliation) process. If a coder's work contains frequent errors, that coder's work shall be verified and suitable training shall be given. For automatic coding, an analogous quality control will be made at least once every third year, in an appropriate survey. If the error rate is unacceptable, revisions of the dictionary should be carried out.

Moreover, an instruction for developing code frames (when not having standard classifications) has been established. All classifications and code frames are stored in a common repository. The different staff, in a wide sense, involved in the coding process have to cooperate in order to quality assure the process. To that end, the different roles have been pointed out clearly. Because the coding of open-ended responses is often subjective in nature, proper training is essential. The coders at Statistics Sweden are trained on how to use classifications, reference files *etc.* through awareness of rules regarding what should be included or excluded from a given code. The training is done through a web-based application connected to a database with exercises, which are given in three degrees of difficulty.

A modern IT tool for computer-assisted coding has been developed, in close cooperation with the coders. Support for coding decisions is given, working with a user-friendly interface. Functionality for handling access rights for original coders, verification coders and 'adjudicators' is also available.

Information on the coding process, required quality assurance measures and the IT tool for computer-assisted coding is given in the intranet process information system. Appliance will be followed up via the directors of the subject matter and data collection departments.

Key Words: Coding; Computer-assisted coding; Coding errors; Code frames; Standardization; Verification coding; Reconciliation.

1. Introduction

Coding is in this context defined as follows: to use codes to classify survey objects into different categories according to an established classification or another predefined code frame. This paper is about the endeavors of Statistics Sweden in recent years to assure the quality of the coding process. There have been several reasons for focusing on this process. One reason is that the process of coding survey responses is error-prone. This has been shown by evaluations in 2007 of the occupation coding at Statistics Sweden. Regular quality controls are needed. Another reason is that coding is time-consuming and thus costly. Rationalization through modern IT support is required. The director-general decided in 2008 that Statistics Sweden shall work towards certification according to the international standard ISO 20252 for market, opinion and social research, see International Organization for Standardization (2006). The standard requires quality assurance and quality control of the coding process. The requirements are quite specific in terms of percentage of records that shall have verification coding. How these requirements could be fulfilled at Statistics Sweden has been investigated thoroughly. The overall objective for the broad work described in this paper is to achieve continuous improvements for the coding process.

¹Jörgen Svensson, Statistics Sweden, Process Department, Örebro, SE-701 89, Sweden, e-mail: jorgen.svensson@scb.se.

2. Quality control of coding

Statistics Sweden has established a standard routine for quality control of the coding process, which is described in the intranet process information system. The main operations are to conduct *independent verification coding* after the original coding and – if the original code and the verification code differ – to decide upon the ‘correct’ code by a *reconciliation* (adjudication) process. The survey manager, in consultation with the process owner for the coding process, then has to stipulate what is to be considered *unacceptable error rates*, whereupon measures are to be taken to achieve better quality. This routine was implemented in a majority of the relevant surveys during 2011.

Automatic coding using a dictionary shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for at least five percent of the automatically coded records in a suitably selected paper questionnaire survey. This quality control shall be performed at least once every third year, starting 2011. If the error rate is unacceptable, the dictionary has to be revised. (The amount of automatic coding is relatively small at Statistics Sweden.)

Coding at telephone interviews using a list of occupations within a CATI system shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for the Labour Force Survey for a three months period (restricted to records the first time they appear in this longitudinal survey). This quality control shall be performed at least once every third year, starting 2011. If the error rate is unacceptable, the survey manager, in consultation with the process owner, has to decide on what measures should be taken to improve the results.

Computer-assisted manual coding using a comprehensive reference file shall be quality controlled by conducting an independent computer-assisted manual verification coding and reconciliation for at least five percent of the coded records in each relevant survey. This quality control is a form of acceptance sampling and shall be performed continuously, starting 2011. Often the percentage will be higher than five percent, since it must be possible to measure the coding results for each coder. If a coder’s work contains frequent errors, that coder’s work shall be verified and suitable training shall be given. Efforts might be needed to raise the competence both on the individual level and the group level. Apart from training, the reference files and instructions can be improved in order to reduce the error rates in coding. Note that the limit of five percent is not statistically motivated, but a requirement for certification according to ISO 20252.

External coding is a common alternative at Statistics Sweden. Either the data providers perform the coding themselves and send us the coded records, or the data providers code the records according to their own nomenclatures and send us the records, whereupon translation keys are utilized to set the classification codes. For external coding it is not possible to conduct quality control according to the routines above. However, evaluations and quality assurance are recommended for the concerned surveys.

Sampling for verification coding can be designed in different ways. One easy alternative is simple random sampling. Another alternative is stratified simple random sampling with strata equaling the coders. A third alternative is stratified systematic sampling, which is relevant to use when verification coding should be conducted simultaneously with the ordinary coding within the survey production period.

The quality control has a twofold *purpose*. On one hand, the coding process should be improved so that the errors will be less frequent the next survey round. On the other hand, the survey data should (time permitting) be adjusted directly, either through corrections of the erroneous records by using the final, possibly reconciled codes from the sampled verification records or possibly through adjustments to the population level using these corrected codes. To start with, the first purpose will be aimed at in most cases.

Deficiencies in the coding results can be shown in two dimensions. First, error rates for *each coder* are calculated directly, using unweighted data on the final codes from the sampled verification records. The result is used for follow-up of coders and groups of coders with frequent errors. Secondly, *gross errors* and *net errors* are calculated in order to show the effects on microdata (important for analyses of statistical associations and flows between classes) and macrodata (descriptive statistics), respectively. These errors can be produced both unweighted for the sampled verification records and weighted for the population. An issue here, yet to be resolved in practice, is the definition of an unacceptable error rate. Product-specific features must then be taken into account. Different levels of

for instance the classification of occupations ISCO, *i.e.*, different number of digits in the codes, would lead to different requirements on error rates. A process control approach could be adequate. The error level is then monitored through the acceptance sampling. A control chart might be useful; it shows when the error rate exceeds normal random variation or drifts away to too high levels.

3. Prisma – a new IT tool for computer-assisted coding

A *modern IT tool*, named Prisma, for computer-assisted coding was developed at Statistics Sweden during the period November 2010 – April 2011. The development was necessary, since the IT applications used for the Labour Force Survey and several other surveys have been almost out of date and too dependent on a few persons. Another necessity was to build functionality for verification coding and reconciliation, according to the new standard routine described above. The functionality needed could not be found in any commercial IT tool or in any tool used at statistical offices.

Prisma is programmed in C#.NET and will support coding for all different classifications and (almost) all surveys. Classifications are easily updated and loaded again. The tool provides a user-friendly interface for the coders, which is important for their working environment. Prisma will rationalize the work of the coders by giving support for coding decisions. When a new record is opened, an automatic search through reference files *etc.* is done. If there is an adequate hit, the coder just proceed by a simple click on a suitable category. Then all codes for different classifications within an area, such as occupation, are set automatically. However, sometimes there are no hits or a lot of hits, and the coder has to search for more information about the individual and the occupation. Difficult records can be placed on a waiting list. Scanned images of questionnaires are possible to import. Functionality for handling access rights for original coders, verification coders and ‘reconcilers’ is also available. A few sampling techniques are supported, to start with. Process data are generated. Configuration for different classifications and surveys is done through a module within Prisma.

The development project has been carried out in close cooperation with the coding staff, which has been testing the software in different early stages. Their wishes and requests have been integrated to a large extent in the specification of requirements for Prisma. Seminars have been arranged to present the IT tool and the standard coding process. In May 2011, the director-general approved Prisma as a standard tool for computer-assisted coding.

The plan was and is still to *implement* Prisma in all relevant surveys. Among these 10–15 surveys, the Labour Force Survey is the most important one and also the first one where Prisma was implemented (in April–May 2011). The classifications used in this survey are ISCO for occupations (in two versions), the Swedish socio-economic classification, the Swedish standard industrial classification, sector and county.

4. Training and other quality assurance measures

Due to the fact that coding of open-ended responses often is subjective in nature, proper *training* is essential. The coders at Statistics Sweden are trained on how to use classifications, reference files *etc.* through awareness of rules regarding what should be included or excluded from a given code. The training is done through a small web-based application connected to a database with exercises, which are given in three degrees of difficulty.

The different co-workers, in a wide sense, involved in the coding process have to cooperate in order to assure the quality of the process. To that end, the different *roles* have been pointed out clearly via a decision by the director-general. Some of the roles are: coder, interviewer, survey manager, classification owner, process owner (of Process & Analyse) and IT staff responsible for Prisma. The classification owner, for example, will have to revise the classification dictionary if the error rate is unacceptable according to the quality control.

A short *instruction for developing code frames*, when not having standard classifications, has been established. Among other things, it points out how to construct categories and instructions, and how to treat ‘other’ (catch-all) categories. The survey manager is responsible for producing, documenting and possibly revising the code frame in accordance with the instruction. All classifications and code frames shall be stored in a common repository.

The coding staff is mainly *centralized* to one of the two data collection departments. A full centralization is almost finalized. The rationale for the centralized approach is to have a strong group of coders that work in a similar fashion with a common IT tool. Dialects among coders in different surveys should be avoided as far as possible, so that consistency and comparability can be achieved. Moreover, ‘coder variance’ should be minimized by as far as possible employing many coders for each survey instead of a divided coding group with few coders per survey.

Information on the coding process and required quality assurance measures is given in the intranet process information system. Application will be followed up annually via the directors of the subject matter and data collection departments.

5. Conclusions and future work

Extensive work has been done throughout the last years in order to improve the process of coding survey responses at Statistics Sweden. The main task now is to implement the adopted methodology and routines as well as Prisma, our new IT tool. Analyses of error rates per coder and gross and net errors in statistics will be made, and measures must be taken to improve the coding process. Awareness and competence regarding quality assurance of the process have to be increased among all parties concerned. Adopted routines have to be evaluated.

In 2012, there will hopefully be a new project for further development of Prisma. The whole specification of requirements could not be met in the recent project. The use of Prisma during 2011 has led to new functionality requests from the coders. Automatic coding using a dictionary might be included in the next version of Prisma. There is also a need for better connections between Prisma and other standard IT tools at Statistics Sweden. Monitoring of telephone interviews is scheduled to be implemented in 2012 and could lead to a partially different routine for quality control of coding at telephone interviews.

References

International Organization for Standardization (2006), Market, opinion and social research – Vocabulary and service requirements (ISO 20252:2006, IDT).

SESSION 8B

QUESTIONNAIRE DESIGN AND COLLECTION MODE EFFECTS

Survey data quality provisions in Statistics Canada E-Questionnaire solution: Retrospective and perspectives

Yamina Abiza¹

Abstract

The online data collection mode is relatively new and the full potential, benefits and methodological issues of using it, alone or in multi-mode (*i.e.*, different modes for different respondents), are still being explored. Nevertheless, with regards to survey data quality this collection mode presents a strong potential to effectively efficiently exceed or at least match what is achieved in the other well established modes. This potential is only going to increase in the future with the advances in technology and survey's methods research.

During the last three or four years, Statistics Canada Collection has developed and operated a fully functional online data collection solution that was used in collecting respondent data for Statistics Canada surveys, where the online collection mode was used as the main data collection channel, in combination with other modes (*e.g.*, paper and computer-assisted telephone surveys).

In this presentation, we describe the Statistics Canada e-Questionnaire Solution architecture and its features that contribute to different dimensions of survey's data quality. We also describe how we could further enhance our solution by integrating new survey's methods research findings on the best practices about the use of online surveys in multi-mode data collection.

¹Yamina Abiza, Statistics Canada.

Designing a questionnaire to examine the Canadian Forces sports program

Krystal K. Hachey¹

Abstract

Physical fitness is one of the essential requirements for serving in the Canadian Forces (CF). Given its importance, there are programs offered for CF personnel that enable them to meet and maintain the standards set by the CF (Hillier, 2009). One of these programs is the CF Sports Program; which despite its importance, has yet to be evaluated. The current paper is part of a larger study that will examine the satisfaction and participation of CF personnel with regards to the CF Sports Program. Both participants and non-participants will be sampled. There are five overall purposes of the project: (1) To examine the type of individual who participates in the CF Sports Program; (2) To examine the reasons why individuals participate in the CF Sports Program; (3) To examine the overall benefits of the CF Sports Program; (4) To examine if the CF Sports Program adheres to its main objectives; and (5) To examine participants' overall satisfaction with the CF Sports Program. The intent of this report is to present the methodological approach for addressing these five goals including the development of the questionnaire. The presentation will cover the questionnaire development, sampling, as well as any methodological issues that arose.

1. Introduction

Physical fitness is an integral aspect of military life. Previous research has indicated that participation in sport provides a basis for individuals to work on personal characteristics (*e.g.*, leadership and team cohesion; Alimo-Metcalf & Alban-Metcalf, 2001), in addition to the benefits to physical (*e.g.*, flexibility) and psychological health (Fentem, 1994). Although there are benefits associated with regular physical fitness, there are also other unique advantages associated with participation in group sports (Pate, Trost, Levin, and Dowda, 2000). Research has revealed that participation in organized group sports promotes fair play, competitiveness, achievement (Pate et al., 2000) and social connectivity (Long, 2004; Sherry, 2010). This is particularly important for developing bonds with other individuals working in the same area or on the same emergency task (*e.g.*, sea evacuation) (Defence Administrative Orders and Directives [DAOD] 5023-2, 2010).

The Canadian Forces (CF) Sports Program is an essential part of the CF, as it is an avenue for the training and development of CF members (Canadian Forces Administrative Orders [CFAO] 50 3, 2010). CF personnel have the opportunity to participate in sports at the base (*i.e.*, intramural), regional, national and international levels. Some of the main objectives of the program are to develop unit cohesion, team work, as well as individual attributes such as self-esteem, self-sacrifice and leadership, while also promoting physical fitness (CFAO 50 3, 2010). Overall, the CF Sports Program provides a means for CF members to participate in competitive sports, while also giving them the opportunity to further develop important characteristics that are integral to the CF.

As there are necessary physical requirements that CF personnel must maintain, it is vital to be able to offer physical activity programs. Thus, the aim of the current project was to examine the significance of the CF Sports Program for recruitment, retention, fitness levels and physical and psychological benefits, as well as to assess satisfaction with the program.

¹Krystal K. Hachey, Department of National Defense and the Canadian Forces, Canada.

2. Methodology

2.1 Research Questions

The primary research questions included the following:

1. Can the CF Sports Program be used as a recruitment tool?
2. Is there any link between participation in the CF Sports Program and retention of CF personnel?
3. Can participation in the CF Sports Program enable CF personnel to maintain their fitness levels?
4. Does participation in the CF Sports Program enable CF personnel to develop personal characteristics sought after by the CF (leadership, team cohesion, *etc.*)?
5. Is there any link between the participation in the CF Sports Program and health benefits for CF personnel?
6. Are participants satisfied with the CF Sports Program?

2.2 Participants and Survey

The stratified random sample included both individuals who had participated in the CF Sports Program and those who had not, in order to provide a basis for comparison of health measures and influences on recruitment and retention. The data were collected by means of an electronic survey, in which participants received an email that included an information sheet, consent form, as well as a link to the survey. The survey was available online to participants from May 2011 until September 2011, with a final response rate of 34%.

2.3 Survey Development

In order to establish the main variables that needed to be addressed, an extensive literature review on participation in sports and physical activity was conducted. Survey items already in use at the Department of National Defence (DND) (*e.g.*, Recruit Health Questionnaire [RHQ]), which were applicable to the CF Sports Program study, were used where possible. This was important in order to have a basis for comparison with other surveys. As no study had conducted an examination into satisfaction and participation of the CF Sports Program, several items were developed to examine these areas. Finally, once a bank of survey items was established, they were categorized under specific sections based on the research questions.

The following sections will review the main components of the questionnaire and describe the items that were chosen to address the specific goals of the project.

2.4 Demographic Variables

The demographic variables included standard variables used in questionnaires administered to CF samples, including age, sex, official language, rank, environmental uniform and Base/Wing support center.

2.5 Research Questions 1 and 2: Can the CF Sports Program be used as a recruitment tool, and is there any link with attrition?

In order to assess whether participation in sports could be further used as a recruitment/retention tool and to determine whether the CF Sports Program has an impact on attrition, six items from the 2010 CF Retention Survey (Holden, 2011; Howe, 2006) were used. Items were measured on a five-point Likert scale ranging from definitely not, to definitely yes (*e.g.*, “I intend to stay in the CF to complete my terms of service”). In addition, other items were developed by the primary researcher to examine recruitment (*i.e.*, whether respondents felt that the CF Sports Program could be used as a recruitment tool), how they found out about the program (*e.g.*, through base newspaper, civilian media, friends), reasons for joining the CF (*e.g.*, moving away from home, CF pay and benefits) reasons for joining the CF Sports Program (*e.g.*, to develop/enhance camaraderie), and barriers to participating (*e.g.*, too busy).

2.6 Research Question 3: Can participation in the CF Sports Program enable CF personnel to maintain their fitness levels?

In order to assess whether the improvement/maintenance of CF personnel fitness levels could be promoted through participation in the CF Sports Program, several items from the RHQ, a baseline health surveillance tool of CF recruits, the Canadian Community Health Survey (CCHS) (Statistics Canada, 2008), and the Health and Lifestyle Information Survey of CF Personnel (HLIS; Decima Research inc., 2002, Defence Force Health prevention, 2003), were used. The items included physical fitness test results, height and weight for the calculation of body mass index (BMI) (CCHS, 2001), self-reported health/mental health (CCHS, 2001), physical injuries (including repetitive injuries) (CCHS, 2001), and physical fitness maintenance strategies. These areas were included because they addressed general physical health, including injuries that may impede the improvement or maintenance of CF personnel fitness levels.

2.7 Research Question 4: Does participation in the CF Sports Program enable CF personnel to develop personal characteristics sought after by the Canadian Forces (e.g., leadership)?

In order to assess whether personal characteristics such as leadership are fostered by the CF Sports Program, two items were developed: (1) “Do you feel that your participation in the CF Sports Program contributes to how you interact with the members of your unit?” (yes/no response); and (2) “Do you participate in the CF Sports Program to improve your leadership skills?” (yes/no response). Both questions provided space for additional comments.

2.8 Research Question 5: Is there any link between participation in the CF Sports Program and perceived health benefits for CF personnel in areas such as depression, anxiety, smoking and alcohol behaviours?

Items from both the RHQ and CCHS were used in order to assess mental health and personal health behaviours. Mental health, including depression and anxiety were measured by the 10-item Kessler Psychological Distress Scale (K10; Kessler *et al.*, 2002). Respondents rated each of the items on a 5-point Likert scale ranging from none of the time to all of the time (e.g., “Did you feel nervous?”). Questions regarding alcohol use and smoking were adopted from the CCHS (CCHS, 2001). Items addressed current smoking status (*i.e.*, never smoked, ex-smoker, or current smoker) as well as frequency of drinking alcoholic beverages.

2.9 Research Question 6: Are participants satisfied with the CF Sports Program?

In order to assess the satisfaction of CF personnel with the CF Sports Program, items were developed based on previous research (Sigrist *et al.*, 2005; Chin, White, Howel, Harland, & Drinkwater, 2006). Respondents were asked to rate their overall satisfaction with different levels of the CF Sports Program (*i.e.*, base sport, regional sport, national sport, international sport, outservice/extreme sport, and other) on a five-point Likert scale ranging from completely satisfied to completely dissatisfied.

3. Conclusion

The goal of the project was to provide a general overview of the CF Sports Program including its influence on retention, recruitment, fitness levels, personal characteristics, and mental health. Although the survey addressed all of the stated goals, there were some limitations. One of the main limitations was that the survey was cross-sectional, which limits the ability to make cause and effect statements. In addition, the qualitative component of the survey was limited on account of constraints on survey length. Despite these limitations, the results of the current study will provide information on the CF Sports Program as it relates to attrition/retention, physical fitness, and personal characteristics that are in line with those sought after by the CF. Since no other study has examined the participation in or objectives of the CF Sports Program, the present study will serve as a basis for future studies on the benefits of the program.

References

- Alimo-Metcalfe, B. and R.J. Alban-Metcalfe (2001), "The development of a new transformational leadership questionnaire", *Journal of Occupational and Organizational Psychology*, 74, 1-27.
- Canadian Forces Administrative Orders (CFAO) 50-3 (2010), "CFAO 50-3 – Sports", unpublished manuscript.
- Decima Research Inc. (2002), "*CF Health and Lifestyle Information Survey 2000 Regular Force Report*" (Report prepared for Canadian Forces Department of National Defence), Ottawa, Canada: Decima Research Inc.
- Defence Administrative Orders and Directives (DAOD) 5023-2 (2008), "Physical fitness program", retrieved online August 13th from: http://admfincs.mil.ca/admfincs/subjects/daod/5023/2_e.asp.
- Directorate of Force Health Protection (2003), "*Canadian Forces Health and Lifestyle Information Survey 2004 Regular Force Report (A-MD-015-FHP/AF-001)*", Ottawa, Canada: Department of National Defence, Directorate of Force Health Protection. Retrieved August, 2010 from: <http://www.dnd.ca/health-sante/pub/hlissv/pdf/AMD015FHPAF001-20030901-eng.pdf>
- Fentem, P.H. (1994), "ABC of sports medicine: Benefits of exercise in health and disease", *British Medical Journal*, 308, 1291-1295.
- Howe, D. (2006), "Building and sustaining a retention culture in the Canadian Forces", DGMPRA TR 2006-006, Ottawa, Canada: Director General Military Personnel Research and Analysis.
- Hyams, K.C., Barrett, D.H., Duque, D., Engel, Jr., C.C., Friedl, K., Gray, G., Hogan, B., Kaforski, G., Murphy, F., North, R., Riddle, J., Ryan, M.A.K., Trump, D.H., and J. Wells (2002), "The Recruit Assessment Program: A program to collect comprehensive baseline health data from U.S. military personnel", *Military Medicine*, 167 (1), 44-47.
- Kessler, *et al.* (2002), "Short screening scales to monitor population prevalences and trends in non-specific psychological distress", *Psychological Medicine*, 32, 959-976.
- Long, J. (2004), "The social benefits of sport: Measurement and evaluation", retrieved online July 16th, 2010 from: http://www.cpl.biz/isrm/infonotesite/recreation/documents/REmay04_16_17.pdf.
- Pate, R.R., Trost, S.G., Levin, S., and M. Dowda (2000), "Sports participation and health related behaviours among US youth", *Archives of Paediatric and Adolescent Medicine*, 154, 904-911.
- Sherry, E. (2010), "(Re) engaging marginalized groups through sport: The homeless world cup", *International Review for the Sociology of Sport*, 45 (1), 59-71.
- Sigrist, L.D., Anderson, J.E., and G.W. Auld (2005), "Senior military officer's educational concerns, motivators and barriers for healthful eating ad regular exercise", *Military Medicine*, 170 (10), 841-845.
- Spitzer, R.L., Kroenke, K., Williams, J.B., and the Patient Health Questionnaire Primary Care Study Group (1999), "Validation and utility of a self-report version of PRIME-MD: The PHQ primary care study: Primary Care Evaluation of Mental Disorders: Patient Health Questionnaire", *JAMA*, 282, 1737-1744.
- Statistics Canada (2001), "Canadian Community Health Survey (CCHS) questionnaire for cycle 1.1", retrieved on 12 December, 2007, from: http://www.statcan.ca/english/sdds/instrument/3226_Q1_V1_E.pdf
- Statistics Canada (2008), "Canadian Community Health Survey (CCHS) 2008 questionnaire", retrieved online June 24, 2010 from <http://www.statcan.gc.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=3226&lang=en&db=imdb&adm=8&dis=2>.

Thompson, M.M. and L.S. Smith (2002), "Peace Support Operations Predeployment Survey: Scale reliability analysis", DRDC CORA TR 2002-190, Toronto, Canada: Defence Research and Development Canada – CORA.

Ware, Jr. J.E. and C.D. Sherbourne (1992), "The MOS 36-item short-form health survey (SF-36), I. Conceptual framework and item selection" *Medical Care*, 30 (6), 473-483.

Harmonized content: The new paradigm in developing surveys at Statistics Canada

Richard Nadwodny and Pamela Best¹

Abstract

Statistics Canada like many other statistical organizations around the world undertakes a wide range of government and special surveys, designed at different times to meet different needs. As a result, over time, a lack of cohesion developed in concepts, definitions, classifications and the survey questions themselves, resulting in disharmonies between survey sources—specifically, the difficulty in comparing data from different survey sources for the same theme/variable.

Improving the cost-effectiveness and flexibility of the household survey program has been recognized by Statistics Canada over the past three decades. In 2005, the *New Household Survey Strategy*—a vision- and priority-setting white paper—was adopted by Statistics Canada to begin the process of transformation to become more responsive to client needs, to be more competitive financially and to be able to provide survey data faster and for a wider range of outputs through greater integration and harmonization of survey content and processes.

Working groups were established to look at various survey operation methodologies such as collection platforms, load levelling, continuous interviewing, redesign of some flagship surveys, common processing tools and the standardization of question modules most commonly used in household surveys.

This paper will describe how two important components of the *New Household Survey Strategy*—standardization of question modules and the development of common processing tools—helped change the paradigm in developing surveys at Statistics Canada.

Key Words: Data harmonization; Question standardization; Common processing tools.

1. Introduction

1.1 Description

There is an old saying that says “Good ideas never go away; some just take longer to come to fruition than others.” Although different titles have been used since the 1970s to describe the topic (data coordination, harmonization, standardization, data/statistical integration) the intent has always been the same: “To reduce the duplication of processing functionality and storage; reduce system maintenance costs, reduce respondent burden and harvest efficiencies that come from using standardized modules of concepts, questions, processes and classifications”.² ‘Data harmonization’ is the current terminology used at Statistics Canada to describe this concept.

In 2005, the *New Household Survey Strategy* (NHSS)—a vision and priority setting white paper—was adopted by Statistics Canada to begin the process of transformation to become more responsive to client needs, to be more competitive financially and to be able to provide survey data faster and for a wider range of outputs through greater integration and harmonization of survey content and processes.

In the decades preceding the NHSS, social surveys conducted by Statistics Canada had been increasingly ambitious, simultaneously seeking breadth of content, depth of content and geographical specificity. The consequence of this

¹Richard Nadwodny, Senior Project Manager, Data Harmonization Project, Special Surveys Division, Statistics Canada, Ottawa, Ontario, Canada, KIA 0T6 richard.nadwodny@statcan.gc.ca; Pamela Best, Assistant Director, Special Surveys Division, Statistics Canada, Ottawa, Ontario, Canada, KIA 0T6 pamela.best@statcan.gc.ca.

²Priest, G (1996) *The Issue of Harmonization of Data from Diverse Sources*, invited paper at the EUROSTAT Workshop on Harmonization, London, England, November 1996.

was that the agency was pushing the limits of respondents' abilities to provide sufficiently precise information and it was making many of our surveys costly and ponderous, difficult to administer and slow to yield results.

The NHSS offered opportunities to position ourselves to conduct simple, small-scale surveys with rapid turnaround by taking into consideration the evolving collection environment demanded by clients to get their data faster and cheaper than in the past. The NHSS business plan architecture was endorsed by senior management and working groups were quickly set to work to analyze and recommend changes to Statistics Canada's household survey development program.

This paper looks at how two important components of the NHSS (**data harmonization** and **common tools**) were developed, their results and their implementation.

1.2 Data Harmonization

At the Methodology Symposium in 1995, Gordon Priest identified a problem in that "methods, systems, concepts, definitions, classifications, products and services were developed independently, resulting in inefficiencies, redundancies, disharmonies and some client frustration."³ He proposed using data/statistical integration as a way to reduce these problems.

Today, we use the term 'data harmonization,' which refers to the process of developing standardized questionnaire modules for cross-cutting household survey variables. These modules contain standard concepts, definitions, classifications and wording for multiple collection modes. It was expected that by using standardized question modules, efficiencies and timeliness would improve through the re-use of Computer Assisted Interviewing (CAI) specifications, testing, processing, documentation and dissemination.

Over time, as a result of questions being asked in different ways, a lack of cohesion developed in concepts, definitions and classifications, resulting in disharmonies between survey sources—specifically, the difficulty in comparing data from different survey sources for the same theme or variable.

The current disharmonies in data analysis could be reduced if clients were assured that different Statistics Canada surveys were using the same questions for a particular theme and the questions were gathered and processed in the same way so the results could be analyzed and compared with confidence.

An example of this disharmony can be seen in this analysis of the dwelling tenure question which was asked in different ways by previous Statistics Canada surveys conducted in 2006. The 2011 National Household Survey used the standardized version of the question, and is shown below for reference.

³Priest, G (1995) Data Integration: The View from the Back of the Bus. *Proceedings of Statistics Canada Symposium 95, From Data to Information – Methods and Systems*, November 1995, Ottawa, Catalogue no. 11-522-XPE.

2006 CENSUS	GENERAL SOCIAL SURVEY (GSS)	SURVEY OF HOUSEHOLD SPENDING (SHS)	LFS, CCHS, SLID	2011 NATIONAL HOUSING SURVEY (NHS)
<p>Is this dwelling:</p> <p>Owned by you or a member of this household (even if it is still being paid for)?</p>	<p>Is this dwelling owned by a member of your household?</p> <p>1. Yes 2. No DK, R</p> <p>Do you pay rent to live in this dwelling?</p> <p>1. Yes 2. No DK, R</p>	<p>On December 31, 2007 was your dwelling:</p> <p>1. Owned without a mortgage by your household? 2. Owned with a mortgage by your household? 3. Rented by your household? 4. Occupied rent free by your household (that is, where no member owns the dwelling and no rent is charged)?</p>	<p>Is this dwelling owned by a member of this household?</p> <p>1. Yes 2. No DK, R</p>	<p>Is this dwelling:</p> <p>1. Owned by you or a member of this household (even if it is still being paid for)? 2. Rented (even if no cash rent is paid)?</p>

LFS: Labour Force Survey, **CCHS:** Canadian Community Health Survey, **SLID:** Survey of Labour and Income Dynamics

As a result, the effect of differences in how the dwelling tenure questions (dwellings owned and dwellings rented) were asked, affected the comparability of data between the different survey sources. Although these differences could be attributed to differences in sample size, collection methodology, interviewer instruction *etc.* (the usual reasons for comparability problems between surveys) eliminating incongruities in the way the questions were worded even though they were asking the same information was an obvious starting point to improving the data quality.

Owner Hhlds	Count	Percent
SHS	8,215,000	66.1%
SLID	8,603,731	68.7%
Census	8,381,125	68.5%
Renter Hhlds	Count	Percent
SHS	4,218,331	33.9%
SLID	3,922,799	31.3%
Census	3,861,155	31.5%

Harmonized content provided off-the-shelf pre-tested question modules, accompanied by pre-programmed, approved and tested processing specifications and BLAISE code, which would allow surveys to be more nimble and flexible in data collection approaches and be able to respond to clients more quickly by reducing the time and cost to get a survey out into the field. Standardized question blocks also proved to be of benefit to computer-assisted telephone interviewing (CATI)/computer-assisted personal interview (CAPI) interviewers who would become accustomed to the wording and sequencing of standard questions, thereby speeding up the response time per question and reduce training time.

The NHSS also created the opportunity to work more closely with the census to share experiences “on a wide range of areas including content, coding and classification systems, frames, technological expertise, management tools, media and respondent relations, and studies on mode effect (the census has a multi-modal response design—mail,

interviewer and internet), which could be a model for Head Office/Regional Office collaboration on household surveys in general.”⁴

1.3 The Process of Developing Harmonized Content

Special Surveys Division was mandated to manage the Harmonization Project, which began with an internal review of our current practices, specifically to identify and document the household surveys (including the census) that asked questions or had question modules on the most commonly used themes, such as marital status, health, education, languages, *etc.* Similarities and/or differences in how these questions were asked, grouped and documented, were noted. The project also looked at any international standards developed by the United Nations and other international bodies. Once the analysis of our current practices and international recommendations were documented, working groups of subject matter experts and stakeholders were assembled to review the documentation and begin the process of recommending standardized question modules for 18 of the most common cross-cutting themes used in household surveys. Rounding out the working group members were analysts from Standards Division who were responsible for developing the standard concepts, definitions and classifications for the recommended standardized questions. The Standards Division metadata was a very important component of this project because, without the metadata, survey or census data could not be disseminated.

Once the subject matter working groups finished their analysis, there were some questions that had several variations for the, which, therefore, required respondent testing. Working with the Questionnaire Design Resource Centre (QDRC), different question formats and variations were qualitatively tested across Canada in focus group and one-on-one testing. After the working group reviewed the results of the testing and made their decision on which questions to recommend for inclusion in the harmonized content program, these recommendations were presented to various levels of senior management, until approval was given to adopt the standards.

2. The Creation of Common Tools

Standardized question modules and its associated metadata, by itself, does not introduce the efficiencies required to make the *New Household Survey Strategy* a success. True savings and efficiencies are introduced by developing a new processing system that is universally used by all divisions within the agency that undertake household surveys.

In this context, the Common Tools Project was created to harmonize the business processes of household surveys to develop the common tools that will allow survey areas to efficiently create, process and disseminate social survey data. The Social Survey Metadata Environment (SSME) was developed around the following four common tools: the Questionnaire Development Tool (QDT), the Processing and Specifications Tool (PST), the Data Dictionary Tool (DT) and the Derived Variables Tool (DVT).

Along with improvements to timeliness and data quality by making use of proven questions, it also promotes coherence across surveys. Eventually, a repository of associated metadata will be made accessible through a component of the interface to gain access to pre-coded blocks.

The SSME was designed around the following user requirements:

- To create questionnaire specifications in a structured environment regardless of collection mode.
- To provide access to a common repository of questionnaire specification blocks for all collection modes (CAPI, CATI, e-questionnaires and paper).
- To promote the use of harmonized content.
- To promote the use of the latest CAI standards.
- To create various types of questionnaire outputs that will improve the timeliness and standardize the process of disseminating metadata for multiple uses in all aspects of the Survey Life Cycle, from pre-collection to dissemination.

⁴ Statistics Canada, (2005). *New Household Survey Strategy: Summary Report* (internal document). Ottawa, Ontario. Oct. 5, 2005.

The common tool most closely associated with the harmonized content project is the Questionnaire Development Tool (QDT), which takes the standardized question modules and allows subject-matter staff to specify and disseminate questionnaires in a timely fashion using a standard approach.

Specifically, the QDT function allows survey developers to

- consolidate, manage, and standardize the work involved in the development of a survey questionnaire;
- enable users to access and develop a questionnaire specification repository;
- create questionnaire specifications in a structured environment regardless of the collection mode;
- provide access to a common repository of questionnaire specification blocks for all collection modes (CAPI, CATI, e-collection and paper);
- create various types of questionnaire outputs to improve the timeliness and standardize the process of disseminating metadata for multiple uses in all aspects of the Survey Life Cycle from pre-collection to dissemination;
- promote the use of harmonized content;
- promote the use of the latest CAI standards;
- monitor the progress of development of surveys;

The QDT is currently equipped to create Blaise specifications for CATI/CAPI collection, but in the future, the QDT will be able to create specifications for Electronic Questionnaires as well as for Paper Questionnaires.⁵

3. Future work in the Development of Harmonized Content and Common Tools

3.1 Standardized Question Modules for EQs

The initial exercise included the development of standardized questions for paper questionnaires and CATI/CAPI surveys. Statistics Canada has already indicated the desire to introduce further cost-savings by developing future surveys in the Internet environment as EQs. In 2006, 18% of census questionnaires were completed using the internet option. This increased to 54% for the 2011 Census/National Housing Survey—a strong indication that, with an Internet promotion strategy, EQs are becoming the preferred survey mode for Canadians. Other surveys, like the Public Service Employee Survey and Survey of Staffing, are already Internet-based with the Labour Force Survey currently pilot testing a monthly LFS EQ and plans are in development for an Internet version of the General Social Survey.

Internet-based surveys are a hybrid of paper and CATI/CAPI question formats. As a result, it is imperative that future work with harmonized content include the development of Internet-based standardized question modules. Finding the balance between maximizing the strength of the mode, i.e. providing interviewer guidance during an interviewer-assisted survey, using interactive help in an EQ and weighing this against the comparability with paper will be one of the main challenges in developing this content.

3.3 Five-year Review and Updating of Harmonized Content

Statistics Canada made the decision that standardized questions should be reviewed for their continued relevance and that this review should be done on a quinquennial basis concurrent with the census program.

This review could add, change or drop standardized questions from the database. Any changes or additions would require the same methodology of investigation as employed in the first round of developing harmonized content, that is, establishment of subject matter working groups, international comparisons, qualitative testing and hierarchical decision making committees, *etc.*

⁵Statistics Canada, (2011). *Common Tools Project for Social Surveys: Bringing it all together* (internal document). Ottawa, Ontario, June 2011.

References

- Colledge M. (1999), *Statistical Integration through Metadata Management*, Statistical Directorate, Organization for Economic Cooperation and Development, International Statistical Review, Paris France.
- Moore. T., Bailie. L. and G. Gilmour (2009), Building a Business Case for Census Internet Data Collection, *Proceedings of Statistics Canada Symposium 2008, Data Collection: Challenges, Achievements and New Directions*, Ottawa, Catalogue no. 11-522-X.
- Priest, G. (1995), Data Integration: The View from the Back of the Bus. *Proceedings of Statistics Canada Symposium 95, From Data to Information – Methods and Systems*, Catalogue no. 11-522-XPE, Ottawa, Ontario. November 1995.
- Priest, G. (1996), The Issue of Harmonization of Data from Diverse Sources, *invited paper at the EUROSTAT Workshop on Harmonization*, London, England, November 1996.
- Priest, G. (1998), Report on the Progress on the Harmonization of Social Statistics - Working Paper 3. *Conference of European Statisticians, Statistical Commission and Economic Commission for Europe*, Geneva, Switzerland, February 18-20, 1998.
- Statistics Canada (2004), *Policy on Standards (Revised)*, internal document. Ottawa, Ontario, July 14, 2004.
- Statistics Canada (2005), *New Household Survey Strategy: Summary Report*, internal document, Ottawa, Ontario, Oct. 5, 2005.
- Statistics Canada (2009), *Questionnaire Development Tool, Business Requirements Version 1.1*, internal document, Social, Health and Labour Statistics Common Tools, Ottawa, Ontario, Sept. 14, 2009.
- Statistics Canada (2011), *Common Tools Project for Social Surveys: Bringing it all together*, internal document. Ottawa, Ontario, June 2011
- Statistical Commission – Economic and Social Council (1999), *Draft standards of the United Economic and Social Information System for data structure and metadata in international data exchange and dissemination*, Catalogue no. 98-35662 (E), United Nations, New York, March 1999.
- Statistical Office of the European Communities (EUROSTAT) and the United Nations Economic Commission for Europe (2006), *Conference of European Statisticians Recommendations for the 2010 Censuses of Population and Housing*, New York and Geneva.

Calibrating mode effects in the Dutch crime survey

Bart Buelens and Jan van den Brakel¹

Abstract

In the Dutch crime and victimization survey, a mix of data collection modes is employed in a sequential fashion: when no response is obtained in one mode, a different mode is used for reapproach. Year-to-year instabilities in the GREG-estimates of this survey were observed. Analysis suggested that these were at least partly due to changes in the response mode composition, indicating the presence of mode effects. At the same time, selection effects cause the subpopulations that are reached by the various modes to differ. Mode effects and selection effects are confounded, and are impossible to isolate. A practical approach to reducing the problem of temporal instabilities is presented. The model underlying the GREG-estimator is assumed to be fully correct for selectivity of the response. Based on this assumption, differences in the response mode composition between years are not a sign of unremoved selectivity. They must have other causes, and may be responsible for mode effects. These effects cannot be removed, but they can be leveled out by calibrating the response modes to fixed benchmarks. Such calibration is achieved by a straightforward extension of the GREG regression model. As sequential mixed-mode data collection has become the de-facto standard for social surveys at Statistics Netherlands, particular attention is paid to issues concerning the introduction of mode calibration as a standard in the generic statistical process of social surveys.

¹Bart Buelens and Jan van den Brakel, Statistics Netherlands, The Netherlands.

SESSION 9A

**STANDARDIZATION IN INTERNATIONAL COMPARATIVE STUDIES:
BENEFITS AND CHALLENGES**

Designing, standardizing and monitoring survey operations in international large-scale educational research

Ralph Carstens¹

Abstract

Large-scale educational research is complex and this is even more evident in international studies that involve multiple languages and cultures. The need for data to be comparable across countries calls for the design of strong, standardized survey operations, the provision of comprehensive manuals and guidelines, the organisation of effective training sessions, the implementation of quality control at key stages of the survey process, the definition of common data protocols, as well as validity and reliability analyses. However, national organizational structures, expertise, and idiosyncrasies in survey research necessitate a somewhat flexible approach to implementing international sampling and data collection plans while maintaining overall quality and comparability. A related consideration is whether to centralize specific tasks, say, translation or its verification at the international level or allocate the responsibilities at the national level. Drawing on strategies and experiences from IEA's International Computer and Information Literacy Study (ICILS) and the OECD's Programme for the International Assessment for Adult Competencies (PIAAC), the paper will discuss some of the goals and limitations of standardisation and centralisation.

Key Words: Standardization; Survey operations; Technical standards; Large-scale assessments; International.

1. Design and standardization of survey operations

1.1 Quality goals

For each new cross-national survey or assessment, a large number of often diverse goals, expectations, needs, inputs yet also budget and time constraints need to be considered. In combination, they affect a survey's conceptual frameworks, design, operations, and eventually quality. The overarching goal during the design phase of a survey, and additionally as revisions become necessary during its implementation, is to maximize a survey's overall quality or, as it is sometimes called, *total* quality. Some of the common quality dimensions (Brackstone, 1999; Biemer and Lyberg, 2003; Biemer, 2010) can be described as 'user' dimensions as they are predominantly set by stakeholders and sponsors. They relate to the relevance, richness, or completeness of the data collected; to the timeliness with which data, analysis, and reports become available; and the accessibility and usability of a public-use data product to foster secondary analysis. Furthermore, the way in which cross-national surveys are operationally implemented at the international coordinating level and within participating countries—jointly considered the data 'producers'—is a critical component in a survey's quality chain as it can directly influence the previously mentioned and other quality dimensions, most importantly the accuracy of the estimates derived from the collected data at the national level and the comparability of these estimates across countries, languages, demographic domains, and survey cycles. Jointly, the degree to which these quality dimensions were satisfied determines whether the data can be meaningfully used to produce estimates, comparisons, and inferences.

Another way of expressing quality dimensions is by means of a 3-tier model of *product*, *process* and *organizational quality* (for a discussion, see Lyberg and Stukel, 2010). The quality of a data product is set by the respective stakeholders (in large-scale assessments, these are the participating countries and the commissioning organization) and specified in terms of analytical potential, included variables, and the required precision of estimates. Naturally, the quality of a product is conditional on the quality of the processes that are generating it. In practical terms, the choice of methods, the definition of standard operating procedures, and their monitoring are critical aspects to ensure that the processes can yield an accurate and comparable data product. Finally, the organizations and teams that

¹Ralph Carstens, IEA Data Processing and Research Center, Mexikoring 37, 22297 Hamburg, Germany (ralph.carstens@iea-dpc.de).

implement a survey at the local as well as international level are critical to the success and overall quality of a survey.

1.2 General approach

At the heart of most large-scale studies in education is a cross-national sample survey of achievement in one or more content domains, enriched by the collection of contextual information at various levels (*e.g.*, system, school, teacher, classroom, parents) that are believed to be associated with the variation in achievement and, for example, the effectiveness of individual types of schools by relating inputs (antecedents) with outputs. Here, a study across countries can provide insights that are not attainable from the study of a single country and attempts to identify the malleable factors that could be manipulated to bring about improvements in attitudes, achievement, or efficiencies in the educational system.

Implementing such a study operationally in, for example, more than 60 countries² is a hugely complex and demanding enterprise in political, financial, and operational terms (see also Tamassia, 2005). Clearly, every attempt is made to standardize the survey operations across all participating countries in order to minimize or eliminate special process variability, which could and often does lead to errors that are difficult or impossible to fix. The general approach to standardization in many, if not all, educational surveys therefore is input or *ex ante* harmonization, that is, a strategy where all countries use the same definitions and uniform processes in contrast to output or *ex post facto* harmonization that relies on finding common denominators with respect to the data produced as well as on assessing the quality and comparability of it after data has been collected using diverse methodologies. For example, participating countries in IEA studies start their translation work from one, sometimes two, source instrument versions and all adaptations to questions and concepts are documented, reviewed, and approved prior to implementation, ensuring that structural adaptations can be reasonably mapped back to an international layout and concepts are translated appropriately. Given a particular survey design, often no universally valid or accepted “gold standard” exists for certain parts of the survey process. Rather, the coordinators of large-scale assessments aim to use or adapt the current best survey methods and practices (*cf.* Harkness *et al.*, 2010; De Leeuw, 2008; Statistics Canada, 2010; Survey Research Center, 2010) and, where applicable, its legacy, *i.e.*, the approaches and methodology used in previous cycles. However, the sophistication of contemporary survey methodology (*e.g.*, sampling or interviewing) can be striking and usually methods are selected that can provide the desired quality yet be implemented realistically and reliably by all national teams.

The key benefit of standardizing methods and operations upfront is the predictability of the process outputs under usually very tight timelines, which simply do not allow for variation with respect to, for example, the data files formats delivered by countries as all available time and budget must be spent on substantial work rather than reconciling and/or harmonizing one-off idiosyncrasies. While globally optimal approaches can usually be identified for areas such as sample selection or data capture and editing, the standardization does not have to be strict and prescriptive in other areas. It could even be detrimental to the quality goals. For example, the involvement of appropriate local bodies to endorse and promote the research or the specifics if the school/respondent contact strategies can be allowed to vary within certain limits if a better and defensible local optimum exists that does not contradict the global one. This often follows a ‘whatever works best’ philosophy and ‘soft’ or ‘guiding’ standardization in this context means that countries are required to develop and document concrete plans for such aspects.

Survey coordinators strive to take known error sources into account such as coverage, sampling, and non-response error, the cross-cultural validity of constructs, measurement non-invariance, data collection error, and processing error. In addition, they attempt to anticipate new error sources in case of redesigned or newly employed methods (more recently with respect to the advent of computer-based data collections). Within each area, the goal is to minimize natural (random) and eliminate systematic process variation (and hence bias or excessive variance; *cf.* Biemer, 2010) triggered by accidental—yet sometimes negligent and even willful—failure to comply with a specified procedure. Obviously, an almost infinite amount of time and resources could be spent on identifying and studying about any error source. In reality though, cost-error trade-offs have to be made, *i.e.*, the balance between identifying and reducing errors and the cost and time of doing so. In this line of thinking, planners give some priority to the most consequential error and bias sources (such as non-response or erroneous data collection) or notorious

² For example, refer to the IEA TIMSS 2011 country list at <http://timss.bc.edu/timss2011/countries.html>.

ones (such as natural human scoring variation) yet without ignoring other areas completely and basing priorities on the experiences from previous cycles. Subsequently, coordinators develop plans for quality control monitoring that are designed to collect sample information as to how uniformly processes were implemented (more on this later).

1.3 Allocation and share of responsibilities and workload

Any attempt to develop standards for a cross-national study must take into consideration that it operates on essentially two levels, national and international, and any set of operations must accommodate this reality and the cooperative nature. In theoretical terms, all tasks should be allocated at the international level to maximize communality and standardization, but this is not advisable from a quality perspective. Equally, it is also not realistic and advisable to decentralize them completely. Instead, certain factors and constraints usually determine how best to share responsibilities and workload.

One constraining factor is the complexity of the survey task versus the organizational capacity at the national level. Large variation exists with respect to the survey-taking traditions, the professional capacity, the experience, and the financial resources at the national level. While many national research organizations have developed an impressive ability to plan and implement international (and national) assessments over time—thinking of IEA TIMSS, which started in 1995, or the Organisation for Economic Co-operation and Development (OECD) PISA survey, which surfaced in 2000—it is still not realistic to expect all countries to implement survey operations with the high level of sophistication used by specialized survey organizations. One example of complex and, hence, predominantly centralized work is sampling and weighting. The design and selection of probability samples is routinely carried out at the international level, using national analytical needs as inputs (*e.g.*, the wish to report by domain or region), yet requiring countries to compile a recent and complete sampling frame according to specification defined at the international level or carry out non-response bias analysis, if necessary. In the case of sampling/weighting, it is usually both more effective (thinking of, say, documentation) and cost efficient to carry out the task at the international level than to verify the work done by countries. In school-level assessments, usually all samples are drawn centrally. In other survey settings though, countries may elect to carry out the sample selection themselves. In the OECD PIAAC assessment, this applies to about half of the countries and the work is mostly done by the respective national statistical offices. The key reasons for this include, but are not limited to, the confidentiality of the information on the frame and the intricacy of some multi-stage sample designs.

Another key factor for determining the share of responsibilities is the knowledge of the field and the access to the local educational (eco)system. With respect to the translation of survey instruments, local knowledge and expertise are required that cannot be provided reasonably at the international level, implying that a fully centralized translation process is not advisable. Hence, the translation work is shared such that national teams are responsible for ensuring that appropriate translations and adaptations are made (*e.g.*, with respect to local definitions of standardized educational attainment levels) as to maximize to overlap with the concepts and definitions stipulated by the frameworks and source instrument versions. The responsibility for the provision of translation guidelines, review and approval of adaptations, linguistic translation verification, and optical layout verification is then allocated at the international coordinating level. For the data collection itself, clearly, national teams are in a better position to judge how best to access and contact sampled units, secure endorsement by relevant local bodies, organize communication, and schedule the test administration within a set of guidelines and allowances applicable to all countries.

Last but not least, workload and costs are constraining factors and tasks need to be shared at peak times of the survey to limit expenditures at the international level. For example, translation work for a larger volume of materials (*e.g.*, tests, questionnaires, manuals) in a usually short timeframe has to be shared not only for the sake of quality as it puts a high demand on all involved. Here, the share of responsibilities between the national and international level also helps to keep the work manageable for either side. In contrast, the preparation and execution of the actual field work and post-collection tasks such as data capture, coding of occupational response, or the scoring of student's constructed responses are usually the sole responsibility of the national teams under strong guidance from the international level.

2. Quality assurance and quality control monitoring

As argued above, the design of a survey or assessment must be guided by quality goals. With regard to the project's operations, quality assurance (QA) means that the organizations implementing a survey have the capacity and experience to do so and that the defined and standardised processes can yield a data product that meets the needs and expectation of the stakeholders. In everyday language, QA means: "building the right thing." Since implementation can negate a good design, QA is merely input-oriented and could be treated as 'fiction.' The role of quality control (QC) is to check the 'reality,' generate verifiable documentation, and gauge compliance, *i.e.*, whether the processes actually work and products are actually good. In everyday language, QC means: "building the thing right" (or fixing it where needed).

2.1 Quality assurance approaches and artifacts

As described above, the central idea of quality assurance is that the quality requirements or goals for the product are fulfilled. In practical terms, QA are the planned and systematic activities implemented to achieve this. In educational surveys, this relates to the detailed and exhaustive documentation of the survey operation procedures that each country is required to carry out. Besides the conceptual aspects that provide the national teams with the framework that justifies a study, a number of resources and artifacts are commonly used to convey the necessary information about the study implementation. The central ones are:

- overviews describing the roles, responsibilities, and required skills of national research coordinators and other key staff;
- unambiguous and straightforward manuals, guidelines, and accompanying checklists for all operations on the critical path (*e.g.*, sampling manuals, translation and adaptation guidelines, test administration / interviewer manuals, data management manuals, scoring manuals);
- centrally provided software systems, tools, and services where complexity and the need for consistency dictates this;
- technical standards summarizing key product and process parameters (discussed later);
- regular face-to-face meetings to report progress, discuss observed obstacles, lessons learned, areas of improvement, and upcoming tasks (community of practice); and
- trainings for key tasks such as data collection in the field ('train the trainer' strategy), data capture and processing, or analysis.

Together, these approaches provide national coordinators with exhaustive, focused and sufficiently detailed information to implement the survey locally, assemble national teams with the required expertise and capacity, find in-house support or sub-contractors for specific areas (such as translation, data collection, or occupational coding), plan respondent contact strategies, implement the data collection, and prepare the required data files. In short, they receive all the information, materials, and help necessary to carry out the study without the need to develop their own materials and tools for specific aspects. For example, the IEA provides each participating country with a copy of its Within-school Sampling Software to facilitate the enumeration of classrooms, students, and teachers (depending on design), flag exclusions and enter sampling-relevant auxiliary information, draw samples according to the international design, assign hierarchical identification codes, produce tracking lists and questionnaire labels, and record the participation of respondents as reported by test administrators and given the receipt of instruments.

For some areas—contact strategies were mentioned earlier as an example—national coordinators are required to develop a plan for their local context by customizing and sometimes extending generic recommendations. It is then necessary to systematically collect information on how national coordinators are planning to implement guidelines, for example with respect to the number of interviewers (and supervisors) given a target number of completed cases and a fixed data collection period. As an example, the OECD PIAAC Consortium has used a so-called "National Survey Planning and Design Report" to systematically collect information on key implementation aspects well ahead of the field work and to gauge whether countries are likely to implement the study soundly or require additional support.

The organizations coordinating international large-scale assessments (this is at least true for the IEA), however, do not assess a country's capacity to implement a study prior to admitting them to it. Rather, they are initially accommodating regarding a country's intent to participate. The ultimate test of the overall functioning of the QA

approaches and the local compliance with standardized operations in most, if not all, cross-national studies is a field trial, that is, a dress rehearsal that serves multiple goals: i) testing, validating, and refining survey instruments and/or their translations, ii) testing and revising survey operations, QA approaches in general, and iii) providing national and international coordinators with information regarding each country's capacity to implement the work. Where necessary, international coordinators analyze problems in depth and provide countries with additional support.

2.2 Quality control approaches and programs

Quality control is the systematic generation and collection of information and evidence that the survey operations have been implemented in compliance with standards and plans and that the collected data is fit for the intended use. The quality control efforts must cover the entire survey process, regardless of whether international or national coordinators were responsible. Ideally, the QC efforts should not encounter any 'black boxes,' *i.e.*, unverifiable processes, and be conducted not only by individuals close to those who have implemented the process but additionally by independent monitors. Usually, QC programs are focused on the processes under national responsibility (*e.g.*, translation, field work, or data capture). Nevertheless, also the work conducted by the international coordinators is documented in great detail such that it can be reviewed and cross-checked by national coordinators or technical advisors.

QC work strives to quantify the amount of error and, where applicable and feasible, even use statistical process control methods. This is relatively easy with respect to, say, monitoring collection production, projected response rates, scoring reliability, or double data capture accuracy. In most cases though, it is impossible to cover each and every action and quality control programs must routinely be limited to samples of the work. As an example, IEA studies routinely implement an ambitious international quality control program to document data collection activities by nominating an international Quality Control Monitor (QCM) in each of the participating countries. This person, following comprehensive training, monitors the test administration in approx. 10% of schools (*i.e.*, 15 out of 150 schools under the canonical design) among other things. In TIMSS 2007, 248 QCMs and their assistants monitored a total of 1,371 testing sessions (Olson, Martin, Mullis, 2008). These programs generate important evidence regarding the achieved uniformity of the data collection and are usually further extended by similar, national programs based on the international model. Sometimes though, it may be impossible to collect information independently and directly. In the OECD PIAAC assessment, the deployment of independent international monitors to verify interviewers' work during and/or after the fact turned out to be unfeasible on account of contact data confidentiality. In this case, quality control is mostly done by the organizations also responsible for the interviewing work.

For all survey activities, national coordinators are expressly asked to systematically report their compliance and their (or the respondents') experiences with pre-defined operations in field work reports, survey activity questionnaires, checklists, and/or phone calls, seemingly different artifacts that, however, serve a common purpose. Again, the field trial and the quality control evidence collected during its conduct are used to check whether a country is likely to complete the main data collection successfully or whether substantial improvements, support or even the reconsideration of participation should be advised. Following the main data collection, all available quality control evidence (*e.g.*, achieved response rates, scoring reliability, and reports by quality control monitors) are compiled and used to adjudicate the national samples and as a result, make recommendation for unconditional, conditional, or non-inclusion of data in the international analysis, scaling and reporting.

2.3 Technical standards

All large scale cross-national studies in education define technical standards, that is, a set of requirements to be satisfied by a product or process. For example, technical standards define key aspects of national responsibilities, management, sample design, translation process, data collection procedures, data file formats and contents, coding/scoring, or minimum response rates. In this way, technical standards can be seen as a synopsis of key quality assurance/control aspects and embody the shared and agreed quality expectations for a survey.

A great deal of variation exists though regarding how technical standards are documented and circulated. In IEA studies, where the national stakeholders and the national research coordinators are identical, operational manuals such as a sampling or data management manual are sufficient to communicate standards, quality expectations, mandatory processes, and checkpoints along with the detailed instructions. This approach is beneficial as it avoids redundant documentation. In OECD surveys and assessments, technical standards are typically developed as a

dedicated document that holds standards across areas, further detailed in operational manuals. The key rationale for this probably relates to the way that OECD projects are managed, *i.e.*, by a i) board of participating countries as the main decision making body, which isn't concerned with operational documentation but with high-level quality specifications from a user perspective, and ii) a group of national project managers that implement the work locally and hence require information on standards as well as detailed instructions. Another benefit of a technical standards document, along with timeline and milestone documents, is that they can be used as important input in the national tendering for survey organizations. Still, the level of detail in technical standards can vary strikingly. The OECD PIAAC Consortium has produced a set of technical standards (not yet published) that comprise almost 200 pages, largely owed to the study's methodological complexity and rooting in the standards model of the previous IALS and ALL work. While this admittedly reduces the likelihood of the document actually being read, the 20+ pages of the PISA 2009 assessment (OECD, 2011) appear relatively sparse in contrast. The main difference between the documents is that PIAAC includes a large amount of supporting information in rationale, guideline, recommendation, and quality assurance/control sections for each area besides the actual standards. The OECD's Teaching and Learning International Survey (TALIS) adopted some middle grounds and the standards document there (not yet published either) spans about 40 pages yet leaves, just as PISA does, more of the technical details to operational manuals.

Clearly though, the key challenges in producing technical standards are not related to the level of detail. Rather, it is important to develop a comprehensive set and indicate where standards will be strictly enforced (*e.g.*, with respect to response rate requirements or permissible data editing) or where some latitude could be exercised and countries may apply for derogations if they prefer a different or modified approach to some survey work subject to the approval of international coordinators (*e.g.*, the use of scanning and optical character recognition as opposed to manual key data entry). Another important consideration relates to pre-existing local standards as they are used in an increasing number of survey organizations and many, if not all, national statistical offices. Here, much time is usually spent on the review of proposed derogations to match local practices and, in some cases, their non-approval.

3. Examples: IEA/ICILS and OECD/PIAAC

Table 3.1 below provides key design aspects of two current large-scale international assessments, the IEA's International Computer and Information Literacy Study and the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) for the sake of illustrating the previously described rationale for the standardization of operations given a particular design. This is not so much a valid comparison as they serve very different purposes yet they also have a high degree of similarity as both strive to produce authoritative international data for comparison and benchmarking. Additionally, both studies are interesting as they standardly involve computer-based administration (CBA) of tests and contextual instruments in natural connection with the domains and constructs under investigation. While ICILS extends the proven model of IEA studies to CBA, PIAAC is likely the most ambitious, complex, and costly cross-national survey in education and adult competencies to date given its combination of household survey, educational assessments, and CBA methodology.

As can be seen from the table, the studies differ significantly in the organizations responsible for the international coordination and national implementation, in how and by whom samples are selected, in data collection procedures (classroom test administration *vs.* face to face interviews), in the corresponding quality control programs and whether QC evidence is available directly (international observers/monitors) or indirectly (phone verification by supervisors), and naturally the key challenges that each survey faces. However, both projects also share certain approaches and best practices although the specific approaches, tools and systems vary. This is true for the mixed instrument modes, the fact that both studies provide countries with the software systems they need to implement the data collection, the translation and adaptation processes, and the post-collection processes such as data capture from paper and scoring/coding of open-ended responses, which follow very similar rules and goals.

Table 3-1 Key design aspects of the IEA ICILS and OECD PIAAC studies

	IEA International Computer and Information Literacy Study (ICILS)	OECD Programme for the International Assessment of Adult Competencies (PIAAC)
Domains	Information and computer literacy	Literacy, numeracy and problem-solving skills
Targets	Grade 8 students and their teachers	16-65 year-old adults
Coordination	Jointly managed by 3 partners: ACER (lead), IEA HQ and IEA DPC	Led by ETS; 7 partners: Westat, IEA, cApStAn, ROA, DIPF, CRP, and GESIS
Participants	~20 countries	25 countries in round 1, ~8 in round 2
Schedule	2010-2014, field trial in 2012, main collection in 2013	2008 to 2013, field trial in 2010, main collection from late 2011 until early 2012
National centers	Mostly universities, ministries and/or affiliated educational research institutes	Mostly national statistical offices supported by commercial survey organizations, some educational research institutes
Samples	Two-stage PPS of 150 schools with 20 students and 15 teachers each; all centrally selected	Large variation from registry samples to multi-stage samples + household screeners; about half centrally selected/weighted; other half under country responsibility
Instruments	Student computer-based (CBA) test + electronic questionnaire, school and teacher questionnaires (choice of paper or online)	Adaptive CBA test, paper fallback, CAPI background questionnaire
Translation	Adaptation/translation, expert review, revision, layout checks (national level); Verification, layout checks, documentation (international level)	Adaptation/translation (double recommended), reconciliation, layout checks (national level) Verification, layout checks, documentation (international level)
Systems	Translation, within-school sampling, student assessment (USB), on-line data collection, and data capture systems (all centrally provided)	Translation, CAPI+CBA (laptop-based) and data capture systems (centrally provided) Study/case management system (local responsibility)
Data collection	Proctored and timed student sessions, all questionnaires self-administered (online, default, paper fallback)	CAPI background questionnaire, non-timed cognitive CBA or paper-based (PBA) portion self-administered (controlled by interviewer)
Post-collection tasks	Coding of parental occupation, scoring of constructed responses, data capture from paper (local responsibility with international reliability studies)	Coding (occupation, industry, language, country, region), scoring of PBA responses, paper data capture (local responsibility with international reliability studies)
Quality control program	National quality monitors; international monitors observing test administration in 10% of schools (direct evidence)	Quality control phone calls, 10% verification of each interviewer's work (100% if suspect) by national centers, reported to Consortium (indirect evidence)
Key challenges	School participation, school ICT infrastructure, uniform test conditions given CBA systems and test administrator skills	Uniform interview and test conditions, complex CAPI+CBA system (also: roll-out and patching), initial contact, response rates, overall schedule

4. Standardizing standards

It was argued earlier that it is essential to standardize definitions and operations across the countries participating in a single survey. The same can be argued for the multi-cycle studies (*e.g.*, trend indicator programmes such as TIMSS/PIRLS and PISA) and studies implemented by the same organization (*e.g.*, IEA or OECD). Contemporary large-scale assessment share a relatively large number of features, approaches and procedures at the operational level while the populations, constructs, and designs can vary greatly. Clearly, this is a welcome development but not everything works everywhere.

At the national level, ISO 20252 accreditations (International Organization for Standardization, 2006) are emerging and relate to the standardization of survey planning and conduct within a survey organization. Standardization of survey practices can also be found in national statistical offices (*e.g.*, Statistics Canada, 2009) or other governmental agencies (*e.g.*, National Center for Education Statistics, 1991) although the nature of such documents can range from recommendations/guidelines to fairly firm instructions that leave relatively little or no room for derogations. As Susan Linacre said during her keynote, “if something is worth standardizing, it is also worth enforcing it.” This was applied to the Australian context and the challenge to standardize, say, editing and imputation across ABS’ various

business areas. Besides the likely gains in quality and consistency, provided that the global optimum is also the local optimum for each and every survey, standardization across surveys can also save resources in the medium or long term as generalized systems are created, resources are pooled, and new projects (hopefully) require very little customization. Often though, standardization first generates additional costs.

For cross-national surveys, the balance between the global and local optimum is the topic of a highly useful resource developed and published by the Survey Research Center (2010) at the University of Michigan. The Cross-cultural Survey Guidelines (CCSG) apply to social research in general yet primarily relate to cross-sectional surveys of households and individuals. They identify and promote the use of best practices, *e.g.*, a team translation approach, and include numerous literature references, examples, and a fine glossary. The guidelines are phrased in general terms though and they are neither prescriptive nor do they impose specifics, implying that each survey with its own goals and design requires specific approaches, procedures and operations to be defined. In this sense, the CCSG are not firm technical standards themselves but more a framework of ‘meta-standards’ for developing them by international coordinators.

The standardization and harmonization of similar cross-national studies and for the same study over time remains a challenge though. For the current OECD programmes (PISA, TALIS, PIAAC, AHELO), projects are managed by different departments at the OECD Secretariat and usually implemented by different international contractor(s), typically a consortium of internationally operating survey organizations. OECD projects, at least at the school-level, strive for some conceptual alignment, the re-use of questions, or the harmonization with joint UNESCO-OECD-EU data collection approaches. Nevertheless, technical standards and operations are defined independently for each project and cycle yet a relatively high degree of communality exists through the use of the same contractor for consecutive cycles and the more general use of current best survey practices by the contractors. Also, countries sometimes ask for operational alignment between multiple OECD projects they participate in (Can we do this as in PISA?).

In the IEA’s cross-national assessments, many parts of the survey process are fairly standardized and follow a common model (*e.g.* with respect to sampling, translation, data capture and processing) while others are not or cannot be given the specific design of the study. All IEA projects follow the guidelines laid down in the organization’s technical standards (Martin, Rust and Adams, 1999; also Gregory and Martin, 2001), which are tailored to the unique setting in which IEA is operating, that is, they have a clear focus on educational assessments. Similar to the CCSG, no specifics such as a fixed response rate minimum are imposed on all studies. Rather, these meta standards serve as a framework that requires international coordinators to work though each part of the survey process – from setting up an international study center to releasing data and technical documentation – to define study-specific standards. Potentially, this could lead to very different designs as also the IEA’s studies (currently: TIMSS, PIRLS, TEDS, ICCS, ICILS) are led by different organizations and individuals. However, a large degree of communality exists as projects learn from each other and adopt tried and proven approaches and practices from each other. Further, many survey tasks are not only standardized but also centralized across studies and this applies, to various degrees, to sampling, translation/verification, operations, quality control of data collection, data management, (secondary) analysis and dissemination. Additionally, a single Technical Executive Group (TEG) oversees all studies’ plans, progress, and quality. Together, these structures result in high between-study communality for many, yet not all, standards, processes, and their quality control.

References

- Biemer, P.P. (2010), “Total survey error: Design, implementation, and evaluation”, *Public Opinion Quarterly*, 74(5), pp. 817–848.
- Biemer, P.P. and L.E. Lyberg (2003), *Introduction to Survey Quality*, Hoboken, NJ: John Wiley & Sons.
- Brackstone, G. (1999), “Managing data quality in a statistical agency”, *Survey Methodology*, 25(2), pp. 139-149.
- De Leeuw, E.D., Hox, J.J. and D.A. Dillman (eds.) (2008), *International Handbook of Survey Methodology*, New York: Lawrence Erlbaum Associates, Taylor & Francis Group.

- Gregory, K.D. and M.O. Martin (2001), *Technical Standards for IEA Studies: An Annotated Bibliography*, Amsterdam: IEA.
- Harkness, J.A., Braun, M., Edwards, B. Johnson, T.P., Lyberg, L.E., Mohler, P.P. Pennell, B.-E. and T.W. Smith (eds.) (2010), *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Hoboken, NJ: John Wiley & Sons.
- International Organization for Standardization (2006), ISO 20252, *Market, opinion and social research - Vocabulary and service requirements*, Geneva, Switzerland: ISO.
- Lyberg, L.E. and D.M. Stukel (2010), "Quality assurance and quality control in cross-national comparative Studies", in Harkness, J.A. et al. (eds.) *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, Hoboken, NJ: John Wiley & Sons, pp. 227-249.
- Martin, M.O., Rust, K. and R.J. Adams (1999), *Technical Standards for IEA Studies*, Amsterdam: IEA.
- National Center for Education Statistics (1991), *SEDCAR Standards for Education Data Collection and Reporting*, Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.
- OECD (2011), *PISA 2009 Technical Standards*, Appendix 7 of PISA 2009 Technical Report (preliminary version), Paris: OECD. Retrieved January 15, 2012 from http://www.oecd.org/document/19/0,3746,en_2649_35845621_48577747_1_1_1_1,00.html
- Olson, J.F., Martin, M.O. and I.V.S. Mullis (2008), *TIMSS 2007 Technical Report*, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Statistics Canada (2009), *Quality Guidelines*, 5th edition, Catalogue Number 12-539-XWE, Ottawa: Statistics Canada.
- Statistics Canada (2010), *Survey Methods and Practices*, Catalogue Number 12-578-X, Ottawa: Statistics Canada.
- Survey Research Center (2010), *Guidelines for Best Practice in Cross-Cultural Surveys*, Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved January 6, 2012, from <http://www.ccsr.isr.umich.edu/>
- Tamassia, C. (2005). "Implementing Surveys In An International Context: An Overview", paper presented at the OECD Programme on Educational Building (PEB) Expert Group Meeting on Evaluating Quality in Educational Facilities, Lisbon, Portugal.

Summarizing item responses in large scale assessment

Eugenio Gonzalez and Matthias Von Davier¹

Abstract

Large scale assessments and other survey programs generally administer multiple questions to the surveyed participants. These questions can be diverse in nature, or can attempt to measure a common construct from complementary or different angles. When multiple questions are administered to measure a common construct, it is generally useful to summarize the responses to these multiple questions into a single variable or index. The process of combining the responses to these variables into a single number is what is known as scaling. There are multiple ways to make this combination of items, from taking the simple sum of points on the items, to using more complex methods such as item response theory (IRT) and making multiple draws from an expected posterior distribution of possible outcomes. This paper will present advantages and disadvantages of different methods of scaling in the context of large scale educational assessments. Particular methods presented and discussed include sum scores, average scores, percent scores, factor scores, IRT scores, and plausible values.

¹Eugenio Gonzalez and Matthias Von Davier, Educational Testing Service, USA.

Standardization of sampling plans and quality assurance in comparative surveys

Marc Joncas and Sylvie LaRoche¹

Abstract

The growing interest in international comparative education studies presents many challenges in standardizing statistical methods, especially sampling plans. In this paper, we examine the various steps in defining a standardized sampling plan for comparative studies of this type: defining the target and survey populations; constructing the sampling frames; choosing the sample selection method; determining the sample sizes and estimate precision; and evaluating the implementation. We conclude with a discussion of the lessons learned in the many years of participating in various international surveys.

Key Words: Standardization; Sampling plan; International comparative studies; Education.

1. Introduction

International comparative surveys in the field of education have been conducted for many years. They look to measure the effectiveness of education systems as a whole and are generally administered to students and/or teachers. A number of these surveys assess the skills acquired by students and thus provide data for comparing the performance of participating countries' education systems. Among the best known such surveys are the Programme for International Student Assessment (PISA)², the Progress in International Reading Literacy Study (PIRLS)³ and the Trends in International Mathematics and Science Study (TIMSS)⁴. In addition, since 2008, the Organisation for Economic Co-operation and Development (OECD) has administered the Teaching and Learning International Survey (TALIS), which focuses on teachers' working conditions and pedagogical environment.

Since most of these studies rank participating countries' education systems on the basis of student performance, their findings receive heavy media coverage and are often politically sensitive. Because of their comparative nature, international studies tend to be controversial. Consequently, their managers face many challenges in ensuring that every aspect of the surveys is checked and tested so that the results are comparable and credible.

Sampling is certainly an important component. The sampling methodology used in previous surveys has often been criticized. We can always ask ourselves the following questions: Is a country's sample representative of the target population? Are the exclusion levels comparable between countries? Were the samples properly selected? Are the participation rates acceptable? These questions and many others led to the need for a series of controls and standards specifically for comparative education surveys; these controls and standards are reviewed in the sections below. Note that our review is limited to elements that relate to the sampling plan in particular. More specifically, we first describe the conditions for an ideal context leading to an appropriate sampling plan and then move on to the difficulties of implementing such a plan in the field. That section is followed by a brief discussion of the reasons for standardization in an international context. Section 4 presents a number of examples of the most widely used controls and standards in this type of survey, and section 5, the conclusion, contains a list of the lessons learned from many years of experience working on international comparative education surveys.

¹Marc Joncas, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (marc.joncas@statcan.gc.ca); Sylvie LaRoche, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6 (sylvie.laroche@statcan.gc.ca).

²Conducted by the Organisation for Economic Co-operation and Development (OECD).

³Carried out by the TIMSS and PIRLS International Study Center and funded by the International Association for the Evaluation of Educational Achievement (IEA).

⁴Also carried out by the International Study Center and funded by the IEA.

2. Contexts

2.1 Ideal context

In an ideal context, a sampling plan should lead to unbiased, accurate and internationally comparable results. Consequently, three conditions appear to be essential in designing international surveys.

First, the survey population has to be the same as the target population. This condition is crucial because the inferences relate to the survey population. Following collection, it is not always possible to make adjustments in the estimation process to remedy coverage flaws. It is often easier to convince people of the importance of fully covering the population of interest when you are conducting a census. What good would a census be with just 80% coverage?

In the case of a sample survey, however, people tend to think that this condition is not as critical, though it actually is. It is even more important in an international context, since it is the first thing that can undermine the survey's credibility. Coverage rates that differ between countries inevitably invite discussions that cast doubt on the validity of the comparisons and can limit the potential for analysis. When it comes to education surveys, this element is particularly important because it can leave the impression that the results will be biased if the portion of a country's population that is not covered includes the worst students.

The second essential condition is a valid sampling plan. It is our connection with the survey population. Every unit in the survey population must have a chance of being selected (a non-zero inclusion probability), or the exclusion level will increase (and coverage will be incomplete). The probability of inclusion must be known and calculable to eliminate a risk of bias. It is also important to ensure that the sampling plan is properly implemented (because there is no point in developing good sampling plans if they are not executed correctly in the field). Without a valid sampling plan, it is risky to use the survey results to make inferences about the survey population.

The third and final condition is that sampling errors should be as small as possible. This will ensure that the survey data are as close as possible to the values that would have been obtained with a census.

In an ideal context, where all these conditions are met, with perfect implementation of the sampling plan and the collection procedures, and with 100% response rates, there would be no need for standards such as those set out in section 4.1.

2.2 International context (in the field)

In survey projects involving a number of countries, it is rare to find an ideal context in which the conditions described in the previous section are fully satisfied. Just about everything differs from one country to another, and education is no exception. No matter what attribute one examines—the education systems themselves, geographic or cultural characteristics, the availability of and access to pertinent administrative data, the capacity to conduct surveys, the response burden or the survey culture,⁵ to name only a few—it is impossible to find two perfectly comparable countries. Invariably, coverage rates, response rates, the quality of the implementation—in fact, everything that relates to the sampling plan—is affected to differing degrees when a survey goes from one country to another. This observation does not mean that all attempts at international comparison are doomed to failure, but rather that such projects require the establishment of minimum standards and norms below which comparisons become suspect.

3. Why standardize

The use of standards and controls in this type of study validates comparisons of results and increases their credibility for users. The aim is to be able to attribute a significant difference (statistically speaking) in the results to a real difference in the populations being compared, and not to a combination of uncontrolled errors (inadequate response

⁵Survey culture refers to a country's openness to surveys. In some countries, participation in surveys may be mandatory, while in others, surveys are voluntary and are sometimes perceived as an invasion of privacy.

rate, poor coverage, inferior implementation of field procedures, measurement errors, *etc.*). Users' motivation to analyze the results is heavily influenced by the quality of the data and the relevance of results comparisons.

Establishing standards also makes it much easier to implement the sampling plan. In the particular domain of international education surveys, standardization often means unification of procedures. For example, in prestigious surveys such as PISA or TIMSS, the sampling plan, the sample size, the collection method or even the wording of questionnaires is nearly the same for all participants (examples of standards will be provided in the next section). Such unification ensures more balanced workloads among the participating countries and makes it possible to introduce effective, uniform minimum measures for data quality control.

In summary, the establishment of standards and controls pays off for all parties involved: the sponsors are reassured by valid, comparable results that their investment was worthwhile; the survey managers can guarantee the quality of the procedures and the validity of the results; and the participating countries obtain results of assured quality for what is essentially an equivalent amount of work regardless of their individual conditions, constraints and environment.

We now present a brief overview of the established standards and typical controls applied to international comparative education surveys to ensure the comparability of the results.

4. Examples of standards and controls

4.1 Introduction

We distinguish between standards and controls as described below. Standards are norms established to ensure data quality. Those norms are described and documented. For countries that fail to comply with standards, the consequences usually range from notes in published tables to relegation of all their results to an appendix.⁶ Controls are less formal and often take the form of more flexible criteria. They are more commonly referred to as quality control procedures. Such control procedures were defined and applied on the basis of years of experience. However, some procedures that were controls early in the history of international comparative studies are now established standards.

4.2 Target and survey populations

The first examples of standards relate to the target population and the survey population. Each participating country is required to provide a description of its national population. At the very least, it must describe its education system (including the age at which children start school, the school structure, and the ISCED level⁷). The survey population must cover at least 95% of the target population (this standard is probably the most widely recognized by the education community). Any divergence between survey population and target population must be documented (type and magnitude of exclusions). If a country has less than 95% coverage of the target population, this will automatically be noted in international publications. A 95% standard may seem high, but it is important to keep in mind that in international publications, there is one row per country in the results tables. The user naturally assumes that the results are representative of the country as a whole, and comparisons between countries are made on that basis. As mentioned previously with regard to education surveys, there is often a strong presumption of correlation between coverage and measured performance. It would be difficult to lower those requirements and still claim that the comparisons are credible.

Certain controls are commonly used to ensure that the definition of the national target population matches the definition of the international target population. Additional checks of the information provided by the country (such as age, years of schooling, and school attendance⁸) are performed against external data sources. For repeated surveys,

⁶The results of non-compliant countries are omitted from the main tables and placed in an appendix at the end of the reports.

⁷The International Standard Classification of Education (ISCED) is a UNESCO standard for classifying education systems.

⁸School attendance is defined as the percentage of the age cohort in a given school year that is attending school.

we check that the definition of populations is comparable between cycles so that estimates reflect valid trends. If the target population changes from one cycle to the next, the portion of the population that is common to both cycles must be identified so that a trend analysis can be carried out. The latter situation is not so unlikely. For example, it may arise following a change in a country's education system that pushes the school year representing four years of formal education (often used as a basis for defining a population) from the fourth year to the fifth year at the start of the new cycle (which generally extends over three to five years). A reform may also change the age at which children start school. To our knowledge, there are currently no norms that require a specific percentage in those situations (such as a minimum percentage of the population common to the two cycles).

4.3 Sampling plan

This subsection covers standards applied to the sampling frame, the sample selection method, the sample size and the implementation of the sampling plan.

4.3.1 Sampling frame

In education surveys, the sampling frame used is often composed of a list of schools, and the size measure is the number of units in the target population (teachers or students). To our knowledge, there are no established, published standards for checking the frame's quality. However, good practices (controls) are followed that are similar to the checks usually performed on all sampling frames. First, we check that the frame is as up to date as possible. Second, we make sure that the frame supplied by participants provides complete coverage of the survey population and that it contains no erroneous data, duplicates or elements extraneous to the survey/target population. In addition, wherever possible, an up-to-date size measure for each unit in the frame is required. We also insist that the sampling frames supplied by the countries provide access to the entire target population. This makes it possible to estimate and document exclusions more effectively. Moreover, we use tools such as the Web, information from previous cycles, and information from other countries to validate the information supplied by country representatives. For example, a number of countries have international schools. A country may have omitted these schools because they are not considered part of the education system.

4.3.2 Selection method

With regard to the standards associated with sample selection, the norm is to require a single selection method for all participating countries. Adaptations and/or deviations are permitted, but they must be approved by the survey managers before implementation, and they must be documented. Having a single selection method allows us to develop and use generalized sample selection and weighting programs, thereby minimizing the risks of error. This approach facilitates the equitable distribution of work among the participants, resulting in much more uniform sample sizes across the various countries. It also facilitates the validation of selections and minimizes the number of control programs required for implementation. In addition, with the adoption of a single method, collection operations can follow uniform procedures, which limits the number of operations manuals required. This reduces the risks of error, preventing differences in instructions from affecting data quality and comparability. The use of a single selection method also helps reassure participants of the comparability of the results: non-sampling errors are expected to be comparable. Sampling errors too are expected to be of similar magnitude, which is not necessarily the case; nevertheless, the perception remains. Note that there are consequences to not meeting standards. The risk that the data will not be published or will be annotated in the tables increases substantially if the plan is not approved or irregularities are observed.

4.3.3 Sample size

Invariably in this type of survey, there is a standard for the minimum size of the sample. Most of the time, the minimum size is based on the desired margins of error and the study's inherent constraints. Another commonly used standard is the advance identification of so-called replacement schools. In general, there is a maximum of two replacement schools for each school originally selected. Again, any deviation must be documented and approved. Note that replacement schools cannot be used to replace eligible schools that refuse to participate. The use of replacements satisfies requirements concerning the sample size and may help minimize the risk of bias. Nevertheless,

we maintain strict requirements regarding the minimum participation rate of originally selected schools (see the next section).

4.3.4 Implementation

The most widely recognized standard in this area is the requirement that all countries take part in a trial. The purpose of the trial is to test procedures in the field and, in particular, take corrective measures (which are often needed) before the survey begins. In principle, participation in the trial is compulsory; non-compliant countries will have their data omitted from the international publications.

Another recognized standard is the minimum response rate. In general, we can define three zones:

- (1) There is the absolute minimum zone, or red zone. If a country fails to achieve these minimum rates, its data will simply be excluded from all international publications.
- (2) At the other end of the spectrum is the green zone. A country is in the green zone if its participation rates are above a certain threshold, without the use of replacement schools. In this case, the risk of bias in the statistics derived from that country's data is considered negligible. If the threshold is attained only after the replacement schools are brought in, the country's results are included in the international publications, but they are annotated to alert users to the increased risk of bias.
- (3) Then there is the grey zone. If a country's participation rates are above the threshold for the red zone but below the threshold for the green zone, even after the replacement schools are used, a decision is usually made on a case-by-case basis. The results may be placed at the bottom of the tables or in appendices, or they may not be published.

It is essential that response rates (sometimes referred to as participation rates) be documented so that analysts can assess the quality of the inferences and analyses based on the data.

With regard to controls, it is worth noting that during implementation, countries are often instructed to contact the survey managers when they encounter an unusual situation. This makes it possible to check and take action before data collection is complete and no further corrective measures can be taken. The presence of a measure of school size in the sampling frame provides another control: comparison of that size with the size observed in the field. It is possible to request explanations and more detailed documentation in cases where the differences are substantial (omission of classes, an error in identifying the school, a change in the school's structure, *etc.*). In addition, it is not uncommon to validate the status of non-participating schools following collection to determine whether a more appropriate status should be considered (for example, classify some refusals as exclusions). Standard errors are calculated in part to detect outliers, influential values, and influential or abnormal weights on the basis of the key variable of interest. It is also possible to compare observed and expected estimates (for example, exclusion rates in schools compared with rates in previous cycles, population totals compared with known totals from previous cycles). All these controls help detect potential violations of the rules set out in the sampling plan. Participants are usually required to provide written explanations for any abnormalities detected.

Lastly, we would like to point out the importance of conducting an evaluation of the implementation. Sampling plans and their implementation are usually reviewed in the presence of an expert from outside the survey's management circle. This independent evaluation and approval lend important credibility to the survey. In addition, it is essential to wrap up the project by preparing a technical report describing all the procedures affecting the sampling plan and its implementation.

5. Conclusion

From our experience with international comparative surveys, we have learned the following:

- (1) It is difficult to have standards that meet every need and retain some flexibility. In every survey or cycle, we have to deal with unusual situations. It is therefore important to have a technical team responsible for supporting the participants and to invite them to consult the team before and during the survey's implementation to address the various unforeseen problems.
- (2) The establishment of standards is necessary and critical to dispel any doubts about the relevance of the analyses based on the survey.
- (3) It is important to check, at a reasonable cost, all the procedures followed to improve the quality of the data collected.
- (4) It is also important to quantify and document the actions taken to ensure the quality and comparability of the data and build confidence in those responsible for implementing the survey.

It is safe to say, of course, that there is always room for improvement. Some of the above-mentioned controls could easily be beefed up and turned into standards. As survey managers, however, we have to exercise caution and maintain a degree of flexibility, always with a view to guaranteeing an acceptable level of quality. Standards are a constraint for the participating countries. Should we therefore aim for more standards to the detriment of flexibility and accommodation in the field, or should we take the opposite course, with the risks of possible abuse? The debate is still open.

References

- TIMSS & PIRLS International Study Center (2007), *TIMSS 2007 Technical Report*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.
- TIMSS & PIRLS International Study Center (2005), *TIMSS 2007 school sampling manual*, Chestnut Hill, MA; TIMSS & PIRLS International Study Center, Boston College.

SESSION 9B
STANDARDIZED SOFTWARE

Harmonisation of seasonal adjustment practices through the development of the DEMETRA+ software

Jean Palate and Pascal Jacques¹

Abstract

Seasonal adjustment (SA) is an important step of the official statistics business architecture and harmonisation of practices has proved to be key element of quality of the output. In this spirit, since the 90s, Eurostat has been playing a role in the promotion, development and maintenance of a software solution (Demetra) freely available for seasonal adjustment in line with established best practices.

In 2008, the ESS (European Statistical System) guidelines on SA have been endorsed by the CMFB and the SPC (Statistical Programme Committee) as a framework for seasonal adjustment of PEEIs (Principal European Economic Indicators) and other ESS and ESCB economic indicators.

The ESS guidelines cover all the key steps of the seasonal and calendar adjustment process and represent an important step towards the harmonisation of seasonal and calendar adjustment practices within the ESS and in Eurostat. A common policy for the seasonal and calendar adjustment of all infra-annual statistics will improve the quality and comparability of the national data as well as enhance the overall quality of European to the extent that proper SA tools exist and are available.

The SA Steering Group (the Eurostat-ECB high level group of experts from NSIs and NCBs which has produced the ESS Guidelines for seasonal adjustment) is promoting the development of a flexible software solution for SA to be used within the ESS.

The group has drawn its attention on the object oriented technologies used by the R&D Unit of the Department of Statistics of the National Bank of Belgium to develop a series of prototype tools for SA. This has been considered as an adequate framework for the cooperative development of a new generation of sustainable SA tools, enabling the implementation of the ESS guidelines and replacing the previous Demetra software.

The new software for SA (Demetra+) has now been released in its .NET, C# version as the official tool to sustain the implementation of the guidelines. The work is now continuing to redevelop the Demetra+ tool in JAVA including the core engines Tramo/Seats and X12/Arma and to release it as Open Source on the OSOR platform. This will lead to the deployment of a flexible, multi-platform ESS tool for seasonal adjustment enabling the implementation of ESS guidelines on seasonal adjustment, meeting the requirements from the users for the entire benefit of the SA community and for any future national account production system.

¹Jean Palate, National Bank of Belgium and Pascal Jacques, Eurostat, Luxembourg.

How does CANCEIS work & can it benefit more users?

Chunxiao (William) Liu, Sean Crowe and Asma Alavi¹

Abstract

The CANadian Census Edit and Imputation System (CANCEIS) provides users a flexible, efficient and data-driven edit and imputation methodology which:

- allows users to specify large numbers of edits through Decision Logic Tables,
- can perform both donor and deterministic imputation, as well as derive new variables,
- can work with different types of variables simultaneously,
- can work with data files at the unit or subunit level,
- allows users to customize its functions through parameters,
- is portable for use on most computational environments, and
- allows new features to be continually added to fulfill the growing needs of users.

CANCEIS is very efficient and effective in finding minimum change imputation actions (IAs) in a highly data-driven fashion by first identifying the most similar donors (nearest-neighbours) to the unit needing imputation and then determining best IAs from them. Users have very fine control over how nearest-neighbours are defined through matching variables, and IAs are determined from the nearest-neighbours.

CANCEIS is based on the Nearest-neighbour Imputation Methodology (NIM), which was first developed by Michael Bankier of Statistics Canada in 1992. CANCEIS has been used for the Canadian Census since 1996. It has also been used by some non-Census surveys in Statistics Canada and by several statistical agencies from other countries.

This paper talks about the CANCEIS methodology, its application, and its potential for more general use.

Key Words: Nearest-neighbour imputation; Minimum change; Data-driven; Decision Logic Table; Imputation Actions.

1. Introduction

The CANadian Census Edit and Imputation System (CANCEIS) was originally developed for dealing with Edit and Imputation (E&I) tasks on the 1996 Canadian Census data. It was used for five subject matter areas in the 2001 Canadian Census and for nearly 100% of the Census variables in 2006. It will again be used for 100% of 2011 Census and 2011 National Household Survey (NHS) variables. CANCEIS was designed primarily to work with text files and it has been rewritten in C# (C sharp) computer language within the .NET environment for the 2011 Census and NHS. The new developments include the capability of reading in EXCEL format files and generating HTML format output files.

Traditionally, the Canadian Census consisted of two types of questionnaire: one (called short form) collected answers to a few demographic questions and another (called long form) asked over 50 questions, including the short form questions, on a variety of diverse topics. The short forms were sent to 80% of Canadian households while the long forms were sent to a one in five sample of Canadian households. Both types of questionnaire were mandatory, except in 2011 when the long form Census became a voluntary separate survey of one in three Canadian households, and was named National Household Survey.

Currently, the E&I processing of Canadian Census and NHS data is partitioned into several processes according to characteristics and internal relationships of the questions. In CANCEIS, an E&I process is a sequence of two types of modules: derive and donor. Each of these derive and donor modules accomplishes a particular task in order to complete the E&I strategy developed for the process.

¹Chunxiao (William) Liu, Sean Crowe and Asma Alavi, Statistics Canada, R.H. Coats Bldg., 15th floor, Tunney's Pasture, 100 Tunney's Pasture Drive, Ottawa, Ontario, Canada, K1A 0T6, asma.alavi@statcan.ca.

In this paper, we explain how CANCEIS works in implementing the minimum change and data-driven approach to impute all types of variables simultaneously with easy specification of large numbers of edits. Based on this understanding, we will discuss its potential to benefit more users.

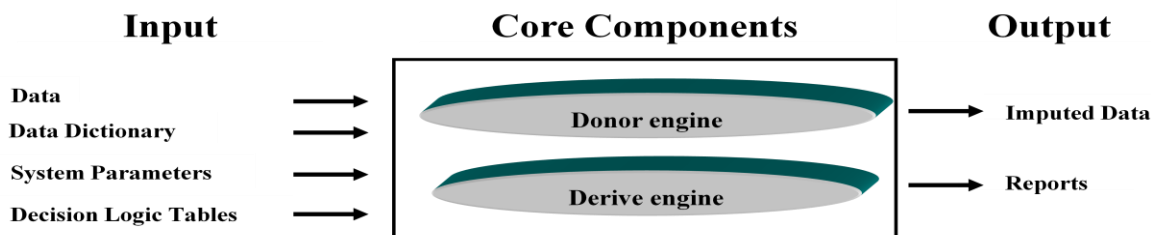
2. CANCEIS Overview

CANCEIS is based on the NIM methodology (Bankier, 2011). As an alternative to the well known Fellegi-Holt methodology (Fellegi, I.P. and Holt, D., 1976), NIM finds potential donors first and then decides the minimum number of variables to be imputed based on each potential donor. The reversal of the imputation steps fulfills the data-driven approach and gives NIM significant computational advantages, while still meeting the objectives of imputing the fewest variables possible and preserving subpopulation distributions as much as possible. Note that the phrase “fewest variables possible” is an over-simplification as, for example, CANCEIS would prefer, in general, to change the values of two variables by a very small amount than to change the value of one variable by a very large amount (more details in Section 2.4).

Due to its computational advantages, CANCEIS can process very big data files, *i.e.*, a great number of records and variables, and allows users to specify a large number of edits. CANCEIS can work with categorical, numerical and alphanumerical variables simultaneously, which gives the widest scope for finding the most suitable donors.

The system consists of two components: derive engine and donor engine. The derive engine performs deterministic imputation and creates new variables, and the donor engine performs the donor imputation. Around the two core components, there are four types of input files provided by users and several output files produced by the system, as shown in Figure 1 below.

Figure 1: CANCEIS Components



The input and output files of CANCEIS are introduced first and the core components of the system are described next.

2.1 Input files

The input data file contains records to be processed. The object to be processed in CANCEIS is the unit. A unit could be a single record or it could consist of several records. When the latter is the case, records under a unit are called subunits, such as, household members within a household. Note that CANCEIS can process only units containing identical numbers of subunits together.

In order to increase the effectiveness and efficiency of donor searching, similar units will appear on the file as closely together as possible based on subject matter expertise. Using the processing of the Canadian Census as an example, in the demography process, households are arranged according to their geographical proximity. This is based on the long-standing assumption that similar families live in the same neighbourhood. However, in other situations, a different sort order may be preferable. For example, in the income process, units are ordered by their total income when imputing for the income components. In turn, the subunits can be ordered within a unit according to a chosen order, and the evaluation of the similarity of two units is accomplished by evaluating the similarity of the corresponding subunits.

The data dictionary provides information to be used in the module, such as the variable names and types, and all possible and valid values for variables. It is quite straightforward for users to construct a data dictionary, either in text or Excel format, using the templates provided. CANCEIS works with user-defined text labels for categorical variables instead of the associated numeric codes, *e.g.*, “MALE” and “FEMALE” rather than “1” and “2”.

The whole E&I process is controlled through the system parameters. The system parameters play an important role in the flexibility, adaptability and efficiency of CANCEIS by providing users full control over the vast array of choices available. To name a few, users can have control on editing, searching for donors, reordering of subunits, auditing of the E&I process, *etc.*

In CANCEIS, Decision Logic Tables (DLTs) are used to specify edit rules for both derive and donor modules. The edit rules defined in CANCEIS are called conflict edit rules as the system is looking for anomalies in the data and not the norms. In other words, a conflict arises when there are inconsistencies between the variables. In derive modules some edit rules specify conditions, which are the conflict situations in data with respect to the variables of interest, and actions, which are the remedies to resolve the conflict. These actions are how derive DLTs perform deterministic imputation. Other edit rules in derive DLTs specify the creation of new variables under certain given conditions. On the other hand, in donor modules, typical DLTs specify conflict situations that have to be resolved without a specified action provided by the user. There are no actions specified, since the purpose is to find a donor in order to resolve the inconsistencies.

Figures 2 and 3 are examples of DLTs to be used in derive and donor modules, respectively. In the two Figures, the propositions (*i.e.*, the rows in the Figures) of the first edit rule (*i.e.*, second column in the Figures) identify the inconsistency (*i.e.*, conflict) of someone who is less than 15 years old and is defined having the MARST (*i.e.*, marital status) of EVER_MARRIED (which is defined as either married or divorced or widowed). The derive DLT prescribes the solution of deterministically changing the marital status to “single”. On the other hand, the donor DLT would prescribe that a donor must be found so that either AGE or MARST, or both variables, will be imputed in order to resolve the conflict. Note that both AGE and MARST will, generally, not be imputed together since that would violate the minimum change principle.

In these DLTs, we can see two further features of CANCEIS. First is the expression CLASS(EVER_MARRIED) which refers to the group of some values of the variable MARST, defined in the data dictionary by users. If this feature is not used then it would be necessary to have separate edit rules for each of these values. A second feature pertains to writing common edits applicable to all subunits within a unit, only once. The “#1” represents a “subunit position” and states that the rule is applied to all subunits. Without this feature, it would be necessary to repeat the same edit rules for each subunit.

Figure 2: Derive DLT

Propositions	Edit Rule 1	Edit Rule 2
AGE(#1) < 15	Y	Y
MARST(#1) = CLASS(EVER_MARRIED)	Y	
INCOME(#1) > 0		Y
MARST(#1) = SINGLE	X	
INCOME(#1) = 0		X

2.2 Output files

At the end of an E&I process, edited/imputed data as well as some reports are generated. In particular, as one of the output files, the audit trail from a donor module provides a very useful step-by-step report showing how the best donor for a given unit was found. Another output file provides information on how many times a given donor was used. The number of output files and extent of the details provided is controlled by the user through system parameters.

Figure 3: Donor DLT

Propositions	Edit Rule 1	Edit Rule 2
AGE(#1) < 15	Y	Y
MARST(#1) = CLASS(EVER_MARRIED)	Y	
INCOME(#1) > 0		Y

2.3 Derive Engine

The derive engine has the capacity to either create new variables or perform deterministic imputation to correct systematic errors based upon subject matter experience. For this purpose, CANCEIS provides the same functionality, in its DLTs, as most computing languages, such as functions (*e.g.*, assigning random values, finding maximum value etc.), “GO TO” statements, “DO loops” and “calling other DLTs”.

2.4 Donor Engine

The donor engine is responsible for donor imputation and also has data editing features to help in performing the imputation. Editing is conducted and controlled through the specification of validity classes (acceptable values for variables) and conflict edits defined within the DLTs. Each unit is evaluated to identify invalid values (values not in the validity class) and inconsistencies (unit matching conflict edits). Those units that have no invalid values and no inconsistencies are said to pass the edits and are referred to as passed units, whereas those units that have one or more invalid values and/or one or more inconsistencies are said to fail the edits and are referred to as failed units. Conceptually, CANCEIS processes failed units successively in the order of their appearance, but in practice, CANCEIS is capable of independently processing several failed units simultaneously.

While, for user convenience, the edits can be specified in many small decision logic tables, CANCEIS creates a unified version of the edits by combining all the individual edits and deleting the redundancies, in order to make sure that records after imputation pass all edits. When imputing a particular record, it is often possible to drop many edits because the potential donor is identified first. For example, if there are no grandparents in the failed household and the potential donor household being examined, CANCEIS will drop all the edits referring to grandparents because they don’t apply. Customizing the set of edits to each failed-donor pair decreases the processing time by a lot.

For each failed unit, many passed units are considered and nearest-neighbours are identified from them. The nearest-neighbours are determined on multiple dimensions by comparing a given set of matching variables. As mentioned earlier, an effort is made to group similar units together in the input data file beforehand. This allows the closest nearest-neighbours to be found in the least time. Identification of the nearest neighbours is based on the similarity of the values between the failed and passed units for each matching variable and the weight (relative importance) of each matching variable. These concepts are made quantitative through the measurement tool D_{fp} , the distance measure between the failed and a passed unit:

$$D_{fp} = \sum_i w_i D_i(V_{fi}, V_{pi}) \quad (1)$$

where V_{fi} and V_{pi} are the values of the i^{th} matching variable for the failed and passed unit, respectively, w_i is the weight assigned to the i^{th} matching variable in the data dictionary and D_i is the distance function selected for the matching variable that evaluates the similarity of V_{fi} and V_{pi} . The weight assigned to a matching variable is usually reflective of its value as a predictor for the variables to be imputed. The choice of different weights for different variables is primarily based on subject matter expertise. CANCEIS currently has ten distance functions which enable CANCEIS to process all types of variables. Each distance function also has user-defined parameters to allow even more flexibility. The distance measure in Equation (1) allows all types of variables to be brought together on the same scale, as the values of D_i are between 0 (values “essentially equal”) and 1 (values “totally dissimilar”), and this

is why CANCEIS can treat various types of variables simultaneously. For a given failed unit, a list of the closest nearest-neighbours is retained for further evaluation.

For a given donor module, each variable appearing in the input file is defined by the user to either be imputable or non-imputable. Both imputable and non-imputable variables could be used in the module as matching variables and/or in the DLT edits specification through propositions. During the imputation procedure, imputable variables that have invalid values will always be imputed immediately since this is necessary for any successful imputation action (IA), which specifies which values will be taken from the potential nearest-neighbour. In the simplest situation, the failed unit would then pass all edits and a potential IA has been found. Otherwise, if inconsistencies are still present, then for each remaining imputable variable (for which $V_{pi} \neq V_{fi}$) the choice remains whether to impute or not. In order to facilitate an efficient search for potential IAs, a binary tree is employed where each node yields two branches representing the decision to impute or not impute for a particular variable. It is to be noted that it is possible for a given nearest-neighbour to generate more than one potential IA or none at all, for example, IAs consisting of imputing either age or marital status to resolve the conflict in the first edit of the Figure 3 DLT. Each nearest-neighbour from the list of best nearest-neighbours is in turn evaluated, by ascending D_{fp} , for potential IAs.

For the purpose of deciding the best potential IAs, CANCEIS uses the quantitative measurement tool D_{fpa}

$$D_{fpa} = \alpha D_{fa} + (1 - \alpha) D_{ap} \quad (2)$$

where, 'a' represents the imputed unit, D_{fa} and D_{ap} are defined as in Equation (1), and α is a user-defined system parameter in the range (0.5, 1]. In general, we expect that the passed units most similar to the failed unit, implying lowest D_{fp} , are the most likely to yield the best or minimum-change IA. The D_{fa} value measures the similarity of the imputed unit to the failed unit, thus measuring the minimum change aspect of the IA. On the other hand, the D_{ap} value measures the similarity of the imputed unit to the passed unit, thus measuring the plausibility aspect of the IA because the potential donor has entirely real data. Note that to enforce the minimum change principle, the parameter α must be greater than 0.5.

CANCEIS keeps a running list of the best potential IAs, *i.e.*, those with the lowest D_{fpa} , for a failed unit as the nearest-neighbours are examined. The user specifies the maximum number of IAs to retain on the list and how much worse an IA can be compared to the absolute best IA. Once all nearest-neighbours have been looked at for all potential IAs, an IA is randomly chosen from the list of all potential IAs, as the actual imputation action for the failed unit.

3. Further Control and Efficiencies

The user has fine control over most aspects of the E&I process through various parameters. These aspects include staged nearest-neighbour searching, outlier detection and control, reordering of subunits, using failed units as donors, and the ability to customize the weights and distance functions for different types of failed units, to name a few.

Donor searching is usually done in stages for two reasons. First, it may not be necessary or practical to evaluate all passed units for potential IAs. Second, it is often true that the best donor units are physically close to the failed unit, especially when units in the data file have been appropriately ordered in advance, with respect to a desired characteristic. Through system parameters, users have control over how many units are considered as nearest-neighbours and potential donors in each stage and the maximum number of stages that can be done. Before going on to an additional stage, CANCEIS will evaluate whether there has been significant improvement in the quality of potential IAs in the previous stage. If not, no further stages will be processed. The user can also specify a minimum standard for the nearest-neighbour used to generate the IA so that passed units with too high a D_{fp} relative to the failed unit will not even be considered thus saving time by not analyzing unacceptable donors for IAs, while ensuring that the final chosen IA is of a minimum standard.

Based on the best IAs found so far, CANCEIS can conclude that there is no further use generating additional branches from a node of a binary tree, since any IAs obtained from these branches would not be good enough. Similarly, CANCEIS can terminate its evaluation of the nearest neighbours for IAs in a given stage if it concludes

that the D_{fp} of the nearest neighbour units relative to the failed unit, has become too high to generate IAs of acceptable quality compared to the best found so far. Both of these mechanisms and many others, increase efficiency substantially.

4. Can CANCEIS Benefit More Users?

Based on the features described in the previous sections and our experience, it seems that CANCEIS can potentially be used by more and diverse users, beyond censuses and other social surveys. It is important to remember that at this point, CANCEIS can only do deterministic and nearest-neighbour donor imputation and not any other type of imputation such as direct regression or historical imputation. However, note that CANCEIS can use variables that are correlated to the variables of interest, which would enter a regression model, and historical versions of the variables of interest, as auxiliary matching variables in finding a donor.

CANCEIS does have limitations in a computing environment with small memory and number of CPUs, say a personal computer. However with sufficient resources, say on a multiple-processor server of reasonable size, CANCEIS, through multi-threading, can handle situations with a very large number of records, variables and edits, and do it in a reasonable time. To re-iterate, CANCEIS can be a huge benefit for E&I users who require:

- a system that can perform deterministic and donor imputation as well as derive new variables,
- the ability to process categorical, numerical, and alphanumeric variables simultaneously,
- the facility to easily define large numbers of edits,
- the capability of processing large data files quickly and efficiently,
- the flexibility to have fine control over all aspects of the process through simple user-defined parameters, and
- a software which can be used immediately on most computing platforms without the need of complex installation of custom programs.

Acknowledgements

The authors would like to thank Marcel Bureau, Mike Sirois, and Daniel Finch; all from Statistics Canada, for their useful comments and suggestion that helped improve the paper.

References

- Bankier, M., Lachance, M. and P. Poirier (1999), "A generic implementation of the nearest neighbour imputation method", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 548-553.
- Bankier, M., Poirier, P. and M. Lachance (2001), "Efficient Methodology Within the Canadian Census Edit and Imputation System (CANCEIS)", ASA Joint Statistical Meetings, Atlanta.
- Bankier, M. (2011), "Imputing Numeric and Qualitative Variables Simultaneously", A Technical Report Detailing the Methodology of CANCEIS, Internal report, Statistics Canada.
- Fellegi, I.P. and D. Holt (1976), "A systematic approach to automatic edit and imputation", *Journal of the American Statistical Association*, March 1976, Volume 71, No. 353, 17-35.

Development of the social survey processing environment

Larry MacNabb¹

Abstract

In 2009, Statistics Canada commenced development of a suite of tools supporting the major steps of the survey life cycle from pre-collection to dissemination. Built on the principal that metadata should drive the process, these tools facilitate the efficient use and sharing of information between surveys and allow for the efficient creation of survey questionnaires, survey documentation and fully processed survey data. This paper will cover the following key points: history and rationale for development of the generic processing environment; overall objectives of the project; the principles and supporting best practices that guided the development of the environment and supporting tools. An overview of the generic processing environment will be presented including a description of the flow of metadata between the various processing steps. The methodological benefits of the new environment will be discussed as well as the many challenges encountered by the development team. Finally, results to date will be presented, as will next steps for the project.

Key Words: Processing; Generic Tools; Survey.

1. Background

In 2005, a Statistics Canada cross-divisional working group conducted a review of processing activities across a variety of household based surveys within the social statistics domain. Results of the review demonstrated that business processes and supporting tools were optimized within individual processing areas. While these results appeared optimal when viewed from within a particular division this within processing area optimization presented several challenges when viewed from an overall corporate perspective.

The major disadvantages of this development approach included: the development of several tools to perform similar tasks; difficulties in updating technology; multiple systems to maintain; inefficiencies in the training of staff and finally difficulties in managing priorities for system development across program areas.

As a result of this review, the working group concluded that the development of a generic processing approach and supporting tools would make considerable strides towards resolving many of the observed challenges.

2. Guiding Development Principles

In early 2009 the Social, Health and Labour Statistics Field of Statistics Canada launched an ambitious project to develop a suite of generic processing tools to support all activities of the survey life-cycle. The survey life-cycle is defined as all activities related to pre-collection, collection, processing and dissemination. At the onset of this project, several development principles were established to guide the development of these systems.

Specifically, the developed system should:

- maximize the reuse of existing software and systems;
- minimize inefficient processes;
- fully incorporate metadata as part of the process;
- allow for the integration between and across business processes and
- facilitate sharing across survey areas.

¹Larry MacNabb, Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6. (larry.macnabb@statcan.gc.ca).

3. Metadata Driven Processes Explained

A metadata driven process is one in which the process is in essence self-documenting. One of the major challenges associated with previous processing approaches has been that the survey documentation step is usually not performed until all processing activities have been completed. This occurs despite the fact that much of the information required to prepare survey documentation is an integral component of the business processes associated with the task of processing collected survey information. The end result of this approach is that documenting surveys is very inefficient and the preparation of survey documentation has involved going back through the processing steps to extract and assimilate the relevant survey information.

To further explain the concept one can observe that the creation of questionnaire specifications for the development of a collection instrument is in essence the initial step in processing a survey. The information required to specify a questionnaire includes question text, response codes (1 = male, 2 = female), non-response codes (Not Applicable, Don't Know, Refusal, Not Stated), flows through the questionnaire and field edits applied to the data. This information is in turn used to prepare record layouts for files received from the field. With this basic information, one is able to prepare the relevant inputs necessary to complete the later stages of processing once data begins returning from the field. It also serves as the minimal information required to describe collected variables.

Once collected information begins moving through the system, metadata is used to define subsequent processing activities related to the verification of returning data, application of consistency edits and the correction and creation of derived variables. All of this information ultimately serves as further information used to describe a collected piece of information. Once processing is completed, the system can then extract the relevant pieces of metadata starting with questionnaire development to fully document a collected data set. This is generally accomplished with the production of a comprehensive data dictionary or codebook.

To achieve the ideal scenario of enter once and reuse as appropriate a metadata processing system must facilitate the capture of metadata at the appropriate stage of processing and allow this information to flow with the data and be available for use at subsequent processing steps. The optimal way to achieve this is to use the metadata to control the actual processing of collected survey information.

4. Existing Best Practices

Development of an overarching vision for a generic processing system within the Social, Health and Labour Statistics Field involved an initial review of existing best practices and tools from areas involved in the processing of household surveys. This review identified two areas with processing systems and approaches that showed considerable promise in the development of a generic processing environment.

Health Statistics Division (HSD), using Microsoft Access, had developed a comprehensive system for the creation and maintenance of survey questionnaire specifications that had been linked to their processing systems. This ultimately served as the prototype for the development of a metadata driven processing system.

Special Surveys Division (SSD) had developed the Generalised Processing System (GPS), consisting of a suite of SAS based processing modules and macros, using excel based input templates, supported by a common directory structure. This system had proven itself capable of processing a variety of different survey instruments and modes of collection. To that end, it possessed demonstrable advantages in terms of representing a very robust and efficient processing methodology.

By combining the metadata driven approach developed within HSD, with the strong processing methodology present in the GPS the development team was provided with a strong foundation upon which to begin development of a generic suite of metadata driven processing tools.

5. System Overview

5.1 Generic System Overview

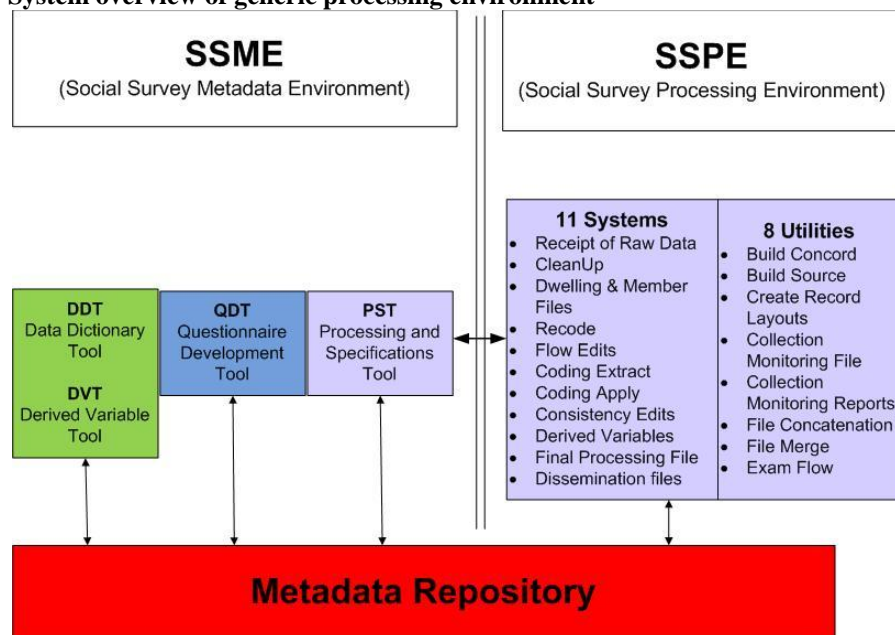
Figure 5.1-1 provides a schematic representation of the vision for the suite of tools being developed to support the survey life-cycle. The foundation of the system is the metadata repository, which is the database where all metadata is stored and is used by the various supporting tools to access and store information required for survey processing. This repository is supported by common routines used to manage metadata and ensure internal consistency of the database.

The Social Survey Metadata Environment (SSME) consists of a suite of tools used to develop and manipulate survey metadata. It consists of:

- the Questionnaire Development Tool (QDT) used for questionnaire development,
- the Data Dictionary Tool (DDT) used for the creation of survey code books,
- the Derived Variable Tool (DVT) used to manage and document the creation of derived variables and
- the Processing Specification Tool (PST), which serves as the bridge between the metadata environment and the processing system.

The Social Survey Processing Environment (SSPE) is the heart of the processing system. In its current iteration it is comprised of a suite of eleven individual processing systems associated with specific processing tasks and eight supporting utilities available for use by all systems within the SSPE.

Figure 5.1-1
System overview of generic processing environment



5.2 Social Survey Processing Environment

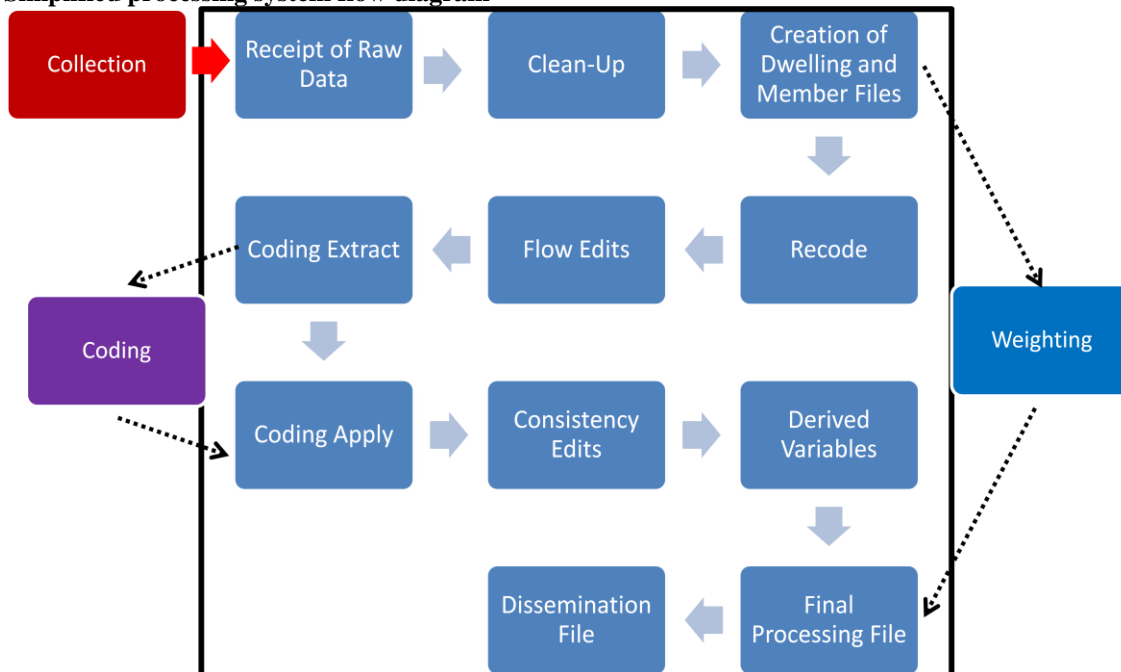
Figure 5.2-1 provides a basic schematic of a simplified processing model using the suite of modules contained within the SSPE. Though the description of each processing step is beyond the scope of this paper, there are several key aspects of the processing system that warrant further explanation.

First, the system is designed in such a way that the outputs for one processing step serve as the inputs for subsequent steps. This approach ensures consistency between processing steps and allows for consistent error checking between one step and the next.

Second, the system is designed so that the integration of one step to the next can be customized based on individual processing requirements and each step can be run more than once depending on the requirements. For example in some circumstances it may be beneficial to run flow edits before moving to the coding steps and then reapply the flow edit step to ensure coding did not impact the integrity of the flows through the questionnaire.

Finally, the system is designed to allow for the movement of data into and out of the SSPE to utilize other systems such as those used in survey weighting or coding. Over the long term, this approach will ultimately allow for a smooth and efficient integration of new and improved generic systems as they evolve within Statistics Canada. During the short and medium term, it will allow individual survey areas to gradually transition to the new environment and allow them to utilize any existing tools that have not been identified as part of the initial generic tool set.

Figure 5.2-1
Simplified processing system flow diagram



6. Methodological Benefits

There are many methodological advantages associated with the use of the generic SSPE suite of systems.

The system supports the use of a common business process, thereby facilitating the development of best practices and comparisons across survey areas. This in turn will allow all areas to adopt revised and updated approaches created within individual subject matter areas. For example, the creation of edit specification for the treatment of household income data as designed by the Income Statistics Division can easily be made available to all survey areas, thereby ensuring a consistent treatment of income variables across all social survey programs.

The practice of having a specific processing step performing only one task allows for improved focus and reduced potential for error. The end result is that by ensuring data has been processed correctly at each individual step, before moving on to the next, processing areas will spend less time revisiting previous processing steps to search for errors that they may encounter during the later stages of processing.

The system is in essence self-documenting using a common directory structure and standard input templates. This approach will facilitate the movement of staff between survey areas and reduce training requirements allowing for

greater staff mobility. It will also ensure the consistent application of processing steps over time, resulting in consistent outputs being generated by survey areas for subsequent cycles of a given survey, with common content.

The system also makes extensive use of automated error reporting, simplifying the task of monitoring data quality and ensuring consistent outputs. This will allow survey areas to spend less time on processing and more time on analysis and dissemination. It will also allow more junior staff to complete a given processing step, further improving the overall operational efficiency of processing activities within Statistics Canada.

7. Development Challenges

Development of a generic processing environment is not without challenge. During the requirement gathering process considerable effort was required in understanding the needs of individual areas and determining whether differences were the result of different terminology or indeed varying requirements. For example: upon further analysis the use of terms Clean Up and Verify to describe processing steps within two different processing areas turned out to represent the same processing activity.

In developing business rules contained within the system, the development team continually needed to balance the need to enforce standards and add complexity to the system, with the need to keep the system simple to use and maintain. Managing this balance necessitated the need to assess the cost-benefits of adding any subsequent functionality and complexity to the system. The team was also required to ensure that business rules enforced during the initial stages of development did not negatively influence their ability to efficiently implement new business rules at later stages of development.

Finally, once the system had demonstrated its effectiveness within the user community, considerable attention to the management of scope creep was required by the development team. This was required to not only manage user expectations but to also ensure planned enhancements were appropriately prioritized. In essence the system had become a victim of its own success and the subsequent enthusiasm generated within the user community required strict management and prioritization of new enhancements in order to meet the original development objectives within the allotted budgetary envelope.

8. Conclusion

Though development is ongoing, considerable progress has been made within the Social, Health and Labour Statistics Field of Statistics Canada towards the goal of creating truly generic processes and systems in support of the survey life-cycle. The system has already begun to prove itself in terms of generating efficiencies in the development of questionnaires and is facilitating the ability of Statistics Canada to further integrate processing systems across corporate business lines.

Moving forward efforts will focus on assisting survey areas with the successful transition and implementation of the new tools within their respective programs. In the future development will be expanded to include other modes of collection, beyond those associated with computer-assisted telephone and personal interviews, such as electronic and paper questionnaires.

References

Internal Report (2005), "Household Surveys Processing Working Group, Generic Household Surveys Processing, Initial Report", unpublished report, Ottawa, Canada: Statistics Canada

Internal Report (2009), "Report of the Task Force on Corporate Business Architecture", unpublished report, Ottawa, Canada: Statistics Canada.

Implementing methodology changes or enhancements in a standard processing system or how can an Advisory Group help facilitate change?

Katherine J. Thompson¹

Abstract

This paper provides background on the StEPS Methodology Advisory Group (SMAG), an ongoing team of methodologists that supports the U.S. Census Bureau's Standard Economic Processing System (StEPS). I primarily focus on operating procedures, demonstrating how this cross-divisional group of statisticians has developed and overseen the implementation of several major enhancements to the existing system, briefly touching upon this group's role in developing requirements for the re-engineered StEPS II system.

Key Words: Advisory group; Technical requirements; Methodology enhancements.

1. Introduction

In May 1995, the Economic Directorate of the U.S. Census Bureau began development of a generalized standard economic processing system (StEPS) to be used for post-sampling processing activities including data collection, post-collection processing, and tabulation. Over the past decade, over 100 monthly, quarterly, and annual surveys use or have used this processing system (Ahmed and Tasky, 2001). These surveys cover several different sectors of the U.S. economy, including construction, retail, transportation, wholesale, services, and manufacturing.

Developing a single processing system to accommodate such a variety of programs creates a host of challenges. Survey units differ. For example, the Survey of Construction (SOC) samples permits for residential homes; the Annual Retail Trade Survey samples companies or tax entities; and the Quarterly Plant Capacity Utilization Survey samples manufacturing plants. The number of collected items per form varies, as does the expected maximum number of items per form. In some cases, the program requires that the survey unit provide a roster of subunits, each of which provides a complete set of data. Data collection needs differ, ranging from mail-out/mail-back to personal interview to mail, fax, or internet collection. Programs may need to distinguish between reporting units—established for data collection by the survey unit – and tabulation units – established for estimation by the program managers. With business surveys, the survey unit composition can change over time due to mergers, divestitures, and acquisitions. And of course, the methodological treatment of tabulation unit data -- editing, imputation, weighting, outlier adjustment, estimation, and variance estimation – differs as well. Despite these challenges, there are many advantages of a consolidated processing system. Using the same system facilitates knowledge sharing among the different areas and reduces the need for specialized staff. It is easier to maintain one centralized set of programs than to maintain several separate processing systems. Multiple programs can take advantage of a single processing enhancement.

Three surveys used StEPS when it was introduced in 1998. Since then, over 100 surveys use or have used this system; we are currently forecasting that twenty-three surveys will be using StEPS in the upcoming fiscal year. Each new survey's migration facilitated revisions to existing StEPS modules. Proposed revisions to StEPS are vetted by a change control board (CCB) comprised of subject matter experts. Proposed changes to data collection and auditing processes often have little impact on other survey processes. However, this is rarely the case with proposed changes to post-collection modules, *i.e.*, editing, imputation, interactive data review and correction, estimation, variance estimation, and disclosure avoidance. Consequently in 2006, the directorate established the StEPS Methodology Advisory Group (SMAG). The SMAG's purpose is threefold:

¹Katherine J. Thompson, Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau, Washington, DC 20233 (Katherine.J.Thompson@census.gov).

1. Review any proposed changes to StEPS from a methodological perspective;
2. Recommend changes to StEPS to enhance methodology. Develop and validate detailed non-functional (technical) requirements related to methodology; and
3. Develop standard procedures from a best practices perspective that addresses all Research and Methodology area requirements for given procedures.

This paper provides background on the SMAG's composition, primarily focusing on the SMAG's operating procedures. This paper does not describe the specific functions of StEPS. For this, see Ahmed and Tasky (2001), Sigman (2000), and Sigman (2001).

2. Evolution of StEPS Administration

Within the Economic Directorate program areas, **subject matter divisions** have been established with the mission of producing accurate, efficient, and timely production of estimates for *specific* economic programs. Their responsibilities include survey design, data collection/processing, tabulation, and dissemination. There are three subject matter divisions that use StEPS: Company Statistics Division (CSD), Manufacturing and Construction Division (MCD), and Services Sector Statistics Division (SSSD). **Support divisions** provide directorate-wide support in specified service area. The Economic Planning and Coordination Division (EPCD) provides directorate-wide StEPS support and administration of survey migration, change control, testing, requirements development, training, and project management. The Economic Programming Division (EPD) houses the directorate's programmers. Lastly, the Office of Statistical Methods and Research for Economic Programs (OSMREP) supports the directorate's programs by conducting general research in statistical methods, by collaborating in the implementation of recommended methods to ongoing programs, by offering technical training, and by providing "expert" advice upon request. In the subject matter divisions, mathematical statisticians provide research and methodology support for designated programs. In contrast, OSMREP staffs are organized by specialty areas, such as sample survey methods, disclosure avoidance, and time series methods.

StEPS originated with a small dedicated team of programmers, subject matter experts, and methodologists located in EPCD. The development process was agile, and the project's scope was flexible. Initially, changes in StEPS modules were incorporated as requested on a flow basis. As the number of programs using StEPS increased, this responsive approach became impossible. The first version of a formal change control process for StEPS was introduced in 1998, with all change requests (CRs) being entered into the Census Bureau's Remedy system and the complete list of change requests maintained in EPCD. Once a month, the User Review Board -- a group of program managers and StEPS development programmers -- met to prioritize the change requests and to track resolution.

By 2001, 90 surveys were using StEPS (Ahmed and Tasky, 2001). Enhancements to StEPS were generally produced with survey migration; fixes to StEPS procedures were introduced concurrently. In the meantime, the user community was establishing specialized "user groups": the StEPS Users Group was established in 2000; the StEPS Estimation User Group was established in 2001; and the StEPS Imputation Users Group was established in 2004. These groups provided forums for discussion and resolution of processing problems and tended to focus on the usage and enhancement of existing software/modules. In 2004, the StEPS Change Control Board (CCB) was established, implementing a formal change control process. The StEPS CCB is chaired and coordinated by EPCD. Members are assigned to the CCB by appointment, and each division has a fixed number of representatives.

The StEPS Governing Board (SGB) was established in November 2006 to "provide systematic and representative approach to the guidance and direction of StEPS" (Russell, 2011). The SGB comprises upper level program managers (Assistant Division Chiefs). The SGB "provide(s) high-level guidance to and oversight of the development and maintenance of StEPS, legacy system and future releases...determine(s) plans and priorities for StEPS improvements and resolve(s) issues and develop(s) policies to improve StEPS operations." By establishing the SGB, the Economic Directorate designated an "owner" for this very large scale processing system and for associated projects. From a change control perspective, it ended the issue of resolution of disputed CRs. The "buck stops here" with the SGB. The SGB requires documentation and research for all major decisions. Annual operating plans outline initiatives and provide justification for chartered projects. And presentation of disputed CRs to the SGB must be accompanied by supporting evidence for and against the CR (*e.g.*, technical benefits, number of programs involved, requirements-development and programming effort).

The establishment of the SGB indirectly led to the establishment of the SMAG. In general, the StEPS CCB members are subject matter experts, not methodologists. Methodology-related CRs were not easily addressed in this forum. Each CR requires thorough justification, and the CCB must determine the level-of-effort needed to implement the CR. Recognizing the “ripple effect” of making a change in a general module applied to several programs, no one wanted to “rubber stamp” methodological CRs. However, the expertise did not reside with the StEPS CCB. Consequently, the Associate Director of Economic Programs established a separate StEPS Methodology Advisory Group (SMAG) in 2006. All divisions that use StEPS for data processing are represented on this group. Table 2-1 lists the responsibilities of the SMAG members by division. The SMAG partners with the StEPS CCB, reviewing and providing recommendations and supporting documentation for any proposed change to StEPS that involves a computation or methodology change. Representatives are responsible for consulting the appropriate methodologists and for apprising the subject-matter experts of potential implications for their divisions’ program of methodological changes/enhancements in StEPS. The SMAG members are appointed by their respective Assistant Division Chiefs and are managers or are authorized by managers to make decisions on topics under discussion, and are generally at a supervisory or middle management level.

**Table 2-1
SMAG Composition**

Division	Responsibilities
OSMREP	<ul style="list-style-type: none"> • Chair • Administrative functions • Provide leaders for subgroups and user groups
CSD, MCD, SSSD	<ul style="list-style-type: none"> • Provide subject-matter expertise with respect to application of methods to ongoing programs
EPCD and EPD	<ul style="list-style-type: none"> • Serve as liaisons to their divisions and to CCB • Provide resources for requirements development and design approval

As an advisory board, the SMAG develops “best practices.” When established in 2006, the SMAG included representatives from subject matter divisions that did not use StEPS for processing. Over time, these divisions elected not to participate in the SMAG: once a best practice has been established, the SMAG focuses on implementation in StEPS. The SMAG’s documentation is available to the directorate, and other divisions can and have participated on SMAG subgroups when the topic is of broad interest.

3. SMAG Operating Procedures

The SMAG charter establishes the purpose, in-scope activities, resource requirements, and operating assumptions for the SMAG; the SMAG bylaws outline the SMAG operating procedures and include detailed information on membership, committees/subteams, meetings, voting, and documentation. Although the SMAG has established formal operating procedures, it is important to note that the operating procedures are the skeleton. The flesh of the advisory group – how SMAG operates – is considerably more dynamic. Because members are appointed, SMAG membership tends to be consistent. More important, SMAG members share a common vision. The SMAG objective is always to implement methodology that is beneficial to all areas and detrimental to none. We strive to obtain unanimous approval of proposed methods from the division representatives. Group discussion encouraged on any topic, with group email is used as a supplement, not as a replacement for discussion. That said, group discussion time is limited due to the SMAG’s membership composition.

The operating procedures that were codified developed over time through trial and error. The SMAG utilizes three forums for initiatives: group decision making, subgroups, and committees/user groups.

3.1 Group Decision Making Process

The group decision making process utilizes scheduled meetings to develop and approve all deliverables. Generally, a facilitator leads the group discussion, and one person is designated as record-taker. Outside-meeting assignments are limited and are generally restricted to information gathering, although the facilitator may have additional tasks. The following examples illustrated the usage of this process.

Unit Response Rates The first SMAG initiative was mandated by the Office of Management and Budget (OMB) Standards and Guidelines for Statistical Surveys (2006), which requires that ongoing programs produce and publish response rates using standard formulas. Working with existing draft documentation, the SMAG responsibilities included:

- Establishing an accepted directorate-wide definition of “respondents.” This included overseeing the revision of the program-level respondent definitions and determining how to implement these definitions in StEPS
- Reviewing and modifying StEPS item and unit-level flagging rules
- Writing requirements for all computations and collaborating in the development, testing, and documentation of these requirements
- Developing and conducting directorate-wide training
- Developing use cases and testing the StEPS code

This project took approximately five months of weekly meetings to complete. In addition, small teams of SMAG members convened outside of the meetings to develop proposals and the SMAG chair worked directly with the StEPS development programmers to implement and test the new unit and item flagging rules.

Weighted Item Response Rates (Total Quantity and Quantity Response Rates) These mandated item level response rates measure the weighted proportion of a key estimates reported by responding units and from equivalent quality sources; total quantity response rates allow both types of data in the numerator, whereas quantity response rates are restricted to directly-obtained respondent data. The SMAG responsibilities for developing these metrics were nearly identical to those for the unit response rates. However, there were additional technical considerations, such as determining which weighting adjustments needed to be included in computations, developing rules for “eligible” data items, and determining appropriate treatment of real-valued items in computations. Consequently, as part of the requirements development process, the SMAG held a series of directorate-wide seminars to gather subject matter expertise. This project took approximately six months to complete.

Historic Percentage Change (Unit Level) An indicator survey requested the ability to view the percentage change for a selected item at the survey unit level in review and corrections. This metric is quite useful for positively-valued data items such as sales or employment, but not useful for real-valued items such as income, and it cannot be calculated for increases or decreases from zero. The SMAG developed detailed requirements for computation and display of this metric. The implemented metric is available to all programs that use StEPS.

Resistant Fences Bechtel (2011) describes how the SMAG developed requirements for and implemented the resistant fences outlier detection methodology into StEPS.

The advantages of a completely democratic process are obvious. The open discussion leads to thoroughly vetted requirements with all division perspectives considered. The consensus making decision process facilitated unanimous SMAG recommendations. For the projects described above, all recommendations were adopted by the StEPS CCB, with the majority of CCB discussion focusing on timing and resource allocation for implementation.

Although the group decision making process is thorough, it can be slow, especially when subject-matter experts are not in the same room. The process can also “drag down” meetings when topics are restricted to one area of expertise that may not be shared by all SMAG members. Finally, there are concerns that the technical requirements should not be developed without accompanying functional (usage) requirements, but that latter responsibility was assigned to a separate team of subject matter experts. Consequently, after completing the resistant fences initiative (Bechtel, 2011), the SMAG decided to establish “subgroups” (subteams) for short term projects.

3.2 Subgroups/Subteams

SMAG subgroups/subteams are committees that are established for single projects with an established timeframe. These are chartered projects with specific milestones and pre-approved deliverables. Membership on a SMAG subgroup is determined on a case-by-case basis by the project-sponsoring divisions and is not restricted to SMAG members, although OSMREP always provides the subgroup chair and performs all administrative support functions. SMAG subgroups meet independently and provide regular reports on progress to the SMAG. Subgroup recommendations must be approved by the SMAG before being forwarded for any further action. The following examples illustrate the SMAG usage of subteams.

Link Relative Estimation The link-relative estimator combines a periodically obtained benchmark with a restricted survey estimate of change to produce a current period estimate (Madow and Madow, 1978). This estimator is used by two economic indicators: the Manufacturing, Shipments, and Inventories (M3) program, a non-probability sample; and the Advance Monthly Retail Trade Survey (MARTS), a probability sample. MARTS migrated to StEPS in 2010; the M3 program had migrated in 2003. Consequently, there were existing data-entry and estimation programs in StEPS needed for MARTS processing. A team comprised of two OSMREP members and subject matter experts from each program (methodologists and program managers) convened for three months to develop comprehensive requirements for link relative estimation, variance estimation, and review and correction of input data. The team wrote the technical requirements for modifications to existing software, performed coding and conducted testing, and developed training for methodologists, programmers, and analysis.

Audit Table Team Ongoing programs in the Economic Directorate must participate in the Quality Audit program, which checks for compliance with the OMB Standards. The purpose of this chartered project was to develop requirements for automated production in StEPS of selected tables that provide evidence of compliance with OMB Quality Standards. For this, the SGB established a team comprised of three OSMREP members, three methodologists from CSD, MCD, and SSSD, and four subject matter experts (from the same divisions). This team met bi-weekly for one year. The team's purpose was to develop proposals for automatically generated tables and graphs. Each proposed table was accompanied by high level requirements for implementation in StEPS. The SGB endorsed the team's final recommendation, although the inclusion of the tables and graphs has been postponed until StEPS completes its redesign.

Imputation Base Team As a response to several CRs from different sources requesting changes to the StEPS imputation module, the SMAG convened a team comprised of expert methodologists to document best practices for constructing an imputation base or selecting source data for imputation methods available in StEPS. This group met for fourteen months. The team's recommendations were presented to the directorate in January 2010, and the formal document was issued on April 20, 2010.

In general, the usage of dedicated subgroups has been quite successful. These groups provide professional development opportunities to non-SMAG members and ensure complete and informed representation from stakeholders. Team members are assigned to the team with the understanding that they are working on a short term project with clearly outlined project obligations and a non-negotiable scope. Moreover, the subgroup members have a vested interest in the specific project outcome. Managing these projects can be, however, quite challenging. OSMREP provides subgroup chairs. The team members are from other divisions and have competing production duties. "Matrix management" is the rule, not the exception, and the team leader must be very careful in work assignments to ensure that the projects are completed within the established timeframe.

3.3 User Groups

SMAG user groups are "permanent" committees with a single topical mandate. There are two SMAG user groups: the General Imputation Users Group and the Time Series Analytical Repository (TSAR) Methodology Advisory Group. Membership in both groups is determined by the sponsoring divisions. OSMREP provides the chair of the user groups, and the chairs provide regular reports to the SMAG. Like the SMAG, both user groups have a project charter; bylaws are optional. User groups are responsible for reviewing all topic-related CRs. If the user group recommends implementing the change, then the group is also charged with developing the technical requirements. User groups also propose topic-related methodological improvements to StEPS, and have testing responsibilities including developing test cases and test scenarios and conducting user testing of new or enhanced StEPS code.

Like the SMAG subgroups, there are professional advantages to belonging to a SMAG user group. More important, the committees' outputs benefit the directorate by ensuring that recommendations are developed by "experts." Membership on a committee is less restricted than membership on the SMAG or on a SMAG subgroup; as long as the participant has division approval, they can attend and participate in committee meetings. The OSMREP chair is challenged by "matrix management" when making assignments. In addition, there are some concerns about back-up membership, since these committees are permanent, and staff is not.

4. Conclusion

The SMAG was established in 2006 under the authority of the Associate Director for Economic Programs with the full and consensual support of the sponsoring divisions. At the time, SMAG filled a vacuum, providing a formal forum for vetting methodological CRs to StEPS. Membership on the SMAG also created quite a bit of additional work for its members. Managing that work in such a way that initiatives are accomplished in a reasonable time-frame without overly burdening participants is a constant challenge.

The SMAG has proved effective for many reasons. First, this advisory group has greatly benefitted from the endorsement from upper management, including the Associate Director, the SGB, and the sponsoring divisions. Second, the organizational structure of the Economic Directorate allows the SMAG to operate smoothly. OSMREP provides organizational stability to the SMAG and its subteams/committees, providing chairs and performing all administrative functions. This allows other SMAG and SMAG-affiliate participants to restrict their obligations to reviewing the relevant material and providing input. Lastly, the SMAG has an excellent group dynamic. In 2009, the Associate Director for Economic Programs authorized the redesign of the StEPS. This redesign will include a new database structure and a new user interface. Although the number of StEPS processed surveys has shrunk to twenty-three, this is by no means a small project. The SMAG's input into the development of the new system is, however, a constant.

Acknowledgements

This paper is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. I thank Anne Russell and Xijian Liu for their careful review and thoughtful comments on earlier versions of this manuscript, and the members of the SMAG for their comments and suggestions on the presentation.

References

- Ahmed, S.A. and D.L. Tasky (2001), "Are generalized systems the way of the future: A case study on the Standard Economic Processing System (StEPS)?" *Proceedings of the Section on Survey Research Methods*, American Statistical Association.
- Bechtel, L.T. (2011), "Retro-fitting a simpler outlier detection procedure into a complex generalized system", *Proceeding of the 2011 Statistics Canada Symposium*.
- Federal Register Notice (2006), OMB Standards and Guidelines for Statistical Surveys.
- Madow, L.H. and W.G. Madow (1978), "On link relative estimation", *Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 534-539.
- Russell, A.S. (forthcoming in 2011), "StEPS Governing Board Charter", unpublished U.S. Census Bureau memorandum, available upon request.
- Sigman, R.S. (2000), "Estimation and variance estimation in a standardized economic processing system", *Proceedings of the Second International Conference on Establishment Surveys*, pp. 677-686.

Sigman, R.S. (2001), "Editing and imputation in a standard economic processing system", *Proceedings of Statistics Canada Symposium 2001*.

SESSION 10A
FRAMEWORK

Developing a methodology for the Canadian Framework for Culture Statistics

Mary K. Allen¹

Abstract

For at least twenty years, there have been efforts, both at the national and international levels, to develop standard measures for the arts and culture. This has led to the UNESCO framework for culture statistics, and, in Canada, to the 2004 Framework for Culture Statistics. These frameworks were developed to meet the specific statistical needs of culture researchers who were unable to derive comparable and standard measures from existing data based on standard classification systems.

The 2004 Framework provided a structure for measuring culture, but was defined by and dependent upon existing classification systems such as NAICS 2002 and the SCG. This meant that, while useful for research purposed, the framework was restricted to existing data and standards.

In 2011, the Canadian Framework for Culture Statistics was redesigned with a more conceptual approach. Instead of defining culture in terms of existing standard classifications, the 2011 Framework defined culture and its components conceptually and then developed a methodology to guide the mapping of existing and future classification systems to the framework. This new methodology supports the creation of standard, replicable lists of classification codes to support culture research and guide future data development, including exploring opportunities to address data gaps. It provides a standard, objective method for determining what is, for example, a culture product, or a culture industry.

This paper presents the method developed and published as the 2011 Canadian Framework for Culture Statistics.

Key Words: Culture; Framework.

References

Statistics Canada (2011), “Conceptual Framework for Culture Statistics 2011”, Statistics Canada Catalogue no. 87-542-X No. 001, Ottawa, Ontario, Canadian Framework for Culture Statistics.

Statistics Canada (2011), “Classification Guide for the Canadian Framework for Culture Statistics 2011”, Statistics Canada Catalogue no. 87-542-X No. 002. Ottawa, Ontario, Canadian Framework for Culture Statistics.

¹Mary K. Allen, Statistics Canada, 100 Tunney’s Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6.

How many Canadians live in a city? Conceptualization, definition and proposed dissemination for alternative standards

Ray D. Bollman and Peter Murphy¹

Abstract

Statistics Canada does not assemble its data on cities in a format to answer the question “How Many Canadians Live in a City?” The objectives of this paper are:

1. To describe three alternative ways to delineate a ‘city.’
2. To present a time-series of Canada’s population living in a ‘city’ for
 - a. each of the three delineations; and
 - b. various possible cut-off sizes to be classified as a ‘city.’
3. To discuss the level of ‘urbanization’ implied by the share of Canada’s population living in cities of different sizes for different delineations.
4. To show, for each delineation of a city, which part of the urbanization spectrum was growing the fastest? (*i.e.*, which size class of a city was growing, or urbanizing, at the greatest rate?)
5. To show whether this rate of urbanization was due to demographic factors (births, deaths, immigration or net migration within Canada) or due to reclassification or amalgamation (*i.e.*, did the locality change size classes between two census periods and/or did the boundaries of the locality change between two census periods?).

Key Words: City; Urban; Rural; Population.

1. Introduction

Data users ask Statistics Canada “How many Canadians live in a city?” There is no simple way to find the answer to this question as Statistics Canada does not assemble and publish data on levels of population and rates of urbanization for (a) alternative definitions of ‘cities;’ and (b) alternative size cut-offs for each definition of a ‘city.’

The objectives of this paper are:

1. To describe three alternative ways to delineate a ‘city.’
2. To present a time-series of Canada’s population living in a ‘city’ for
 - a. each of the three delineations; and
 - b. various possible cut-off sizes to be classified as a ‘city.’
3. To discuss the level of ‘urbanization’ implied by the share of Canada’s population living in cities of different sizes for different delineations.
4. To show, for each delineation of a city, which part of the urbanization spectrum was growing the fastest? (*i.e.*, which size class of a city was growing, or urbanizing, at the greatest rate?)
5. To show whether this rate of urbanization was due to demographic factors (births, deaths, immigration or net migration within Canada) or due to reclassification or amalgamation (*i.e.*, did the locality change size classes between two census periods and/or did the boundaries of the locality change between two census periods?).

¹Ray Bollman (RayD.Bollman@sasktel.net) recently retired as Chief, Rural Research Group, Statistics Canada and Peter Murphy (Peter.Murphy@statcan.gc.ca) is a Chief in the Geography Division, Statistics Canada, 170 Tunney's Pasture Driveway, Ottawa, Ontario, K1A 0T6.

2. Three ways of conceptualizing a ‘city’

Demographers have used three alternative ways of conceptualizing a ‘city’ (see, for example, Puderer (2009)). These three measures always start with the population density of a locality and the boundaries are determined in different ways. The three classifications are presented below.

2.1 Form or morphology

This concept describes the built-up area of a population settlement and may be simply described as the ‘windshield view.’ On your Sunday afternoon drive, when are you out of the city? Arguably, this is the target clientele for an urban transportation planner.

2.2 Administrative unit

This concept represents the view of the mayor. How many people live in my administrative unit? From the point of view of the resident—to whom do I pay my taxes and which administration is responsible for delivering local services. On your Sunday afternoon drive, you usually cannot see the boundary of the administrative city—although there is often a sign to say “Welcome to our city of xxx,xxx residents.”

2.3 Functional area

This concept is based on the idea that “We are all in this together.” Thus, from the point of view of citizens and from the point of view of investors, we all share in the outcomes of good development projects and we all share in the outcomes of bad development projects. In this sense, the agglomerated population operates together as a functional area, regardless of form and regardless of administrative boundaries.

3. Implementation of these concepts at Statistics Canada

3.1 Form or morphology

At Statistics Canada, form or morphology of a built-up area is delineated for all population centres with built-up cores, with a population density of 400 inhabitants per km² or more and with a total population of 1,000 more. Population centres were formerly known as ‘census urban areas.’ For details, see Statistics Canada (2007). A similar analysis of the way to define and delineate built-up areas was reported by Hofmann *et al.* (2010a, 2010b). Their methodology fine-tuned the amount of land delineated in built-up areas but made a relatively small adjustment to the population delineated within population centres.

3.2 Administrative unit

Within Statistics Canada, a census subdivision (CSD) is defined for each incorporated town and each incorporated municipality. In cooperation with each province and territory, land (and the residents) that are not part of an incorporated municipal government are delineated into CSDs for statistical purposes. For details, see Statistics Canada (2007).

3.3 Functional area

At Statistics Canada, census metropolitan areas (CMAs) and census agglomerations (CAs) are delineated for any built-up core of a population of 10,000 or more and includes all neighbouring CSDs where 50% or more of the employed residents commute to the built-up core of the CSD. Thus, commuting rates are used to measure or proxy the areas around the built-up core that ‘function together.’

4. Results

4.1 Population centres that represent form or morphology

In terms of the level of urbanization, three population centres (Montreal, Toronto and Vancouver) had a population of 1 million or more and they represented 32% of Canada's population (Table 4-1.1). In 2006, there were 895 population centres with a population of 1,000 or more.

**Table 4.1-1
Population by size of population centre, Canada, 1991 to 2006**

Population size class of population centre ¹	Total population and number of settlements ¹												Percent distribution of the population								Percent change of the population within constant boundaries and constant classification						
	Within 1991 boundaries and classification			Within 1996 boundaries and classification			Within 2001 boundaries and classification			Within 2006 boundaries and classification			Within 1991 boundaries and classification		Within 1996 boundaries and classification		Within 2001 boundaries and classification		Within 2006 boundaries and classification		1986 to 1991	1991 to 1996	1996 ³ to 2001	2001 to 2006			
	1986	1991	Number of population centres ¹	1991	1996	Number of population centres ¹	1996 ³	2001	Number of population centres ¹	2001	2006	Number of population centres ¹	1986	1991	1991	1996	1996	2001	2001	2006							
1 million and over	7,158,005	7,865,789	3	7,962,741	8,539,938	3	...	9,412,027	3	9,372,715	10,022,987	3	28	29	29	30	...	31	31	32	9.9	7.2	...	6.9			
500,000 to 999,999	3,623,967	3,905,298	6	3,934,353	4,079,970	6	...	4,369,921	6	4,378,914	4,660,213	6	14	14	14	14	...	15	15	15	7.8	3.7	...	6.4			
100,000 to 499,999	2,705,621	2,888,861	15	3,135,825	3,274,783	17	...	3,728,333	20	3,774,438	3,973,453	20	11	11	11	11	...	12	13	13	6.8	4.4	...	5.3			
50,000 to 99,999	1,345,360	1,479,768	21	1,561,942	1,662,590	24	...	1,481,831	22	1,549,340	1,653,109	23	5	5	6	6	...	5	5	5	10.0	6.4	...	6.7			
30,000 to 49,999	956,787	1,015,237	26	957,602	992,847	26	...	966,319	26	1,113,989	1,197,050	31	4	4	4	3	...	3	4	4	6.1	3.7	...	7.5			
10,000 to 29,999	1,445,804	1,515,559	93	1,480,773	1,555,786	97	...	1,597,682	100	1,496,398	1,563,022	100	6	6	5	5	...	5	5	5	4.8	5.1	...	4.5			
5,000 to 9,999	820,029	866,470	127	897,370	946,935	136	...	946,880	133	935,749	960,734	136	3	3	3	3	...	3	3	3	5.7	5.5	...	2.7			
2,500 to 4,999	725,056	745,143	209	720,444	748,677	210	...	786,829	222	694,921	695,905	198	3	3	3	3	...	3	2	2	2.8	3.9	...	0.1			
1,000 to 2,499	605,212	612,630	380	648,645	659,684	410	...	618,389	381	629,495	624,270	378	2	2	2	2	...	2	2	2	1.2	1.7	...	-0.8			
Under 1,000 ²	5,920,489	6,402,101	...	5,997,161	6,385,548	6,098,883	...	6,061,135	6,262,154	...	23	23	22	22	...	20	20	20	8.1	6.5	...	3.3			
All areas	25,306,330	27,296,856	880	27,296,856	28,846,758	929	...	30,007,094	913	30,007,094	31,612,897	895	100	100	100	100	...	100	100	100	7.9	5.7	...	5.4			
Population within population centres with a population size of:																											
1 million and over	7,158,005	7,865,789	3	7,962,741	8,539,938	3	...	9,412,027	3	9,372,715	10,022,987	3	28	29	29	30	...	31	31	32	9.9	7.2	...	6.9			
500,000 and over	10,781,972	11,771,087	9	11,897,094	12,619,908	9	...	13,781,948	9	13,751,629	14,683,200	9	43	43	44	44	...	46	46	46	9.2	6.1	...	6.8			
100,000 and over	13,487,593	14,659,948	24	15,032,919	15,894,691	26	...	17,510,281	29	17,526,067	18,656,653	29	53	54	55	55	...	58	58	59	8.7	5.7	...	6.5			
50,000 and over	14,832,953	16,139,716	45	16,594,861	17,557,281	50	...	18,992,112	51	19,075,407	20,309,762	52	59	59	61	61	...	63	64	64	8.8	5.8	...	6.5			
30,000 and over	15,789,740	17,154,953	71	17,552,463	18,550,128	76	...	19,958,431	77	20,189,396	21,506,812	83	62	63	64	64	...	67	67	68	8.6	5.7	...	6.5			
10,000 and over	17,235,544	18,670,512	164	19,033,236	20,105,914	173	...	21,556,113	177	21,685,794	23,069,834	183	68	68	70	70	...	72	72	73	8.3	5.6	...	6.4			
5,000 and over	18,055,573	19,536,982	291	19,930,606	21,052,849	309	...	22,502,993	310	22,621,543	24,030,568	319	71	72	73	73	...	75	75	76	8.2	5.6	...	6.2			

1. A population centre has a minimum population of 1,000 persons and a population density of at least 400 persons per square kilometre. In 2001 and 2006, this was based on the current census population count and in previous censuses, this was based on the population count of the previous census.

2. In 1991, contains 12,119 individuals in 13 settlements with a population of 834 to 999. In 1996, contains 16,477 individuals in 17 settlements with a population of 858 to 999.

3. Population data for 1996 within 2001 boundaries are not available. *Since 2001 blocks did not necessarily respect 1996 enumeration areas, it was not possible to recreate 1996 urban areas based on 2001 blocks with 100% accuracy.* See Matier, Kelly. (2008) *Delineation of 2006 Urban Areas: Challenges and Achievements* (Ottawa: Statistics Canada, Geography Working Paper Series No. 2008001, Catalogue no. 92F0138), p. 5.

Source: Statistics Canada, Census of Population, 1986 to 2006.

Following the observation of Mendelson and Lefebvre (2003), who noted localities with a core population of 50,000 or more exhibited 'metro' functions, we may also note that there were 52 population centres in 2006 with a population of 50,000 or more and they represented 64% of Canada's population.

Alternative views of the level of urbanization (for analysts preferring 'population centres' as their preferred delineation of a city) are shown in Table 4.1-1. The share of population living in population centres of 1 million or more has increased from 29% in 1991 to 32% in 2006. For population centres of 50,000 or more, they represented 59% of Canada's population in 1991 and 64% in 2006. Given that some jurisdictions (e.g., Saskatchewan) assign the term 'city' to localities with 5,000 or more population, we show that, in 2006, there were 319 population centres with 5,000 or more residents and they represented 76% of Canadians, up from 72% in 1991.

The final four columns of Table 4.1-1 show the rate of population change (i.e., the rate of urbanization) due to demographic factors (births, deaths and net migration), which is calculated within constant boundaries and using a constant size classification (both based on the situation at the end of the 5-year period). Much of the difference in growth rates across population size classes is due to net migration. Thus, we are seeing where people preferred to live or to move to. Between 2001 and 2006, population increased fastest in population centres with a population of 30,000 to 49,999 (an increase of 7.5%). Population centres of 1 to 4,999 showed virtually no change. In the 1991 to 1996 period, population centres of 1 million or more grew the fastest. In the 1986 to 1991 period, centres with a population of 50,000 to 99,999 grew the fastest. Thus, rate of urbanization due to demographic factors does not always occur in the biggest population centres.

However, for the total change in the population living in a given population centre, size class is also determined by reclassification. Reclassification might be due to (a) demographic change causing the locality to be reclassified to a larger or a smaller class of population centre or (b) two population centres may be merged which may cause the

population of the new population centre to be reclassified to another size class. Thus, between 't' and 't+1,' there is a change in the number of localities and a change in the number of Canadians enjoying the benefits and costs of a living in a locality of a given size. Due to these reclassifications, we see the change in the total number of Canadians residing in a population of a given size class (*i.e.*, the rate of urbanization in terms of the rate of change in the number of individuals living in a city of a given size, whether due to demographic change or reclassification). A calculation based on the data in Table 4.1-1 (but not shown) indicates that the size class with the largest increase in population was in the 30,000 to 49,999 class from 2001 to 2006 and in the 100,000 to 499,999 class in 1991 to 1996. Importantly, most (three-quarters or more) of the total change was due to demographic change—less than one-quarter of the total change was due to reclassification.

4.2 Census subdivisions that represent administrative areas

In 2006, in terms of the level of urbanization, there were two CSDs with 1 million or more residents (Montreal and Toronto) and they accounted for 13% of Canada's population (Table 4.2-1). This was up from one CSD (Montreal) in 1991, representing 4% of Canada's population. In 2006, there were 48 CSDs with a population of 100,000 or more, representing 52% of Canada's population (up from 36 CSDs representing 38% of Canada's population).

In general, larger CSDs (but not the largest CSDs) have a larger rate of urbanization due to demographic growth than smaller CSDs (Table 4.2-1). Demographic growth from 2001 to 2006 was the largest in CSDs with 250,000 499,999 inhabitants (up 11.6%); from 1996 to 2001, the largest growth was in the 500,000 to 999,999 class (up 8.3%); from 1991 to 1996, the largest growth was in the 250,000 to 499,999 class and from 1986 to 1991, the largest growth was in the 100,000 to 249,999 class.

However, due to considerable municipal amalgamations over recent decades, from half to three-quarters of the total change in population living in a CSD of a given size is due to reclassification. From 2001 to 2006, the largest increase in urbanization in terms of change in total population was in the 100,000 to 249,999 class; from 1996 to 2001, the largest increase was in the 1 million and over group (as Toronto CSD surpassed this threshold in this period) and in the 1991 to 1996 period, the largest increase was in the 500,000 to 999,999 group.

**Table 4.2-1
Population by size of census subdivision, Canada, 1991 to 2006**

Population size class of census subdivision (incorporated city, town or municipality)	Total population and number of census subdivisions (incorporated city, town or municipality)											Percent distribution of the population								Percent change of the population within constant boundaries and constant classification							
	Within 1991 boundaries and classification			Within 1996 boundaries and classification			Within 2001 boundaries and classification			Within 2006 boundaries and classification			Within 1991 boundaries and classification		Within 1996 boundaries and classification		Within 2001 boundaries and classification		Within 2006 boundaries and classification		1986 to 1991	1991 to 1996	1996 to 2001	2001 to 2006			
	1986	1991	Number of census subdivisions	1991	1996	Number of census subdivisions	1996	2001	Number of census subdivisions	2001	2006	Number of census subdivisions	1986	1991	1991	1996	1996	2001	2001	2006							
1 million and over	1,015,420	1,017,666	1	1,017,669	1,016,376	1	3,401,797	3,521,028	2	4,065,084	4,123,974	2	4	4	4	4	12	12	14	13	0.2	-0.1	3.5	1.4			
500,000 to 999,999	3,458,638	3,666,765	6	4,601,246	4,863,602	8	3,782,391	4,097,182	6	4,587,587	4,915,294	7	14	13	17	17	13	14	15	16	6.0	5.7	8.3	7.1			
250,000 to 499,999	2,268,988	2,495,274	7	2,048,115	2,203,177	7	2,039,440	2,202,176	6	2,396,848	2,675,280	7	9	9	8	8	7	7	8	8	10.0	7.6	8.0	11.6			
100,000 to 249,999	2,888,246	3,236,429	22	3,509,845	3,740,267	28	3,613,238	3,862,586	27	4,520,530	4,827,672	32	11	12	13	13	13	13	15	15	12.1	6.6	6.9	6.8			
50,000 to 99,999	3,192,641	3,527,719	49	3,306,687	3,505,619	50	3,471,260	3,606,808	51	2,974,961	3,212,519	45	13	13	12	12	12	12	10	10	10.5	6.0	3.9	8.0			
30,000 to 49,999	1,692,013	1,880,042	46	1,860,459	1,972,461	50	1,732,283	1,782,950	46	1,488,090	1,582,105	41	7	7	7	7	6	6	5	5	11.1	6.0	2.9	6.3			
10,000 to 29,999	3,645,807	4,024,391	249	3,921,631	4,211,771	261	4,094,033	4,230,425	262	3,762,857	3,979,858	247	14	15	14	15	14	14	13	13	10.4	7.4	3.3	5.8			
5,000 to 9,999	2,154,167	2,316,074	333	2,135,152	2,274,139	328	2,257,857	2,303,986	327	2,095,159	2,171,479	311	9	8	8	8	8	8	7	7	7.5	6.5	2.0	3.6			
2,500 to 4,999	1,853,697	1,957,740	554	1,869,067	1,958,326	550	1,710,631	1,700,611	479	1,526,596	1,546,831	444	7	7	7	7	6	6	5	5	5.6	4.8	-0.6	1.3			
1,000 to 2,499	1,823,180	1,874,269	1,200	1,783,300	1,844,056	1,174	1,549,505	1,535,309	980	1,441,071	1,453,236	922	7	7	7	6	5	5	5	5	2.8	3.4	-0.9	0.8			
500 to 999	848,576	847,852	1,174	793,549	811,781	1,121	748,339	733,174	1,023	697,045	692,586	958	3	3	3	3	3	2	2	2	-0.1	2.3	-2.0	-0.6			
250 to 499	333,976	331,271	874	326,863	329,392	880	323,461	312,991	843	318,270	308,209	832	1	1	1	1	1	1	1	1	-0.8	0.8	-3.2	-3.2			
Under 250	135,982	121,367	1,491	123,276	115,794	1,526	122,526	117,868	1,548	133,196	123,854	1,570	1	0	0	0	0	0	0	0	-10.7	-6.1	-3.8	-7.0			
All areas	25,309,331	27,296,859	6,006	27,296,859	28,846,761	5,984	28,846,761	30,007,094	5,600	30,007,094	31,612,897	5,418	100	100	100	100	100	100	100	100	7.9	5.7	4.0	5.4			
Population within census subdivisions with a population size of:																											
1 million and over	1,015,420	1,017,666	1	1,017,669	1,016,376	1	3,401,797	3,521,028	2	4,065,084	4,123,974	2	4	4	4	4	12	12	14	13	0.2	-0.1	3.5	1.4			
500,000 and over	4,474,058	4,684,431	7	5,618,915	5,879,978	9	7,184,188	7,618,210	8	8,652,671	9,039,268	9	18	17	21	20	25	25	29	29	4.7	4.6	6.0	4.5			
250,000 and over	6,743,046	7,179,705	14	7,667,030	8,083,155	16	9,223,628	9,820,386	14	11,049,519	11,714,548	16	27	26	28	28	32	33	37	37	6.5	5.4	6.5	6.0			
100,000 and over	9,631,292	10,416,134	36	11,176,875	11,823,422	44	12,836,866	13,682,972	41	15,700,049	16,542,220	48	38	38	41	41	45	46	52	52	8.1	5.8	6.6	6.2			
50,000 and over	12,823,933	13,943,853	85	14,483,562	15,329,041	94	16,308,126	17,289,780	92	18,545,010	19,754,739	93	51	51	53	53	57	58	62	62	8.7	5.8	6.0	6.5			
30,000 and over	14,515,946	15,823,895	131	16,344,021	17,301,502	144	18,040,409	19,072,730	138	20,033,100	21,336,844	134	57	58	60	60	63	64	67	67	9.0	5.9	5.7	6.5			
10,000 and over	18,159,753	19,848,286	380	20,265,652	21,513,273	405	22,134,442	23,303,155	400	23,795,757	25,316,702	381	72	73	74	75	77	78	79	80	9.3	6.2	5.3	6.4			
5,000 and over	20,313,920	22,164,360	713	22,400,804	23,787,412	733	24,392,299	25,607,141	727	25,890,916	27,488,181	692	80	81	82	82	85	85	86	87	9.1	6.2	5.0	6.2			

A census subdivision (CSD) is the general term for municipalities (incorporated cities, towns and rural municipalities, as determined by provincial/territorial legislation) or areas treated as municipal equivalents for statistical purposes (e.g., Indian reserves, Indian settlements and unorganized territories).

Source: Statistics Canada, Census of Population, 1986 to 2006.

4.3 Census metropolitan areas and census agglomerations that represent functional areas

In 2006, in terms of the level of urbanization, 45% of Canadians were living in a CMA with a population of 1 million or more—up from 29% in 1991 (Table 4.3-1).

In each five-year period from 1981 to 2006, CMAs with a population of 1 million or more showed the highest rate of urbanization in terms of demographic growth. However, in some periods, other sizes classes of CMAs or CAs recorded the largest change in urbanization in terms of change in total population: from 1996 to 2001, CMAs of 100,000 to 249,999; from 1986, CAs of 50,000 to 99,999; and from 1981 to 1986, CAs of 10,000 to 49,999. However, most (half to three-quarters) of the rate of urbanization was due to demographic growth.

Table 4.3-1
Population by size class of functional labour market area, Canada, 1981 to 2006

Size class of functional labour market area	Population										Percent distribution of population										Percent change									
	within 1981 boundaries		within 1986 boundaries		within 1991 boundaries		within 1996 boundaries		within 2001 boundaries		within 2006 boundaries		within 1981 bound-aries	within 1986 bound-aries	within 1991 bound-aries	within 1996 bound-aries	within 2001 bound-aries	within 2006 bound-aries	1976 to 1981	1981 to 1986	1986 to 1991	1991 to 1996	1996 to 2001	2001 to 2006						
	1976	1981	1981	1986	1986	1991	1991	1996	1996	2001	2001	2006	1976	1981	1986	1991	1996	2001	2006	1976	1981	1986	1991	1996	2001	2006				
	1976	1981	1981	1986	1986	1991	1991	1996	1996	2001	2001	2006	1976	1981	1986	1991	1996	2001	2006	1976	1981	1986	1991	1996	2001	2006				
Larger CMAs (1 million and over)	6,771,966	7,095,479	7,260,861	7,729,254	7,734,067	8,622,790	9,652,307	10,432,430	10,420,589	11,159,876	13,078,028	14,110,317	28	29	30	31	32	35	36	37	44	45	4.8	6.5	11.5	8.1	7.1	7.9		
Mid-sized CMAs (500,000 to 999,999)	3,370,701	3,670,790	3,822,645	4,061,654	4,050,342	4,412,478	3,500,925	3,647,683	3,647,567	3,905,672	2,025,564	2,103,094	14	15	16	16	16	13	13	13	13	7	7	8.9	6.3	8.9	4.2	7.1	3.8	
Smaller CMAs (100,000 to 499,999)	2,767,796	2,892,675	3,224,726	3,364,585	3,364,195	3,630,092	3,633,886	3,784,533	4,110,441	4,231,378	5,017,869	5,295,164	11	12	13	13	13	13	13	14	14	17	17	15.4	4.3	7.9	4.1	2.9	5.5	
Census metropolitan areas (subtotal)	12,910,463	13,658,944	14,308,232	15,155,493	15,148,604	16,665,360	16,787,118	17,864,646	18,178,597	19,296,926	20,121,461	21,508,575	53	56	59	60	61	61	62	63	64	67	68	5.9	10.0	6.4	6.2	6.9		
Larger CAs (50,000 to 99,999)	1,447,751	1,523,607	1,903,808	1,973,431	2,069,418	2,277,832	2,407,087	2,578,276	2,349,659	2,423,726	1,847,219	1,947,917	6	6	8	8	8	9	9	8	8	6	6	5.2	3.7	10.1	7.1	3.2	5.5	
Mid-sized CAs (30,000 to 49,999)	1,069,788	1,128,815	1,035,003	1,057,094	1,137,284	1,194,341	1,018,807	1,056,633	1,057,158	1,064,817	1,118,738	1,157,978	4	5	4	4	4	4	4	4	4	4	4	5.5	2.1	5.0	3.7	0.7	3.5	
Smaller CAs (10,000 to 29,999)	701,844	711,176	1,038,437	1,029,089	903,473	929,681	927,144	950,300	1,069,278	1,053,617	997,280	1,017,087	3	3	4	4	4	3	3	3	4	4	3	3	1.3	-0.9	2.9	2.5	-1.5	2.0
Census agglomerations (subtotal)	3,219,383	3,363,598	3,977,248	4,059,614	4,110,175	4,401,854	4,353,038	4,585,209	4,476,095	4,542,160	3,963,237	4,122,982	13	14	16	16	16	16	16	15	13	13	13	4.5	2.1	7.1	5.3	1.5	4.0	
Larger urban centres (subtotal)	16,129,846	17,022,542	18,285,480	19,215,107	19,258,779	21,067,214	21,140,156	22,449,855	22,654,692	23,839,086	24,084,698	25,631,557	66	70	75	76	76	77	77	78	79	79	80	81	5.5	5.1	9.4	6.2	5.2	6.4
Strong Metropolitan Influenced Zone	1,435,028	1,574,359	1,458,448	1,564,700	1,470,493	1,524,579	1,289,265	1,350,098	6	6	5	5	5	4	9.7	7.3	3.7	4.7	
Moderate Metropolitan Influenced Zone	2,280,052	2,335,157	2,289,911	2,365,175	2,307,387	2,285,538	2,203,563	2,224,347	9	9	8	8	8	7	-2.4	3.3	-0.9	0.9	
Weak Metropolitan Influenced Zone	1,952,122	1,951,974	2,041,871	2,078,342	2,027,488	1,969,211	2,077,950	2,049,199	8	7	7	7	7	7	6	0.0	1.8	-2.9	-1.4
No Metropolitan Influenced Zone	334,560	315,813	316,281	332,604	330,616	333,847	296,785	297,984	1	1	1	1	1	1	1	-5.6	5.2	1.0	0.4
RST Territories	48,790	52,342	50,192	56,085	56,085	54,833	54,833	59,712	0	0	0	0	0	0	-7.3	11.7	-2.2	8.9
Rural and small town (RST) areas (subtotal)	6,862,759	7,320,635	6,057,697	6,094,222	6,050,552	6,229,645	6,156,703	6,396,906	6,192,069	6,168,008	5,922,396	5,981,340	28	30	25	24	23	22	21	21	20	19	6.7	0.6	3.0	3.9	-0.4	1.0		
Total	22,992,605	24,343,177	24,343,177	25,309,329	25,309,331	27,296,859	27,296,859	28,846,761	28,846,761	30,007,094	30,007,094	31,612,897	100	100	100	100	100	100	100	100	100	100	100	100	5.9	4.0	7.9	5.7	4.0	
Population within functional labour markets with a population size of:																														
1 million and over	6,771,966	7,095,479	7,260,861	7,729,254	7,734,067	8,622,790	9,652,307	10,432,430	10,420,589	11,159,876	13,078,028	14,110,317	29	29	30	31	32	35	36	37	44	45	4.8	6.5	11.5	8.1	7.1	7.9		
500,000 and over	10,142,667	10,762,269	11,083,506	11,790,908	11,784,409	13,035,268	13,533,232	14,080,113	14,068,156	15,065,548	15,103,592	16,213,411	44	44	46	47	47	48	48	49	50	50	51	6.1	6.4	10.6	7.0	7.1	7.3	
100,000 and over	12,910,463	13,658,944	14,308,232	15,155,493	15,148,604	16,665,360	16,787,118	17,864,646	18,178,597	19,296,926	20,121,461	21,508,575	56	56	59	60	61	61	62	63	64	67	68	5.8	10.0	6.4	6.2	6.9		
50,000 and over	14,358,214	15,182,551	16,212,040	17,128,924	17,218,022	18,943,192	19,194,205	20,442,922	20,528,256	21,720,652	21,968,680	23,456,492	62	62	67	68	68	69	70	71	71	72	73	74	5.7	10.0	6.5	5.8	6.8	
30,000 and over	15,428,002	16,311,366	17,247,043	18,186,018	18,355,306	20,137,533	20,213,012	21,409,555	21,585,414	22,785,469	23,087,418	24,614,470	67	67	71	72	73	74	75	75	76	77	78	5.7	5.4	9.7	6.4	5.6	6.6	
10,000 and over	16,129,846	17,022,542	18,285,480	19,215,107	19,258,779	21,067,214	21,140,156	22,449,855	22,654,692	23,839,086	24,084,698	25,631,557	70	70	75	76	76	77	77	78	79	79	80	81	5.5	5.1	9.4	6.2	5.2	6.4

Census Metropolitan Areas (CMAs) have 100,000 or more in the urban core and includes all neighbouring towns and municipalities where 50 percent or more of the workforce commutes to the urban core.
Census Agglomerations (CAs) have 10,000 to 99,999 in the urban core and includes all neighbouring towns and municipalities where 50 percent or more of the workforce commutes to the urban core.
Metropolitan Influenced Zones (MIZ) are assigned on the basis of the share of the workforce that commutes to any CMA or CA (Strong MIZ: 30 to 49 percent; Moderate MIZ: 5 to 29 percent; Weak MIZ: 1 to 5 percent; No MIZ: no commuters).
The data for the 1991 and 1996 MIZ have been adjusted to be consistent with the 2001 protocol whereby non-CMA/CA towns and municipalities in the Territories were not allocated to a MIZ classification.
The designation of MIZ for 1991 and 1996 were obtained from Sheila Rambeaux and Kathleen Todd. (2000) **Census Metropolitan Area and Census Agglomeration Influenced Zones (MIZ) with census data** (Ottawa: Statistics Canada, Geography Working Paper Series No. 2000-1, Catalogue No. 92F0138MIE) (www.statcan.ca/cgi-bin/downpub/ftp.cgi?catno=92F0138MIE). Note that the Rambeaux and Todd designation of MIZ for 1991 used the preliminary 1996 CMA/CA delineations, but still using 1991 boundaries. For this table, we have re-imposed the 1991 CMA/CA delineation and we have assigned "strong MIZ" in 1991 for towns or municipalities that had been coded into a CMA/CA for 1996.
The designation of MIZ for 2001 was obtained from Statistics Canada, **GeoSuite, 2001 Census** (Ottawa: Statistics Canada, Catalogue No. 92F0085XCB).
Source: Statistics Canada, Census of Population, 1981 to 2006.

5. Discussion: So how many Canadians live in a city?

The answer to this question is clearly “It depends!”

If you define a city as having 100,000 or more inhabitants, the answer is (in 2006):

- 59% in built-up areas (population centres) of 100,000 or more inhabitants ;
- 52% in administrative units (census subdivisions) of 100,000 or more inhabitants; and
- 68% in functional areas (CMAs) of 100,000 or more inhabitants (Table 5.1-1).

The answer also depends upon to whom you are talking:

- 59% if you are talking to public transit planners;
- 52% if you are talking to mayors; and
- 68% if you are talking to economic development analysts.

For a given size class, we can rank our definitions of cities in terms of which definition provides the highest level of urbanization. The ranking is (a) functional areas (CMAs & CAs), (b) built-up areas (population centres), and (c) administrative units (census subdivisions).

Similarly, for a given size class, we can rank our definition of cities in terms of which definition produces the highest rate of urbanization. The highest is (a) administrative areas (census subdivisions)—due to amalgamations; and (b) built-up areas (population centres) and functional areas (CMAs and CAs).

Table 5.1-1
How many Canadians live in a city?

.. For alternative population size classes to be a "city"							
... For alternative ways of defining a "city"							
Alternative population size classes to be a "city"	Alternative ways of defining a "city"	1981	1986	1991	1996	2001	2006
		percent of total population					
1 million and over	Population centres ¹	29	30	31	32
	Census subdivisions ²	4	4	12	13
	CMAs and CAs ³	29	31	32	36	37	45
500,000 and over	Population centres ¹	43	44	46	46
	Census subdivisions ²	17	20	25	29
	CMAs and CAs ³	44	47	48	49	50	51
100,000 and over	Population centres ¹	54	55	58	59
	Census subdivisions ²	38	41	46	52
	CMAs and CAs ³	56	60	61	62	64	68
50,000 and over	Population centres ¹	59	61	63	64
	Census subdivisions ²	51	53	58	62
	CMAs and CAs ³	62	68	69	71	72	74
30,000 and over	Population centres ¹	63	64	67	68
	Census subdivisions ²	58	60	64	67
	CMAs and CAs ³	67	72	74	75	76	78
10,000 and over	Population centres ¹	68	70	72	73
	Census subdivisions ²	73	75	78	80
	CMAs and CAs ³	70	76	77	78	79	81

1. Population centres: form or morphology or built-up area (any locality with a population density of 400 inhabitants per km² or more – delineated for localities with a total population of 1,000 more).

2. Census subdivision: administrative unit (an incorporated city, town or municipality).

3. CMA or CA: functional labour market unit (Census Metropolitan Areas and Census Agglomerations -- with an urban core of 10,000 or more and includes all neighbouring census subdivisions where 50% or more of the employed residents commute to the urban core).

6. Summary

Using three alternative ways of showing density (or population size), we see the following:

1. The answer to the question “How many Canadians live in a city?” depends upon to whom one is speaking and the size cut-off of what comprises a ‘city’ for the topic under discussion.
2. Demographic growth drives the rate of urbanization for built-up areas (population centres) and for functional areas (CMAs and CAs).
3. Reclassification (including amalgamations) drives most of the urbanization for administrative areas (census subdivisions).

References

- Hofmann, N. *et al.* (2010a), “*Introducing a New Concept and Methodology for Delineating Settlement Boundaries: A Research Project on Canadian Settlements*”, Ottawa: Statistics Canada, Environmental Accounts and Statistics Analytical Technical Paper No. 11, Catalogue no. 16-001, <http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=16-001-M>.
- Hofmann, N. *et al.* (2010b), “A new research project on Canadian settlements: initial geographic results”, *EnviroStats* Vol. 4, No. 1, Ottawa: Statistics Canada, Catalogue no. 16-002, <http://www.statcan.gc.ca/bsolc/olc-cel/olc-cel?catno=16-002-X&chroprg=1&lang=eng>.
- Mendelson, R. and J. Lefebvre (2003), “*Reviewing Census Metropolitan Areas (CMA) and Census Agglomerations (CA) in Canada According to Metropolitan Functionality*” Ottawa: Statistics Canada, Geography Working Paper Series No. 2003-001, Catalogue no. 92F0138MIE, www.statcan.gc.ca/cgi-bin/downpub/listpub.cgi?catno=92F0138MIE.
- Puderer, H.A. (2009), *Urban Perspectives and Measurement*, Ottawa: Statistics Canada, Geography Working Paper, Catalogue no. 92F0138M — No. 2009001, <http://www.statcan.gc.ca/pub/92f0138m/92f0138m2009001-eng.htm>.
- Statistics Canada (2007), *2006 Census Dictionary*, Ottawa: Statistics Canada, Catalogue no. 92-566, <http://www12.statcan.gc.ca/english/census06/reference/dictionary/index.cfm>.

The role of data quality standards in the standardization of survey methods and tools

John L. Eltinge¹

Abstract

This paper explores the interface between data quality standards and the standardization of survey methods and tools. First, this paper considers statistical survey methodology as a form of technology, and uses the resulting conceptual framework to explore several ways in which to evaluate the prospective costs and benefits of quality standards, and standardization of methods, for statistical programs. The conceptual framework places primary emphasis on types of standards; methods for calibration; methods for application and enforcement; and special issues for government-sponsored statistical programs. This framework leads to discussion of prospective benefits of standards, including improved data quality; reduction of quantifiable costs for stakeholders; and reduction of risks for stakeholders. In parallel with benefits, the paper also reviews potential costs and risks associated with survey standards, including direct costs, indirect costs and inefficient allocation of resources.

Second, this paper explores the impact of standardization of methods and tools on the survey process. The impact potentially includes changes (both positive and negative) in: (a) the fixed and variable components of the overall survey cost structure; (b) the robustness of the survey to changes in stakeholder needs, resource availability, or the external operating environment; and (c) the extent to which the resulting survey may meet certain data quality standards.

The paper closes with comments on practical implications of these general ideas for development, implementation and enforcement of standards; development, implementation and maintenance of standardized tools and methods; and communication with external stakeholders.

¹John L. Eltinge, Bureau of Labor Statistics, USA.

Enterprise architecture work at Statistics Sweden

Martin Axelson, Jakob Engdahl, Ylva Fossan, Eva Holm, Ingegerd Jansson, Boris Lorenc
and Lars Göran Lundell¹

Abstract

The paper presents current work on enterprise architecture and its components at Statistics Sweden. It outlines architectural foundations, namely drivers, standards, architecture frameworks, and a vision for the production of statistics. Modelling is presented as the main contribution of business architects, in general and through two main areas of application. These are the Triton platform for design, data collection and post-processing, and an elaborated vision of a data warehousing strategy and platform for integrated statistics production from administrative data and data obtained through primary data collection. Modelling outside the business process level is also mentioned. The paper ends with some general remarks.

Key Words: Enterprise architecture; business architecture; TOGAF; information modelling; process modelling.

*Would you tell me, please, which way I ought to go from here?’
‘That depends a good deal on where you want to get to,’ said the Cat.
‘I don’t much care where—’ said Alice.
‘Then it doesn’t matter which way you go,’ said the Cat.
‘—so long as I get SOMEWHERE,’ Alice added as an explanation.
‘Oh, you’re sure to do that,’ said the Cat, ‘if you only walk long enough.’
Lewis Carroll: “Alice’s Adventures in Wonderland”*

1. Foundations

1.1 Drivers

The work of Statistics Sweden on enterprise architecture and its components is driven by the goals of

- increasing efficiency of the production process and reducing costs;
- increasing quality of the processes and the statistics produced;
- reducing administrative burden of data providers caused by primary data collection;
- reacting faster to changes in user needs and the environment (including the IT environment);
- increasing transparency and simplifying governance of the change management process; and
- improving long-term planning and the overview and management of development work.

Internally, we view these drivers, shared with many other national statistical institutes (NSIs), as components of a simpler equation that includes only quality and cost. We see as the goal of a producer of statistics to produce statistics of as high quality as possible given a certain fixed cost; or, alternatively, to minimise the cost of producing statistics of a certain fixed quality. Standardisation (*e.g.*, Hofman *et al.* 2011; Lopdell and Dunnet 2011) is a means for reaching this goal, but it is not a goal in itself; it is the goal only as long as it leads to an overall improvement (under some evaluation) of the joint cost/quality function.

1.2 Standards

In working on enterprise architecture and its components, Statistics Sweden is taking into consideration standards and common approaches to regulating some of the main issues related to production of statistics, as well as taking part in developing some of them. These include: GSBPM (METIS 2009), GSIM (OCMIMF 2011), SDMX (SDMX 2011), DDI (Vardigan *et al.* 2008), and CORE (Scannapieco 2011).

¹Martin Axelson, Jakob Engdahl, Ylva Fossan, Eva Holm, Ingegerd Jansson, Boris Lorenc, and Lars Göran Lundell, Statistics Sweden, 701 89 Örebro, Sweden.

Importance of the existence of these standards and of reliance on them lies in improving communication between the different NSIs, reducing duplication of work, and finding a potential for cost reduction by division of the work. Additionally, comparability between countries (*e.g.*, in the European Union), or across different producers of statistics within a decentralised national statistical environment (*e.g.*, in the United States), requires high similarity of processes and definitions at least at the output level, so as to enable comparability of the produced statistics.

1.3 Enterprise architecture framework

In order to structure Statistics Sweden's work on EA, we rely on Zachman's (1987) framework's notion of levels of architecture, consisting of: I. Enterprise architecture (EA), II. Business architecture (BA), III. Solutions architecture, IV. Applications architecture, V. IT infrastructure architecture including hardware. While the majority of our practical work takes place on the BA level or below, it is our understanding that this does not suffice for achieving the goals in Section 1.1. A broader perspective is required, involving costs (budgetary considerations), personnel competence, a quality framework, *etc.* Therefore, Statistics Sweden considers an enterprise architecture approach necessary to be able to—in a long-term perspective—achieve a sensible statistics production system. By 'enterprise architecture' we mean "a coherent whole of principles, methods and models that are used in the design and realisation of an enterprise's organisational structure, business processes, information systems, and infrastructure" (Lankhorst *et al.* 2009).

In particular, we are exploring TOGAF (The Open Group 2009) as a framework able to support the holistic approach outlined in the preceding paragraph. An attractive aspect of TOGAF is that it incorporates a method for developing architecture components, called Architecture Development Method.

1.4 Vision

We acknowledge that it is neither trivial nor perhaps wise to specify in too much detail the exact properties that a system for statistics production will have in say ten years. However, working towards a vision has clear advantages. Taking as an example the goal of better integration of administrative and survey data in statistics production, it is likely not the case that this goal will be achieved within a foreseeable future by simply approving development proposals as they come. This goal needs to prominently exist and steps towards its realisation be actively taken so that it has a reasonable chance to be accomplished.

In particular, the European Commission's communication "On the production method of EU statistics" (EC 2009), known as Vision 2020, is noteworthy. Some components of the Vision apply only to the supra-national level, but it in general calls for a transition from a stovepipe model of statistics production to systems for statistics production, increased use of administrative data in the production, and so on, and thus likely is inspirational to any producer of official statistics.

While Statistics Sweden is still working on its long-term vision, some components of it are implied in the two platforms presented in Section 4. Some further components of it also relate to achieving the goals of Section 1.1. One, for instance, involves providing services that are designed so as to put only light requirements on human IT resources. This and the other goals highlight the need to put in place quantitative measures of progress towards achieving components of the vision, including tracking of costs (more precisely, of return on investment).

2. Organisational setup

An *architecture group* was established in 2008 within the R&D department of Statistics Sweden. The group consists of about ten senior methodologists and senior IT experts, dividing their time between architecture work and other duties. The immediate environment of the architecture group, within the R&D department, consists of the *quality group* and the *project leader group*.

With the goal of standardising the processes used for statistics production, a Process department was established also in 2008, collecting most of Statistics Sweden's IT and methodology competence. Also, a new group within Process department was established: process owners for main phases of the GSBPM.

In 2011, an evaluation of the changes made in 2008 has led to a reorganisation of the Process department. A new IT department was created by taking the IT units out of it, with the aim of improving governance of the IT sector. As a consequence, two architectural forums are about to be put in place to coordinate architectural activities between the IT, Process, and R&D Departments.

A Project Management Group (PMG), chaired by the Deputy Director General in the capacity of the executive with responsibility for R&D, exists since 2008 assuring that development efforts are prioritised from a holistic view point and that local initiatives do not depart from the overall goals. A recent evaluation of PMG has led to a proposal for a *development projects portfolio management* resource.

3. Implementation of EA at Statistics Sweden

Statistics Sweden's business architects use four basic components in structuring the work on BA and EA: business goals, processes, information and applications. Business goals span a space in which both business processes and information for carrying them out are realised. A relation of mutual compatibility needs to exist between processes and information. An exact specification of the goals, processes, and information provides the context for application solutions. A consequence is that business rules, process flows, and information structures become represented in the models in a systematic and consistent way.

3.1 Process modelling

Formal description of the business process model at Statistics Sweden largely coincides with that contained in the GSBPM. However, a complete EA will provide corresponding models also for processes supporting the business, for instance human resources management and legal framework management.

Process modelling is carried out at a level that enables identification of common business subprocesses. The goal of the modelling is to produce a representation of the process flows. Currently, we have models for the following subprocesses in place:

- Within the Design phase: i) Choose data source(s) and data collection method(s), ii) Choose contact strategy and identify relevant population groups, iii) Decide on level of editing during data collection and choose data input method, iv) Design the production flow, v) Verify administrative routines, vi) Plan and book resources;
- Within the Collect phase: i) Update sample, ii) Prepare questionnaire distribution, iii) Prepare scanning, iv) Prepare web data collection, v) Provide support to data provider, vi) Manage double submissions, vii) Manage reminders, viii) Scan and check;
- Within the Process phase (Edit subprocess): i) Error alert, ii) Automatic edit, iii) Manual check

3.2 Information modelling

Information modelling helps formalise the description of the information used in the enterprise, with the aim of achieving compatibility with the IT infrastructure that is processing the information. The models are conceptual, formal, and independent of and able to be unequivocally interpreted by, the lower layers of application and infrastructure. We distinguish between two levels of models: object group models and (detailed) object models.

Object group models may be used for assigning ownership of information. Object models are the basis for a common concept repository at the enterprise level. The models are also part of—and help specify—the general requirements space. For instance, no business goals are allowed to be implemented in physical data models without first being put into object models. By doing so, interpretation of the business goals is formalised and recorded, rather than left to arbitrary (and undocumented) interpretation of the personnel implementing the applications.

Among the groups of information models developed thus far at Statistics Sweden is, for instance *Statistical survey* (with object models *survey round*, *sample object*, *reporting object*, *sending material*, *data item*, *comment*, etc); outside the GSBPM, we have object group models for e.g., *personnel* and *organisational structure*. Among architectural principles there is one stating that specified types of information shall not be created and/or stored in more than one place, such as *statistical survey*, *edit rule*, and *status code*. This, however, needs to be weighed against other architectural concerns, like the risk of creating too strong system dependencies (and thereby an unstable system).

3.3 Concept modelling

In order for a system to be shared by many, the concepts used need to be well defined. Concepts are implicitly present in process models and information models; however, crucial concepts may need to be formalised into well-defined concept models. We have done such models for Editing and Disclosure control subprocesses of the GSBPM.

An invaluable result of concept modelling and information modelling performed at Statistics Sweden thus far consisted in making explicit the different ways concepts and information were interpreted in the organisation; this in turn has led, through discussions, to adoption of unified definitions of the terms used.

We plan to investigate whether formulation of the models needs to be dependent on the architectural level (of Section 1.3) on which they are to be applied.

3.4 Architecture capability

Regarding the need for continued work on EA and structuring the different architectural groups involved, Statistics Sweden is implementing the following architecture-related roles: i) business architect, ii) solution architect, and iii) software architect. Additional competencies include requirements analysts, test personnel, infrastructure specialists and specialists for specific technical platforms, among others.

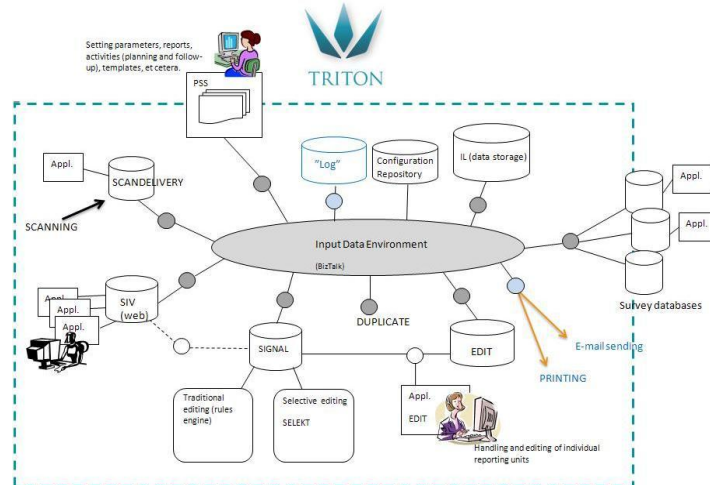
4. Examples of EA work

4.1 General data collection platform

Statistics Sweden is in the process of developing a general platform for data collection and editing (Engdahl 2010; Ericson 2011). It supports the Build and Collect phases of GSBPM, as well as subprocesses 2.3 and 2.6 of the Design phase and 5.1 - 5.4 of the Process phase. In contrast to traditional solutions, which focus on data storage, the platform is built to support business processes with necessary data and metadata (Figure 4.1-1). An event-driven approach is used, where data and metadata are transferred between processes using business objects (as parts of a Business Information Model). In practice, a value chain—as a mapping between the process models and the information models—is created for each business object type, taking care of where the business object is created and what processes/services should be used to add value to the object. A communication platform supports the information flow.

A component related to Triton is the Production Support System. It originated as an HTML-linked set of guidelines for performing common business processes. However, it is successively being transformed into an interactive production environment. In this environment the common methods, tools, and approaches will be available as services through common interfaces where parameters set for each survey will determine which functionality is applied (Bergdahl and Blomqvist 2011).

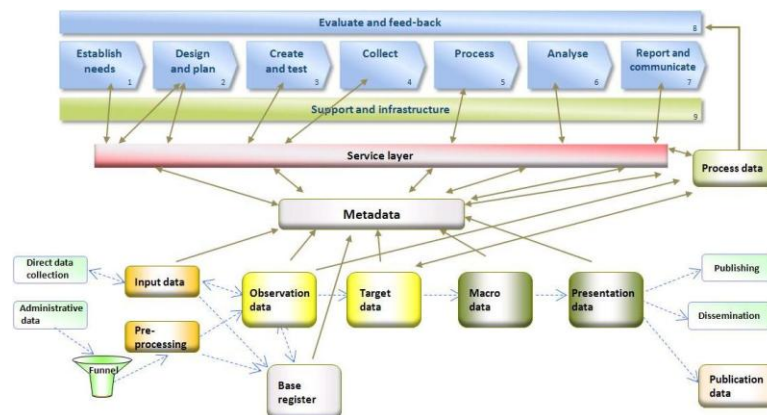
Figure 4.1-1
Triton, a platform for survey design, data collection and post-processing



4.2 Data warehousing strategy for integrated statistics production

Statistics Sweden is currently in the initial stages of designing a comprehensive statistical platform using an approach involving a consistent, structured and well documented data warehouse (Figure 4.2-1). It aims to have registers as its centrepiece, and its design is guided by the principles of minimised data transport and no storage of duplicate or near duplicate data, among others. It is designed so as to provide support for process oriented production methods, efficient data collection, and flexible dissemination.

Figure 4.2-1
A complete conceptual view of the envisioned system based on the data warehouse approach



The work on the platform is expected to shed new insights into the relation between GSBPM and a data warehouse approach, as well as to highlight the role of metadata in modern production of statistics. Another task to be included in this work concerns establishing more precisely capabilities of a service layer that can serve the function in both Triton and the data warehouse system.

4.3 Further architectural work

We are currently modelling the Specify Needs phase and Disseminate phase of the GSBPM. In the course of redesign of the Business Register, we are modelling data collection from administrative sources. Further and outside

the GSBPM, we have made process models for the payroll process and for the enterprise-wide processes of planning and monitoring. Together with three other agencies, we have worked on common models of concepts and information, in order to facilitate sharing of information that an agency collects.

5. Some concluding remarks

We encourage NSIs to simply start by applying any EA framework that they may find appropriate for their needs: Federal Enterprise Architecture Framework (or another government EA), Gartner's, TOGAF, or something else. The way to work with EA is not as a paper product but rather by directly involving it in projects and relevant areas of development.

We strongly believe that NSIs should use existing architectural frameworks rather than invent their own, as the former have in general been developed with a far larger EA competence than what an NSI usually has at its disposal. Further, we propose that open frameworks (*e.g.*, TOGAF) be preferred to proprietary ones, and accepted standards for statistics production be used (*e.g.*, GSBPM), as this will enable easier integration of production systems across NSIs.

Acknowledgements

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of Statistics Sweden. The authors are grateful to Anders Holmberg and Hans Irebäck, both at Statistics Sweden, for constructive comments on an earlier draft of the paper.

References

- Bergdahl, M., and K. Blomqvist (2011), "National Implementation of the GSBPM: The Swedish Experience". Workshop on Statistical Metadata. (Geneva, Switzerland, 5-7 October 2011).
- EC (2009), "On the production method of EU statistics: a vision for the next decade". COM(2009) 404. <http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2009:0404:FIN:EN:PDF>.
- Engdahl, J. (2010), "An Event-Driven Architecture for Data Collection", presented at MSIS 2010 meeting, Daejeon, Republic of Korea, 26-29 April 2010.
- Ericson, J. (2011), "Triton: a general tool for data collection and micro editing", presented at Statistics Canada's 2011 International Methodology Symposium, Ottawa, Canada.
- Hofman, F., Renssen, R. and A. Camstra (2011), "Standardisation of Processes", presented at Statistics Canada's 2011 International Methodology Symposium, Ottawa, Canada.
- Lopdell, J. and G. Dunnet (2011), "Statistics New Zealand's Standard Methodology Toolbox", presented at Statistics Canada's 2011 International Methodology Symposium, Ottawa, Canada.
- Lankhorst, M. *et al.* (2009), *Enterprise Architecture at Work: Modelling, Communication and Analysis*, Berlin, DE: Springer-Verlag.
- METIS (2009), Generic Statistical Business Process Model: Version 4.0: April 2009, UNECE Secretariat, <http://www1.unece.org/stat/platform/download/attachments/8683538/GSBPM+Final.pdf?version=1>.
- OCMIMF (2011), "Generic Statistical Information Model (GSIM), Common Reference Model, Version 0.1: June 2011", http://www1.unece.org/stat/platform/download/attachments/62751291/GSIM+Common+Reference+Model+V0_1.docx?version=1

Scannapieco, M. (2011), “ESSnet CORE Intermediary Report”, available online: http://www.essnet-portal.eu/sites/default/files/79/Dell1.2-intermediary_report_v4.doc.

SDMX (2011), “Framework For SDMX Technical Standards, Version 2.1”, http://sdmx.org/wp-content/uploads/2011/04/SDMX_2-1_SECTION_1_Framework.pdf.

The Open Group (2009), *The Open Group Architecture Framework (TOGAF), Version 9*, The Open Group: San Francisco, CA.

Vardigan, M., Heus, P. and W. Thomas (2008), “Data Documentation Initiative: Toward a Standard for the Social Sciences”, *The International Journal of Digital Curation* 3, 1.

Zachman, J.A. (1987), “A Framework for Information Systems Architecture”, *IBM Systems Journal*, Volume 26, Number 3.

SESSION 10B

CALENDAR EFFECTS AND TEMPORAL COHERENCE

Benchmarking and forecasting: A top-down approach for combining forecasts at multiple frequencies

Michele A. Trovero, Ed Blair and Michael J. Leonard¹

Abstract

Forecasters often deal with data accumulated at different time intervals (for example, monthly data and daily data). A common practice is to generate the forecasts at the two time intervals independently so as to choose the best model for each series. That practice can result in forecasts that do not agree.

This paper shows how the SAS® High-Performance Forecasting HPFTEMPRECON procedure uses the lower frequency forecast as a benchmark to adjust the higher-frequency forecast to take the best advantage of both forecasts.

Key Words: Forecasting; Benchmarking; Multiple Frequencies; SAS/HPF; PROC HPFTEMPRECON.

1. Introduction

Forecasters often need to produce forecasts for a certain time series at more than one frequency. For example, a company that provides warranty repairs for appliances might want to forecast the number of daily calls for staffing and operational planning, such as ordering supplies. The company might also want to forecast service calls at a monthly frequency to plan long-term expansion and to plan for financial concerns such as the purchase of more vehicles or the hiring of new staff. This paper deals with the problem of forecasting one time series at different frequencies, with a focus on stock variables. For a stock variable, the low-frequency series is the temporal aggregation of the high-frequency series. The term ‘accumulation’ indicates temporal aggregation, and thus distinguishes it from other forms of aggregation, such as the aggregation of series within a subclass that can take place in a hierarchical forecasting context. The problem of forecasting at multiple frequencies is easily solved in an ideal world where data are plentiful, series are well behaved (meaning they have mostly nonzero values and are easily transformed to a covariance stationary series), and the correct model is chosen for each series. In this case, the accumulation of the high-frequency forecasts is at least as efficient as the forecasts generated by modeling the low-frequency series, in the sense that the mean squared error of the h-step-ahead prediction of the former is less than or equal to the mean squared error of the h-step-ahead prediction of the latter. A formal outline of this argument for seasonal ARIMA processes can be found in Wei (1990, Chapter 16). The idea is simple: a forecast (prediction) is the linear projection onto the Hilbert space generated by the observed series. The space spanned by the low-frequency data is a subset of the space spanned by the high-frequency data. Therefore, the accumulation of the projection on the finer space generated by the high-frequency data is at least as close to the actual future value as the projection on the coarser space spanned by the low-frequency data. Another way to express the same concept that is simpler and does not require any mathematical jargon is that the accumulation process is a form of compression that involves loss of information. The original high-frequency data cannot be regenerated using only the accumulated data. Therefore, forecasts generated with the restricted information contained in the accumulated data cannot be better than forecasts generated with full information of the non-accumulated data. Reality, however, rarely comes in textbook format. Consider the following real-life examples (the name of the companies are retained for confidentiality reasons):

¹Michele A. Trovero, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Michele.Trovero@sas.com); Ed Blair, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Ed.Blair@sas.com); Michael J. Leonard, SAS Institute Inc, 100 SAS Campus Drive, Cary, NC, USA, 27513 (Michael.Leonard@sas.com).

Example 1. The spare-parts branch of a large company operates nationwide and manages more than 40,000 spare parts. Three-months-ahead daily forecasts are needed for each ZIP code for replenishing the repair trucks and for making staffing decisions. Very few parts are needed with regularity. Approximately only 10% of the parts show a somewhat regular demand for each ZIP code. For the remaining parts, the daily demand is almost always zero. Long-term monthly forecasts are needed for part production, hiring purposes, and long-term investments.

Example 2. A large retail store chain collects POS (point-of-sale) data in each store. Hourly forecasts are needed in the medium term for staffing purposes. The hourly data are kept for three months, after which they are discarded due to the cost of storing such a large amount of data. Only data accumulated at daily intervals are kept. Long-term monthly forecasts are needed for expansion and financial planning.

In both examples, forecasts are needed at different frequencies for different purposes. However, there are good reasons to believe that the accumulation of the high-frequency forecasts will not lead to good forecasts for the low-frequency data. In the first example, most series show intermittent behavior. Intermittent series consist mostly of a single value, usually zero. Models for intermittent data, such as the popular Croston (1972) model, cannot capture important features such as trend, seasonality, and dependency on events or other external variables. Additionally, multiple seasonal components might be present in the high-frequency data, whether they are intermittent or not. Modeling and estimating multiple seasonal components simultaneously can be complex and computationally intensive.

In the second example, the duration of the hourly (high-frequency) data is not sufficient to produce monthly (low-frequency) forecasts of any value. Indeed, you can reasonably argue that the information contained in the longer history of the daily data can be used with benefit to forecast the hourly data. For example, when making staffing decisions about the very important winter holiday season, the retailer should use the information contained in the daily data, which covers the previous holiday seasons, and not rely solely on the hourly data forecasts which are based only on the previous three months. In practice, the forecasts for the two or more frequencies of interest are often derived independently from each other by selecting at each frequency a model that provides the best results according to criteria, such as minimizing the MAPE (mean absolute percentage error). However, when the forecasts are derived independently, the accumulation of the high-frequency forecasts is generally different from the forecasts generated by the model for the low-frequency data.

Additionally, as in Example 2, you might want to use the low-frequency forecasts to improve the high-frequency forecasts. This paper shows a method for revising the high-frequency forecasts such that their accumulation at the low frequency is equal to the forecasts generated by the model selected for the low-frequency data. The first section details the method. The second section introduces the HPFTEMPRECON procedure in SAS® High Performance Forecasting and shows how it can reconcile monthly forecasts to daily forecasts for the Box and Jenkins' airline data. The third section presents the results of applying the method to a data set that consists of several time series that exhibit intermittent behavior. Finally, the last section draws the conclusions.

2. Method

The combination of a series of high-frequency data with a series of more reliable but less frequent data is seen often in business statistics. For example, surveys are conducted at quarterly intervals on a subsample of the population of interest to determine the inter-annual variations, while comprehensive surveys on the whole population are conducted only on a yearly basis. The process of adjusting the more frequent data to match the less frequent but more reliable data is known in the literature as benchmarking. Denton (1971) provided the first general framework for benchmarking based on minimizing a quadratic function. A recent and comprehensive review on the topic can be found in Dagum and Cholette (2006). The lower-frequency forecasts are also referred to as the benchmark forecasts. The higher-frequency forecasts are also referred to as the indicator forecasts. Benchmarking procedures can be applied more generally to any two series that are measured at different time intervals. Therefore, this paper more generally refers to the benchmark series and indicator series to indicate the forecasts involved in the benchmarking. Denote the indicator series by x_t with $t = 1, \dots, T$, where t is associated with a date. Denote the benchmark series by a_m , $m = 1, \dots, M$. The benchmarks have a starting date $t_{1,m}$ and ending date $t_{2,m}$, such that $1 \leq t_{1,m} < t_{2,m} \leq T$.

You want to find an optimal benchmarked series $\theta_t, t = 1, \dots, T$ such that the accumulation of benchmarked series at the frequency of the lower-frequency forecasts is equal to the benchmark series. That is,

$$\sum_{t=\ell_{1,m}}^{t_{2,m}} \theta_t = a_m.$$

For $m = 1, \dots, M$.

The bias is defined as the expected discrepancy between the benchmark and the indicator series. You can decide whether to adjust the original indicator series to account for the bias. Denote the bias-adjusted indicator series by s_t . When no adjustment for bias is performed, $s_t = x_t$. The additive bias correction is given by:

$$b = \frac{\sum_{m=1}^M a_m - \sum_{m=1}^M \sum_{t_{1,m}}^{t_{2,m}} x_t}{\sum_{m=1}^M \sum_{t_{1,m}}^{t_{2,m}} 1}.$$

In this case, the bias-adjusted indicator is $s_t = b + x_t$.

The multiplicative bias correction is given by:

$$b = \frac{\sum_{m=1}^M a_m}{\sum_{m=1}^M \sum_{t_{1,m}}^{t_{2,m}} x_t}.$$

In this case, the bias adjusted-series is $s_t = bx_t$. Note that the multiplicative bias is not defined when the denominator is zero.

Let $\mathbf{s} = [s_1, \dots, s_T]'$ be the vector of the bias-corrected indicator series, and let $\boldsymbol{\theta} = [\theta_1, \dots, \theta_T]'$ be the vector of its reconciled values. Let \mathbf{D} be the $T \times T$ diagonal matrix whose main-diagonal elements are $d_{t,t} = |s_t|^\lambda, t = 1, \dots, T$. Indicate by \mathbf{V} the tridiagonal symmetric matrix whose main-diagonal elements are $v_{1,1} = v_{T,T} = 1$ and $v_{t,t} = 1 + \rho^2, t = 2, \dots, T - 1$, and whose sub- and super-diagonal elements are $v_{t,t+1} = v_{t+1,t} = -\rho, t = 1, \dots, T - 1$. Define $\mathbf{Q} := \mathbf{D}^+ \mathbf{V} \mathbf{D}^+$ and $\mathbf{c} := -\mathbf{Q} \mathbf{s}$, where \mathbf{D}^+ indicates the Moore-Penrose pseudo-inverse of \mathbf{D} . The benchmarked (reconciled) series is given by the values $\theta_t, t = 1, \dots, T$, that minimize the quadratic function

$$f(\boldsymbol{\theta}; \lambda, \rho) = \frac{1}{2} \boldsymbol{\theta}' \mathbf{Q} \boldsymbol{\theta} + \mathbf{c}' \boldsymbol{\theta}$$

under the constraints

$$\sum_{t=\ell_{1,m}}^{t_{2,m}} \theta_t = a_m, \quad m = 1, \dots, M$$

where $0 \leq \rho \leq 1$ and $\lambda \in \mathbb{R}$ are parameters that you select. When \mathbf{s} does not contain zeros, the target function is equivalent to the one proposed by Quenneville *et al.* (2006).

Two issues are considered when benchmarking. The first one is to preserve the movement in the high-frequency series as much as possible (movement preservation). The second is to account for the timeliness of the benchmarks, in the sense that the benchmark for the last period might not be available if the indicator series extends beyond the last benchmark value. Bias correction is a way to improve the timeliness of the benchmark in that it attempts to reduce the expected discrepancies between the benchmark and the indicator function. The parameter ρ is a smoothing parameter that controls the movement preservation. The closer ρ is to one, the more the original series movement is preserved. The parameter λ usually takes values 0, 0.5, or 1. For $\lambda = 0$, you have an additive benchmarking model. For $\lambda = 0.5$ and $\rho = 0$, you have a prorating benchmarking model.

In the traditional application of benchmarking, the goal is to regain the additivity of some seasonal adjusted series with respect to the benchmark. In the context of this paper, the goal is to find the optimal forecasts for the high-frequency series that respect the accumulation constraint. Therefore, it is suggested that you select the bias correction and values of the parameters ρ and λ in such a way as to optimize the selection criteria that was originally used to select the models for the high-frequency data. For example, if the model for the high-frequency data was selected by

minimizing MAPE, likewise the parameters ρ , λ , and the bias correction should be chosen to minimize MAPE for the benchmarked forecasts.

When $0 \leq \rho < 1$, the constrained minimization problem can be derived from the constrained regression problem

$$\begin{aligned} s_t &= \theta_t + c_t e_t & t &= 1, \dots, T \\ e_t &= \rho e_{t-1} + \epsilon_t & t &= 1, \dots, T \\ \sum_{t=t_{1,m}}^{t_{2,m}} \theta_t &= a_m, & m &= 1, \dots, M \end{aligned}$$

where ϵ_t is a white-noise process with variance σ_ϵ^2 , and c_t are weights proportional to $|s_t|^\lambda$. Therefore, when $\lambda = 0$, the minimization problem is equivalent to a constrained regression problem where the error between the bias-adjusted indicator and the benchmarked series follows an AR(1) process with an autoregressive parameter proportional to ρ .

Let $\mathbf{a} = [a_1, a_2, \dots, a_M]'$. The constraint equation can be rewritten as

$$\mathbf{J}\boldsymbol{\theta} = \mathbf{a}$$

where \mathbf{J} is a matrix of zeros and ones such that $\mathbf{J}\boldsymbol{\theta}$ is the accumulation of the benchmarked series at the frequency of the benchmark. The solution of the minimization problem then becomes

$$\hat{\boldsymbol{\theta}} = \mathbf{s} + \mathbf{C}\boldsymbol{\Sigma}_e \mathbf{C}' (\mathbf{J}\mathbf{C}\boldsymbol{\Sigma}_e \mathbf{C}')^{-1} (\mathbf{a} - \mathbf{J}\mathbf{s})$$

where \mathbf{C} is a diagonal matrix whose main-diagonal elements are c_t , and $\boldsymbol{\Sigma}_e$ is the covariance matrix of e_t . When benchmarking can be interpreted as a regression problem, it is also possible to derive the covariance of the reconciled forecasts. See Quenneville *et al.* (2006) for the details.

A further interpretation of this method is as a way to combine the forecasts at the two frequencies to produce forecasts for the higher frequency. The weights for the combination are derived using the solution of the minimization problem. The lower-frequency forecasts are assigned unit weights since they provide the right-hand side of the constraint equations.

3. The HPFTEMPRECON Procedure

Using the method outlined in the preceding section, the HPFTEMPRECON procedure reconciles high-frequency forecasts to low-frequency forecasts in such a way that the accumulation of the reconciled high-frequency forecasts is equal to the low-frequency forecasts. PROC HPFTEMPRECON reconciles forecasts for the same item at two different time frequencies whose intervals are nested in one another. In other words, it reconciles a two-level hierarchy of forecasts in the time dimension. For example, it reconciles monthly forecasts for the Box and Jenkins airline passenger data (in the Sashelp.Air data set) to the quarterly forecasts for the same series. For this reason, the HPFTEMPRECON procedure not only requires two input data sets for the predictions, but also it requires that the two frequencies of the forecasts be specified in two separate statements: the ID statement for the high-frequency data, and the BENCHID statement for the low-frequency data.

SAS High Performance Forecasting procedures are used to generate the forecasts at monthly and quarterly frequencies. These forecasts become the inputs to PROC HPFTEMPRECON. A full discussion about the SAS High Performance Forecasting system is outside the scope of this paper. Details can be found in SAS High-Performance Forecasting: User's Guide.

First, the HPFESMSPEC procedure generates an exponential smoothing model specification which is then selected by the HPFSELECT procedure:

```

proc hpfesmspec
  rep=work.repo
  specname=myesm;
esm;
run;

proc hpfselect
  rep=work.repo
  name=myselect;
spec myesm;
run;

```

Then, forecasts are generated with PROC HPFENGINE at the monthly and the quarterly frequencies using the selected model specification:

```

proc hpfengine
  data=Sashelp.Air
  rep=work.repo
  globalselection=myselect
  out=OutMon
  outfor=OutForMon
  outmodelinfo=OutMod;
id date interval=month;
forecast air;
run;

proc hpfengine
  data=Sashelp.Air
  rep=work.repo
  globalselection=myselect
  out=OutQtr
  outfor=OutForQtr
  outmodelinfo=OutModQtr;
id date interval=qtr accumulate=total;
forecast air;
run;

```

Note that the variable `air` appears in the FORECAST statement of both PROC HPFENGINE instances. The INTERVAL= option in the ID statements are different. In the first instance, the time ID interval is month; in the second instance, it is quarter. The monthly forecasts are stored in the PREDICT variable of the OutForMon data set, and the quarterly forecasts are stored in the PREDICT variable of the OutForQtr data set.

Finally, the monthly forecasts are reconciled to the quarterly forecasts using PROC HPFTEMPRECON:

```

proc hpftemprecon
  data=OutForMon
  benchdata=OutForQtr
  outfor=BenFor
  outstat=BenStat
  exp=0.5
  smooth=0.5;
id date interval=month;
benchid date interval=qtr;
run;

```

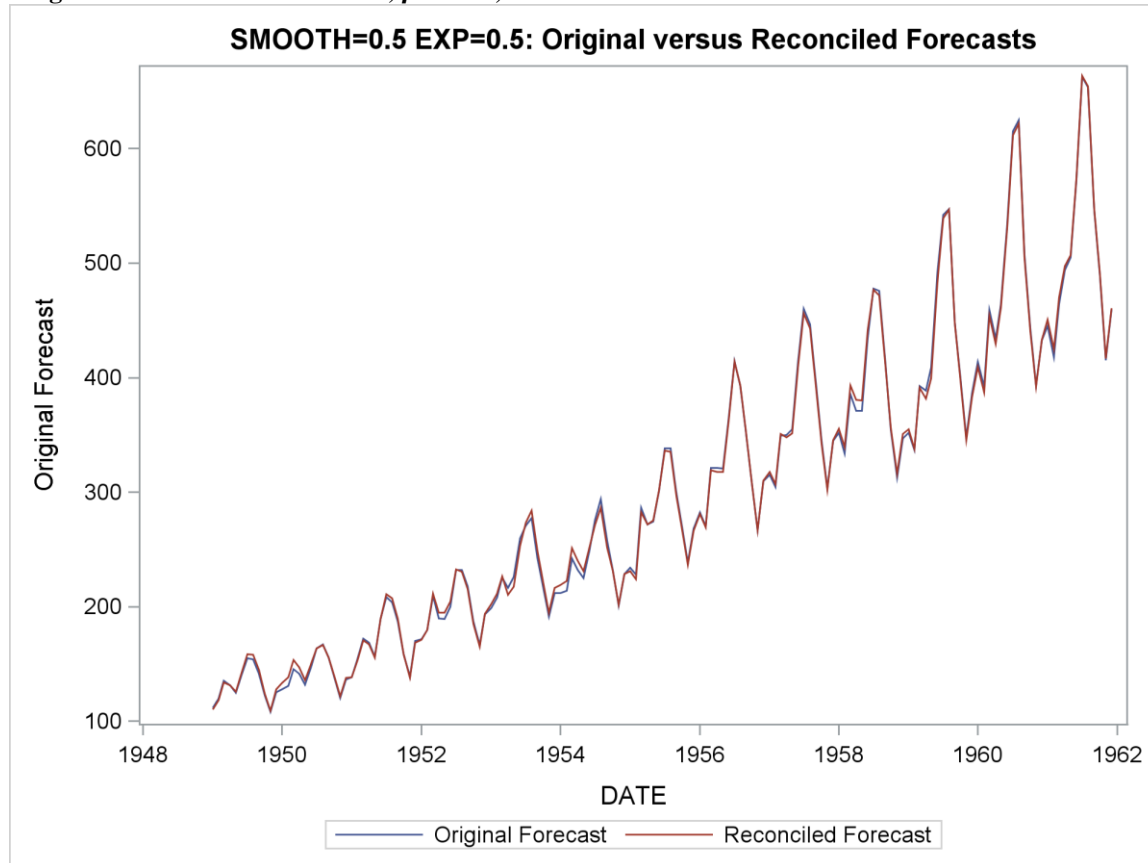
First, notice that the data set of the monthly forecasts is the argument of the DATA= option in the HPFTEMPRECON statement, and the quarterly forecasts data set is the argument of the BENCHDATA= option.

Second, notice that there are two statements to specify the frequency of the data, one for each input data set that contains the predictions. The ID statement is associated with the DATA= data set and specifies the variable that contains the time index of the indicator predictions and its relative frequency (interval). The BENCHID statement is associated with the BENCHDATA= data set and specifies the variable that contains the time index of the benchmark predictions and its relative frequency. Remember that the interval of the ID variable needs to be fully nested in the interval of the BENCHID variable. For example, months are fully nested in quarters. On the contrary, weeks are not

fully nested in months, since a week can span two months. Therefore, the frequency of the indicator series cannot be weekly when the benchmark series has a monthly frequency.

The ρ and λ parameters are set by the EXP= and SMOOTH= options, respectively, in the HPFTEMPRECON statement. You can vary the reconciled forecasts by selecting the values of the SMOOTH= and EXP= options. Figure 3-1 shows the original forecasts versus the reconciled forecasts when both parameters are equal to 0.5.

Figure 3-1.
Original vs. Reconciled Forecasts, $\rho = 0.5$, $\lambda = 0.5$

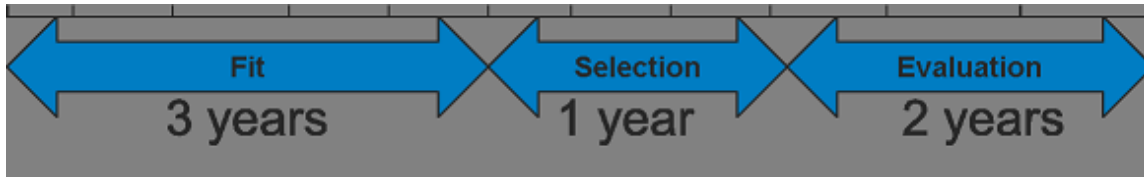


4. Data Analysis

This section applies the method discussed in the preceding sections to a data set of real data that consists of several time series, most of which show intermittent behavior. The data represent six years of monthly demand for 753 parts at the British Royal Air Force (RAF), between July 1992 and June 1998, for a total of 72 observations. Demand for spare parts is a typical example in which intermittent demand is usually encountered. And, indeed, a majority of the series in this collection exhibit intermittent behavior.

First, forecasts are generated independently at the monthly and quarterly intervals. Two years of data are used to fit the model. One year is used for out-of-sample model selection. After model selection, the model parameters are estimated again to use the full three years of data. That leaves two years of data for evaluation of the performance of the forecasts. SAS Forecast Server is used to perform model selection. The full details of the model selection procedure it uses can be found in Leonard (2002).

Figure 4-1.
Model Selection and Evaluation.



RMSE is chosen as selection criterion because it can be computed unequivocally regardless of the value of the series. The most common selection criterion in the forecasting practice, the mean absolute percentage error (MAPE), is not meaningful with intermittent series.

Figure 4-2 and Figure 4-3 display the model family selected for the monthly and the quarterly data, respectively. You can see that for approximately 50% of the monthly series, a model for intermittent data is selected. This proportion is dramatically reduced for the quarterly data.

Figure 4-2.
Model Family Distribution for Monthly Data.

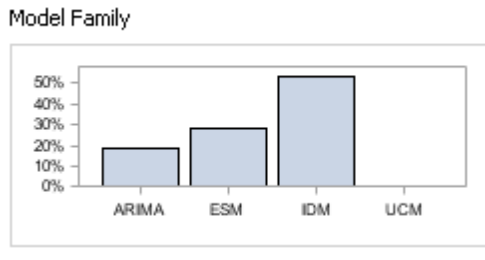
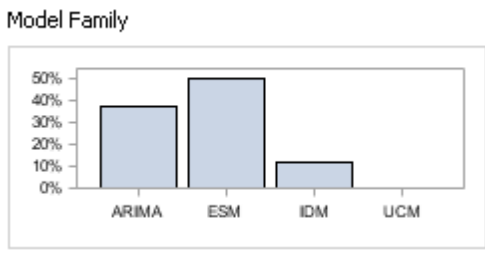


Figure 4-3.
Model Family Distribution for Quarterly Data.



The monthly forecasts are reconciled to the quarterly forecasts for a grid of values of ρ and λ , with $\rho \in (0, 0.1, 0.2, \dots, 0.9, 1)$ and $\lambda \in (0, 0.5, 1)$. For each series the set of values of ρ and λ is selected as those that minimize the out-of-sample RMSE in the selection interval. Finally, the RMSE of the reconciled forecasts is compared to the RMSE of the original model forecasts in the two-year evaluation period.

The RMSE of the reconciled monthly forecasts for the selected values of ρ and λ is improved for 562 of the 753 series when compared to the original model RMSE. The average improvement for these 562 series is 52%.

5. Conclusions

This paper presents a method for reconciling higher-frequency forecasts to lower-frequency forecasts for a time series accumulated in a hierarchy of time intervals. The method is based on the minimization of a quadratic loss function subject to the constraint that the reconciled lower-frequency forecasts accumulate to the higher-frequency

intervals. Under certain circumstances, the problem can also be interpreted as a regression problem. This method is implemented in the SAS HPFTEMPRECON procedure. The target function depends on two parameters whose selection can depend on the same criteria that are used to select the models for the forecasts at the two frequencies. The application of this method can lead to more accurate forecasts when the data at higher frequency are mostly intermittent and therefore are not suitable for models that include features such as input variables, events, and seasonal components.

References

- Box, G.E.P. and G.M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, revised edition, San Francisco: Holden-Day.
- Croston, J.D. (1972), "Forecasting and stock control for intermittent demands," *Operations Research Quarterly*, 23, No. 3.
- Dagum, E.B. and P.A. Cholette (2006), "Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series", volume 186 of *Lecture Notes in Statistics*, Springer Verlag.
- Denton, F. (1971), "Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization," *Journal of the American Statistical Association*, 82, 99-102.
- Leonard, M.J. (2002), "Large Scale Automatic Forecasting: Millions of Forecasts," paper presented at the International Symposium of Forecasting.
- Quenneville, B., Fortier, S., Chen, Z.-G. and E. Latendresse (2006), "Recent developments in benchmarking to annual totals in X-12-ARIMA and at Statistics Canada," in *Proceedings of the Eurostat Conference on Seasonality, Seasonal Adjustment, and Their Implications for Short-Term Analysis and Forecasting*, Luxembourg.
- Wei, W.W.S. (1990), *Time Series Analysis: Univariate and Multivariate Methods*, Redwood: Addison-Wesley.

Improving calendarization using X-12-ARIMA: Application to GST data

Rossana Manríquez¹

Abstract

The Canada Revenue Agency shares information on enterprises with Statistics Canada in the form of Goods and Services Tax remittances. Enterprises can remit annually, quarterly, monthly or more frequently. A calendarization process is needed so that transactions are on a comparable monthly basis. One of the inputs in calendarization is the indicator series, which shows the monthly movement. This paper will discuss the construction of indicator series, improvements made and studies conducted to validate indicator series.

Key Words: Administrative data; Calendarization; Indicator series; X-12-ARIMA; Revisions.

1. Introduction

Administrative data are becoming increasingly important in the various survey programs at Statistics Canada. The Canada Revenue Agency (CRA) collects information on the Goods and Services Tax (GST) from incorporated and unincorporated enterprises. The GST is a value-added tax that applies to most goods and services provided in Canada. Enterprises collect the GST and remit it to the CRA, which has been sharing this information with Statistics Canada since 2003. The information includes data on the enterprise's revenue, the amount of tax and the input tax credit.

Enterprises have to report their revenue at a given frequency: enterprises with annual revenues less than \$1.5 million must report their revenue at least annually; those with annual revenues between \$1.5 million and \$6 million must report quarterly; those with revenues greater than \$6 million, monthly. Each month, the CRA sends Statistics Canada new transactions as well as updates to the file for the previous month. The file can contain transactions dating back four years, but most transactions are recent. Transactions are of different lengths and frequencies and do not necessarily coincide with the beginning or end of the month. Statistics Canada applies different processes to the data, including calendarization, which is used to obtain monthly transactions. GST data are a monthly source of administrative data, offering an alternative to the cost and response burden associated with business survey activities. They are used by a number of sub-annual business surveys and the National Accounts. For internal needs, the data must be available quickly and be of good quality, while reducing the number of revisions.

Section 2 defines calendarization and describes how indicator series are constructed, comparing two methods of doing this. Section 3 reports on a study that assesses whether the quarterly movement of indicator series based on enterprises reporting their revenue on a monthly basis corresponds to the movement for enterprises reporting their revenue on a quarterly basis. Section 4 presents an empirical measure of the stability of indicator series. The article ends with a conclusion.

¹Rossana Manríquez, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6 (rossana.manriquez@statcan.gc.ca).

2. Calendarization and indicator series

2.1 Calendarization - general

Calendarization is a process of converting the values in a time series of flows observed over different time intervals into values that cover calendar intervals, such as day, month, quarter or year. Take the example of an enterprise in retail trade that reports its revenue on a quarterly basis. In particular, we are interested in a transaction going from October 1 to December 31, for a given total revenue. We want to divide this total revenue into three transactions covering respectively the months of October, November and December. Dividing the total revenue into three equal amounts is not the best approach, since retail trade generates more revenue in December. We therefore need a temporal distribution that tells us how to divide revenue according to calendar months. This temporal distribution will be what we will call an indicator series in the rest of the article.

We employ two methods: one that uses regression techniques (Dagum and Cholette, 2006) and the other, linear interpolation using a spline (Quenneville et al., 2010). The equation that follows defines calendarization in terms of a minimization problem. Assume that ρ and λ are given and a_m is the revenue from the m^{th} transaction; we want to find the values $\theta_t, t = 1, \dots, T$ that minimize this function of θ :

$$(1 - \rho^2) \left(\frac{s_1^* - \theta_1}{|s_1^*|^\lambda} \right)^2 + \sum_{t=2}^T \left\{ \left(\frac{s_t^* - \theta_t}{|s_t^*|^\lambda} \right) - \rho \left(\frac{s_{t-1}^* - \theta_{t-1}}{|s_{t-1}^*|^\lambda} \right) \right\}^2$$

subject to $\sum_{t \in m} \theta_t = a_m$. The series s_t^* is the indicator series giving the movement that we want for the calendarized series.

2.2 Constructing indicator series

One of the calendarization inputs is the temporal distribution that gives the industry's monthly movement, that is, the indicator series. The only movement observable and available is obtained as a function of certain enterprises that report their revenue monthly. These enterprises are a subset (approximately 5%) of the total population. They are mainly high-revenue enterprises. On average, over the past four years, they have had annual revenue 25 times higher than that of the population to be calendarized. These high-revenue enterprises are used to calendarize enterprises with medium or low revenue. It seems unlikely that the economy behaves the same for high-revenue enterprises as for those with low or medium revenue. However, the proportion of revenue that is calendarized in the database is only 20%, and therefore the calendarization process does not have a major influence overall.

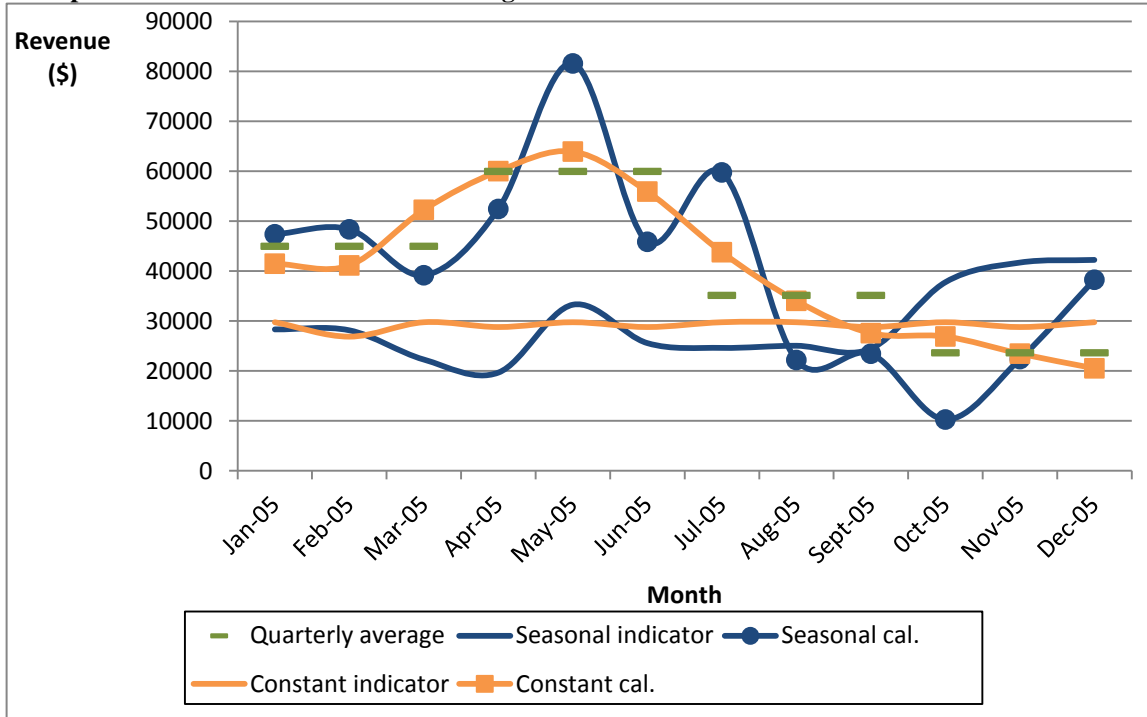
It is also possible that for a given industry, the monthly movement will be solely the effect of the number of days in the month. In that case, a constant indicator series that depends solely on the length of the month will be used. Figure 2-1 provides an example of calendarization using a constant indicator series, with a seasonal indicator series for a quarterly enterprise. It shows, for example, that the March revenue in the calendarized series using the constant indicator series is above the average for the quarter, with the revenue for the next quarter being higher. On the other hand, with the seasonal indicator series, the March value is below the average, since the indicator series declines between February and March.

Figure 2-1 shows that it is important to be careful when choosing indicator series for the calendarization process. We use the X-12-ARIMA software to explore them. This software is used to estimate the components of time series: trend-cycle, irregular component, seasonality, the Easter effect and the trading-day effect.

Below we present two methods of constructing indicator series: the one used in production until September 2010 and the one currently used in production. We describe the advantages and disadvantages of each method. These series are produced at the national level, for each industry, for a total of approximately 1,000 industries.

Figure 2-1

Example of fictitious data calendarized using a constant indicator series and a seasonal indicator series



2.2.1 Constructing indicator series – old method

For the old method, a study was conducted to ascertain the presence of seasonality in each industry. The decision to use a constant or seasonal indicator series was made using the X-12-ARIMA software, since its many diagnoses serve to determine whether or not seasonality is present in a time series.

In the case of non-seasonal industries, a constant series is used in calendarization. Otherwise, the indicator series are deemed to be seasonal and are built each month. The indicator series thus obtained are up to date, since they take every change into account. However, they give rise to revisions that are not always justifiable.

The indicator series are comprised of the average revenue, per industry code, of certain enterprises that report their revenue monthly. To be one of the contributors to the indicator series, the enterprise must exist in the most recent reference month; from one month to another, enterprises disappear or are added to the contributors, thereby contributing to the variability of the indicator series. Also, the most recent industry code is used, which causes the indicator series to vary when an enterprise changes from one industry to another following an update of the classification. The updating of contributors' past remittances is another source of revision of indicator series.

By their nature, these indicator series include all the components of time series. We can hypothesize that the irregular component is the result of an atypical value coming from a single enterprise and not an actual case of an irregular component of the industry. The trend-cycle allows us to track the economy in real time, provided that the variations that it represents are not due solely to a single contributor.

The enterprises that contribute to the indicator series are high-revenue. There is some question as to whether the economy will affect medium- and low-revenue enterprises in the same way. The information on enterprises that are calendarized using these indicator series are revised every month and over their entire history, since the indicator series changes every month. The past is therefore constantly revised (Beaulieu and Quenneville, 2008).

2.2.2 Constructing indicator series – current method

Once again, we derive the indicator series using certain enterprises that report their income monthly. However, with the current method, we impose the same annual revenue—a benchmark—to all these enterprises. More specifically, we calendarize the benchmark revenue using the enterprise's monthly revenue. In this way, enterprises' monthly movements are of comparable magnitude and thus it is no longer the enterprise with the highest revenue that drives the movement. An average, which we call the “democratic mean”, is then calculated.

The population of enterprises contributing to the indicator series is not limited to those that are active in the current month. The number of enterprises is thereby increased, since enterprises contribute every month in which they are active. The choice of classification by industry is also different. It corresponds to the classification in December of the year targeted. Finally, if the series is seasonal, we extract the components of the series obtained on the basis of the democratic mean. Otherwise, we use a constant series. We retain the calendar factors, that is, we exclude the trend-cycle and the irregular component, but we include with the seasonal factors the Easter component and the trading-day effect where applicable.

Under the current method, indicator series are revised once a year rather than monthly. Consequently, what we lose in timeliness, we gain in stability. For the months not covered by the data, we use the forecasts generated by X-12-ARIMA. With respect to enterprises to be calendarized, the revision due to calendarization is conducted less often; it is now annual rather than monthly.

3. Study conducted to validate the calendarization of quarterly enterprises

3.1 Calendarization of quarterly series

Taking an indicator series obtained from enterprises that report their income monthly, we use its movement to calendarize enterprises that report their income on a quarterly basis. We want to validate this practice. The monthly distribution of enterprises that report their income quarterly is not available, and therefore we must make our comparisons on a quarterly basis. We approached the problem in two ways. We chose six industry codes corresponding to food services and drinking places (France, 2010).

3.1.1 Calendarization with quarterly indicator series

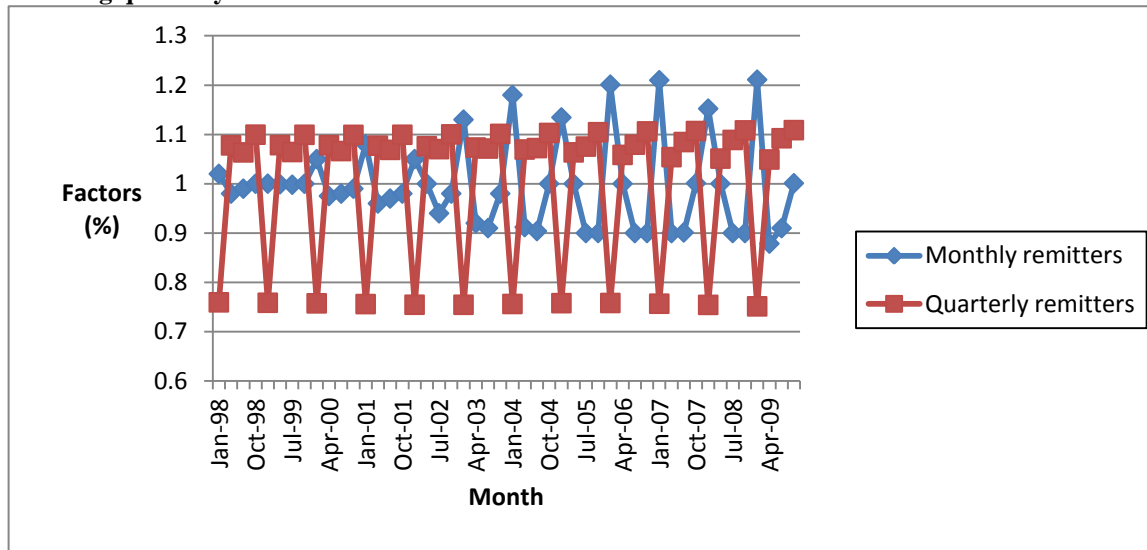
We constructed new indicator series on a quarterly basis. Then we considered enterprises reporting according to fiscal quarters. We extracted the calendar factors using X-12-ARIMA, and we calendarized enterprises that had to report their income quarterly. Then we compared these results to the production data for these same enterprises by summing the monthly revenues in order to recreate the corresponding quarters. The results were comparable. The largest differences were observed for enterprises with fewer than eight quarters reported.

3.1.2 Comparison of seasonal factors

We looked at monthly contributors in the indicator series used in production. We added together the months in fiscal quarters and we extracted the calendar factors. Then we made a comparison with the calendar factors obtained in the previous study.

In general, we obtained very good results. The only questionable result concerns the series for industries not identified as seasonal on the basis of monthly contributors. An example is given in Figure 3-1. As may be seen, it might be essential to have auxiliary information or the knowledge of an expert in the field in order to construct an indicator series that reflects the seasonality of enterprises reporting quarterly.

Figure 3-1
Sample comparison of quarterly calendar factors obtained with enterprises remitting monthly and those remitting quarterly



4. Empirical measurement of indicator series' stability and representativeness

4.1 Measurement objectives

An indicator series is produced on the basis of a subset of the population. We cannot verify whether it provides a good representation of the monthly movement of the entire population. However, we can verify whether it represents the sub-population from which it is drawn and whether the seasonality is similar for each contributor (Delavaquerie, 2011). We can also verify whether it is influenced by some of the elements that comprise it. To do this, we chose to develop an empirical measure based on the jackknife method (Girard, 2009).

We hypothesize that the greater the number of contributors, the more stable the indicator series will be, meaning that the suppression of a contributor will not influence the resulting indicator series. To validate this hypothesis, we will break down the results by the number of contributors in the industry. We are already using a seasonal adjustment measure (X-12-ARIMA's M7 statistic), since we want to make sure that the empirical measure of stability does not measure the same thing.

We have approximately 1,000 industries to check in a fairly short time period, so we want to identify those industries that require further analysis before deciding whether to use a seasonal or a constant indicator series. The measure must also meet this need.

4.2 Method

We calculate the democratic mean for each replicate and we extract the calendar factors, using X-12-ARIMA, imposing the multiplicative model and using automated models. We obtain the indicator series s_r^* composed of the calendar factors for the replicate r . The original series s_o^* is the one with the calendar factors extracted from the series constructed with all contributors. We then compare all the replicates in the original series and present two ways of summarizing the information. Some indicator series for particular industries do not meet all the requirements of X-12-ARIMA. These industries were set aside. For reasons of computing time, we set aside industries with more than 350 contributors.

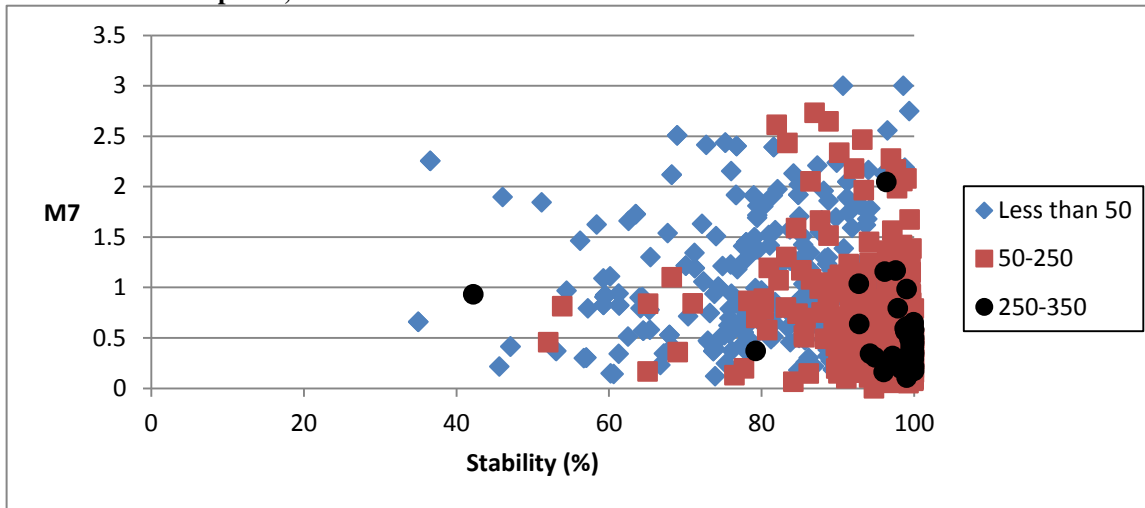
4.2.1 Empirical measure based on X-12-ARIMA's Sliding Span spec

For each replicate r and each reference month t , we calculate the following measure: $|s_{t,r}^* - s_{t,o}^*|/s_{t,o}^*$ and we use the thresholds suggested by X-12-ARIMA's Sliding Span spec (U.S. Census Bureau, 2009). If this distance is greater than 3%, we will say that the measure is unstable; otherwise it is stable. If, for a given industry, the ratio of the number of stable measures to the total number of measures is greater than 85%, we will say that the series for the industry is stable.

Figure 4-1 shows that the M7 statistic is related to stability, but the relationship is not perfect. It also shows that stability depends on the number of contributors in the industry and an industry with more than 250 contributors is almost always stable. Now we can set thresholds below which stability is considered unsatisfactory and examine these cases more carefully. When this measure is implemented in production, we will easily be able to determine the thresholds for identifying industries requiring further analysis. For example, in the 250-350 category, the two industries with stability below 85% will require follow-up.

Figure 4-1

Adjustment of seasonal adjustment based on percentage of stable measures – Industries classified according to number of enterprises, three classifications



4.2.2 Empirical measure based on coefficient of variation

We define a measure that is, to all intents and purposes, a coefficient of variation (c.v.):

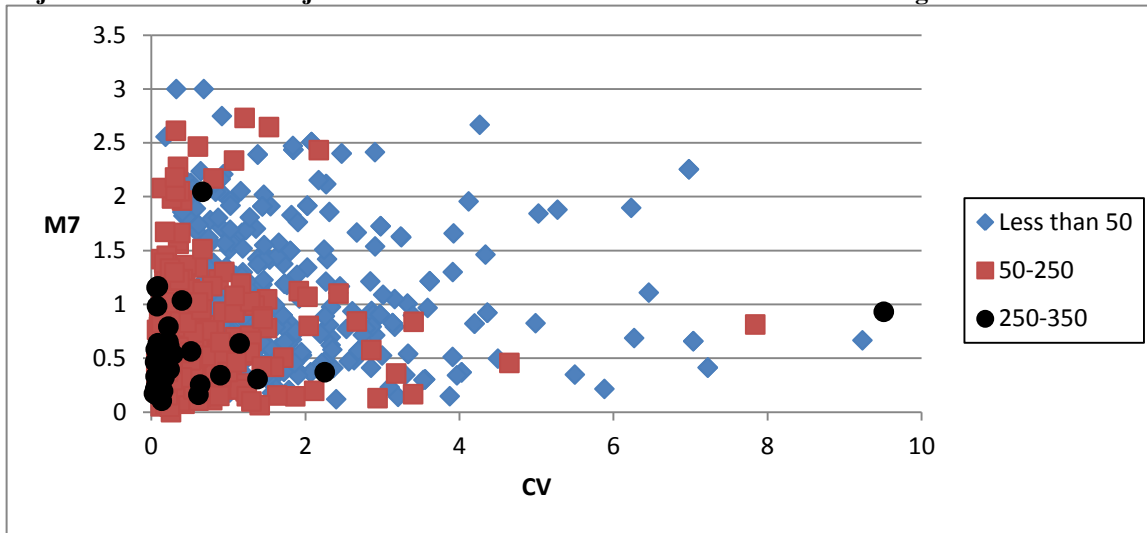
$$CV_t = \frac{1.4286 * \text{Median}_r |s_{t,r}^* - s_{t,o}^*|}{s_{t,o}^*} \quad (2)$$

The correction of 1.4286 allows us to say that in a normally distributed sample, the value returned should, on average, be approximately equal to the standard deviation. Thus, the measure may be seen as an unbiased estimator of the standard deviation in the population. We obtain one measure per month. We can therefore combine these measures to study a specific period, such as the recent past of the indicator series. For the rest of this article, we will take the average over all the months available. We thus obtain one c.v. per industry.

In Figure 4-2, we can see that the number of contributors affects the measure of the c.v., just as it affected the percentage of stable measures in Figure 4-1. We will determine the different thresholds according to the number of contributors in order to identify the industries to be analysed in greater detail.

Figure 4-2

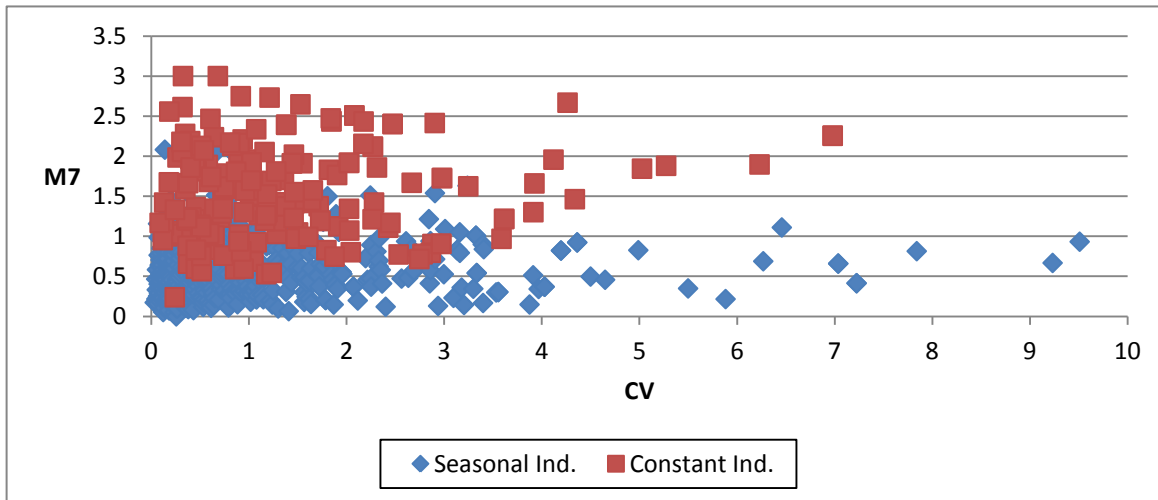
Adjustment of seasonal adjustment based on c.v. – Industries classified according to number of enterprises



In Figure 4-3, we classify industries according to whether or not they are seasonal. We can see that the c.v. measured seems smaller for non-seasonal industries. For seasonal indicator series, we use the M7 values obtained in production. The decision as to whether or not to use a seasonal indicator series for an industry depends on several factors, including not only seasonal adjustment but also the prior knowledge of an expert in the field, the number of contributors and potential revisions.

Figure 4-3

Final adjustment of seasonal adjustment based on c.v. – Classification of industries according to whether the indicator series is seasonal or constant



5. Conclusion

We recently improved calendarization by changing the way indicator series are constructed. The new series have a larger number of contributors and the classification is more stable for purposes of revising indicator series. They are less influenced by high-revenue enterprises, since the calculation is done using the democratic mean. These series are now revised annually, which makes it possible to limit the revisions due to indicator series. We are aware of the loss of timeliness for the trend-cycle component. Efforts must be made to improve this aspect.

We verified that for six industries, the result of calendarization seemed valid. Studies on the calendarization of quarterly enterprises are encouraging. It would be interesting to repeat the study for all industries. The jackknife-based measure seems highly promising for gaining a better understanding of contributors and improving efficiency in the efforts that we devote to each industry. Now we want to implement the measure in production. We also want to use the measure corresponding to a c.v. in order to study the stability of indicator series in the recent past.

Acknowledgments

The author would like to thank Susie Fortier, Benoit Quenneville and Christian Wolfe for the advice that they provided throughout this project. Thanks are also extended to Martin Beaulieu, Johanne Boivin, François Brisebois, Pierre Cholette, Estela Bee Dagum, Hugo Delavaquerie, Louis-Baptiste France, Claude Girard, Joanne Leung, Marie-Eve Mainville, Frédéric Picard and the members of the Time Series Research and Analysis Centre for their contribution to the project.

References

- Beaulieu, M. and B. Quenneville (2008), “Calendarization of the Goods and Services Tax (GST) Data: Issues and solutions”, *Proceedings of the Joint Statistical Meetings*, Section on Survey Research Methods, JSM 2008.
- Dagum, E.B. and P.A. Cholette (2006), “Benchmarking, temporal distribution, and reconciliation methods for time series”, *Lecture Notes in Statistics*, New York: Springer, vol. 186.
- Delavaquerie, H. (2011), “Une mesure de la volatilité des facteurs saisonniers”, unpublished report, Ottawa, Canada, Statistics Canada.
- France, L.-B. (2010), “Rapport de stage effectué à Statistique Canada”, unpublished report, Ottawa, Canada, Statistics Canada.
- Girard, C. (2009), “Un guide d’estimation de la variance pratiquement intelligible”, working paper, Ottawa, Canada, Statistics Canada.
- Latendresse, E., Djona, M. and S. Fortier (2007), “Benchmarking Sub-Annual Series to Annual Totals – From Concepts to SAS[®] Procedure and SAS[®] EnterpriseGuide[®] Custom Task”, *Proceedings of the 2007 SAS Global Forum*.
- Quenneville, B., Picard, F. and S. Fortier (2010), “Interpolation, benchmarking and temporal distribution with natural splines”, *Proceedings of the Joint Statistical Meetings*, Business and Economic Section, JSM 2010.
- U.S. Census Bureau (2009), *X-12-ARIMA Reference Manual*, Version 0.3, Washington, D.C.

The error in business cycle estimates obtained from seasonally adjusted data

Tucker McElroy¹

Abstract

Economists have great interest in measuring the business cycle inherent to economic time series, but they typically base their analysis on published seasonally adjusted data. We study the question of how the estimation of the cycle is affected by model-based seasonal adjustment, and how one can quantify the additional signal extraction error due to prior seasonal adjustment.

We give a precise theoretical description of the asymptotic values of maximum likelihood estimates obtained by fitting misspecified models, and use these so-called pseudo-true parameter values to quantify the asymptotic mean squared error of cycle estimates computed from seasonally adjusted data. A full illustration is provided on an employment series.

¹Tucker McElroy, U.S. Census Bureau, USA.

SESSION 11A

BUILDING AND USING GENERALIZED SYSTEMS

Generalized systems: The Statistics Canada experience

Yves Deguire, Laurie Reedman and Michael Wenzowski¹

Abstract

Statistics Canada has a long history of developing generalized software solutions. Our suite of generalized systems spans the complete spectrum of the typical survey processing cycle: from sampling and estimation, to edit and imputation, through to tabulation and disclosure avoidance.

The current set of generalized systems at Statistics Canada includes many systems that have evolved over multiple generations, spanning multiple decades. In short, this has presented us with many opportunities for 'lessons learned,' which has resulted in a progressively more refined and more efficient set of offerings.

We present a brief history of the motivation, specification, engineering and use of generalized systems at Statistics Canada; the methods and techniques routinely used to specify them and prescribe their use; the architectural and engineering challenges inherent in the process of building them; and the feedback and control mechanisms implemented in collaboration with each product's respective user community.

Key Words: Generalized systems; Engineering; Architecture.

1. Background

Statistics Canada has a long history of creating generalized statistical processing software. Our first systems began to appear in the late 1970s, and initial offerings focussed on editing, table preparation and statistical database management. Since then, we've produced generalized systems to address requirements such as sampling, estimation, edit and imputation, coding, time series analysis and disclosure avoidance.

It is probably very safe to say that the Generalized Systems Program at Statistics Canada stands out as one of the most comprehensive sets of general-purpose statistical processing software available within a national statistical agency. While our current set of products represents many decades of progressive refinement and new development, it must be said that the initial impetus for their creation lies with the nature of the agency itself (Kovar, Jeays and Poirier, 1999). The Government of Canada's statistical systems are highly centralized and most of the processing of this vast assortment of disparate data collections is performed by Statistics Canada. Very early on, this resulted in the recognition of significant re-use and generalization opportunities, as statisticians were repeatedly exposed to similar requirements across many different data processing applications.

The earlier set of generalized systems resided on a single processing architecture (an IBM mainframe computer), and were quite monolithic in their nature. The current offering of our generalized systems includes software that operates largely within a Microsoft processing environment, but also across a wide range of Unix-based processing environments. In addition, our current offerings are significantly modularized, which allows for users to select only the functionality and operating modes required for a particular application. Many of our systems are available for both batch and interactive modes of operation, as well as the capability of being embedded within some other hosting software environment. This latter mode of operation completely removes any outward indication that a particular generalized system is being used, and offers the user only the interface presented by the hosting application.

The decades of technical evolution of this software set has also been complemented by a significant evolutionary trend in increasing and enhancing overall usability (Otrata and Chinnappa 1989). The result is that our systems do not require the intervention of technical staff in order to be used within a given application. Instead, what is required is subject matter and methodological expertise in order to ensure that processing is performed in an appropriate manner.

¹Yves Deguire, Laurie Reedman and Michael Wenzowski, Statistics Canada, Ottawa, Ontario, K1A 0T6.

Benefits that have accrued from this program include the provision of very robust software implementations. This is because the potential to apply a generalized solution to many processing applications helps to justify higher initial development costs. The result is that the software is more fully featured and better tested. In addition, because the software is used by many applications, the code within the software is executed much more frequently, and with greater variance than a typical custom-built application. This results in problems, bugs and limitations being found very early in the product’s life cycle, which ensures that downstream applications ultimately benefit.

Another significant benefit that has been realized is that of ‘standardizing’ the methodology used within a particular problem domain. This, in turn, yields the benefit of being able to move staff from project to project with greater ease, since they are already conversant with the methodology and software in use.

The following table lists the generalized systems which comprise the current suite of Statistics Canada’s offerings.

**Figure 1-1
Current Suite of Statistics Canada Generalized Systems**

System Name	Function
Banff	Edit and Imputation
G-Code	Coding
G-Confid	Disclosure Control
GES	Weighting and Estimation
G-Link	Record Linkage
GSAM	Sampling
G-Series	Time Series
G-Tab	Tabulation
LogiPlus	Editing (Decision Tables)

Despite the fact that, in general, it simply costs more and takes more time to develop a generalized system than a custom system, significant economies are achieved through their use (Poirier 2011). The primary economy is achieved when the software is used to address the multiple requirements of many different applications. Secondary to this is the significant reduction in ongoing support costs, which are realized through a reduced staff allocation to support the many different applications. (That is, since many applications use the same software, the software can be supported by a single support group.) Another economy achieved is that of abbreviated planning times. New survey applications can take much less time to plan if they are based upon widely-known and intimately-understood generalized systems.

For such a program to be realized it is essential that funding and support be provided at a high level, and that a certain degree of mandated use be prescribed. Centralized funding ensures that no single application bears the cost of the software development and prescribed use ensures that the agency is able to reap the intended level of benefit from the capital outlay and development effort. Of course, no single generalized system can offer absolutely everything to any and all applications in which it is deployed. The success of the program relies heavily on both ensuring that all significant processing functionality is included in an easy-to-use manner and that the end-user of the software is willing to accept a certain amount of compromise. We believe that we have somehow managed to strike such a balance with our current offerings and we continue to work closely with all users of the software—both current and potential—to ensure that shortcomings and opportunities for needed enhancements are identified early, prioritized and included in the longer-term support plans for our products.

2. Roles and Responsibilities

2.1 Development Roles

The development phase is driven by a well-structured Rational Unified Process (RUP) (Kroll and Kruchten). During the development phase, there are three distinct groups who contribute. Statistical researchers develop the

methodology and build prototypes, methodology developers generalize the methodology and write detailed specifications, and system engineers determine the system architecture and write the programming code.

First, statistical researchers recognize the need for a particular methodology in a statistical application. They develop the idea, expressing it in terms of concepts, algorithms and algebraic expressions. The statistical researchers build a prototype, as proof that the concept will work. The prototype is tested to make sure it produces the expected results. For example, they test to make sure the results are accurate when the prototype is used on real data. They also test for acceptable performance under realistic conditions, using both typical as well as extreme data. The statistical researchers document what the prototype does, the theory behind it and how it works.

Methodology developers consider the work done by statistical researchers and the business requirements of the subject matter clients, and recommend to management which methods should be included in the generalized systems. Some methods are clearly good candidates for inclusion, such as simple random sampling, calibration weighting and donor imputation. However, in domains such as disclosure control and sample coordination, research continues towards finding optimal methods. It is not obvious at what point a method is considered to be sound and mature enough or applicable to enough different survey programs that it should be incorporated into a generalized system. Often, a compromise is struck, balancing resources and demand (Poirier 2004). Ultimately, management gives their support for the use of a particular methodology and therefore its development as a generalized module to serve global (across several programs) rather than local needs. The prototype with its documentation is given to the methodology developers who use it in the testing phase and who use the documentation as an aid in writing the detailed specifications.

The methodology developers document how users will interact with the modules, what parameters are needed, what inputs will be required and what outputs will be produced. They are the link between the statistical researchers and the system engineers. These methodologists understand how the prototype works and how it will be used. They understand the overall concepts and are skilled at writing detailed specifications. The deliverables produced by the methodology developers include the business requirements as well as detailed specifications. These specifications describe the inputs and outputs in plain language. Mathematical formulas as well as descriptions in words are used to describe what manipulations to perform on the inputs in order to get the desired outputs. The specifications do not include pseudo-code.

The system engineers analyze the specifications to determine how best to do the implementation. They meet regularly with the methodology developers to clarify the needs and to identify how the different modules will interact with each other. They investigate implementation options and determine the most appropriate system architecture. It is only once the specifications are fully analyzed and the complexity of the programming task is well understood that the system engineers can accurately predict how long the programming step will take.

2.2 Testing and Certification

The system engineers test each module to ensure that it performs according to their interpretation of the specifications. The methodology developers test each function individually and also integrated with the other modules to ensure that it performs as expected in typical as well as extreme situations. Often, the modules are made available to a set of experienced users for beta testing. These are often methodologists who plan to use the modules when they are finalized and who can test them in a real environment. The beta testers provide very useful feedback to the methodology developers and the system engineers. Beta testing provides a very welcome opportunity for system architects from the subject matter area to see how the modules integrate into their production systems.

The development and testing phases are iterative. Once a module has successfully cycled through the various testing phases, the methodology developers certify its completeness and the module is formally released to the user community.

2.3 Ongoing User Support

The methodology developers are responsible for maintaining two-way communication with the user community throughout the entire development cycle. With support from management, they promote the use of the generalized systems over other 'local solution' systems. They give seminars to increase awareness of what methods are available

and they provide training courses and workshops for hands-on experience. On an on-going basis, the methodology developers find solutions to user problems that relate to the methodology itself. They communicate back to the statistical researchers additional needs that are expressed by users. This can initiate a change request or additional development (Kozak 2005). They produce and maintain a user guide, tutorial and methodology documentation to accompany each module. The intended audience for the user guide is the end user. This could be a methodologists or subject matter specialists. The user guide describes how to make the module perform the various methods, for example, what parameters need to be set and what inputs are required. The tutorial is also written primarily for the benefit of the end user. It is self-guided learning through the use of examples. The methodology documentation describes the statistical theory and provides more detail than the user guide.

The system engineers solve user problems that are related to the software itself. They are responsible for maintaining the programming code and its accompanying documentation. Bug fixes and enhancements are implemented as the need arises. The system engineers can initiate a change request based on software considerations, such as re-writing parts of the code to take advantage of improvements in later versions of the foundation software.

3. Software Engineering Considerations

3.1 Software Engineering and Generalized Systems

The development of generalized systems yields software that implements complex algorithms and processes large amounts of survey data. Such software must be high quality and can only be developed using strong software engineering practices (Pressman, 2005).

Software engineering brings a disciplined and systematic approach that is enforced by the use of modern development tools. Software engineering is used during the entire development lifecycle, which includes the analysis, design and implementation of the software. Some of the practices used at Statistics Canada in the development process include: software modeling, coding guidelines, version control, code reviews and unit tests.

3.2 Four Important Software Characteristics

The application of sound software engineering practices must be supplemented with a clear understanding of the outcome of the process. In the context of generalized system, it can be described with four characteristics that the resulting software must possess.

Adaptability

Adaptability is the ability to adapt to different requirements. A generalized system, as its name implies, is to be used by a large number of surveys and, as such, must be developed to accommodate runtime processing specifications. To achieve this, generalized systems should be developed in a highly cohesive modular fashion whereas each module implements a specific statistical function (Veryard, 2001). By not overloading a module, it is easier to build the software in a flexible manner such that the user can easily alter the constraints and the assumptions (in other words, the module must be 'parameter driven').

Reliability

Reliability is the ability to produce accurate results in a timely fashion. Because of their status as corporate standards and their widespread use, generalized systems must implement sound, well understood and defensible statistical methods. As such, a generalized system is not a playground for research. Users also expect the software to be robust and that its execution produces results that can be trusted. The software is therefore built with a continuous quality assurance process.

The software must also be efficient at processing large amounts of data and be able to produce results in a reasonable timeframe. Efficiency should not be an afterthought. It starts at the specification stage by questioning inefficient

methods. Statistical researchers play a key role by prototyping those methods before a decision is made on their implementation. The software is also designed and built with efficiency in mind throughout the development process.

Maintainability

Maintainability is the ability to enhance existing functions or add new functions, and to adapt to new operating environments. The development of generalized systems is a long and expensive process. The resulting software must be in production for many years to justify such an important investment on the part of a statistical agency. To do so, the software must be built in such a way that it will survive many changes in the operating environment and be able to entertain enhancements to its functionality.

As already mentioned, the software should be built as a collection of highly cohesive modules. Each of these modules can be enhanced as long as the enhancements pertain to the statistical function it implements. New statistical functions should be built as new modules and hence augment the collection.

The software should also be built with a layered approach in order to isolate core components from the operating environment. The adoption of an application virtual machine such as SAS® or Microsoft .Net is critical in this regard. The next section will expand on software layers.

Interoperability

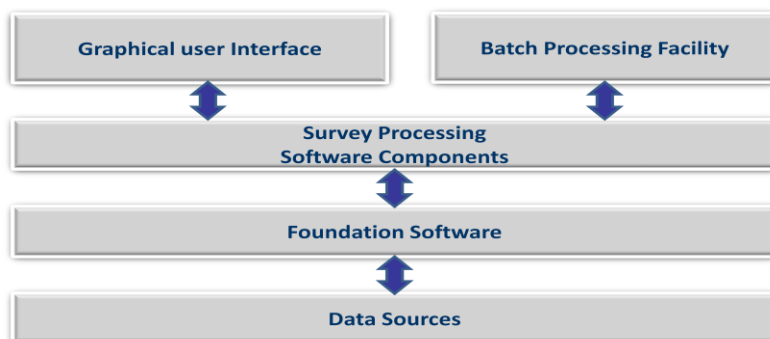
Interoperability is the ability to interoperate with other systems and software. Several generalized systems modules are typically assembled to implement a specific application for a survey. Furthermore, custom and commercial software typically supplement the generalized system modules. This reality implies that these must not only be cohesive but also loosely coupled so that they don't have interdependencies with other modules. The modules must have a clean and well-defined programming interface. This interface allows the output of a module be the input of another with no knowledge of their respective internal implementation.

When modules are invoked across a network as would be the case in a Service Oriented Architecture (SOA) environment, the use of the Extensible Markup Language (XML) is desirable. XML is an industry standard and enables messaging between modules. As a side effect, because of its capability to represent arbitrary data structures, it is a good candidate to define complex processing instructions.

3.3 A Proposed Software Architecture for Generalized Systems

So far, we have discussed software development practices that should yield software with four important characteristics. We will now propose a software architecture: it is essentially a blueprint that organizes the different software elements composing generalized systems. More specifically, we are presenting the multi-tier component-based architecture that has been adopted by Statistics Canada for the development of generalized systems. The following diagram is a graphical representation of the architecture.

Figure 3.3-1
Graphical Representation of Generalized Systems Multi-tier Component-based Architecture



As its name implies this architecture revolves around software components and is organized into layers called tiers (Heineman and Councill 2001). We have already introduced the notion of software components by describing highly cohesive loosely coupled modules as a means to obtain adaptable, reliable, maintainable and interoperable software. In the context of the software architecture, these are called survey processing software components. They are the high level modules that implement the statistical methods. As such, they are central to this architecture.

The notion of software layers has already been briefly introduced. We discussed the importance of isolating the software components with the use of an application virtual machine. A multi-tier architecture goes much further since it separates the presentation, the application processing and the data management. By breaking a system into tiers, the software can be deployed across multiple machines to leverage the computing power of servers and enable distributed systems. The software can also be developed one tier at the time. Even more important, a specific tier can be re-developed without affecting the other tiers, making the overall system less costly to maintain over time. Specifically, the proposed architecture includes five tiers.

Data Sources

Generalized systems must access data in various formats across disparate systems. This layer simply represents the various data storage vehicles such as databases and sequential files.

Foundation Software

This is where the application virtual machine is materialized. It provides a rich set of functions and modules for statistical processing and analysis so a generalized system simply enriches this functionality. This is a runtime environment highly scalable and suitable for processing high volume of data in batch—a typical scenario in survey data processing. It enables the creation of ‘assembly lines’ that run a number of steps serially and/or in parallel. A good example of foundation software with these characteristics is SAS®.

Survey Processing Software Components

Software components are typically implemented using a well-know, well-supported programming language such as C and SAS®.

Graphical User Interface

This is a presentation layer that enables the visual and interactive use of the software components. This tier is optional and is usually developed in the second phase of the development project since it requires the basic software components to be in place before it can be implemented. The development and maintenance of user interfaces command an important effort. Interfaces are prone to frequent maintenance because they are heavily dependent on a rapidly changing technology and are also highly subjective. In this regard, the most recent generalized systems have developed graphical user interfaces (GUI) as plug-ins to existing interfaces. This minimizes the scope of the interface and reduces the cost of its development and maintenance. A good example is the development of SAS® Enterprise Guide® custom tasks using the VB.Net or the C# programming language.

Batch Processing Facility

This optional presentation layer aims at facilitating the creation of production jobs using pre-fabricated software components (generalized, customized, commercial and open-source software) as well as survey-specific processing steps. One approach is to define a series of jobs that call the various components and processing steps using a metadata repository. The jobs themselves only exist for the duration of a run. They are generated at runtime from the metadata. This approach allows the user to create and modify production jobs with minimum intervention on the part of IT.

Overall, this proposed architecture emphasizes flexibility, reusability, low cost and ease of use. It maximizes the use of generalized systems by allowing each survey to adapt the components to its own needs and to insert any other components that it sees fit. The cost is relatively low because most of the effort is spent building the software

specific for the generalized systems while allowing the resulting to be deployed in a distributed computing environment.

4. Conclusions

There is no doubt that it is a long and challenging process to develop the methodology, write the specifications, write the programming code and do the testing for generalized systems. However, the benefits are long-term and more efficient use of resources and more robust tools that can be maintained and added to over time. A strong partnership between methodology, informatics and subject matter is the key to success in developing generalized systems to perform complex statistical functions. Governance is essential to ensure adoption of these systems.

References

- Heineman, G.T. and W.T. Councill (2001), “*Component-Based Software Engineering: Putting the Pieces Together*”, Addison-Wesley Professional, Reading 2001.
- Kovar, J., Jeays, M. and C. Poirier (1999), “Generalized Systems: Where are we at and where are we going”, internal document presented to the Advisory Committee on Statistical Methods, meeting #28, April 1999.
- Kozak, R. (2005), “The Banff System for Automated Editing And Imputation”, *Proceedings of the Survey Methods Section*, Statistical Society of Canada Annual Meeting, June 2005.
- Kroll, P. and P. Kruchten, *The Rational Unified Process Made Easy*, Addison-Wesley, ISBN 0-321-16609-4.
- Outrata, E. and B.N. Chinnappa (1989), “General survey function design at Statistics Canada”, *Bulletin of the International Statistical Institute*, 53: 2, 219-238, 1989.
- Poirier, C. (2004), “The Processing Environment Behind a Statistical Program”, *Proceedings of the Survey Methods Section*, Statistical Society of Canada Annual Meeting, June 2004.
- Poirier, C. (2011), “The Impact of a Changing Business Architecture on Editing”, UNECE Work Session on Statistical Data Editing, Slovenia, May 2011.
- Pressman, R.S. (2005), *Software Engineering: A Practitioner’s Approach* (6th Ed.), Boston, Mass: McGraw-Hill.
- Veryard, R. (2001). *Component-based business: Plug and play*, London: Springer.

Triton: A general tool for data collection and micro editing

Johan Erikson¹

Abstract

The Triton project is an ongoing project with the goal of building a general but flexible production environment for data collection and micro editing. The aim is to cover most kinds of surveys, but in a first stage it is directed at surveys with direct data collection through questionnaires (web and paper). Even though a version of the platform is already in use, a new and significantly improved version is under development. The new version will be released at the end of June 2011. The aim is that the new platform will replace many of the old survey specific IT systems, be usable for a majority of the surveys at Statistics Sweden, integrate the common tools already in place and eliminate as much manual work as possible. Some of the most important expected gains of the platform will be that metadata will have an actual effect on the production process, that quality assurance will be a built-in part of the production process and that there will be a standardised and common look and feel to the production of many surveys, facilitating resource pooling. Besides integrating existing common tools such as the web collection tool and the scanning system, the platform will have three main new parts: an administration/design tool for setting parameters for a specific survey and monitoring the survey progress, a tool for working with individual objects and a communication platform that connects all the parts of the platform. The paper will present the different parts of the platform and how they are used in daily work, using an example survey.

Key Words: Generalised systems; Standardisation.

1. Introduction (and update)

The above abstract was written in the spring of 2011, while the paper was written at the end of 2011. This means that some of the wording in the abstract is outdated. The new version was released at the end of June 2011 as planned, and was implemented in six surveys in the second half of the year. For 2012, around 20 surveys plan to start using the platform.

2. Background

The Triton project was initiated and driven based on the following factors

1. Statistics Sweden is moving to a process oriented statistical production, using common tools, methods and routines whenever possible. These tools, methods and routines are described in the Process Support System (PSS) which was created in 2008 and has, up to now been mostly a bank of information on what to do in different situations. The long-term goal for the PSS is for it to become a more interactive tool that is used to actually drive the production process.
2. Statistics Sweden also aims to move towards a metadata-driven statistical production where design choices have a direct impact on the IT tools used and where more of quality assurance will be built into the production system itself.
3. The data collection process is already using a number of common tools, for example a web data collection tool, a telephone interview collection tool and a tool for optical scanning of paper questionnaires. However, before the Triton project, these tools were – at least for surveys directed at enterprises and the public sector - used as add-ons to survey-specific production systems. There were poor bridges between the different tools, meaning that much manual data shuffling and data transformation was necessary. The survey-specific production systems in turn are getting old, are often built using technique that is now outdated, and in many cases the

¹Johan Erikson, Statistics Sweden, Process Department, Örebro, SE-701 89, Sweden, johan.erikson@scb.se.

maintenance of the systems are dependent on single persons, often the person who built the system a number of years ago.

4. The data collection (including micro editing) for enterprise and public sector surveys has been centralised only in the last few years, before that it was carried out at the different subject matter departments. The centralisation in itself has raised a need to use more common tools, to realise the possible gains of the centralisation, *i.e.*, more efficient and streamlined production, pooling of resources et cetera.

The third and fourth of these factors were the main reasons to start a project with the aim of building a generalized production environment for data collection and micro editing—the Triton project. As the project moved on, it became apparent that this was also a chance to take a large step in reaching the aims specified in the first two points above. The latest part of the project, which was to move the Triton platform from a prototype version to a more readily available one that could be used for a larger number of surveys, therefore decided to place the new version within an updated version of the PSS system, thereby taking a large step towards all the aims described above.

The paper has the following contents: section 3 describes the PSS at Statistics Sweden and what has been done within the Triton project to expand the PSS. Section 4 describes the different parts of the Triton platform, how they are working today and some thoughts for the future and what happens next. Section 5 summarises the results.

3. The PSS at Statistics Sweden

In moving towards a process-oriented statistical production, Statistics Sweden has, as most other countries, decided on a process model that describes the statistical production process. Like in many other countries, the process model is similar but not identical to the General Statistical Business Process Model that has been developed by the United Nations. The process model at Statistics Sweden divides the statistical production process into eight sub-processes:

1. Specify needs
2. Design and plan
3. Build and test
4. Collect
5. Process
6. Analyse
7. Disseminate and communicate
8. Evaluate and feed back

A ninth sub-process, ‘Support and infrastructure,’ is also defined but is somewhat apart from the statistical production.

Five process owners (for processes 1+7, 2+3, 4, 5+6 and 8+9) have been appointed, and are responsible for providing the surveys with agency wide methods, tools and work routines based on the different needs and filling the PSS with information. The information in the PSS covers common tools, methods and routines for how to run the statistical production process as efficiently as possible. This is described in the following manner: Each sub-process is divided further into sub-processes at finer and more detailed level (the level of breakdown varies from one to five additional levels). Each sub-process is described in a template using four main parts: a short description of the purpose of the process, the input needed, the main part that describes how to carry out the sub-process and the output that comes from the sub-process. The main part includes detailed information on tools and methods to use and when to use them—when necessary, even more detailed information, instructions, user’s guides et cetera are placed in additional documents that can be accessed from the sub-process page—and often also includes a step-by-step description of the moments you need to go through to run the process. It can also include templates or checklists where such documents are relevant. If there are several methods and routines that can or should be used in different situations, all the different methods/routines and when to use each one of them is described. In total, the PSS can be said to contain the recommended standards of Statistics Sweden.

The PSS contains a lot of information. This has meant that many people at Statistics Sweden find it hard to find all the relevant and necessary information, and one question often posed by specific surveys is “Which parts are relevant for us?” The developments made within the Triton project in cooperation with additional expertise in a PSS project

has come up with a solution to this problem. A new part of the PSS, called Process areas, has been created. These areas are survey specific areas, where the different surveys gain access to the information, tools and documents that are relevant to them. The process areas are built in Microsoft Sharepoint, which also gives the opportunity to use the built-in functions in that software. The process area is the dashboard from which the survey can run the production. At the moment, the process areas only cover data collection and micro editing, since those were the processes covered by the Triton project, but in the future, the concept of process areas and dashboard functionality will be expanded to other processes within the statistical production process. To allow for ongoing surveys of different types, and for design choices to vary over time within the same survey, the process areas are set up in a hierarchical way with three levels; survey, survey round and collection round. Each collection round can then be monitored on its own, while there is also some possibilities to re-use information from one round to another (further possibilities in re-using information is one of the highest priorities for future development). When the process area concept is expanded to other processes, further areas like 'processing round' and 'publication round' are envisioned.

When it comes to providing each survey with the information that it needs, based on the properties of the survey, this is what has been done:

1. Based on the information in the 'ordinary' PSS, a number of activities have been created. The activities are the tasks that have to be carried out for the production process (data collection and micro editing) to be successful. Each activity has been given a description (the specific small part of the ordinary PSS that covers that specific activity). If necessary, a document (such as a template, a checklist or a more detailed instruction) is attached to the activity. If there are different documents for different types of surveys (for example different templates to be used for an enterprise survey and a household survey) two different but similar activities have been created, with the relevant template attached to each.
2. A form (it can also be described as a questionnaire) to be filled in by the survey that wants to use a process area has been created. The survey fills in its properties and design choices (at a fairly high level).
3. Each activity has been connected to the form in a simple format (*i.e.*, if the answer to question x is y, then display this activity). The statement can be made more complicated using and/or statements.
4. The survey creates a process area with its survey rounds and collection rounds, and fills in the design and properties form. The form is interpreted by the underlying engine of the process areas and Triton, and the relevant activities are copied to the new collection round area. This means that the survey now has all the relevant information and documents necessary for its collection and micro editing activities.
5. The activities can be assigned to a specific person, and they can also be ticked off when they are finished. This gives the production manager a good possibility to both steer and monitor the activities. The activities are an important feature of the dashboard functionality of the process areas.

The dashboard functionality of the process areas also include a number of other functionalities that are important in driving the survey (these are described further in section 4):

- Possibility to assign persons to different roles in the production and to give them access to the survey and its tools.
- Defining and describing collection variables.
- Access to paradata reports on the collection process et cetera.
- Access to the necessary tools for setting up data collection and micro editing.

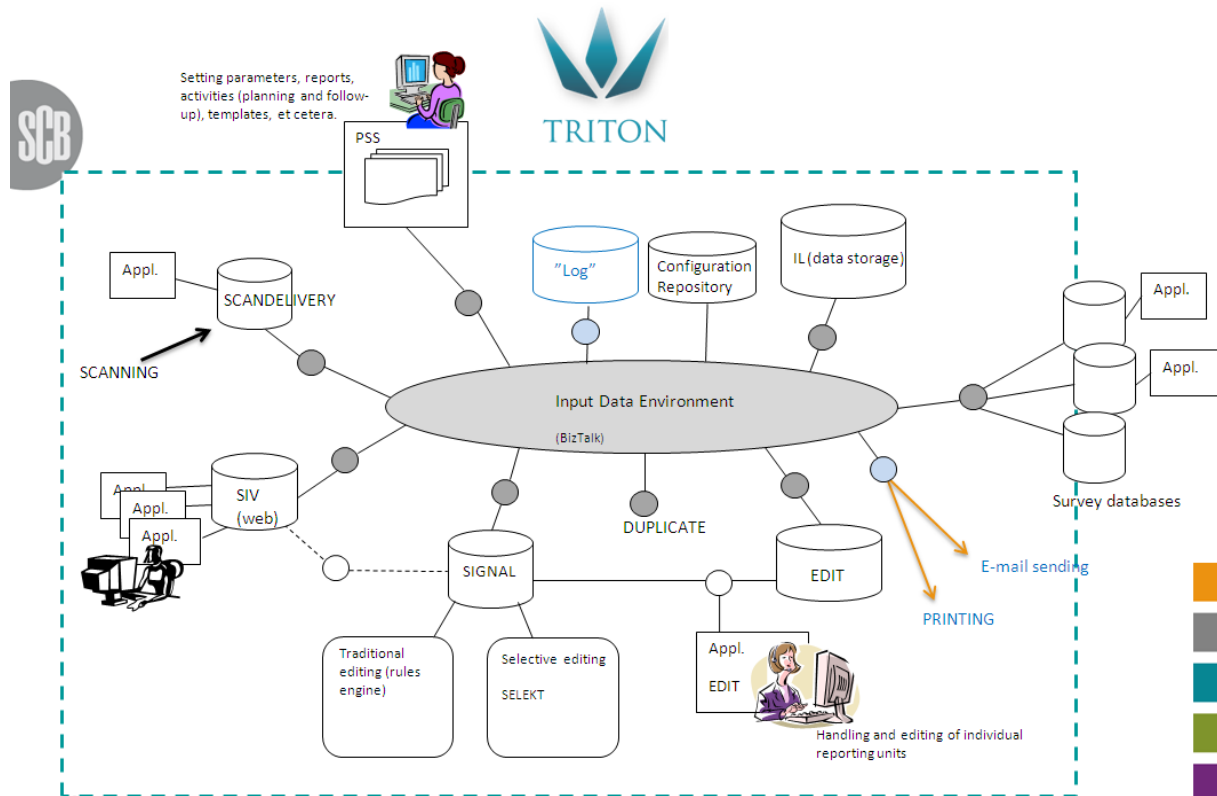
4. The Triton platform

As shown in section 3, the process areas within the PSS are the starting point, the dashboard, for surveys using the Triton platform. But the platform is much more than just the process areas; it is more than just a production system or a number of tools. It is the whole concept of working in an integrated process oriented production environment that can be called Triton. However, for the staff at Statistics Sweden who are more interested in having a functional production system than worrying about architectural principles or grand concepts, it is specifically a production system that integrates some tools that already existed with some new tools into an intelligent whole.

4.1 The different parts of Triton: An overview

The different parts included in the platform and how they are connected are shown in the picture below:

Figure 4.1.1
The Triton platform: An overview



- The centre (or the heart) of the platform is the communication platform itself, the Input Data Environment (IDE) which is based on Microsoft BizTalk. This part distributes all information and data to and from the other parts of the platform.
- The PSS part at the top is the process areas described in section 3.
- The configuration repository is the part where all the information about the rules on how to transport data within the platform based on the choices made is stored.
- The IL (data storage) is the database where the actual collected data is stored as long as it is treated within the platform.
- SIV and SCANDELIVERY are the systems for the two collection tools connected to the platform at the moment; web data collection and optical reading of paper questionnaires. Here, the web questionnaire and the electronic format of the paper questionnaire are defined. The systems have databases where they store the data that is transmitted from respondents. From these databases, data is transported into the platform for storage in IL.
- SIGNAL and DUPLICATE are two services that handle error signalling and treatment of duplicates. SIGNAL can use both traditional editing and selective editing using the SELEKT tool.
- EDIT is the interface where the collection staff handles the reporting units, make reminders and investigate possible errors.
- The survey databases are the databases where surveys carry out activities before and after data collection and micro editing. As long as the platform only covers these processes, the survey databases are still necessary.

4.2 Working in the Triton platform: Step by step

This is a short overview of what a survey does when it does its collection and micro editing using Triton:

1. Define the survey, its survey rounds and collection rounds. This is done in the process area in the PSS and is described in section 3 above.
2. Define persons working on the survey and their roles. This is done within the process area of PSS. At the moment, Statistics Sweden is making an in-depth review of the handling of authorisation, so this functionality is very basic at the moment.
3. Load sub-systems. When the parameters are sent from the survey's process area, the information is sent to the relevant tools that there will be a collection round and which tools it will use. The data flows based on the design choices are stored in the configuration repository.
4. Define variables. There is a specific tool accessible from the process area for doing this. The survey defines, describes and names the variables to be used in collection. If the survey is ongoing and there is already a web questionnaire, the variables from that can be imported into the variable defining tool and used as a starting point. The information is sent to the platform and stored in the configuration repository.
5. Build collection instrument. This is still done in the SIV system (for web forms) and in Microsoft Word or Crystal Reports for paper questionnaires. Paper questionnaires are also defined electronically in the scanning system. In this step, each cell that is defined in either of these systems is connected to a variable defined in step 3. No other variable names can be used than the ones defined.
6. Define edit checks. Editing can be made at several stages in the production process, both at the respondent side while filling in a questionnaire (applicable only for web questionnaires, not paper ones) and afterwards, when data reaches Statistics Sweden. For defining edit checks at the respondent side, there was already an "edit check builder" within the SIV system. This has been expanded so that all edit checks using traditional editing are defined there, and a specific parameter is set for each edit check, whether to run at the respondent side, afterwards, or both. If a survey wants to use selective editing, all editing parameters for this are set within SELEKT. All information about edit checks are sent to the platform and stored in the configuration repository.
7. Define value sets. For questions which use a fixed set of response alternatives, these alternatives are defined and sent to the platform. At the moment, there is a lot of manual work (cutting and pasting in Microsoft Excel) involved in this step, future developments envisioned are re-using response alternatives defined in the web questionnaire (for small value sets) and connecting to the central metadata storage for larger classifications.
8. Configure EDIT settings. The interface for manual investigation of possible errors and data collection activities is set up according to the survey needs. This is done within the EDIT tool, using the variables defined in step 3 above.
9. Load sample and background information. There is an interface within the process area for defining when the sample is to be collected from the survey database, where it is set up at the moment. Future development will include direct access from sampling tools to the platform. Background information (such as register information, information on contact persons, data from previous rounds et cetera) is also defined and loaded into the platform.
10. The survey is sent out. At the moment, this is done outside the platform, but a future development will be a connection to the printing and sending functions, a graphic design on how this is going to work has been created but the functionality is not yet implemented in the platform.
11. Respondents send data. The architecture of the platform is based on treating each reporting unit separately, so as soon as a respondent sends data to Statistics Sweden, the platform will notice this and get the data from the web collection or scanning database and store it in the IL.
12. Treatment of duplicates. The rule for how to treat duplicates is set by the survey in the form in step 1 above, the rule is applied automatically whenever a set of data is received where there is already one. At the moment, three simple rules are available: choose the first one, choose the last one or decide manually. A graphic interface for choosing manually is planned but not yet implemented.
13. Edit data. If the survey has editing, the incoming data is sent to the SIGNAL service which runs the pre-defined edit-checks and presents a result. If there are no errors, the data is transferred to the survey database, which means that the data set in the survey database is filled with data for incoming reporting units gradually, as they pass the editing stage. If there are possible errors that have to be treated manually, the data

is sent to the EDIT interface, where Statistics Sweden staff can investigate the possible errors. Reporting units appear in EDIT gradually too, as they send data and are flagged as having possible errors by SIGNAL.

14. Reports on the status of the survey, the number of incoming reporting units by stratum et cetera, is gathered in paradata reports and sent daily to the process area. At the moment, four pre-defined status reports have been prepared, and after migrating to Sharepoint 2010 in February 2012, this functionality will be implemented. The number of paradata reports and the coverage are expected to increase rapidly over time.

4.3 The future: What happens next?

The Triton platform including the process areas in the PSS was released in the version described in sections 3 and 4 at the end of June 2011. After summer, the first stage of implementation in surveys started. In the autumn of 2011, six surveys have used the new version of the system with good results. As always, there have been some things that have had to be fixed that became apparent only when the system was used in practice. Some minor improvements to existing functionality have been added, but no new major parts. For 2012, there is a plan to implement around 20 additional surveys in the system. To cope with this, specific implementation teams will be set up to aid the surveys when they begin using the system. These teams contain both IT support and support from the data collection departments at Statistics Sweden. Besides this, a maintenance group has been set up, together with routines for handling reporting incidents (parts of the platform not working) and suggestions for improvements.

Besides maintenance and implementation, broadening and deepening of the platform is also planned. This will be done in several steps, but two projects are already now on their way. The first will be dedicated to improving the existing parts of the platform, expanding the functionality as described in the different points in section 4. The other will look at the platform itself and the process areas, and how they can be expanded to cover other sub-processes in the statistical production process. Both these projects will run in 2012.

5. Conclusions

The Triton project has been a successful project in taking the first step in moving Statistics Sweden to a process-based production environment. There has been significant progress by developing a generalised tool for data collection and micro editing, while also developing the concept of process areas for distributing information and standards to the surveys using the platform. The platform for data collection and micro editing has integrated existing and new tools into a coherent and intelligent whole. Six surveys have implemented the existing version and 20 more will do so in 2012. There are still many steps to take before a full process-oriented production platform is developed, and functionality is still missing from the existing version. But the Triton project shows much promise in the area. Further projects to both broaden and deepen the contents of the platform are already in the planning stage.

Statistics New Zealand’s standard methodology toolbox

John Lopdell and Gary Dunnet¹

Abstract

Statistics New Zealand has recently developed a “standardisation roadmap”, which provides a pathway towards increased standardisation of our methods, processes, data management and technology. In response to the roadmap, we have developed a Standard Methodology Toolbox (SMT). The SMT provides a definitive and validated list of the methodological tools that are used in the statistical process. The SMT aims to reduce costs and gain efficiencies by promoting greater use of standard tools across collections, infrastructure and platforms. A streamlined and complementary set of tools can be expected to decrease the level of maintenance, support and training, and increase capacity. It also facilitates the use of best practice methods. The SMT incorporates information on tool ownership and versioning, and provides a mechanism for managing the status of each tool (such as discovery, current, or legacy), as well as being a central point of reference for documentation of tools.

This paper provides an overview of the SMT, and how it contributes to Statistics New Zealand’s standardised generic business processes model (gBPM). It also describes upcoming work to further develop the toolbox.

¹John Lopdell and Gary Dunnet, Statistics New Zealand, New Zealand.

SESSION 11B
MODELING AND ESTIMATION

Model-based and semi-parametric estimation of time series components and mean square error of estimators

Michail Sverchkov, Richard Tiller and Danny Pfeffermann¹

Abstract

This paper will focus on time-series analysis and, more specifically, estimation of seasonally adjusted and trend components and the mean square error (MSE) of the estimators. It will compare the component estimators obtained by application of the X-11 ARIMA method with estimators obtained by fitting state-space models that account more directly for correlated sampling errors. The component estimators and MSE estimators are obtained under a different definition of the target components. By this definition, the unknown components are defined to be the X-11 estimates of them in the absence of sampling errors and if the time series under consideration is long enough for application of the symmetric filters imbedded in this procedure. New MSE estimators are proposed with respect to this definition. The performance of the estimators is assessed by using simulated series that approximate a real series produced by the Bureau of Labor Statistics in the U.S.A.

¹Michail Sverchkov and Richard Tiller, Bureau of Labor Statistics, U.S.A; Danny Pfeffermann, Hebrew University of Jerusalem, Israel, and University of Southampton, U.K.

Challenges and issues in weighting the Travel Survey of Residents of Canada

Félix Labrecque-Synnott¹

Abstract

This paper is about the weighting of the recently redesigned Travel Survey of Residents of Canada (TSRC). We begin with an overview of the survey and the concepts behind the weighting methodology and then take a closer look at the various weighting adjustment factors. Most of those factors also play a role in the weighting methods used in many household surveys at Statistics Canada. The TSRC has some special complexities, such as multiple units of analysis (trips, persons and person-trips), the rostering and subsampling of trips, and a two-month recall for overnight trips.

Key Words: Weighting; Non-response; Household survey; Response propensity; Calibration.

1. Introduction

This paper describes the weighting system used for the Travel Survey of Residents of Canada (TSRC). The system was revised following a redesign of the survey in 2011. As the survey's title suggests, the focus is on travel, therefore both trips and respondents must be weighted. Following an overview and a brief explanation of the survey's background, we provide a detailed description of trip and respondent weighting. We conclude with a discussion of the various challenges encountered throughout the weighting process, most of which are due to the TSRC's particular features and require some flexibility in our weighting approach.

2. Overview of the TSRC

2.1 Survey background and overview

The TSRC, a monthly supplement to the Labour Force Survey (LFS), deals with the travel habits and travel spending of Canadian residents within Canada. The LFS, Statistics Canada's flagship household survey, has about 55,000 respondent households per month (Statistics Canada, 1998). Each selected household is interviewed six times (once a month for six months). The LFS sample is divided into six rotation groups, and each month, one rotation group is removed from the sample and replaced with a new group. The survey's sample design is complex and includes stratification and multiple stages (Statistics Canada, 1998).

Each month, one adult from each household in the rotation group that has just completed the second of its six LFS interviews is selected at random to take the TSRC. About 9,100 people are asked each month to participate in the survey.

As the survey's title suggests, users are most interested in travellers, their trips and their travel spending. Non-traveller respondents are needed to obtain travel incidence rates, but they provide comparatively little information. Unfortunately, they make up the majority of the sample. The proportion of travellers varies from month to month (not surprisingly, it peaks in July and August), but it is nearly always between 25% and 35%.

Historically, the TSRC's sample was roughly twice its current size. Until 2009, households were asked to participate in the TSRC after completing the second and sixth LFS interviews. This sample reduction was particularly

¹Félix Labrecque-Synnott, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Canada, K1A 0T6 (felix.labrecque-synnott@statcan.gc.ca).

problematic since users want estimates of travel spending, trip volumes and traveller activities in tourist areas. The cut made it necessary to produce such estimates using only a subset of the sample.

To address this challenge and increase the number of trips and travellers in the sample, the survey was redesigned. The new version of the survey has been collecting data since January 2011.

2.2 The 2011 redesign

The 2011 redesign had two main goals: increase the number of trips collected and limit the length of the interviews. Increasing the number of trips collected improves the precision of the estimates of trip volumes and travel spending. It would also be useful to increase the proportion of respondents reporting one or more trips in order to get maximum benefit from the work of contacting respondents. Since the TSRC is an LFS supplement, the interview can be no more than 15 minutes long. It is also desirable to limit the interview's length to avoid increasing the response burden. While the majority of respondents do not travel and the majority of travellers report only one trip, a substantial minority travel frequently and are able to provide particularly useful data. It would be a pity to lose those respondents because the interview is too long.

One of the key changes made in the redesign is the addition of a second recall month for overnight trips. In the November collection period, for example, respondents are asked about overnight trips they took that ended in October or September and same-day trips that ended in October. This results in the collection of more trips and increases the proportion of "productive" interviews, *i.e.*, interviews with at least one trip reported.

To limit interview length, instead of collecting detailed information about all trips reported, we now ask respondents to list the trips that ended in the reference months and provide basic information about them (the main reason for the trip, the principal means of transportation used, the trip's duration and the number of adults on the trip). This information is used to determine whether the trips are in scope, and if so, to weight them. For each respondent, the trip roster serves as a sampling frame from which to select one or more trips for more detailed questions (municipalities visited during the trip, accommodations, detailed expenditures and activities). Thus, while the addition of a second recall month increases the number of trips collected and the proportion of traveller respondents, the rostering and sampling of trips help control interview length.

TSRC users receive two files: a person-level file, used to derive travel rates and determine the demographic profiles of travellers and non-travellers; and a trip-level file, used to produce estimates of spending and activities. Since detailed information is not collected for all rostered trips, two options are currently under consideration for the trip file: weight the selected trips to reflect the subsampling and include only those trips, or include all the rostered trips and impute the details of trips that are not selected.

3. Weighting

3.1 Weighting overview

Because of the TSRC's structure, multiple sets of weights have to be produced to accompany the files provided to users. In addition to person weights, used to calculate travel incidence rates, two sets of weights have to accompany the trip data: person-trip weights, used to estimate trip volumes and characteristics; and household-trip weights, needed to estimate travel spending.

The addition of a second recall month for overnight trips also has a substantial impact on the weighting system. In the redesigned TSRC, the data collected in a given month relate to two reference months. For example, respondents surveyed in November are asked about trips that ended in September and October. Conversely, there are two collection months for each reference month. Hence, the data collected in October (first recall month) and November (second recall month) are needed to produce tourism estimates for September. However, since the second recall month applies only to overnight trips, the definition of a "traveller" is different for the two recall months.

Consequently, two person-level files (and two sets of weights) are needed. One file, containing only respondents for the first recall month, is used to calculate “overall” travel rates (including same-day and overnight trips). The other file contains both recall months and is used when the client is interested only in overnight trips. The two sets of weights are necessary because the trip weights for same-day and overnight trips are based on different person weights.

3.2 Weighting of persons

In the weighting process for a given reference month, the two collection months are first processed separately. The data for the two months are weighted as described in this section and calibrated so that both months are representative of the Canadian population in the reference month. To produce a complete file containing all respondents interviewed about the reference month, the two files are combined and the person weights are divided by two. Hence, the full file also represents the Canadian population in the reference month. The full file and the file for the first collection month are both used to produce estimates and trip weights.

The starting point of the TSRC’s weighting system, and its primary unit of analysis, is the respondent, an adult selected at random from the members of the LFS respondent household. The base weight used by the TSRC is the LFS subweight, which reflects the LFS sample design and non-response. Next we apply adjustment factors for the selection of one of the six LFS rotation groups and for the random selection of one adult member of the household.

Then we apply an adjustment factor for TSRC non-response. To that end, the response propensity of each individual in the sample is modelled with a logistic regression model based on paradata and demographic information from the LFS. Homogeneous response propensity classes are created, and within each class, respondent weights are adjusted with the inverse of the weighted observed response rate. This type of non-response adjustment is mentioned frequently in the scientific literature and is used in many surveys (Haziza and Beaumont, 2007; Little, 1986; Eltinge and Yansaneh, 1997).

A particular characteristic of the TSRC is the presence of non-respondent travellers, individuals who reported at least one in-scope trip but provided incomplete data and therefore cannot be considered respondents. A special adjustment is made for those non-respondents: their weight is redistributed among respondent travellers within classes based on age, sex and province of residence.

Lastly, calibration by age group, sex and census metropolitan area is performed to ensure that the weight totals align with census-based population estimates. This too is a proven method (Deville and Särndal, 1992).

3.3 Weighting of trips

To weight the rostered trips, adjustment factors are applied to the respondent’s person weight. The base weight is the person weight for the first collection month for same-day trips and the person weight from the full file for overnight trips. This distinction is necessary to ensure that same-day trips are not systematically under-represented in the final weighting (and in turn underestimated in analyses). Respondents are asked about same-day trips in only one of the two collection months. Since the weights in the full file are weights for both collection months divided by two, the distinction adjusts for the fact that overnight trips have the opportunity to be reported by twice as many respondents as same-day trips.

The first adjustment involves multiplying the weight by the number of trips that are identical to the rostered trip. In the rostering process, for each trip reported, respondents can specify how many identical trips they have taken for each rostered trip. The identical trips are not listed separately, which helps limit the length of the interview and mitigate the response burden for respondents who travel frequently (for example, people who go to their cottage every weekend). Thus, multiplication of the trip weight reflects that fact that the one rostered trip actually represents other, unrostered trips.

Another adjustment factor is applied to correct for disparities between the number of rostered trips and the number of trips reported at the beginning of the interview.

Some trips that would normally be in-scope for the survey have missing values for certain non-imputed variables. For example, a tourism trip whose main destination is in Canada but unknown would fall into this category. If a respondent lists at least one trip of this type and no valid trips, he or she is considered a non-respondent traveller. If a respondent lists a trip of this type plus at least one valid trip, the weight of the valid trip or trips is adjusted to offset the in-scope trip with missing data.

Non-response may also be present at the trip level. Some trips selected for the more detailed questions are missing from the detailed files. Those trips are not imputed and do not appear in the files provided to users; they are removed from the trip roster. If all of a respondent's selected trips are missing, the respondent is considered a non-respondent traveller. If not, an adjustment is applied to the weights of the rostered trips (selected and non-selected trips) in the same non-response class. Non-response classes are based on trip duration and province of origin and destination.

With the weights of the rostered trips, we can produce weights for the selected trips. No matter which option is chosen for the files provided to users (only the selected trips, or the selected trips and imputed trips), those weights will be necessary: in one case, as weights accompanying the selected trips, and in the other, to validate the imputation system.

Those weights are obtained by multiplying the rostered weights by a factor that adjusts for trip subsampling. The trips are selected by sampling with unequal probability, and the selection probabilities depend on the number of identical trips, intraprovincial or interprovincial status, and trip duration. The weights obtained are then calibrated on the totals of the weights of the rostered trips using province of destination, interprovincial status and number of nights as calibration variables.

Lastly, (household) trip weights (rostered and selected) are calculated by dividing the person-trip weights (rostered and selected) for each trip by the number of adults in the household who went on the trip. The person-trip weights are used to estimate trip volumes and characteristics, and the (household) trip weights are used to estimate travel spending.

4. Conclusion

In summary, some of the TSRC's characteristics made the development of the weighting system more complicated. The survey has various units of analysis and requires the production of several different sets of weights: person weights for the first collection month, person weights for the full sample, person-trip weights for the rostered and selected trips, and (household) trip weights for the rostered and selected trips. The various adjustment factors applied during weighting are summarized in Table 4-1.

Non-response takes a number of forms in the survey: refusal and no contact, respondents with incomplete trip rosters, selected trips missing, and non-respondent travellers. The latter constitute a person-level response status determined by the data collected at the trip level. Hence there is a relationship between the various units of analysis present in the TSRC.

Table 4-1
Adjustment factors required to produce the TSRC weights

Weight	Adjustment factor
Person	Rotation group, adults in household, non-response (persons), non-response (travellers), calibration
Person-trip (rostered)	Identical trips, initial number reported, missing non-imputed variables, non-response (trips)
Person-trip (selected)	Selection probability, calibration
Trip	Adult members of household on the trip

Having a second recall month for some trips only means that the samples and weights of two successive collections must be combined. This results in a duplication of response statuses, as an individual can be respondent for one reference month and non-respondent for the next month.

To overcome these challenges, common methods are applied throughout the weighting process. For example, logistic regression is used to model response propensity; non-response is addressed with homogeneous response classes; and calibration is used to ensure consistency between the weights of rostered and selected trips and agreement between person weights and population estimates.

References

- Deville, J.-C. and C.-E., Särndal (1992), "Calibration estimators in survey sampling", *Journal of the American statistical association*, 87, p. 376-382.
- Eltinge, J.L. and I.S., Yansaneh (1997), "Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. expenditure survey", *Survey Methodology*, 23, p. 33-40.
- Haziza, D. and J.-F., Beaumont (2007), "On the construction of imputation classes in surveys", *International Statistical Review*, 75, p. 25-43.
- Little, R.J.A. (1986), "Survey nonresponse adjustments for estimates of means", *International Statistical Review*, 54, p. 139-157.
- Statistics Canada (1998), "*Methodology of the Canadian Labour Force Survey*", Catalogue N° 71-526-X.

Statistical methods for evaluating secular trends using estimates from annual independent cross-sectional complex probability sample surveys

Philip J. Smith and Zhen Zhao¹

Abstract

When data are available from cross-sectional complex probability sample surveys that are conducted annually, trends are often analyzed by using regression methods that pool the data across the survey years. In this case, the estimated slope from the regression is used to summarize the trend across the years, and the standard error of the estimated slope is estimated from all of the observations from all survey years. This method fails to recognize that (i) only the annual estimates are relevant in estimating the trend and (ii) the entire sample across survey years is not relevant for estimating the precision (*i.e.*, the standard error) of the estimated trend. Moreover, the sample for a survey year only contributes to the estimation of the precision of survey outcomes, and not the precision of slope of the regression. Other methods use the annual estimate of survey outcomes regressed on time. These approaches ignore the uncertainty of the survey estimates in assessing statistical significance of the slope.

The method we propose accounts for the uncertainty of annual estimates of survey outcomes and recognize that the entire sample across survey years is not relevant for estimating the precision of the estimated trend. Our method consists of three steps: (i) bootstrapping the regression of bootstrap replicate estimates of the annual survey outcomes on the permuted survey year values to obtain the distribution of the estimated slope under the null hypothesis of no relation between the survey outcome and survey year; (ii) bootstrapping the regression of bootstrap replicate estimates on the survey year values to obtain the distribution of the slope under the alternative hypothesis of a non-zero relation between the outcome and survey year; and (iii) using the Wilcoxon statistic to test whether the distributions of the estimated slope under the null and alternative are significantly different. Within this framework, the bootstrap replicate estimates account for uncertainty in the estimated survey outcome, the survey weights, and the complex probability sampling design; the use of the bootstrap to conduct the permutation tests generate the distribution of the slope under the null and alternative hypotheses; and the Wilcoxon statistic provides a method for detecting significant differences between these two distributions. This method is easily extended to account for multiple change points in the secular trend over the range of survey years.

¹Philip J. Smith and Zhen Zhao, Centers for Disease Control and Prevention, USA.