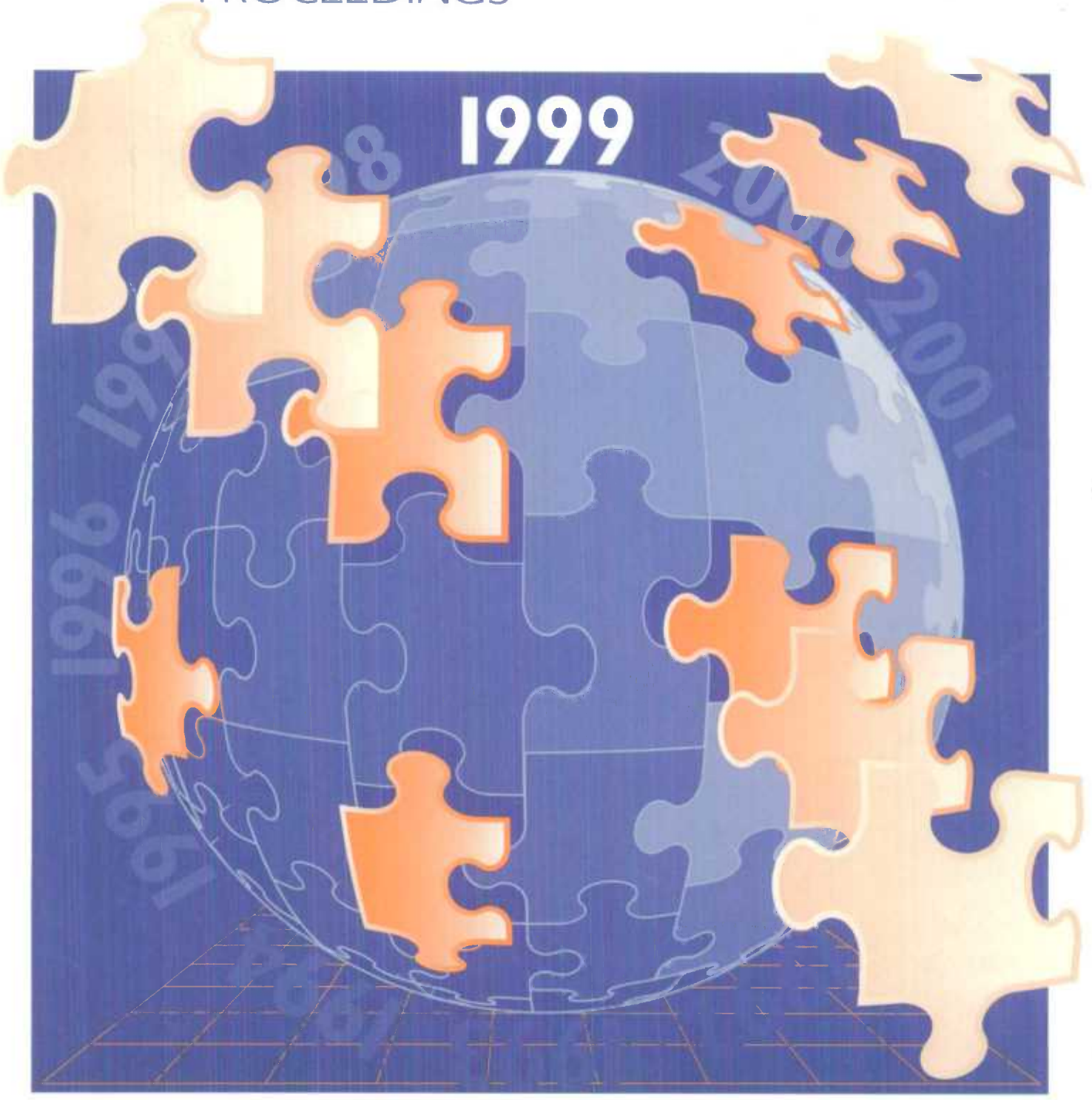
C.3

Catalogue no. 11-522-XPE

SYMPOSIUM 99

Combining Data from Different Sources

PROCEEDINGS



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: (613) 951-8615).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our Web site.

National inquiries line	**1 800 263-1136**
National telecommunications device for the hearing impaired	**1 800 363-7629**
Depository Services Program inquiries	**1 800 700-1033**
Fax line for Depository Services Program	**1 800 889-9734**
E-mail inquiries	**infostats@statcan.ca**
Web site	**www.statcan.ca**

Ordering and subscription information

This product, Catalogue no. 11-522-XCB, is published annually on a CD-ROM at a price of CDN $60.00. The following additional shipping charges apply for delivery outside Canada:

	Single issue
United States	CDN $6.00
Other countries	CDN $10.00

All prices exclude sales taxes.

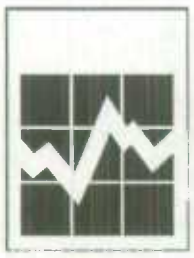
This product can be ordered by

- Phone (Canada and United States) **1 800 267-6677**
- Fax (Canada and United States) **1 800 287-4369**
- E-mail **order@statcan.ca**
- Mail Statistics Canada
  Dissemination Division
  Circulation Management
  120 Parkdale Avenue
  Ottawa, Ontario K1A 0T6
- And, in person at the Statistics Canada Regional Centre nearest you, or from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136.



Statistics Canada
Methodology Branch

SYMPOSIUM 99

Combining Data from Different Sources

Proceedings

Published by authority of the Minister responsible for Statistics Canada

October 2000

Catalogue no. 11-522-XPE
ISBN 0-660-18174-6

Frequency : Annual

Catalogue no. 11-522-XCB
ISSN 1481-9678

Frequency : Annual

Ottawa

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Canadian Cataloguing in Publication Data

Symposium 99, Combining Data from Different Sources
(1999 : Ottawa, Ontario)
Symposium 99, Combining Data from Different Sources: proceedings

CR-ROM bilingual version also issued under title:
Symposium 99, Combining Data from Different Sources: proceedings
ISSN 1481-9678
CS11-522-XCB

Paper version issued also in French under title:
Symposium 99, Combiner des données de sources différentes: recueil.
ISBN 0-660-96364-7
CS11-522-XPF

1. Combining data – Congresses. 2. Statistics –
Methodology – Statistics. I. Statistics Canada. Methodology
Branch. II. Title.

HA12 S95 1999 300.7'27
C99-988034-9

PREFACE

Symposium 99 was the sixteenth in the series of international symposia on methodological issues sponsored by Statistics Canada. Each year the symposium focuses on a particular theme. In 1999, the theme was: "Combining Data from Different Sources".

The 1999 symposium was held from May 5 to May 7 1999 in the Simon Goldberg Conference Centre in Ottawa and it attracted over 300 people from 11 countries. A total of 29 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 99 would like to acknowledge the contribution of the many people, too numerous to mention individually, who helped make it a success. Over 30 people volunteered to help in the preparation and running of the symposium, and 22 more verified the translation of papers submitted for this volume. Naturally, the organizers would also like to thank the presenters and authors for their presentations and for putting them in written form. Finally, they would like to thank Lynn Savage for processing this manuscript.

In 2000, the symposium will be replaced by the *International Conference on Establishment Surveys,* as was the case in 1993. This conference will be held in Buffalo, New York from June 17 to June 21 2000. The next Statistics Canada symposium will be held in Ottawa in 2001.

Symposium 99 Organizing Committee

Christian Thibault
Jean-Marie Berthelot
Milorad Kovacevic
Pierre Lavallée
Jackie Yiptong

STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data	**1993** - International Conference on Establishment Surveys
1985 - Small Area Statistics	**1994** - Re-engineering for Statistical Agencies
1986 - Missing Data in Surveys	**1995** - From Data to Information: Methods and Systems
1987 - Statistical Uses of Administrative Data	**1996** - Nonsampling Errors
1988 - The Impact of High Technology on Survey Taking	**1997** - New Directions in Surveys and Censuses
1989 - Analysis of Data in Time	**1998** - Longitudinal Analysis for Complex Surveys
1990 - Measurement and Improvement of Data Quality	**1999** - Combining Data from Different Sources
1991 - Spatial Issues in Statistics	**2000**- International Conference on Establishment Surveys II
1992 - Design and Analysis of Longitudinal Surveys	

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

Use this two page order form to order additional copies of the proceedings of Symposium 99: Combining Data from Different Sources. You may also order proceedings from previous Symposia. Return the completed form to:

SYMPOSIUM 99 PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 6th FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada" - Indicate on cheque or money order: Symposium 99 - Proceedings Canada).

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	_____ @ $10
1987 -	Les utilisations statistiques des données administratives - FRENCH	_____ @ $10
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	_____ @ $12
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	_____ @ $10
1989 -	Analysis of Data in Time - BILINGUAL	_____ @ $15
1990 -	Measurement and Improvement of Data Quality - ENGLISH	_____ @ $18
1990 -	Mesure et amélioration de la qualité des données - FRENCH	_____ @ $18
1991 -	Spatial Issues in Statistics - ENGLISH	_____ @ $20
1991 -	Questions spatiales liées aux statistiques - FRENCH	_____ @ $20
1992 -	Design and Analysis of Longitudinal Surveys - ENGLISH	_____ @ $22
1992 -	Conception et analyse des enquêtes longitudinales - FRENCH	_____ @ $22
1993 -	International Conference on Establishment Surveys - ENGLISH (available in English only, published in U.S.A.)	_____ @ $58
1994 -	Re-engineering for Statistical Agencies - ENGLISH	_____ @ $53
1994 -	Restructuration pour les organismes de statistique - FRENCH	_____ @ $53
1995 -	From Data to Information - Methods and Systems - ENGLISH	_____ @ $53
1995 -	Des données à l'information - Méthodes et systèmes - FRENCH	_____ @ $53
1996 -	Nonsampling Errors - ENGLISH	_____ @ $55
1996 -	Erreurs non dues à l'échantillonnage - FRENCH	_____ @ $55
1997 -	New Directions in Surveys and Censuses - ENGLISH	_____ @ $80
1997 -	Nouvelles orientations pour les enquêtes et les recensements - FRENCH	_____ @ $80
1998 -	Longitudinal Analysis for Complex Surveys - ENGLISH	_____ @ $60
1998 -	L'analyse longitudinale pour les enquêtes complexes - FRENCH	_____ @ $60
1998 -	Longitudinal Analysis for Complex Surveys -BILINGUAL on CD-ROM	_____ @ $60
1999 -	Combining Data from Different Sources - ENGLISH	_____ @ $60
1999 -	Combiner des données de sources différentes - FRENCH	_____ @ $60
1999 -	Combining Data from Different Sources - BILINGUAL on CR-ROM	_____ @ $60
	PLEASE ADD THE GOODS AND SERVICES TAX (7%) (Residents of Canada only)	$_____
	TOTAL AMOUNT OF ORDER	$_____

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME _____

ADDRESS _____

CITY _____

PROV/STATE _____

COUNTRY _____

POSTAL CODE _____

TELEPHONE (____) _____

FAX (____) _____

For more information please contact John Kovar: Telephone (613) 951-8615, Facsimile (613) 951-5711.

COMBINING DATA FROM DIFFERENT SOURCES

TABLE OF CONTENTS

SESSION III: THE USE OF META ANALYTICAL TECHNIQUES IN POPULATION HEALTH RISK ASSESSMENT

Chairperson: D. Krewski

SESSION IV: RECORD LINKAGE

Chairperson: M. Fair

SESSION V: STATISTICAL MATCHING

Chairperson: J.-L. Tambay

OPENING REMARKS

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

OPENING REMARKS

Gordon Brackstone, Statistics Canada

Good morning and, on behalf of Statistics Canada, a very warm welcome to Symposium 99. This is the 16th in the series of annual methodology symposia organized by Statistics Canada. We are pleased to see a strong international character to the Symposium again this year with speakers from Australia, France, Germany, the Netherlands, New Zealand, the United Kingdom, and the United States. Altogether we have participants from 11 countries on 3 continents. We also have participants from a wide range of institutions including national statistical agencies, universities, industry, and several organizations in the health field.

The Statistics Canada Methodology Symposium series started in 1984 as an annual international event aimed at addressing issues of statistical methodology - issues which were, on the one hand, relevant to the work of a government statistical agency, but which would, on the other hand, benefit from exposure to a broader spectrum of interested individuals working in other sectors including universities, industry, and other government organizations. A wide range of topics has been covered at previous symposia. They have included specific topics within survey methodology: design of establishment surveys, longitudinal surveys; dealing with missing data; non-sampling errors. But they have also included broader topics such as the impact of high technology on survey taking; re-engineering; and quality improvement.

The theme of this year's symposium is: combining data from different sources. The program of this symposium is, to say the least, impressive, and will be looking at the problems and challenges, new theoretical developments and their practical applications, and a range of other subjects related to the need to combine data from different sources. The symposium will also touch on concerns about privacy protection and confidentiality when combining data.

Several of our symposia have broached subjects related to the theme of this symposium - for example, *Statistical Use of Administrative Data* in 1987, and *Data to Information - Methods and Systems* in 1995, but we have never before focussed expressly on this theme. Allow me to explain briefly the reason for choosing this subject, which has evidently drawn so much interest.

Combining data from different sources is not a new idea. Compiling the National Accounts, a task which has been around for many decades now, is an example *par excellence* of combining data from many different sources. Producing population estimates between censuses is another old example of utilizing many sources of data to produce estimates. In fact, several countries are now using several sources of information to produce population estimates in census years too - and our keynote talk will elaborate on that. So what's new about Combining data from different sources?

Well, I think there has been a quite significant change, over the past two decades, in the way we think of a survey, or for that matter any other data collection exercise, as a source of information. The traditional view was perhaps that each survey produced a set of outputs - publications, tables, special requests, maybe a public use microdata file. Each survey was a stovepipe (to overuse a cliché) that transformed a set of responses into a set of statistical outputs and spewed them out separately from every other survey. And that was the end of the survey. Some systems, such as the national accounts, existed for taking aggregate results from surveys and integrating them into a coherent picture of broader phenomena than those measured by a single survey. But the survey microdata themselves were rarely seen as an asset to be used in combination with other datasets.

Today there is much more emphasis on the survey as a supplier of microdata to be used for further analysis in combination with data from other sources. The survey - and I use the term "survey" to denote any data collection exercise - is seen as a feeder system that will add to an existing stock of data that can be analysed more broadly, or that will add data within a framework that integrates information pertaining to a common aspect of society or the economy. I do not claim that we have made that transition successfully yet, but that is the direction of change.

Integration of the outputs of different surveys into a common analytic framework is one example of *Combining data from different sources*. This integration may be aimed at reconciling different datasets to produce improved information, or at obtaining deeper analytic insights than can be obtained from any of the datasets individually, or at assessing the quality of one source of data against another. The techniques that are now referred to as meta analysis also fit in here.

These are examples of combining **aggregate** data from different surveys after the surveys are completed. But there are also many examples of combining **microdata** from different sources, both within and between surveys. The use of record linkage techniques to combine data about the same individual from different sources can provide a richer database than any of the individual sources. Or it can provide a basis for assessing one source against another. Record linkage raises some important privacy issues, as well as technical challenges, some of which will be covered during this Symposium. Different aspects of record linkage, statistical matching (or data fusion as it is called in some places), and imputation techniques will be tackled, including comparisons of existing methods and related software. Linking data from different surveys is a special challenge in the presence of survey weights. Several speakers will address this problem.

In addition to direct record linkage where we are trying to find matches between records that relate to the same individual, there are more recent methods of statistical linkage or synthetic linkage where we try to combine records that typically relate to different individuals, but in a way that results in a set of composite records that in total reflect the statistical properties of the population under study. The statistical properties of data resulting from such linkages deserve attention.

So we have examples of combining data from surveys both at the micro level and at the macro or aggregate level.

But there are also many examples of combining data from different sources during the design and execution of surveys. Multiple frames; substitution of administrative data for survey responses; use of auxiliary data in imputation or estimation; assessment of survey data, - these all involve combining data from different sources in different ways.

Most of these examples are driven by one or both of two complementary motivations: to derive maximum information from data already collected, and to optimize the use of resources spent on new data collection

The theme of this year's symposium is very timely for Statistics Canada, since there is a growing demand for and an increasing supply of a wide range of rich information on the one hand, and growing concerns regarding respondent burden and privacy problems on the other. In response to these demands, the practice of combining available data has become very common, especially with the availability of fast and versatile software for record linkage. The number of papers at this symposium containing examples from real applications reflects the increasing popularity of combining data from different sources.

However, certain pre-conditions and requirements have to be satisfied to combine data meaningfully and to benefit from their joint use. The use of comparable variables and classification systems, consistent methods of editing and processing, full understanding of quality and methodology underlying the data, attention to privacy concerns are all needed if one is to combine data from different sources in a way that is going to inform and not mislead. Although we have a session specially devoted to the issues of prerequisites and benefits, these preconditions will become evident in many of the applications to be discussed.

We included in the program methods and techniques such as meta analysis, which essentially combines the outcomes of different statistical tests across studies believed to be homogeneous, and combined analysis, where, besides the test results, the data from many experiments are still available for possible pooling and combined analysis. Although these methods were mainly developed in epidemiology and biostatistics, they inspired further development and modifications to applications elsewhere.

Although the symposium covers many important issues, there are plenty of methodological problems that still need to be addressed in the context of combining data. One such issue is linear regression modelling of data assembled via record linkage or statistical matching techniques. The problem is how to evaluate the matching error and its impact on the estimation of the coefficients in the subsequent regression model. That is, how does one account for the fact that some variables in the combined data set have never been observed together on the same subject?

Needless to say, many other topics related to combining information, such as benchmarking of time series, hierarchical modelling, data augmentation, and small area estimation, could not be included due to time constraints. They may perhaps be a challenge for a future symposium. For now, we have a rich list of stimulating statistical issues.

I believe that Christian Thibault and his organizing committee have developed an impressive program which, over the next three days, will prompt us to examine many aspects of the subject of this symposium.

Thank you for attending this symposium, and I hope that these three days will be interesting and useful. Every year, I say that we, at Statistics Canada, benefit inestimably from the presentations and exchanges that take place in the context of these symposia. I hope you will all find that you and your organization benefited greatly from your time spent here.

I wish you all well in these discussions. I am sure that we in Statistics Canada will benefit from the exchanges, and I trust that by the conclusion of the symposium, everyone will have gained useful new insights and ideas.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

COMBINING CENSUS, SURVEY, DEMOGRAPHIC AND ADMINISTRATIVE DATA TO PRODUCE A ONE NUMBER CENSUS

Ray Chambers, Ian Diamond[1] and Marie Cruddas[2]

ABSTRACT

The focus of Symposium'99 is on techniques and methods for combining data from different sources and on analysis of the resulting data sets. In this talk we illustrate the usefulness of taking such an "integrating" approach when tackling a complex statistical problem. The problem itself is easily described - it is how to approximate, as closely as possible, a "perfect census", and in particular, how to obtain census counts that are "free" of underenumeration. Typically, underenumeration is estimated by carrying out a post enumeration survey (PES) following the census. In the UK in 1991 the PES failed to identify the full size of the underenumeration and so demographic methods were used to estimate the extent of the undercount. The problems with the "traditional" PES approach in 1991 resulted in a joint research project between the Office for National Statistics and the Department of Social Statistics at the University of Southampton aimed at developing a methodology which will allow a "One Number Census" in the UK in 2001. That is, underenumeration will be accounted for not just at high levels of aggregation, but right down to the lowest levels at which census tabulations are produced. In this way all census outputs will be internally consistent, adding to the national population estimates. The basis of this methodology is the integration of information from a number of data sources in order to achieve this "One Number".

In this paper we provide an overview of the various stages in such a "One Number Census", and how they allow this integration of information

KEY WORDS: Census; Underenumeration; Coverage Survey;

1. INTRODUCTION TO THE ONE NUMBER CENSUS METHODOLOGY

A major use of the decennial UK census is in providing figures on which to rebase the annual population estimates. This base needs to take into account the level of underenumeration in the census; this has traditionally been measured from data collected in a post-enumeration sample survey (PES) and through comparison with the national level estimate of the population based on the previous census. In the 1991 Census, although the level of underenumeration was comparable with that observed in other developed countries (estimated at 2.2 per cent), it did not occur uniformly across all socio-demographic groups and parts of the country. There was also a significant difference between the survey-based estimate and that rolled forward from the previous census. Further investigation showed that the 1991 PES had failed to measure the level of underenumeration and its degree of variability adequately.

For the 2001 Census maximising coverage is a priority. To help achieve this a number of initiatives have been introduced, for example:

- the Census forms have been redesigned to make them easier to complete;
- population definitions for the Census have been reviewed;

[1] Southampton University, Highfield, Southampton, Hampshire, SO17 1BJ, UK. Email: rc@alcd.soton.ac.uk
[2] Office for National Statistics, Segensworth Road, Titchfield, Fareham, Hampshire, PO15 5RR, UK. Email: marie.cruddas@ons.gov.uk

- postback of Census forms will be allowed for the first time; and
- resources will be concentrated in areas where response rates are lowest.

Despite efforts to maximise coverage in the 2001 Census, it is only realistic to expect there will be some degree of underenumeration. The One Number Census (ONC) project was established as a joint research project between the Office for National Statistics (ONS) and the University of Southampton. The aim of the project is to develop a methodology which measures this underenumeration, provides a clear link between the Census counts and the population estimates, and adjusts all Census counts (which means the individual level database itself) for underenumeration.

The main aspects of the ONC methodology are:

1. The design of the PES in 2001. This will be in the form of a very large area-based household survey (known as the Census Coverage Survey or CCS) which will re-enumerate a sample of postcodes (area-based units which average around 15 households) and collect data on a small number of key variables central to estimating underenumeration;
2. The matching of the CCS data and the census data using both clerical and probability matching methods;
3. The use of combined dual system/regression-based methods to produce population estimates by age and sex for regions, each with a population of around one-half million;
4. The integration of demographic information from administrative registers and small area estimation methods to "cascade" these regional estimates down to Local Authority District (LAD) level;
5. The use of donor imputation methods which are calibrated to the LAD estimates to create a census microdata file that includes imputes for people missed by the census. All "One Number Census" tabulations will then be based on this file.

Figure 1. A schematic overview of the one number census process

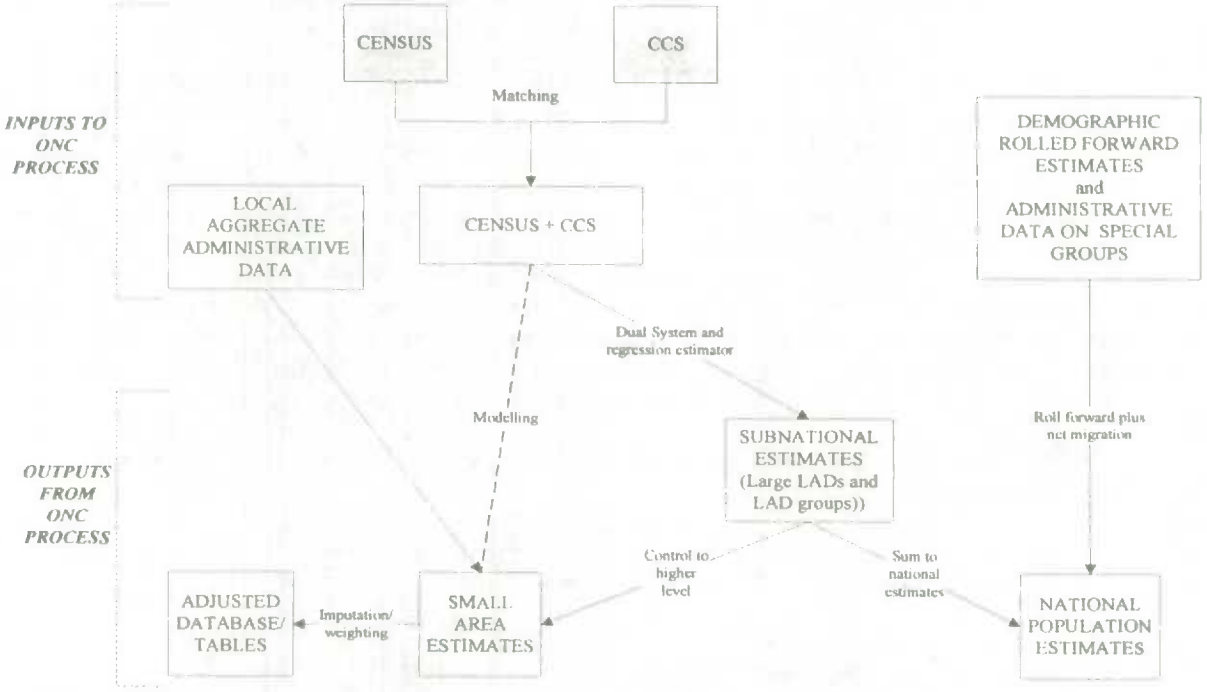


Table 1 sets out the various data sources used in the ONC process. Figure 1 illustrates the various stages of this process and shows how the information provided by these data sources are integrated. In the remainder of this paper we describe each of these stages in more detail.

Table 1. The various data sources used in the ONC process.

Data Source	Use of data source in ONC process
Demographics: • Birth/Death data • Migration estimates	• Making 'rolled forward' population estimates
Previous Censuses	• Making 'rolled forward' population estimates • Design information for the CCS
2001 Census	• The main individual data source • Provision of aggregate level data for benchmarking estimates • Imputation data
2001 CCS	• Population estimates • Undercount models
Administrative records, e.g. • Health Service Records • ONS Birth Registration data	• Quality assurance of population estimates

2. THE CENSUS COVERAGE SURVEY

Following the 1991 Census, a Census Validation Survey (CVS) was carried out in England, Scotland, and Wales. This survey aimed to estimate net underenumeration and to validate the quality of Census data (Heady *et al.*, 1994). The second of these aims required a complete re-interview of a sample of households that had previously been enumerated in the Census. This requirement was costly, due to the time required to fill out the complete census form, resulting in a small sample size. It also meant that the ability of the CVS to find missed households was compromised, since no independent listing of households was carried out.

An alternative strategy was required for 2001. Administrative records were found not to be accurate enough to measure Census quality to the required precision. It was therefore concluded that a PES was needed with a clear objective and different design. The CCS (as the PES will be known in 2001) is designed to address coverage exclusively. Focusing on coverage allows for a shorter, doorstep questionnaire. Savings in time can be translated into a larger sample size. Information on question response error in the Census data will be obtained from other sources, particularly the question testing programme, the 1997 Census Test and through a separate quality survey carried out in 1999.

The CCS will be a postcode-unit based survey, re-enumerating a sample of postcode units rather than households. It is technically feasible to design a household-based CCS by sampling delivery points on the UK Postal Address File, but the incomplete coverage of this sample frame makes it unsuitable for checking coverage in the Census. Consequently, an area-based sampling design has been chosen for the CCS, with 1991 Census Enumeration Districts (EDs) as primary sampling units and postcodes within EDs as secondary sampling units. Sub-sampling of households within postcodes was not considered since coverage

data from all households in a sampled postcode is necessary for estimation of small area effects in the multilevel models proposed for the final stage of the ONC.

Subject to resource constraints, the CCS sample design will be optimised to produce population estimates of acceptable accuracy within sub-national areas called Design Groups. Each such area will correspond to a group of contiguous local administrative areas (Local Authority Districts or LADs) with an expected 2001 population of approximately half a million people. The population estimates will be for the 24 age-sex groups defined by sex (male/female) and 12 age classes: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-79, 80-84, 85+. The ages 45-79 have been combined since there was no evidence of any marked underenumeration in this group in 1991. This age grouping will be reviewed prior to finalising the CCS design.

Underenumeration in the 2001 Census is expected to be higher in areas with particular characteristics. For example, people in dwellings occupied by more than one household (multi-occupancy) have a relatively high probability of not being enumerated in a census. In order to control for the differentials, EDs within each Design Group are classified by a 'Hard to Count' (HtC) score. A prototype version of this score has been produced. This represents social, economic and demographic characteristics that were found to be important determinants of 1991underenumeration by the Office for National Statistics (ONS) and the Estimating With Confidence Project (Simpson *et al.*, 1997). The variables making up the HtC score will be reviewed prior to finalisation of the CCS design. The prototype HtC score was used in the CCS Rehearsal, which was undertaken as part of the 1999 Census Rehearsal, and was based on the following variables from the 1991 Census:

- percentage of young people who migrated into the Enumeration District in the last year;
- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented.

For sample design purposes, the HtC score has been converted to a five point HtC index, with each quintile assigned an index value from 1 (easiest to count) to 5 (hardest to count). At the design stage, the role of the HtC index is to ensure that all types of EDs are sampled. Further size stratification of EDs within each level of the HtC index, based on a derived size measure which reflects key 1991 age-sex counts in EDs improves efficiency by reducing within stratum variance.

The second stage of the CCS design consists of the random selection of a fixed number of postcodes within each selected ED. The proposed sample size for the CCS was investigated through a research project more fully described in Brown *et al* (1999). The research used anonymised individual level 1991 Census data which was assumed to constitute the 'true' population. Using a sequence of Bernoulli trials based on 1991 underenumeration probabilities, individuals were randomly removed from the data set. CCS samples were simulated and estimates of the 'true' population count for each age-sex group in each HtC stratum made. The results of these simulations suggest that a sample of around 20,000 postcodes (approximately 300,000 households) with five postcodes per ED and splitting the country up into Design Groups each with a population of around 500,000 would give an acceptable degree of precision. For England and Wales with a population of 52 million these simulations indicate a 95% confidence interval for the true population count has a width of +/- 0.13%.

3. MATCHING THE CENSUS COVERAGE SURVEY AND CENSUS RECORDS

The ONC estimation strategy requires the identification of the number of individuals and households observed in both the Census and CCS and those observed only once. Underenumeration of around two to three percent nationally means that, although absolute numbers may be large, percentages are small. Thus the ONC process requires an accurate matching methodology. This is more fully described in Baxter (1998).

The independent enumeration methodologies employed by the Census and CCS mean that simple matching using a unique identifier common to both lists is not possible. Furthermore, simple exact matching on the variables collected by both methods is out of the question as there will be errors in both sets of data caused by incorrect recording, misunderstandings, the time gap between the Census and the CCS, errors introduced during processing etc. The size of the CCS also means that clerical matching is not feasible. Thus a largely automated process involving probability matching is necessary.

Probability matching entails assigning a probability weight to a pair of records based on the level of agreement between them. The probability weights reflect the likelihood that the two records correspond to the same individual. A blocking variable, e.g. postcode, is used to reduce the number of comparisons required by an initial grouping of the records. Probability matching is only undertaken within blocks as defined by the blocking variables.

Matching variables such as name, type of accommodation and month of birth are compared for each pair of records within a block. Provided the variables being compared are independent of each other, the probability weights associated with each variable can be summed to give an overall probability weight for the two records. Records are matched if, for the Census record that most closely resembles the CCS record in question, the likelihood of them relating to the same household or individual exceeds an agreed threshold.

A key use of the CCS data will be to enable the characteristics of underenumeration to be modelled in terms of variables collected in the Census. This will allow adjustments based on such models to be applied to the whole population. In order that this process is not biased by application of matching rules, the variables underlying matching and modelling should be as independent as possible.

The initial probability weights used in 2001 will have been calculated from the data collected during the 1999 Census Rehearsal. These weights will be refined as the 2001 matching process progresses.

4. ESTIMATION OF DESIGN GROUP AGE SEX POPULATIONS

There are two stages to the estimation of Design Group populations within age-sex groups, see Brown *et al* (1999). First, Dual System Estimation (DSE) is used to estimate the number of people in different age-sex groups missed by both the Census and CCS within each CCS postcode. Second, the postcode level population estimates obtained via DSE are used in regression estimation to obtain final estimates for the Design Group as a whole.

For matched Census/CCs data, the simplest estimate of the total count for a postcode is the union count (i.e. the total of those counted in the Census and/or CCS within the postcode). However, this will be biased low because it ignores the people missed by both counts. DSE methodology gives an estimated count which adjusts the union count for this bias. DSE assumes that:

(i) the Census and CCS counts are independent; and
(ii) the probability of 'capture' by one or both of these counts is the same for all individuals in the area of interest.

When these assumptions hold, DSE gives an unbiased estimate of the total population. Hogan (1993) describes the implementation of DSE for the 1990 US Census. In this case assumption (i) was approximated through the operational independence of the Census and PES data capture processes, and assumption (ii) was approximated by forming post-strata based on characteristics believed to be related to heterogeneity in the capture probabilities.

In the context of the ONC, DSE will be used with the Census and CCS data as a method of improving the population count for a sampled postcode, rather than as a method of estimation in itself. That is, given matched Census and CCS data for a CCS postcode, DSE is used to define a new count which is the union

count plus an adjustment for people missed by both the Census and the CCS in that postcode. This DSE count for the sampled postcode is then used as the dependent variable in a regression model, which links this count with the Census count for that postcode.

The regression model is based on the assumption that the 2001 Census count and the DSE count within each postcode satisfy a linear regression relationship with a zero intercept (a simple ratio model). However, for some age-sex groups there is the possibility of a non-zero intercept, as in some postcodes the Census can miss all the people in particular age-sex groups. In such cases an intercept term α_d will be added to the ratio model described below. This issue is currently being researched.

It is known from the 1991 Census that undercount varies by age and sex as well as by local characteristics, therefore a separate regression model within each age-sex group for each HtC category within each Design Group is used. Let Y_{id} denote the DSE count for a particular age-sex group in postcode i in HtC group d in a particular Design Group, with X_{id} denoting the corresponding 2001 Census count. Estimation is based on the simple zero intercept regression model:

$$E\left\{Y_{id} \middle| X_{id}\right\} = \beta_d X_{id}$$
$$Var\left\{Y_{id} \middle| X_{id}\right\} = \sigma_d^2 X_{id}$$
$$Cov\left\{Y_{id}, Y_{jf} \middle| X_{id}, X_{jf}\right\} = 0 \quad for\ all\ i \neq j; d, f = 1,...,5$$

Substituting the weighted least squares estimator for β_d into (1), it is straightforward to show (Royall, 1970) that under this model the Best Linear Unbiased Predictor (BLUP) for the total count T of people in the age-sex group (within the Design Group) is the stratified ratio estimator of this total given by:

$$\hat{T} = \sum_{d=1}^{5}\left\{T_{Sd} + \sum_{i \in R_d}\left(\hat{\beta}_d X_{id}\right)\right\} = \sum_{d=1}^{5}\hat{T}_d$$

where T_{Sd} is the total DSE count for the age-sex group for CCS sampled postcodes in category d of the HtC index in the Design Group; and R_d is the set of non-sampled postcodes in category d of the HtC index in the Design Group. Strictly speaking, the ratio model above is known to be wrong as the zero covariance assumption ignores correlation between postcode counts within a ED. However, the simple ratio estimator remains unbiased under this mis-specification, and is only marginally inefficient under correlated postcode counts (Scott and Holt, 1982).

The variance of the estimation error $\hat{T} - T$ can be estimated using the ratio model specified above. However this variance estimator is sensitive to mis-specification of the variance structure (Royall and Cumberland, 1978). Consequently, as the postcodes are clustered within EDs, the conservative ultimate cluster variance estimator will be used. This is given by:

$$\hat{V}(\hat{T} - T) = \sum_{d=1}^{5}\frac{1}{m_d(m_d - 1)}\sum_{e=1}^{m_d}\left(\hat{T}_d^{(e)} - \hat{T}_d\right)^2$$

where $\hat{T}_d^{(e)}$ denotes the BLUP for the population total of category d of the HtC index based only on the sample data from ED e and m_d is the number of EDs in HtC group d.

The above estimation strategy represents a regression generalisation of the Horvitz-Thompson DSE estimator proposed in Alho (1994). As a postcode is a small population in a generally small geographic

area, and with the counts split by age and sex, the DSE homogeneity assumption should not be seriously violated. In the situation where people missed by the Census have a higher chance of being missed by the CCS than those counted by the Census, one would expect the regression estimator based on the DSE count to underestimate, but to a lesser extent than the regression estimator based on the union count. When the reverse happens and the CCS is very good at finding the missed people (the requirement for getting unbiased estimates when using the union count in the regression estimator) one would expect the DSE count regression estimator to overestimate. However, unless these dependencies are extremely high, one would not expect a gross error.

5. ESTIMATION FOR LOCAL AUTHORITY DISTRICTS

Direct estimation using the CCS produces estimates by age and sex for each Design Group. In the case of a LAD with a population of approximately 500,000 or above this will give a direct estimate of the LAD population by age and sex. However, for the smaller LADs clustered to form Design Groups, this will not be the case – even though all LADs will be sampled in the CCS. For these LADs it will be necessary to carry out further estimation, and to allocate the Design Group estimate to the constituent LADs.

Standard small area synthetic estimation techniques will be used for this purpose. These techniques are based on the idea that a statistical model fitted to data from a large area (in our case the CCS Design Group) can be applied to a smaller area to produce a synthetic estimate for that area.

A potential problem with this approach is that, while the estimators based on the large area model have small variance, they may be biased for any particular small area. A compromise, introduced in the 1980s, involves the introduction of random effects for the small areas into the large area model. These allow the estimates for each small area to vary around the synthetic estimates for those areas. This helps reduce the bias in the simple synthetic estimate for a small area at the cost of a slight increase in its variance (Gosh and Rao, 1994).

As described in the previous section, direct Design Group estimation is based on the linear regression model (1) linking the 2001 Census count for each postcode with the DSE-adjusted CCS count for the postcode. This model can be extended to allow for the multiple LADs within a Design Group by writing it in the form

$$Y_{idl} = \beta_d X_{idl} + \delta_{dl} + \varepsilon_{idl}$$

where the extra index l = 1...L denotes the LADs in a Design Group, δ_{dl} represents an LAD 'effect' common to all postcodes with HtC index d, and ε_{idl} represents a postcode specific error term. The addition of the δ_{dl} term above represents differences between LADs that have been grouped to form a Design Group.

This two level regression model can be fitted to the CCS data for a Design Group, and the LAD effects δ_{dl} estimated. For consistency, LAD population totals obtained in this way will be adjusted so that they (i) sum to the original CCS Design Group totals; and (ii) they are always at least as large as the 2001 Census counts for the LAD. Research is currently underway to determine whether the potential bias gains from this approach outweigh the variance increase relative to that of the simple synthetic estimator.

6. DEMOGRAPHIC ESTIMATES AND QUALITY ASSURANCE

A key aspect of the One Number Census process is the availability of the best possible comparable demographic estimates as well as data from other administrative sources which can serve as a reliable check on the national level estimates that will be produced by aggregating the Design Group estimates produced from the CCS. How these sources will be used is part of an on-going development of a quality

assurance strategy for the 2001 Census that acknowledges the qualitative rather than quantitative nature of the data for this purpose.

Comparable demographic estimates will be obtained by 'rolling forward' information from a previous census, using registration data on births and deaths, and migration information from a number of sources. Different levels of error are associated with these sources. Thus in year t the population P_t is given by:

$$P_t = P_0 + \sum_i (B_i - D_i + I_i - E_i),$$

where P_0 is the base population and B, D, I and E are respectively the Births, Deaths, Immigrants and Emigrants in each subsequent year.

Cohort analysis carried out by the Office for National Statistics (Charlton *et al*, 1998) indicated that the national level population estimates produced by "rolling forward" the 1991 Census counts, even when these are adjusted for undercount by the Census Validation Survey, were less reliable than the demographic estimates rolled forward from the 1981 Census counts. This means that in 2001, the rolled forward estimates, which will be used as a check on the census count, will still be based on the 1981 Census.

Further work is being done to estimate the 'margin of error' in these rolled forward estimates of the population taking into account sampling and non-sampling errors in the source data (Charlton and Chappell, 1999). This will provide the 'plausibility range' for assessing the adjusted 2001 census figures. Whilst death registration data are believed to be virtually complete, the sources of migration data are subject to error.

6.1 Migration data sources

International migration data are produced from the International Passenger Survey (IPS), a sample survey of international arrivals and departures at ports and airports. The IPS covers all travellers and, as only a small proportion of these are migrants, the number of migrants included in the sample is relatively small. Therefore, relatively high error rates are attached to these international migrant figures. Persons travelling between the UK and the Irish Republic are not currently included in the IPS. Net Irish migration is estimated from the UK and Irish Labour Force Surveys (which include a question on address 12 months previously) and counts of Irish citizens registering with the National Health Service.

Migration within and between the constituent countries of the UK (England and Wales, Scotland, Northern Ireland) is estimated from the reporting to the NHS of re-registrations with General Practitioners. However some people, particularly young adults, do not register promptly (or at all) following a change of address. The geographical information from this source is limited, showing only the equivalent of the county or borough of origin or destination.

It is thought that the majority of asylum seekers are not recorded in the IPS migrant sample. Improved data on these persons are being received from the Home Office to be incorporated in the estimates. Similarly the Home Office is supplying improved information on the numbers of persons who enter the UK as short term visitors but later apply to remain as residents.

6.2 Using administrative records

As an additional independent source against which to compare the rolled forward estimates, use of Department of Social Security data on the number of Retirement Pension and Child Benefit claimants is being investigated. This administrative source is believed to offer almost complete coverage of the elderly and of young children - these two groups have been relatively poorly enumerated in past censuses.

7. IMPUTATION OF MISSED HOUSEHOLD AND INDIVIDUALS

This final stage of the ONC process creates a ONC database which includes all individuals and households enumerated in the Census as well as imputed individuals and households missed by the Census. This process is described more fully in Steele *et al* (1999). The imputation procedure is based on the fact that there are two processes that cause individuals to be missed by the Census. First, when there is no form received from the household and therefore all household members are missed. Second, when contact with the household fails to enumerate all household members and therefore some individuals are omitted from the form. These two processes are treated separately by the methodology.

The first step in the process models the probability of either an individual or a household being counted in the Census in terms of the characteristics of that individual or household. This is possible in CCS postcodes where there are matched counts of the population. The second step estimates a Census coverage probability for every individual and household counted in the Census by applying the fitted model from step 1 to all Census enumerated individuals and households. Each such probability is then inverted to obtain a weight that reflect the corresponding individual's or household's propensity to be enumerated. These coverage weights are calibrated to agree with the total population estimates by age-sex group and by household size for each LAD. Finally a controlled imputation step that generates imputed individual and household level data consistent with the weighted counts of these data is carried out.

The aim of this process is to eventually create an individual level database that will represent the best estimate of what would have been collected had the 2001 Census not been subject to underenumeration. Tabulations derived from this database will automatically include compensation for underenumeration and therefore all add to the 'One Number'.

8. SUMMARY

The One Number Census process is an example of what is possible when data from a number of statistical sources are integrated. In particular, integration of data from the 2001 Census data collection and the subsequent Census Coverage Survey will provide a base for creating a final UK Census in 2001with a "built-in" underenumeration adjustment. Integral to this process is the need to match the data collected from these two sources, as well as to quality assure the estimates subsequently generated using data from alternative demographic and administrative systems. This leads to extra complexity, but also to a much more useful Census product, and one expected to be more acceptable to the majority of Census data users in the UK.

In this paper we have attempted to distil the "flavour" of the ONC process, focussing in particular on the data inputs to the various steps in this process. These lead from a traditional Census collection at the start to the final creation of a ONC Census individual and household level data base that includes an underenumeration adjustment consistent with all the different data sources that have been used as inputs at earlier stages in the process.

However, such integration comes at a cost – and in this case this cost is reflected in the complex nature of the ONC process and the need to ensure that all input data sources are accurately matched, particularly where the methodology is based on a comparison of how good these sources are at enumerating individuals and households in the UK. A considerable amount of research has been carried out into assessing the robustness of the methodology, and this work is continuing. In particular, data obtained from the 1999 Census Dress Rehearsal will allow further refinement of this methodology and further assessment of its potential reliability in 2001. Provided the results of this analysis continue to indicate that the methodology described in this paper offers a significant improvement over a "standard" Census collection, we expect that the Office for National Statistics will carry out a One Number Census in the UK in 2001.

REFERENCES

Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics*, **10**, 245 - 256.

Baxter, J. (1998) One Number Census Matching. One Number Census Steering Committee Working Papers ONC(SC)98/14. Available from the Office for National Statistics, Titchfield. (onc@ons.gov.uk)

Brown, J., Diamond, I., Chambers, R and Buckner, L(1999). The role of dual system estimation in the 2001 Census Coverage Survey of the UK. *Statistics Canada Symposium, Ottawa, May 4-7, 1999.*

Charlton, J., Chappell, R. and Diamond, I. (1998). Demographic Analyses in Support of a One Number Census. *Proceedings of Statistics Canada Symposium 97*, 51-57.

Charlton, J. and Chappell, R. (1999) Uncertainty Intervals for National Demographic Estimates. One Number Census Steering Committee Working Papers ONC(SC)99/05. Available from the Office for National Statistics, Titchfield. (onc@ons.gov.uk)

Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: An appraisal. *Statistical Science*, **9**, 55-93.

Heady, P., Smith, S. and Avery, V. (1994) *1991 Census Validation Survey: Coverage Report*, London: HMSO.

Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *J.A.S.A.*, **88**, 1047-1060.

Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.

Royall, R. M. and Cumberland, W. G. (1978) Variance estimation in finite population sampling. *J.A.S.A.*, **73**, 351-361.

Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *J.A.S.A.*, **77**, 848-854.

Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends*, **90**, 31-39.

Steele, F., Brown, J. and Chambers, R. (1999) A donor imputation system to create a census database fully adjusted for underenumeration. *Statistics Canada Symposium, Ottawa, May 4-7, 1999.*

SESSION I

PREREQUISITES AND BENEFITS

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

STATISTICAL PROCESSING IN THE NEXT MILLENNIUM

Wouter Keller, Ad Willeboordse, Winfried Ypma[1]

ABSTRACT

This paper states how SN is preparing for a new era in the making of statistics, as it is triggered by technological and methodological developments. An essential feature of the turn to the new era is the farewell to the stovepipe way of data processing. The paper discusses how new technological and methodological tools will affect processes and their organization. Special emphasis is put on one of the major chances and challenges the new tools offer: establishing coherence in the content of statistics and in the presentation to users.

KEY WORDS: metadata; stovepipe; integration; EDI; StatLine; output data warehouse.

1. INTRODUCTION

The rapid developments in information technology have a huge impact on statistical production processes of statistical institutes. Traditionally, the latter carry out a wide range of surveys. Implementation of new developments exceeds the mere replacing of paper questionnaires and publications by electronic ones, or the shift from manual to automated editing and imputation. There is a fundamental change in the organization of statistical production processes as well. Two types of integration can be observed: integration of the various activities within a survey process, vs integration of various survey processes as a whole. The benefits of integration are manifold and affect different players involved: a lower response burden for respondents, cheaper, faster and more sophisticated processing for statistical institutes, and more reliable, consistent and accessible information for users.

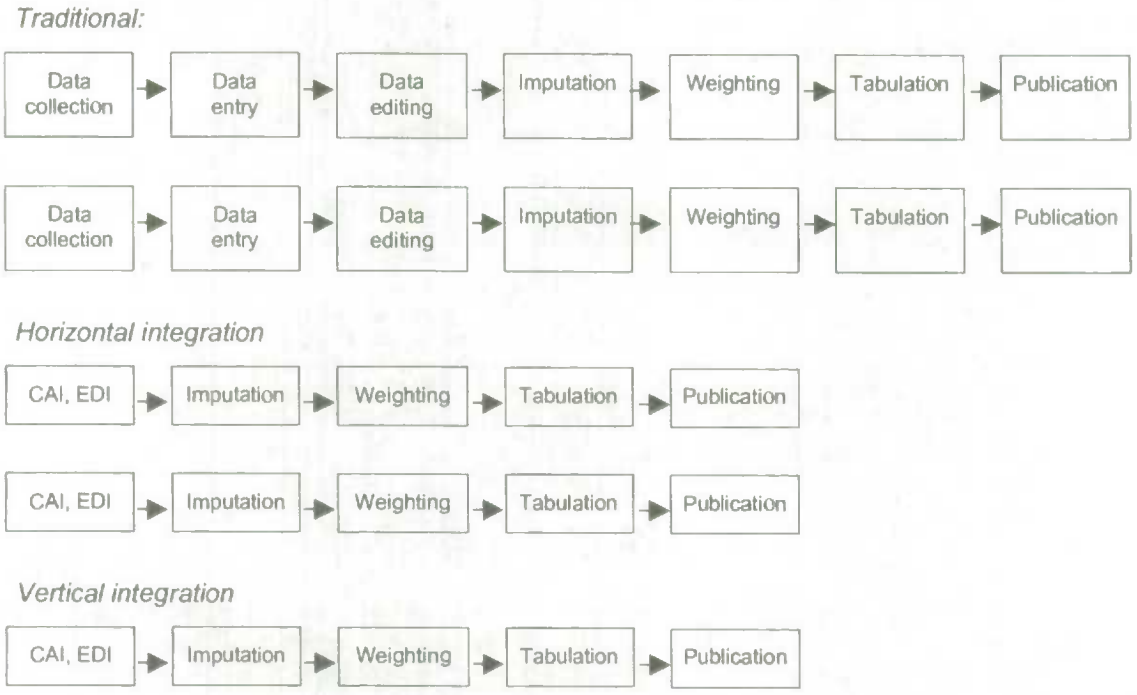
This paper states how SN is preparing for this new era in the making of statistics. By passing through all phases of the statistical process, it will be described how new technological and methodological tools affect the processes and their organization. Special emphasis is put on one of the greatest challenges for SN's: establishing coherence in the content of statistics and in the presentation to users.

2. NEW TECHNOLOGIES FOR DATA COLLECTION

The traditional approach towards making statistics is based on data collection by means of (sample) *surveys* using *paper* questionnaires. Developments in IT have triggered a change from *paper data streams* to *bit data streams*. The paper form is gradually being replaced by an electronic one, or data are collected electronically without questionnaires and interviewers (EDI, i.e. Electronic Data Interchange). This new way of data collection has a fundamental impact on the way statistical agencies operate. This impact reveals itself in two directions, as figure 1 shows. First there is *task integration* (here denoted as *horizontal integration*), and secondly survey integration, i.e. *vertical integration*.

[1] Statistics Netherlands, PO Box 400, 2270 JM Voorburg, The Netherlands

Figure 1. Horizontal and vertical integration of survey processes



After successfull introduction in household surveys, vertical integration is now being implemented in business surveys as well. SN gives priority to integrate data collection for business surveys that use EDI. This approach focuses on *data sources* (i.e. book keeping systems) instead of *surveys*: there is *one* (electronic) questionnaire for *one* book keeping system. Data collected from these systems may serve several statistics. Vertical integration means an important step towards the realization of one of the future goals of Statistics Netherlands, i.e. the *one-counter concept*: a respondent communicates with only one unit in SN.

In persons/households surveys the share of EDI has evolved form 24% in 1987 to 90% in 1997. For 2007 this percentage will approach 100, while two third of the EDI will refer to *secondary* sources, i.e. existing administrative registers. In business statistics, the paper share is higher: 70% in 1987 and an estimated 25% in 2007. Here, we expect that the secondary sources will account for less than half of the EDI-share.

The introduction of EDI triggers the tendency towards integration of the organizational units involved in data collection. Anticipating on this development, SN created in 1993 a separate data collection division. This division takes care of all surveys among persons and households. Integration of the major part of data collection for business surveys will take a number of years. The speed of this development heavily depends on the pace with which the implementation of primary and secondary EDI evolves.

It should be stressed that EDI has beneficial effects on the *quality* of statistics as well, in particular with respect to consistency of concepts and data. Concentration of data collection leads to *standardization* of questionnaire concepts, and by that of statistical concepts. And it leads to a more *consistent* reporting by respondents, which in turn creates better chances for consistent and relatable statistics. This is especially true for large and complex businesses.

3. NEW METHODOLOGIES FOR DATA PROCESSING

3.1 Editing

The introduction of EDI has a substantial effect on data processing activities of statistical institutes. Some kind of *administrative editing* is needed at the input stage. This type of editing relates to detecting and correcting *conceptual* and *processing errors*. It is characterized by batch-oriented error detection procedures. Detected errors are analyzed and corrected on a form-by-form basis. Correction may involve contact with the information supplier. The result of administrative editing is an administratively clean input database. This input database is the data source for the subject-matter departments. For them, the database can be viewed as a set of *virtual respondents*. The subject-matter departments focus on *statistical editing*. Statistical editing focuses on finding distribution errors in the data, like outliers, and unusual yearly shifts. Typically, graphical macro-editing and output editing will be applied. Errors are corrected without contact with the information supplier. Statistical editing may very well be integrated in the imputation and estimation procedures.

3.2 Imputation: the micro-database

As stated earlier, more and more electronic data will become available as secondary sources for statistics. Important sources will be administrative (public) records in registers like those of the tax authorities and the social security administration, but also records from private sources. It is to be expected that in the future those registers will become the main data sources for statistical institutes. Combining this *abundant* amount of (cheap) register data with the *sparse* amount of (expensive) survey data in such a way that the results meet high quality standards will become a major challenge for survey methodology.

Figure 2. Survey and register data

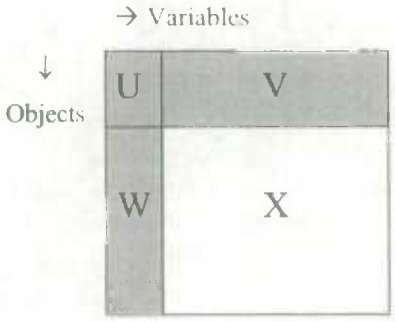


Figure 2 explains the differences between register and survey data. The columns denote the variables to be observed and the rows are the units (persons, businesses, etc.) or *objects* for which the variables are measured. Most surveys are samples that tend to cover *many* variables and a *limited* number of objects. U and V denote the sets of variables observed where U refers to variables that also appear in registers. Both sets of variables are defined by the statistical institute. For administrative registers, the opposite tends to be the case: they cover *many* objects (most often the whole target population) and a *limited* number of variables. This part is denoted by the sets U and W. Neither definition nor observation is under control of the statistical institute.

Combining survey data with register data means linking the survey records to the corresponding register records. Links must be established on the basis of the common part of the records, i.e. the measurements on the variables in the set U. Due to definition differences of the variables appearing both in the register and in the survey, this may prove to be not an easy task.

After linking of registers, statistics can be compiled on the basis of the complete dataset. For statistics relating to variables in the set U + W data are available on all population elements. So, there is no sampling

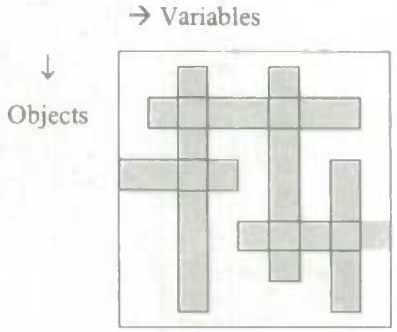
error. The situation is more complex with respect to the variables in the set V. Here, only partial information is available. It is the challenge of survey methodology to predict the missing information in X using a model for the relationship between the variables in U and V. This prediction can take two forms:

- Mass imputation. Based on models, synthetic values are computed for the missing data in X, and these values are actually included in the data file.
- Weighting. Based on models, weights are computed and assigned to the complete records in U + V. In the computation of statistics only the weighted records in U + V are used.

Modeling of the relationship is a vital part of both approaches. Methodologies like *record linkage, imputation, synthetic estimation,* and *calibration estimation* will become dominant once the potential of this approach is better understood.

The example above illustrates the idea of combining *one* register with *one* survey. In practice, focus should be on the more general case of combining *many* surveys with *several* registers, all containing measurements on the same objects. This situation is graphically shown in figure 3. For each object-type (person, business), the corresponding population is enumerated. Starting with an empty file or database table, all available data from surveys and registers is filled in. Since the database table covers all objects in the population, the resulting fully imputed or weighted table will be called a *synthetic population* or *micro-database*.

Figure 3. The Synthetic Population



Two micro-databases are important: one for the object-type *person* (with households as an aggregate), and one for the object-type *establishment* (with enterprise as an aggregate). Besides these two obvious object-types, also others, like dwellings, cars, jobs, etc can be considered.

3.3 Aggregation: the reference database StatBase

Whatever approach is used to construct a micro-database (mass imputation or weighting), the resulting database cannot be used at the micro level: imputation might result in nonsense micro data. This occurs when the number of records in surveys (corresponding to U and V) is small compared to the population, while the correlation between the variables in U and V is weak, even with multiple surveys/registers available. To avoid these problems, the data must be aggregated to a suitable higher level. Also, there will be aggregates that fail to meet confidentiality standards. Therefore, an additional database is needed, containing all *publicable* aggregates of the micro- database. It is a *Reference Database* and we call it StatBase.

In the future, surveys should be designed such that the result of combining survey data with secondary source data is optimal. Again, it should be noted that developments towards integration have positive effects on the quality of the statistics. Needless to say that the combined treatment of various data sources, both surveys and registers, enhances relatability of statistical results. Matching operations reveal inconsistencies between different sources and provide the opportunity to solve them in an early stage of the process, instead of having to wait until the integration in National Accounts or other integration frames.

Moreover, bringing together all variables for a specific object-type in one micro database and subsequently in one reference database requires standardization of concepts, in particular classifications. The major tool for achieving this standardization is a set of so-called *meta-servers* surrounding StatBase. This is the topic of the next subsection.

3.4 Meta-servers around StatBase

No data without metadata. Therefore, the reference *data*base StatBase is surrounded by a set of *meta-servers*, which label and define all data that enter StatBase. The meta-servers are designed according to a data model based on Sundgren (1992). Very simplified it goes as follows: Statistical data describe (sets of) *objects*, e.g. enterprises. For these objects certain *count variables* are observed (e.g. turnover). The population of objects may be subdivided into sub-populations using a *classification variable* that constructs a classification (e.g. size). This is generally done for a certain period of *time*. There is a meta-server for each of the components of the data model: a classification server (including time), an object server and a (count) variable server. Data retrieved from a micro-database or directly from a survey can only be read into StatBase by reference to the meta-servers. This means that each data item in StatBase is exclusively and exhaustively described in terms of an object-type, a class and a count variable of the set of meta-servers.

There is no reason to confine the use of the data model as described above to the meta-servers surrounding StatBase. Indeed, meta-servers can serve input and throughput processes as well. In order to establish a perfect link between micro-databases and StatBase, the two should make use of the same metadata. This uniformity will become general policy and in the end there will be no boundaries as to its scope. Eventually, *all* metadata throughout the whole statistical process should be drawn from the meta-servers described above.

The metadata discussed so far are still rather static. They describe the *meaning* of the statistical concepts, not the way the data have been obtained. Therefore, SN aims at *activating* meta-data as well. Metadata will then describe operations on data in a way that can be understood by the processing tools. The tools are *steered* by this meta-data and the resulting documentation is *generated* automatically. See also Bethlehem et al (1999).

4. TECHNOLOGIES AND DATA DISSEMINATION

4.1 Introduction

The reference database is the starting point for the dissemination stage. As StatBase contains *all* publicable aggregates produced by whichever survey, it is in principle possible to derive *all* publications, both paper and electronic, both general and specific, from StatBase. The scope of these publications is never confined to the boundaries of individual surveys or organizational units. This enables the design of tables and publications that really cover an area of interest from the user perspective, instead of merely describing the outcomes of a specific survey. Indeed, the new concept will definitely finish with the *stove-pipe* era, where users looking for consistent information on a topic, say housing, were forced to consider many different (paper) publications (e.g., on production, labor, use of houses). There is, however, one important condition for successfully combining data from different sources: their concepts, i.e. their object-types, classifications and variables should be harmonized, in such a way as to allow data from different surveys to fit in one and the same table. The meta-servers around StatBase are the appropriate tools to fulfill this harmonization task.

This section discusses how the data in StatBase are used for dissemination. The output data warehouse *StatLine* is at the heart of the dissemination process.

4.2 Content of StatLine: datacubes

StatBase is the one and only source for StatLine, which contains a number of multi-dimensional datacubes, each covering a theme and together providing a comprehensive and coherent picture of the society. As themes may overlap, the same data may appear in several cubes under different themes. StatLine can be

characterized as a set of standard views on StatBase. These views are defined in the *tabulation server*. Designing a datacube such that it both covers a whole area of interest *and* shows a minimum number of blank cells *and* is easily accessible by users, requires new and specific skills of statisticians, which SN refers to as the *art of cubism*.

Datacubes can be very extended, both in length, width and depth. Unlike paper tables, this does not lead to inconvenient presentations to the user. For, he is not confronted with a (huge) table, but with a *menu*, from which he can select the ranges of the table he is interested in. The challenge for the "artists" comes down to designing and structuring the menu, in such a way that the user can easily find and select what he looks for.

4.3 Publications based on StatLine

StatLine is the spider in the dissemination web of SN, be it alone because it is deemed to become the one and only source from which *all* SN publications are derived, either electronic, telephone or paper.

Most publications will be *selections* from StatLine. For the major part they coincide with the content of one or more datacubes, but they may combine parts of datacubes as well. There are two publications that cover the whole content of StatLine, i.e. contain all datacubes: StatWeb on the Internet en StatLine on a CD-ROM.

5. TECHNOLOGIES AND COHERENCE: TOOLS, NO RULES

5.1 Introduction

It was repeatedly stressed that new technologies trigger consistency of data and co-ordination of concepts. This final section provides a synopsis and a further elaboration of this issue, which is one of the most outstanding goals in SN's long term business survey strategy. Since StatLine we are able to define more precisely and more concrete than before what is to be understood by the rather vague notion of "coherence". It denotes a situation where StatLine provides a comprehensive and coherent picture of society. This implies that there is consistency *within* datacubes and *among* datacubes. There is another difference with the past, and that refers to the way coherence is to be achieved. Before StatLine we used to enforce co-ordination by (paper) *rules*, which did not work very well. Now we try to obtain the goal by offering the statistical divisions attractive *tools*.

5.2 Logical steps towards coherence

The state of ultimate coherence has a number of features that can be described as logical steps on the way to the ideal, where each step represents a certain level of ambition:

1. The first step is as trivial as important: establish *well-defined concepts*. It makes no sense to discuss the comparability between concepts if one does not know where they stand for.
2. The second step relates to *uniform language*: if we know what our terms mean, we have to make sure that the same terms have the same meaning throughout StatLine, and conversely, that the same concepts are named with the same terms.
3. The third step concerns *co-ordination*, which comes down to attuning (well-defined) concepts in such a way that they can be meaningfully related. Somewhat simplified, and with the StatLine datacube in mind, there are mainly two "directions" of relatability:

- *Horizontally*: for two count variables (e.g. turnover and number of staff) to be relatable, they have to refer to the same populations, and thus to the same object-type and the same classification(s);
- *Vertically*: for a count variable (e.g. number of staff) to be addible over the full range of a classification (e.g. for agriculture, manufacturing industry, trade etc.), it must be equally defined for all classes of the classification.

The co-ordination step results in a set of coordinated concepts, but it does not prohibit statistics to maintain their own "nearby-concepts".

4. The fourth step therefore involves *standardization*. In order to protect users against a too broad and subtle - and hence confusing - assortment of nearby-concepts, these are eliminated as far as justified.

Now that the concepts are clear, coordinated and sparse, we move to *data*:

5. The fifth step consists of establishing *consistency* among data throughout StatLine. The most eye-catching expression of inconsistency occurs when for the very same concept different figures show in StatLine. But also more hidden forms of inconsistency are possible.
6. The final step relates to the *presentation* of the data. They should be offered to the user in such a way that their relationships become maximally apparent. This step is implemented by the *structuring* of datacubes such that they describe areas of interest and consequently in the *ordering* of the cubes in a thematic tree structure describing society.

These are *logical* steps indeed, each next step standing for a higher and more difficult to achieve ambition level. In practice the order may be less rigid than suggested here, depending on specific circumstances.

5.3 Organizational steps towards coherence

Different places in the organization play different parts in the coordination process. Leading principle is that those who cause inconsistency are responsible for solving the problems.

The *Division for Presentation and Integration* maintains StatLine. However, the statistical divisions that supply the data remain fully owners of and thus responsible for their data: they "hire" rooms in the warehouse where they display *their* products. A *Council of Editors*, in which all statistical divisions are represented, advises with respect to matters of common interest, such as editorial guidelines, the thematic tree structure and a centrally maintained list of synonyms. The responsibility for co-ordination of concepts and consistency of data has recently been assigned to *Directors of Statistical Divisions*, in the understanding that the total field of statistical concepts has been distributed over them. E.g., the Director of the Division for Labor Statistics has been assigned responsibility for (the co-ordination of) all concepts and data relating to labor, wherever in the Bureau such concepts may occur.

REFERENCES

Altena, J.W. and Willeboordse, A.J. (1997): Matrixkunde of "The Art of Cubism" (Dutch only). Statistics Netherlands, Voorburg.

Bethlehem, J.G. (1995): Improving the Quality of Statistical Information Processing. Proceedings of the 6-th Seminar of the EOQ committee on Statistical Methods, Budapest, 1995, pp. 87-114.

Bethlehem, J.G. and Van der Pol, F. (1998): The Future of Data Editing. In: M.P. Couper et al.: Computer Assisted Survey Information Collection. Wiley, New York, pp. 201-222.

Bethlehem, J.G., Kent, J.P., Willeboordse, A.J. and Ypma, W. (1999): On the use of metadata in statistical processing. Third Conference on Output Databases, Canberra, March 1999.

De Bolster, G.W. and Metz, K.J. (1997): The TELER-EDISENT Project, Netherlands Official Statistics, Autumn 1997, Statistics Netherlands, Voorburg.

Hammer, M. and Champy, J. (1993): Reengineering the Corporation, A Manifesto for Business Revolution. London: Nicholas Brealey Publishing.

Ijsselstein, H. (Ed.) (1996): Proceedings of the Conference on Output Databases. Statistics Netherlands, Voorburg.

Keller, W.J. and Bethlehem, J.G. (1998): The impact of EDI on Statistical data processing. Meeting on the Management and Information Technology, Geneva, February 1999.

Sundgren, B. (1992), Statistical Metainformation Systems, Statistics Sweden, Stockholm

Ypma, W. (1997) Remarks on a Classification Server. Second Conference on Output Databases, Stockholm, November 1997.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

THE CHALLENGES OF USING ADMINISTRATIVE DATA TO SUPPORT POLICY-RELEVANT RESEARCH: THE EXAMPLE OF THE LONGITUDINAL IMMIGRATION DATABASE (IMDB)

Jane Badets and Claude Langlois[1]

ABSTRACT

The Longitudinal Immigration Database (IMDB) links immigration and taxation administrative recordsinto a comprehensive source of data on the labour market behaviour of the landed immigrant population in Canada. It covers the period 1980 to 1995 and will be updated annually starting with the 1996 tax year in 1999. Statistics Canada manages the database on behalf of a federal-provincial consortium led by Citizenship and Immigration Canada.

The IMDB was created specifically to respond to the need for detailed and reliable data on the performance and impact of immigration policies and programs. It is the only source of data at Statistics Canada that provides a direct link between immigration policy levers and the economic performance of immigrants.

The paper will examine the issues related to the development of a longitudinal database combining administrative records to support policy-relevant research and analysis. Discussion will focus specifically on the methodological, conceptual, analytical and privacy issues involved in the creation and ongoing development of this database. The paper will also touch briefly on research findings, which illustrate the policy outcome links the IMDB allows policy-makers to investigate.

1. INTRODUCTION

Detailed, reliable and policy-relevant information is needed to assess the impact and performance of the immigration program governing admissions to Canada. The Longitudinal Immigration Database (the IMDB) was created in response to a growing need on the part of policy-makers for such information. The following paper will examine the issues related to the development of a longitudinal database combining administrative records to support policy-relevant research, analysis and evaluation. Discussion will focus specifically on the methodological, conceptual, analytical and privacy issues involved in the creation and ongoing development of this database. The paper will also touch briefly on research findings, which illustrate the policy outcome links the IMDB allows policy-makers to investigate.

2. BACKGROUND

2.1 What is the IMDB?

The longitudinal Immigration Database links administrative records of immigrants at landing with their subsequent taxfiles into a comprehensive source of data on the economic behaviour of the immigrant population (including refugees) in Canada. It covers the tax and landing years of 1980 to 1995 and will be updated annually.

The IMDB was created to respond to the need for detailed and reliable data on the performance and the impact of the immigration program in Canada. It allows, for the first time, the analysis of relative labour market behaviour of different categories of immigrants over a period long enough to assess the impact of immigrant characteristics, such as education and knowledge of Canada's official languages, on their

[1] Jane Badets, Housing, Family & Social Statistics Division, Statistics Canada, Jean Talon, 7th floor, Ottawa, Ontario. Claude Langlois, Citizenship & Immigration Canada , Jean Edmonds Tower South, 18th Floor, 365 Laurier Avenue West, Ottawa, Ontario.

settlement pattern. It permits the investigation and measurement of the use of social assistance and (un) employment insurance by different categories of immigrants. The database also permits the analysis of secondary inter-provincial and inter-urban migration of newcomers to Canada. It shows promise as a source of data on job creation by immigrant businesses.

The IMDB is the only source of data at Statistics Canada , which links economic outcomes to immigrant policy levers. For the researcher, the IMDB is the only source of labour market data which permits the user to distinguish between categories or classes of immigrants, and to distinguish between cohort, period of arrival, aging, location and program effects when analyzing immigrant economic behaviour.

2.2 Management of the Database

The IMDB is managed by the Housing, Family & Social Statistics Division, Statistics Canada (SC), on behalf of a federal-provincial consortium led by Citizenship & Immigration Canada (CIC). CIC has been the lead department in terms of the development of the database.

The IMDB consortium was created to ensure that all government departments with a direct interest in immigration policy have access to a shared body of information to support research and analysis on the performance of the immigration program. In addition to CIC, the founding consortium members are the federal departments of Canadian Heritage, Human Resources Development Canada, and Industry Canada and the provincial departments with immigration and/or immigrant settlement responsibilities from Quebec west to British Columbia. The consortium meets annually to discuss the ongoing development of the database, status of projects and to share members' research on the immigration program based on results from the database.

2.3 Contents

The IMDB combines two data sources: (1) administrative records on immigrants obtained from the landing record of immigrants entering Canada each year (the Landed Immigrant Data System a.k.a. LIDS); and (2) selected fields from the personal income tax return (a.k.a. the T1). The landing record contains:

- demographic data on immigrants (age at landing, gender, countries of origin, mother tongue, intended destination);
- program data (e.g. immigrant category or class, special program codes principal applicant or spouse – dependent status);
- personal attributes at the time of landing (e.g. intended occupation, years of schooling, level of education, self-assessed knowledge of an official language).

The database includes information from the tax return, such as province of residence at the time of taxation, employment earnings, (un)employment insurance benefits, earnings from self-employment.

Geographical variables based on postal code are available to permit geo-coding. Also, information on the points assessed against selection criteria of immigrants entering Canada under the economic class (known as the CIC—Immigration Data System Overseas -IDSO) have been recently added to the database. Future development plans include adding the industry of employment information (SIC from the income tax form know as the T4 supplementary), and information from the tax T4 summary form to provide information on immigrant businesses. Other data elements from the federal immigration files may also be added.

3. COVERAGE AND REPRESENTATIVENESS

3.1 Coverage and Capture Rates

Currently, the database contains data on over 1.5 million records, or 58%, of the 2.6 million immigrants who landed in Canada between 1980 and 1995. In terms of age, over 66% of immigrants 18 years of age

or more and 21% of those under 18 who were admitted from 1980 to 1995 are captured in the IMDB. According to the categories of admission, the lowest overall capture rate is 63% for the family reunification category, followed by 64% for economic spouses and dependants. The highest rate is 74% for the humanitarian (refugees) category, while 70% of economic principal applicants who came between 1981-1995 are captured in the IMDB at least once.

Individuals included in the database are immigrants who were landed during the period and for whom at least one tax return was filed since their arrival in Canada. All immigrants do not necessarily file tax returns in each year after their arrival. Some do not enter the labour market immediately and some leave the labour market temporarily or permanently. For example, if an individual arrived in 1980, but did not file a return until 1990, they would be included in the 1990 tax year. Others may be absent from the database because they have left the country or because of mortality. Overall, for the 1995 tax year, the capture rate for all immigrants who were landed from 1980 to 1994 and were at least 18 years of age at landing was 58%.

It should be also noted immigrants who filed taxes prior to receiving their landed or permanent resident status in Canada will be also included in the database. Over 17% of immigrants in the IMDB filed tax returns prior to their year of landing, although this figure varies and in some years is as high as 43%. The database also allows for the monitoring of the labour market behaviour of persons who immigrate as children. As these children age, enter the labour force and begin filing taxes, they will be included in the database.

Table 1 shows that the percentage of working age immigrants 'ever present' for the 1981 to 1987 cohorts is about six points higher than the percentage for the 1989 to 1993 cohorts. The reason for this differential is the recourse to a supplementary SIN retrieval process which was applied to the earlier group: all doubtful matches were submitted to the Social Insurance Register for confirmation and the results incorporated in a subsequent re-extraction process. For the latter period, all doubtful matches were written to a separate file that is yet to be processed.

Table 1

Number of immigrant taxfilers present by landing year and tax year

landyr	tyr80	tyr81	tyr82	tyr83	tyr84	tyr85	tyr86	tyr87	tyr88	tyr89	tyr90	tyr91	tyr92	tyr93	tyr94
1980	44,479	58,030	58,167	57,855	56,123	55,[illegible]	[illegible]	57,156	57,331	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
1981	4,646	[illegible]	54,227	54,696	55,080	55,174	[illegible]	57,896	58,759	[illegible]	59,911	60,470	61,773	62,607	62,943
1982	2,772	[illegible]	49,[illegible]	51,386	53,253	53,773	[illegible]	55,916	56,933	[illegible]	57,736	58,000	59,283	60,395	60,763
1983	2,123	3,577	7,6[illegible]	[illegible]	37,952	39,594	[illegible]	42,380	43,034	[illegible]	43,659	43,896	44,573	45,588	45,791
1984	2,005	3,752	5,263	[illegible]	[illegible]	39,049	[illegible]	42,607	43,212	[illegible]	43,789	43,902	44,553	45,065	45,260
1985	1,568	2,328	3,868	5,054	[illegible]	[illegible]	[illegible]	41,050	41,929	42,[illegible]	42,186	42,293	42,867	43,375	43,230
1986	1,143	1,513	2,311	4,002	6,683	11,9[illegible]	[illegible]	50,879	52,614	53,105	53,252	53,196	53,773	54,361	54,246
1987	1,007	1,309	1,800	2,757	5,931	9,816	[illegible]	[illegible]	77,001	78,776	79,223	79,118	80,042	80,760	80,651
1988	834	723	785	906	1,339	2,855	4,635	[illegible]	[illegible]	74,498	76,177	76,774	77,579	77,948	77,931
1989	369	409	452	522	728	1,078	2,137	4,199	[illegible]	[illegible]	84,018	85,270	86,701	87,478	87,092
1990	296	338	359	429	555	807	2,247	7,062	9,365	[illegible]	[illegible]	92,176	95,073	96,732	96,530
1991	282	330	380	480	668	964	2,498	9,970	17,890	28,007	[illegible]	[illegible]	106,846	110,259	111,006
1992	253	283	307	343	469	660	1,832	6,573	11,803	23,946	31,382	[illegible]	[illegible]	116,493	119,427
1993	222	249	261	292	353	460	949	2,717	4,714	10,663	17,480	23,334	[illegible]	[illegible]	118,794
1994	152	188	190	219	264	319	470	956	1,707	3,461	5,848	10,539	15,092	20,759	[illegible]

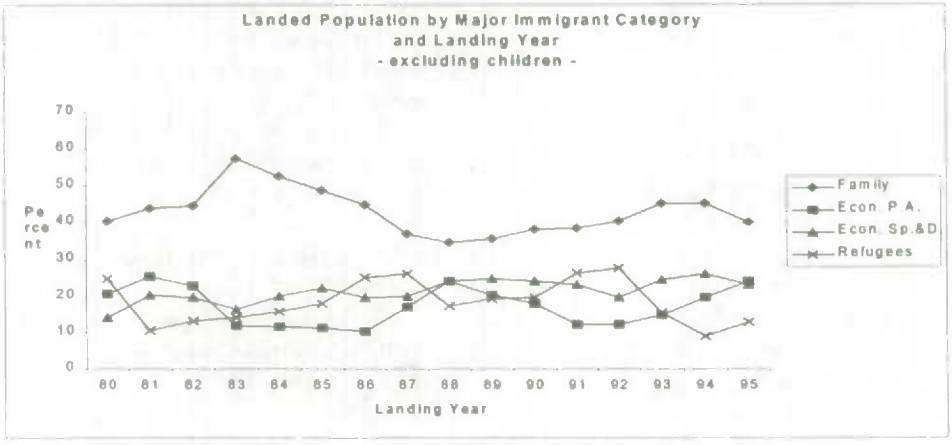
	Landings - All ages					Landings - 20 to 64 years old				
landyr	landings	evrpr	pcntepr	allpr	pcntapr	land_20	evrpr_20	pepr_20	allpr_20	papr_20
1980	143,464	83,450	58.17	35,172	24.52	88,270	59,600	67.52	31,735	35.95
1981	128,783	79,345	81.62	32,574	25.30	82,758	59,499	71.90	30,196	36.49
1962	121,287	75,918	62.59	32,001	26.38	79,668	58,617	73.39	29,981	37.54
1983	89,242	56,568	63.39	24,190	27.11	58,702	43,550	74.19	22,284	37.96
1984	88,380	55,858	63.20	25,921	29.33	58,888	44,384	75.36	23,946	40.68
1985	84,414	53,187	63.01	26,688	31.62	56,236	42,734	75.99	24,369	43.33
1986	99,668	66,141	86.36	35,594	35.71	67,536	54,112	80.12	32,486	48.10
1987	152,824	97,102	63.54	58,003	36.65	102,702	80,507	78.39	51,326	49.98
1988	162,104	92,540	57.09	57,503	35.47	104,094	75,983	72.99	52,165	50.11
1989	192,150	101,482	52.80	68,534	35.67	127,568	85,940	67.37	62,830	49.25
1990	215,101	111,096	51.65	80,260	37.31	145,785	96,652	66.30	74,110	50.84
1991	231,404	127,505	55.10	97,304	42.05	163,698	114,058	69.68	90,247	55.13
1992	252,847	132,976	52.59	110,842	43.84	177,348	119,662	67.48	101,770	57.38
1993	255,162	127,082	49.80	118,794	46.56	174,590	115,147	65.95	107,510	61.58
1994	222,573	88,545	39.78	0	0.00	149,349	61,525	54.59	0	0.00

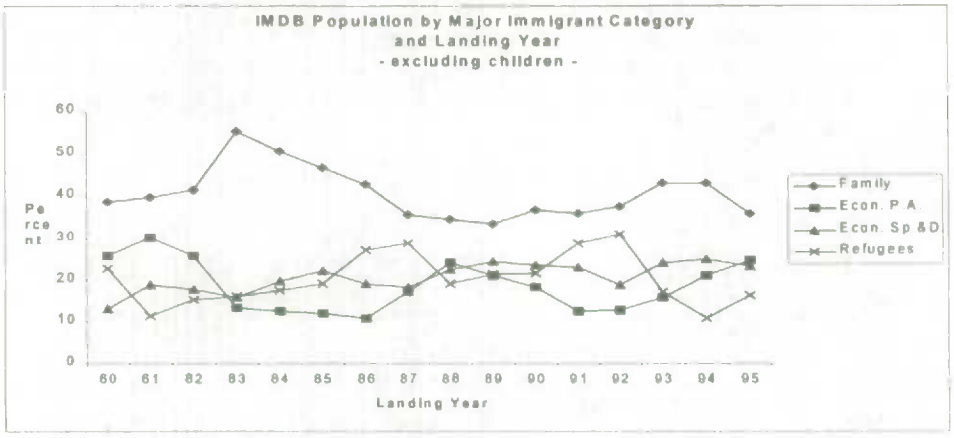
evrpr: present at least once
allpr: present in all years
pcntapr: % of landings always present from year after landing
pcntepr: % of landings ever present

Note: tax year less than landing year shows number who filed prior to landing

Charts 1 & 2 illustrate the overall representativeness of the population captured in the IMDB. The upper panel shows the percentage breakdown by immigrant category of the target population: adult immigrants

by immigrant category by year of landing. The lower panel shows the same breakdown for the immigrant taxfiler population captured in the IMDB. The fact that one is the mirror image of the other 'demonstrates' representativeness. The fact that the same correlation of capture rates to target population - with allowances for differential labour force participation - holds true for other characteristics such as age at landing, gender, level of education, knowledge of official languages, etc further reinforces this conclusion.





3.2 Representativeness

An analysis of the representativeness of the population included in the IMDB, 1980-1988, was carried out by the Quebec Ministry responsible for citizens and immigration (MRCI). The study concluded that the IMDB "appears to be representative of the population most likely to file tax returns. Therefore, the results obtained from the IMDB should not be inferred to the immigrant population as a whole, but rather to the universe of tax-filing immigrants". (Carpentier & Pinsonneault, 1994)

The characteristics of the immigrant taxfiler population will differ from those of the entire foreign-born population because the tendency or requirement to file a tax return will vary in relation to a person's age, family status, and other factors. One would expect a higher percentage of males to file a tax return, for example, because males have higher labour force participation rates than females. The extent to which immigrants are "captured" in the IMDB will also be influenced by changes to the income tax. For example, the introduction of federal and provincial non-refundable tax credit programs encourage individuals with no taxable income to file a return to qualify for certain tax credits.

The formal study of the representativeness of the 1980 to 1988 version of the IMDB was done before the incorporation in the database of additional records of doubtful matches were confirmed by the SIN Register (see description of this activity below). A descriptive analysis was then carried out for the entire timeframe of the current database. This analysis has shown that the degree of representativeness apparent for the earlier period is maintained in the later period. The conclusion of these later studies remains that the population captured in the IMDB is representative of the immigrant taxfiling population.

4. METHODOLOGY & DEVELOPMENT

The IMDB is created by a file linkage process, which matches individuals from the landing files to individuals on the personal income tax forms. As no individual identifier is common to both sets of files, personal data must be used to link these two files. The IMDB linkage process matches the first and last names, date of birth and gender of immigrants from the landing file to records on the taxation file for a particular tax year. All four personal attributes must be successfully matched a single time only on both files for an individual record to be "captured" in the database. Once an individual is "successfully" found in both files, then his/her corresponding social insurance number is retained for referencing in future tax years.

The IMDB is built on a foundation of administrative data from two separate sources produced over a sixteen-year period, which witnessed numerous policy, program, and administrative changes. This evolution and the nature of the programs - taxation and immigration - have major methodological implications for both the linkage and extraction processes underlying the construction of a longitudinal database.

One major peculiarity is the multiplicity of Social Insurance Numbers (SINs) which can be assigned to a foreign born resident. Temporary residents are assigned temporary SINs. Permanent residents are assigned permanent SINs. And, temporaries who become permanent are assigned new permanent SINs to replace their temporary SIN. The Social Insurance Number is therefore not unique.

As a result, the linkage process - which serves first to obtain a SIN - must be followed by a search for alternate SINs before the appropriate records can be extracted. And, the extraction of records must be done retrospectively in the event that the SIN obtained through linkage was not the first SIN assigned to that person.

The linkage process is also complicated by the occurrence of multiple records for the same individual within one of the data sources and the occurrence of records for the same individual with different names (married vs. single females) within both data sources.

Fortunately, it has been observed that foreign-born taxfilers have a strong tendency to use the name spellings recorded on their visas to fill out other official documents. It is because of this practice that the linkage and SIN assignment processes have been largely successful.

On the extraction front, the situation is complicated by the evolutionary formatting of the tax records. One of the major challenges was to standardize the formatting of the tax records over the entire 1980 to 1996 period, taking into account the changes in position as well as the changes in definition for tax purposes of the fields that were either changed or introduced over time.

5. ISSUES & CHALLENGES

The following are some of the ongoing challenges in developing and managing a database of this sort, as well as issues researchers should consider when contemplating to use information from the database. Some of the issues involved in constructing the database based on a record linkage process have already been discussed.

5.1 Conceptual limitations of the database

One of the conceptual weaknesses of the database for analytical purposes is the lack of a reference or comparison group of non-immigrant taxfilers. Limited comparison of certain categories of earnings broken down by gender, age groups and province of residence for the entire Canadian taxfiling population are possible based on the annual statistics published by Revenue Canada. It may also be possible to use Statistics Canada's longitudinal administrative database (LAD) as a reference group of resident (including immigrant) taxfilers.

For researcher who wish to compare the labour market behaviour of immigrants to their Canadian –born counterparts, however, it is not possible to do this with the IMDB. The creation of a reference group of taxfilers is being considered for future development of the database. However, it should be noted that information on the taxation files does not permit the clear identification of foreign-born individuals from the entire taxfiling population. If a reference group of taxfilers were created for the IMDB, then it would necessarily include both non-immigrants and immigrants who came to Canada prior to 1980.

The IMDB also does not contain information on the skills, education or language ability of immigrants acquired since arrival. There is no way of knowing this from the personal taxation form. Nor is information on the degree of labour force attachment known. This can only be inferred from the level of earnings. Finally, the IMDB is constituted of records and data on individual immigrants, so there is no family or household level information in the database, as of yet.

5.2 Technical Issues

Increasingly powerful personal computers and software have made it technically easier to build, store, and manipulate a database such as the IMDB, as well as to extract information from it for dissemination and research purposes. Personal computers have also made it possible to create the required source files to the IMDB, in particular the landing files. In the initial stages of the database, all of the processing, storage and extraction were done on the mainframe – which greatly added to the cost of the project, not to mention the limitations in extracting and processing information. Fortunately, enhanced desktop-computing power now make it possible to create anonymized copies of microdata files of the database and store this off the mainframe. This has also meant that requests for data from the IMDB can be done now at relatively reasonable costs to the researcher.

Another technical issue has been how to structure the files that make-up the database in order that cross-sectional and longitudinal analyses can be supported. At the same time, a structure was needed to support future expansion and development, and to have a database relatively easy to store, manipulate and extract information. The present structure of one file for all landing years and a file each for taxation year, with a unique record identifier, seems to be fulfilling these requirements.

5.3 Confidentiality & Privacy of the Information

The very idea of linking administrative data from different sources – even for the strict purposes of research – raises concerns about the protection of privacy. Any proposal to create and maintain a longitudinal database of information from administrative data sources about a target sub-population in Canada must weigh the benefits to be gained from this statistical information with adequate controls to protect this sensitive information.

To create a database such as the IMDB, the database must satisfy the requirements of the Statistics Act, the Privacy Act and the Income Tax Act. The confidentiality provisions of the Statistics Act provide strong protection of any personal or sensitive information in the IMDB. The Statistics Act prohibits the disclosure of identifiable information from the IMDB to anyone. No one, other than Statistics Canada's employees, may have direct access to the microdata records. All aggregate data (for example, in the form of cross-tabulations) from the database are screened for confidentiality and subject to random rounding. In addition, one of the conditions of approval of the linkage proposal to create the IMDB was that no public access micro-data file be created from the database.

The proposal to create the IMDB by means of linking two administrative files was reviewed and approved within Statistics Canada, according to the Agency's policy on record linkage. For example, it had to be demonstrated that the purpose of the linkage activity resulting in the IMDB was solely for statistical/research purposes and consistent with the mandate of Statistics Canada as described in the Statistics Act. It had to be demonstrated that the record linkage activity for the IMDB would not be used for purposes that could be detrimental to the individuals involved and that the benefits to be derived from such a linkage had to be clearly in the public interest. As well, the record linkage activity had to have demonstrable cost or respondent burden savings over other alternatives, or be the only feasible option; and, it had to be judged that the record linkage activity would not jeopardize the future conduct of Statistics Canada's programs. The IMDB project will be reviewed in the year 2000 to ensure that the continued benefits of developing such a database meet these conditions.

5.4 Accessibility & Dissemination

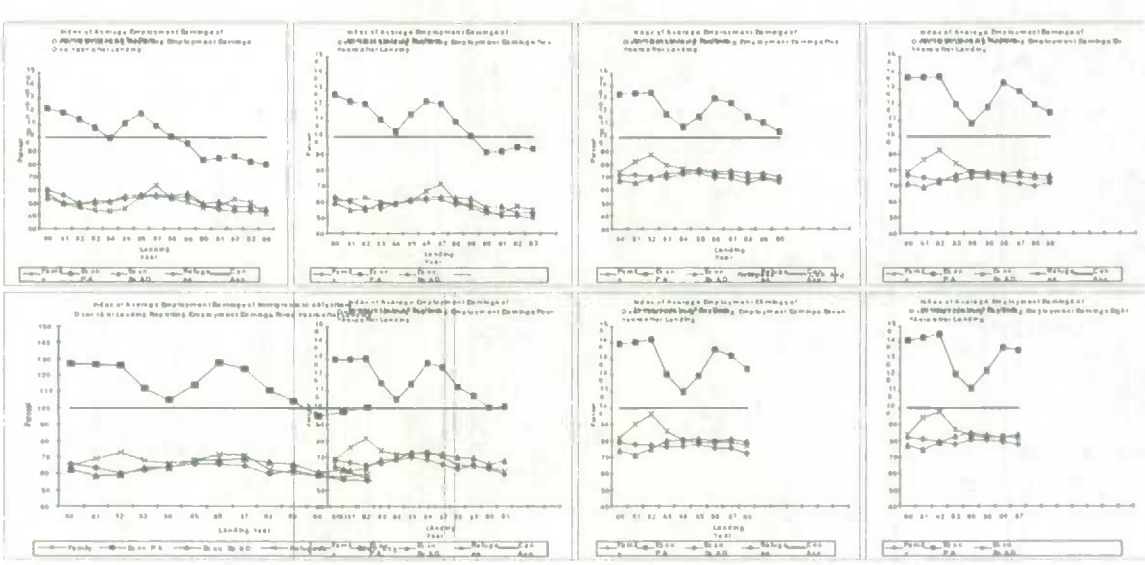
The database has been publicly available since August 1997. Requests can be made to Statistics Canada for cross-tabulations or other statistical queries to the database on a cost-recovery basis. As noted previously, the use of the IMDB is subject to the privacy and confidentiality constraints of the Statistics Act to prevent the release of personal information. All data released from the IMDB is screened for confidentiality and subject to random rounding.

One of the challenges for Statistics Canada is to make the information accessible to researchers. The database is relatively new among the research community. There is a need on the part of researchers to become familiar with both immigration and taxation policies and programs in order to exploit the database to its potential. Statistics Canada is also currently developing analytical capabilities and services for researchers who wish to do more in-depth analysis with the IMDB, for example, in the form of multivariate regression analysis. One possibility is to create a test file replicating the structure of the variables in the database, but with dummy data, in order for researchers to test their regression code.

6. POWER OF THE IMDB FOR POLICY-RELEVANT RESEARCH

The IMDB is the only source of information that links immigration policy levers with labour market or economic outcomes. It allows the researcher to 'target' a cohort admitted under a specific program and to compare that cohort's behaviour to other cohorts – at a point in time or over time. A researcher can identify immigrant specific characteristics that contribute to settlement success and identify immigrant groups that are encountering barriers..

This policy-related cohort targeting is illustrated in the series of charts showing the employment earnings patterns for different categories of immigrants *(Charts 3)*.



7. CONCLUSION

The Longitudinal Immigration Database is an example of administrative files combined to provide detailed and policy-relevant information on the economic behaviour of immigrants in Canada's labour market. It serves as a model for other possible projects using administrative files to inform the policy-making process, as well as a challenge to researchers who wish to undertake longitudinal analysis based on administrative files.

REFERENCES

Carpentier, Alain and Pinsonneault, Gérard (1994). *Representativeness Study of Immigrants included in the Immigrant Data Bank (IMDB Project).* Ministère des Affaires internationales, de l'Immigration et des Communautés culturelles, Québec

Langlois, Claude and Dougherty, Craig (1997). *Disentangling Effects: A Glimpse of the Power of the IMDB.* Presented at the CERF-CIC conference on Immigration, Employment and the Economy, Richmond, British Columbia.

Langlois, Claude and Dougherty, Craig (1997). *Preliminary Profile of Immigrants Landed in Canada over the 1980 to 1984 Period.* Presented at the CERF-CIC conference on Immigration, Employment and the Economy, Richmond, British Columbia.

Langlois, Claude and Dougherty, Craig (1997). *The Longitudinal Immigration Database (IMDB): An Introduction.* Presented at the CERF-CIC conference on Immigration, Employment and the Economy, Richmond, British Columbia.

Pinsonneault, Gérard (1998). *The Methodological, Technical and Ethnical Issues involved in using Administrative Data to Statistically Monitor the Immigrant Population: A few examples* Presented at "Séminaire conjoint France-Canada sur l'immigration", Montreal, Québec.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

COMBINING ADMINISTRATIVE DATA WITH SURVEY DATA: EXPERIENCE IN THE AUSTRALIAN SURVEY OF EMPLOYMENT AND UNEMPLOYMENT PATTERNS

Mel Butler[1]

ABSTRACT

One method of enriching survey data is to supplement information collected directly from the respondent with that obtained from administrative systems. The aims of such a practice include being able to collect data which might not otherwise be possible, provision of better quality information for data items which respondents may not be able to report accurately (or not at all) reduction of respondent load, and maximising the utility of information held in administrative systems. Given the direct link with administrative information, the data set resulting from such techniques is potentially a powerful basis for policy-relevant analysis and evaluation. However, the processes involved in effectively combining data from different sources raise a number of challenges which need to be addressed by the parties involved. These include issues associated with privacy, data linking, data quality, estimation, and dissemination.

KEY WORDS Administrative data: Linking: Longitudinal: Survey.

1. THE SURVEY OF EMPLOYMENT AND UNEMPLOYMENT PATTERNS

The main purpose of the longitudinal Survey of Employment and Unemployment Patterns (SEUP) conducted by the Australian Bureau of Statistics (ABS) was to assist in the evaluation of the government's labour market assistance programs and the survey design was therefore focused towards this aim. However, the design also allowed for the provision of general information about the dynamics of the labour market.

The main features of the survey were:

annual collection of data from the same respondents by personal interview over three years;

an initial total sample size of 8600 individuals aged 15 to 59 years;

a multi-panel design - jobseekers, people known to have participated in labour market assistance programs, and a cross section of the general population;

a strong focus on episodal (spell) data describing respondents' labour market activities (that is, whether they were working, looking for work, or absent from the labour market) and, most importantly in the context of this paper;

with respondents' consent, the supplementation of data collected direct from them with information from the administrative systems of two other government departments.

[1] Australian Bureau of Statistics

More detailed information about the SEUP can be found in *Australians' Employment and Unemployment Patterns 1994 - 1997* (ABS Catalogue no. 6286.0) or on the Internet at http://www.abs.gov.au.

2. DATA LINKING - WHY DO IT IN THE SETUP?

Given the main purpose of the SEUP, it was important that the data set contain high quality information about respondents' participation in labour market assistance measures (referred to as Labour Market Programs, or LMPs). Also, there was strong interest in the relationship between a person's receipt of income support (government benefits) and their labour market behaviour.

At the time of the survey, administrative data bases on these two topics were maintained by the Department of Employment, Education, Training and Youth Affairs (DEETYA) and the Department of Social Security (DSS). Given known difficulties in collecting such information from respondents, it was decided that, with individual respondents' consent, the required information about LMP participation and income support would be best sourced from these administrative systems. There were three main reasons for this:

to minimise both respondent load (particularly given that SEUP was longitudinal) and interview costs;

to provide more comprehensive and more accurate data than would be possible with a normal survey approach; and

a desire to increase the use of administrative data for official statistical purposes.

The proposed combination of data collected direct from the respondent in a household survey with that from the administrative systems of government departments was a new approach for the ABS to adopt and it was to be expected that many substantive issues of policy, logistics, and methodology would be raised. However, given the context of this paper it is appropriate to focus mainly on two issues - the success rate in obtaining respondents' consent to such data linking, and some of the implications of combining data for estimation and analysis.

3. PROCEDURES

3.1 Level of respondent consent

There was naturally some concern about the likely reaction from respondents to the ABS seeking their consent to data linking of this type, which could range from refusal to agree to release information required to effect a link, through non-participation in the SEUP, to public distrust of the ABS's confidentiality pledges which might potentially affect all collections. However, following careful consideration of privacy issues, and appropriate interviewer training, by the third and final wave of the survey these worries proved to be unfounded and consent rates for the various panels were as follows.

Table 1: Consent rates

		Percentage consenting to linking with:
Panel	DSS records	DEETYA records
Jobseekers	84.8	83.5
LMP participants	90.4	93.4
General population	50.7	39.7

Two observations can be made about these data. First, for the Jobseeker and LMP panels, the consent rates were very high. The significantly lower rate of consent from the General population panel was almost certainly because many of the people in this panel had never been clients of either department (for example, because they were 'long term employed'), and hence they perceived the whole idea as irrelevant and an invasion of privacy. Conversely, those in the LMP panel had all been clients of DEETYA (and probably a high proportion had been clients of DSS too), and saw the appropriateness of giving consent to allow access to information about them. Second, the higher rate of consent from two panels (and particularly from the General population) to DSS linking is interesting, given that DSS records contain financial information, which is normally a sensitive area for respondents, leading to higher item non-response in direct questioning.

3.2 Linking to administrative data

The information collected from respondents in order to effect a link with administrative data was: full name, date of birth, sex, address and, if known, the identification (client) number used by each department. Linking with the relevant data bases (both very large) was undertaken by each of the two departments. The success rates (ie, client records located within a defined three year period) are shown below.

Table 2: Link rates for consenters

		Percentage linked with:
Panel	DSS records	DEETYA records
Jobseekers	65.6	90.0
LMP participants	66.3	98.8
General population	40.3	48.6

The absence of a link being made could be either because of imperfections in the linking routine (a concern) or because the respondent had not been a client within the reference period (not a concern). While the availability of the client reference number was a very useful linking tool, it was not foolproof - for example, instances of transposed digits, and respondents providing their partner's number were not uncommon. However, given the combination of variables used as a linking key, and clerical intervention to resolve 'near links' (such as misspelling of non-english names), it is considered that the procedure used resulted in a very high link rate for people who had been clients of the departments.

The almost 100% link of the LMP group to DEETYA records was to be expected given that these records were the original source of the sample for this panel (the small amount of non-link is because this panel was identified before the start of the three year reference period). Similarly, a high proportion of people in the Jobseeker panel would have been DEETYA clients because of registration with the Commonwealth Employment Service administered by that department. However, because they might not meet eligibility requirements for income support, there are lower link rates with DSS for both these panels. The much lower link rates for the General population were also to be expected given the relatively low incidence in the wider population of receipt of both DEETYA and DSS services.

4. A SUMMARY OF ISSUES ARISING

4.1 Logistics and data quality

Obtaining the information from administrative sources proved to be a time consuming and non-routine process, which was largely outside the control of the ABS (an added complication for any statistical agency).

Typically such data bases are not designed to facilitate the extraction of statistical information. They are subject to changes in service delivery policies and practices over time (such as eligibility for benefit payments), to changes in information technology platforms (necessitating reprogramming of linking routines), and the priority ascribed to the task was not always high, given more mainstream demands.

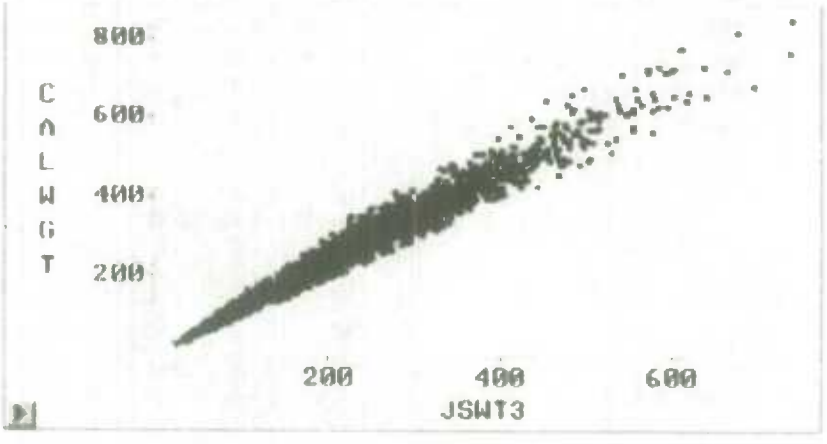
As a result, administrative data received by the ABS from the departments was far from perfect. For example, it was relatively common to receive, initially, records that had episode start dates after the end dates, incomplete records, invalid codes, implausible values, and theoretically impossible combinations of episodes. This may be due to turnover of staff, data recording not being the principal objective in workers' jobs, and changes in policy and administrative procedures. Also, because of lags in administrative processing, a link undertaken at one point in time may yield different outcomes to one undertaken at a subsequent point. Traditional editing by the ABS, together with cooperation in solving queries by the departments involved, was essential in 'cleaning up' the data, including the generation of a 'seamless' data set over a contiguous three year reference period.

4.2 Treatment of non-consent

An important statistical issue was the question of how to deal with records where there was no consent to link with administrative data, a key question being whether these people had different characteristics to those who did consent. From the point of view of being able to reliably and simply incorporate the data sourced from DEETYA and DSS in tabulations and analyses it was most desirable that estimates using the administrative data could be made in respect of the total population, rather than having a 'not known' category because of a lack of consent.

In order to achieve this, new weights were attached to the records of respondents who consented to account for those who did not consent. These were derived by a two-stage process. First, modelling was undertaken to establish which variables were most related to consent - all variables available were correlated against the consent flag, and those with the highest correlations (that were suitable) were put through a stepwise logistic regression. Second, calibration was used to adjust the original or "input" longitudinal weights so that they met a number of benchmark constraints. This was done in a way that minimised the change in the input weights, through the SAS macro CALMAR. This process was undertaken for each of the panels, and for each type of consent (that is, consent to linking to DSS, to DEETYA, and to both). As an example of the result, the following graph shows how much the calibration changed the input weights; the x-axis are the initial weights, and the y-axis are the new calibrated weights.

Figure 1: Plot of DSS consent weights (CALWGT) and input weights (JSWT3) for the jobseeker panel



Estimates using administrative data therefore aligned, at broad benchmark levels, with population estimates produced from other data collected in the survey. However, when analysing variables other than those used in the calibration exercise some relatively minor differences in estimates will be observed for individual characteristics.

However, analysts wishing to work with unweighted observations in modelling applications need to make their own decisions as to how to treat non-consent.

4.3 Data confrontation

Combining the data obtained from the respondent with that sourced from administrative systems raised a number of data confrontation issues which, as well as posing statistical challenges, enabled new light to be shed on old issues.

One example is the ability to compare, at an individual level, labour force status (collected at each interview according to the International Labour Organisation standard) and LMP participation (obtained from administrative data) for the same point in time - in effect, to answer the often asked question 'How do people on labour market programs report their labour force status?'. Summary results of this comparison are shown below.

Table 3: Relationship between LMP participation and labour force status

	Labour force status (%)			
LMP type	Employed	Unemployed	Not in labour force	Total
Training programs	25.6	50.6	23.8	100.0
Jobskills	95.1	0.0	4.9	100.0
New work opportunities	98.3	1.7	0.0	100.0
Jobstart	98.8	0.8	0.6	100.0
National training wage	100.0	0.0	0.0	100.0
Other	78.8	4.5	16.7	100.0

Close to 100% of participants in the major employment programs reported that they were employed while on the program and the majority of participants in other employment programs also stated that they were employed (79%). However, participants in training programs had quite different labour force status patterns - half stated they were unemployed, a quarter were not in the labour force and the remaining quarter were employed, mostly part-time. Intuitively, this is the type of result that would be expected, and hence there is no conflict between the two data sources.

Related to this was the extent to which the self-reported episodal data about labour market activity (LMA - 'working', 'looking for work', or 'absent from the labour market') could be related to the data on LMP participation obtained from DEETYA. This was important from at least two perspectives - being able to distinguish between 'normal' work episodes and those work episodes that were in some way associated with a LMP (such as a subsidised job), and because of a requirement, for some analyses, to be able to treat as non-concurrent those episodes of LMP participation which were undertaken within a spell of looking for work.

Exact matches, date-wise, between directly collected data and administrative data were infrequent. This was partly because participants often did not report a change in their labour market activity while on an LMP; it may also have been because many respondents did not recall the precise dates when they started or

finished a particular activity, and/or because the dates on the administrative files were not the exact dates of participation.

An algorithm was therefore developed to assess the degree of relationship between the differently sourced data. Two ratios were calculated. The first expressed the period of overlap as a percentage of the participant's time in the LMA, and the second expressed the overlap as a percentage of the participant's time on the LMP. The algorithm used these percentage figures to determine which labour market activity best matched the LMP, and identified two kinds of association between LMP and LMA.

A **direct association** was established when the LMP and the LMA coincided very closely in time. For example:

Case 1: A person in wage subsidised employment stopped looking for work while in that LMP. Here the participant was on a program in January, February and March and reported in the survey that they were working for those months. The working episode could therefore be directly associated with the LMP. The respondent looked for work both before and after they were on the program.

Figure 2: Case 1

Episode	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
LMP					XXX	XXX	XXX					
Looking for work	XXX	XXX	XXX	XXX				XXX	XXX	XXX	XXX	XXX
Working					XXX	XXX	XXX					

Case 2: A person was on a wage subsidised employment LMP in January, February and March and reported that they were working for those three months. The working episode was directly associated with their LMP. However this person also reported that they were looking for work for the whole twelve month period (the model adopted in the SEUP allowed overlap of 'working' and 'looking for work' episodes).

Figure 3: Case 2

Episode	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
LMP					XXX	XXX	XXX					
Looking for work	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX
Working					XXX	XXX	XXX					

An **indirect association** was established where the respondent did not report a change in LMA while on an LMP. For example:

Case 3: A respondent had a long period of looking for work during which they participated in two month training program. The respondent did not report any change in their 'looking for work' labour market activity.

Figure 4: Case 3

Episode	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Training program			XXX	XXX								
Looking for work	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX	XXX
Working											XXX	XXX

This indirect association was the most commonly reported pattern for training programs.

Only a minority of respondents had LMA/LMP patterns which did not fit these typologies.

4.4 Impact on microdata release

The combination of data from different sources also had an impact on the dissemination of microdata from the survey. The concern from the ABS was that the administrative data would be able to be readily matched with the source data bases held by the two departments (which were likely to be the major clients for the survey data!), and, further, that the potential existed for third parties to also match the data if they had access to these departments' data bases. These concerns were addressed by:

excluding all the administrative data items from the main microdata file, but providing a set of 500 records with synthesised administrative data items to facilitate analysts developing and testing models, which could subsequently be run by the ABS against the main file;

notwithstanding the above step, not allowing either DEETYA or DSS access to the public microdata file on their own premises; but

providing a 'data laboratory' facility so that DEETYA and DSS officers could access the equivalent of the public microdata file, and the data sourced from their departments, on ABS premises under secure conditions.

5. SUMMING UP

Notwithstanding the issues that arose in combining the data from three different sources, and some lengthy delays in the availability of data from administrative sources, the resultant longitudinal data set is unique in Australia and offers considerable scope for analysis of a range of social and economic topics over and above what would have been possible with a 'single source' approach.

At the same time, the data set is complex, and the relationships (and inconsistencies) between the different types of data require considerable thought on the part of analyst in order to maximise its potential.

Two specific research projects are presently benefiting from the availability of the combined longitudinal data set:

an investigation into labour market programs, unemployment and employment hazards (by Thorsten Stromback and Michael Dockery, Centre for Labour Market Research); and

a study of the dynamics of welfare receipt and labour market transitions (by Guyonne Kalb, Social Policy Research Centre).

However, it is only in recent times that analysts are starting to intensively use the data so it will be some time before there can be a fuller evaluation of the exercise.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

PROJECT OF LINKAGE OF THE CENSUS AND MANITOBA'S HEALTH CARE RECORDS

Jean-Marie Berthelot[1], M.C. Wolfson[1], C. Mustard[2]

ABSTRACT

The current economic context obliges all partners of health-care systems, whether public or private, to identify those factors that determine the use of health-care services. To increase our understanding of the phenomena that underlie these relationships, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation have established a new database. For a representative sample of the province of Manitoba, cross-sectional micro-data on the level of health of individuals and on their socioeconomic characteristics, and detailed longitudinal data on the use of health-care services have been linked. In this presentation, we will discuss the general context of the linkage of records from various organizations, the protection of privacy and confidentiality. We will also present results of studies which could not have been performed in the absence of the linked database.

1. INTRODUCTION

Socio-economic status - as measured by income, education or occupation - is a complex phenomenon used to describe social inequalities. Since the end of the 1980s, it has been shown that people in lower socio-economic categories experience higher mortality rates and poorer health than those further up the social ladder (House, 1990 ; Mackenbach, 1992 ; Adler, 1994). To improve our understanding of the phenomena underlying the relationships between SES and health, Statistics Canada and the Manitoba Centre for Health Policy and Evaluation (MCHPE) took the necessary steps in 1989 to create a new database with administrative health records from different sources and data from the 1986 Census of population. From 1991 to 1993, matching operations allowed the linkage, for a representative sample of Manitoba's population, of detailed longitudinal data on the use of health care services with cross-sectional microdata on the health status and socioeconomic characteristics of individuals. This initiative also satisfies an important recommendation of the 1991 report of the National Task Force on Health Information (Wilk report), produced by the National Council on Health Information: "Capacities to link data elements are of crucial importance for health information development and those capacities must be expanded and exploited".

2. CONFIDENTIALITY AND PRIVACY

In creating a database from both administrative and survey data, it is of the utmost importance to ensure the confidentiality of the data and prevent any invasion of individual privacy. In accordance with the policies of the collaborating organizations, several procedures were undertaken prior to matching these data sets. They included consultations held in 1989-90 with the Privacy Commissioner of Canada, the Faculty

1 Jean-Marie Berthelot, Statistics Canada, R.H.Coats building, 24th floor, Ottawa(Ontario), Canada K1A 0T6 ; e-mail : berthel@ statcan.ca
2 Dr Cameron Mustard, Institute for Work and Health, 250-Bloor Street East, Toronto (Ontario), M4W 1E6

Committee on the Use of Human Subjects in Research at the University of Manitoba, and Statistics Canada's Confidentiality and Legislation Committee. In addition, Manitoba Health's Committee on Access and Confidentiality was made aware of the project.

Following these consultations, the proposal was submitted to the Minister responsible for Statistics Canada, as required by the agency's official policies[2], and ministerial approval was obtained in 1991. It was to be a pilot project to assess the feasibility and analytical benefits of matching. Individuals' names and addresses were not used for matching purposes and were not included in the database. The matching process was carried out entirely on Statistics Canada premises by persons who had taken the Statistics Act oath. Only a sample of 20,000 matched units was used for research and analysis. Access to the final data is tightly controlled in accordance with the provisions of the Statistics Act. All uses of the new file generated by the matching process are covered by a tripartite agreement signed in 1993 between the University of Manitoba, the Manitoba Ministry of Health and Statistics Canada. Consequently, solid safeguards to prevent privacy violations and release of confidential information are entrenched directly in the pilot project's *modus operandi*.

3. DATA SOURCES

The detailed questionnaire (questionnaire 2B) of the 1986 Census of Population contains extensive socio-economic information including variables such as dwelling characteristics, tenure, ethnic origin and mother tongue, as well as a number of variables relating to income and educational attainment. Data about other members of the household, family structure and neighbourhood are also available. The computer file used for matching purposes consisted of 261,861 records. For a subset of 5,342 people, it was possible to obtain health information from the 1986-87 Health and Activity Limitations Survey. Once weighted, these files can provide reliable provincial cross-sectional estimates as of June 3rd, 1986.

All contacts between an individual and the Manitoba public health care system are recorded for administrative purposes. The Manitoba Health longitudinal files contain information on visits to physicians, stays in hospital, diagnoses, surgical procedures, admission to personal care (nursing) home, health care received at home, the date and cause of death, and other data on health care utilization. For this pilot project, a register of persons covered by Manitoba health insurance was identified as of June 1986, using the date of commencement of health insurance coverage and the date of cancellation of coverage around the target date. The register contained around 1,047,000 records.

---

2 This footnote and the one on the following page describe briefly Statistics Canada policies about linkage:

The purpose of the linkage must be consistent with Statistics Canada's mandate.
The products of the linkage will be subjected to the confidentiality provisions of the Statistics Act with any applicable requirements of the Privacy Act.
The linkage process will reduce costs or burden, or is the only feasible option.
The linkage activity will not be used for purposes that can be detrimental to the individuals involved.
The analytical benefits to be derived from such a linkage are clearly in the public interest.
The record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs
The linkage satisfies a prescribed review and approval process

4. CREATION OF THE ANALYTIC DATABASE

Data from the 1986-87 Health and Activity Limitation Survey (HALS), the 1986 Census and the files of Manitoba Health were linked (without using names or addresses) by a software package named CANLINK, using postal code and family structure. The linkage rate was 74% for private households. A quality assessment based on names and addresses for a small sub-sample showed that the overall rate of agreement among the matched pairs was 95.5%. The analytic database is made up of 20,000 units selected from the set of matched households using modern sampling techniques. Because extending the HALS complex sample design would force the development of specialized analytical tools and delay the analysis itself, two different sub-bases were created.

One of these sub-bases consists of all HALS respondents who reported an activity limitation in 1986 (4,434 individuals). This database is used primarily for analysis of disability-related questions.

The other sub-base, known as the "general population database", was selected independently of the HALS sub-base. Since the upper limit was 20,000 households, there remained 15,566 units available to form the general population database for Manitoba, plus the expected number of units overlapping both sub-bases. The final database contains data about approximately 48,000 people for analysis purposes. Estimates based on the sample accurately reflect the socio-demographic profile, mortality, hospitalization, health care costs and health care services utilization of Manitoba residents (Houle, 1996). Following these steps, the analytical phase of the project began in 1994.

5. OPERATIONALIZATION OF THE ANALYSIS

The analytical team for this project consists of members of the MCHPE at the University of Manitoba (Winnipeg) and members of the Health Analysis and Modeling Group (HAMG) at Statistics Canada (Ottawa). The MCHPE is a unit of the Department of Community Health Services in the University of Manitoba Faculty of Medicine. Research activities are focused on health services and health policy analysis and evaluation, for the most part using the Manitoba Health Data Base to describe and explain patterns of care and profiles of health and illness. Like other academic institutions, the MCHPE is funded partly through a general budget and partly through grants for specific research projects.

The HAMG is a unit of Statistics Canada's Social and Economic Studies Division which is block funded to undertake health-related data analyses, one of which is the Census/Manitoba linkage project.

Communication between the two groups is ongoing and makes use of electronic mail, conference calls, regular mail and ad hoc meetings. Every research proposal is reviewed by at least one member of each research group; this enables the groups to share their expertise and avoid duplication. Special procedures assure that no confidential data is exchanged by public electronic networks in order to protect confidentiality. A copy of the final analytical database is stored at the Ottawa premises of Statistics Canada, and another copy is stored at the Winnipeg regional office premises of Statistics Canada. Access is restricted to authorized users who have been sworn under the Statistics Act. All authorized users, after taking the oath set out in the Statistics Act have a duty of secrecy. If they divulge directly or indirectly any confidential information, they become liable on summary conviction to a fine or to imprisonment.

6. RESULTS

The aims of the project were, in the first phase, to determine the feasibility of matching census data with Manitoba Health files without using names and addresses, and then in the second, to initiate research into

the relationships among socioeconomic status, health, and health care utilization.

Phase one was a success, with a match rate of 74% and an accuracy rate of over 95%. In phase two, which is still in progress, we undertook analytical studies that exploited the database's unique features; they could not have been attempted without the database. Some of the studies confirmed hypotheses concerning the relationship between socioeconomic status and use of health care services, while others quantified the range of these inequalities for the first time, and still others produced new findings.

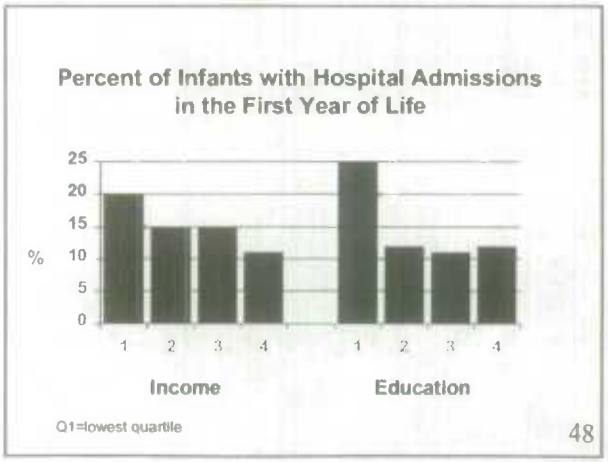
We studied a wide range of issues, but our findings are centered on two main themes. First, there is the potential for improving the existing health information system. Our research yielded a profile of the socioeconomic gradients of mortality and health care service consumption at various stages in individuals' lives, and enabled us to quantify non-medical factors (socioeconomic status, marital status, pension income) and disability as predictors of health care use. For example, children of poorly-educated women consume twice as much health care in their first year as children of well-educated women; asthma is more common in the socio-occupational groups defined by teachers, electricians and labourers ; life expectancy is lower for individuals with less education; the risk of being admitted to a health care institution is lower for homeowners, for men who have a spouse and for women who have investment income; and people with severe disabilities are three times more likely to die and four times more likely to be admitted to a nursing home than individuals with no disabilities, even after controlling for age, gender and number of chronic conditions.

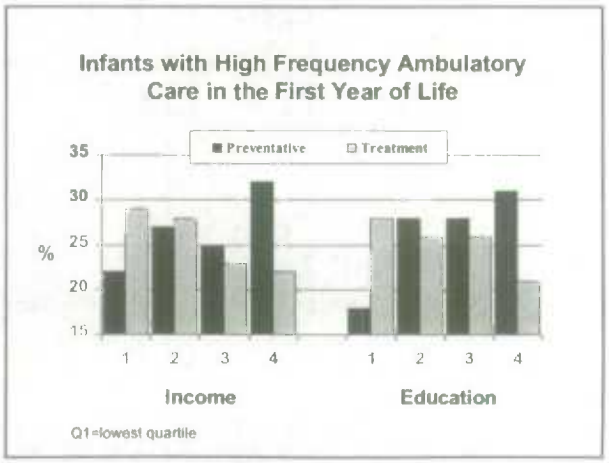
This theme raises questions about existing health care management databases which contain little or no non-medical information. For example, could the predictive value of "case mix groupings" (CMGs) produced by the Canadian Institute for Health Information be improved by including disability level ? If a second phase of this project lead us to a larger sample, a more detailed analysis concerning the treatment of specific diseases could be undertaken to determine more precisely the effects disability has on health care service consumption.

The second theme is that while access to health care is universal when people are sick, the health care system may not be universal for preventive care. For example, our studies have shown that access to or use of preventive medicine in the first year of life is less frequent among children of poorly-educated women while hospitalization is more frequent (indicating poorer health for these children). We also found that in a four-year period, 27% of women in the lowest income or education quartile did not have a Pap test, compared with 16% of women in the highest quartile. These differences precede the intervention of the "formal" health care system. One of today's biggest public health challenges is how prevention, effected through public health policy or by changing the way patients and doctors interact, can be used to reduce the differences. Our studies help identify some of the groups at risk.

The following summaries describe in more details the most important analytical results emerging from this project.

- *Socioeconomic gradient in the use of health care services in the first year of life. (Knighton et al., 1998)*





This is the first Canadian study to examine the relationship between health care use in a child's first year of life and the mother's socioeconomic status. With the exception of low-birth-weight babies, no significant differences were observed at birth. Yet, health care costs in the first year of life are about twice as high for children of women in the first income or education quartile as for children of women in the fourth quartile. If health care services are divided into two broad groups, "curative" hospital and ambulatory care is inversely related to socioeconomic status, with a gradient effect relative to income and a threshold effect relative to education. On the other hand, individuals of higher socioeconomic status make greater use of "preventive" health services, with the same type of income gradient effect and education threshold effect. These findings are important for three reasons. First, since the mother's socioeconomic status is determined prior to the birth of her children, the study supports a cause-and-effect relationship. Second, the study suggests an underuse of preventive care by mothers in the low socioeconomic group. Third, the study identifies women with a low level of education as a likely target group for public-health intervention, both during pregnancy and after delivery.

- *Use of public health care services by socioeconomic status (Mustard et al., 1995;Mustard et al., 1997)*

The detailed information that would be needed to prepare a comprehensive description of the differences in the use of state-funded health care based on socioeconomic status is almost non-existent in Canada. This study produced a report containing a wealth of information about mortality, morbidity and health care costs. In an article summarizing the report, ambulatory medical care and hospital visits are examined in relation to education and income quartiles. The results indicate that the lower the income or education level, the greater the use of hospital services, though paradoxically, there is no similar relationship between income or education level and ambulatory care. The reasons for this paradox and the implications for the organization, funding and delivery of public health care in Canada are also discussed in the report.

- *Factors associated with senior citizen admissions to health care institutions*

In this analysis, admissions are examined in relation to Andersen's conceptual framework, which groups variables on the basis of three factors: predisposing factors (age, education and marital status), enabling factors (income, rural area, social support, available services) and needs-based factors (limitation, disease, care consumption). Using Andersen's framework, the authors take a multivariate, structured approach to the study of factors related to admissions. In addition to repeating known findings – i.e. age, functional limitations, health care use and disease play a dominant role in admissions – the study reveals factors associated with social support and socioeconomic status. The risk of admission is lower for homeowners (the relative risk (RR) is 0.51 for men and 0.62 for women), men who have a spouse (RR=0.61) and women who have investment income (RR=0.69). The study identifies potential intervention points for public-sector decision-makers: improving the social support network, diversifying retirement income and controlling certain diseases could reduce the need for institutionalization. The analytical team has initiated a follow-up study to take a closer look at the social support network's impact on admissions using information about "social services" for disabled persons in the Health and Activity Limitation Survey.

- *Socioeconomic status and Pap test use among women between 21 and 65 years of age.*

Early detection of cervical cancer by means of a Pap test is a major success story for disease prevention programs in Canada. Cross-sectional data on Pap tests reported by Canadian women reveal, however, that the test is underused by economically-disadvantaged groups. Very few analytical studies have been conducted with longitudinal data on the actual use of Pap tests by socioeconomic status. In this study, we analyzed longitudinal use of Pap tests over a four-year period by a sample of 10,868 women. We found that during the observation period, 21% of the women did not undergo a Pap test, and that the percentage was inversely related to socioeconomic status, ranging from 27% for women in the first income or education quartile, to 16% for women in the fourth quartile. Use of the Pap test is also associated with frequency of contact with the health care system and with area of residence (rural vs urban). The findings suggest that an awareness campaign for economically disadvantaged women might be effective in improving uptake for this group. In addition, disseminating this information to the medical profession might help doctors to develop a proactive approach to women who have irregular contact with the health care system.

- *The incidence of asthma by socio-occupational group*

There is currently a major gap in Canada in the monitoring of diseases associated with or caused by particular occupations. Because of the size of the sample, only relatively common conditions can be studied by socio-occupational group. In this study, the analytical team examined the relationship between asthma and socio-occupational group within the labour force. After age, education, area of residence, marital status, income and health care service consumption are taken into account, teachers, electricians and labourers have a significantly higher rate of treatment for asthma. The study demonstrates the value of monitoring diseases whose causes may be linked to occupational exposures.

- *Life expectancy by education quartile and income quartile*

Death certificates in Canada contain no socio-economic information about the deceased. The database was used to produce the first estimates ever made in Canada of life expectancy by education quartile. The difference in life expectancy between the first and fourth education quartiles is 3.2 years for males and 2.2 years for females. In addition, for the first time in Canada, individual data were used to generate estimates of life expectancy by income quartile. These estimates were compared with estimates calculated using ecological data. The results show a surprising level of agreement between the individual and ecological data. Life expectancies by education and income quartile can be used to calculate composite indicators such as disability-free life expectancy by education and income level. Measured on a regular basis, these indicators provide a picture of socioeconomic inequalities in health over time. The reduction of these inequalities may become an objective for health policy.

- *Exploratory use of hierarchical models in the health field*

The level of agreement for life expectancy between individual and ecological data motivated us to find analytical tools to quantify and document the neighbourhood effect. This effect, which has been studied in some detail in education, has been largely ignored in the health field. In this exploratory study, we used hierarchical models to study admissions to health care institutions in an attempt to answer the following question: Does the socioeconomic environment (neighbourhood) in which a person lives have an effect on his/her health independent of his/her individual characteristics? In this particular case, it is clear that the neighbourhood, as we have defined it, has little or no effect on admissions. To gain a better understanding of the neighbourhood effect on the health of individuals, we need to examine other events, such as health care consumption in the first year of life, mortality, and accidents. Because of the sample size, the current analysis is limited – i.e. we cannot define small size neighbourhoods. If a significant and independent neighbourhood effect on health care utilization is ever demonstrated, data collection and dissemination would have to be redesigned to allow health analysts to take this phenomenon into account.

- *Profile of health care use by individuals with disabilities*

The 1986 Health and Activity Limitation Survey (HALS) produced a highly detailed profile of the health of persons with disabilities. However, it did not provide any information about the effects of disabilities on health care consumption. This study is the first in Canada to document health care use and mortality among persons with disabilities, thus complementing the HALS. It shows that annualized and age-standardized rates for mortality, admissions, ambulatory care and hospitalization days are 30% to 150% greater for persons with disabilities, after controlling for the other important variables. The rates vary substantially by sex and level of disability. Surprisingly, men with slight disabilities have a significantly smaller chance of being institutionalized or dying than men with no disabilities. No similar phenomenon was observed among women. Severe disability shows a threshold effect, tripling mortality and quadrupling institutionalization. Disability level has a significant effect on health care consumption, even when the number of chronic conditions present at the beginning of the study period are controlled for. A more detailed analysis will be possible if the sample size is increased, particularly concerning the treatment of specific diseases to determine more precisely the effect that disability has on health care use. This study raises important questions about the allocation of health care system resources. If the findings of the detailed study are conclusive, should we collect raw data on disabilities at a detailed geographic level, so that decision-makers can take the population's disability profile into account in allocating resources to the regional or local level?

- *Unemployment and health status.*

An important question in labour market analysis is whether the unemployed subpopulation of the labour force is identical in health status to the population of workers. Using census data, the analytical team compared health care use by the employed and the unemployed before, after and at the time of employment status measurement. If the hypothesis is true that the unemployed group is identical to the labour force as a whole, then health care consumption should be similar for unemployed people and employed people before a period of unemployment. Preliminary results contradict the hypothesis. The unemployed appear to be chronically in poorer health than the employed. This kind of information is of interest, not only to health policy makers concerned with the "socio-medical" aspect of the unemployed, but also, by labour analysts involved in employment insurance.

7. DISCUSSION

We are convinced that the two original goals described at the beginning of section 6 for this pilot project have been met and that the analytical results are important. Furthermore, there is the possibility, in a second stage of this pilot project, of using all matched individuals instead of the 20,000 sampled units. A larger sample size would generally allow for the study of association between socioeconomic status, activity limitations, and health care utilization at a much finer level and would consequently improve the capacity of the database to identify target groups for health policy interventions.

Following a new round of consultations with privacy commissioners and considering a process similar to the pilot project, a new project is proposed. We would like a) to acquire the longitudinal health data from 1991 to 1997 ; b) to increase the sample size for the province of Manitoba to 15% of the total population which would result in a final sample size of 150,000 individuals; c) to obtain the HALS records for the additionnal individuals in the sample; d) to produce a similar database by linking health utilization data, HALS data and Census data for a 15% (450,000 individuals) sample of British Columbia population; e) to protect confidentiality and privacy by applying the same process as of the pilot project.

If the increase in sample size is approved, many innovative research projects could be initiated. Study of neighbourhood effects on health care, using ecological and individual SES measures simultaneously, could be performed with a larger sample size, allowing an adequate definition of neighbourhood. Double jeopardy questions of the type "Is it worse for an individual's health to be poor in a poor neighbourhood ?" could be answered.

Detailed analyses of socioeconomic differences in treatment of specific conditions could be done. Question of the type "To what extent is Canada's health care system colour blind with respect to SES ?" could be answered.

Other questions of interest like "Small-area SES markers and individual needs should or should not and can or cannot be used for implications for regional health funding formulae ?" and "What are the non-financial barriers to "medically necessary" health care ?" will be of interest.

Whatever the research projects that will be achieve, the analytical benefits of linkage projects well managed with a legal and institutional frame have been demonstrated by the pilot project. When the process is well defined and transparent, the public does not feel threatened by unjustified intrusion into private life.

REFERENCES

Adler, N.E., Boyce, W.T., Chesney, M.A., Cohen, S., Folman, S., Kahn, R.L. and Syme, L. (1994). Socioeconomic Status and Health: The Challenge of the Gradient. *American Psychologist,* **49**, 15-24.

Houle, C., Berthelot, J.-M., David, P., Mustard, C., Roos, L. and Wolfson, M.C. (1996). Le projet d'appariement du Recensement et des fichiers de soins de santé du Manitoba: Composante des ménages privés. Document de recherche No. 91, *Statistique Canada*, Direction des études analytiques.

House, J.S., Kessler, R.C. and Herzog, A.R. (1990). Age, Socioeconomic Status, and Health. *Milbank Quarterly,* **68**, 383-411.

Knighton, T., Houle, C., Berthelot, J.-M. and Mustard, C. (1998). Incidence de l'héritage économique et social sur l'utilisation des soins de santé durant la première année de vie. *Statistique Canada*, No 89-553-xpb au catalogue, 155-166.

Mackenbach, J.P. (1992). Socio-economic Health Differences in the Netherlands: A Review of Recent Empirical Findings. *Social Sciences and Medicine,* **34**, 213-26.

Mustard, C.A., Derksen, S., Berthelot, J.-M., Carrière, K.C., Wolfson, M. and Roos, L.L. (1995). Socioeconomic Gradients in Mortality and use of Health Care Services at Different Stages in the Life Course. *Manitoba Centre for Health Policy and Evaluation*. Rapport 95.04, Winnipeg, Manitoba.

Mustard, C.A., Derksen, S., Berthelot, J.-M., Carrière, K.C., Wolfson, M. and Roos, L.L. (1997). Age-Specific Education and Income Gradients in Morbidity and Mortality in a Canadian Province. *Social Science and Medicine* , **45**, 383-397.

LIST OF PAPERS RELATING TO THE CENSUS-MANITOBA PROJECT

Chun, B., Coyle, D., Berthelot, J.-M. and Mustard, C.A. (1996). Estimating the cost of coronary heart disease in Manitoba. *Proceedings of the American Statistical Association Joint Annual Meeting,* Chicago.

David, P., Berthelot, J.-M. and Mustard, C.A. (1996). Linking Survey and Administrative Data to Study Determinants of Health (Part 1). *The Record Linkage Resource Centre (RLRC) Newsletter, No. 8, Statistics Canada*

David, P., Berthelot, J.-M. and Mustard, C.A. (1996). Linking Survey and Administrative Data to Study Determinants of Health (Part 2). *The Record Linkage Resource Centre (RLRC) Newsletter, No. 9, Statistics Canada*

Houle, C., Berthelot, J.-M., David, P., Mustard, C., Roos, L. and Wolfson, M.C. (1997). Project on matching Census 1986 database and Manitoba health care files. Proceedings of the Record Linkage Workshop - Washington.

Houle, C., Berthelot, J.-M., David, P., Mustard, C., Roos, L. and Wolfson, M.C. (1996). Project on Matching Census 1986 Database and Manitoba Health Care Files : Private Households Component. Research document No. 91, *Statistics Canada*, Analytical Studies Branch.

Knighton, T., Houle, C., Berthelot, J.-M. and Mustard, C. (1998). Health Care Utilization during the First Year of Life : The Impact of Social and Economic Background. *Statistics Canada*, No 89-553-xpb, 145-155.

Knighton, T., Houle, C., Berthelot, J.-M. and Mustard, C. (1997). Socioeconomic status and pap use among Manitoba Women aged 21-65. (Not published but available on request).

Kraut, A., Mustard, C.A. and Walld, R. Health care utilization by self-reported work-related chronic respiratory disability. (Submitted for publication).

Kraut, A., Walld, R. and Mustard, C.A. (1997). Prevalence of Physician Diagnosed Asthma by Occupational Groups in Manitoba, Canada. American Journal of Industrial Medecine, **32(3)**, 275-282.

Mustard, C., Derksen, S., Berthelot, J.-M. and Wolfson, M.C. (1999). Assessing ecologic proxies for household income: a comparison of household and neighbourhood level income measures in the study of population health status. *Health & Place,* **5**, 157-171.

Mustard, C.A., Derksen, S., Berthelot, J.-M., Carrière, K.C., Wolfson, M. and Roos, L.L. (1997). The Use of Insured Health Care Services By Education and Income in a Canadian Province. (Submitted for publication).

Mustard, C.A., Finlayson, M., Derksen, S. and Berthelot, J.-M. (1997). Income and education as predictors of nursing home admission in a universally insured population. (Submitted for publication).

Mustard, C.A., Shanahan, M., Derksen, S., Barber, M.B., Horne, J. and Evans, R.G. (1997). Consumption of insured health care services in relation to household income in a Canadian province. (Proceedings of the IHEA conférence).

Mustard, C.A., Derksen, S., Berthelot, J.-M., Carrière, K.C., Wolfson, M. and Roos, L.L. (1995). Socioeconomic Gradients in Mortality and use of Health Care Services at Different Stages in the Life Course. *Manitoba Centre for Health Policy and Evaluation*. Report 95.04, Winnipeg, Manitoba.

Mustard, C.A., Derksen, S., Berthelot, J.-M., Carrière, K.C., Wolfson, M. and Roos, L.L. (1997). Age-Specific Education and Income Gradients in Morbidity and Mortality in a Canadian Province. *Social Science and Medicine,* **45**, 383-397.

Tomiak, M., Berthelot, J.-M. and Mustard, C. (1997). A profile of health care utilization of the disabled population in Manitoba. (Submitted for publication).

Tomiak, M., Berthelot, J.-M. and Mustard, C. (1997). Factors associated with nursing home entry for elders in Manitoba, Canada. (Accepted for publication in *Journal of Gerontology: Medical Sciences*).

SESSION II

METHODOLOGICAL ISSUES: LONGITUDINAL PURPOSES

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

MODELING LABOUR FORCE CAREERS FOR THE LIFEPATHS SIMULATION MODEL

G. Rowe and X. Lin[1]

ABSTRACT

We estimate the parameters of a stochastic model for labour force careers involving distributions of correlated durations employed, unemployed (with and without job search) and not in thdabour force. If the model is to account for sub-annual labour force pattems as well as advancement towards retirement, then no single data source is adequate to inform it. However, it is possible to build up an approximation from a number of different sources. Data are being assembled for the period leading up to 1991-96:

- Short term labour force dynamics – as reflected in weekly labour force states from the 1988-90 Labour Market Activity Survey.
- Long-term employment dynamics – as reflected in the job tenure distributions of older (i.e., aged 55+) employed respondents to the 1991-96 Labour Force Survey.
- Distributions of labour force state by single year of age -- from both the 1991 and 1996 Censuses for the reference week and during the 1990/1995 calendar years.

Estimation is carried out by embedding initial estimates of the model parameters in the LifePaths simulation model; parameter estimates are then iteratively updated to improve simulated approximations of observed data structures. The LifePaths model provides a suitable platform for estimating effects of educational attainment on labour force dynamics, because it already contains a detailed model of secondary and post-secondary educational careers based on survey, Census and administrative data. Similarly, the fertility and marriage modules of the LifePaths model will, in the future, provide for estimation of effects of family status onlabour force dynamics.

KEY WORDS: microsimulation, education, labour force, correlated durations

1. INTRODUCTION

LifePaths is a longitudinal microsimulation model in which individual lifetimes are simulated. The model initializes a case by randomly generating an individual's sex, province of residence, age at immigration and year of birth. The year of birth can range from 1892 to 2051; but by fixing mortality and immigration assumptions, births occurring in 1892-1991 will reproduce provincial age-sex structures as enumerated in the 1991 Census. The set of variables that describe the demographic, social and economic circumstances of the individual undergoes changes as he/she ages. These changes are dictated by the events in each individual life. The model comprises a set of events that can affect each case at appropriate points in their lifetime. These events include: entering the education system, education progression, graduation, entering the job market, gaining or losing a job, common law union formation, marriage, having children, separation, divorce and death. The model chooses which event will take place by randomly generating a waiting time to the next event for all possible types of event. The event type with the shortest waiting time is selected. Competing waiting times are conditioned upon the individual's current set of characteristics; for example, a married couple is much more likely to have a child in the near future than a 18 year old unattached male. In this way, an individual's unfolding lifetime is not driven merely by the passage of calendar time, but by the time intervals between events. Ultimately, the individual's lifetime concludes when the death event occurs.

[1] Geoff Rowe and Xiaofen Lin, Statistics Canada , Socio-Economic Modeling Group, 24th Floor R.H.Coats Building, Ottawa, Ontario, Canada, K1A 0T6; e-mail – rowegt@statcan.ca

LifePaths is, and will remain, a work in progress, since the goal of the model is to encapsulate as much detail as possible on socio-economic processes in Canada as well as the historical patterns of change in those processes: thus, LifePaths must undergo continuous updating and refinement. Nevertheless, to date, LifePaths has been employed in a broad range of policy analysis and research activities. Examples of LifePaths applications include: analysis of Canada Student Loan policy options (under contract to HRDC and the Government of Ontario), study of returns to education (Appleby, Boothby, Rouleau and Rowe 1999), examination of time use over the life course (Wolfson and Rowe 1996; Wolfson, 1997; Wolfson and Rowe 1998a) and simulating future prospects for the tax-transfer system and pensions (Wolfson, Rowe, Gribble and Lin 1998; Wolfson and Rowe 1998b). In addition, the task of assembling data for LifePaths has required new research into, for example, educational careers (Chen and Oderkirk 1997; Rowe and Chen 1998; Plager and Chen.1999).

2. OBJECTIVES

In addition to accounting for life course events, LifePaths also imputes various individual characteristics: in particular, hourly equivalent wage rates (conditional on sex, education level, field of study and duration since graduation) and usual hours of work (conditional on education level, sex, age and hours last year). These characteristics together with employment status determine earnings. In the long term, one of our objectives is to put together a model that will simulate realistic family income distributions and trajectories over recent decades. A necessary component of such a model is a representation of lifetime labour force activity which represents differences in activity patterns/dynamics by educational level and by marital status/family composition as well as incorporating sub-annual/seasonal effects and which tracks business cycles in the recent past.

The work presented here is the first stage in the construction of a comprehensive labour force model. Our starting point involves modeling links between educational attainment and subsequent labour force careers because these are largely separable stages in the life course. Subsequent refinements to the model will take family level factors into account – bearing in mind interactions between family and labour force careers.

3. LABOUR MARKET ACTIVITY SURVEY

The Labour Market Activity Survey (LMAS, 1992) was a longitudinal survey designed to collect information on the labour market activity patterns of the Canadian population over multiple years. We draw our data from the 1988-90 longitudinal microdata files. These data were collected from retrospective interviews conducted in January-February of 1989, 1990 and 1991 respectively. The reference period for the data was January 1988 through December 1990 – the period just prior to the 1991 Census. We made use of about 55,000 responses representing all individuals who were either not full time students as of January-June 1988 or who ceased being full time students in that period.

We base our analysis of short-term labour force dynamics on the LMAS weekly composite labour force status vector (i.e., variables CLFSV88, CLFSV89 and CLFSV90). The coding of these variables distinguishes among (1) employed, (2) unemployed and engaged in job search (or on temporary lay-off), (3) unemployed and not engaged in job search and (4) not in the labour force. We make use of this detailed categorization anticipating that it will lead to a more appropriate model, despite the fact that our only immediate need is for simulation of Work and Non-Work states.

Figure 1 displays scatter plots -- with plotting symbols' area proportional to survey weight -- of completed Work and Non-Work spells for males having exactly two such completed spells in the 1988-90 interval. In order that two completed spells be identified within the three-year observation period, they must be bracketed between censored spells at the beginning of 1988 and the end of 1990.

Figure 1

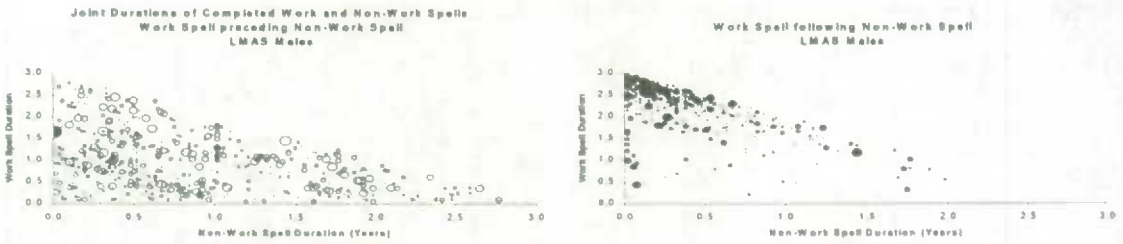


Figure 1 shows a marked difference in distribution depending on whether the completed Work spell preceded or followed the Non-Work spell: in the former case, respondents – first observed in a censored Non-Work spell -- are widely scattered; while in the latter case, respondents -- first observed in a censored Work spell -- are concentrated in the region representing Non-Work spells of less than six months and Work spells of two or more years. The concentration of relatively short Non-Work and relatively long Work spells corresponds to cases with known labour force attachment at the beginning of the observation period, while the remaining cases have unknown labour force attachment at the outset.

Tentative conclusions drawn from these observations set the stage for subsequent analysis of the LMAS data: (1) there is considerable heterogeneity in the group with unknown labour force attachment and (2) there may be 'correlation' in labour force attachment within individuals. We expect to account for the heterogeneity by distinguishing among the states unemployed with and without job search and not in the labour force. However, as a consequence we need to consider 12 event types – transitions between pairwise combinations of the four states – and to deal with the added complexity of a competing risk model.

4. DATA ANALYSIS AND MODEL STRUCTURE

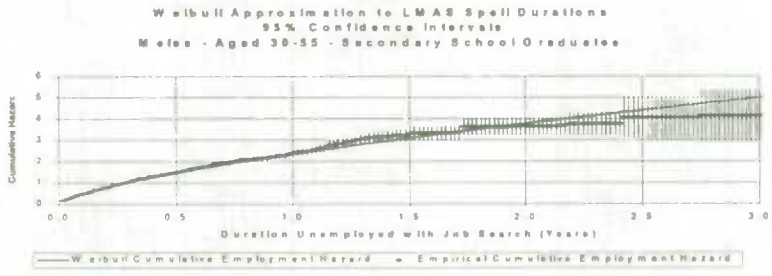
4.1 Weibull Approximations of LMAS Spell Duration Distributions

Our analysis of the LMAS spell duration data began with a partition of the data into categories determined by the intersection of 1988 fulltime student status, age group (i.e., <30, 30-55, 56+ in 1988), education level (i.e., Below Secondary [<SSG], Secondary Graduate [SSG], Some Post-Secondary [SPS], Post-Secondary Non-University [PSE-NonU], Post-Secondary University [PSE-U] in 1990) and sex. Within each of the resulting categories, we attempt to find a parsimonious approximation to the empirical hazard function. In order to obtain meaningful estimates from the LMAS data, we had to address the problem of left censored, or both left and right censored spells (i.e., spells that had begun before the observation period and so had unknown duration). The problem was 'resolved' in two steps: (1) a duration was imputed to the beginning of each left censored spell by randomly drawing from the right censored spells of respondents in the same category and three years younger and (2) the focus of analysis was narrowed to a description of the dynamics of spells initiated within the interval 1987-88 (i.e., 1987-88 labour force state entry cohorts). The latter was accomplished by assigning weights to each spell determined by the calendar date of the first week of the spell. These weights had a value of 1.0 for spells with observed or imputed initial week in the interval 1987-88. The weights for initial weeks outside of the interval were reduced progressively to 0.0 – the value assigned to the last week of December 1990.

Calculation of empirical hazards indicated that a Weibull distribution would provide a reasonable approximation to spell durations. The Weibull distribution is characterized by a log-linear relationship between duration (**t**) and cumulative hazards (**H(t)**), which is expressed in terms of a Scale parameter α and a Shape parameter β: $\mathbf{H(t) = (t/\alpha)^{\beta}}$. Figure 2 illustrates the fit by comparison of empirical and Weibull cumulative hazards for Unemployed Male Secondary School Graduates Aged 30-55 who were actively engaged in a job search. The hazards, in this case, represent the chance of finding a job and imply a survival probability of 0.227 at six months duration (i.e., the probability of continuing in the unemployed

state: **exp(-H(t))**). Corresponding results for Job Loss indicate median job durations in excess of 3 years. Note that, in common with nearly all of hazard functions we estimated, hazards for short durations appear to be larger than hazards for long durations – implying Weibull Shape parameters $\beta < 1.0$.

Figure 2



4.2 Correlated Weibull Spell Durations

The Weibull models outlined above are independent. Hougaard (1986a, 1986b, 1987) describes a generalization of the Weibull that can account for correlated spells. In common with Clayton and Cuzick (1985), Hougaard introduces correlation by postulating an unobserved random variable **Z** having a common influence on all durations, rather than by explicitly introducing lagged regression parameters. **Z** is intended to represent unmeasured personal and/or labour market characteristics that remain fixed over time. In contrast to Clayton and Cuzick, Hougaard exploits some unique advantages of specifying **Z** to be drawn from a Positive Stable distribution with parameter θ. If an individual's spell durations given **Z** were Weibull with Scale α and Shape β, then population spell durations are also be Weibull:

$$E_z(e^{-H(t)Z}) = e^{-H(t)^{\theta}} = e^{-\left(\frac{t}{\alpha}\right)^{\beta\theta}}$$

Thus, given appropriate values of θ, the Weibull models already estimated might be transformed to a parsimonious correlated duration model. Hougaard demonstrates that θ can be estimated by $\sqrt{(1 - r)}$, where **r** is the product moment correlation between log spell durations. Note that the relation between θ and **r** implies that only positive correlations can be accounted for in this model (i.e., $0 < \theta < 1$). Estimates of median (bias corrected) correlations suggest values of θ of about 0.87 and 0.86 for males and females respectively. The results are not precise enough to allow an evaluation of differences among education levels.

4.3 Model Implementation

Given the estimates obtained from the LMAS, implementation of a labour force model in LifePaths is straightforward using randomly generated Weibull waiting times. A random waiting time **t** from a Weibull distribution with Scale α and Shape β may be generated directly given a Uniform (0,1) pseudorandom number **U**:

$$t = \exp\left(\frac{\ln(-\ln(U))}{\beta} + \ln(\alpha)\right)$$

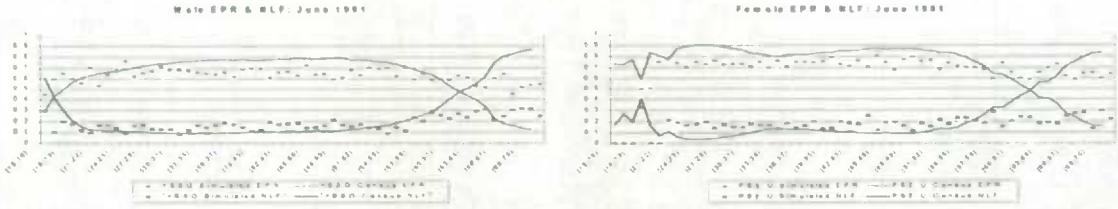
However, if correlated waiting times are to be generated; we must first generate a Positive Stable random number **Z** (Chambers, Mallows and Stuck, 1976) which will remain fixed, and then make an adjustment to each subsequent random Weibull waiting time. In that case, we make use of the conditional shape parameter $\beta' = \beta/\theta$:

$$t = \exp\left(\frac{\ln(-\ln(U))}{\beta} + \ln(\alpha)\right) \Big/ Z^{\frac{1}{\beta\theta}}$$

5. INITIAL SIMULATION RESULTS AND VALIDATION

Figure 3 compares simulated employment to population ratios (EPR) and proportions not in the labour force (NLF) to 1991 Census data by age for selected sex and educational attainment categories. The simulation results were obtained by tabulating the simulated person years lived and simulated person years in the 'Employed' or 'Not in the Labour Force' states for each age interval and computing the proportion. In both cases, independent waiting times were used (i.e., no adjustment for θ). The left panel shows results for Males with less than Secondary School (<SSG), while the right panel shows results for Female University Graduates (PSE-U). It is evident in both cases, that level biases exist and that an auxiliary special case model is required to generate retirement times.

Figure 3



In order to allow retirement, an additional step was added to the simulation module. Age specific 'permanent retirement' hazards are not known or clearly estimable from any available data. Part of the difficulty is that although the event clearly occurs, we can not know that an individual has permanently retired until death. We propose to estimate the unknown parameters by calibrating the LifePaths simulations to Census data. Age specific 'permanent retirement' hazards are expressed as proportional to (approximate) age specific Long-Term Disability hazards estimated from the 1991 Census. The estimates of relative risk of 'permanent retirement' are specific to educational attainment categories.

6. INITIAL CALIBRATION RESULTS

Calibration is accomplished by first evaluating the lack of fit calculated by comparing auxiliary data with the corresponding simulated values. To this end, we read auxiliary data into the LifePaths program as parameters, then produced an output table containing only the measure of lack of fit. The auxiliary data that we are currently working with includes:

- 1991 and 1996 Census students (full time attenders) and non-students by sex, single year of age, and educational attainment in each labour force state, employed, unemployed and engaged in job search, unemployed and not engaged in job search and not in the labour force.
- 1991 and 1996 enumerated non-students by sex, single year of age and educational attainment by weeks worked in 1990 and 1995, where weeks are categorized as 0, 1-4, 5-8, ..., 40+.
- 1991-1996 LFS annual average job tenure frequencies for employed persons aged 55+ where years of job tenure are as <2.5, 2.5-5, 5-10, ..., 45+.
- LMAS summary data comprising weighted events and population at risk by spell type, spell duration, sex, educational attainment and age group.

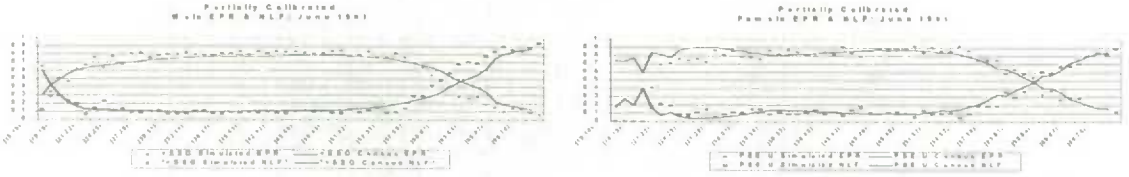
Working with these auxiliary data, we use a composite measure of lack of fit:

$$-2*\text{LMAS Log Likelihood}$$
$$+ 6*\Sigma_{AGE}\,[\text{Simulated count}*\Sigma_{STRATUM}\,(\text{Census proportions} - \text{Simulated proportions})^2]$$
$$+ 6*\Sigma_{YEAR}\,[\text{Simulated count}*\Sigma_{STRATUM}\,(\text{LFS proportions} - \text{Simulated proportions})^2]$$

Calibration continues using a special purpose computer program (Calibrator) that implements general (no derivative) optimization strategies, including the Nelder-Mead Simplex algorithm and the Kiefer-Wolfowitz random downhill search algorithm, and attempts to scale parameters appropriately. Calibrator determines adjustments to LifePaths' input parameters that are appropriate given the algorithm chosen, launches new simulation runs and collects the results of simulation runs to determine further updates to input parameters leading to further runs.

Figure 4 presents initial calibration results corresponding to results in Figure 3. These results were obtained by estimating relative risks and probabilities of retirement by education level and estimating θ and an adjustment to α for each education level. Improvements due to the retirement model are evident. The reductions in level bias – for example, the EPR for Males <SSG aged 40 goes from a peak under-estimate of about 15% points to an over-estimate of about 2% points – are in part due to the use of correlated waiting times and calibrated estimates of θ. Plugging in the initial estimates of θ – section 4.2, giving each education category the same initial value – produced an immediate improvement. Calibration gave further improvement with estimates of θ being reduced from about 0.86 to values ranging between 0.67 and 0.78. These estimates imply that the bias-corrected correlations were themselves negatively biased.

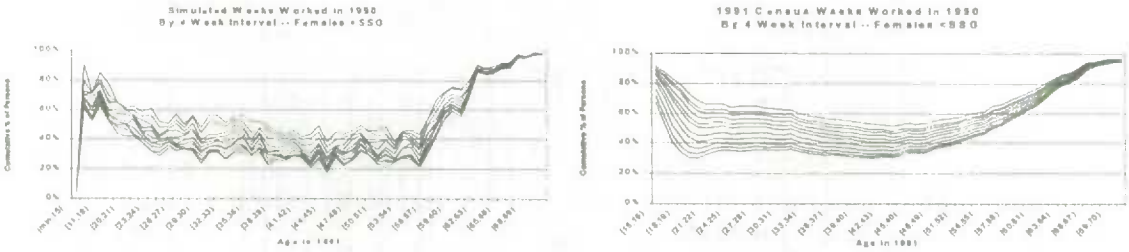
Figure 4



7. NEXT STEPS

The results of partial calibration do not look as good when viewed in terms of either simulated weeks worked in a year or job tenure. Figure 5 illustrates the comparison of simulated and Census weeks worked in 1990.

Figure 5



It is clear that we simulate too few persons with 0.0 weeks worked in a year (except at young ages) and too many persons who work full-year. The net result is that the dispersion of simulated weeks worked is too small. Similarly, the results for simulation of 'job' tenure, Figure 6, indicate that far too many employed 55+ year olds have short tenures (i.e., less than 2.5 years). Thus, although there are too many persons working full year, there are also too few continuously employed for more than two years.

Figure 6

At this stage, further calibration must address the problems posed by large numbers of parameters (i.e., about 750 LMAS based Weibull parameters) and by the fact that the 'error' being minimized is subject to Monte Carlo variability. Dealing with these two problems requires efficient algorithms that do not involve partial derivatives – which would be expensive to approximate numerically and be prone to Monte Carlo error – and appropriate tests of convergence. Chan and Ledolter (1995) discuss the need for a special

stopping criterion for Monte Carlo EM estimation (i.e., an estimation approach similar to ours). Booth and Hobert (1999) discuss both stopping criteria and high dimensionality in Monte Carlo EM estimation.

More parsimonious models would be easier to fit. Consequently, we will explore ways to reduce the number of parameters in the model by, for example, estimating common Weibull shape parameters for males and females (section 4.1). However, at the same time, we need to consider ways of extending the model to involve estimating: 1) Minimum age at retirement (currently 55 for all, but may be as low as 45 for some age-sex groups), 2) Marriage/Family effects, 3) Business Cycles/Seasonality/Secular Trends and 4) Improvements to representation of the transition from Education to Work.

Finally, given the complexity required of models of Labour Force careers, it is clearly useful to be able to embed them in a microsimulation model. This gives us the means to derive implications from the model that might not be possible otherwise (e.g., job tenure results) and also gives us the means to carry the estimation process beyond a single data source. The value of the approach is demonstrated by the deficiencies in our direct Weibull estimates from the LMAS. The biases revealed in Figure 6 may be due to intrinsic limitations of short period longitudinal data, our efforts to deal with left censoring and/or deficiencies in the regressions we chose to fit. Regardless, we would likely not have been aware of them had our assessment of our model been limited to an evaluation of measures of agreement with LMAS data.

REFERENCES

Appleby, J., D. Boothby, M. Rouleau and G. Rowe. (1999) "Level and Distribution of Individual Returns to Post-Secondary Education: Simulation Results from the LifePaths Model" to be presented at the 1999 meetings of the Canadian Economics Association.

Booth, James G. and James P. Hobert (1999). "Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm", Journal of the Royal Statistical Society B, **61(1)**, pp.265-285.

Chambers, J.M., C.L. Mallows and B.W. Stuck (1976). "A Method for Simulating Stable Random Variables", Journal of the American Statistical Association, **71(354)**, pp.340-344.

Chan, K.S. and Johannes Ledolter (1995). "Monte Carlo EM Estimation for Time Series Models Involving Counts", JASA, **90(429)**, pp.242-252.

Chen, E.J. and Oderkirk, J. (1997). "Varied Pathways: The Undergraduate Experience in Ontario", Feature article. Education Quarterly Review, Statistics Canada, 81-003-XPB, **4**, No. 3, 1998.

Clayton, David and Jack Cuzick (1985). "Multivariate Generalizations of the Proportional Hazards Model", Journal of the Royal Statistical Society.A, **148(2)**, pp.82-117.

Hougaard, Philip (1986a). "Survival models for heterogeneous populations derived from stable distributions", Biometrika, **73(2)**, pp.387-96.

Hougaard, Philip (1986b). "A class of multivariate failure time distributions", Biometrika, **73(3)**, pp.671-78.

Hougaard, Philip (1987). "Modelling Multivariate Survival", Scandavian Journal of Statistics, **14**, pp.291-304.

Labour Market Activity Survey Profiles (1992), "Canada's Women: A Profile of Their 1988 Labour Market Experience", Statistics Canada, Catalogue No. 71-205, Ottawa.

Rowe, G. and E.J. Chen (1998). "An Increment-Decrement Model of Secondary School Progression for Canadian Provinces", Symposium on Longitudinal Analysis for Complex Surveys, Statistics Canada, Ottawa.

Plager, L. and E.J. Chen (1999). "Student Debt in the 1990s: An Analysis of Canada Student Loans Data". Education Quarterly Review, Statistics Canada, 81-003-XPB, **5**, No. 4, 1999.

Wolfson, M.C. (1997). "Sketching LifePaths: A New Framework for Socio-Economic Statistics" in R. Conte, R. Hegselmann and P. Terna (Eds.), *Simulating Social Phenomena, Lecture Notes in Economics and Mathematical Systems 456*, Springer.

Wolfson, M.C. and G. Rowe (1996). "Perspectives on Working Time Over the Life Cycle", Canadian Employment Research Forum Conference on "Changes to Working Time", Ottawa.

Wolfson, M.C. and G. Rowe (1998a). "LifePaths – Toward an Integrated Microanalytic Framework for Socio-Economic Statistics", 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Wolfson, M.C. and G. Rowe (1998b). "Public Pension Reforms – Analyses Based on the LifePaths Generational Accounting Framework", 26th General Conference of the International Association for Research in Income and Wealth, Cambridge, U.K.

Wolfson, M.C., G. Rowe, S. Gribble and X. Lin (1998). "Historical Generational Accounting with Heterogeneous Populations", in M. Corak (Ed), Government Finances and Generational Equity, Statistics Canada Cat. No. 68-513-XPB, Ottawa.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

THE U.S. MANUFACTURING PLANT OWNERSHIP CHANGE DATABASE: RESEARCH POSSIBILITIES

Sang Van Nguyen[1]

ABSTRACT

The U.S. Manufacturing Plant Ownership Change Database (OCD) was constructed using plant-level data taken from the Census Bureau's Longitudinal Research Database (LRD). It contains data on all manufacturing plants that have experienced ownership change at least once during the period 1963-92. This paper reports the status of the OCD and discuss its research possibilities. For an empirical demonstration, data taken from the database are used to study the effects of ownership changes on plant closure.

Keywords: Manufacturing; Plant-level Data; Ownership Change; Data Matching.

1. INTRODUCTION

Recently, the Center for Economic Studies (CES), U. S. Census Bureau, has constructed the U.S. Manufacturing Plant Ownership Change Database (OCD), using plant-level data taken from the Census Bureau's Longitudinal Research Database (LRD). The OCD contains nine variables (listed below) for all manufacturing establishments that changed owners at least once during the period 1963-92. Because each plant in the OCD and LRD is assigned a unique permanent plant number (PPN), researchers can use this PPN to merge the OCD with the LRD. In this way, data on many other variables available in the LRD --describing economic activities of the plants that changed owners as well as those that did not experience any ownership change (the control group)— may be obtained for research. The data file will be updated when new data become available.

This paper reports the status the OCD and discusses research possibilities using the data. For an empirical demonstration, plant-level data taken from the OCD and LRD are used to study the effects of ownership changes on plant closing. The empirical results confirm previous finding that acquired plants are less likely to be closed than other plants.

2. THE DATA AND METHOD FOR IDENTIFYING OWNERSHIP CHANGE

2.1 Contents of the OCD

The OCD contains data for all establishments that changed their firm IDs at least once during the period 1963-92. There are 86,700 such establishments in the database. Because one can use PPNs and IDs to merge the OCD with LRD to obtain data on many variables available in the LRD, the OCD is designed to contains only

---

1 Sang V. Nguyen, Center for Economic Studies, U.S. Bureau of the Census, 4700 Silver Hill Road, Stop 6300, Washington D.C. 20233. Any opinions, findings or conclusions expressed here are those of the author and do not necessarily reflect the views of the U.S. Bureau of the Census.

the following nine variables: (1) BUYID: ID number of the buying (acquiring) firm; (2) CC: Coverage code; (3) ID: Firm identification number; (4) IND: 4-digit SIC industry code; (5) PPN: Permanent plant number; (6) SELLID: ID number of the selling (acquired) firm; (7) STATUS: This variable indicates whether an ID change is an ownership change; (8) TE: Total employment (number of employees); (9) TVS: Total value of shipment (See Nguyen ,1999 for a detailed description of these variables).

2.2 Data sources

The data used to construct the OCD are taken from the LRD. At present, the LRD is an unbalanced panel that links seven CMs for the years 1963, 1967, 1972, 1977, 1982, 1987 and 1992 and 16 ASMs for the years between census years, starting in 1973. The LRD contains data on output, inputs, and production costs of individual U.S. manufacturing establishments. The output data include total value of shipments, value added and related variables such as inventories of work-in-process and finished goods. Data on inputs include information on capital, labor, energy, materials, and selected purchased services. The employment data include total employment, production workers, non-production workers, production worked hours as well as salaries and wages. (For a more detailed description of the CM, ASM and LRD, see McGuckin and Pascoe, 1988).

An important feature of the LRD is its plant identification information, including firm affiliation, location, products, industries, and various status codes which identify, among other things, birth, death, and ownership changes. Two important identification numbers used in developing both the longitudinal plant linkages and ownership linkages among plants are PPN and ID numbers.

2.3 Assessing the Data

Plant and company level data in the LRD and OCD are collected by the Census Bureau under the authority of Title 13 of the United States Code. To protect confidentiality of the data, Title 13 and the rules and regulations of the Census Bureau prohibit the release of micro data that could be used to identify or closely approximate the data on a plant or a company. Thus, only sworn Census employees have direct access to these microdata.

To gain access to microdata at the CES or its Research Data Centers (RDC), non-Census researchers would have to become *special sworn employees* (SSE) It is important to emphasize that while SSEs can directly access CES' microdata, perform analysis on the data, and include research results in papers and reports, they cannot remove any portion of these data files from the premises of CES or its RDCs. In addition, CES reviews all materials produced by SSEs for evidence of disclosure of sensitive information. In general, no information that could be used to identify, or to approximate the level of activity of a plant or a company can be included in papers and reports.[2]

2.4 Method of Identifying Ownership Change

The following three-step procedure is used to target ownership changes: (1) Using PPN and firm ID numbers to identify plants that changed firm IDs between two census years; (2) Within this set of plants, using CC codes to identify directly reasons for ID changes (e.g., ownership change, a "multi-unit" plant becomes a single-unit firm, reorganization, etc.); and (3) From the remaining plants, identifying further ownership

---

2 Most SSEs working with CES or its RDCs are employees of Federal agencies engaging in statistical work and related activities, or are individuals affiliated with universities, research, or research-related organizations. These individuals can provide expert advice or assistance on CES projects. For information on access to microdata at the CES, contact Mary L. Streitwieser, Center for Economic Studies, U.S. Bureau of the Census, Washington D.C., 20233, (301) 457-1837, e-mail: mstreitw@ces.census.gov.

changes indirectly by matching the firm IDs of acquired firms and acquiring firms. (All programs for identifying and matching data on ownership changes were written in SAS, version 6.12.).

Identifying true ownership changes (mergers and acquisitions) -- step (2) -- requires additional information. The main additional information is in the census CC codes assigned to establishments in the CM or ASM. The CC codes are two-digit numbers that indicate the status of the establishment in the survey. For example, a CC code equals 18 indicates that the establishment was sold to another company.[3] Ideally, all new firm ID and CC codes would be recorded during the actual year that the establishment changes its status so that it would be easy to identify ownership changes. In practice, except for a set of large ASM establishments, neither changes in ID nor proper CC codes are systematically recorded during the years of status change. In many cases, particularly for small establishments, a change in the firm ID of a plant appears one or more years before a proper CC code is assigned. The reverse is also possible: the CC code can indicate an ownership change before the ID changes.

To address these issues, for the years when ASM data were available, I examined CC codes in the years before and after the ID change. However, this procedure leaves two unresolved problems. First, in non-census years, not all plants are in the survey sample and, in particular, when the ASM panel changes (in years ending in 4 and 9), the set of non-certainty cases (the smaller plants) turns over completely. Secondly, for a number of establishments, proper CC codes are not assigned at all. Nevertheless, using CC codes allows identification of a large portion of the establishments that have ID changes due to true ownership changes.

Finally, in step 3 it is necessary to bring together initial and ending firm IDs for all plants that were owned by the firm in question. For example, assume that the LRD shows that plant A belonged to firm X in 1977 and to firm Y in 1982, but the 1982 CC code for plant A does not show this as an ownership change. But assume, also, that firm Y acquired at least one other plant from firm X between 1977 and 1982, as confirmed by the CC codes. In this case, it seems likely that firm Y bought plant A as well, and we code plant A accordingly. (For a more detailed discussion on how the OCD was constructed, see Nguyen ,1999).

3. RESEARCH POSSIBILITIES: WHAT CAN WE LEARN FROM THE LRD-OCD?[4]

3.1 Causes and Consequences of Mergers and Acquisitions

Economists have long been interested in determining the causes and consequences of merger and acquisition activities. However, the empirical findings based on aggregate data are controversial (see Muller, 1993). Thus research based upon new microdata is imperative to arrive at more definitive results on this important topic. In what follows, I review some typical studies using the OCD/LRD to examine issues related to ownership changes.

Lichtenberg and Seigel (1992a) are among the researchers who first used LRD data to study the causes and consequences of ownership changes. Their empirical work is based on a matching model that is closely

---

3 For a complete list of CC codes, see the LRD documentation (U.S. Bureau of the Census, Center for Economic Studies, 1992, a revised version of this documentation is forthcoming).

4 This review is by no means exhaustive. There are other studies that used the data, but are not discussed here due to the space limit. Also, currently there are a number of researchers using the database in their research.

related to the theory of job turnover developed by Jovanovic (1979), using an LRD panel of 20,493 plants owned by 5,700 firms for the period 1972-81. For this sample, they identified nearly 21 percent of the 20,493 plants that experienced at least one ownership change. With these data, they found that plant TFP is negatively related to ownership change and that ownership change is positively related to TFP growth. They, therefore, concluded that ownership change is primarily motivated by lapses in efficiency (bad matches).

McGuckin and Nguyen (1995) argue that Lichtenberg and Seigel's version of the matching model is "too restrictive" because it does not recognize the importance of the demand side of the market: purchase of a plant or firm will be undertaken if the buyer (acquiring firm) places a higher value on the plant than does the seller (selling firm). However, there is no reason to believe that acquiring firms purchase only poorly performing plants. Indeed, there are many possible motives for mergers and acquisitions: monopoly power, synergies and tax incentives are potential motives of mergers that do not require purchase of low productivity plants. They also point out that Lichtenberg and Seigel's study is likely to be subject to sample selection bias because their sample includes mostly large surviving plants. To avoid this potential bias, McGuckin and Nguyen used an unbalanced panel taken from the OCD and LRD, covering the entire population of the U.S. food manufacturing sector (SIC 20). This panel consists of 28,294 plants owned by all food producing firms operated during 1977-87. With these data they found that plants with high productivity were the most likely to experience ownership change. However, for a subset of large surviving plants, they found — consistent with Lichtenberg and Seigel — that initial productivity is inversely related with ownership change. Finally, they found that plant productivity growth is positively related to ownership change. They concluded that "gains from synergies between the buying and selling firms are the most important motive for ownership change during 1977-82 period" (p. 259). Managerial discipline and matching motives appear to be applicable only to mergers and acquisitions that involved large poorly performing plants.

3.2 The Impact of Ownership Change on Labor

Despite strong opposition from labor unions and widespread, often negative, press reports on ownership changes through mergers and acquisitions, there are few studies of the impact of ownership change on labor. The reason for this is the lack of appropriate data for empirical studies. Since the LRD became available to researchers, economists have conducted several studies on the impact of ownership change on labor. Lichtenberg and Seigel (1992b) used plant level data taken from the LRD to examine the effects of ownership change on wages and employment on both central offices and production establishments. They found that ownership change is associated with reduction in both wages and employment at central offices, but it has little effect at production establishments. These results suggest that managers and white collar workers suffer the most after ownership change; but overall, the effects of ownership change on labor, particularly on production workers, appear to be small.

McGuckin, Nguyen and Reznek (1998) examined the effects of ownership changes on employment, wages and productivity using an unbalanced panel of more than 28,000 establishments taken from the LRD and OCD. The study covers the entire U.S. food producing industry (SIC 20). They found that (1) five to nine years after acquisition, the growth rates of wages, employment and labor productivity for typical acquired plants (as well as originally owned plants of acquiring firms) are higher than the typical plants of non-acquiring firms; (2) to a lesser extent, the typical worker in either category of plants owned by the acquiring firms also enjoyed higher growth rates of wages, employment and productivity after acquisitions; and (3) plants that changed owners show a greater likelihood of survival than those that did not. McGuckin and Nguyen (1999) extended this work to include data on the entire U.S. manufacturing sector for the same period and found similar results.

In brief, the OCD/LRD data have been proven to be useful in economic research. It is important to

emphasize that uses of the OCD are not limited to the above issues. It is a potentially valuable database that can be used to study many aspects of the economy where mergers and acquisitions could have a real impact.[5]

4. AN EMPIRICAL DEMONSTRATION: OWNERSHIP CHANGE AND PLANT CLOSING

For demonstration, I use OCD/ LRD data to analyze the effect of ownership changes on plant closure. To do so, following McGuckin and Nguyen (1995), I run probit regressions in which plant closing is the dependent variable. Explanatory variables include a dummy variable representing plants having ownership change, initial relative labor productivity, initial employment, a variable identifying whether the plant was originally owned by an acquiring firm (the omitted category is plants that were owned by non-acquiring firms). Other control variables include type of plant (plants owned by a single-unit firm vs a multi-unit firm), plant ages, regions, and industry (4-digit). Finally, non-linear effects of initial productivity and employment size on plant closure are also included. (See McGuckin and Nguyen (1995) for a detailed description of the model).

To better assess the impact of plant type on the probability of plant closure, we used the parameter estimates of the probit models to estimate the probabilities of plant closure for plants that experienced ownership change, plants originally owned by acquirers, and plants owned by non-acquirers in 1977. Table 1 shows the estimated probabilities of plant closing. Column (1) reports the results based on a simple probit regression, while column (2)
shows the probabilities based on a probit model in which ownership changes are assumed to be endogenous. To
show the effect of plant size, I use different employment sizes (lnE_{77}) in the evaluation of the probabilities. These

Table 1: Probabilities of plant closure

Types of Plants	Model I (1)	Model II (2)
Acquired Plants		
Case 1[a]	.3009	.3279
Case 2[b]	.3676	.4786
Case 3[c]	.4122	.5814
Acquirers' Own Plants		
Case 1	.5185	.5424
Case 2	.5909	.5990
Case 3	.6354	.6380
Non-Acquirers' Plants		
Case 1	.4088	.4291
Case 2	.4812	.4786
Case 3	.5275	.5146

[a] Case 1: The probabilities are estimated by setting lnE77 = 4.60. (E77 = 99)

5 As an example, these data can be used to study the impact of ownership changes on job creation and job destruction and their variation within and between firms.

[b] Case 2: The probabilities are estimated by setting lnE77 = 3.90. (E77 = 49)

[c] Case 3: The probabilities are estimated by setting lnE77 = 3.50. (E77 = 33)

Note: The simple means of lnE77 (log of employment in 1977) for acquired plants, acquirers' plants, and non-acquirers' plants are 4.60, 4.56, and 2.13, respectively.

sizes are: $lnE_{77} = 4.6$ (case 1), $lnE_{77} = 3.9$ (case 2), and $lnE_{77} = 3.5$ (case 3). The table shows that the probability of closing of acquired plants is smaller than that of non-acquired plants. As for Model II, in case 1 where lnE_{77} is set equal to the average size of acquired plants ($lnE_{77} = 4.60$), the probability of closing for the typical acquired plant is .3279 which is more than 10 percentage points lower than that for a non-acquirer's plant of a similar size with an estimated probability of closing of .4291. When size is reduced to 3.9, both acquired plants and non-acquirers' plants had the same probability of closing at .4786. Finally, when setting $lnE_{77} = 3.5$ the probability of closing for acquired plants becomes larger than that for non-acquirers' plants. Finally, the estimated probabilities of closing for acquirers' own plants are greater than those for the other two types of plants. However, as the size gets larger, the difference in these probabilities become smaller. This result indicates that acquirers are more willing to close their small plants than non-acquirers.

In summary, the results strongly suggest that plants changing owners had a much greater chance to survive than plants not changing owners. Acquirers' own plants, particularly small ones, are more likely to be closed than those originally owned by non-acquirers. This latter finding contradicts the results obtained for the food industry where acquirers' own plants were less likely to be closed than those owned by non-acquirers.

5. CONCLUDING REMARK

The OCD data are unique and valuable. A major strength of this database is that it contains plant level data for the entire U.S. manufacturing sector over a long period (1963-92). These data allow researchers to take a close look inside the firm and to observe the contribution of each individual component of the firm before and after ownership change occurs. The OCD has been proven to be valuable in empirical studies such as research on causes and consequences of owner changes that involve "control" of the firm. In particular, the data have provided some convincing evidence on the effects of mergers and acquisitions on productivity, employment and wages.

The OCD also has its shortcomings. First, it is biased toward large, surviving establishments. In particular, during the periods 1963-67 and 1967-72, when ASM data were not available, among the plants that changed owners, only those survived to the next census year could be identified. With ASM data available beginning in 1974, additional plants that changed owners in the years between two censuses were identified in the later periods. However, even in these periods a large number of smaller non-ASM establishments that had ownership changes in the years between the two censuses and were closed before the second census year could not be identified. Similarly, a new non-ASM plant that was opened after the first census year, and changed owner before the next census year, were identified as a "new" plant of the acquiring firm. Thus, the OCD under-counts ownership changes and closed acquired plants, and overstates the number of acquiring firms' new plants. Second, the OCD covers only manufacturing. For companies that operate in both manufacturing and non-manufacturing, the database contains information only on the manufacturing portions of these companies. Finally, the OCD does not contain information about certain types of ownership change, such as tender offers or hostile takeovers. These types of ownership changes are believed to perform differently after mergers.

REFERENCES

Jovanovic, B. (1979), "Job Matching and the Theory of Turnover," *Journal of Political economy*, **87**, pp. 972-90.

Lichtenberg, F.R. and Siegel, D. (1992a), "Productivity and Changes in Ownership of Manufacturing Plants," in *Corporate Takeovers and Productivity,* F. Lichtenberg (ed.), Cambridge: The MIT Press, pp. 25-43.

Lichtenberg, F.R. and Siegel, D. (1992b), "Takeovers and Corporate Overhead," in *Corporate Takeovers and Productivity*, F. Lichtenberg (ed.), Cambridge: The MIT Press, pp. 45-67.

McGuckin, R.H. and Nguyen, S.V. (1995), "On Productivity and Plant Ownership Change: New Evidence from the LED," *The RAND Journal of Economics*, **26**, Number 2, pp. 257-76.

McGuckin, R.H. and Nguyen, S.V. (1999), "The Impact of Ownership Changes: A view from Labor Markets" *International Journal of Industrial Organization* (forthcoming).

McGuckin , R.H, Nguyen, S.V., and Reznek, A.P. (1998), " On Measuring the Impact of Ownership Change on Labor: Evidence from U.S. Food Manufacturing Data," in *Labor Statistics Measurement Issue,* John Haltiwanger, Marilyn Manser and Robert Topel (eds.), Chicago: NBER, pp. 207-46.

McGuckin, R.H. and Pascoe, G. (1988), "The Longitudinal Research Database: Status and Research Possibilities," *Survey of Current Business*, **68**, N°. 11, pp. 30-37.

Mueller, D. C. (1993), "Mergers: Theory and Evidence,' in Gallium Mussati, ed., *Mergers, Markets, and Public Policy*, Dordrecht; Kluwer.

Nguyen, S.V. (1999), "The manufacturing Plant Ownership Change Database: Its Construction and Usefulness," *Journal of Economic and Social Measurement*, **23**, pp. 1-24.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

CREATION OF AN OCCUPATIONAL SURVEILLANCE SYSTEM IN CANADA
COMBINING DATA FOR A UNIQUE CANADIAN STUDY

Maureen Carpenter[1], Kristan J. Aronson[2], Martha E. Fair[3], Geoffrey R. Howe[4]

ABSTRACT

Objective: To create an occupational surveillance system by collecting, linking, evaluating and disseminating data relating to occupation and mortality with the ultimate aim of reducing or preventing excess risk among workers and the general population.

Methods: Data were combined: (i) to establish the cohort; (ii) to create longitudinal individual work histories (the time in each occupation by occupation codes from 1965-71); (iii) to establish socio-economic status (white-collar, blue-collar occupations); (iv) to establish the cause of death (by linkage to the Canadian Mortality Data Base); and (v) to create cause of death code groupings across time from 1965-91 (by converting historic death codes into seventy cause of death classifications). Analyses were conducted of these combined data and the extensive results are being prepared for dissemination.

Results: A cohort was created of approximately 700,000 men and women employed between 1965 and 1971, constituting approximately ten percent of the employed Canadian labour force covered by unemployment insurance at that time. There were almost 116,000 deaths among males and over 26,800 deaths among females up to 1991. Over 27,000 comparisons were made between 670 occupations and 70 specific causes of death according to sex, age, and calendar period groupings. Relative risks and 95% confidence intervals have been determined. Results have been produced in the form of tables both by occupation and by cause of death, and as an ASCII file, and are being made available by Statistics Canada as a CD-ROM product.

Conclusions: The establishment of this occupational surveillance system has been accomplished through the ability to combine data from different sources such as survey, national indexes and administrative data. This has the benefit of minimising response burden, minimising cost and maximising the use of information to produce a public good. Where excess mortality risk is apparent, it is intended to spark further research to gain confirmation and etiologic knowledge. Ultimately, the hope is that this work may lead to action to reduce or prevent excess mortality risk associated with specific occupations.

KEY WORDS: occupation; occupational health; health surveillance; mortality; record linkage; epidemiology

1. INTRODUCTION

1.1 Introduction and Outline

Surveillance for health has evolved from watching for serious outbreaks of communicable diseases, such as smallpox, into other areas of public health such as chronic diseases, e.g. stroke, cardiovascular disease, cancers, occupation safety and health and environmental health. Surveillance has assumed major significance in disease control and prevention.

1 Maureen Carpenter, Statistics Canada, Health Statistics Division, 18th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6
2 Kristan Aronson, Community Health and Epidemiology, Faculty of Health Sciences, Queens University, Kingston, Ontario, Canada K7L 3N6
3 Martha Fair, Statistics Canada, Health Statistics Division, 18th Floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6
4 Geoffrey Howe, Division of Epidemiology, School of Public Health, Presbyterian Hospital, Columbia, University, New York, New York, U.S.A.

Work-related deaths stand out because of their large financial and personal cost. Recently, it has been projected that the "full" costs of occupational injuries and illnesses (excluding environmental) are on a par with each of cancer and heart disease, with the numbers around U.S.$170 billion (Leigh et al, 1997). This would also be true for Canada normalised for the population. The development of an *occupational surveillance system* is, therefore, highly needed.

This paper will describe the developments towards creating such a system for Canada. Some background includes a current definition of occupational surveillance and describes recent interest in this subject. This is followed by the objectives of this particular project and the methods used to create this surveillance system in Canada. Selections from the numerous results are given to show some of the various ways the results can be extracted, set up and viewed when the CD-ROM is available.

1.2 Definition and Demand for Occupational Surveillance

A current definition of occupational surveillance is the systematic collection, evaluation and dissemination of data relating to workplace exposures, to disease or mortality among workers, with the ultimate aim of reducing and preventing excess risks (Langmuir, 1963; Baker et al, 1989).

There have been many demands for more systematic information on occupational health outcomes. In the 1990's, especially, there has been a growing interest in women's occupational health with their increasing participation in the labour force, both in Canada and internationally (Statistics Canada, 1995). A number of national and international conferences and meetings have recently been increasing these demands for improvements in the quality and quantity of data collected regarding occupational health. In discussions of current issues, all these areas have emphasised the establishment of appropriate databases in occupational health as a priority (Health and Welfare Canada, 1992; JOEM 1994/1995).

Workplace injuries and occupational illnesses exact a large toll on the health of workers and most are preventable. Previous occupational studies show that rates of cancer mortality attributable to occupational deaths vary from 4% to 25% for specific causes (Doll and Peto, 1981; Siemiatycki, 1991; Silverman et al., 1990). These percentages translate into large numbers of people. Occupational injury rates rose by about one-third in Canada from 1955 to 1987, while rates were declining in most other OECD countries (OECD, 1989).

Canadian adults spend about one-quarter of their lives at work (Health Canada, 1994). The workplace is a major physical and social environment, so initiatives to make the workplace a safe and healthy setting is a key element for ensuring population health.

2. OBJECTIVES

The objectives are:

- to create an occupational surveillance system in Canada;
- to collect, link, evaluate, and disseminate data relating to occupation and mortality;
- to detect unsuspected associations between workplace situations and specific causes of death in Canada; and
- to encourage further studies to identify the cause, when excess risk of mortality is found.

Ultimately the aim is to reduce or prevent any excess risk of death among workers.

3. UNIQUE FEATURES ABOUT THIS OCCUPATIONAL SURVEILLANCE SYSTEM

There are several unique features that allowed the construction of this Canadian occupational surveillance system (Newcombe, 1974). First, the presence and recognition of important **existing centralised files** of machine-readable records for: a) job history survey information for 1942-1971 (from 1965 with Social

Insurance Number (S.I.N.)) for a working population; b) death registrations for the whole of Canada developed into the Canadian Mortality Data Base (CMDB) from 1950; and c) tax filing information, developed into a summary historical file from 1984.

Second, the availability of the **Generalized Record Linkage System (GRLS)**, a probabilistic system at Statistics Canada, enabled these records to be brought together. Without this system there would have been a problem, as there is no unique number available on all records. Development of computer record linkage methods enabled the records to be brought together.

Third, the plan to make available **all the results on a CD-ROM**, rather than selected results only in a publication. This will allow access to all who wish to examine and follow-up these data further. The CD-ROM will include all the results: i) in an ASCII file and ii) in prepared tables. (See section 6. The Future and Summary, for more details).

The major features contributing to this *Occupational Surveillance System* are given in Figure 1.

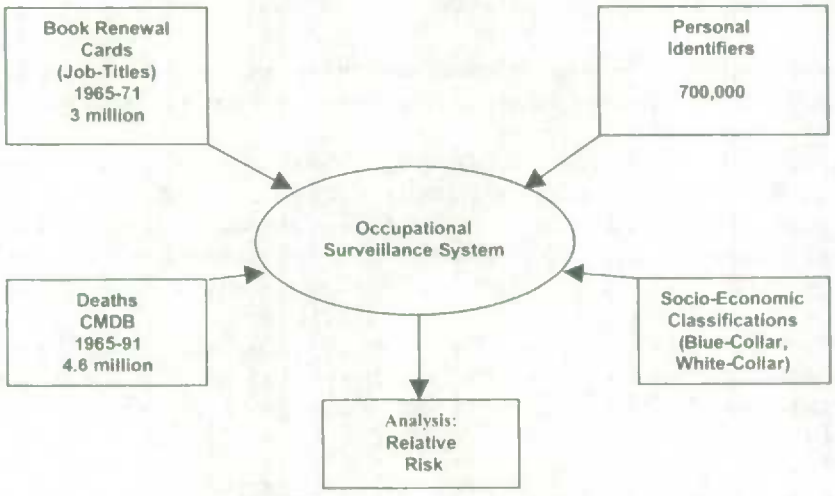


Figure 1. Combining data for a Canadian Occupational Surveillance System

4. METHODS

4.1 Establishing the Cohort

Employment Canada (formerly the Canadian Employment and Immigration Commission, CEIC) undertook an annual national survey of employers between 1942 and 1971. Survey results were available at Statistics Canada in the form of "book renewal cards" giving occupation codes. Social Insurance Numbers (S.I.N.'s) were first allocated in Canada in 1964. It was, therefore, only feasible to use these data from 1965.

Employers provided information for each employee whose Social Insurance Number ended in the digit "4" for the years 1965-1968; and ending in the digit "4" preceded by an odd number for the years 1969-1971. The data, therefore, constituted a 10% sample of Canadian workers who were covered under the Unemployment Insurance Act for 1965-1968 and 5% for 1969-1971. The information collected was S.I.N., surname, occupation code and industry code, year, province and area code in which that individual was currently employed. Additional identifiers needed to link this cohort to the national death file in Canada, the Canadian Mortality Data Base were obtained by Statistics Canada from the S.I.N. Index file, also held at Employment Canada.

4.2 Creating Longitudinal Individual Job Title Histories

Job title histories were collected across Canadian industry on unemployment insurance book renewal cards. Approximately 3 million records were collected during the Employment Canada surveys for the years 1965-1969 and 1971 data (data for the year 1970 was lost before the start of this study). Records referring to the same person needed to be brought together, since one individual could possibly have up to 6 records on the file. This was done using S.I.N., and then cross-checking using the name and sex, to ensure the correct records were combined for the same person. This yielded a large cohort and job histories for 699,420 individuals: 457,224 men and 242,196 women.

An individual was included in a particular occupation if employed during at least one year of the survey in that occupation. Individuals working in more than one occupation were, therefore, counted again for each occupation in which they worked for one year or more. There was an average of 1.06 jobs per person. Person-years were calculated from entry into the cohort until the time of death, or to the end of 1991, if no death occurred. Two age-at-death groups were considered: i) all ages 15 and over, and ii) ages 15 to 64.

4.3 Establishing Socio-Economic Status

The cohort was divided into two broad groups in an effort to take socio-economic status into consideration. This was done to help control for lifestyle confounding factors, e.g. smoking, diet or alcohol consumption.

First, each occupation was classified into one of 6 levels of the Pineo scale (a scale with up to 16 levels for classifying occupations according to income and status) (Pineo, 1985). The first 3 levels were grouped into "white-collar" jobs. These included professional and high level managerial; semi-professional, technical and middle managerial; supervisors, foremen and forewomen. The other 3 levels were grouped into "blue-collar" jobs. These included skilled; semi-skilled; and unskilled workers and employees.

Each standard occupation code (SOC) included in the study was allocated a "blue" or "white" classification code. All comparisons were made between one occupation and all occupations within the same sex and occupational classification code (i.e. blue- or white-collar) to produce relative risks.

4.4 Combining Cohort, Death and Summary Tax Records to Establish Fact and Cause of Death

The cohort was linked with the Canadian Mortality Data Base using the Generalized Record Linkage System (GRLS) to determine the mortality experience of the cohort. The **CMDB** is a computer file of deaths for all of Canada with date, place and cause of death from 1950 as notified by all the Vital Statistics Registrars in each Canadian province and territory. It has been used for many record linkage health follow-up studies in Canada since its establishment within Statistics Canada, Health Statistics Division in the late 1970's (Smith and Newcombe, 1982). **GRLS** is a probabilistic record linkage system, based on methods originally developed in Canada by Newcombe (Newcombe et al., 1959) and later developed further to form a generalized record linkage system by Howe and Lindsay (Howe and Lindsay, 1981). The underlying theory for these ideas was firmly established by Fellegi and Sunter (Fellegi and Sunter, 1969).

Two previous linkages have been undertaken for this cohort: i) to the end of 1973 (Howe and Lindsay, 1983), and ii) to the end of 1979 (Lindsay, Stavraky and Howe, 1993). This new linkage follow-up covered the whole period from 1965 to the end of 1991. This ensured consistent quality of the linkage procedure over the entire study period and used improved computer matching techniques developed in the interim period. The smaller size of the group of women and their relatively young age has prohibited meaningful analysis until now (Aronson and Howe, 1994).

Another innovative procedure was also undertaken within Statistics Canada. To both complement and verify the death search results, an "alive" search was undertaken. For this, the occupational cohort was matched to a **summary historical tax file** from 1984-91 to determine if the person was alive, dead or

emigrated. This methodology reduces the number of cases lost to follow-up and verifies the results of the death linkage.

4.5 Combining Causes of Death Across Time from 1965-1991

Historic individual cause of death codes were combined into 70 cause of death code groups. The cause of death on the Vital Statistics death registrations at Statistics Canada is coded and classified according to the International Classification of Diseases (ICD). These ICD codes are classified for 1965-1967 in Revision 7 (WHO, 1957), for 1968-1978 in Revision 8 (National Center for Health Statistics, 1968) and for 1979 - 1991 in Revision 9. (WHO, 1977);

Conversions were made from these three ICD classifications to combine them into 70 grouped causes of death for use in this study. The 70 grouped causes include infectious diseases; chronic diseases including cancers, heart, asthma, etc.; accidents, including external causes of homicide and suicide; and combined groups which include all leukemias, all cancers and all causes.

5. RESULTS

The results from combining the data has created a large cohort of approximately 700,000 individuals (457,224 men; 242,196 women i.e. two-thirds men; one-third women). Individual job titles were assembled by occupation codes to give 261 - 1961 codes and 409 - 1971 codes. Socio-economic status was allocated for each occupation code within two classifications to produce 432 blue-collar occupations and 238 white-collar occupations.

These individuals were followed up for up to 27 years from 1965-1991. Men contributed over 11 million person-years and women more than 6 million person-years. A total of 142,800 deaths with cause of death (~19% of the cohort) were found through the CMDB up to the end of 1991; 116,000 deaths among males, 26,800 deaths among females.

In addition to the ~19% deaths, the "alive" follow-up verified that an additional ~70% of the cohort was still alive in 1991. The remainder of ~11% were not found to be dead, or among the tax filers for 1991. This remaining percentage would be a mixture of those dying outside Canada, non-residents/emigrants who are still alive, and/or living persons in Canada who did not file a tax return in 1991. Elsewhere, evaluations of linking a specific cohort to the CMDB and comparing computerised linkage with more traditional manual follow-up methods have reported about 4% of deaths not found (Shannon et al., 1989; Goldberg et al., 1993). This is mainly due to deaths occurring outside Canada.

As mentioned, 70 causes of death groupings were created across time, for 1965-1991 from combining ICD-7, ICDA-8 and ICD-9 codes. There were over 27,000 comparisons made between 670 occupations and 70 specific causes of death. Relative risks and 95% confidence intervals were determined.

Selected results have been published elsewhere (Aronson et al., 1999). Some examples are shown in Figures 2 - 4 below. Figure 2 shows 15 results that meet very stringent reporting criteria. Figure 3 shows an example of how results can be viewed by occupation. Figure 4 shows examples of results viewed by selected causes of death.

Figure 2: Fifteen Potential Associations Between Causes of Death and Occupations Deaths ≥5, Relative Risk >1.50 and P-Value <0.0001

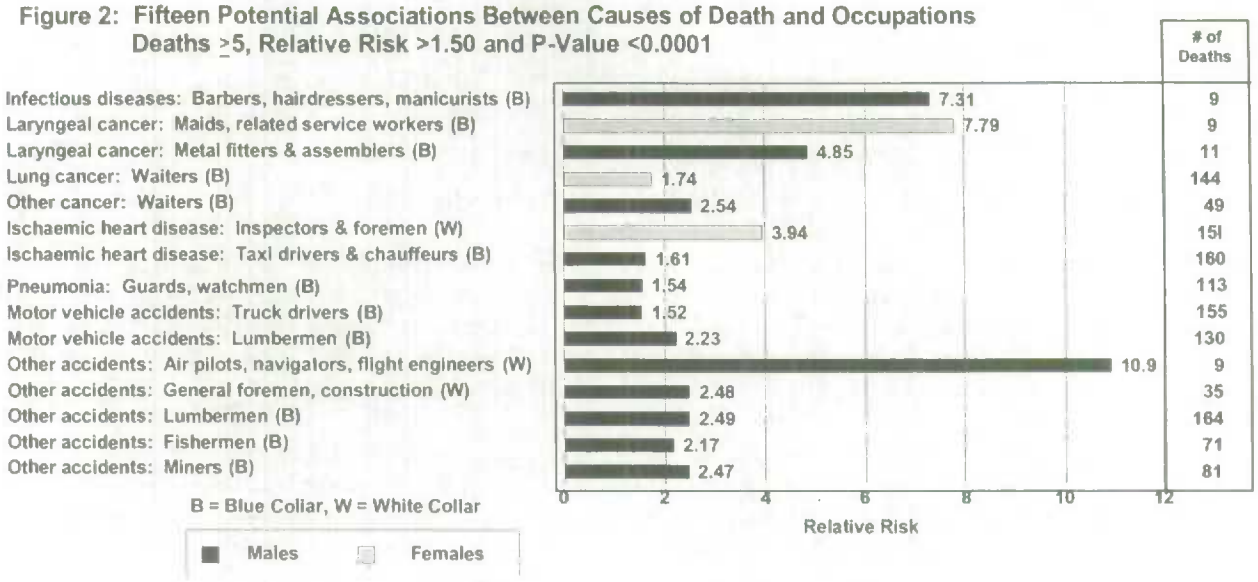


Figure 3: Potential Associations With the Occupation of Truck Driver Among Men (All Ages) Deaths >5, Relative Risk ≥1.30 and P-Value <0.05

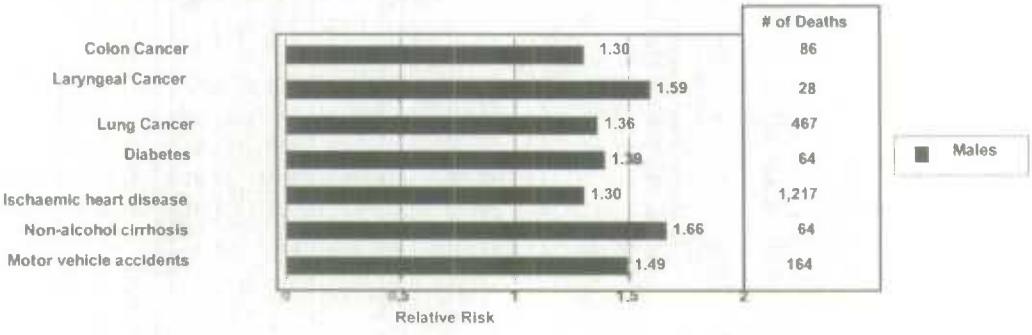
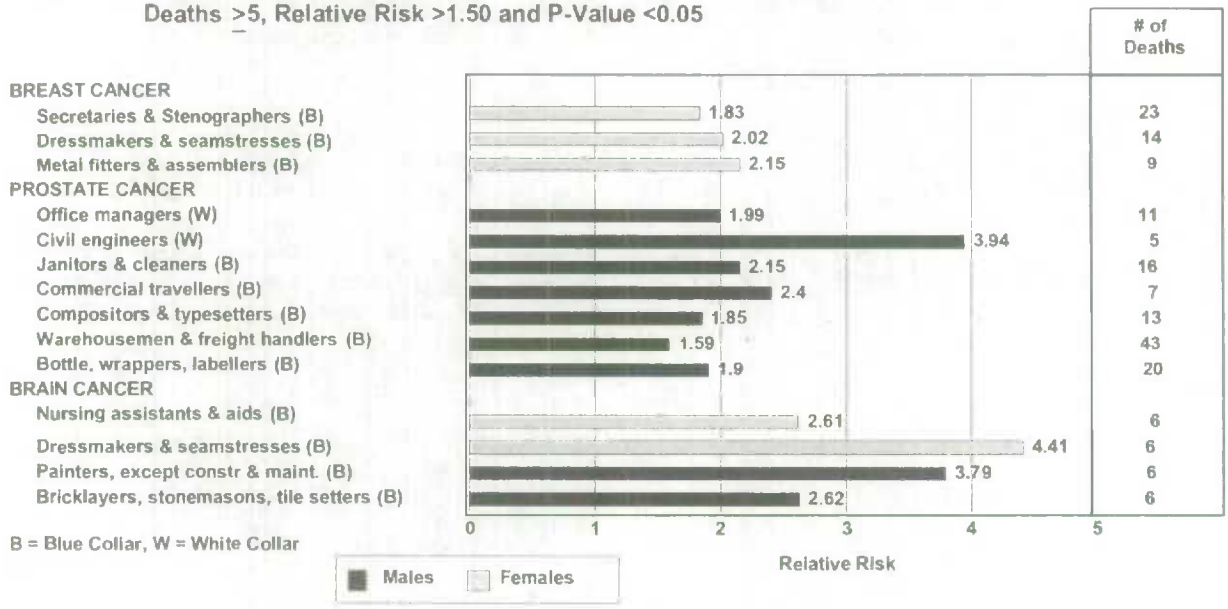


Figure 4: Potential Associations Between Causes of Death and Occupations Deaths ≥5, Relative Risk >1.50 and P-Value <0.05



6. THE FUTURE AND SUMMARY

The full results are being produced on a CD-ROM product to be released by Statistics Canada (Aronson et al., 2000, CD-ROM). This CD-ROM will include: i) an ASCII file containing 14 variables; ii) full tables by occupation (670) and by cause of death (70) giving all the results including the Relative Risk, the 95% confidence intervals and the p-value for each case; iii) selected tables by occupation and by cause of death; selected on the basis of Relative Risk greater than 1.0; observed deaths greater or equal to 5; and p-value less than 0.05; iv) documentation giving details of the study; v) 8 Appendices including a full list of the 1961 and 1971 Standard Occupational Codes by code number and alphabetically; cause of death codes by code group number and alphabetically; a record layout and data dictionary for the ASCII file; and a glossary of terms.

Through the use of a unique epidemiological study, we have consolidated information in a systematic way to help identify both known and previously unsuspected potential associations between occupations and cause-specific mortality risk. This is an examination of the "big picture" to help guide the generation of hypotheses for more detailed studies for areas where excess risks are found here. Plans include providing the information for all those who need to know, particularly those who can take further action, e.g. occupational health researchers, management, unions and governments.

ACKNOWLEDGEMENTS

The authors wish to especially thank Pierre Lalonde, Russell Wilkins, Lucille Poirier, Sevag Serimazi and Tanya Weston. The National Cancer Institute of Canada, through their Epidemiology Unit, originally provided funding to establish this cohort. Health Canada, through the National Health Research and Development Program (NHRDP), funded the record linkage and analysis. KJA was supported in part through a Research Scholar Award from NHRDP and is now supported through a Career Scientist Award from the Ontario Ministry of Health.

REFERENCES

Aronson, K.J. and Howe, G.R. (1994). Utility of a surveillance system to detect associations between work and cancer among women in Canada, 1965-1991. *Journal of Occupational Medicine,* **36**:1174-1179.

Aronson, K.J. and Howe, G.R., Carpenter, M., Fair, M.E. (1999). Surveillance of potential associations between occupations and causes of death in Canada, 1965-1991. *Occupational and Environmental Medicine,* **56**: 265-269.

Aronson K.J., Howe G.R., Carpenter, M. and Fair, M.E. (2000). Occupational Surveillance in Canada: Cause-Specific Mortality Among Workers, 1965-1991. Statistics Canada, Ottawa, Ontario CD-ROM Cat. No. 84-546-XCB 91001 (in preparation).

Baker, E.L., Honchar, P.A. and Fine, L.J. (1989). I. Surveillance in occupational illness and injury: Concepts and content. *American Journal of Public Health,* **79** Suppl: 9-11.

Doll, R. and Peto, R. (1981) The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today, *Journal of the National Cancer Institute*, **66**: 1191-1308.

Fellegi, I.P. and Sunter, A.B. (1969) A theory of record linkage. *Journal of the American Statistical Association,* **40**, 1183-1210.

Goldberg, M.S., Carpenter, M., Thériault, G. and Fair, M.E. (1993). The accuracy of ascertaining vital status in a historical cohort study of synthetic textile workers using computerized record linkage to the Canadian Mortality Data Base. *Canadian Journal of Public Health,* **84**, No. 3: 201-204.

Health and Welfare Canada (1992). *Proceedings of the Research Round Table on Gender and Workplace Health, June 22-23, 1992, Ottawa, Canada.* Office of the Senior Adviser Status of Women.

Health Canada (1994). *Strategies for Population Health, Investing in the Health of Canadians*. Prepared by the Federal, Provincial and Territorial Advisory Committee on Population Health for the Meeting of the Ministers of Health, Halifax, Nova Scotia, September 1994. Ottawa, 1994.

Howe, G.R. and Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Comput Biomed Res*, **14**: 327-340.

Howe, G.R. and Lindsay, J.P. (1983). A follow-up study of a ten-percent sample of the Canadian labour force. I. Cancer mortality in males, 1965-73. *Journal of National Cancer Institute,* **70**: 37-44.

Journal of Occupational and Environmental Medicine (JOEM) (1994/1995). *Proceedings, International Conference on Women's Health: Occupation and Cancer*, Baltimore, Maryland, November 1993. August 1994, November 1994, March 1995 issues.

Langmuir, A.D. (1963). The surveillance of communicable diseases of national importance. *New England Journal of Medicine,* **268**: 182-192.

Leigh, J.P., Markowitz, S.B., Fahs, M.,Shin, C. and Landrigan, P.J. (1997) Occupational injury and illness in the United States. Estimates of costs, morbidity and mortality. *Arch Intern Medicine*, **157**:1557-1568.

Lindsay, J.P., Stavraky, K.M. and Howe, G.R. (1993). Cancer mortality among males 1965-1979, in a 10-percent sample of the Canadian labour force. *Journal of Occupational Medicine,* **35**: 408-414.

National Center for Health Statistics (1968). *International Classification of Diseases, Adapted 8th Revision.* U.S. Dept. of Health, Education and Welfare, Public Health Service, Washington, D.C.

Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959). Automatic linkage of vital records. *Science*, 130: 954-959.

Newcombe, H.B. (1974). *A Method of Monitoring Nationally for Possible Delayed Effects of Various Occupational Environments*. NRCC No. 13686. National Research Council of Canada, Ottawa.

Organization for Economic Co-operation and Development (1989). *The OECD Employment Outlook*, Paris.

Pineo, P.C. (1985). *Revisions of the Pineo-Porter-McRoberts Socio-Economic Classification of Occupation for the 1981 Census.* QSEP (Quantitative Studies in Economics and Population) Research Report No. 125. McMaster University, Hamilton, Ontario.

Shannon, H., Jamieson, E., Walsh, C., Julian, J., Fair, M.E. and Buffet, A. (1989). Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. *Canadian Journal of Public Health,* **80**: 54-57.

Siemiatycki, J. (ed.) (1991). *Risk factors for cancer in the workplace*. Boca Raton: *CRC Press*.

Silverman, D.T., Levin, L.I. and Hoover R.N. (1990). Occupational risks of bladder cancer among white women in the United States. *American Journal of Epidemiology*, **132** (3): 453-461, 1990.

Smith, M.E. and Newcombe, H.B. (1982). Use of the Canadian Mortality Data Base for Epidemiological Follow-up. *Canadian Journal of Public Health*, 73: 39-46.

Statistics Canada (1995). *Women in Canada: A Statistical Report*. 3rd edition, 1995. Statistics Canada Cat. No. 89-503XPE. Ottawa.

Statistics Canada (1999). *GRLS – Generalized Record Linkage System, Concepts*. Research and General Systems, Systems Development Division, Statistics Canada.

World Health Organization (1957). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death, 7th Revision*, Geneva.

World Health Organization (1979). *Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death*. Based on the Recommendations of the Ninth Revision Conference, 1975, Geneva.

SESSION III

THE USE OF META ANALYTICAL TECHNIQUES IN POPULATION HEALTH RISK ASSESSMENT

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

META ANALYSIS OF BIOASSAY DATA FROM THE U.S. NATIONAL TOXICOLOGY PROGRAM

Kenny S. Crump,[1] Daniel Krewski,[2] Cynthia Van Landingham[1]

ABSTRACT

A meta analysis was performed to estimate the proportion of liver carcinogens, the proportion of chemicals carcinogenic at any site, and the corresponding proportion of anticarcinogens among chemicals tested in 397 long-term cancer bioassays conducted by the U.S. National Toxicology Program. Although the estimator used was negatively biased, the study provided persuasive evidence for a larger proportion of liver carcinogens (0.43, 90% CI: 0.35, 0.51) than was identified by the NTP (0.28). A larger proportion of chemicals carcinogenic at any site was also estimated (0.59, 90% CI: 0.49, 0.69) than was identified by the NTP (0.51), although this excess was not statistically significant. A larger proportion of anticarcinogens (0.66) was estimated than carcinogens (0.59). Despite the negative bias, it was estimated that 85% of the chemicals were either carcinogenic or anticarcinogenic at some site in some sex-species group. This suggests that most chemicals tested at high enough doses will cause some sort of perturbation in tumor rates.

KEY WORDS: Meta analysis; bioassay data; trend test; carcinogenesis

1. INTRODUCTION

The National Toxicology Program (NTP) has been testing chemicals for carcinogenic potential for about 30 years. Of the approximately 400 chemicals that have been tested, about 50% have been identified as carcinogens. About 25% of the tested chemicals have been found to be carcinogenic to the liver, which is the most frequent site of carcinogenesis in NTP bioassays (Huff et al., 1991). Results from these bioassays are used by regulatory agencies charged with protection of public health and by the scientific community concerned with carcinogenic risk assessment. Consequently, these studies have important health and economic consequences. Chemicals identified as carcinogenic in animal bioassays are generally regulated much more stringently than chemicals not so identified.

In this paper, data from 397 NTP bioassays are used to estimate the portion of the chemicals that were carcinogenic to the liver, and the proportion that were carcinogenic at any site, in any experimental animal group. Estimates are also developed of the proportion of the 397 chemicals that were anticarcinogenic, or either carcinogenic or anticarcinogenic. A chemical that caused a dose-related increase in tumor incidence at one site and a dose-related decrease at a different site would be considered both a carcinogen and an anticarcinogen. More detailed accounts of this work may be found elsewhere (Crump et al., 1998, 1999).

---

[1]ICF Kaiser, 602 East Georgia Avenue, Ruston, LA 71270 USA
[2]University of Ottawa, Department of Epidemiology and Community Medicine, Ottawa Ontario CANADA

2. METHODS

Data from 397 long-term carcinogenicity bioassays were obtained from NTP data archives (Crump et al., 1998, 1999). Most of these studies involved mice and rats, and our analysis was restricted to these two species. Generally, males and females of a species were tested in separate experiments, which involved two or three dose groups of about 50 animals each, in addition to a control group of that same size (although control groups in a few of the earlier studies contained as few as 10 animals).

The procedure used to estimate the number of carcinogens in the NTP data base is based on the empirical distribution of p-values obtained from a statistical test applied to the individual studies (Bickis et al., 1996; Crump and Krewski 1998), and is now briefly described. If none of the chemicals had any effect upon tumor rates, these p-values would be uniformly distributed between zero and one, which means that the cumulative distribution of p-values would graph as a straight line from the point (0,0) to the point (1,1), as illustrated in Figure 1A. However, if, e.g., 60% of the chemicals were carcinogenic—so highly carcinogenic that the corresponding p-values were essentially zero (Figure 1B)—the cumulative distribution would still plot as a straight line, and this line would intersect the y-axis at 0.6, the proportion of carcinogens. Figure 1C depicts a more realistic case in which the proportion of carcinogens is still 60% but the carcinogenic responses are weaker, so that the p-values from the carcinogens are not all zero. In this case a tangent line drawn at any point on the theoretical cumulative distribution of p-values will intersect the y-axis at a point that is less than 0.6 (the proportion of carcinogens), but the intersection point will be nearer to 0.6 if the tangent line is drawn at a point closer to $p = 1$. Figure 1D is a modification of Figure 1C in which the proportion of carcinogens remains at 0.6, but, in addition, 20% of the chemicals are anticarcinogens. In this case, any tangent line will intersect the y-axis at a value that is less than 0.6. However, as suggested by Figure 1D, the point of intersection will be largest and consequently nearer to the true proportion of carcinogens, when the tangent line is drawn through the inflection point of the curve.

This discussion suggests estimating the proportion of carcinogens from the empirical distribution function, $\hat{F}(p)$ (defined as the proportion of studies with p-values $\leq p$), by drawing a secant line through two points, a and b, of the graph of $\hat{F}(p)$, and using as the estimator the y-intercept of this secant line. This estimator is $[b\,\hat{F}(a) - a\,\hat{F}(b)]/(b - a)$, which under fairly general conditions, has the following properties (Crump and Krewski, 1998), all of which are suggested by the above discussion:

1) The estimator is biased low no matter how a and b are chosen.
2) This negative bias is smallest when a and b are selected near the inflection point of $F(p)$, the expected value of $\hat{F}(p)$.
3) For a given value of b, the value of a that minimizes the bias is on the opposite side of the inflection point from b.
4) The variance of the estimator becomes large when a and b are close together.

These properties suggest selecting a and b near to, and on opposite sides of, the inflection point of the graph of $\hat{F}(p)$, but not so near to the inflection point that the variance becomes excessively large.

To use this procedure to estimate the proportion of NTP chemicals that were liver carcinogens, we first calculated p-values from the POLY3 test (Bailar and Portier, 1988) applied to the liver tumor data from all dose groups of each sex-species-specific experiment for each tested chemical. The POLY3 test is an age-adjusted test of trend, and has recently been adopted by the NTP as the statistical test of choice for the NTP carcinogenesis bioassays. The p-value for a test of the hypothesis that a chemical was carcinogenic to the liver in at least one sex-species group was then defined as

$1 - (1 - p_{min})^k$, where p_{min} was the minimum POLY3 p-value from all sex-species experiments and k was the number of such experiments for a chemical (for the majority of chemicals, $k = 4$, corresponding to the fact that separate experiments had been performed in both sexes of mice and rats). This p-value can be shown to be theoretically uniformly distributed in the null case (when the exposure does not affect tumor rates in any sex-species group).

Calculation of p-values needed to estimate the proportion of chemicals that were carcinogenic at any site in any sex and species group was somewhat more complicated. First the POLY3 test was applied to each of 93 tumor categories defined so that they were similar to the categories routinely analyzed by the NTP. The test statistic, T, for an effect in a sex-species experiment was the largest of these POLY3 test statistics, after application of a continuity correction, derived from any tumor category. In order to insure that the test based on T had the proper false positive rate, its p-value was determined using a randomization procedure (Farrar and Crump, 1990).

3. RESULTS

Figure 2A gives the results of the analysis used to estimate the number of liver carcinogens. The empirical distribution of POLY3 p-values is considerably above the line y = x for small values of p, which indicates the presence of liver carcinogens. This graph also shows some evidence of anticarcinogenesis, as the graph lies slightly below the graph y = x for p-values close to 1.0. This evidence for anticarcinogenesis was somewhat surprising, since this graph was based on p-values determined from the minimum of generally four p-values, and it might be expected that, e.g., anticarcinogenicity in one sex-species group would have relatively little effect upon the minimum of four p-values.

Figure 2A also shows estimates of the number of liver carcinogens among the NTP-tested chemicals obtained for various values of the parameter a between 0.2 and 0.55, with b selected as $1.2 - a$. This graph provides evidence for more liver carcinogens than were identified by the NTP as both the point estimate and the 95% lower confidence bound are above the NTP estimate (0.28) for all values of a.

Figure 2B presents results of the analysis to estimate the proportion of chemicals that were carcinogenic at any site in any sex-species group. The graph of the estimate of the proportion of carcinogens shows the estimate for all values of a between 0 and 1, with b fixed at $b = 1$. Although the point estimate of the proportion of carcinogens is higher than the proportion identified by the NTP (0.51) for all values of a except the smallest (where the estimator has the greatest amount of negative bias) and largest (where the variance of the estimator is greatest), the 95% lower bound on our estimate is close to or below the proportion of carcinogens obtained by the NTP for most values of a. Thus, evidence for an excess proportion of chemicals that were carcinogenic at any site over the proportion obtained by the NTP is weaker than the corresponding evidence for liver cancer.

Figures 2C and 2D contain results of the analyses used to obtain estimates of the proportion of chemicals that were anticarcinogenic, or either carcinogenic or anticarcinogenic, respectively, at any site in any sex or species. Figure 2C indicates the presence of a considerable amount of anticarcinogenesis, and Figure 2D suggests that most chemicals were either carcinogenic or anticarcinogenic at some site in some sex-species group.

Table 1 contains representative estimates of the proportion of liver carcinogens and the proportion of chemicals that were carcinogenic, anticarcinogenic or either. The estimated proportion of liver carcinogens was 0.43 (90% CI: 0.35, 0.51), which was significantly larger than the proportion identified by the NTP (0.28). Although the estimated proportion of chemicals carcinogenic at any

site was 0.59 (90% CI: 0.49, 0.69), compared to the NTP estimate of 0.51 this excess was not statistically significant. The estimated proportion of anticarcinogens was 0.66, which was higher than the estimate of the proportion of carcinogens. The proportion of chemicals that was either carcinogenic or anticarcinogenic was estimated as 0.85 (90% CI: 0.78, 0.91).

4. DISCUSSION

This study estimated that 43% of NTP chemicals were liver carcinogens and 59% were carcinogenic at some site. Although both proportions are greater than the NTP estimate, only the excess of liver tumors is statistically significant. We estimated that there were more anticarcinogens (0.66) than carcinogens (0.59) among NTP chemicals. An analysis that used a conventional significance level of 0.05 to detect effects would not have discovered this, since the proportion of chemicals having a p-value <0.05 for anticarcinogenesis was smaller than the corresponding proportion for carcinogenesis. Estimating a larger proportion of anticarcinogens than carcinogens was unexpected because chemicals were selected for study by the NTP on the basis of suspected carcinogenicity, and also because anticarcinogenesis should be inherently more difficult to detect than carcinogenesis due to the relatively low carcinogenic background.

It was estimated that 85% of the chemicals studied by the NTP were either carcinogenic or anticarcinogenic at some site in some sex-species group of rodents. It should be kept in mind that the estimator used to obtain this estimate is inherently negatively biased. This suggests that most chemicals, when given at sufficiently high doses, may cause perturbations that affect tumor responses, causing increases at some sites and decreases at others.

Acknowledgment. The authors would like to thank Dr. Joe Haseman for his invaluable help with this project.

REFERENCES

Bickis, M., Bleuer, S., and Krewski, D. (1996). Estimation of the proportion of positives in a sequence of screening experiments. *Canadian Journal of Statistics* 24, 1-16.

Crump, K.S., Krewski, D., and Wang, Y. (1998). Estimates of the number of liver carcinogens in bioassays conducted by the National Toxicology Program. *Risk Analysis* 18, 299-308.

Crump, K.S., Krewski, D., and Van Landingham, C. (1999). Estimates of the proportion of chemicals that were carcinogenic, anticarcinogenic or either in bioassays conducted by the National Toxicology Program. *Environmental Health Perspectives* 107, 83-88.

Crump, K.S., and Krewski, D. (1998). Estimation of the number of studies with positive trends when studies with negative trends are present. *Canadian Journal of Statistics* 26, 643-655.

Farrar, D., and Crump, K.S. (1990). Exact statistical tests for any carcinogenic effect in animal bioassays. II. Age-adjusted tests. *Fundamental and Applied Toxicology* 15, 710-721.

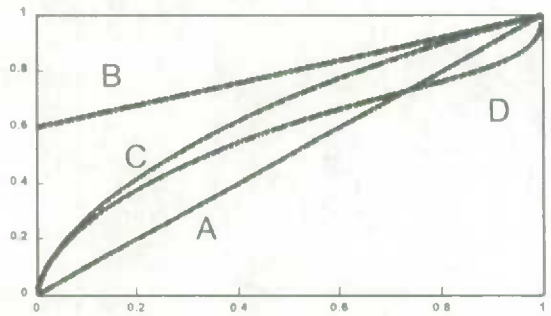
Huff J., Cirvello, J., Haseman, J., and Bucher, J. (1991). Chemicals associated with site-specific neoplasia in 1394 long-term carcinogenesis experiments in laboratory rodents. *Environmental Health Perspectives* 93, 247-270.

Table 1
Representative Estimates of the Proportion of Chemicals That Were Carcinogenic to the Liver of Any Sex-species Group, Carcinogenic Overall (At Any Site in Any Sex-species Group), Anticarcinogenic Overall, or Either Carcinogenic or Anticarcinogenic Overall

	Estimated Proportion	90% C.I.	NTP Estimate
Liver Carcinogenic[a]	0.43	(0.35, 0.51)	0.28
Carcinogenic Overall[b]	0.59	(0.49, 0.69)	0.51
Anticarcinogenic Overall[c]	0.66	(0.56, 0.75)	
Carcinogenic or Anticarcinogenic Overall[d]	0.85	(0.78, 0.91)	

[a] Obtained from Figure 2A using a = 0.375.
[b] Obtained from Figure 2B using a = 0.75.
[c] Obtained from Figure 2C using a = 0.75.
[d] Obtained from Figure 2D using a = 0.75.

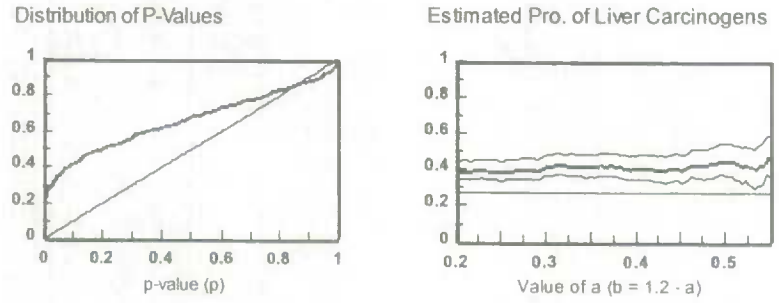
Figure 1
Theoretical Distribution of P-values Under Various Conditions



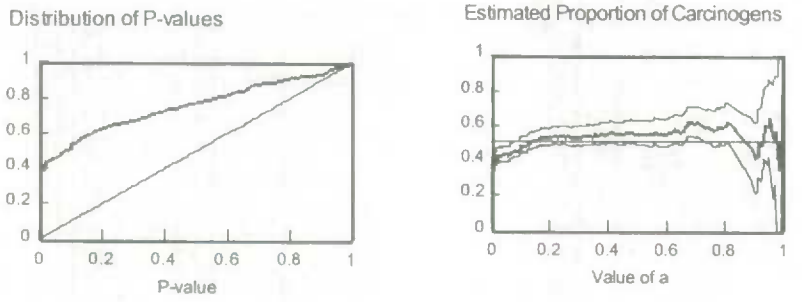
A No carcinogens or anticarcinogens.
B 60% of chemicals are highly carcinogenic.
C 60% are carcinogenic, but not so extremely as in Curve B.
D 60% of chemicals are carcinogenic (as in C), and an additional 20% are anticarcinogenic.

Figure 2
Results of Analysis to Estimate Proportion of Chemicals Tested by NTP
That Were Carcinogenic or Anticarcinogenic in Any Sex-species Group

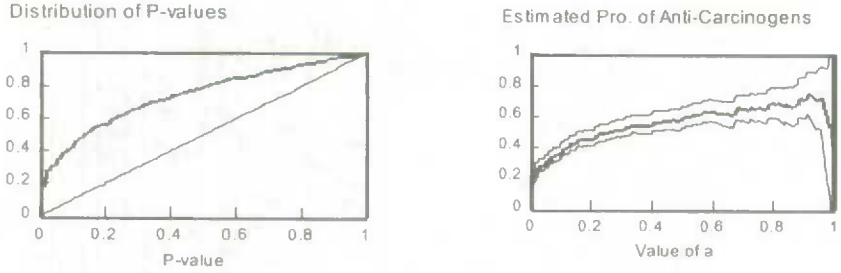
A. Carcinogenesis, Liver



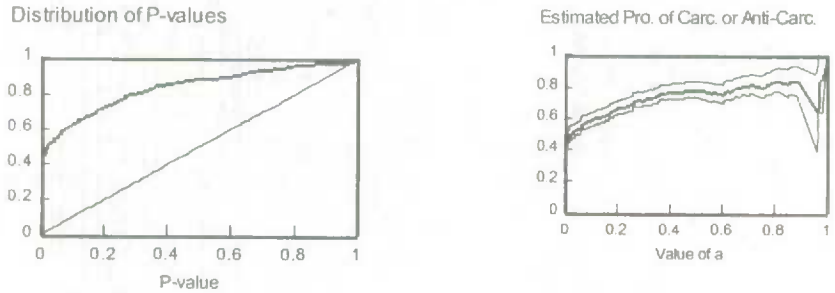
B. Carcinogenesis, Any Site



C. Anticarcinogenesis, Any Site



D. Carcinogenesis or Anticarcinogenesis, Any Site



Solid wavy line is point estimate; narrower wavy lines define 90% C.I. on point estimate; horizontal line indicates NTP estimate.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

PARTICULATE MATTER AND DAILY MORTALITY: COMBINING TIME SERIES INFORMATION FROM EIGHT US CITIES

Francesca Dominici, Jonathan M. Samet and Scott L. Zeger[1]

ABSTRACT

Time series studies have shown associations between air pollution concentrations and morbidity and mortality. These studies have largely been conducted within single cities, and with varying methods. Critics of these studies have questioned the validity of the data sets used and the statistical techniques applied to them; the critics have noted inconsistencies in findings among studies and even in independent re-analyses of data from the same city.

In this paper we review some of the statistical methods used to analyze a subset of a national data base of air pollution, mortality and weather assembled during the National Morbidity and Mortality Air Pollution Study (NMMAPS). We present log-linear regression analyses of daily time series data from the largest 8 U.S. cities and use hierarchical regression models for combining estimates of the pollution-mortality relationship for particulate matter less than 10μ in aerodynamic diameter (PM_{10}).

Our analyses demonstrate that there is a consistent association of particulate air pollutionPM_{10} with daily mortality aeross the eight largest US cities and that the effect ofPM_{10} is robust respect to the inclusions of other pollutants.

KEY WORDS: Air pollution; Longitudinal data; Hierarchical models; Markov Chain Monte Carlo; Log-linear regression; Mortality; Relative rate

1. INTRODUCTION

1.1. Description of the Problem

In spite of improvements in measured air quality indicators in many developed countries, the health effects of particulate air pollution remain a regulatory and public health concern. This continued interest is motivated largely by recent epidemiologic studies that have examined both acute and longer-term effects of exposure to particulate air pollution in different cities in the United States and elsewhere in the world (Dockery and Pope, 1994; Schwartz, 1995; American Thoracic Society, 1996a; American Thoracic Society, 1996b; Korrick *et al.*, 1998). Many of these studies have shown a positive association between measures of particulate air pollution – primarily total suspended particles (*TSP*) or particulate matter less than 10μ in aerodynamic diameter (PM_{10}) – and daily mortality and morbidity rates. Their findings suggest that daily rates of morbidity and mortality from respiratory and cardiovascular diseases increase with levels of particulate air pollution below the current National Ambient Air Quality Standard (NAAQS) for particulate matter in the United States. Critics of these studies have questioned the validity of the data sets used and the statistical techniques applied to them; the critics have noted inconsistencies in findings among studies and even in independent re-analyses of data from the same city (Lipfert and Wyzga, 1993; Li and Roth, 1995).

---

[1]Francesca Dominici, Jonathan M. Samet and Scott L. Zeger, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, 21205 Baltimore (MD) USA.

These controversial associations have been found using Poisson time series regression models fit to the data using the generalized estimating equations (Liang and Zeger, 1986) or generalized additive models (Hastie and Tibshirani, 1990) methods.

Statistical power of analyses within a single city may be limited by the amount of data for any location. Consequently, in comparison to analyses of data from a single site, pooled analyses can be more informative about whether or not association exists, controlling for possible confounders. In addition, a pooled analysis can produce estimates of the parameters at a specific site, which borrow strength from all other locations (DuMouchel and Harris, 1983; DuMouchel, 1990; Breslow and Clayton, 1993).

In this paper we develop a statistical approach that combines information about air pollution/mortality relationships across multiple cities. We illustrate this method with the following two-stage analysis of data from the largest 8 U.S. cities:

1. Given a time series of daily mortality counts in each of three age groups, we use log-linear generalized additive models (Hastie and Tibshirani, 1990) to estimate the relative change in the rate of mortality associated with changes in PM_{10}, controlling for age-specific longer-term trends, weather, and other potential confounding factors, separately for each city;

2. We then combine the pollution-mortality relative rates across the 8 cities using a Bayesian hierarchical model (Lindley and Smith, 1972; Morris and Normand, 1992) to obtain an over-all estimate.

See Samet *et al.* (1995, 1997) and Kelsall *et al.* (1997) for details on methods for the first stage analyses, summarized in section 3. See Dominici *et al.* (1999) for details on methods for the second stage analyses, summarized in section 4.

1.2. Organization of the Paper

In section 2, we describe the database of air pollution, mortality, and meteorological data from 1987 to 1994 for the 8 U.S. cities in this analysis. In section 3, we fit generalized additive models with log link and Poisson error to produce relative-rate estimates for each location. The Poisson regression was conducted using the previous day's (lag 1). In section 4, we present the hierarchical regression model for combining the estimated regression coefficients. Results of our analysis are summarized in section 5, followed by a discussion in section 6.

2. DESCRIPTION OF THE DATABASE

The database includes mortality, weather, and air pollution data for the 8 largest metropolitan areas (Los Angeles, New York, Chicago, Dallas, Houston, San Diego, Santa-Ana-Anaheim, Phoenix) in the U.S. for the 7-year period 1987-1994 (data not shown).

The mortality data, aggregated at the level of county, were obtained from the National Center for Health Statistics. We focused on daily death counts for each site, excluding non-residents who died in the study site and accidental deaths. Because mortality information was available for counties but not smaller geographic units to protect confidentiality, all predictor variables were aggregated to the county level.

Hourly temperature and dew point data for each site were obtained from the Earth-Info CD[2] database. After extensive preliminary analyses that considered various daily summaries of temperature and dew point as predictors, such as daily average, maximum, and eight-hour maximum, we have used the 24-hour mean for

each day. If there was more than one weather station in a city, we took the average of the measurements from all available stations.

3. CITY-SPECIFIC ANALYSES

In this section, we summarize the model used to estimate the air pollution/mortality relative rate separately for each location, accounting for age-specific longer-term trends, weather, and day of the week. The core analysis for each city is a generalized additive model with log link and Poisson error that accounts for smooth fluctuations in mortality that potentially confound estimates of the pollution effect and/or introduce autocorrelation in mortality series.

To specify our approach more completely, let y_{at}^c be the observed mortality for each age group $a = (\leq 65, 65-75, \geq 75 \text{ years})$ on day t at location c, and x_{at}^c be the air pollution variable. Let $\mu_{at}^c = E(y_{at}^c)$ be the expected number of deaths and $w_{at}^c = \text{var}(y_{at}^c)$. We use a log-linear model $\log \mu_{at}^c = x_{at}^{c'} \beta^c$ for each city c, allowing the mortality counts to have variances w_{at}^c that may exceed their means (*i.e.*, be overdispersed) with the overdispersion parameter φ^c also varying by location so that $w_{at}^c = \varphi^c \mu_{at}^c$.

To protect the pollution relative rates β^c from confounding by longer-term trends due, for example, to changes in health status, changes in the sizes and characteristics of populations, seasonality, and influenza epidemics, and to account for any additional temporal correlation in the count time-series, we estimated the pollution effect using only shorter-term variations in mortality and air pollution. To do so, we partial out the smooth fluctuations in the mortality over time by including arbitrary smooth functions of calendar time $S^c(\text{time}, \lambda)$ for each city. Here, λ is a smoothness parameter which we pre-specified, based upon prior epidemiologic knowledge of the time scale of the major possible confounders, to have seven degrees of freedom per year of data so that little information from time-scales longer than approximately two months is included when estimating β^c. We also controlled for age-specific longer-term temporal variations in mortality, adding a separate smooth function of time with eight degrees of freedom for each age-group.

To control for weather, we also fit smooth functions of the same day temperature (temp_0), average temperature for the three previous days (temp_{1-3}), each with six degrees of freedom, and the analogous functions for dew point (dew_0, dew_{1-3}), each also with three degrees of freedom. Since there were missing values of some predictor variables on some days, we restricted analyses to days with no missing values across the full set of predictors.

In summary, we fit the following log-linear generalized additive model to obtain the estimated pollution relative rate $\hat{\beta}^c$ and the sample variance v^c at each location:

$$\begin{aligned} \log \mu_{at}^c = {} & x_{at}^{c'} \beta^c + \gamma^c \text{DOW} + S_1^c(\text{time}, 7/\text{year}) + \\ & + S_2^c(\text{temp}_0, 6) + S_3^c(\text{temp}_{1-3}, 6) + S_4^c(\text{dew}_0, 3) \\ & + S_5^c(\text{dew}_{1-3}, 3) \\ & + \text{intercept for age group } a \\ & + \text{separate smooth functions of time} \\ & (8 \text{ df}) \text{ for age group } a. \end{aligned} \qquad (1)$$

where DOW are indicator variables for day of week. Samet *et al.* (1995, 1997), Kelsall *et al.* (1997) and

Dominici *et al.* (1999) give additional details about choices of functions used to control for longer-term trends and weather. The estimates of the coefficients at lag 1 and their 95% confidence intervals for PM_{10} are shown in Figure 1 (top left). Cities are presented in decreasing order by the size of their populations.

4. POOLING RESULTS ACROSS CITIES

In this section, we present the hierarchical regression model designed to pool the city-specific pollution relative rates across cities to obtain summary value for the 8 largest U.S. cities.

Let β^c_{PM10} be the relative rate associated with PM_{10} at city c. We consider the following hierarchical model:

$$\begin{aligned} \hat{\beta}^c_{PM10} \mid \beta^c_{PM10} &\sim N(\beta_{PM10}, v^c) \\ \beta^c_{PM10} &\sim N(\alpha_{PM10}, \sigma^2) \end{aligned} \qquad (2)$$

This model specification implies independence between the relative rates of city c and c'. See Dominici *et al.* (1999) for additional details on hierarchical strategy for combining pollution relative rates of mortality across locations.

Inference on the parameters α_{PM10} and σ^2 represents a synthesis of the information from the 8 cities; they determine the overall level and the variability of the relative change in the rate of mortality associated with changes in the PM_{10} level on average over all the cities.

The Bayesian formulation is completed by specifying prior distributions on all the unknown hyper-parameters. We assume that the joint prior is the product of a normal distribution for α_{PM10} and an Inverse Gamma distribution for σ^2. The prior hyper-parameters have been selected to do not impose too much shrinkage of the study-specific parameters toward their overall means, while at the same time specifying a reasonable range for the unknown parameters a priori (see Dominici *et al.* (1999) for details). To approximate the posterior distribution of all the unknown parameters, we implement a Markov chain Monte Carlo algorithm with a block Gibbs Sampler (Gelfand and Smith, 1990) in which the unknowns are partitioned into the following groups: β^c_{PM10} 's, α_{PM10}, and σ^2.

5. RESULTS

Figure 1 (top right) shows the boxplots of samples from the posterior distributions of city-specific regression coefficients, β^c_{PM10}, associated with changes in PM_{10} measurements. The vertical scale can be interpreted as the percentage increase in mortality per 10 $\mu g / m^3$ increase in PM_{10}. The results are reported using the previous day's (lag 1) pollution levels to predict mortality.

The marginal posterior distribution of the overall regression effect combines and synthesizes the information from the 8 locations. Figure 1 (bottom left) shows the marginal posterior distributions of the overall pollution relative rates at the current day, one-day, and two-day lags. At the top right are summarized the posterior probabilities that the overall effects are larger than zero for each lag-specification.

In addition to univariate analyses, we have also estimated the health effects of particulate matter in the presence of exposure to O_3. Figure 1 (bottom right) shows the pairwise plots of the estimated relative rates of PM_{10} versus the estimated relative rates of PM_{10} adjusted for O_3. All the relative rates are calculated with the

pollutant variables at one day lagged. The empty circle is plotted at the posterior means of the corresponding overall effects. The effect of PM_{10} is robust respect to the inclusions of O_3 in the model. Additional findings on the effects of PM_{10} on total mortality are reported elsewhere (Samet *et al.*, 1999).

6. DISCUSSION

We have developed a statistical model for obtaining an overall estimate of the effect of urban air pollution on daily mortality using data for the 8 largest US cities. The raw data comprised publicly available listings of individual deaths by day and location, and hourly measurements of pollutants and weather variables.

These analyses demonstrate that there is a consistent association of particulate air pollution PM_{10} with daily mortality across the 8 largest US cities leading to an overall effect, which is positive with high probability. While only a first step, the modeling described here establishes a basis for carrying out national surveillance for effects of air pollution and weather on public health. The analyses can be easily extended to studies of cause-specific mortality and other pollutants (Samet *et al.*, 1999). Monitoring efforts using models like the one described here are appropriate given the important public health questions that they can address and the considerable expense to government agencies for collecting the information that forms the basis for this work.

ACKNOWLEDGEMENTS

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the Environmental Protection Institute (EPA R824835) and automotive manufacturers. The contents of this article do not necessarily reflect the views and policies of HEI, nor do they necessarily reflect the views and policies of EPA, or motor vehicles or engine manufacturers.

REFERENCES

American Thoracic Society, Bascom, R. (1996a). Health effects of outdoor air pollution, Part 1. *American Journal of Respiratory and Critical Care Medecine* 153, 3-50.

American Thoracic Society, Bascom, R. (1996b). Health effects of outdoor air pollution, Part 2. *American Journal of Respiratory and Critical Care Medecine* 153, 477-498.

Breslow, N., and Clayton, D. (1993). Approximation inference in generalized linear mixed models. *Journal of the American Statistical Association* 88, 9-25.

Dockery, D., and Pope, C. (1994). Acute respiratory effects of particulate air pollution. *Annual Review of Public Health* 15, 107-132.

Dominici, F. (1999). Combining contingency tables with missing dimensions. *Biometrics* (to appear).

Dominici, F., Samet, J., and Zeger, S. (1999). Combining evidence on air pollution and daily mortality from the twenty largest US cities: A hierarchical modeling strategy. *Royal Statistical Society*, Series C, read paper (to appear).

DuMouchel, W. (1990). *Bayesian Metaanalysis*. Marcel Dekker.

DuMouchel, W. H., and Harris, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (c/r: P308-315). *Journal of the American Statistical Association* 78, 293-308.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85(410), 398-409.

Hastie, T. J., and Tibshirani, R. J. (1990). *Generalized additive models*. Chapman and Hall, New York.

Kelsall, J., Samet, J., and Zeger, S. (1997). Air pollution, and mortality in Philadelphia, 1974-1988. *American Journal of Epidemiology* 146, 750-762.

Korrick, S., Neas, L., Dockery, D., and Gold, D. (1998). Effects of ozone and other pollutants on the pulmonary function of adult hikers. *Environmental Health Perspectives* 106.

Li,Y., and Roth, H. (1995). Daily mortality analysis by using different regression models in Philadelphia county, 1973-1990. *Inhalation Toxicology* 7, 45-58.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73,13-22.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society*, Series B, Methodological 34,1-41.

Lipfert, F., and Wyzga, R. (1993). Air pollution and mortality: Issues and uncertainty. *Journal Air Waste Manage. Assoc* 45, 949-966.

Morris, C. N., and Normand, S.-L. (1992). Hierarchical, models for combining information and for meta-analysis. In: J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), *Bayesian Statistics* 4, 321-344, Oxford. Oxford University Press.

Samet, J., Dominici, F., Xu, J., and Zeger, S. (1999). *Particulate Air Pollution and Mortality: Findings from 20 U.S. Cities*. Technical report, Johns Hopkins University.

Samet, J., Zeger, S., and Berhane, K. (1995). *The Association of Mortality and Particulate Air Pollution*. Health Effect Institute, Cambridge, MA.

Samet, J., Zeger, S., Kelsall, J., Xu, J., and Kalkstein, L. (1997). *Air pollution, weather and mortality in Philadelphia, In Particulate Air Pollution and Daily Mortality: Analyses of the Effects of Weather and Multiple Air Pollutants. The Phase IB report of the Particle Epidemiology Evaluation Project*. Health Effect Institute, Cambridge, MA.

Schwartz, J. (1995). Air pollution and daily mortality in Birmingham,, Alabama. *American Journal of Epidemiology* 137, 1136-1147.

FIGURE CAPTIONS

Figure 1. Top left Results of regression models for the eight cities: $\hat{\beta}^c_{PM10}$, and 95 % confidence intervals of $\hat{\beta}^c_{PM10} \times 1000$ for PM_{10}. Cities are presented in decreasing order by population living within their county limits. The vertical scale can be interpreted as the percentage increase in mortality per 10 $\mu g / m^3$ increase in PM_{10}. The results are reported, using the previous day's (lag 1) pollution levels.

Figure 1. Top right Marginal posterior distributions of the overall effects, α_{PM10} for different lags. At the top right are specified the posterior probabilities that the overall effects are larger than zero.

Figure 1. Bottom left Boxplots of samples from the posterior distributions of city-specific regression coefficients, β^c_{PM10} associated with changes in PM_{10} measurements. The results are reported using, using the previous day's (lag 1) pollution levels.

Figure 1. Bottom right Pairwise plots of the estimated relative rates of PM_{10} versus the estimated relative rates of PM_{10} adjusted for: O_3. All the relative rates are calculated with the pollutant variables at one day lagged. The empty circle is plotted at the posterior means of the corresponding overall effect.

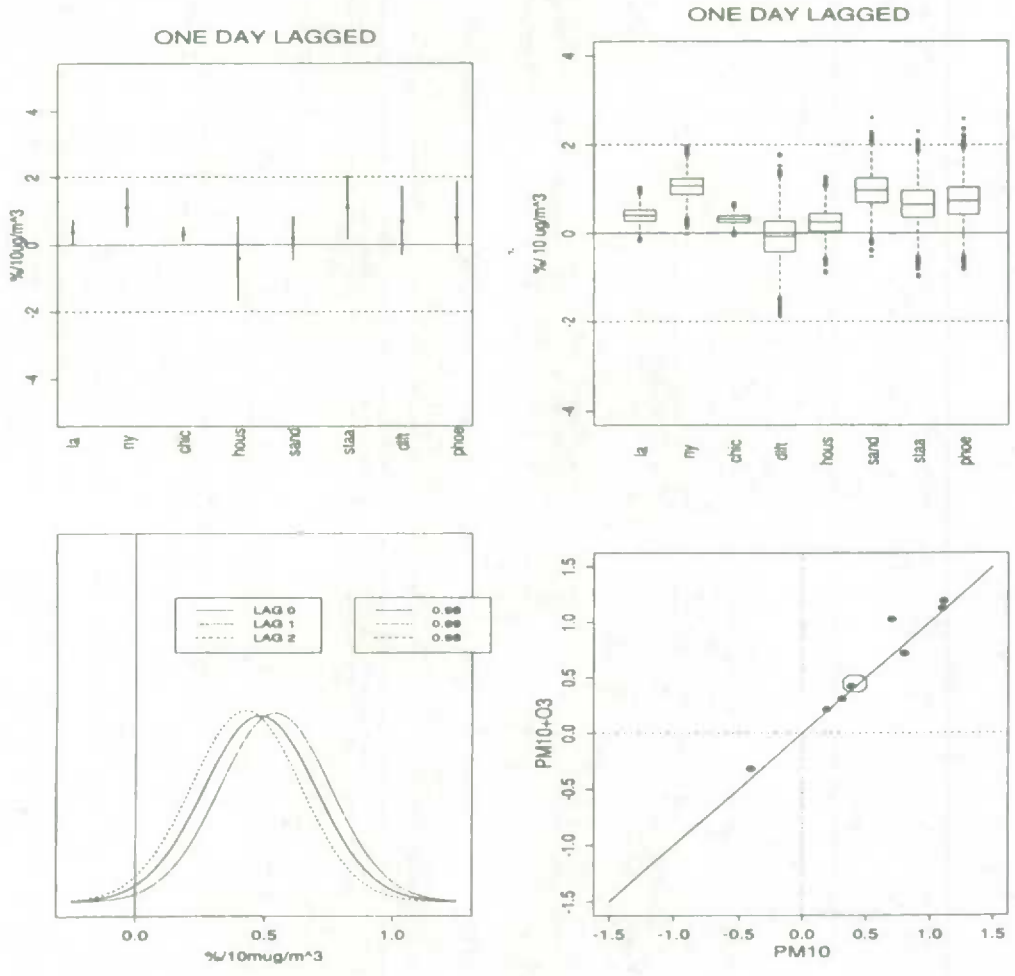


Figure 1:

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

UNCERTAINTIES IN ESTIMATES OF RADON LUNG CANCER RISKS

S.N. Rai, Zielinski, J.M. and D. Krewski[1]

ABSTRACT

Radon, a naturally occurring gas found at some level in most homes, is an established risk factor for human lung cancer. The U.S. National Research Council (1999) has recently completed a comprehensive evaluation of the health risks of residential exposure to radon, and developed models for projecting radon lung cancer risks in the general population. This analysis suggests that radon may play a role in the etiology of 10-15% of all lung cancer cases in the United States, although these estimates are subject to considerable uncertainty. In this article, we present a partial analysis of uncertainty and variability in estimates of lung cancer risk due to residential exposure to radon in the United States using a general framework for the analysis of uncertainty and variability that we have developed previously. Specifically, we focus on estimates of the age-specific excess relative risk (ERR) and lifetime relative risk (LRR), both of which vary substantially among individuals.

KEY WORDS: Attributable risk; excess relative risk; lifetime relative risk; lognormal distribution; Monte Carlo simulation; multiplicative risk model; radon progeny; uncertainty; variability.

1. INTRODUCTION

Radon, an inert gas naturally present in rocks and soils, is formed during the radioactive decay of uranium-238 (Nation Research Council, 1999). Radioactive decay products, or radon daughters, emit alpha particles, which can damage intracellular DNA and result in adverse health outcomes. In particular, underground miners exposed to high levels of radon in the past have been shown to be at excess risk of lung cancer, raising concerns about potential lung cancer risks due to the presence of lower levels of radon in homes (Letourneau et al., 1994; Pershagen et al., 1994). The U.S. Environmental Protection Agency (1992) estimated that 7,000 - 30,000 cases of lung cancer in the United States might be attributable to residential radon exposures each year, and promoted voluntary testing of homes and exposure mitigation whenever the Agency's guideline of 4 pCi/L was exceeded. These estimates are subject to considerable uncertainty, which can be described using current methods for uncertainty analysis (National Research Council, 1999).

There are many sources of uncertainty and variability in health risk assessment (Bartlett *et al.*, 1996). Many epidemiological investigations are based on occupational groups with higher levels of exposure than the general population, requiring extrapolation from occupational to environmental exposure conditions. For example, lung cancer risks experienced by underground miners exposed to high levels of radon gas in the past may be used to predict the potential risks associated with lower levels of radon present in homes (Lubin, 1994). Retrospective exposure profiles can be difficult to construct, particularly with chronic diseases such as cancer for which exposure data many years prior to the onset of disease are needed. Radon measurements in homes taken today may not reflect past exposures because of changes in dwelling address, recent building renovations, changes in lifestyle such as sleeping with the bedroom window open or closed, or inherent variability in radon measurements.

[1] Shesh N. Rai, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.M. Zielinski, Environmental Health Directorate, Health Canada, Ottawa, Canada; D. Krewski, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada

It is important to distinguish clearly between uncertainty and variability in risk assessment. Uncertainty represents the degree of ignorance about the precise value of a particular variable, such as the body weight of a given individual (Bogen, 1995; National Research Council, 1994; Hattis and Burmaster, 1994). In this case, uncertainty may be due to systematic or random error associated with simple scale or more sophisticated mass balance used to measure the body weight. On the other hand, variability represents inherent inter-individual variation in the value of a particular parameter within the population of interest. In addition to being uncertain, body weight also varies among individuals.

In this paper, we present a partial analysis of uncertainty and variability in cancer risk due to residential radon exposures using the general methods developed by Rai *et al.* (1996) and Rai and Krewski (1998). We briefly describe our measures of uncertainty and variability in the special case of multiplicative risk models in section 2. In the penultimate section, we formulate models for the age specific excess relative risk (ERR), the lifetime relative risk (LRR) and the population attributable risk (PAR), for residential radon exposure, and we apply our analysis of uncertainty in the ERR and LRR. The implications of these results for radon risk management are discussed briefly in section 4.

2. UNCERTAINTY ANALYSIS

2.1 Multiplicative Risk Models

Rai *et al.* (1996) developed a general framework for characterizing and analyzing uncertainty and variability for arbitrary risk models. In this section, we briefly outline the framework for the analysis of uncertainty and variability in the special case of multiplicative risk models described previously by Rai and Krewski (1998). Specifically, suppose that the risk R is defined as the product

$$R = X_1 \times X_2 \times \ldots \times X_p \quad (1)$$

of p risk factors $X_1, \ldots, X_p$. Each risk factor X_i may vary within the population of interest according to some distribution with probability density function $f_i(X_i \mid \theta_i)$, conditional upon the parameter θ_i. Uncertainty in X_i is characterized by a distribution $g_i(\theta_i \mid \theta_i^0)$ for θ_i, where θ_i^0 is a known constant.

If θ_i is a known constant and the distribution f_i is not concentrated at a single point, X_i exhibits variability only. On the other hand, X_i is subject to both uncertainty and variability if both θ_i and f_i are stochastic. When f_i is concentrated at a single point θ_i, and θ_i is stochastic, X_i is subject to uncertainty but not variability. Consequently, the variables $X_1, \ldots, X_p$ can be partitioned into three groups: variables subject to uncertainty only, variables subject to variability only, and variables subject to both uncertainty and variability.

2.2 Measures of Uncertainty and Variability

After logarithmic transformation, the multiplicative model (1) can be re-expressed as an additive model

$$R^* = \log(R) = \sum_{i=1}^{p} X_i^*, \quad (2)$$

where $X_i^* = \log X_i$. The expected risk on a logarithmic scale is $E(R^*)$, with variance $Var(R^*)$. Although the distribution of R^* is of interest, calculation of the first two moments of this distribution requires only mean $E(X_i^*)$ and covariance $Cov(X_i^*, X_j^*)$.

Let $W(R^*) = Var(R^*)$ be the total uncertainty/variability in R^*. This variance can be decomposed into two components: $W(R^*) = Var(E(R^* \mid X_i)) + E(Var(R^* \mid X_i))$. We define $W_i(R^*) = Var(E(R^* \mid X_i))$ as the

uncertainty in R^* due to X_i. Note that when X_i is neither uncertain nor variable, $W_i(R^*)$ is zero. With this definition, $W_i(R^*)$ effectively represents the uncertainty/variability in $E(R^* \mid X_i)$. If $X_1, \cdots, X_p$ are independent, $W_i(R^*) = Var(X_i^*)$.

We now introduce a measure $U(R^*)$ of uncertainty in R^* due to uncertainty in all of the risk factors combined. Similarly, we define a measure $V(R^*)$ of variability in R^* which encompasses variability in all the risk factors. One further decomposition of $U(R^*)$ and $V(R^*)$ is needed to determine the degree to which variables subject to both uncertainty and variability contribute to the uncertainty in X_i separately from the variability in X_i. Let $U_i(R^*)$ and $V_i(R^*)$ denote the contribution of X_i to the total uncertainty in R^* due to the uncertainty and variability in X_i, respectively. When the risk factors are independent, the total uncertainty/variability in R^* due to X_i partitions into two components: $W_i(R^*) = U_i(R^*) + V_i(R^*)$, the first due to uncertainty in X_i and the second due to variability in X_i. Explicit expressions for these measures are given in Rai and Krewski (1998).

2.3 Risk Distributions

In order to provide the fullest possible characterization of risk, consider the distribution of R^*. Allowing for both uncertainty and variability in risk, the distribution of R^* is given by

$$H^*(R^* \le r^*) = \int_{x_1^* + \cdots + x_p^* \le r^*} \int h(X_1, \cdots, X_p \mid \theta^0) dX_p \cdots dX_1$$

$$= \int_{x_1^* + \cdots + x_p^* \le r^*} \int f(X_1, \ldots, X_p \mid \theta^0) g(\theta \mid \theta^0) d\theta \, dX_p \cdots dX_1 \quad (3)$$

where $h(.)$ is a joint density function of all the risk factors; the parameter of this joint distribution is θ^0. This joint density function $h(. \mid \theta^0)$ can be partitioned into two parts: $f(. \mid \theta)$ represents the joint density function due to variability in the risk factors and $g(. \mid \theta^0)$ represents the joint density function due to uncertainty in the risk factors. The distribution of R^* has mean $E(R^*)$ and variance $W(R^*)$.

If all of the risk factors in (1) are lognormally distributed, the distribution of R^* is normal with mean $E(R^*)$ and variance $W(R^*)$. If all of the risk factors are approximately lognormally distributed, the distribution of R^* can be approximated by a normal distribution. If one or more of the risk factors is not well approximated by a lognormal distribution, the distribution of R^* can be approximated by Monte Carlo simulation. Although straightforward, Monte Carlo simulation can become computationally intensive with a moderate number of risk factors.

In addition to examining the distribution of R^* taking into account both uncertainty and variability in the $\{X_i\}$, it is of interest to examine the distribution of R^* considering only uncertainty in the $\{X_i\}$ or only variability in the $\{X_i\}$. By comparing the distributions of R^* allowing both uncertainty and variability, only variability, and only uncertainty, it is possible to guage the relative contribution of uncertainty and variability to the overall uncertainty/variability in risk. The density of R^* based on only uncertainty in the $\{X_i\}$ is the density of $\sum_{i=1}^{p} E_{f_i}(X_i^*)$, where $E_{f_i}(X_i^*) = \int_{-\infty}^{\infty} X_i^* f_i(X_i \mid \theta_i) dX_i$. This density can be approximated by a normal distribution with mean $E(R^*)$ and variance $U(R^*)$. The density of R^* based on

only variability in the $\{X_i\}$ is the density of $\sum_{i=1}^{p} \tilde{X}_i^*$, where $\tilde{X}_i^*$ has distribution f_i with known parameter $E_{g_i}(\theta_i) = \int_{-\infty}^{\infty} \theta_i \, g_i(\theta_i \mid \theta_i^0) d\theta_i$. This density can also be approximated by a normal distribution with mean $E(R^*)$ and variance $V(R^*)$.

3. CANCER RISKS MODELS FOR RESIDENTIAL RADON EXPOSURE

3.1 Age Specific Excess Relative Risk

The U.S. National Research Council (1999) recently conducted a comprehensive review of the potential health effects of residential exposure to radon. Of primary concern is the excess lung cancer risk demonstrated in 11 cohort studies of underground miners conducted in a number of countries around the world. Since radon appears to exert its carcinogenic effects through DNA damage to lung tissue caused by alpha particles emitted by radon daughters, it is thought that even low levels of exposure to radon confer some increase in risk. After considering different possible approaches to risk estimation, the BEIR VI Committee elected to base in risk models on epidemiological data derived from studies of the mortality experience of miners. The committee conducted a combined analysis of updated data from the 11 miner cohorts, and developed two models for projecting lung cancer risks in the general population. These two models are referred to as the exposure-age-concentration and exposure-age duration models, based on the major risk factors included in the models. We consider the exposure-age-duration model for demonstrating the uncertainty analysis.

If e_t denotes the excess relative risk at age t, the exposure-age-duration is expressed as

$$e_t = \beta \times \omega(t) \times \phi(t) \times \gamma_{dur}(t) \times K \quad .(4)$$

The factor β ([Bq/m^3]$^{-1}$) reflects the carcinogenic potency of radon, as modified by the other risk factors in model (4). The last term in these models is the dosimetric K-factor (dimensionless), used to extrapolate from occupational to environmental exposure conditions. The factor $\omega(t)$ represents a time-weighted average exposure to radon, expressed in Bq/m^3, within exposure-time windows 5-14, 15-25, and 25+ years prior to disease diagnosis. The factor $\omega(t) = \omega \times \eta_t$, where

$$\eta_t = \omega \times [\Delta_{[5,14]}(t) + \vartheta_2 \Delta_{[15,24]}(t) + \vartheta_3 \Delta_{[25,\infty]}(t)] \text{ with } \Delta_{(a,b]}(t) = \begin{cases} 10 & for \quad t > b \\ (t-a) & for \quad a \le t \le b \\ 0 & otherwise. \end{cases}$$

The factor $\phi(t)$ (y^{-1}) indicates the effects of attained age and the factor $\gamma_{dur}(t)$ (y^{-1}) reflects the effects of duration of exposure to radon (in years); these factors are categorized into broad groups:

$$\phi(t) = \begin{cases} \phi_1 & for \quad t \le 54 \\ \phi_2 & for \quad 55 \le t \le 64 \\ \phi_3 & for \quad 65 \le t \le 74 \\ \phi_4 & for \quad t \ge 75. \end{cases}$$

and

$$\gamma_{dur}(t) = \begin{cases} \gamma_{d1} & for \quad t < 10 \\ \gamma_{d2} & for \quad 10 \le t < 20 \\ \gamma_{d3} & for \quad 20 \le t < 30 \\ \gamma_{d4} & for \quad 30 \le t < 40 \\ \gamma_{d5} & for \quad t \ge 40. \end{cases}$$

Our interest is in risks associated with low (residential) radon exposures and for mid to old age population. We demonstrate the uncertainty analysis only for any person of age more than 40 years.

The risk function (4) represents the age specific excess relative risk (ERR) model which is represented as a product of five factors: the carcinogenic potency of radon $\beta = X_1$, the exposure to radon $\omega\eta_i = X_2$, the effect of age $\phi = X_3$, $K = X_4$ and $\gamma_{d5} = \gamma$. Thus,

$$ERR = X_1 \times X_2 \times X_3 \times X_4 \times \gamma, (5)$$

For simplicity, we assume that γ is constant (=10.18), exhibiting neither uncertainty nor variability. Note that the risk factors X_1, X_2 and X_3 are correlated. In this application, X_1, and X_3 are assumed to be subject to uncertainty only, and X_2 and X_4 are assumed to be subject to both variability and uncertainty. For a detailed description of uncertainty and variability distributions in these risk factors, see Tables A-8b and A-9b in the National Research Council (1999).

Our analysis allows for correlation among the risk factors in the multiplicative model (5). The correlation structure is based on the covariance matrix obtained when estimating the model parameters. Other than statistical correlation between X_1 and X_3 induced due to the estimation procedure, the risk factors are assumed to be independent.

Table 1

Components of uncertainty and variability in the ERR (x100)

Risk Factor	Uncertainty			Variability			Both
	$\frac{U_i}{W}$	$\frac{U_i}{U}$	$\frac{U_i}{W_i}$	$\frac{V_i}{W}$	$\frac{V_i}{V}$	$\frac{V_i}{W_i}$	$\frac{W_i}{W}$
I. 40 ≤ Age ≤ 54 years							
Potency	5.6	64.3	100	0	0	0	5.6
Exposure	0.8	8.7	1.0	77.2	84.6	99.0	77.9
Age	0.5	6.2	100	0	0	0	0.5
K-factor	1.8	20.8	11.5	14.1	15.4	88.5	15.9
All	8.8	100	NA	91.2	100	NA	100
II. 55≤ Age ≤64 years							
Potency	3.9	39.6	100	0	0	0	3.9
Exposure	0.8	10.3	1.0	78.2	84.6	99.0	79.0
Age	3.4	34.3	100	0	0	0	3.4
K-factor	1.8	18.4	11.5	13.9	15.4	88.5	15.7
All	9.8	100	NA	90.2	100	NA	100
III. 65≤ Age ≤74 years							
Potency	3.6	29.2	100	0	0	0	3.6
Exposure	0.7	6.0	1.0	74.2	84.6	99.0	74.9
Age	6.2	50.5	100	0	0	0	6.2
K-factor	1.8	14.3	11.5	13.5	15.4	88.5	15.3
All	12.3	100	NA	87.7	100	NA	100
IV. Age ≥75 years							
Potency	2.4	6.1	100	0	0	0	2.4
Exposure	0.5	1.3	1.0	51.5	84.6	99.0	52.1
Age	34.9	89.4	100	0	0	0	34.9
K-factor	1.2	3.1	11.5	9.4	15.4	88.5	10.6
All	39.1	100	NA	60.9	100	NA	100

NA: Not applicable

Measures of uncertainty and variability in the *ERR* are given in Table 1. Several conclusions can be drawn from this table. A comparison of columns 3 and 6 in the case of all risk factors (X_1, ..., X_4) indicates that the variability tends to account for the majority of the total uncertainty and variability in the *ERR*. Whereas potency and age tend to be more uncertain than variable, exposure to radon and the *K*-factor are more variable than uncertain. The last column in the table indicates that exposure is the most influential variable overall (contributing most to uncertainty and variability), followed by the *K*-factor, potency, and then age in all but the last age group, where age is the most influential factor.

The distributions of the *ERR* can be obtained using either the lognormal or Monte Carlo approximations described in section 2.3. We find that these two approximations are in close agreement regardless of whether uncertainty and variability, only uncertainty, or only variability in the *ERR* is considered. Preliminary results indicate that, whereas uncertainty in the *ERR* for an individual of a given age precludes determining the *ERR* to less than a 10-fold range, variability in *ERR* exceeds 100-fold.

3.2 Lifetime Relative Risk

To determine the lifetime relative risk of radon-induced lung cancer, we require the hazard function for the death rate in the general population. Deaths can be classified into two categories: those due to lung cancer and due to other competing causes. For simplicity, we assume that the death times are recorded in years and identify the range of T as {1,2,...,110,...}. Let h_t and h_t^* be the lung cancer and overall mortality rates for age group t respectively in an unexposed population, and e_t be the excess relative risk for lung cancer mortality in an exposed population for age group t. Then, the death rates in the exposed population are given by $h_t + h_t e_t$ for lung cancer and $h_t^* + h_t e_t$ for all causes including lung cancer.

Let R_t be the probability of death due to lung cancer for a person of age t, and S_t be the survival probability up to age t in the exposed population. Following Kalbfleisch and Prentice (1980), the survival function can be expressed as

$$S_t = \Pr\{T \geq t\} = \prod_{i=1}^{t-1} \{1-(h_i^* + h_i e_i)\}$$

$$\approx \exp\{-\sum_{i=1}^{t-1}(h_i^* + h_i e_i)\}$$

Table 2
Quantiles of the Distribution of the LRR

Distribution	Min	Quantiles									Max
		2.5*	5	10	25	50	75	90	95	97.5**	
Uncertainty	1.029	1.014	1.045	1.048	1.055	1.065	1.078	1.094	1.107	1.123	1.365
Variability	1.001	1.006	1.008	1.013	1.028	1.062	1.144	1.296	1.470	1.673	6.895
Both	1.001	1.005	1.008	1.013	1.027	1.067	1.161	1.353	1.578	1.829	6.679

* Lower 95% limit; ** Upper 95% limit.

and the probability of death due to lung cancer, R_t can be written as

$$R_t = \Pr\{T = t\} = (h_t + h_t e_t)S_t \approx (h_t + h_t e_t)\exp\{-\sum_{i=1}^{t-1}(h_i^* + h_i e_i)\} \qquad (6)$$

The approximation in (6) is highly accurate and will be treated as exact in what follows. This differs slightly from the expression for R_t used by BEIR VI committee (National Research Council, 1999). Our

expression for R_i is not only more accurate, but considerably simplifies the calculation of the lifetime relative risk and the population attributable risk.

Assuming a maximum lifespan 110 years, the lifetime lung cancer risk is given by the sum of the annual risks:

$$R = \sum_{i=1}^{110} R_i = \sum_{i=1}^{110} (h_i + h_i e_i) \exp\left(\sum_{l=1}^{i-1} (h_l + h_l e_l)\right) \quad (7)$$

In the unexposed population, the $\{e_i\}$ are assumed to be zero, so that $R_0 = \sum_{i=1}^{110} h_i \exp\{-\sum_{l=1}^{i-1} h_l^*\}$, where R_0 is the lifetime risk in an unexposed population. Note that the evaluation the lifetime risk R in (7) depends on the excess relative lung cancer risks $\{e_i\}$ within each of the age groups.

The lifetime relative risk (*LRR*) is the ratio of the lifetime risk in the exposed population relation to that in the unexposed population:

$$LRR = R/R_0. \quad (8)$$

Like the *ERR*, the *LRR* is subject to both uncertainty and variability. Although the *LRR* is summed over all age groups, the uncertainty in the modifying effects of each age group in the sum is taken into account in our analysis.

Since the *LRR* is not of the multiplicative form, however, the risk distributions for uncertainty/variability, uncertainty only and variability only are obtained by Monte Carlo methods.

The results of uncertainty analysis in the *LRR* are given in Table 2. These results indicate that the variability in the *LRR* (due to interindividual variation in the levels of radon exposure and the dosimetric *K*-factor) is much greater than uncertainty in the *LRR* (due to the uncertainty in the model parameters, the levels radon exposure and the *K*-factor).

3.3 Population Attributable Risk

The attributable risk of lung cancer mortality due to exposure to radon is defined as the excess lung cancer risk in a population due to exposure as a fraction of total lung cancer risk. Thus, the population attributable risk (*PAR*) is given by

$$PAR = \frac{E(R) - R_0}{E(R)}. \quad (9)$$

Here, $E(R)$ is the average value of R, the lifetime risk of death due to lung cancer across individuals in the exposed population. The quantity R_0 is the lifetime risk in an unexposed population.
As noted previously, the inter-individual variability in the *LRR* is due to variability in radon exposure and the dosimetric K-factor. In order to evaluate $E(R)$ we need to evaluate $E(R_i)$. For evaluating $E(R_i)$, we need to know the distribution of $b = K\ \omega$ in model (4). Suppose b has a distribution $f_b(b)$. After some algebraic manipulation, it can be shown that

$$E(R_i) = h_i e^{-\sum_{l=1}^{i-1} h_l^*} \left\{ E(e^{-bm_i}) + \frac{m_i}{h_i} E(be^{-bm_i}) \right\}, \quad (10)$$

where $m_i = h_i\phi_i\eta_i\gamma_i\beta$ with $n_i = \sum_{i=1}^{i-1} m_i$. The expectation in (10) is then with respect to the distribution f_b(b). Further simplification of (10) depends on the distributions of ω and K. Consider first the most general case in which the distributions for these two risk factors are arbitrary. If these two risk factors are statistically independent, it is straightforward to find $E(e^{-bn_i})$ and $E(be^{-bn_i})$, the former being the moment generating function at all values of (-b) for which the expected value exists. Note that $E(be^{-bn_i}) = dE(e^{-bn_i})/db$ for the exponential family of distributions of b. Secondly, if these two factors are independent and log-normally distributed, computation of the moments is simplified since the product of two log-normally distributed variables has a lognormal distribution; which is the case in our example. To evaluate $E(b^l)$ in (10), one can use the fact that *log(b)* has a normal distribution with mean μ and standard deviation σ. After some algebraic manipulation, it can be shown that $E(b^l)=\exp(l\mu+\frac{1}{2}l^2\sigma^2)$.

Due to space limitation, a discussion of uncertainties in the *PAR* is not possible; however, some discussion can be found in Krewski *et al.* (1999).

4. DISCUSSION

In this article, we have applied new methods for the analysis of uncertainty and variability to evaluate the lung cancer risks due the presence of radon gas in homes. This important population health issue was recently examined by the National Research Council (1999). The NRC evaluation included not only an assessment of the most likely risk estimates, but also an analysis of the uncertainties in such estimates. Such stochastic analyses take into account both uncertainty and variability in the risk factors affecting risk. In our analyses, each risk factor is assumed to follow a distribution with one or more parameters reflecting variability within the population of interest; uncertainty in the values of each risk factor is characterized by an appropriate distribution for the parameter values. Within this framework, overall measures of uncertainty and variability and relative contributions of individual risk factors to overall uncertainty and variability in lung cancer risk due to radon are computed. Influential factors that contribute most to uncertainty and variability may then be targeted for further study.

The results of this analysis of uncertainty and variability are informative in several ways. Estimates of the age-specific excess relative risk of lung cancer are highly variable among individuals, largely due to the substantial variability in radon levels in U.S. homes. Indeed, radon exposure appears to be the most influential factor affecting individual risk of the four factors included in the BEIR VI risk models. Although uncertainty in these factors leads to about a 10-fold range in risk, variability in the ERR exceeds 100-fold. Our analysis indicates that like the *ERR*, the *LRR* is more variable than uncertain.

Despite their limitations, these initial attempts at applying general methods for the analysis of uncertainty and variability in individual and population risks have proven useful in evaluating the reliability of radon risk estimates. These results may be refined as more information about the factors affecting radon related lung cancer risk becomes available, including the extent of uncertainty and variability in these factors.

REFERENCES

Bartlett, S., Richardson, G.M., Krewski, D., Rai, S.N. and Fyfe, M. (1996). Characterizing uncertainty in risk assessment - conclusions drawn from a workshop. *Human and Ecological Risk Assessment*, 2, 217-227.

Bogen, K.T. (1995). Methods to approximate joint uncertainty and variability in risk. *Risk Analysis*, 15, 411-419.

Hattis, D. and Burmaster, D.E. (1994). Assessment of variability and uncertainty distributions for practical risk analysis. *Risk Analysis*, 14, 713-730.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Krewski, D., Rai, S.N., Zielinski, J.M., Hopke, P.K. (1999). Characterization of uncertainty and variability in residential radon cancer risks. *Annals of the New York Academy of Sciences*. To appear.

Letourneau, E.G., Krewski, D., Choi, N.W., Goddard, M.J., McGregor, R.G., Zielinski, J.M. and Du, J. (1994). Case-control study of lung cancer and residential radon exposure in Winnipeg, Manitoba, Canada. *American Journal of Epidemiology*, 140, 310-322.

Lubin, J.H. (1994). Invited commentary: lung cancer and exposure to residential radon. *American Journal of Epidemiology*, 140, 223-232

National Research Council. (1999). *Health Effects of Exposure to Low Levels of Ionizing Radiation BEIR VI*. Washington DC: National Academy Press.

National Research Council. (1994). *Science and Judgment in Risk Assessment*. Washington DC: National Academy Press.

Pershangen, G., Akerblom, G., Axelson, O., Clavensjo, B., Damber, L., Desai, G., Enflo., A., Lagarde, F., Mellander, H., Svartnengren, M. and Swedjemark, G.A. (1994). Residential radon exposure and lung cancer in Sweden. *New England Journal of Medicine*, 330, 159-164.

Rai, S.N., Krewski, D., and Bartlett, S. (1996). A general framework for the analysis of uncertainty and variability in risk assessment. *Human and Ecological Risk Assessment*, 2, 972-989.

Rai, S.N. and Krewski, D. (1998). Uncertainty and variability analysis in multiplicative risk models. *Risk Analysis*, 18, 37-45.

U.S Environmental Protection Agency. Technical support document for the 1992 citizen's guide to radon. *EPA* 1992; 400-R-92-011.

SESSION IV

RECORD LINKAGE

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

OVERVIEW OF RECORD LINKAGE

Julie Bernier and Karla Nobrega[1]

ABSTRACT

There are many different situations in which one or more files need to be linked. With one file the purpose of the linkage would be to locate duplicates within the file. When there are two files, the linkage is done to identify the units that are the same on both files and thus create matched pairs. Often records that need to be linked do not have a unique identifier. Hierarchical record linkage, probabilistic record linkage and statistical matching are three methods that can be used when there is no unique identifier on the files that need to be linked. We describe the major differences between the methods. We consider how to choose variables to link, how to prepare files for linkage and how the links are identified. As well, we review tips and tricks used when linking files. Two examples, the probabilistic record linkage used in the reverse record check and the hierarchical record linkage of the Business Number (BN) master file to the Statistical Universe File (SUF) of unincorporated tax filers (T1) will be illustrated.

KEY WORDS: Record Linkage; Exact Matching; Statistical Matching; Hierarchical Exact Matching.

1. INTRODUCTION

Record linkage is the process in which records or units from different sources are combined into a single file. There are two categories of record linkage: exact matching and statistical matching. The purpose of an exact match is to link information about a particular record on one file to information on a secondary file, thus creating a single file with correct information at the level of the record. This is achieved by using deterministic or probabilistic record linkage procedures. However, the purpose of a statistical match is to create a file reflecting the correct population distribution, records that are combined do not necessarily correspond to the same physical entity (e.g. a person, business or farm). Statistical matching involves modeling and imputation techniques to complete a file (Armstrong, 1989).

1.1 Selecting the Linkage Procedure

The type of record linkage implemented depends on the purpose of the linkage, type of data, cost, time, confidentiality, acceptable level and type of error, and response burden (FCSM, 1980). To determine the method to use, first assess the purpose of the linkage. For example, in linking a vital statistics file with a survey you would want the data linked at the record level. On the other hand, if what was of interest is the relationship between variables it is not necessary (or often possible) to link two files exactly. In this case statistical matching can be used to create a single file.

Along with the purpose, the type of data in each file is an important factor when deciding on a linkage method. Ideally each data set would have a unique identifier or 'key', to link efficiently several sources of data. However, it is often the case that there is no unique identifier, that not all records contain a unique identifier or that the unique identifier is subject to error. If any of these are the case a method must be employed that will compensate for the specific deficiencies in the file. If the purpose is to get an exact match at the record level, then the two files must contain the same records or individuals. If the relationship between variables is of interest, the same types of data must be available on the two files, but the same individuals or records need not be on both files. With statistical matching one assumes the relationship between variables on the file used for modeling or imputation represents the 'true relationship' in the population. With either exact or statistical matching the two files must have variables in common.

1 Julie Bernier and Karla Nobrega, Statistics Canada, Ottawa Ontario

Cost and time are also important considerations that need to be weighed against acceptable error rates and response burden. Secondary sources of data, for example administrative, census or other survey data offer a wealth of information that need not be duplicated, and so, record linkage offers the possibility of reducing response burden and cost. The time and cost are also dependent on the knowledge and software needed for each type of matching. Since statistical matching covers a wide range of techniques it can be implemented using standard statistical computer software or specialized software. Deterministic exact linkage will only need a standard software program, whereas probabilistic linkage will require a specialized software package. Probabilistic linkage and statistical matching require specific methodological knowledge, whereas deterministic matching does not.

The type of error and the ability of the user to quantify the error depend upon the type of record linkage. With an exact match there are two types of errors, false pairs and missed pairs. False pairs are created when one links information related to two different units or records. Missing a true pair is also an error. These are analogous with type I and type II errors in tests of statistical hypothesis. With statistical matching an error is defined not at the level of a record but in terms of distribution. If the joint distribution of the variables is incorrect after the match then this is an error (Kovacevic, 1998). The technique implemented may not be the most 'efficient' method based on any single criteria listed above, but overall should be the 'optimum' choice.

2. EXACT MATCHING

An exact match is when files are linked at the level of the record (FCSM, 1980). Exact matching can be either deterministic or probabilistic. Generally, when doing a two-file exact linkage, one of the files is much smaller than the other file, for example, a link of mortality records to the Census or of a business sample to tax records.

2.1 Hierarchical Exact Matching

Hierarchical exact matching is a deterministic exact matching technique. It is often the case in a linkage project that no single variable exists that is essentially free of errors, present on the majority of data and discriminating. Often, only a combination of variables will discriminate between two records. It is possible that the 'best' combination of variables in terms of error rates and discriminating power is not available on the whole file, or that there are several possible combinations of variables that will equally discriminate between records. In this case, the files can be linked by a series of exact matches done in a hierarchy.

There are five main steps for a hierarchical exact match:
1. Identify the matching variables.
2. Pre-process the files.
3. Determine the hierarchy.
4. Match the files.
5. Review the linkages.

First decide on which variable(s) to use for matching. It is desirable to have a variable(s) that is as free of errors as possible since a deterministic exact matching procedure cannot compensate for errors. As well, the variable(s) should be present on the majority of records and be as unique or discriminating as possible. For example if full name, address, postal code, date of birth and gender appear on both files, it may seem that there is more than enough information in common to link the files directly. However, after analyzing the files one may find that the postal code is the most reliable variable, the full name is only recorded as the first initial and the last name, the birth date on the file is often missing and the address has spaces and variations in spelling. In order to perform a deterministic exact match, the variables must be complete and exactly the same on both files.

Since the variables need to be exactly the same on both files and the procedure can not compensate for errors, the second step pre-processes the file. This reduces the impact of errors (i.e. spelling mistakes) and

makes the variables consistent on the files. Several packages, such as NYSIIS, are available in order to re-code names and addresses. Combinations of full name, first initial, first name, middle name and last name are created.

Next, decide on the hierarchy to be used in matching. List all of the potential variables that will be used in the match. Then, construct a series of matches starting with the subset of these variables that has the fewest errors and the highest discriminating power. The purpose is to maximize the discriminating power while minimizing the impact of missing values and errors. Using the previous example the first step could be to match using re-coded full name, address, birth date and gender. The second match could be to match on first initial and last name, address, birth year and gender. The third match could be on last name, address, and gender. The fourth could be on last name, postal code and birth year and so on.

The fourth step is to link the files; this can be done using software such as SAS or MATCH 360. Then the linkages are reviewed. It is possible, since each step in the linkage is generally not mutually exclusive of the others, that the same link may be found at more than one step. It is also possible to have a different link found at each step. When there is more than one link for a specific record on the primary file the multiple linkages need to be reviewed manually.

2.1.1 An Example of a Hierarchical Exact Match

Taxation files provide an administrative data source that is used to reduce response burden in economic surveys. Using tax data in the Unified Enterprise Survey (UES) required tax records linked to each enterprise selected in the sample. The frame used for this survey was Statistics Canada's Business Register; which is a structured list of businesses engaged in the production of goods and services in Canada. The frame includes a unique Business Number (BN) for each enterprise.

Revenue Canada collects tax data for incorporated businesses using a T2 tax form and for unincorporated businesses using a T1 personal income tax form. The incorporated businesses indicate their unique BN on the tax form, whereas unincorporated businesses do not. As well, there are other Revenue Canada sources that allow for links between businesses and their business numbers. Whenever there exists a link this information is collected on an accumulated links file. For cases without a link Armstrong and Faber (1998) developed an alternate approach.

When an enterprise did not have a link to a business number the only other variables available for linking were the address (including postal code) and the owner's name. The files were pre-processed to standardize the name and address. Twenty-four different matching passes were performed on the pre-processed files. These passes started with the most discriminating set of variables, and then the discriminating power was diminished with each subsequent pass. The last step was to review the linkages manually.

One advantage of this methodology was that it was possible to confirm links by comparing them with links that were already on the accumulated links file. Also, it was possible to determine what proportion of links at each pass conflicted with the links on the accumulated links file. Since a business can link to more than one tax record for various reasons, these conflicts are not necessarily errors. If the rate of conflict was small, less than 5%, then the links on that pass were kept. After the 11^{th} step the conflict rate jumped to over 15% and grew exponentially so only links found in the first 11 steps were included (Armstrong and Faber, 1998).

2.2 Probabilistic Record Linkage

Probabilistic record linkage is another form of exact matching. In this case one has the same individuals on the two files, but there are no or few unique identifiers. Unlike deterministic matching, probabilistic linkage can compensate if the information is incomplete and/or subject to error. Furthermore, with this method you assume that two records could be identical but not the same individual. For example it is possible to have two men named John Smith but they are not the same person.

In order to use a probabilistic record linkage there are two important theoretical assumptions. First an error in one variable is assumed to be independent of an error in the other variables and second, accidental agreement of a variable must be independent of accidental agreement in other variables (Fellegi and Sunter, 1969).

There are several steps involved in probabilistic record linkage:

1. Identify matching variables.
2. Determine rules
3. Pre-process the files.
4. Decide on the variables for the 'pockets'.
5. Calculate the initial weights.
6. Link pairs
   a. calculate the total weight
   b. calculate frequency weights (optional)
   c. decide on the threshold.
7. Review the pairs.

As with deterministic exact linkage, we first decide which variables are available for matching. The second step is creating the rules associated with each variable. Probabilistic linkage uses rules to exploit the discriminating power of variables and allows for a spectrum of agreement with variables (Newcombe et al, 1987). For example, if names are being compared it is possible to have one rule for total agreement, one for partial agreement, and one for disagreement. By having levels of agreement, the procedure can still link pairs when there are errors or missing values in some (but not all) of the fields. The rules should be built so that the agreement is ordered according to the probability of being a true pair.

The third step is to pre-process the files in order to eliminate as many differences on the files as possible. The names and addresses can be re-coded and spaces can be removed. Although the rules can be determined to compensate for errors, re-coding is still an important step.

Next the file can be split to make it more computationally efficient to do the linkage. Unlike an exact match that will only compare records that have exactly the same variables, probabilistic linkage looks at every possible combination of pairs within the two files (Newcombe, 1988). It is important for computational reasons to limit the number of possible pairs. For probabilistic linkage the file is subdivided into groups called pockets in order to decrease the number of possible pairs. Only the pairs within a matched pocket value will be assessed.

Within each pocket one calculates initial weights for each rule. Then for each rule and agreement a numerical weight is assigned, directly related to the probability of getting agreement when the record pair is a true link compared to the probability of agreement when the pair is not a true link (Newcombe, 1988). These probabilities can be estimated. To do this, link deterministically a sample from the two files. The proportion of agreement within each category can be used to estimate the true probabilities of agreement given one has a true pair. Estimating the probability of linking, given one does not have a true pair is approximately equivalent to estimating the probability of a chance agreement. For example gender will agree half the time by chance or birth month will agree 1 in 12 by chance (Armstrong, 1990).

The sixth step is to use software designed for record linkage (e.g. Statistics Canada's Generalized Record Linkage Software (GRLS) or Automatch). These programs will calculate the overall weights. The overall weight is the sum of all the weights for all the rules within a pocket. In this step one will have to decide upon thresholds. There are two thresholds, an upper and a lower one. If the overall weight of a pair is above the upper threshold then the pair is considered to be a true match. Likewise, if a pair has an overall weight below the lower threshold then the pair considered true non-match. The area between the two thresholds represent a region of uncertainty; these may or may not be true links. Pairs in this region will need to be manually reviewed.

Frequency weights are used to refine overall weights to reflect the probability of agreement within a rule. For example, a rule for agreement on a name could have, a high positive weight for total agreement, a small

positive weight for partial agreement and a negative weight for disagreement. Now consider a match on Smith compared to a match on Zarbatany. The probability of a true agreement and chance agreement are very different within this one rule (Newcombe, 1988). Frequency weights can be calculated by GRLS or Automatch and used to refine the weights.

2.2.1 An Example of Probabilistic Record Linkage

The Reverse Record Check (RRC) is used to estimate coverage errors for the Canadian Census of Population. For this project, a sample is selected from sources independent of the 1996 Census and each selected person is surveyed; in order to do so we need an updated address for the selected person or a family member. The RRC uses both deterministic and probabilistic record linkage techniques. There were many sources used for the sample; the oldest of these were the previous Census (1991), a file of persons missed by the Census, and the 1991-93 file of births. For the sample taken from these files, we did a probabilistic linkage to the older tax records in order to get more up to date personal information. The other sources were 1994-96 file of births, immigration files, and a file of non-residents. Samples from these files were linked deterministically to the 1995 tax file.

The RRC used the following variables for the probabilistic record linkage: year of birth, month of birth, day of birth, marital status, municipality, first name, second name, surname, and postal code. The file was then split using gender and region (Quebec, Ontario, East and West) and NYSIIS code of the surname was used as a pocket. Altogether 15 rules were created for these variables, each of them involving several levels of agreement.

Considering the linkage of the person or a family member as being a success, the success rate was 92% for records coming from the Census data base and 80% for those coming either from the previous Census missed file or the files of births. More work was done in order to estimate the reliability of those links. It has been estimated that the reliability is over 90% for the Western region and as high as 98% for the three other regions (Bernier, 1997).

3. STATISTICAL MATCHING

Unlike exact matching, statistical matching is done when one is interested in the underlying distributions of the whole population. The files that are being matched can have different individuals on them but one assumes that the relationship of the variables in the population will be similar to the relationship on the files (Kovacevic, 1998). Generally, the recipient file has X and Y, while the second or donor file has X and Z and we are interested in the relationship of X, Y, Z (Armstrong, 1989).

In general these are the steps used in statistical matching:

1. Pre-process the files.
2. Identify similar records.
3. Impute and/or model missing data.
4. Review the model and the distributions on the new file.

As previously noted, the files must be pre-processed. The type of pre-processing will depend on the methods used to decide similar records and the method used to impute the data on the donor file. If one is using categorical constrained matching then if X, Y or Z is continuous it will have to be categorized (Liu and Kovacevic, 1996). If one is using a hot deck method for the imputation then it is necessary to remove records that have missing values. If a regression method is used then outliers will need to be removed.

There are two methods of imputing the data onto the second file: regression and hot deck (Armstrong, 1989). With the regression method a model of the relationship between X and Z is constructed from the donor file and then it is applied (with or without an error term) to the recipient file. With the hot deck method the imputation can be random or based on some metric distance between observations.

With statistical matching there is no review or assessment of specific pairs for errors. An error occurs when the distribution of the data is not correct. The distribution of Z on the new file can be compared with the distribution of Z on the donor file. As well, the distribution of Z given X can be compared on the two files for errors. Unfortunately the joint distribution can not be assessed for errors.

4. CONCLUSION

The type of linkage procedure depends on what is available and on the intended use of the new file. Issues such as the purpose, time, type of data, cost, confidentiality, acceptable level and type of error, and response burden need to be considered when deciding what type of linkage procedure is appropriate.

In general, exact matching is less computer intensive but it can involve more manual review if more than one linkage pass is performed. Probabilistic record linkage is more computer intensive and will need specialized software. However, it will generally produce better results than exact matching. This is especially true if the variables used to match are subject to errors and there are few or no unique identifiers on the two files. Statistical matching is dissimilar to exact methods. It is not a linkage of two files with the intention of linking files to identify the same individuals. With this type of matching it is not even necessary to have the same individuals on the two files. The most important concept is the distributions of the data.

REFERENCES

Armstrong, B, and Faber, G. (1998), T1-BN Linkage Methodology, *Methodology Branch Working Paper,* Statistics Canada

Armstrong, J. (1989). An Evaluation of Statistical Matching Methods. *Methodology Branch Working Paper*, BSMD, 90-003E. Statistics Canada.

Armstrong, J.B. (1990). Weight Estimation for Record Linkage. *Methodology Branch Working Paper*, BSMD-90-015E. Statistics Canada

Bernier, J. (1997). Quantitative Evaluation of the Linkage Operation of the RRC96, *Proceedings from Workshop on Record Linkage*, Washington.

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

Liu T.P., and Kovacevic M.S. (1996). Categorically Constrained Matching. *Proceedings of the Statistical Society of Canada Annual Meeting*, Waterloo.

Kovacevic M.S. (1998). Record Linkage and Statistical Matching – How much do they differ? *Record Linkage Resource Center News,* No. 16., Statistics Canada, Internal document.

Newcombe, H. B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business.* Oxford University Press.

Newcombe, H.B., Fair, M. and Lalonde, P. (1987). Concepts and Practices that Improve Probabilistic Linkage. *Proceedings of the Statistics Canada Symposium,* Ottawa.

United States. Federal Committee on Statistical Methodology: Subcommittee on Matching Techniques. Edited by D. Radner (1980). *Report on exact and statistical matching techniques.* U.S. Dept. of Commerce, Office of Federal Statistical Policy and Standards.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

CREATING AND ENHANCING A POPULATION-BASED LINKED HEALTH DATABASE: METHODS, CHALLENGES, AND APPLICATIONS

Green, B.[1], McGrail, K., Hertzman,C., Barer, M.L., Chamberlayne, R., Sheps, S.B., Lawrence, W.J.

ABSTRACT

As the availability of both health utilization and outcome information becomes increasingly important to health care researchers and policy makers, the ability to link person-specific health data becomes a critical objective. This type of linkage of population-based administrative health databases has been realized in British Columbia. The database was created by constructing an historical file of all persons registered with the health care system, and then by probabilistically linking various program files to this 'coordinating' file. The first phase of development included the linkage of hospital discharge data, physician billing data, continuing care data, data about drug costs for the elderly, births data and deaths data. The second phase of development has seen the addition data sources external to the Ministry of Health including cancer incidence data, workers' compensation data, and income assistance data. The individual datasets each presented a unique challenge. That is, the specific fields in each dataset were in some ways limiting for linkage, but could sometimes also be used opportunistically to improve the linkage rate. We achieved a high rate of linkage of health services and other contacts with the social system to person-specific registration records. This success has allowed research projects to be proposed which would otherwise not have been feasible, and has initiated the development of policies and procedures regarding research access to linked data. These policies and procedures include a framework for addressing the ethical issues surrounding data linkage. The linked data have now been requested and used by several dozen 'external' researchers for a variety of research projects, with considerable success. With continued attention to confidentiality issues, this data presents a valuable resource for health services research and planning.

[1] Green, B., McGrail, K., Hertzman, C., Barer, M.L., Chamberlayne, R., Sheps, S.B., Lawrence, W.J.
University of British-Colombia, Canada

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

A COMPARISON OF TWO RECORD LINKAGE PROCEDURES

Shanti Gomatam, Randy Carter and Mario Ariet[1]

ABSTRACT

Much work on probabilistic methods of linkage can be found in the statistical literature. However, although many groups undoubtedly still use deterministic procedures, not much literature is available on these strategies. Furthermore there appears to exist no documentation on the comparison of results for the two strategies. Such a comparison is pertinent in the situation where we have only non-unique identifiers like names, sex, race *etc.* as common identifiers on which the databases are to be linked. In this work we compare a stepwise deterministic linkage strategy with the probabilistic strategy, as implemented in AUTOMATCH, for such a situation. The comparison was carried out on a linkage between medical records from the Regional Perinatal Intensive Care Centers database and education records from the Florida Department of Education. Social security numbers, available in both databases, were used to decide the true status of the record pair after matching. Match rates and error rates for the two strategies are compared and a discussion of their similarities and differences, strengths and weaknesses is presented.

KEY WORDS: Exact matching; Hierarchical methods; Deterministic strategies; Probabilistic matching; AUTOMATCH.

1. INTRODUCTION

One is often required to combine information on people or items from various sources. In a medical cohort follow-up study for example, a cohort or group of individuals is followed and their medical information is combined with subsequent morbidity or mortality experience. One way such followup studies could be carried out is by "physically" following the group of interest. However such resource-consuming methods limit the size and type of the groups that can be followed. Another way of following cohorts is through surveillance of data sets containing subsequent outcomes and the use of record linkage. So, for example, files containing medical records of individual members of a cohort may be linked with records from files of morbidity and mortality data. Thus the linkage process involves two data sources or "files", say file A, containing medical information on individuals in the cohort, and file B, containing the morbidity or mortality data for members of the cohort. In this paper we consider methods of "exact matching" *i.e.* we are interested in linking information on the same (as opposed to "similar") individuals in the two files.

Suppose file A has n_a records and file B has n_b records, then the file AxB contains $n_a \times n_b$ record pairs. Each of the n_b records in file B is a potential match for each of the n_a records in file A. Thus there are $n_a \times n_b$ potential record pairs whose match/non-match status is to be determined. When we consider the case of "unique matching" *i.e.* we expect at most a single record in one file to be a correct match for any record in the other, then at most $\min(n_a, n_b)$ pairs can be matches. When $n_a = n_b = n$, for example, we see that there are at most n matches, and at least $n(n-1)$ non-matches. Thus we would have to search for $O(n)$ matches among $O(n^2)$ non-matches. Clearly, files A and B must contain some common identifiers that

---

[1] S.V. Gomatam, 4202 East Fowler Av., PHY 114, University of South Florida, Tampa, FL 33620; R.C. Carter, PO Box 100212, Division of Biostatistics, University of Florida, Gainesville, FL32610-0372; M. Ariet, Department of Medicine, Division of Computer Science, University of Florida, Gainesville, FL 32610-0372.

can be used to match the records. The linkage process uses these to classify the $n_a \times n_b$ record pairs in the file AxB as either matches or non-matches.

When unique and reliable identifiers exist for every individual in the databases to be linked, the linkage task is simplified. However, such unique identifiers rarely exist, and it is often necessary to use identifying characteristics such as last name, first name, middle name, and date of birth, in order to link records from the two files. Due to unreliable reporting, non-uniqueness and changes over time, these fields are not enough to specifically identify records, thus making the linkage task difficult.

Two major classes of linkage strategies exist – deterministic and probabilistic. Both deterministic and probabilistic methods of record linkage can be employed in this situation. While Roos and Wajda (1991) have a brief discussion of deterministic and probabilistic methods and suggestions on when their use is appropriate, the literature appears to contain no studies that compare how the methods perform. Here we consider two different methods of record linkage – one a stepwise deterministic method, and the other the probabilistic method implemented in AUTOMATCH – and compare their performance on linkage of files for which the truth is known.

A description of the two linkage strategies compared is given in Section 2, and Section 3 contains the results of their implementations to the data. Section 4 contains a discussion.

2. STRATEGIES BEING COMPARED

In this section we first describe the methodology of probabilistic linkage that AUTOMATCH is based on. The stepwise (or hierarchical) deterministic strategy (SDS) is described next.

Fellegi and Sunter (1969) were one of the first to present the mathematical model and theoretical foundation for probabilistic record linkage rigorously. Under their model record pairs can be classified as either matches (A_1), possible matches (A_2), or non-matches (A_3). For record a from file A and record b from file B, information on the records available in their respective source files is denoted $\alpha(a)$ and $\beta(b)$ respective. A comparison or agreement vector, γ, for a record pair $(\alpha(a), \beta(b))$ represents the level of agreement between record pairs. When record pairs are compared on k identifying fields the γ vector has k components.

$$\gamma = \left(\gamma^1(\alpha(a),\beta(b)), \gamma^2(\alpha(a),\beta(b)), ..., \gamma^k(\alpha(a),\beta(b))\right)$$

is a function on the set of all $n_a \times n_b$ record pairs. Let the set of true matches be denoted by M and that of true non-matches be denoted by U. For an observed agreement vector γ in Γ, the space of all possible comparison vectors,

$$m(\gamma) = P\{\gamma \mid (a,b) \in M\}$$

gives the conditional probability of observing γ given that the record pair is a true match, and

$$u(\gamma) = P\{\gamma \mid (a,b) \in U\}$$

gives the conditional probability of observing γ given that the record pair is a true non-match. Thus the ratio $m(\gamma) / u(\gamma)$ gives the likelihood ratio of match versus non-match for the record pair with vector γ.

The probabilities of the two kinds of errors are: the probability of a false match, also known as a false positive, $P(A_1 \mid U) = \sum_{\gamma \in \Gamma} u(\gamma) P(A_1 \mid \gamma)$, and the probability of a false non-match, also known a false negative or a missed match, $P(A_3 \mid M) = \sum_{\gamma \in \Gamma} m(\gamma) P(A_3 \mid \gamma)$.

For fixed values of the false match rate (μ) and the false non-match rate (λ), Fellegi and Sunter (1969) define the optimal linkage rule on Γ at levels μ and λ, $L(\mu, \lambda, \Gamma)$ as that rule for which $P(A_1 \mid U) = \mu$, $P(A_3 \mid M) = \lambda$ and $P(A_2 \mid L) \leq P(A_2 \mid L')$ for all other rules L' which have error levels μ and λ. Essentially the method fixes the probabilities of the two kinds of errors, and finds an "optimal" rule. Optimal here is defined as the rule that minimizes the probability of classifying a pair as belonging to A_2. Such a rule is a function of the likelihood ratio defined above. Let $m(\gamma)/u(\gamma)$ be ordered to be monotonically decreasing (with ties broken arbitrarily) and the associated γ be indexed by $1, 2, ..., N_\Gamma$. If $\mu = \sum_{i=1}^{n} u(\gamma i)$ and $\lambda = \sum_{i=n'}^{N_\Gamma} m(\gamma i)$, $n < n'$, then Fellegi and Sunter (1969) show that the optimal rule is given by

$$
\begin{aligned}
(\alpha(a), \beta(b)) \in A_1 \ \ & if \ T_\mu \leq m(bold\gamma)/u(\gamma), \\
\in A_2 \ \ & if \ T_\lambda < m(\gamma)/u(\gamma) < T_\mu, \\
\in A_3 \ \ & if \ m(\gamma)/u(\gamma) \leq T_\lambda,
\end{aligned}
$$

where $T_\mu = m(\gamma_n)/u(\gamma_{n'})$, and $T_\lambda = m(\gamma_{n'})/u(\gamma_{n'})$. Under the assumption of conditional independence of the components of the γ vector, the decision rule above can be written as a function of $\log(m(\gamma)/u(\gamma)) = \sum_{j=1}^{k} w_j$, where $w_j = \log(m(\gamma_j)/u(\gamma_j))$.

Jaro (1989) discusses a method of implementing the decision rule proposed by Fellegi and Sunter (1969). He estimates the m probabilities by using the EM algorithm. In his formulation of the problem, the complete data vector is given by (γ, g), where γ is as defined above, and g indicates the actual status (match or non-match) of the record pair. g takes a value corresponding to a match with probability p and that corresponding to a non-match with probability $1-p$. As g cannot be observed, the EM algorithm is used to estimate the parameters m, u and p under the assumption of conditional independence of the components of the agreement vector. Only the m probabilities from this solution are used in the implementation of the decision rule. The u probabilities are estimated as the probabilities of chance agreement between records, by using a frequency analysis. For unique matching the optimal choice of record pairs is obtained by solving a linear sum assignment problem. His methodology is implemented in the commercially available software AUTOMATCH discussed in Jaro (1995). The software requires the specification of: blocking variables, which help reduce the number of comparisons actually carried out; initial values for the m and u probabilities for each of the identifiers considered; and the cutoffs (to determine the three decisions).

The second method used in the comparison is a stepwise deterministic strategy (SDS) constructed for unique matching. In the simplest deterministic record linkage, matches are determined by "all-or-nothing" comparisons, *i.e.* unique agreement on a collection of identifiers called the "match key". In this kind of matching when comparing two records on first and last name, for example, the records are considered matches only if the names on the two records agree on all characters. In the SDS the records are linked in a sequence of steps each of which decides the linkage status (either match or non-match) of the record pair by considering exact agreement on a particular subset of identifiers. The unique matches are extracted, the duplicates are discarded, and the remaining observations in each of the two datasets (the residuals) form the input to the next step in the data linkage process, which continues the process with a different subset of identifiers. Steps that are implemented earlier in the procedure use collections of identifiers that are considered more reliable than those in later steps. The choice of the sequence enables the informal incorporation of some prior knowledge in the subject area.

The advantage of this strategy is that, although each identifying variable is still tested for total agreement at each step, by implementing the succession of steps described above the effect is similar to that of considering some partial agreement matches. Including New York State Intelligence Information System (NYSIIS) phonetic codes for names as alternative identifiers is also likely to increase this effect. The first step implemented would have the lowest false match (or false positive) rate and the highest false non-match rate.

Each successive step would increase the overall false match rate and decrease the false non-match rate. This allows the user some indirect control over error rates via a choice of the number of steps to be executed. One implements the strategy by specifying a sequence of subsets of variables in decreasing order of reliability. Here the strategy is implemented via SAS (Statistical Analysis Systems, Cary, NC) macros.

3. APPLICATION TO RPICC – DOE LINKAGE

Medical records from children born in the years 1989-1992 and treated in Florida's Regional Perinatal Intensive care centers (RPICC) were to be combined with their subsequent educational performance recorded by Florida's Department of Education (DOE). The objective was to model the association between school outcome with the medical and sociodemographic conditions at birth. 49862 RPICC records and 628860 DOE records were available in total. However in order to compare the strategies discussed above we considered only the subset of 1203 records in the RPICC database for which unique social security number matches were available in the DOE database. A subset of the DOE data, that included the 1203 records that were matches to these RPICC records, was combined with an additional 1% of the remaining DOE records to create a DOE file with 7483 records.

Table 1
Breakup of matches obtained in 5 AUTOMATCH passes into true and false matches

Pass Number	1	2	3	4	5	Total
True Match	648	167	2	6	154	977
Total	651	170	2	6	224	1053

The common identifiers used in the linkage were the last name, first name, middle name, date of birth, race, and sex.

A county code was also present in each data set – however, the county of mother's residence at child's birth was available in the RPICC dataset while that of current enrollment was available in the DOE data. NYSIIS codes created for first, middle and last names were also available. Social security numbers were used, after the linkage on other variables was conducted, to judge the true status of the record pair.

For the AUTOMATCH linkage the identifiers last name, first name, middle name, date of birth, race, county number and sex were input with initial m values of 0.9, 0.5, 0.5, 0.97, 0.60, 0.90, 0.95 and u values of 0.0001, 0.0003, 0.0003, 0.0001, 0.02, 0.0001, 0.05 respectively. The m values were revised to 0.86, 0.4, 0.1, 0.81, 0.81, 0.72, 0.94 by the MPROB routine, and the u values were revised to 0.0002, 0.0002, 0.0002, 0.0006, 0.1667, 0.0143, 0.5 by the frequency analysis in AUTOMATCH. Five blocking variables were used: NYSIIS code of last name; NYSIIS code of first name; NYSIIS code of the middle name; County number and sex; County number and date of birth. Based on a study of the histograms of weights that was generated, the following pairs of cutoffs were used for each of the five passes: (20, 15), (15, 11), (8, 6.5), (6, 3.5), (5, 4). This linkage resulted in a total of 1053 matches of which 977 were true matches. Clerical matches, involving extra manual inputs, were ignored for the purposes of this comparison. The breakup of matches for each of the five passes is given in Table 1. A two-way table indicating the breakup for the record pairs is given in Table 2. This linkage gave a sensitivity of 0.8121, a specificity of 0.9999 and a positive predictive probability of 0.9278.

For the SDS the two data sets created above were matched in four stages borrowed from an earlier linkage process on similar data (See Gomatam and Carter, 1999). In the first stage (subsets 1-20) all reasonable matches including county code as one of the identifiers were tried. At this stage we expect to find all children who were treated in the RPICCs and remained in the county of birth when they first enrolled in school. This

resulted in a total of 318 unique matches (*i.e.*, 26.43% of the initial RPICC observations were matched). See Table 3 (Codes used in tables: L = last name, Nl = NYSIIS code of the last name, F = first name, Nf = NYSIIS code of the first name, M = middle name, M1 = first initial in the middle name, Nm = NYSIIS code of the middle name, Dn = County code, D = Date of birth, S = Sex, R = race.

Table 2

Two-way tables indicating breakup of record pairs under the two strategies

AUTOMATCH				Stepwise Deterministic Strategy			
Decision \ Truth	Match	Non-match	Total	Decision \ Truth	Match	Non-match	Total
Match	977	76	1053	Match	759	3	762
Non-match	226	9000770	9000996	Non-match	444	9000843	9000084
Total	1203	9000846	9002049	Total	1203	9000846	9002049

Table 3

First and second stage of RPICC linkage operation: Matches with and without county number

Variable subset	Unique matches	True matches	Variable subset	Unique matches	True matches
1) LFDnDSRM	85	85	21) LFDSRM	12	12
2) LFDNDSRNm	6	6	22) LFDSRNm	0	0
3) LFDnDSRM1	93	93	23) LFDSRM1	14	14
4) LNfDnDSRM	0	0	24) LNfDSRM	0	0
5) LNfDnDSRNm	0	0	25) LNfDSRNm	0	0
6) LNfDnDSRM1	9	0	26) LNfDSRM1	0	0
7) NlFDnDSRM	0	0	27) NlFDSRM	0	0
8) NlFDnDSRNm	0	0	28) NlFDSRNm	0	0
9) NlFDnDSRM1	1	1	29) NlFDSRM1	1	1
10) NlNfDnDSRM	0	0	30) NlNfDSRM	0	0
11) NlNfDnDSRNm	0	0	31) N1NfDSRNm	0	0
12) NlNfDnDSRM1	0	0	32) NlNfDSRM1	0	0
13) LFDnDSR	81	81	33) LFDSR	16	16
14) LNfDnDSR	5	5	34) LNfDSR	4	4
15) NlFDnDSR	1	1	35) NlFDSR	0	0
16) NlNfDnDSR	0	0	36) NlNfDSR	0	0
17) LFDnDS	39	39	37) LFDS	9	9
18) NlNfDnDS	3	3	38) NlNfDS	1	1
19) LFDnDR	3	3	39) LFDR	0	0
20) NlNfDnDR	1	1	40) NlNfDR	0	0
Total	318	318	Total	57	57

At the second stage (subsets 21-40) we tried matches that were similar to those tried in the first stage, except that county number was not used. At this stage we hoped to find children treated in the RPICC who moved from their county of residence at birth to another county within Florida before they enrolled in school. Fifty-seven additional unique matches (*i.e.*, 4.73% (All match percentages in this section are given as a fraction of the initial 1203 RPICC observations)) were found. See Table 3.

At the third stage (subsets 41-50) we attempted to find children who changed their last names between birth and school enrollment by considering merges that did not include last name as one of the variables. A total of 171 observations, or 14.21%, were linked at this stage. See Table 4.

Of the unmatched observations remaining in the RPICC data 593 records (49.29% of the original 1203 observations!) had either missing first names or middle names, or first or middle names of "Baby," "Boy," "Girl," "Male," "Female", or first and middle name combinations of "BB" or "BG". In order to deal with this problem, this subset of RPICC data was matched with residual DOE data. Combinations of variables other than the first and middle name (see Table 4) were used. Of these, 216 (17.96%) had unique matches. See Table 4 (subsets 51-54.).

A total of 762, or 63.34%, of the initial 1203 RPICC records were uniquely linked to DOE records by this process; of these 759 were true matches. A two-way table giving the breakup for the record pairs is given in Table 2. The SDS gave a sensitivity of 0.6309, a specificity of 0.9999, and a positive predictive probability of 0.9961.

Table 4

Third and fourth stages of RPICC linkage operation: Matches for children with last name changes, and those with first/middle name problems

Variable subset	Unique matches	True matches	Variable subset	Unique matches	True matches
41) FDSRDnM	25	23	48) FDSRNm	2	2
42) FDSRDnNm	5	5	49) FDSRM1	6	6
43) FDSRDnM1	28	28	50) NfDSRNm	0	0
44) NfDSRDnNm	3	3	51) LDSRDn	176	176
45) FDSRDn	86	86	52) NlDSRDn	6	6
46) NfDSRDn	7	7	53) LDSR	32	31
47) FDSRM	9	9	54) NlDSR	2	2

4. DISCUSSION

From Section 3 above we see that specificities for both methods are practically the same; AUTOMATCH has a much higher sensitivity than the SDS (0.81 vs. 0.63) but its positive predictive probability is lower (0.928 vs. 0.996). AUTOMATCH gives a much higher match rate (87.53%) than the SDS (63.34%.) Clearly some fraction of the higher rate achieved by AUTOMATCH can be attributed to the fact that partial agreements and missing values are both handled by this strategy. It is unclear what fraction of the increase in match rate can be attributed to the probabilistic methodology and how much to the fact that AUTOMATCH considers information-theoretic character comparison algorithms, that provide for random insertions and deletions, replacements and transposition of characters (See Jaro, 1995) to compare character strings. The SDS uses only phonetic codes, which correct for alternative spellings of some names.

The methodology for probabilistic linkage indicates that one could specify false negative and false positive rates and the rule created could use the estimated weights to achieve these rates. AUTOMATCH however, requires the user to specify the cutoffs for each pass used based on weight histograms it generates. It is not clear what error rates are achieved/achievable in the final match. The SDS described provides a means for indirectly varying error rates through the stepwise selection of identifiers to match on. By appropriately limiting the number of steps used in the matching process the practitioner can increase or decrease the false-positive or false-negative rates. Unfortunately, these error rates are tied to the reliability of the identifiers and precise control is not possible.

Picking a good sequence of subsets for the SDS can be laborious and time consuming in some situations, and involves subjective decisions on the part of the implementor. Similarly the choice of blocking variables and cutoffs, that could make quite a difference to the linkage achieved by AUTOMATCH, introduce subjective

elements to AUTOMATCH's linkage scheme.

This study appears to reinforce the contention that probabilistic matching performs better in information-poor situations. Clearly one should not fall into the danger of making generalizations on the basis of one data analysis, as the effects of a large number of factors (choice of subsets, choice of blocking variables and cutoffs, the nature and reliability of the data *etc.*) could all make significant differences to the results. Additional studies, possibly via controlled simulations, are necessary to make broader statements.

Deterministic strategies like the one above have the disadvantage of being *ad hoc*, not being able to handle missing values or partial agreements, and result in lower match rates. They have the advantage of being easier to interpret, and of allowing the practitioner to informally use her/his knowledge of the data sources and variables to arrive at the matching strategy. In spite of the advantages that probabilistic linkage offers, the cost of software for linkage and the investment required to understand the methodology, probably makes it easier for the average practitioner with limited resources to stay with easily programmable methods like the SDS described above.

ACKNOWLEDGEMENT

The authors thank Mike Resnick of the Department of Pediatrics, UF, for providing the data used in the analysis.

REFERENCES

Fellegi, I.F., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64,1183-1210.

Gomatam, S., and Carter R. (1999). A Computerized Stepwise Deterministic Strategy for Linkage. *Technical Report, Department of Statistics, University of Florida.*

Jaro M.A. (1989). Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, 84, No. 406 414-420.

Jaro M.A. (1995). Probabilistic Linkage of Large Public Health Data Files, *Statistics in Medicine*, 14 491-498.

Roos L.L., and Wajda A. (1991). Record Linkage Strategies. *Methods of Information in Medicine*, 30,117-23.

SESSION V

STATISTICAL MATCHING

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

AN EVALUATION OF DATA FUSION TECHNIQUES

Susanne Raessler and Karlheinz Fleischer[1]

ABSTRACT

The fusion sample obtained by a statistical matching process can be considered a sample out of an artificial population. The distribution of this artificial population is derived. If the correlation between specific variables is the only focus the strong demand for conditional independence can be weakened. In a simulation study the effects of violations of some assumptions leading to the distribution of the artificial population are examined. Finally some ideas concerning the establishing of the claimed conditional independence by latent class analysis are presented.

KEY WORDS: Statistical Matching; Conditional Independence; Simulation Study; Latent Class Analysis.

1. INTRODUCTION

Statistical analyses usually require a single source sample, e.g. a sample of households, for which both TV behaviour and purchasing behaviour is measured. Unfortunately the costs and the nonresponse rates of such a single source panel are high. As a powerful attractive alternative data fusion techniques are considered to combine information from two different samples, one usually of larger size than the other, with the number of households appearing in both samples (i.e. the overlap) clearly negligible. Only certain variables, say Z, of the interesting individual's characteristics can be observed in both samples, they are called common variables. Some variables, Y, appear only in the larger sample and others, X, are observed exclusively in the smaller sample. (For generalization purposes X, Y, Z can be treated as vectors of variables.)

Since a single source sample with variables X, Y, Z does not exist, an artificial sample is generated by matching the observations of both samples according to Z. The matching is performed at an individual level by means of statistical matching, which often is called the marriage process. If multiple use of sample units is to be avoided the smaller sample is taken as recipient sample and the larger sample as donor sample. For every unit i of the recipient sample with the observations (x_i, z_i) a value y from the observations of the donor sample is determined and a data set (x_1, y_1, z_1), ..., $(x_{n_R}, y_{n_R}, z_{n_R})$ is constructed with n_R the number of elements of the recipient sample. The main idea is to search for a statistical match, i.e. a donor unit j whose observed data values of the common variables Z are identical to those of the recipient unit i. In common practice the marriage process is carried out using an algorithm based on nearest neighbour techniques calculated by means of a distance measure. For examples see Antoine (1987), Baker (1990), Bennike (1987), Okner (1974) or Roberts (1994). Sometimes multiple use of donors is restricted or penalized.

In the paper presented here the distribution of the artificial population is derived. The conclusion, as already mentioned by other authors, is that this distribution coincides with the distribution of the real population only if the specific variables X and Y are independent conditioned on the values of the common variables Z. It is also shown that this strong requirement is not necessary if only correlations between specific variables X and Y never jointly observed, are of interest. For this purpose a weaker condition will

[1] Susanne Raessler, University of Erlangen-Nuernberg, Institute of Statistics and Econometrics, Lange Gasse 20, D-90403 Nuernberg, Germany, email: susanne.raessler@wiso.uni-erlangen.de; Karlheinz Fleischer, University of Leipzig, Institute of Empirical Economic Research, Marschnerstrasse 31, D-04109 Leipzig, Germany.

be sufficient. A simulation study is presented investigating the effects of violations of the main assumptions. Finally some ideas are mentioned which try to establish the claimed conditional independence by latent class analysis and a single imputation Monte Carlo algorithm is presented.

2. A THEORETICAL CONCEPT OF DATA FUSION

In the following all probability or density functions (joint, marginal or conditional) and their parameters produced by the fusion algorithm will be marked by the symbol ~. To simplify the presentation we will argue only for the discrete case, the results, however, are valid for the continuous case, too. Thus, in case of discrete variables $f_{X,Y,Z}(x_i, y_i, z_i)$ describes the probability to draw a certain unit i with observation (x_i, y_i, z_i) out of the real population and for continuous variables it is the value of the joint density function at the point (x_i, y_i, z_i).

If the units of both samples are drawn independently from each other (independently within and also between both samples) and if the donor units can be chosen many times without penalization, the units of the artificial sample can be treated as being drawn independently with probability $\tilde{f}_{X,Y,Z}(x, y, z)$ each. Out of the fusion process we get an artificial unit with values (x, y, z) if and only if we find a unit (x, z) in the recipient sample which is matched with a unit (y, z) from the donor sample. According to our matching principle the donor is chosen from all possible donors having the same value z as the recipient unit.

Hence we get the joint distribution

$$\tilde{f}_{X,Y,Z}(x,y,z) = f_{X,Z}(x,z) f_{Y|Z}(y \mid z) = f_{X|Z}(x \mid z) f_Z(z) f_{Y|Z}(y \mid z) = f_{X|Z}(x \mid z) f_{Y,Z}(y,z) \quad (1)$$

and the conditional distribution

$$\tilde{f}_{X,Y|Z}(x,y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z) \quad (2)$$

Data fusion as described above will produce a distribution with certain marginals.

$$\tilde{f}_X(x) = f_X(x),\ \tilde{f}_Y(y) = f_Y(y),\ \tilde{f}_Z(z) = f_Z(z), \quad (3)$$

$$\tilde{f}_{X,Z}(x,z) = f_{X,Z}(x,z),\ \tilde{f}_{Y,Z}(y,z) = f_{Y,Z}(y,z), \quad (4)$$

$$\tilde{f}_{X,Y}(x,y) = \int f_{X|Z}(x \mid z)\, \tilde{f}_Z(z) f_{Y|Z}(y \mid z) dz, \quad (5)$$

$$\tilde{f}_{X,Y,Z}(x,y,z) = f_{X,Y,Z}(x,y,z) \frac{f_{Y|Z}(y \mid z)}{f_{Y|X,Z}(y \mid x,z)} \quad (6)$$

$$\Rightarrow \tilde{f}_{X,Y,Z}(x,y,z) = f_{X,Y,Z}(x,y,z) \quad \text{if} \quad f_{Y|X,Z}(y \mid x,z) = f_{Y|Z}(y \mid z).$$

Thus the distribution of (X, Y, Z) after the fusion is equal to the initial distribution if X and Y are independent, conditional on every possible value z of Z, see also Sims (1972) and Rodgers (1984).

As an obvious consequence we obtain for the moments

$$\tilde{E}(X) = E(X), \tilde{E}(X^i) = E(X^i)(i \in N), \tilde{Var}(X) = Var(x) \text{ and so on.} \quad (7)$$

$$E(Cov(X^i Y^j \mid Z)) = E(X^i Y^j) - \tilde{E}(X^i Y^j), \text{ and} \quad (8)$$

$$\widetilde{Cov}(X^i, Y^j) = Cov(X^i, Y^j) - E(Cov(X^i, Y^j \mid Z)) \quad (i, j \in N), \tag{9}$$

$$\widetilde{Cov}(X, Y) = Cov(X, Y) - E(Cov(X, Y \mid Z)). \tag{10}$$

For a more detailed description see Raessler and Fleischer (1998).

If (X, Y, Z) follow a trivariate normal distribution, the covariance and the correlation coefficient of the variables X and Y can be expressed as

$$\widetilde{Cov}(X, Y) = Cov(X, Y)Cov(Y, Z)/Var(Z), \quad \tilde{p}_{X,Y} = p_{X,Z}p_{Y,Z}. \tag{11}$$

In particular, data fusion can produce "precise results" for the true correlation between variables not jointly observed, only if the variables are uncorrelated in the mean conditional on the common variables Z, i.e., $E(Cov(X, Y \mid Z)) = 0$.

3. SIMULATION STUDY

The derivations above assume the existence for every recipient unit of a donor unit having the same z-value. In practice this assumption is violated and the nearest neighbour is taken as a donor substitute. In a simulation study under several conditions the effect of the deviation from this assumption in estimating the covariance $Cov(X, Y)$ is examined. The following adjustments are chosen and a simulation for each of the 40 combinations is run.

- Distributions: Trivariate normal distribution for correlations $p_{X,Z} = p_{Y,Z} = 0, 0.2, 0.4, 0.6, 0.8$.
  - Sample sizes: $n_R = 500$ with $n_D = 500, 1000$.
  - Distance function: $|z_i - z_j|$
  - Marriage algorithms:
    - "Polygamy", i.e., a donor may be used multiply without penalization.
    - "Bigamy", i.e., a donor may be used twice without penalization.
    - "Monogamy", i.e., a donor may be used once, only.
    - "Triple", i.e., the mean of the three nearest neighbours is used.

Furthermore, some cases for transformations of the trivariate normal distribution and also simulations for sample sizes $n_D = 10000$ and $n_R = 5000$ are performed. For each case $k = 100$ computations are started and the following studentized variable is calculated in order to check the accuracy of estimation of the reproduced covariance $\widetilde{Cov}(X, Y)$:

$$t = \frac{\hat{E}(\hat{\widetilde{Cov}}(X, Y) - \widetilde{Cov}(X, Y))}{s(\hat{\widetilde{Cov}}(X, Y))}\sqrt{k}.$$

As an indicator of minor accuracy $|t| > 2$ is chosen which only matches the following cases:

- $p_{X,Z} = p_{Y,Z} = 0, n_D = 2n_R = 1000$, polygamy ($t = -2.29$).
- $p_{X,Z} = p_{Y,Z} = 0.2 n_D = 2n_R = 1000$, polygamy ($t = -2.04$).
- $p_{X,Z} = p_{Y,Z} = 0, n_D = n_R = 500$, bigamy ($t = -2.76$).
- $p_{X,Z} = p_{Y,Z} = 0.6 n_D = n_R = 500$, monogamy ($t = -3.86$).

- $p_{X,Z} = p_{Y,Z} = 0.8n_D = n_R = 500$, monogamy ($t$ = -8.42).

Enormous values of $|t| > 2$ are observed only for monogamy with equal sample sizes and higher correlations. Remember that it is not the accuracy in estimating the covariance Cov(X, Y) in the true population that is considered but that in estimating the covariance $\widetilde{Cov}(X,Y)$ of the artificial population. Despite the assumptions' violations the latter is estimated quite well. Furthermore, in the triple case the true variance of Y does not transfer to the artificial population, where $\widetilde{Var}(Y) = Var(Y)(1 + 2p_{Y,Z}^2)/3$ holds.

4. DATA FUSION USING LATENT STRUCTURES

4.1 Generalized Latent Class Analysis

It seems to be very difficult to identify common variables ensuring the assumption of conditional independence especially in large data sets. Therefore a data matching process relying on latent class analysis is proposed analogous to a procedure suggested by Kamakura and Wedel (1997). This approach has the attractive property that the variables by assumption are (conditional) independent within such discrete latent classes, see Everitt (1984). Since the manifest variables are of very different nature a generalization will be made combining the so-called latent class and latent profile analysis. In the following, as long as the different meaning of X, Y and Z is not important, the set of variables X, Y, Z shall now be denoted by vector U.

Suppose we have a single latent variable S with unordered categories s = 1, 2,..., T and probabilities $P(S = s) = f_S(s;\eta) = \eta_s$ for s =1,2,... ,T. The basic assumption of this model is that within any class s the manifest variables U are independent of each other. Thus, for any p variables U_j the conditional distribution is $f_{U/S}(u \mid s;\theta) = \prod_{j=1}^{p} f_{U_j|S}(u_j \mid s;\theta)$ within each class s = 1,2,..., T. This latent class model implies that the vector $U = (U_1, U_2, ..., U_p)$ of the observable variables has a probability density function given by

$$f_U(u;\theta;\eta) = \sum_{s=1}^{T} f_S(s;\eta) f_{U|S}(u \mid s;\theta) = \sum_{s=1}^{T} \eta_s \prod_{j=1}^{p} f_{U_j|S}(u_j \mid s;\theta). \qquad (12)$$

The density f_U is a finite mixture density. We will now derive the parameter estimates if q components $f_{U_1|S}, f_{U_2|S}, ..., f_{U_q|S}$ are assumed to be multinomial and the latter p - q components $f_{U_{q+1}|S}, f_{U_{q+2}|S}, ..., f_{U_p|S}$ are normal densities arising from the local independence requirement.

If there is a sample of n independently drawn response vectors $u_1, u_2, ..., u_n$, where $u_i = (u_{i1}, u_{i2}, ..., u_{ip})$ then the loglikelihood function, assuming this latent class model, is given by

$$\begin{aligned} L(\theta,\eta;u) &= \sum_{i=1}^{n} \ln \left(\sum_{s=1}^{T} \eta_s \prod_{j=1}^{P} f_{U_j|S}(u_{ij} \mid s;\theta) \right) \\ &= \sum_{i=1}^{n} \ln \left(\sum_{s=1}^{T} \eta_s \prod_{j=1}^{q} \prod_{k=1}^{K_j} \theta_{jk.s}^{I_k(u_{ij})} \prod_{j=q+1}^{p} (2\pi\sigma_{j.s}^2)^{-1/2} e^{-(u_{ij}-\mu_{j.s})^2/(2\sigma_{j.s}^2)} \right) \end{aligned} \qquad (13)$$

where the indicator function $I_k(u_{ij})$ equates one, if category k of variable U_j is observed for the ith individual, and zero otherwise. The parameter $\theta_{jk.s}$ gives the probability that in class s the variable U_j has a value of category k. Analogously, the parameters $\mu_{j.s}$ and $\sigma^2_{j.s}$ belong to the normal densities in each class s.

To find the maximum likelihood estimates of the parameters in the model the loglikelihood is maximized under the restriction $\Sigma_{s=1}^{T}\eta_s = 1$. After a little algebra this leads to the following estimation equations, see Everitt (1984) or Dillon and Kumar (1994), suggesting an iterative scheme for their solution:

$$\hat{\eta}_s = \frac{1}{n}\sum_{i=1}^{n}\hat{\eta}_{s.i}, \quad s = 1,2,...T, \tag{14}$$

$$\hat{\theta}_{jk.s} = \frac{\Sigma_{i=1}^{n} I_k(u_{ij})\hat{\eta}_{s.i}}{\Sigma_{i=1}^{n}\hat{\eta}_{s.i}}, \tag{15}$$

for j = 1,2,..., q, k = 1,2,..., K_j and s = 1,2,..., T. The parameters of the continuous manifest variables are estimated by

$$\hat{\mu}_{j.s} = \frac{\Sigma_{i=1}^{n} u_{ij}\hat{\eta}_{s.i}}{\Sigma_{i=1}^{n}\hat{\eta}_{s.i}}. \tag{16}$$

$$\hat{\sigma}^2_{j.s} = \frac{\Sigma_{i=1}^{n}(u_{ij} - \hat{\mu}_{j.s})^2\hat{\eta}_{s.i}}{\Sigma_{i=1}^{n}\hat{\eta}_{s.i}}, \tag{17}$$

for $j = q+1, q+2,..., p$ and s =1,2,... ,T. Notice that $\hat{\eta}_{s.i}$ is the estimated value of the so-called posterior probability, that observation u_i arises from class s of the latent variable. This estimate is obtained as follows:

$$f_{S|U}(s \mid u_i;\hat{\theta},\hat{\eta}) = \frac{\hat{\eta}_s f_{U|S}(u_i \mid s;\hat{\theta})}{f_U(u_i;\hat{\theta},\hat{\eta})} := \hat{\eta}_{s.i} \tag{18}$$

for each class s = 1,2,..., T and each unit i = 1,2,..., n. A common approach to find the maximum likelihood solutions relies on the so-called EM algorithm, see Dillon and Kumar (1994) and Dempster et al. (1977). Each iteration of the algorithm consists of two steps. In the E-step new estimates of the posterior probabilities are calculated based upon a current set of the parameter values. In the M-step new estimates of the parameter values are obtained on the basis of these provisional posterior probabilities.

The question of identifiability of latent class models and the uniqueness of the EM solutions are discussed in detail by Dillon and Kumar (1994). Let V be the number of unique response patterns formed by the manifest variables, then clearly a necessary condition for the model to be identified is that

$$V > (T-1) + T\sum_{j=1}^{q}(K_j - 1) + 2T(p-q). \tag{19}$$

Thus, overfitting of the model is restricted by the number of response patterns. The number of latent categories T has to be fixed a priori which often is done using well known information criteria like AIC, BIC or CAIC and chi-square goodness-of-fit statistics.

4.2 A Single Imputation Monte Carlo Solution

The following approach is initiated by a multiple imputation procedure published by Kamakura and Vedel (1997). Their algorithm is restricted to the estimation of cell frequencies between two not jointly observed variables. Working together with german market research companies our task is to produce one artificial fusion sample which can be offered to a broad audience of customers for individual data analysis. Thus we are trying to implement a single imputation solution. The following algorithm uses all variables and observations of the two samples, considering them a unique sample of size $n = n_R + n_D$ with some data missing. Now the cell frequencies are estimated using (15) summing over n observations if variable U_j is a common variable and over $n_R (n_D)$ observations if variable U_j belongs to the recipient (donor) sample only. The same is done for the continuous variables in the samples using (16) and (17). Finally, the latent class probabilities $\hat{\eta}_s$ are calculated as mean of all n posterior probabilities $\hat{\eta}_{s.i}$, which are generated by (18) with $f_{U|S}(u_i \mid s;\hat{\theta}) = f_{X.Z|S}(x_i, z_i \mid s;\hat{\theta})$, $f_U(u_i;\hat{\theta}) = f_{X.Z}(x_i, z_i;\hat{\theta})$ for $I = 1,2,\ldots, n_R$ and $f_{U|S}(u_i \mid s;\hat{\theta}) = f_{Y.Z|S}(y_i, z_i \mid s;\hat{\theta})$, $f_U(u_i;\hat{\theta}) = f_{Y.Z}(y_i, z_i;\hat{\theta})$ for $I=1,2,\ldots, n_D$. This algorithm for multinomial distributed variables is described in detail by Kamakura and Wedel (1997). Having found an EM solution for the parameters using all available data we suggest the following algorithm:

(1) Calculate the posteriors $\hat{\eta}_{s.i}$ using the EM solutions as described above.

(2) For a unit i sample the latent class s from $\hat{\eta}_{s.i}$, i.e. draw from a multinomial distribution with probabilities $\hat{\eta}_{1.i},\ldots,\hat{\eta}_{T.i}$.

(3) Impute the missing values of the variables $Y = U_j$ of unit i in the recipient sample conditional upon the values of the parameters θ and the latent group s, i.e. draw from a multinomial distribution with probabilities $\theta_{j1.s}, \hat{\theta}_{j2.s},\ldots,\hat{\theta}_{jKj.s}$ or from a normal distribution with mean $\hat{\mu}_{j.s}$ and variance $\hat{\sigma}^2_{j.s}$.

Like the procedure of Kamakura and Wedel (1997) this imputation process itself is noniterative offering computational advantages. Based on this algorithm there are further variations which still have to be evaluated. To illustrate the proposed imputation procedure a simple numerical example is presented below.

4.3 Numerical Example

Dillon and Kumar (1994) present data from Dash, Schiffman, and Berenson on the shopping behaviour of 412 audio equipment buyers in terms of five dichotomous variables generating 32 unique response patterns. Using the original data set different latent class models were applied leading to the estimated and expected cell frequencies for the response patterns as presented by Dillon and Kumar (1994) in detail. To identify the "best" model, the conventional Pearson (X^2), likelihood ratio (G^2) and Neyman (X^2_{Ney}) chi-square goodness-of-fit statistics are calculated comparing expected $\hat{n}_i$ and observed n_i counts, i = 1, 2,..., V. Moreover, the Hellinger distance measure, with $\sqrt{\Sigma_{i=1}^{V}(\sqrt{n_i/n} - \sqrt{\hat{n}_i/n})^2}$ and a Rao distance measure for multinomial distributions with $2\sqrt{n}\arccos(n^{-1}\Sigma_{i=1}^{V}\sqrt{n_i \hat{n}_i})$ are computed. Notice that, whenever a latent class parameter assumes a boundary solution, i.e. a value of 0.0 or 1.0, these parameters are treated

as fixed in advance and the degrees of freedom *df* are to be corrected for. The following table gives an overview of the goodness-of-fit statistics (see Dillon and Kumar (1994)) for one (i.e. "global" independence), two, three and four latent classes being fitted to the original data:

T	Df	Fixed	$X^2_{0.95}$	p-value	X^2	G^2	X^2_{Ney}	Hell.	Rao
1	26	0	38.8851	0.0000	300.4418	221.6442	243.5420	0.3554	14.5053
2	20	0	31.4104	0.0079	38.4172	40.1695	60.2158	0.1605	6.5246
3	17	3	27.5871	0.1147	24.1755	25.1635	32.7172	0.1262	5.1260
4	13	5	22.3620	0.4924	12.4340	12.1589	12.6271	0.0858	3.4860

We suppose now that 206 buyers were not asked variable 5 and another 206 buyers were not asked variable 4. Applying the EM-algorithm described above to these two data sets leads to cell estimates which could be used for random draws following the proposed imputation procedure. To use all information at hand, in this example all n observations were taken and their missing values of variable 4 or 5 imputed. The averages of the observed frequencies of the response patterns produced by this algorithm being performed k = 1000 times, are compared with the originally observed frequencies leading to the different distance measures tabled below:

T	Df	Fixed	$X^2_{0.95}$	p-value	X^2	G^2	X^2_{Ney}	Hell.	Rao
1	26	0	38.8851	0.0000	66.4282	69.3840	109.3574	0.2113	8.5957
2	20	0	31.4104	0.0957	28.6102	29.3131	37.2174	0.1357	5.5137
3	15	1	24.9958	0.0123	29.8867	29.8996	34.8783	0.1360	5.5248
4	10	2	18.3070	0.0000	65.3571	43.6924	47.3560	0.1569	6.3763

Due to rising identification problems the results for the four-class model in the fusion case are rather bad. Following p-value and distance measures the two-class model seems to be the best choice. Actually, the results of the average of the two-class model imputations are even better than those of the two-class model applied to the original data. According to the chi-square goodness-of-fit statistics for the two-class model a test of homogeneity could not reject the hypothesis that both samples, the original one and the mean imputed one, arise from the same population.

5. SUMMARY

Fusion of data sets via the use of identical common variables or even nearest neighbour matches can reproduce the true correlation between variables X and Y not jointly observed if and only if they are conditionally uncorrelated in the mean, i.e., $E(\mathrm{Cov}(X,Y|Z))=0$. The stronger demand for conditional independence is not necessary if the interest is focused on the correlation between X and Y only.

Another approach using latent structure analysis assuming conditional independence within discrete latent classes was discussed briefly. Some ideas concerning a single imputation method leading to first encouraging results have been presented. There are a lot of questions left, of course. But the latent structure analysis appears to be worth doing further research to establish conditional independence.

REFERENCES

Antoine, J. (1987), A Case Study Illustrating the Objectives and Perspectives of Fusion Techniques, *Readership Research: Theory and Practice*, eds. Henry, H., Amsterdam: Elsevier Science Publishers, 336-351.

Baker, K. (1990), *The BARB/TGI Fusion. Technical Report on the Fusion Conducted* in *February/March 1990*, Ickenham, Middlesex: Ken Baker Associates.

Bennike, S. (1987), Fusion - An Overview by an Outside Observer, *Readership Research: Theory and Practice*, eds. Henry, H., Amsterdam: Elsevier Science Publishers, 334-335.

Dillon, W.R. and Kumar, A. (1994), Latent Structure and Other Mixture Models in Marketing: An Integrative Survey and Review, in *Advanced Methods in Marketing Research*, eds. Bagozzi, R.P., Cambridge: Blackwell Publishers, 295-351.

Dempster, A.P., Laird, N.M. and Rubin, B.D. (1977), Maximum Likelihood Estimation from In-complete Data via the EM Algorithm (with discussion), *Journal of the Royal Statistical Society B*, 39,1-38.

Everitt, B.S. (1984), *Latent Variable Models*, London: Chapman and Hall.

Kamakura, W.A. and Wedel, M. (1997), Statistical Data Fusion for Cross-Tabulation, *Journal of Marketing Research*, 34, 485-498.

Okner, B.A. (1974), Data Matching and Merging: An Overview, *Annals of Economic and Social Measurement*, 3, 347-352.

Raessler, S. and Fleischer, K. (1998), Aspects Concerning Data Fusion Techniques, *ZUMA Nachrichten Spezial*, 4, 317-333.

Roberts, A. (1994), Media Exposure and Consumer Purchasing: An Improved Data Fusion Technique, *Marketing and Research Today*, 150-172.

Rodgers, W.L. (1984), An Evaluation of Statistical Matching, *Journal of Business and Economic Statistics*, 2,1,91-102.

Sims, C.A. (1972), Comments, *Annals of Economic and Social Measurement*, 1, 343-345.

Proceedings of the Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

A DONOR IMPUTATION SYSTEM TO CREATE A CENSUS DATABASE FULLY ADJUSTED FOR UNDERENUMERATION

Fiona Steele[1], James Brown[2] and Ray Chambers[2]

ABSTRACT

Following the problems with estimating underenumeration in the 1991 Census of England and Wales the aim for the 2001 Census is to create a database that is fully adjusted for net underenumeration. To achieve this, the paper investigates weighted donor imputation methodology that utilises information from both the census and census coverage survey (CCS). The US Census Bureau has considered a similar approach for their 2000 Census (see Isaki *et al* 1998).

The proposed procedure distinguishes between individuals who are not counted by the census because their household is missed and those who are missed in counted households. Census data is linked to data from the CCS. Multinomial logistic regression is used to estimate the probabilities thathouseholds are missed by the census and the probabilities that individuals are missed in counted households. Household and individual coverage weights are constructed from the estimated probabilities and these feed into the donor imputation procedure.

The first stage of the imputation procedure uses household coverage weights in conjunction with geographical information to determine the number of households that need to be imputed in the local administrative area. Donor households are selected from counted households with characteristics similar to those of the missed households. The second stage of the procedure imputes individuals missed in counted households, based on individual coverage weights. A donor is selected for each imputed individual who is then placed in a suitable counted household. Finally, certain marginal totals generated from the post-imputation database need to be calibrated to already agreed estimates at the local administrative area. To achieve this for age-sex and household size, some addition and deletion of imputed individuals is necessary.

KEY WORDS: Controlled Imputation, Census Underenumeration, Weighting, Calibration

1. INTRODUCTION

The basic level of underenumeration for the national population estimated by a Post Enumeration Survey (PES) is a useful guide to how well the census has performed and adjustments at the regional level are often used in the re-basing of mid-year population estimates. However, national and local governments use the population census (usually every five or ten years) as the basis for planning and resource allocation down to very small areas. If underenumeration is uniform by geography and by the characteristics of individuals and households then it can effectively be ignored. However, this is generally accepted to not be the case. This paper describes an approach to this problem that adjusts the whole census database to reflect the level of underenumeration recorded at the national level. A similar approach has been investigated for adjustment of the US Census, see Isaki *et al* (1998).

The United Kingdom (UK) has faced the problem of underenumeration for a number of censuses with net underenumeration measured by a PES. In 1991, though, the PES was unable to estimate either the level of net underenumeration at the national level or allocate underenumeration to a regional level. To ensure that this does not occur in 2001 a major research project has been undertaken so that, in 2001, the Office for National Statistics (ONS) will be in a position to **estimate** and **adjust** accurately census outputs for net underenumeration. The overall ONC research is described in Brown *et al* (1999a). The first stage involves

[1] Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, UK.
[2] Department of Social Statistics, University of Southampton, Southampton, Hants SO17 1BJ, UK.

the estimation of underenumeration at sub-national administrative areas called Local Authority Districts or LADs. To achieve this, individuals on the census database in the PES sample areas must be 'matched' to the database generated by the PES, to be called the Census Coverage Survey (CCS) in the 2001 Censuses of the UK. This paper assumes that stage one has taken place and population estimates for these areas are available for a variety of basic demographic characteristics for both households and individuals. In what follows a controlled imputation methodology for stage two is presented. The aim is to adjust the individual level database so that it is consistent with the population estimates already produced at stage one. This is achieved using an imputation system driven by coverage weights estimated for households and individuals, which is described in Section 2. Section 3 presents a simulation study that investigates the performance of the controlled imputation methodology assuming population estimation from the CCS at stage one is achieved without error.

2. CONTROLLED IMPUTATION METHODOLOGY

There are a series of steps in the creation of a database that is fully adjusted for underenumeration.

1) Modelling the census coverage of households and individuals.

2) Imputation of households completely missed by the census.

3) Imputation of individuals missed by the census in counted households.

4) Final adjustments to the database in order to satisfy the consistency requirements for a ONC.

The methodology for each step is outlined in the following sections.

2.1 Step 1: Estimation of household and individual coverage weights

2.1.1 Derivation of household coverage weights

Following the census and the CCS each household within a CCS area can be placed in one of following four categories:

1) Counted in the census, but missed by the CCS
2) Counted in the CCS, but missed by the census
3) Counted in both the census and the CCS
4) Missed in both the census and the CCS

A simplifying assumption is that category four contains no households, that is no household is missed by both the census and the CCS. While an unrealistic assumption, the households missed by both are accounted for in the dual system estimates at the design area[3] level and the final imputed database is constrained to satisfy the totals at the CCS design area level. Excluding category 4, categories 1, 2 and 3 define a multinomial outcome that can be modelled for each LAD as follows:

$$\log\left(\frac{\theta_{jke}^{(t)}}{\theta_{jke}^{(3)}}\right) = \lambda^{(t)} Z_{jke} \qquad t = 1,2 \qquad (1)$$

[3] The design area is a group of LADs for which the CCS can make direct estimates, see Brown *et al* (1999a, 1999b).

where $\theta_{jke}^{(t)}$ is the probability that household j in postcode k in enumeration district (ED) e in an LAD with characteristics defined by Z_{jke} is in category t. (Model (1) uses category 3 as the reference category.) With matched data from the census and CCS, this model is straightforward to fit.

The estimated model for the CCS areas is extrapolated to non-CCS areas within the LAD to obtain predicted probabilities of being in a particular response category for each household. The probabilities for each response category estimated under model (1) are then used to calculate a coverage weight for each household (h/h) counted in the census that can be applied to the household database. The household coverage weight is defined as

$$w_{jke}^{h/h} = \frac{1}{\theta_{jke}^{(1)} + \theta_{jke}^{(3)}}$$

However, the resulting weighted sums of counted households will not, in general, match corresponding totals estimated for the LAD. Therefore the weights are calibrated to the LAD marginal totals for key household variables, such as tenure, using iterative proportional scaling.

2.1.2 Derivation of individual coverage weights

To calculate coverage weights for those individuals counted in counted households, two assumptions are necessary regarding coverage of individuals in CCS areas. If a household is only counted by the census, then no individuals from that household are missed by the census. Similarly, if only the CCS counts the household then no individual from that household is missed by the CCS. These assumptions are necessary because a household counted by only one source has no second list against which counted individuals can be compared. Although this assumption does not hold in general, people missed as a consequence are accounted for through constraining to population totals at the LAD level. In this case the possible categories of counted individuals are:

a) Counted in the census, but missed by the CCS
b) Counted in the CCS, but missed by the census
c) Counted in both the census and the CCS

These categories are then used to define the outcome in another multinomial model:

$$\log\left(\frac{\pi_{ijke}^{(r)}}{\pi_{ijke}^{(c)}}\right) = \beta^{(r)} X_{ijke} + \gamma^{(r)} Z_{jke} \qquad r = a, b \qquad (2)$$

where $\pi_{ijke}^{(r)}$ is the estimated probability that individual i in household j in postcode k in ED e within an LAD with individual characteristics defined by X_{ijke} and household characteristics defined by Z_{jke} is in category r. (Model (2) uses category c as the reference category.) As before this model can also be extended to included random effects terms.

As with the household model the fitted model is then extrapolated to non-CCS areas to obtain predicted probabilities of being in a particular response category for each individual. The probabilities estimated under the model are used to calculate a coverage weight for each individual (ind) that can be applied to the individual database. The individual coverage weights are calculated as

$$w_{ijke}^{ind} = \frac{1}{\pi_{ijke}^{(a)} + \pi_{ijke}^{(c)}}$$

As before the resultant weighted sums of census counted individuals will not be equal to the corresponding LAD totals. At the final stage of the imputation procedure, further adjustments are necessary to meet

agreed LAD totals by age, sex and household size. To minimise the amount of adjustment required at this stage, individual coverage weights are calibrated to the agreed age-sex totals following the household imputation but before the imputation of individuals.

2.2 Step 2: Imputation of households

The household-based file of counted households in an LAD is matched to the file of calibrated household coverage weights (as described in Section 2.1.1). This file is sorted by coverage weight, and by geographical location. For more efficient processing, households are then grouped into impute classes defined by the characteristics on which the household coverage weights are based. Weights are grouped into bands to give impute classes. The processing block is an impute class within an LAD.

Within each processing block, households are processed sequentially and running totals are retained of the unweighted household count and the weighted household count (calculated using calibrated coverage weights). Whenever the weighted count exceeds the unweighted count by more than 0.5, households are imputed into the ED currently being processed until the difference between the weighted and unweighted running totals is less than or equal to 0.5. An imputed household is assigned a household coverage weight of zero.

In order to assign characteristics to the imputed households, a donor imputation method is used. For each imputed household, a donor is selected at random from among the counted households with the same weight and in the same ED as the counted household that was processed immediately before the imputation. Once a donor has been selected, the characteristics of the household and its occupants are copied to the imputed household. The imputed household is then assigned at random to a postcode within the ED.

A further source of information available in the UK is the 'dummy form'. This is a form completed by an enumerator which indicates the presence of a non-vacant household that has not been enumerated in the census. Previous censuses have shown that the quality of information collected in these forms is variable. However, the possibility of using them in the choice of location for imputed households is being investigated. This will require the capture of the information onto a computer database, something that has not been done in the past.

2.3 Step 3: Imputation of individuals into counted households

The individual weights estimated in Section 2.1.2 are not calibrated to population totals when calculated. However, it is necessary to do this to ensure that enough extra individuals with the correct characteristics are added. This is achieved by using iterative scaling to calibrate the weights to population totals that reflect the individuals already imputed by the household imputation described in Section 2.2.

The individual-based file of counted individuals is then sorted by weight, and by geographical location. Impute classes are defined by the characteristics on which the individual coverage weights are based. Individual coverage weights are grouped into bands to give impute classes. Within a processing block (impute class within a LAD), counted individuals are processed sequentially. When the weighted count of individuals exceeds the unweighted count by more than 0.5, individuals are imputed in the current ED until the difference is less than or equal to 0.5.

Individual and household characteristics are assigned to the imputed individuals in two separate stages. Some of an imputed individual's characteristics are determined by the weight of the last counted individual that was processed before the imputation. The remaining individual characteristics are copied from a suitable donor. The search for a donor is carried out in the same way as described above for the household imputation. The donor is selected at random from among the counted individuals with the same coverage weight and in the same ED as the counted individual that was processed immediately before the imputation. When a donor is found, the LAD is searched for a suitable recipient household in which to place the

imputed individual. The household characteristics for an imputed individual come from the selected recipient.

In order to maintain sensible household structures for households into which individuals have been imputed, the type of recipient household sought depends on certain characteristics of the donor. In the simulation study that follows the choice of recipient depends on the age, marital status and houschold structure of the donor. Household structure is defined using both census and CCS information. Therefore, if an individual who was missed by the census is found in the CCS, the structure of their household will be edited accordingly. To illustrate the recipient search, consider an individual that the coverage weights suggest needs to be imputed. Suppose that a married person went missing from a 'couple without children' household. The household structure(s) that would result after exclusion of the imputed person defines the structure required for the recipient household. Thus the recipient for this individual must be a single person household. In this case, the marital status of the single person would be edited to married after the imputed person is added to the household. In a further attempt to maintain sensible households, the age-sex composition of the donor's household is also taken into account in the search for a recipient. After selection of a suitable recipient, the imputed individual is placed in the chosen household and is assigned the recipient's household characteristics.

2.4 Step 4: Final calibration ('pruning and grafting')

Due to the calibration of household coverage weights carried out before the household imputation, the number of households in each impute class will be within one household of the weighted total for that class. Further, the distribution of the household variables to which household weights are calibrated will be almost exactly the same as the target distributions. However, the household size distribution will be incorrect. This is due to individuals being imputed in both Step 2 and Step 3 that leads, in general, to too many larger households. In the final calibration stage, the post-imputation database is adjusted to ensure that the household size distributions and age-sex distributions derived from the ONC database agree with the ONC estimates of their distributions at the LAD level. To achieve this aim some addition and/or deletion of imputed individuals from imputed and counted households will be necessary.

The basic idea of the 'pruning and grafting' procedure is to start at the largest households and work down to houscholds of size one, adding ('grafting') and deleting ('pruning') people to move households up or down in size. The addition of individuals follows the same process as individual imputation while the deletion is at random from a set of possible imputed individuals. This is controlled so that the age-sex distribution after pruning and grafting is exactly calibrated to the control distribution.

3. SIMULATION STUDY

3.1 Generation of the census and CCS data

As with any simulation study the exact nature of the results will depend on the way in which the data have been simulated. For this study ten censuses have been generated from a LAD of 1991 Census records using the same methodology applied to simulating censuses and CCSs in Brown *et al* (1999a and 1999b). The simulation population and CCS design is also the same as the population used for the simulations in Brown *et al* (1999b).

3.2 Generation of the households and individual coverage weights

For the simulated data set three multinomial models, using main effects only, have been estimated: one for household coverage, based on model (1) in Section 2.1.1, and two separate models for individual coverage of adults and children in counted households, both based on model (2) in Section 2.2.2. The explanatory variables used in the household model are tenure, household ethnicity, houschold structure, and the enumeration district's HtC index. In the model for individual coverage within counted households children have been considered separately from adults, as they do not have an economic status (as measured by the

census). The explanatory variables in the model for children are sex, age group at the individual level, a simplified tenure variable and the number of counted adults based on the household structure variable at the household level, along with the enumeration district's HtC index. The model for adults extends the model for children to include economic status, marital status at the individual level and the full household structure variable at the household level. It is important to remember that all variables are based on the joint census-CCS data. If there is a conflict the census value is chosen unless it is due to an individual being missed.

The household coverage weights are calibrated to satisfy marginal distributions estimated at the local authority district level. For this simulation the 'true' marginal distributions have been used, as the aim here is to test the imputation methodology rather than the ability to estimate totals at a higher level. The weights have been calibrated to the true distributions by tenure, household ethnicity, and HtC index. Using the HtC index ensures that, in general, the hardest to count enumeration districts will get more imputed households. The calibration was carried out using an iterative scaling algorithm that converged very rapidly. The individual coverage weights are approximately (see Section 2.3) calibrated to marginal distributions after accounting for the individuals added by the household imputation. As with the household calibration the 'true' marginal distributions have been used.

3.3 Evaluation of the imputation procedure

The methodology described in Section 2 has been applied to the simulated census database and its associated household and individual coverage weights. The computer time taken to run the whole procedure is approximately 48 hours on a 450MHz Pentium II PC with 512 megabytes of RAM.

To evaluate the performance of the imputation methodology on the simulated census database, the marginal distributions of key household and individual characteristics in the unadjusted census and fully adjusted census databases are compared with their true distributions. Standard Pearson chi-square tests are used to test the hypothesis that the distribution of a variable in the adjusted census database is the same as its distribution in the true database. For a categorical variable with C classes the chi-square statistics is calculated as:

$$X^2 = \sum_{c=1}^{C} \frac{(T_c^{(adj)} - T_c)^2}{T_c}$$

where $T_c^{(adj)}$ is the number of households (or individuals) in class c in the fully adjusted census file, and T_c is the true number of households (individuals) in class c. The test statistic is compared to a chi-square distribution on C-1 degrees of freedom.

Chi-square tests are also used to compare distributions of household and individual variables in the unadjusted census database to their true distributions. In the calculation of the chi-square statistic to compare census counts with the truth, the true counts T_c are replaced by the census counts that would be expected if the census had the same percentage distribution as the true database. A comparison of these measures with those obtained for the adjusted census-truth contrast provides an indication of the improvement of the adjusted census database over the unadjusted census database.

Table 1: Evaluation of imputation procedure for selected household and individual variables

	X^2: tests against true distribution (p-value)			X^2: tests against true distribution (p-value)	
Household variable	Adj.	Unadj.	Individual variable	Adj.	Unadj.
*Tenure**	0.73 (0.998)	179.94 (0.000)	*Ethnicity*	8.28 (0.407)	68.93 (0.000)
Building type	0.49 (0.999)	116.11 (0.000)	*Primary activity last week**	6.43 (0.893)	486.12 (0.000)
Number of cars	0.43 (0.934)	37.38 (0.000)	*Tenure**	3.44 (0.842)	267.87 (0.000)

*Note that coverage weights have been calibrated to 'true' LAD marginal totals for these variables.

Results from the chi-square tests for selected household and individual variables are presented in Table1. Before adjustment the marginal distribution of each of these variables differs significantly from the true distributions. However after controlled imputation has been applied to the census the marginal distributions at the LAD level are correct. This is expected for variables such as tenure that have been calibrated for both households and individuals, but not necessarily for variables such as ethnicity that are not calibrated.

While this is not a complete evaluation of the controlled imputation procedure, these initial results are extremely encouraging and demonstrate the feasibility of this methodology to create an adjusted census database. Future analysis will need to consider joint distributions at the LAD level and performance at the ED level.

REFERENCES

Brown, J., Buckner, L., Diamond, I., Chambers, R. and Teague, A. (1999a) A Methodological Strategy for a One Number Census in the United Kingdom. *J. R. Statist. Soc. A*, **162**, 247-267.

Brown, J., Diamond, I., Chambers, R. and Buckner, L. (1999b) The Role of Dual System Estimation in the 2001 Census Coverage Surveys of the UK. Statistics Canada Symposium, Ottawa, 4th to 7th May, 1999.

Isaki, C. T., Ikeda, M. M, Tsay, J. H and Fuller, W. A. (1998) An estimation file that incorporates auxiliary information. Submitted to *J. Off. Statist.*

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

INTEGRATED MEDIA PLANNING THROUGH STATISTICAL MATCHING: DEVELOPMENT AND EVALUATION OF THE NEW ZEALAND PANORAMA SERVICE

James Reilly[1]

ABSTRACT

To reach their target audience efficiently, advertisers and media planners need information on which media their customers use. For instance, they may need to know what percentage of Diet Coke drinkers watchBaywatch, or how many AT&T customers have seen an advertisement for Sprint during the last week. All the relevant data could theoretically be collected from each respondent. However, obtaining full detailed and accurate information would be very expensive. It would also impose a heavy respondent burden under current data collection technology. This information is currently collected through separate surveys in New Zealand and in many other countries. Exposure to the major media is measured continuously, and product usage studies are common. Statistical matching techniques provide a way of combining these separate information sources. The New Zealand television ratings database was combined with a syndicated survey of print readership and product usage, using statistical matching. The resulting Panorama service meets the targeting information needs of advertisers and media planners. It has since been duplicated in Australia. This paper discusses the development of the statistical matching framework for combining these databases, and the heuristics and techniques used. These included an experiment conducted using a screening design to identify important matching variables. Studies evaluating and validating the combined results are also summarized. The following three major evaluation criteria were used; accuracy of combined results, stability of combined results and the preservation of currency results from the component databases. The paper then discusses how the prerequisites for combining the databases were met. The biggest hurdle at this stage was the differences between the analysis techniques used on the two component databases. Finally, suggestions for developing similar statistical matching systems elsewhere will be given.

KEY WORDS: Statistical matching; Media planning; Data fusion; Ratings; Readership; Product usage.

1. INTRODUCTION

1.1 Background

Advertisers and media planners rely on measures of media usage, such as television ratings and magazine and newspaper readership, along with information about product usage to help plan and assess their communications. Traditionally these have been measured in separate surveys. Television ratings have ideally been measured using panels of households with meters attached to their television sets, where the respondents use a remote control to indicate when they are viewing the set. Readership is usually monitored through a large-scale continuous survey, and product usage is often measured in separate customized surveys.

To transfer knowledge gained from one survey to another, the variables of ultimate interest are analyzed by other variables common to all the surveys, particularly demographic characteristics such as age, sex, and income. Some behavioral information is also commonly used, such as the degree of responsibility for household grocery shopping. Often the target audience for communications would ideally relate to consumers of a specific product, but because of the need to bridge the different surveys, this will be translated as best it can be into demographic terms. A media campaign is then targeted to this demographic surrogate group.

[1]James Reilly, Statistics Director, ACNielsen (NZ) Ltd, PO Box 33 819, Auckland 9, New Zealand. Email: reillyj@acnielsen.co.nz or james_reilly@yahoo.com

However, this can introduce inefficiencies due to the demographic group being an imperfect surrogate for product usage (Assael and Poltrack, 1994). More efficient media planning is possible when using integrated data that covers all these areas.

ACNielsen conducts surveys in New Zealand covering all these areas. The national television ratings are collected from a panel of metered households, while the national readership survey also collects product usage information. The two surveys cannot feasibly be combined, partly due to the heavy respondent burden this would impose, at least using current data collection technology. ACNielsen instead developed its Panorama service, which integrates the separate databases using statistical matching. An integrated Panorama database is delivered to media planners each month, allowing them to plan their media campaigns more efficiently than would otherwise have been possible.

1.2 Organization of the Paper

The paper begins with a brief review of the relevant literature. It then outlines the techniques used in the statistical match, and discusses several evaluations of the results. The prerequisites for matching are then covered, with particular emphasis on those most relevant to this matching exercise and those that have not had much exposure in the literature. The final section gives some suggestions for organizations considering undertaking similar projects.

2. LITERATURE REVIEW

2.1 Relevant Literature

There is a recognized need for statistical matching in the field of media research, where it is usually known as data fusion (Weigand, 1986; Dow, 1989; Brown, 1991). It has been successfully applied in this field and others (for example Wolfson *et al*, 1989; Baker *et al*, 1997). Because statistical matching can be misapplied, there is wide agreement that it is important to test the results of any matching exercise (Rothman, 1988). Some authors assert that testing the relationships between the merged data items is particularly important (Barr and Turner, 1990; Brown, 1991). Some attention has also been given to the prerequisites for a good statistical match (Ruggles *et al*, 1977 among others).

3. OUTLINE OF MATCHING FRAMEWORK

3.1 Matching Framework

Three of the main choices to be made when developing a statistical matching framework are how to define the exact matching cohorts, what matching technique should be used within these cohorts, and what distance function should be used in the matching.

We initially defined the exact matching cohorts based on age, sex and area, but found that we also had to include Sky (or pay TV) subscription to give credible results. The credibility of the matching process left little room for maneuver in this area, when the sizes of the different groups are taken into account. Constrained statistical matching (Barr and Turner, 1990) was used within these cohorts, because it directly preserves marginal distributions and has several other advantages over unconstrained matching (Rodgers, 1984).

A weighted Manhattan distance function based on 12 variables was developed. The variables and their weights were chosen based on their strength of association with a range of television viewing measures, and from knowledge of variables commonly used in media planning.

Initially there was substantial leeway regarding how the distance function should be defined, and inconclusive evidence on the best way to define it. We refined the distance function through extensive experimentation and evaluation of the results. This is detailed later in section 4.2, since some discussion of the methods used to evaluate accuracy is needed first.

4. EVALUATION OF RESULTS

4.1 Evaluating Accuracy of Matched Results

To evaluate the accuracy of results from a statistical match, it would be ideal to compare them with the true population figures, or at least an unbiased estimate of these. However the whole point of conducting a statistical match is that these results are not generally available.

One solution to this problem is to collect some information that would usually need to be donated (in our case television viewing) in the recipient survey. Estimates from this restricted single source database do not involve statistical matching. A sub-sample is then taken from this database, and statistically matched to the remainder of the database to derive another set of estimates. By comparing the two sets of estimates, the accuracy of the matched estimates can be measured (Rodgers, 1984; Roberts, 1994). This was carried out using data from the New Zealand readership survey, with encouraging results. However, this approach can only investigate a restricted range of television viewing information due to questionnaire length constraints. The reliability of the television viewing information collected is also open to question.

Another approach was also taken to address these concerns. Some product usage information was collected in a self-completion questionnaire routinely filled out by television meter panelists. Although this approach could not generate the volume of self-completion data required for the Panorama service, it did provide a firm foundation for an evaluation of the accuracy of matched results. The results from this database were compared with matched results, and a regression analysis was conducted to estimate the level of regression to the mean. This showed a fairly strong association between the two sets of results, although there was some room for improvement (Reilly and Thwaites, 1996; Jephcott and Bock, 1998).

4.2 Improving Accuracy by Refining the Distance Function

After the evaluation studies outlined above, a program of experiments was planned to improve the accuracy of the matched results. This included running tests on different distance functions according to a resolution III experimental design, to identify the best matching variables for inclusion in the distance function. Sixty-three variables, previously selected from 176 possible candidates, were included in this experiment. The accuracy of the database resulting from each statistical matching run was evaluated as described in the previous paragraph.

Ultimately ten variables were found to have a useful role in the distance function, some of which had not been included in the original distance function. The experiment also yielded some information about the appropriate weights for these variables. Using the resulting distance function has reduced the level of regression to the mean by over 10%.

4.3 Evaluating Stability of Matched Results

Updated Panorama databases are produced each month containing the latest month's television viewing data. If substantial instability existed in these results, this would quickly become evident. It was expected that the matched results would be less stable than the original television ratings, but they should not be substantially more variable. A study was therefore conducted to evaluate the stability of the matched results.

The variation in ratings from one month to the next month was calculated for both the matched and the original data, across a range of programs and target audiences. The magnitude of these differences was modeled as a function the underlying sample size and the program's rating, with particular attention to the extremes of the distribution. The matched results were found to be 1.3 times more variable than the original television ratings (Reilly, 1996). This is mitigated by the fact that the underlying television sample size is very close to the readership sample size for small samples. This level of variability is manageable for many products.

4.4 Comparing Marginal Distributions

Another vital check on the matched results is whether they closely reproduce the original results. This is especially important in the media research context, where the results are an industry currency used to set rates for buying and selling media space. Small changes can have a direct monetary impact. The results of an initial check also highlighted the inconsistency of analysis systems, discussed further in section 5.3, and provided a useful check on the proposed solutions. This check is now carried out each month when a new Panorama database is created.

5. PREREQUISITES FOR MATCHING

5.1 Aligning Populations

A basic prerequisite for the statistical matching of data from two surveys is that the surveys should cover the same population, or that the populations should be aligned before matching takes place (similar to step B of Ruggles *et al*, 1977). In New Zealand, the television ratings survey covers all people aged 5 or more who live in households with television sets. (Most off-shore islands are excluded; this also applies to the national readership survey.) The readership survey includes households without television sets, which make up slightly over 2% of all households, but only covers people aged 10 or more. Television viewing data for children aged 5 to 9 is omitted, and a dummy record is included that imputes nil viewing in-home to people living in households without television sets. Note that the television ratings include only in-home viewing, and do not include out-of-home viewing such as viewing of sports games in bars.) The matched database then covers all people aged 10 or more.

5.2 Harmonized Definitions for Matching Variables

Due to a long history of cooperation between managers of the two surveys, all of the matching variables used had nearly identical definitions and classifications. There were differences in data collection methodology for some questions (for example self-completion questionnaire versus personal interview), but this did not appear to cause any major differences in the distributions of most matching variables. The one important exception was the amount of time spent viewing television. This is gathered through three questions in a personal interview for the readership survey, and derived from meter records for the television panel. These two approaches give noticeably different results. This was dealt with by using quantiles of the two distributions instead of the raw values (suggested in Ruggles *et al*, 1977, p. 422).

5.3 Analysis Methods and Systems

The above two prerequisites have been discussed in the literature, and did not cause any significant difficulties in developing Panorama. This section discusses another prerequisite that did cause some difficulty, and which does not seem to be addressed in other papers on statistical matching or data fusion (perhaps because this situation is atypical). This is the need for consistent analysis methods.

Consistency is certainly required between the original analysis systems and the new system for analyzing merged data; this relates back to section 4.4. When merging two media research surveys, as was done for

Panorama, a further requirement is added. Since the required analyses on the two component databases are basically identical, the analysis methods used for them must also be consistent. This was not the case here.

This problem may have been exacerbated by an early decision to use our existing software as the analysis package and delivery platform for Panorama. The software package (called MicroScope) allowed flexible analysis of data from cross-sectional surveys, but was not designed to handle longitudinal data like the television viewing information. For instance, many panelists did not provide valid data on every day during a month, and the official ratings system incorporates different weights for each day to accommodate this. MicroScope could not handle this, so the television viewing data had to be re-weighted and imputed to provide a rectangular database.

Another issue was how reach and frequency analyses should be handled. These are effectively simple static micro-simulation models that show how well an advertising performs, and how much it costs. Reach and frequency analyses are handled quite differently for each of the component databases, for the understandable reason that quite different data is available from the two surveys. However this meant that the available models had to be extended to handle both types of data, and to provide results that closely reproduce the original results. Including appropriate checks on the results, this was a fairly substantial undertaking.

6. CONCLUSIONS

6.1 Suggestions for Similar Projects

As other authors have noted, the quality of statistical matching results must be tested, not taken for granted. Although internal diagnostics are important, the matched results deserve more consideration than they often receive. Accuracy is often not directly addressed, and the author is not aware of any articles that demonstrate the stability of matched results. By addressing these aspects directly when testing Panorama results, leverage was gained for further improvement of the matched results.

Matching databases can require harmonization of analysis techniques as well as definitions. It is important to ensure that the analysis systems can handle data from each survey in an appropriate way, both as separate surveys and when combined.

ACKNOWLEDGEMENTS

The author would like to acknowledge the support of many colleagues, including Brian Milnes, Andrew Whitney, Nick Jones, Mike Frisby, Jonathan Jephcott, Mandy Spencer, James Armstrong, Robin Ransom and David O'Neill, as well as the direct contributions of Sharon Thwaites, Phillip Pilkington, Justene Metcalfe and Robert Gentleman at various stages during the development of the Panorama service. This paper was produced with the support of ACNielsen (NZ) Ltd. However the opinions expressed are the author's.

REFERENCES

Assael, H., and Poltrack, D.F. (1994). Can demographic profiles of heavy users serve as a surrogate for purchase behavior in selecting TV programs? *Journal of Advertising Research*, 34(1), 11-17.

Baker, K., Harris, P. and O'Brien, J. (1997). Data fusion: an appraisal and experimental evaluation. *Journal of the Market Research Society*, 39(1), 225-271.

Barr, R.S.and Turner, J.S. (1990). Quality issues and evidence in statistical file merging. pp. 245-313 in *Data*

Quality Control: Theory and Pragmatics, eds. Liepins, G.E. and Uppuluri, V.R.R. : Marcel Dekker, New York.

Brown, M. (1991). Enhancing media survey value through data fusion. *Proceedings of the ESOMAR Congress, Luxembourg, 1991.*

Dow, H.F. (1989). Data fusion – the Canadian experiment. *Television Research – International Symposium, Tarrytown, N.Y.*

Jephcott, J. and Bock, T. (1998). The application and validation of data fusion. *Journal of the Market Research Society*, 40(3), 185-205.

Reilly, J. (1996). Stability of Panorama results; initial findings. *AGB McNair (NZ) Technical Report.*

Reilly, J., and Thwaites, S. (1996). Accuracy of Panorama results. *AGB McNair (NZ) Technical Report.*

Roberts, A. (1994). Media exposure and consumer purchasing – an improved data fusion technique. *Marketing and Research Today*, 22(3), 159-172.

Rodgers, W.L. (1984). An evaluation of statistical matching. *Journal of Business and Economic Statistics*, 2(1), 91-102.

Rothman, J. (1988). Testing data fusion. *ESOMAR Seminar on Media and Media Research, Madrid.*

Ruggles, N., Ruggles, R. and Woolf, E. (1977). Merging microdata: rationale, practice and testing. *Annals of Economic and Social Measurement*, 6(4), 407-428.

Weigand, J. (1986). Combining different media surveys: the German partnership model and fusion experiments. *Journal of the Market Research Society*, 28(2), 189-208.

Woolfson, M., Gribble, S., Bordt, M., Murphy, B., and Rowe, G. (1989). The Social Policy Simulation Database and Model: an example of survey and administrative data integration. *Survey of Current Business*, 69, 36-41.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

FUSION OF DATA AND ESTIMATION BY ENTROPY MAXIMIZATION

Michael Wiedenbeck[1]

ABSTRACT

Data fusion as discussed here means to create a set of data on not jointly observed variables from two different sources. Suppose for instance that observations are available for (X,Z) on a set of individuals and for (Y,Z) on a different set of individuals. Each of X, Y and Z may be a vector variable. The main purpose is to gain insight into the joint distribution of (X,Y) using Z as a so-called matching variable. At first however, it is attempted to recover as much information as possible on the joint distribution of (X,Y,Z) from the distinct sets of data. Such fusions can only be done at the cost of implementing some distributional properties for the fused data. These are conditional independencies given the matching variables. Fused data are typically discussed from the point of view of how appropriate this underlying assumption is.

Here we give a different perspective. We formulate the problem as follows: how can distributions be estimated in situations when only observations from certain marginal distributions are available. It can be solved by applying the maximum entropy criterium. We show in particular that data created by fusing different sources can be interpreted as a special case of this situation. Thus, we derive the needed assumption of conditional independence as a consequence of the type of data available.

Key words: data fusion, statistical matching, conditional independence, maximum entropy estimation

1. INTRODUCTION

In social science research combining data from different sources is mainly applied in situations where the linked units belong to different levels of a hierarchy of observational units. For instance, individual data are supplemented by data of regional aggregates in a natural way. Combining data of such different levels of aggregation requires to use appropriate statistical models, like multilevel models.

Data fusion means combining data of different samples of one population, i.e. the observational units are of the same level of aggregation. Those units are linked which agree with respect to a set of variables contained in both data sets, the fusion variables or fusion criterium. In applications fusion of data is often an asymetric operation: one of the data sets, the recipient sample, gets information from the other, the donor sample. The donor sample therefore has to be large enough, i.e. for each unit of the recipient sample a corresponding unit with an agreeing set of values for the fusion variables has to be available. In a strict sense, this will be impossible in many situations, because in the case of large sets of fusion variables or interval scaled variables there will be hardly one single pair with two perfectly agreeing profiles, i.e. ordered set of values, of fusion variables. This obstacle is usually removed by introducing distances or similarity measures for the fusion variables in conjunction with rules for multiple use of units of the donor sample. This is important for actually fusing data but not for the theoretical aspects considered here. There are in general many alternative fusion procedures which are also not considered here in detail. A summary of fusion procedures is given for instance in Rodgers (1984).

The key point in our approach is to derive conditional independence as a property of the fused data. The conditional independence has a special meaning concerning the average relative frequencies with which the combined values of two not jointly observed variables occur in those sets of samples created by data fusion. The details are given in Proposition 1 on conditional exchangeability, given the fusion criterium (see Appendix).

[1] ZUMA Center for Survey Research and Methodology, B2,1 , Postfach 1221 55 D-68072 Mannheim Germany, e-mail wiedenbeck@zuma-mannheim.de

2. CONDITIONAL INDEPENDENCE AND INTERPRETATION OF FUSED DATA

Mostly, fusion of data is applied when an alternative single source data set cannot be sampled for practical reasons. Therefore, comparisons of results based on single source data with results based on fused data are mainly available within the framework of methodological investigations. They are usually limited to small sets of variables.

In marketing research a typical question for planning the timing of commercial breaks is: which combinations of certain comsumption behavior occur frequently with which habits of watching TV ? Because rather detailed knowledge is needed, the collection of all the relevant information according to a single source design is often impossible. Instead a design with two sources can be applied where one sample contains data of consumption behavior and the other sample the data of TV-habits.

A justification for data fusion is that identity of units of single source data can be substituted by equality of units in fused data, where equality is achieved only within certain limits of accuracy. But as mentioned before, the appropriateness of fused data depends on the degree by which the true common distribution of consumption behavior and TV-habits can be approximated by a distribution satisfying conditional independence given the fusion criterium. Fusion yields data which are close to a sample where all variables are jointly observed, only if the true common distribution itself satisfies the condition of conditional independence. A judgement of this is usually based only on plausibility considerations. Proposition 1 in the Appendix helps to formulate them as follows. When the individuals are conceived as classified according to the fusion criterium, the conditional independence has a special meaning: within those classes the observations of one of the never jointly observed variables can be exchanged arbitrarily between individuals. In the consumption/TV-habits example, for instance, some socio-demographic and life-style variables are used as fusion variables. If then for empirical or theoretical reasons, no association between the consumption behavior and the habits of watching TV is known or thought to be plausible within classes, then an arbitrary or random combination of consumption and TV habits would deliver descriptions appropriate for real individuals. We then could feel justified to adopt conditional independence as a working hypothesis, and it would make sense to collect data from different sources to perform data fusion.

3. MAXIMUM ENTROPY ESTIMATION

As mentioned before fusion of data is motivated by the idea that "identity" can be substituted by "equality" according to fusion variables. The objective of data fusion can be viewed as the analysis of an unknown common distribution [X,Y] of two multivariate variables X and Y. The data are samples from distributions [X,Z] and [Y,Z], where Z is a multivariate covariate. Usual procedures of estimation of [X,Y] cannot be applied without making strong assumptions on [X,Y] as interactions between X and Y are not directly represented by the data and cannot be identified. But there is some additional information contained in the samples of [X,Z] and [Y,Z]. If we want to use this information we have to consider the common distribution of X,Y and Z jointly, i.e. [X,Y,Z]. In general, this information, however, will not fill out the gap completely. For, if we take any estimation [X,Y,Z]^ of [X,Y,Z], there will in general be many alternative estimates which fit the data, i.e. the samples of [X,Z] and [Y,Z], with the same accuracy as [X,Y,Z]^. For example, if we think of the common distribution of a set of categorical variables: in general, marginal distributions do not determine the full distribution, and hence samples from marginal distributions are not sufficient for a unique estimation.

For the uniqueness of a solution we need restrictions. In case we don't have any prior knowledge about [X,Y,Z] itself, restrictions of a general form are appropriate which do not implement any information not contained in the data. Therefore it is plausible to choose among all estimates which fit the data the one containing the least amount of information. In statistical theory the classical entropy measure of Shannon, also known as Kullback-Leibler distance, is often applied as a measure of information. As an example (Golan et al., 1996, p. 1) can be mentioned, who treat estimation based on the maximum entropy criterium

in situations „when the underlying sampling model is incompletely or incorrectly known and the data are limited, partial or incomplete.".

When we apply Kullback-Leibler distance to the described situation, we can show the following (see Proposition 2 in the Appendix): given any estimates [X,Z]^ and [Y,Z]^ of [X,Z] and [Y,Z] where the derived estimates of the marginal distribution [Z] are equal, there exists always a unique estimate [X,Y,Z]* , for which the marginal distributions of (X,Z) and (Y,Z) coincide with [X,Z]^ and [Y,Z]^. Furthermore, for the bivariate distribution of (X,Y), given Z, the pair is conditionally independent.

This opens a new view at samples created by fusion of data. Because fusion implements conditional independence given Z, fused data can be interpreted as samples of [X,Y,Z]* which has [X,Z]^ and [Y,Z]^ as bivariate marginals of (X,Z) respectively (Y,Z). The maximum entropy solution [X,Y,Z]* is of minimum information, and therefore [X,Y]* is of minimum information, too. Therefore, the results of an analysis of [X,Y]* which is essentially equivalent to an analysis based on the fused data will use all information contained in the given data, in the sense of the entropy information measure.

4. CONCLUDING REMARKS

In general, an empirical validation of data fusion for large scale data is nearly impossible. It has been performed only for very small numbers of components of vector-valued variables X, Y and Z, which can be jointly observed. The full set of parameters of the original distributions [X,Y,Z] or [X,Y] can be compared with the corresponding parameters of - estimated - distributions of the fused data only if they are known in advance, i.e. by simulation studies, or if they can be estimated fom samples of [X,Y,Z] or [X,Y], which are not available for those types of variables where data fusion is typically applied. Raessler and Fleischer (1998) have performed simulation studies and discuss the effects of data fusion on estimates of covariances and correlations between X and Y when [X,Y,Z] is trivariate normal or lognormal. One of their main findings is that in the case of a normal distribution of (X,Y,Z) the covariance of X and Y can be estimated without bias if and only if the expected value of the conditional covariance of X and Y given Z equals 0. This is a special case of the result presented here.

REFERENCES

Golan, Amos; Judge, George and Miller, Douglas (1996): Maximum Entropy Econometrics. Robust Estimation with Limited Data. John Wiley, New York.

Luenberger, David G.(1969): Optimization by Vector Space Methods. John Wiley, New York.

Raessler, Susanne and Fleischer, Karlheinz (1998): Aspects Concerning Data Fusion Techniques. In: Nonresponse in Survey Research, ZUMA-Nachrichten Spezial Nr. 4, eds.Koch, A. and Porst, R., Mannheim: ZUMA, 317-333.

Rogers, Willard L. (1984): An Evaluation of Statistical Matching. Journal of Business & Economic Statistics 2, 91-102.

Wiedenbeck, M. (1999, forthcoming): On Conditional Independence of Distributions Fitting Fused Data Samples. ZUMA-Arbeitsbericht, Mannheim: ZUMA.

Appendix[2]

Here we state definitions and the propositions mentioned in the text.

Definition 1: (Invariance with respect to a fusion criterium)
Be $G = \{i = 1,2,3, \ldots, N\}$ a finite population, $x = (x_i)_{i=1,\ldots,N}$, $y = (y_i)_{i=1,\ldots,N}$ and $z = (z_i)_{i=1,\ldots,N}$ be vectors of observations of random variables X,Y and Z.
Let $H_G(x,y)$ be the common relative frequency distribution of x and y in G, and π a permutation of G and πx the vector of x-observations with permuted indices.
Then
a) π is called z-invariant, if $\pi z = z$;
b) $H_G(x,y)$ is called z-invariant, if $H_G(x,y) = H_G(x,\pi y)$

Definition 2: (Conditional independence as reflected in a bivariate distribution)
Be $H_G(x,y)$ as before, $H_G(z)$ the vector of relative frequencies of z and $H_G(x|z)$ and $H_G(y|z)$ the relative frequencies of x and y within the strata defined by z.
$H_G(x,y)$ is called conditionally independent given z, if $H_G(x,y) = \Sigma_z H_G(x|z) H_G(y|z)^T H_G(z)$, where the sum runs over all values of z.

Proposition 1 (Equivalence of conditional independence and conditional exchangeability)
$H_G(x,y)$ is z-invariant, if and only if $H_G(x,y)$ is conditionally independent given z.

Proposition 2 (Existence and uniqueness of probability densities of multivariate random variables, which maximize the entropy measure under the restriction of fixed marginal distributions for two sets of components, whose intersection is non-empty and whose union is the full set of components)
Let X, Y and Z be (multivariate) random variables with bivariate marginal densities $f_1 = f_1(x,z)$ for (X,Z) and $f_2 = f_2(y,z)$, where the marginal densities $f_1(z)$ and $f_2(z)$ coincide almost everywhere (a.e.). Then (1), there exists a unique distribution [X,Y,Z]* of (X,Y,Z) with density $g^* = g^*(x,y,z)$ among all distributions of (X,Y,Z), whose marginal densities of (X,Z) and of (Y,Z) coincide a.e. with f_1 resp. f_2, that has maximal entropy $\Phi(g) = -\int \ln(g^*) dg^* = \max_g -\int \ln(g) dg$, where dg is equal to the probability measure $g(x,y,z)dxdydz$.
Further, (2), for g^* the following conditional independence of X and Y given Z holds

$$g^*(x,y,z) = g^*(x|z)g^*(y|z)g^*(z) \text{ (a.e.).}$$

[2] The propositions of the Appendix are contained in an forthcoming Research Report at ZUMA, (Wiedenbeck, 1999)

SESSION VI

APPLICATIONS IN POPULATION HEALTH

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

SPATIAL STATISTICS AND ENVIRONMENTAL EPIDEMIOLOGY USING ROUTINE DATA

Richard Arnold[1]

ABSTRACT

The effect of the environment on health is of increasing concern, in particular the effects of the release of industrial pollutants into the air, the ground and into water. An assessment of the risks to public health of any particular pollution source is often made using the routine health, demographic and environmental data collected by government agencies. These datasets have important differences in sampling geography and in sampling epochs which affect the epidemiological analyses which draw them together. In the UK, health events are recorded for individuals, giving cause codes, a date of diagnosis or death, and using the unit postcode as a geographical reference. In contrast, small area demographic data are recorded only at thedecennial census, and released as area level data in areas distinct from postcode geography. Environmental exposure data may be available at yet another resolution, depending on the type of exposure and the source of the measurements.

In this paper, we model individual level disease count data using aBayesian hierarchical model allowing for covariates measured (or estimated) at area level, for the purposes of disease mapping. The model requires small area population estimates for non-censal years, area level measures of social deprivation, as well as estimates of environmental exposure. Since (standardised) disease rates are correlated over small scales and are subject to large statistical fluctuations for rare diseases, the model includes a local smoothing structure in which neighbouring areas are smoothed toward a local mean. We discuss the methods used to prepare the data for input to the model and the assumptions that underlie those methods.

KEYWORDS: Epidemiology; Disease mapping; Clustering; Smoothing

1. INTRODUCTION

Small area studies of disease risk seek to describe the associations that may exist between the occurrence of disease and exposure to risk factors which show spatial variation. Although the risk may be from an identified environmental source (for example air pollution, or electromagnetic fields around power lines), demographic and other factors showing significant spatial variation can also usefully be investigated in small area studies.

In this paper we consider small area studies of the type carried out by the UK Small Area Health Statistics Unit (SAHSU). This is a research unit located at Imperial College in London, funded by central government, which uses the national routine registers of health events (e.g. births, deaths, cancers) to test hypotheses about the health effects of environmental risk factors, especially close to industrial installations. Of increasing interest to government and to the funders of health care is disease mapping, where rates and (standardised) relative risks of disease are displayed at small area level. Such maps in principle allow the identification of possible new clusters of disease, and the long term monitoring of the impact on population health of changing demographics and health policy.

The interpretation of disease maps can be problematic. If the areal units are made small enough to resolve the spatial variation in exposure risk, they may be so small that the sampling variability of the disease counts is so large that no significant excess risk can be detected. This difficulty is frequently encountered in rural areas where the population density is low. Worse, a map of raw small area relative risks will be

[1] Richard Arnold, Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College School of Medicine, Norfolk Place, London W2 1PG, UK.

easily misunderstood by the public, who will place too much weight on individual areas with relative risks observed to be high just by chance.

One way round this difficulty is to map the significances (p-values) of the relative risk estimate in each small area, rather than the relative risks themselves, making some appropriate correction for multiple testing. This approach may be overly conservative since neighbouring areas are likely to have highly correlated relative risks. A group of, say, 5 neighbouring small areas may individually have excess relative risks observed with low significance, but had the 5 areas been merged into one a highly significant result would have been reported. Some kind of smoothing procedure which takes advantage of this correlation between areas is therefore a vital part of meaningful disease mapping.

2. ROUTINE DATA IN SMALL AREA STUDIES

Three types of data are required for studies of this kind: demographic data, incidence/mortality and exposure. For studies based on routine data the sources for all three types of information are different. The most important single type of demographic data are population counts, by age and sex. In the UK these are collected at the decennial censuses and released at enumeration district (ED) level, a typical ED containing 400-500 people. Other information relevant to health studies available from the census includes measures of rates of migration and deprivation. Deprivation is often quantified in the form of deprivation indices which are weighed combinations of variables such as unemployment and household overcrowding (eg Carstairs and Morris 1991). Census data can be supplemented by population counts (normally at the large scale level of local authority administrative district) for non-censal years, and other ongoing health and social surveys.

In the UK routine disease incidence data is collected by the providers of healthcare (e.g. hospital admissions), by central government (the national registers of births and deaths), and by regional registries (e.g. cancers and congenital malformations). The postcode of residence of the person at diagnosis or death provides the spatial reference, and appropriate cause codes identify the disease. The reliability of data of this kind is variable. Where there is a statutory requirement to register (e.g. births and deaths) the registers will be substantially complete, although coding errors occur and postcodes may be missing or incorrect. Where there is no compulsion to register a disease (e.g. cancer) significant rates of incompleteness or duplication occur (especially where registration data are collected from a number of different sources). This may lead to large spatial variations in reported counts independent of the true incidence (e.g. Best and Wakefield 1999). Moreover, variations in coding practices, surgical specialities, physical accessibility, public image and the overlapping of catchment areas may all affect the reported cause-specific rates of admission in a given institution.

Data on exposure to risk factors come in a variety of forms such as water quality, traffic flows, electro-magnetic field strengths and industrial chemical releases. Simple measures of exposure, such as distance from a polluting chimney (e.g. Dolk et al. 1999), are commonly used in the absence of other information. In air pollution studies atmospheric dispersion models, combining meteorological and topological data, can provide a more appropriate measure of exposure.

3. COMBINATION OF DATA

For a small area study a common geography for all three types of data must be established, the most appropriate being based on (aggregations of) census EDs. For non-censal years the small area populations can be estimated in various ways. The simplest is to interpolate the small area counts between two censuses and then be rescale the results so that the small area populations correctly sum up to the annual population estimates at local authority district level. Alternatively proxy measures of population change (e.g. school and electoral rolls) can be used to make regression estimates (Simpson et al. 1997).

The incidence data can be connected to census data by postcode lookup tables produced at the censuses. A postcode covers a smaller area than an ED, although postcodes do not in general nest exactly inside EDs.

Exposure data is often referenced by geographical coordinates, which can be linked to census geography through the digitised boundaries of census areas. Some kind of modelling or extrapolation may be required since exposure data may be sparse, collected at just a few locations, aggregated over large areas, or only collected for a limited time period. It is also necessary to assume that all individuals in the smallest geographical unit of the study are equally exposed, which is a long recognised source of bias in environmental studies. A model must be adopted to connect the exposure measure to disease risk: commonly a simple log-linear model, where additive changes in the exposure measure cause multiplicative changes in disease risk. If exposure data are sparse it may be that only a simple binary classification of areas into exposed/unexposed is justified.

An additional complication is that of population migration which tends to lower observed disease rates and make associations more difficult to detect. In areas with high migration rates exposed individuals at elevated risk of a disease move away before being diagnosed, and the population is constantly refreshed with low risk unexposed immigrants. Both effects result in fewer diagnoses in the high risk area. Where the latency period between exposure and disease is short this effect is mitigated. However for a disease with a long latency period study power can be severely affected (Polissar 1980, Armstrong et al. 1996).

4. AN EXAMPLE

As an example, we take the recent study by Maheswaran et al. (1999) and Wakefield and Morris (1999) of the association between protection against acute myocardial infarction and metals in water. Data on calcium (Ca) and magnesium (Mg) concentrations in water were obtained from a water distributing company (North West Water) in a large region in the North West of England. The region includes two large urban centres (Merseyside and Manchester) as well as rural areas, with a total population of 6.1 million in 1991. The Ca and Mg data are repeated measures of concentration throughout various water zones (water supply areas) in the North West of England in the period 1990-1992. Figures 1(a) and (b) show the means of the logarithms of metal concentrations (mg/l) across the 236 water zones in this period. There is overall a 5-fold variation in both Ca and Mg across the map, and the concentrations of the two metals are strongly correlated. Wakefield and Morris (1999) have made a complete analysis of these data, allowing for uncertainty in the measurements of Ca and Mg, and the interested reader should refer to their paper for further detail of that study. Here we use the data for illustrative purposes only.

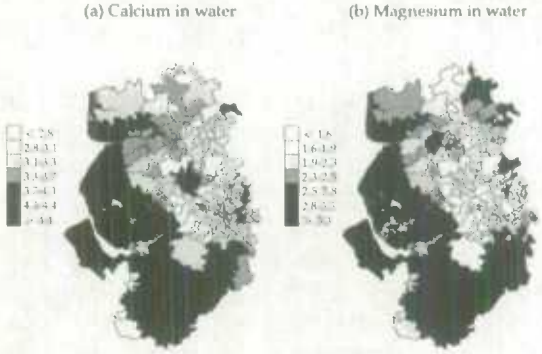


Figure 1. Mean log concentrations (mg/l) Calcium and Magnesium in drinking water in 236 water zones in the North West Water distribution area.

Water zones, each containing 20-25,000 people, are used as the geographical unit of the analysis, i.e. all people within a water zone are considered to be equally exposed. The age and sex standardised mortality

ratios (SMRs) for ischaemic heart disease (IHD) in the period 1990-1992 are shown in figure 2(a). As is typically the case in disease mapping, the map is very heterogeneous: it shows a much greater variability than that expected from a simple Poisson model of disease incidence.

Since disease incidence is often strongly associated with socio-economic deprivation (affluence carrying a lower risk of disease), the relative risks were adjusted using reference rates stratified by quintiles of the Carstairs deprivation index (Carstairs and Morris 1991) as well as by age and sex (Figure 2(b)). This removes some of the spatial variation, and reduces the standard deviation of the relative risks across the map from 0.16 to 0.14. At least some of the variation across the map is due to sampling variability, because in sparsely populated areas the variance of relative risk estimates is large. The remainder of the variation can be thought of deriving from covariates besides age, sex and deprivation. These covariates, measured or unmeasured, will tend to make areas close to one another more similar than areas far apart, and as a result a map of relative risks can be expected to show spatial autocorrelation.

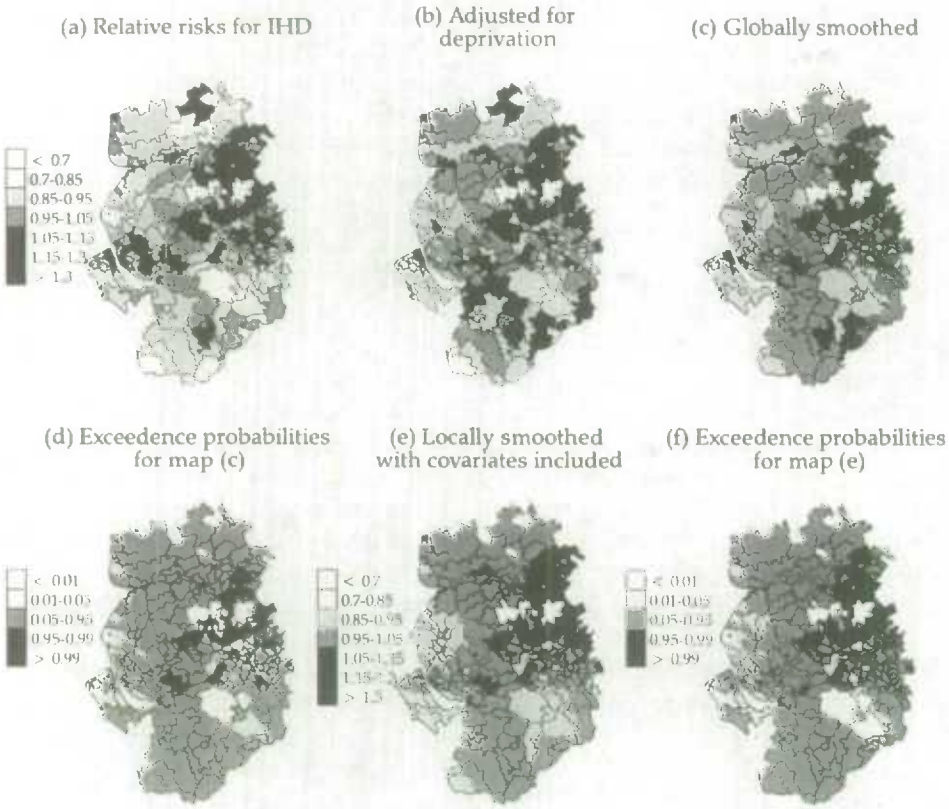


Figure 2. Maps of relative risks and exceedence probabilities (probability that the relative risk exceeds 1) for Ischaemic Heart Disease in the North West. See text for further details.

Various methods of smoothing disease maps are in common use (see e.g. Pickle 1999). These methods can be distinguished according to whether they smooth each area to a global mean (to account for sampling variability), or to some neighbourhood average (to account for spatial correlation). Figure 2(c) shows the effect of smoothing to a global mean using the Empirical Bayes procedure described by Clayton and Kaldor (1987). In this method the relative risks in each area are viewed as random variables drawn from a (gamma) prior distribution, and the smoothing procedure estimates the parameters of that distribution. Areas with low numbers of cases are smoothed more toward the mapwide mean than are highly populated areas. For IHD the standard deviation of the relative risks across the map drops to 0.065 after being smoothed. One of the advantages of taking a Bayesian approach to the problem of disease mapping is that one obtains the posterior probability distribution of the modelled parameters, such as relative risks, after a model has been fitted. We are therefore able to calculate easily interpreted quantities such as the

exceedence probability: the probability that the relative risk in a given area exceeds unity. The exceedence probabilities for IHD are mapped in figure 2(d), and provide a suitable summary of the strength of our belief that the risk in any particular area is raised. As such a map of exceedence probabilities may be preferred as a substitute for (or supplement to) a map of relative risks.

In this paper we wish to allow for spatial correlation, and employ a Bayesian hierarchical model introduced by Besag, York and Mollie (1991). The model is formulated to include separately spatial correlation and heterogeneity. Briefly, the disease counts Y_{ij} in age-sex-deprivation stratum j in area i are modelled as a Poisson distribution $Y_{ij} \sim \text{Poisson}(\varsigma_i E_{ij})$, where E_{ij} are the expected counts and ς_i is the area level relative risk. The relative risks are then modelled by a log-linear function of the area level covariates $\mathbf{X}_i$

$$\log(\varsigma_i) = \vartheta + \mathbf{K}^T\mathbf{X}_i + \theta_i,$$

with an additional random effect term θ_i. This term is the residual log relative risk in area i after we have controlled for measured covariates $\mathbf{X}_i$, and is the sum of two terms: a component u_i to model spatial correlation, and an uncorrelated component v_i to account for excess heterogeneity. The uncorrelated effects v_i follow a normal distribution with mean zero and variance α_v^2. The spatially correlated component u_i in each area i also follows a normal distribution, but with mean and variance conditional on the values of u_j in the neighbouring areas:

$$u_i | u_{j \neq i} \sim \text{Normal}\left(\Sigma_j w_{ij} u_j, \alpha_i^2\right).$$

The set of weights w_{ij} define the neighbourhood relationship between areas i and j. This relationship can be specified in terms of adjacency (w_{ij}=1 if i and j share a boundary, and is zero otherwise), or some function of distance. Other functional and distributional forms are also possible, though they must satisfy a set of regularity conditions (Besag, York and Mollie 1991, see also Best et al. 1999 for a comparison of alternative model specifications). Global spatial trends (e.g. with latitude) can be fitted as measured covariates $\mathbf{X}_i$ separate from the stochastic spatial component.

We have implemented the model in the Bayesian WinBUGS software (Spiegelhalter et al. 1996). We assume vague prior distributions for the coefficients ϑ, K and for the parameters controlling the distributions of u_i and v_i (Best et al. 1999). As a comparison, we also fit log-linear Poisson models with $\log(\varsigma_i) = \vartheta + \mathbf{K}^T\mathbf{X}_i$ allowing for overdispersion. The results are shown in table 1 where the fitted values of the coefficients are given. The data are indeed highly overdispersed. Neither Ca nor Mg were found to have a significant protective effect in this study (Maheswaran et al. 1999). However with or without the fitted covariates the model attributes on average 60% of the variation of the residual relative risks to the spatial component u_i.

The residual relative risks $\exp(\theta_i)$ and are shown in Figures 2(e) and 2(f), in the spatial model with Ca and Mg are included as covariates. The standard deviation of the relative risks in Figure 2(e) is 0.065, the same as the variability across the globally smoothed map in Figure 2(c).

Table 1: Fitted values and 95% confidence intervals for the relative risks (RR=exp(β)) in anoverdispersed Poisson model and the hierarchical spatial model. The coefficient for Ca refers to the change in relative risk for a 2-fold increase in Calcium concentration. The coefficient for Mg refers to a 2.2-fold change.

		Overdispersion or *Spatial Fraction*	RR (Calcium) (95% interval)	RR (Magnesium) (95% interval)
Overdispersed Poisson model	No covariates	2.58		
	+Ca+Mg	2.40	0.96 (0.93,0.99)	1.02 (0.99,1.04)
Hierarchical spatial model	No covariates	*61%*		
	+Ca+Mg	*62%*	0.98 (0.95,1.01)	1.01 (0.98,1.04)

We also simulated a second dataset based on the true IHD counts in the area, including a dependence of the counts on latitude mimicking the overall trend in water hardness across the region. Figure 3(a) shows the underlying relative risk and Figure 3(b) shows the simulated observed relative risks. The globally and locally smoothed residual risks for these data are displayed in Figures 3(c) and 3(d), without including any covariates in the spatial model. The locally smoothed map retains and clarifies the trend with latitude

present in the simulated data, whereas the globally smoothed map is more heterogeneous. The difference between the maps is clearest in the north where low population rural areas in Figure 3(c) are strongly smoothed down to the mapwide mean.

In summary, disease maps can give great insight into small area epidemiological studies which combine routine demographic, incidence and exposure data. However the interpretation of these maps can be obscured by sampling variability and spatially correlated covariates. Bayesian hierarchical models provide a flexible means of including measured covariates and an appropriate local smoothing structure.

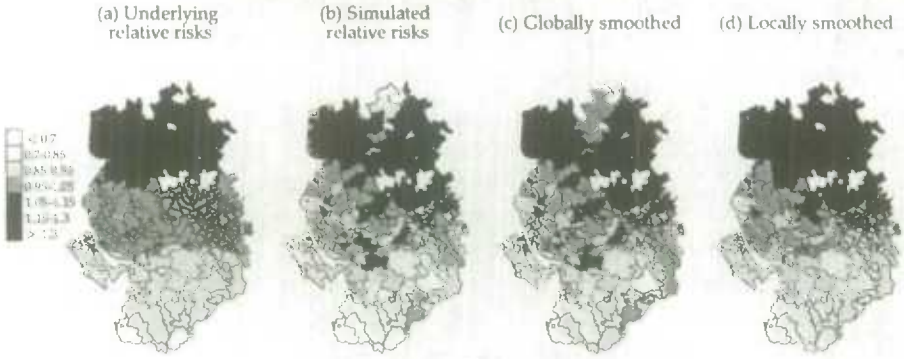


Figure 3. Relative risks for simulated IHD data.

ACKNOWLEDGEMENTS

This work uses data provided with the support of the ESRC and JISC and uses census and boundary material which are copyright of the Crown, the Post Office and the ED-LINE Consortium. This work is supported by ESRC ALCDII award H519255036.

REFERENCES

Armstrong, B.G., Gleave, S, and Wilkinson, P. (1996). The impact of migration on disease rates in areas with previous environmental exposures. *Epidemiology (Suppl.)*, 7, S88

Besag, J., York., J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-59

Best, N. and Wakefield, J., (1999) Accounting for inaccuracies in population counts and case registration in cancer mapping studies. *Journal of the Royal Statistical Society. Series A*. Under revision.

Best, N., Waller, L., Thomas, A., Conlon, E., and Arnold, R. (1999). Bayesian models for spatially correlated disease and exposure data. *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. eds. Oxford: Oxford University Press.

Carstairs, V., and Morris R. (1991). *Deprivation and health in Scotland.* Aberdeen: Aberdeen University Press.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671-681.

Dolk, H., Thakrar, B., Walls, P., Landon, M., Grundy, C., Saez Lloret, I., Wilkinson, P., and Elliott, P. (1999). Mortality among residents near cokeworks in Great Britain. *Occupational and Environmental Medicine*, 56, 34-40.

Maheswaran, R., Morris, S., Falconer, S., Grossinho, A., Perry, I., Wakefield, J., and Elliott, P. (1999). Magnesium in drinking water supplies and mortality from acute myocardial infarction in North West England. *Heart*. Submitted.

Polissar, L., (1980). The effect of migration on comparison of disease rates in geographic studies in the United States. *American Journal of Epidemiology*, 111, 175-182

Pickle, L.W., (1999). Mapping Mortality Data. Disease and Exposure Mapping, Elliott, P., Wakefield, J., Best., N., and Briggs, D. eds. Oxford: Oxford University Press.

Simpson, S., Diamond, I., Middleton, L., and Lunn, D. (1997). Methods for making small area population estimates in Britain. *International Journal of Population Geography*, 3, 265-280

Spiegelhalter, D.J., Thomas, A., Best., N.G., and Gilks, W.R., (1996). BUGS: Bayesian inference using Gibbs Sampling. MRC Biostatistics Unit, Cambridge.

Wakefield, J., and Morris, S., (1999). Spatial dependence and errors-in-variables in environmental epidemiology: investigating the relationship between water constituents and heart disease. *Bayesian Statistics 6*, Bernardo, J.M., Berger, J.O., Dawid, A.P., and Smith, A.F.M. eds. Oxford: Oxford University Press.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 99

FACTORS ASSOCIATED WITH NURSING HOME ENTRY FOR ELDERS IN MANITOBA, CANADA

Monica Tomiak[1], Jean-Marie Berthelot[1], Eric Guimond[1], Cameron A. Mustard[2]

ABSTRACT

Objectives: As the population ages, a greater demand for long-term care services and in particular, nursing homes, is expected. Policy analysts continue to search for altemate, less costly forms of care for the elderly and have attempted to develop programs to delay or prevent nursing home entry. Health careadministrator require information on future demand for nursing home services. This study assesses the relative importance of predisposing, enabling and need characteristics in predicting and understanding nursing home entry.
Methods: Proportional hazard models, incorporating changes in needs over time, are used to estimate the hazard of nursing home entry over a five-year period, using health andsociodemographic characteristics of a representative sample of elderly from Manitoba, Canada.
Results: After age, need factors have the greatest impact on nursing home entry. Specific medical conditions have at least as great a contribution as functional limitations. The presence of a spouse significantly reduces the hazard of entry for males only.
Conclusions: The results suggest that the greatest gains in preventing or delaying nursing home entry can be achieved through intervention programs targeted al specific medical conditions such as alzheimer, musculoskeletal disorders and stroke .
Key words: elderly, nursing home entry, proportional hazard models, chronic conditions, time-varying covariates.

1. INTRODUCTION

As the population ages, a greater demand for long-term care services is expected. Nursing homes have traditionally been the most commonly used form of long-term care. Policy analysts continue to search for alternate, less costly forms of care for the elderly and much research attention has been devoted to developing risk profiles for nursing home admission.[1, 2, 3] Risk profiles are useful for projecting future demands for nursing home care and for developing and targeting programs to delay or prevent nursing home entry. This study contributes to the ongoing research by assessing the relative importance of health and sociodemographic factors in predicting and understanding nursing home entry for a representative sample of elderly residents of the Canadian province of Manitoba.

Many of the previous studies examining factors related to nursing home entry [4, 5, 6] have used Andersen's conceptual framework which considers the use of health services to be a function of an individual's predisposing, enabling and need characteristics. [7] Predisposing factors include demographics, social structure and health beliefs. Enabling factors are those influencing an individual's ability to gain access to health services and include family and community resources. Need factors refer to the functional and health problems which generate the need for health care services. This study assesses the relative importance of predisposing, enabling and need characteristics in predicting and understanding nursing home entry.

[1] Health Analysis and Modelling Group, Statistics Canada
[2] Manitoba Centre for Health Policy and Evaluation

2. DATA AND METHODS

The data used in this study is taken from a unique prototype database created through the collaboration of Statistics Canada, the Government of Manitoba and the University of Manitoba. The database combines information on longitudinally linked individual encounters with the Manitoba health care system through the Manitoba Health Services Insurance Plan (MHSIP) electronic files over a seven-year period (1983-1990), together with detailed demographic, socio-economic and activity limitation information for a sample of individuals taken from the 1986 Census of Population and the 1986 Health and Activity Limitation Survey (HALS). The MHSIP database includes an overall registration file, physician services claims, hospital separations [1], nursing home entry and exit abstracts, and mortality events. Further details are documented in Roos et al. [8]. Individual records from the MHSIP and Statistics Canada files were linked using probabilistic record linkage routines developed by Statistics Canada .The accuracy of the linkage process is estimated to be around 95% for the individuals in the sample. [9] Detailed explanations of the sample design and the record linkage methodology used to create the database are reported elsewhere. [9, 10]

The sample has been shown to provide accurate estimates of mortality rates, number and costs of medical services, and number and duration of short hospital stays, in comparison with those obtained from the entire MHSIP population file. [9] The population studied in this analysis is representative of all individuals aged 65 and above residing in private households in Manitoba on June 3, 1986.

Proportional hazard regression models [11] were used to estimate the hazard of nursing home entry for the five-year time period from June 1986 to June 1991, based on various health and sociodemographic characteristics. Such models are often used to examine the relationship between explanatory variables and survival times, without having to specify any particular form for the distribution of survival times. The dependent variable used in this analysis was the survival time between the start of the study and entry to a nursing home. Observations were censored if either death or the end of the study period occurred before nursing home entry was observed. Two regression models were fitted. The first model, called *Base-Year* model, examines hazard of entry in a nursing home over a five-year period based on predisposing, enabling and need characteristics observed at the start of the study period (June 3 1986). This first model is prospective in that the baseline characteristics are used to predict future nursing home entry. This model in combination with population risk profiles can be used for planning purposes. The second model takes into account changes over time in the need characteristics. Because health status of the elderly population can evolve quite rapidly, it can be argued that when need characteristics are observed only at the start of a five-year study period —the *Base-Year* model—, some characteristics might wrongly be associated or not with nursing home entry. The *Time-Varying Needs* model is assumed to give a more accurate evaluation of associations between contemporaneous need factors and nursing home entry. This model can provide more accurate information to policy analysts in their search for means of delaying or preventing entry into nursing home. For the *Time-Varying Needs* model , needs variables were updated on a yearly basis using the health care records from fiscal year 1986-87 to fiscal year 1989-90. Separate models were fitted for men and women.

3. SELECTION OF INDEPENDENT VARIABLES

In keeping with Andersen's framework, variables were chosen to represent each of the predisposing, enabling and need characteristics groups. Age, marital status and education were included as predisposing characteristics. Enabling factors included household size, an urban/rural indicator, various income measures, home ownership, and the supply of nursing home beds, hospital beds and physicians in the health region of residence. Need factors included functional disability, hospital and physician use during the previous year, excessive comorbidity and 18 specific medical conditions. Since functional disability

---

[1] *Defined as the discharge or death of an inpatient from hospital.*

was only observed on June 3 1986, it was the only need variable not subject to change over time. Summary statistics for these variables are found in Table 1.

Sample weights were used in estimating regression coefficients and in the calculation of summary statistics to make an inference to the Manitoba population. Given that the sample used in this study is based on a complex survey design, analytical weights were used for the estimation of standard errors of the regression coefficients

4. RESULTS

Results from the regression analysis are found in Table 2 and show that there are factors associated (for $p \leq 0.05$) with entry to nursing homes within each of the three categories of variables.

Base-Year model

For both males and females, advanced age is an important predictor of nursing home entry. The only other "predisposing" factors which are significant are marital status for males (RR=1.7 for unmarried) and high education for females (RR=0.4). Several of the "enabling" factors are significantly associated with nursing home entry for females. While home ownership reduces the hazard, living in an urban area increases it. The supply of nursing home beds and physicians in the health care region of residence are also significantly associated with nursing home entry for females. For every additional available nursing home bed (per 1,000 elderly population), the hazard for women increases by 1 % and for every additional physician (per 10,000 population), it decreases by 8 %. The only "enabling" factor associated with entry into a nursing home for males is home ownership (RR=0.4).

Aside from age, the most important variables which are significantly associated with nursing home entry all belong to the "need" factor component. Presence of alzheimer's disease or dementia increases the hazard of nursing home entry by 20.2 times for males and 10.0 times for females. Other specific medical conditions which significantly increase the hazard of nursing home entry are: musculoskeletal disorders (RR=2.8), other mental disorders (RR=2.7) and stroke (RR=2.3) for males, and other mental disorders (RR=1.8) for females. The hazard is significantly decreased for males having arthritis or rheumatism (RR=0.4) or disorders of the digestive system (RR=0.4). For females, the hazard is significantly decreased if they have an history of disorders of the eye or ear (RR=0.7) or of genitourinary disorders (RR=0.6). Turning to more general measures of needs, functional disability is an important predictor variable (RR= 3.2 for males and 1.6 for females). In terms of health care utilization, hospitalization with a stay less than 45 days (RR=1.9) for females and use of more than 5 physician services (RR=1.7) for males are significant.

Time-Varying Needs model

Most predisposing and enabling variables identified with the *Base-Year* model as predictors of nursing home entry remain as such with the *Time-Varying Needs* model: age, education, being a home owner, availability of nursing home beds and physicians. Except for advanced age, risk ratios are of the same magnitude in both models. Marital status for males and urban/rural indicator for females lost their statistical significance while income became significant for females. Women in the third income quartile showed a hazard of nursing home entry 2.1 times greater than women in the lowest income quartile.

The need characteristics associated with nursing home entry are generally the same for both models, but their corresponding risk ratios are in some instances very different. The risk ratio for functionally disabled men went from 3.2 down to 2.8. For females, the previously significant association between disability and nursing home entry disappeared (p=0,07). Hospitalization in the previous year became a much stronger predictor. The risk ratio for hospital stays of less than 45 days increased from 1.9 to 2.7 for females but remained not significant for males. Hospital stays of 45 days or more became significant resulting in a

higher risk of nursing home entry for males (RR=6.5) and females (RR=5.2). Physician utilization, which was only significant for males disappeared. Presence of alzheimer's and dementia have a much lower risk ratio when incorporating dynamic changes to the health status of the elderly population. However, this medical condition remains one of the most important predictor of nursing home entry(males, RR=4.9; females, RR=5.6). Three additional specific medical conditions became significant: musculoskeletal disorders (RR=1.7) for females, ischemic heart disease (RR=0.54) and all other chronic conditions (RR=1.7) for males.

5. DISCUSSION

The *Base-Year* model.

Our findings with the *Base-Year* model, that there are important factors associated with entry to nursing homes in each of the predisposing, enabling and need categories are consistent with the results of many previous studies. The results show that after age, "need" factors are the better predictors of nursing home entry. Other "predisposing" factors and "enabling" factors are of lesser importance. Within the "need" factor category, specific medical conditions have at least as great a contribution as functional limitations.

Marital status results are consistent with previous findings [9] which showed that for males, regardless of the size of their support network, the presence of a spouse was the most important factor in reducing the risk of nursing home admission. The absence of a similar significant association for females may be due to the fact that in a couple the female generally outlives her husband. Eighty-three percent of females aged 85 and over are widowed, compared with thirty-three percent of males. We found no association between the size of the support network and entry to nursing home for both males and females.

In comparing the results for the wealth variables within the "enabling" characteristics category with previous findings, care should be taken in interpreting the results in the Canadian context, where health care is publicly funded. While income does not play the same role in purchasing nursing home care services as it does in the United States, increased wealth can provide the opportunity for purchasing alternate forms of care and for making special adaptations to housing in order to enable individuals to continue residing at home. Although income, per se, was not found to be significant, home ownership significantly reduced the hazard of nursing home entry for both males and females. The lack of association between income and nursing home entry for a population aged 65 and over is not surprising since most individuals have retired by age 65 and income alone should not be used as a sole marker of the availability of material resources.

Higher risk of entry in nursing home for both males and females with alzheimer's, dementia, other mental disorders or musculoskeletal disorders, and males who suffered a stroke were also observed in previous studies. The findings that some medical conditions are associated with a decreased risk of nursing home entry is somewhat counter-intuitive. Two broad explanations come to mind here. First, it is possible that in cases of comorbidity the less severe medical conditions might be slightly under-reported through the medical care system. The consequence of this under-reporting is that the compared sub-populations (e.g., population with arthritis and population without arthritis) are not mutually exclusive therefore causing errors in the calculation of the risk ratios. This is thought to be the case especially with arthritis/rheumatism, which results are nonetheless consistent with the findings of previous studies. A study done by Houle and al (1997) using the 1994-95 Canadian National Population Health Survey found that arthritis was negatively associated with the presence of an individual in nursing homes. Yelin and Katz (1990), using the Longitudinal Study on Age from the United States, found that "persons with arthritis alone had the same rate of institutionalization in a nursing home as did persons without chronic conditions". A second possible explanation is that some medical conditions like ischemic heart disease do not necessarily lead to functional disability which is associated with nursing home entry.

Since comorbidity has been documented to have an important impact on disability [12], the model presented in this study included this variable in addition to specific medical conditions, but to no avail. The lack of association might again be caused by an under-reporting of the less severe medical conditions, resulting in some under-reporting of comorbidity. This can also be resulting from the inclusion of a large number of specific medical conditions in the model. Since no significant association was found for the number of chronic conditions, the model was re-fitted removing these variables to assess the possibility of collinearity effects. The regression coefficients for the 18 medical conditions did not change appreciably, indicating the stability of the results.

The Time-Varying Needs model

The *Time-Varying Needs* model allow us to better understand the impact of changes in health needs on entry to nursing homes. The impact of predisposing and enabling factors were marginally affected by the use of the *Time-Varying Needs* model. This is an interesting result, since it supports the assumption of Andersen's model that needs factors are independent of enabling and predisposing factors. The changes observed in need characteristics are quite important. The effect of functional disability and age at the beginning of the follow-up are reduced, probably due to the changes over time in the chronic conditions. Chronic conditions with long latency have their risk ratio reduced. This is the case for alzheimer's and dementia. Hospital utilization in the year prior to entry to nursing become much more important.

6. CONCLUSION

The base year model presented in this paper can be quite useful for health care planners. This model provides risk ratio of entry to nursing home over the next five year. This information can be used in conjunction with population profiles and average rate of entry to nursing home to help planning needs (e.g.: number of beds) over the next five year. The time-varying needs model provides a more accurate representation of the impact of changes in needs that trigger institutionalization. The findings that certain medical conditions are associated with an increased hazard of nursing home entry can be useful for prevention purposes in order to target interventions to reduce the risk of acquiring specific chronic conditions and for minimizing their disability impact when present. The impact of social support provided by families and friends seems to play a smaller role in entry to nursing home. This study reinforces the need for collecting comprehensive longitudinal data on an on-going basis in order to better understand the impact of changes in predisposing, enabling and needs factors which may occur after baseline.

Table 1. Summary Statistics for Manitoba Residents Aged 65+ on June 3, 1986

		MALES			FEMALES		
		%	% Entering Nursing Home	Sample Size	%	% Entering Nursing Home	Sample Size
PREDISPOSING FACTORS :							
Age	65-74	64.2	2.2	1460	60.7	1.5	1750
	75-84	30.5	7.6	693	31.9	10.4	919
	85+	5.3	31.2	119	7.4	24.4	212
Education	(lowest) Quartile 1	24.1	6.5	548	22.3	9.3	643
	Quartile 2	26.2	5.1	595	27.4	6.5	790
	Quartile 3	23.9	5.5	543	24.2	5.6	696
	(highest) Quartile 4	25.8	4.4	586	26.1	3.1	753
Marital Status	Not married	22.1	10.0	503	57.3	8.0	1651
	Married	77.9	4.0	1769	42.7	3.4	1230
ENABLING FACTORS :							
Rural	Urban	73.7	5.5	1673	78.7	6.1	2268
	Rural	26.4	5.0	599	21.3	5.7	613
Income	(lowest) Quartile 1	24.9	8.3	566	23.7	7.1	682
	Quartile 2	26.0	5.7	591	25.9	7.0	746
	Quartile 3	24.0	4.3	546	24.8	5.5	714
	(highest) Quartile 4	25.0	3.1	569	25.7	4.5	739
Private Pension Income	Yes	40.3	3.6	917	21.9	4.2	630
	No	59.7	6.6	1356	78.2	6.5	2251
Investment Income	Yes	63.7	4.8	1446	58.0	5.6	1671
	No	36.4	6.4	826	42.0	6.6	1210
Income above Low Income Cut-Off	Yes	85.5	4.5	1943	74.0	4.7	2133
	No	14.5	10.6	329	26.0	9.6	748
Home Owner	Yes	77.9	3.7	1769	64.1	4.1	1846
	No	22.1	11.3	503	35.9	9.5	1036
NEED FACTORS :							
Functional Disability [1]	Yes	32.8	9.7	745	34.3	10.5	988
	No	67.2	3.3	1527	65.7	3.7	1893
Hospitalization in Past Year	0 days	82.0	5.1	1863	84.2	4.9	2424
	1-44 days	16.5	6.7	375	14.6	11.9	420
	45+ days	1.5	7.4	34	1.3	10.6	37
Physician Services in Past Year	<5	36.8	3.4	836	32.8	4.4	944
	5 or more	63.2	6.5	1436	67.2	6.8	1937
Number of Chronic Conditions [2]	<5	88.0	4.9	1999	87.6	5.4	2523
	5 or more	12.0	8.7	273	12.4	10.3	358
Specific Medical Conditions							
Arthritis/Rheumatism		15.7	4.7	357	18.3	6.6	528
Musculoskeletal disorders		6.9	10.3	156	9.5	9.9	275
Disorders of eye and ear		30.8	7.1	700	35.1	6.2	1012
Ischemic Heart Disease		18.2	6.1	413	13.6	9.3	391
Diseases of Circulatory System		41.3	7.1	939	48.7	7.5	1402
Chronic Obstructive Pulmonary Disease		9.4	4.1	214	7.9	6.0	227
Disorders of Respiratory System		13.3	5.5	301	7.9	5.5	229
Diabetes		8.8	9.7	200	7.7	8.8	220
Cancer		7.6	4.5	172	5.7	6.2	164
Alzheimer's Disease and Dementia		0.5	46.6	11	0.7	39.5	19
Other Mental Disorders		8.8	12.7	200	12.2	11.2	352
Other Disorders of Nervous System		3.7	13.0	84	4.1	6.3	119
Disorders of Digestive System		20.7	4.4	469	18.8	8.7	543
Disorders of Genitourinary System		16.7	7.3	380	14.7	5.7	424
All Other Chronic Conditions		25.7	6.1	584	28.6	8.6	822
Stroke (ICD9-CM=430-438)		5.5	16.7	124	3.6	12.2	105
Hip Fracture (ICD9-CM=820 or 835)		0.4	13.5	9	1.2	19.0	33
Other Fractures (ICD9-CM=800-848)		5.3	7.6	121	9.1	8.8	262

1. Answered yes to Census disability question (Appendix A)
2. Defined in Appendix B

Table 2. Proportional Hazard Regression Models for Nursing Home Entry, June 1986 - June 1991.

INDEPENDENT VARIABLE		RISK RATIOS+ (RR) FOR HAZARDOF NURSING HOME ENTRY							
		BASE-YEAR MODEL				TIME-VARYING NEEDS MODEL			
		MALES		FEMALES		MALES		FEMALES	
		RR	p	RR	p	RR	p	RR	p
PREDISPOSING FACTORS :									
Age	65-74	(1.00)		(1.00)		(1.00)		(1.00)	
	75-84	**3.56**	<0.001	**6.30**	<0.001	**3.63**	<0.001	**4.79**	<0.001
	85+	**18.55**	<0.001	**14.68**	<0.001	**21.31**	<0.001	**10.60**	<0.001
Education	(lowest) Quartile 1	(1.00)		(1.00)		(1.00)		(1.00)	
	Quartile 2	0.96	0.884	0.73	0.120	1.10	0.719	0.80	0.264
	Quartile 3	1.43	0.193	0.74	0.177	1.35	0.282	0.86	0.493
	(highest) Quartile 4	1.00	0.992	**0.40**	<0.001	0.95	0.851	**0.47**	0.005
Marital Status	Married	(1.00)		(1.00)		(1.00)		(1.00)	
	Not married	**1.67**	0.047	1.11	0.643	1.50	0.110	1.33	0.196
ENABLING FACTORS :									
Household size (#people)		0.78	0.141	0.94	0.545	0.73	0.055	0.97	0.772
Rural	Urban	1.36	0.248	**1.57**	0.042	1.27	0.375	1.35	0.183
	Rural	(1.00)		(1.00)		(1.00)		(1.00)	
Income	(lowest) Quartile 1	(1.00)		(1.00)		(1.00)		(1.00)	
	Quartile 2	0.85	0.571	1.42	0.175	1.00	0.992	1.51	0.126
	Quartile 3	0.92	0.812	1.49	0.244	1.21	0.580	**2.07**	0.038
	Quartile 4	0.60	0.205	1.31	0.491	0.74	0.450	1.57	0.259
Private Pension Income	Yes	0.90	0.653	0.86	0.532	0.93	0.757	0.79	0.312
	No	(1.00)		(1.00)		(1.00)		(1.00)	
Investment Income	Yes	1.16	0.527	0.89	0.520	1.01	0.970	1.02	0.905
	No	(1.00)		(1.00)		(1.00)		(1.00)	
Income above Low Income Cut-Off	Yes	0.88	0.660	0.76	0.313	0.71	0.247	0.69	0.173
	No	(1.00)		(1.00)		(1.00)		(1.00)	
Home Owner	Yes	**0.43**	<0.001	**0.67**	0.026	**0.51**	0.002	**0.67**	0.028
	No	(1.00)		(1.00)		(1.00)		(1.00)	
Nursing Home Beds (per 1,000 elderly)		1.01	0.335	**1.01**	0.039	1.01	0.306	**1.01**	0.048
Hospital Beds (per 1,000)		1.08	0.383	1.04	0.587	1.08	0.394	0.96	0.622
Physicians (per 10,000)		0.95	0.120	**0.92**	0.004	0.97	0.364	**0.95**	0.044
NEED FACTORS :									
Functional Disability	Yes	**3.24**	<0.001	**1.58**	0.008	**2.84**	<0.001	1.36	0.072
	No	(1.00)		(1.00)		(1.00)		(1.00)	
Hospitalization in Past Year	0 days	(1.00)		(1.00)		(1.00)		(1.00)	
	1-44 days	0.65	0.105	**1.85**	0.004	0.89	0.652	**2.65**	<0.001
	45+ days	0.28	0.093	0.96	0.939	**6.47**	<0.001	**5.16**	<0.001
Physician Services in Past Year	<5	(1.00)		(1.00)		(1.00)		(1.00)	
	5 or more	**1.72**	0.039	1.02	0.928	1.34	0.284	1.10	0.709
Number of Chronic Conditions	<5	(1.00)		(1.00)		(1.00)		(1.00)	
	5 or more	1.53	0.307	0.92	0.804	1.64	0.147	1.01	0.986
Specific Medical Conditions									
Arthristis/Rheumatism		**0.41**	0.004	0.76	0.194	**0.49**	0.006	0.94	0.731
Musculoskeletal Disorders		**2.79**	<0.001	1.49	0.085	**2.27**	0.002	**1.70**	0.008
Disorders of eye and ear		1.11	0.628	**0.71**	0.047	1.10	0.641	**0.65**	0.012
Ischemic Heart Disease		0.68	0.145	0.89	0.582	**0.54**	0.010	1.04	0.841
Diseases of Circulatory System		0.92	0.700	1.17	0.378	0.84	0.470	1.37	0.126
Chronic Obstructive Pulmonary Disease		0.74	0.453	0.83	0.545	0.56	0.068	1.14	0.610
Disorders of Respiratory System		0.58	0.107	1.01	0.980	0.91	0.703	0.61	0.079
Diabetes		1.52	0.129	1.30	0.323	1.31	0.328	1.10	0.709
Cancer		0.99	0.978	0.87	0.686	0.61	0.151	0.99	0.966
Alzheimer's Disease and Dementia		**20.21**	<0.001	**10.00**	<0.001	**4.92**	<0.001	**5.62**	<0.001
Other Mental Disorders		**2.73**	<0.001	**1.77**	0.008	**2.97**	<0.001	**2.17**	<0.001
Other Disorders of Nervous System		1.86	0.099	0.74	0.449	1.20	0.555	0.85	0.568
Disorders of Digestive System		**0.35**	<0.001	1.10	0.650	**0.55**	0.009	1.10	0.614
Disorders of Genitourinary System		0.74	0.263	**0.56**	0.019	0.79	0.304	**0.67**	0.036
All Other Chronic Conditions		1.15	0.545	1.41	0.064	**1.66**	0.034	0.91	0.652
Stroke		**2.27**	0.006	0.98	0.956	**2.26**	0.007	1.16	0.584
Hip Fracture		1.89	0.530	1.56	0.326	0.48	0.456	1.05	0.903
Other Fractures		1.10	0.800	1.24	0.389	0.37	0.162	1.59	0.064

+ adjusted for all other variables in the model

Appendix A

Q20 (a) Are you limited in the kind or amount of activity that you can do because of a long-term physical condition, mental condition or health problem:

At home?

No, I am not limited

Yes, I am limited

At school or at work?

No, I am not limited

Yes, I am limited

Not applicable

In other activities, e.g., transportation to or from work, leisure time activities?

No, I am not limited

Yes, I am limited

Q20 (b) Do you have any long-term disabilities or handicaps? No

Yes

Appendix B

Chronic Condition Categories	ICD9-CM Diagnostic Codes
1. Arthritis / Rheumatism	714-715, 721, 725-728
2. Musculoskeletal Disorders(excl Arthritis/Rheumatism)	710, 712-713, 718, 720,722-724, 730-739
3. Disorders of Eye and Ear	360-379, 383-389
4. Ischemic Heart Disease (IHD)	410-414
5. Disorders of Circulatory System (excluding IHD)	393-405, 416, 423-459
6. Chronic Obstructive Pulmonary Disease (COPD)	490-493
7. Disorders of Respiratory System (excluding COPD)	470-478, 494-508, 514-519
8. Diabetes	250
9. Cancer	140-208
10. Alzheimer's / Dementia	290, 331
11. Mental Disorders (excl Dementia)	291-292, 294-299, 300-303,306-307,309-319
12. Disorders of Nervous System (excl Alzheimer's)	326, 330, 332-337, 340-349, 350-356,358-359
13. Disorders of the Digestive System	524-530, 535-537, 550-553,555-558, 562, 564, 565, 568, 569, 571-579
14. Disorders of the Genitourinary System	581-583, 585-589, 591-594, 614-629
15. All Other Chronic Conditions	010-018, 042-044, 137-139, 210-239, 240-249, 251-279, 280-289, 684-686, 694-698,700-709, 740-759, 905-909, E929, E959, E969, E977, E989, E999

REFERENCES

Jette AM, Branch LG, Sleeper LA, Feldman H, Sullivan LM. High-risk profiles for nursing home admission. *The Gerontologist* 1992; 32:634-640.

Shapiro E, Tate R. Who is really at risk for institutionalization? *The Gerontologist* 1988; 28:237-245.

Weissert WG, Cready CM. Toward a model for improved targeting of aged at risk for institutionalization. *Health Services Research* 1989; 24:485-510.

Greene VL, Ondrich JI. Risk Factors for Nursing Home Admissions and Exits: A Discrete-Time Hazard Function Approach. *Journals of Gerontology: Social Sciences* 1990; 45(6):S250-S258.

Wingard DL, Jones DW, Kaplan RM. Institutional Care Utilization by the Elderly: A Critical Review. *The Gerontologist* 1987; 27:156-163.

Wolinsky FD, Callahan CM, Fitzgerald JF, Johnson RJ. The risk of nursing home placement and subsequent death among older adults. *Journals of Gerontology: Social Sciences* 1992; 47:S173-S182.

Andersen RM, McCutcheon A, Aday L, Chiu GY, Bell R. Exploring dimensions of access to medical care. *Health Services Research* 1983; 18:49-73.

Roos LL, Mustard CA, Nicol JP, McLerran DF, Malenka DJ, Young TK, Cohen MM. Registries and Administrative Data: Organization and Accuracy. *Medical Care* 1993; 31: 201-212.

Houle C, Berthelot JM, David P, Mustard C, Roos L, Wolfson MC. Le projet d'appariement du Recensement et des fichiers de soins de santé du Manitoba: Composante des ménages privés. Statistics Canada, Ottawa, November 1995.

Mustard CA, Derksen S, Berthelot JM, Wolfson M, Roos LL. Age-Specific Education and Income Gradients in Morbidity and Mortality in a Canadian Province. In Press, *Social Science and Medicine*.

Cox DR and Oakes D. *Analysis of Survival Data*. New York: Chapman and Hall, 1984.

Verbrugge LM, Lepkowski JM, Imanaka Y. Comorbidity and Its Impact on Disability. *Milbank Q.* 1989; 67(3-4): 450-484.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

COMBINING AGGREGATED SURVEY AND ADMINISTRATIVE DATA TO DETERMINE NEEDS-BASED HEALTH CARE RESOURCE ALLOCATIONS TO GEOGRAPHIC AREAS IN ONTARIO

Vicki Torrance-Rynard[1], Brian Hutchison[1, 2, 3], Jeremiah Hurley[1, 2], Stephen Birch[1, 2], John Eyles[1, 2, 4]

ABSTRACT

A population needs-based health care resource allocation model was developed and applied using age, sex and health status of populations to measure population need for health care in Ontario. To develop the model, provincial data on self-assessed health and health service utilization by age and sex from 62,413 respondents to the 1990 Ontario Health Survey (OHS) were used in combination with provincial health care expenditure data for the fiscal year 1995/96 by age and sex. The model was limited to the services that were covered in the OHS (general practitioner, specialist physician, optometry, physiotherapy, chiropractic and acute hospital). The distribution of utilization and expenditures between age-sex-health status categories was used to establish appropriate health care resource shares for each age-sex-health status combination. These resource shares were then applied to geographic populations using age, sex and health status data from the OHS together with more recent population estimates to determine the needs-based health care resource allocation for each area. Total dollar allocations were restricted to sum to the 1995/96 provincial budget and were compared with 1995/96 allocations to determine the extent to which Ontario allocations are consistent with the relative needs of the area populations.

KEY WORDS: resource allocation, needs-based, health status

1. INTRODUCTION

A basic objective of most publicly funded health care systems is to allocate health care resources among populations according to need (Evans 1996). This objective is consistent with the philosophy underlying the Canada Health Act (Canada House of Commons 1984; Eyles, Birch, Chambers, Hurley and Hutchison 1991; Birch and Chambers 1993; Eyles and Birch 1993). Traditionally, the geographic distribution of health care funding has been based on past allocations and the distribution of health care facilities and providers. Hence the capacity to provide care has not been directly influenced by relative population needs for care. Governments in Canada are increasingly considering or adopting approaches to health care resource allocation based on relative population needs (Nova Scotia Royal Commission on Health Care 1989; Premier's Commission on Future Health Care for Albertans 1989; Saskatchewan Commission on Directions in Health 1990; Integration and Co-ordination Committee, Ontario Premier's Council on Health Strategy 1991; Barer and Stoddart 1991; Comprehensive Health System Planning Commission 1991; British Columbia Royal Commission on Health Care and Costs 1991; Saskatchewan Health 1994).

What is not known is whether the traditional approach to health care resource allocation in Canada has succeeded in distributing resources among populations in keeping with relative needs, despite not being explicitly needs based. We have addressed this question by developing a method to calculate need-based allocations that can be compared to allocations based on the traditional approach. The needs-based approach is *relative* in that the existing level of total health care resources is distributed among

[1] Department of Clinical Epidemiology and Biostatistics
[2] Centre for Health Economics and Policy Analysis
[3] Department of Family Medicine
[4] Department of Geography
McMaster University, HSC-3H1, 1200 Main St. W., Hamilton, ON, Canada, L8N 3Z5

geographically defined populations based on their level of need. For brevity, in the remainder of this paper we use "need" to refer to relative population need for health care.

2. METHODS

2.1 Definition of Need

For computing needs-based allocations, our measure of need was population size adjusted for the distribution of age, sex and self-assessed health status within the population. Variation among populations in age and sex distribution captures a significant portion of variation in need for health care, including preventive and reproductive care (Carter and Peel 1976; Townsend and Davidson 1982; Wilkins and Adams 1983; Waldron 1986; Wyme and Berkman 1986).

Self-assessed health status was used to represent variation in need not accounted for by age and sex. There exists an impressive body of evidence supporting the validity of self-assessed health status as a health status measure. Numerous studies have demonstrated statistically significant relationships, usually of moderate strength, between variants of this single item measure of self-assessed health and other measures of health status (Friedsam and Martin 1963; Maddox and Douglas 1973; LaRue, Bank, Jarvik and Hetland 1979; Linn and Linn 1980; Linn, Hunter and Linn 1980; Tissue 1972; Nagi 1976; Davies and Ware 1981; Fillenbaum 1979; Kaplan and Camacho 1983; Mossey and Shapiro 1982). Idler and Benyamini (1997) recently reviewed 27 longitudinal studies of self-assessed health status and its variants as predictors of mortality and found that self-assessed health status was an independent predictor of mortality in nearly all studies.

2.2 Data Sources

We obtained data on self-assessed health status and self-reported utilization of health care services from the 1990 Ontario Health Survey (OHS). Self-assessed health status was collected as part of the self-complete portion of the OHS (response rate 77.2%) for respondents 12 years of age and over. For children less than 12 years of age we used proxy respondent reports from the personal interview component of the OHS (response rate 87.5%) on the presence or absence of activity-limiting health problems. Complete information for age, sex and health status was available for 62,413 respondents (98.8% of those who responded to both components of the survey).

For this study, we were limited to the categories of health care resources that were examined in the OHS. For health professional services, the interview portion of the OHS contained self-reported information on the number of contacts during the preceding twelve months with general practitioners, specialists, optometrists, physiotherapists and chiropractors. These health professionals account for 97.1% of all Ontario Health Insurance Plan (OHIP) expenditures. For hospital services, the OHS provided the self-reported total number of nights spent in hospital during the preceding twelve months. Acute hospital costs account for 65.7% of the overall operation of hospitals. Together these categories represent 55.7% of the 1995-96 Ministry of Health expenditures for the provision of health care services (Ontario Ministry of Treasury and Economics 1996). We obtained provincial health care expenditures for the fiscal year 1995-96 by age and sex from the Ontario Ministry of Health for the above mentioned categories of health care resources enabling us to compare traditional allocations to needs-based allocations for services that make up over one-half of the total Ministry budget.

2.3 Resource Shares

To use the population distribution of age, sex and health status to allocate resources according to need, we estimated the relationship between these variables and current utilization and expenditures at the provincial level. We assumed that, *at the provincial level of aggregation*, the population in each age-sex-health status category is currently receiving an appropriate *relative share* of health care resources. It is important to note

that our approach did not assume that current allocations to age-sex-health status categories within and among sub-provincial geographic populations are appropriate.

Age, sex and health status-specific resource shares were computed for each category of health care resources. This was done in two steps. First, the age, sex-specific resource shares were calculated as the ratio of the proportion of expenditures accounted for by that age-sex group to the proportion of the population accounted for by that age-sex group. To illustrate, males aged 40-44 years represented 4% of the population but accounted for 3% of the provincial expenditures on general practitioner services. Therefore, the resource share for males aged 40-44 was 0.03/0.04=0.75. The population weighted average resource share for the province was 1.

The second step was to calculate resource shares across health status levels *within* age/sex strata by determining the proportion of the particular health care service used during the preceding 12 months by persons in each of the five levels of self-assessed health. For example, 7% of 40-44 year old males reported fair health but accounted for 13% of all self-reported general practitioner contacts in this age-sex group, so the health status-specific resource share for males aged 40-44 in fair health was 0.13/0.07 = 1.9. Combining this with the previously computed resource share for 40-44 year old males (0.75), resulted in an age, sex, health status-specific share of provincial expenditures for general practitioner services of 1.4 (i.e., 1.9 x 0.75) for males age 40-44 years in fair health. That is, the needs-based share of resources for general practitioner services for a 40-44 year old male in fair health was 1.4 times the provincial per-capita expenditure on general practitioner services. Table 1 shows age, sex, health status-specific resource shares for general practitioner services. Resource shares for each type of service were weighted according to the proportion of the total health care budget that is currently allocated to that service and then summed to compute overall resource shares for each age-sex-health status category.

Population data by age and sex at the census division level were obtained from Statistics Canada CANSIM estimates for July 1, 1995. These data were aggregated to the geographic areas of interest and combined with the OHS data to obtain estimates of age, sex and health status distributions of the geographic areas. The needs-adjusted dollar allocations for each geographic area were calculated by multiplying age, sex and health status specific population estimates for the area by the corresponding resource shares, summing for the area to get total shares, and multiplying by the provincial per-capita expenditure for the health care services included.

2.4 Geographic areas

The OHS was designed "to provide accurate and meaningful information at the Public Health Unit (PHU) level for all major indicators and characteristics" (Ontario Ministry of Health 1992, p. 25). Local areas for this study were defined by the 42 PHUs in existence in Ontario in 1990. These PHUs typically represented regional municipalities or counties. The 1995 population estimates for the PHUs range from 39,832 for Timiskaming to 875,588 for Peel. The local areas can be aggregated into 16 larger administrative areas called health districts which ranged in population from 222,808 for the health district of Muskoka, Nippissing, Parry Sound and Timiskaming to 2,420,054 for Metropolitan Toronto (smaller districts were amalgamated in 1998 to those used in this study). Districts can in turn be aggregated into 6 health planning regions which are the largest administrative area with 1995 populations ranging from 256,547 for North West to 5,015,154 for Central East.

2.5 Current Allocations

From the Ontario Ministry of Health we obtained the actual 1995-96 health care resource allocations to the geographic areas. A critical requirement was that the current allocations be computed on the basis of the place of residence of the recipients of services, rather than on the basis of the location of health care providers and institutions. However, alternate payments for physician services (3.1% of the budget considered here) were only available by location of provider. Expenditure data were linked to an area by

Table 1: Age-sex-health status-specific resource shares for general practitioner services

Sex-Age Group	Measure of Need				
	Limited in Activity due to Health				
Males	Absent	Present			
0-4 years	1.10	2.00			
5-9 years	0.63	1.12			
10-11 years	0.48	1.06			
	Self Assessed Health Status				
	Excellent	Very Good	Good	Fair	Poor
12-14 years	0.43	0.50	0.49	0.49	1.16*
15-19 years	0.45	0.48	0.61	0.58	1.24
20-24 years	0.42	0.49	0.52	0.87	1.07
25-29 years	0.40	0.49	0.62	1.00	1.35
30-34 years	0.41	0.54	0.73	1.28	1.93
35-39 years	0.40	0.56	0.69	2.41	3.99
40-44 years	0.52	0.66	0.86	1.41	2.42
45-49 years	0.42	0.65	0.82	1.63	3.33
50-54 years	0.42	0.68	0.95	1.87	3.61
55-59 years	0.47	0.70	1.12	1.78	3.15
60-64 years	0.55	0.62	1.20	1.70	4.79
65-69 years	0.68	0.91	1.36	2.19	3.28
70-74 years	1.04	1.17	1.52	2.54	2.69
75+ years	1.44	1.66	2.01	3.35	4.68
	Limited in Activity due to Health				
Females	Absent	Present			
0-4 years	1.10	1.73			
5-9 years	0.59	1.91			
10-11 years	0.46	2.23			
	Self Assessed Health Status				
	Excellent	Very Good	Good	Fair	Poor
12-14 years	0.41	0.51	0.49	0.95	3.33
15-19 years	0.60	0.80	0.93	1.06	1.89
20-24 years	0.86	0.85	1.25	1.79	3.34
25-29 years	0.94	0.99	1.45	2.49	2.80
30-34 years	0.92	1.10	1.47	2.12	4.27
35-39 years	0.84	1.02	1.28	1.90	5.89
40-44 years	0.65	0.82	1.31	2.49	5.23
45-49 years	0.55	0.96	1.33	2.86	3.97
50-54 years	0.58	0.79	1.56	2.28	4.52
55-59 years	0.81	0.86	1.36	2.33	3.82
60-64 years	0.43	0.86	1.38	2.32	4.57
65-69 years	0.68	0.98	1.50	2.31	2.92
70-74 years	0.72	1.12	1.68	2.47	2.79
75+ years	1.23	1.75	2.39	3.19	4.30

* imputed as $\frac{\text{gp contacts}_{\text{male,15 - 19, poor health}} \text{ as percent of gp contacts}_{\text{males,15 - 19}}}{\text{pop}_{\text{males,15 - 19, poor health}} \text{ as percent of pop}_{\text{males,15 - 19}}} \times$ per capita resource share for males 12 - 14

the postal code. Recipient addresses were obtained from the hospitalization record for hospital services and from health cards for health professional services. Because the health cards were introduced in 1991/92 with no legal requirement for updates, some geographic classifications may be wrong. The largest impact would be at the smallest geographic level.

3. RESULTS

Table 2 shows the comparison at the local level of needs-based and current allocations. Needs-based allocations ranged from 25.6% higher than current for Northwestern to 18.8% lower for Kingston, Frontenac, Lennox and Addington, with a mean absolute difference of 7.5%. At the larger geographic levels (districts and regions) the percent difference between current and needs-based allocations was smaller. At the district level the mean absolute difference between the allocation methods was 5.4% and at the regional level the mean absolute difference was 2.1%.

Current and needs-based allocations were significantly different for 2 of the 6 regions (33%), 8 of the 16 districts (50%) and 21 of the 42 local areas (50%). We had power to detect a 5% difference in 4 regions (67%), 3 districts (19%) and 3 local areas (7%). We had power to detect a 10% difference in 6 regions (100%), 14 districts (88%) and 17 local areas (40%).

Intraclass correlation coefficients measuring the agreement between needs-based per-capita expenditures and current per-capita expenditures were 0.86, 0.80 and 0.69 for regions, districts and local areas respectively, and were all significantly different from no agreement at the 5% significance level.

4. DISCUSSION

The development of the needs-based allocation method allowed us to establish that current allocations are not fully consistent with relative population needs for health care. However, the high intraclass correlations between needs-based allocations and current per-capita expenditures indicate that in many cases current allocations are aligned with need. Despite this, implementation of needs-based funding would see a substantial redistribution of resources. The potential redistributions are proportionally larger for smaller geographic areas.

The 1990 OHS was the most recent source from which to obtain self-assessed health status at the local level. Consequently, the needs-based resource allocation model would not reflect any subsequent shifts in the distribution of health status among populations. In addition , we are relying on self-reported health care utilization which may not be consistently reported across age, sex and health status categories. The needs-based model is limited in its conclusions to the 55.7% of the total expenditures for the provision of health care services by the Ontario Ministry of Health that we were able to include. We did not adjust the needs-based allocations for differential costs between regions beyond those caused by differences in need. Not all services included in the current allocations were based on place of residence and there may be some misclassification of areas due to out-of-date health cards.

Improvements to this and other health services and population health research projects could be made by a link between population health survey data and health care utilization data. In Ontario, we are currently waiting for access to such a data set from the 1996 National Population Health Survey (Statistics Canada 1998).

Table 2: Relative population (1995) needs-based and current 1995-96 Ministry of Health per-capita allocations *

Region	Local Area	Needs-Based Allocation	Current Allocation	Difference between Needs-Based and Current Allocation	
		Per-capita in dollars (95% CI)	Per-capita in dollars	Per-capita Difference in dollars†	Relative Difference (%)‡
South West	Windsor-Essex	842 (796, 888)	899	-57§	-6.3§
	Kent-Chatham	905 (814, 996)	857	48	5.6
	Middlessex-London	807 (772, 842)	777	30	3.9
	Elgin-St Thomas	845 (754, 936)	837	8	1.0
	Oxford	837 (755, 919)	823	15	1.8
	Bruce-Grey-Owen Sound	857 (787, 926)	899	-42	-4.7
	Sarnia-Lambton	821 (751, 891)	901	-80§	-8.8§
	Huron	849 (753, 945)	939	-90	-9.6
	Perth	816 (734, 897)	721	94§	13.1§
Central West	Niagara	905 (862, 947)	825	80§	9.7§
	Hamilton-Wentworth	862 (827, 897)	956	-94§	-9.9§
	Brant	848 (775, 922)	833	15	1.8
	Wellington-Dufferin-Guelph	810 (758, 862)	712	98§	13.8§
	Halton	737 (705, 770)	747	-9	-1.2
	Haldimand-Norfolk	828 (753, 904)	812	16	2.0
	Waterloo	785 (746, 825)	702	83§	11.9§
Central East	Haliburton Kawartha	933 (863, 1002)	814	119§	14.6§
	Peterborough	847 (778, 916)	869	-22	-2.5
	Durham	779 (741, 818)	752	28	3.7
	Peel	725 (703, 748)	668	57§	8.6§
	East York‖	928 (846, 1011)	1045	-117§	-11.2§
	Etobicoke‖	891 (846, 936)	931	-40	-4.3
	North York‖	865 (833, 896)	988	-123§	-12.5§
	Scarborough‖	833 (800, 866)	821	12	1.5
	Toronto City‖	961 (923, 998)	1027	-66§	-6.4§
	York City‖	1007 (925, 1090)	1004	3	0.3
	York Region	692 (666, 718)	657	36§	5.4§
	Simcoe	838 (795, 881)	757	81§	10.7§
East	Ottawa Carleton	748 (724, 773)	793	-45§	-5.6§
	Leeds, Grenville and Lanark	866 (791, 941)	829	36	4.4
	Eastern Ontario	852 (794, 911)	830	22	2.7
	Hastings Prince Edward	837 (775, 899)	773	64§	8.3§
	Kingston, Frontenac, Lennox and Addington	789 (730, 848)	972	-183§	-18.8§
	Renfrew	947 (852, 1042)	853	95	11.1
North East	Algoma	858 (779, 937)	1012	-155§	-15.3§
	Sudbury	921 (863, 979)	850	71§	8.3§
	Porcupine	880 (788, 971)	910	-30	-3.3
	North Bay	930 (829, 1032)	919	12	1.3
	Timiskaming	959 (798, 1119)	1022	-63	-6.2
	Muskoka-Parry Sound	978 (880, 1077)	847	131§	15.5§
North West	Thunder Bay	862 (795, 928)	892	-30	-3.4
	Northwestern	837 (757, 916)	666	171§	25.6§
	Total	830	830		
	Mean absolute difference			64	7.5

* These allocations include general practitioner, all medical specialists, optometrists, physiotherapists, chiropractors and acute hospital covering 56% of the total expenditures by the Ministry of Health for the provision of health care services in 1995-96.

† Needs-Based – Current based on "unrounded" numbers.

‡ (Needs-Based – Current)/Current based on "unrounded" numbers.

§ Significant at the 5% significance level.

‖ Current data on alternate payments for physician services were not available for the local areas within Metro Toronto. We estimated these data by estimating the Health Service Organization (HSO) program portion of the alternate payments in each local area within Metro Toronto using the number of patients enrolled in HSOs and the average per-capita payment to HSOs. The remainder of the alternate payments were distributed through the local areas according to population.

REFERENCES

Barer, M., Stoddart, G. (1991), *Toward integrated medical resource policies for Canada.* Winnipeg: Manitoba Department of Health, pp. 6B-76-6B-77.

Birch, S., Chambers. S. (1993), "To each according to need: a community-based approach to allocating health care resources," *Canadian Medical Association Journal* 149(5), 607-612.

British Columbia Royal Commission on Health Care and Costs (1991), *Closer to home.* Victoria: Crown Publ; pp. B5-B7.

Canada House of Commons (1984), *Canada Health Act.* Ottawa: Queen's Printer.

Carter, C., Peel J. (1976), *Equalities and inequalities in health.* London: Academic Press.

Comprehensive Health System Planning Commission (1991), *Working together to achieve better health for all.* London, Ontario: Comprehensive Health System Planning Commission, p. 111.

Davies, A.R., Ware, J.E. Jr. (1981), *Measuring health perceptions in the health insurance experiment.* Santa Monica, California: The Rand Corporation.

Evans, R. (1996), "Going for gold: The redistributive agenda behind market-based health care reform," *Health Policy Research Unit Discussion Paper Series.* Centre for Health Services and Policy Research, The University of British Columbia.

Eyles, J., Birch, S., Chambers, S., Hurley, J., Hutchison, B. (1991), "A needs-based methodology for allocating health care resources in Ontario, Canada: Development and an application," *Social Science and Medicine,* 33(4), 489-500.

Eyles, J. Birch, S. (1993), "A population needs-based approach to health-care resource allocation and planning in Ontario: A link between policy goals and practice?" *Canadian Journal of Public Health* 84(2), 112-117.

Fillenbaum, G.G. (1979), "Social context and self-assessments of health among the elderly", *Journal of Health Social Behavior*, 20, 45-51.

Friedsam, H.J., Martin, H.W. (1963), A comparison of self and physician's health ratings in an older population. *Journal of Health and Human Behavior*, 4, 179-183.

Idler, E.L., Benyamini, Y. (1997), "Self-rated health and mortality: a review of twenty-seven community studies," *Journal of Health and Social Behavior*, 38(1), 21-37.

Integration and co-ordination Committee, Ontario Premier's Council on Health Strategy (1991), *Local decision making for health and social services.* Toronto: Premier's Council on Health Strategy, p. 12.

Kaplan, G.A., Camacho, T. (1983), "Perceived health and mortality: a nine-year follow-up of the human population laboratory cohort," *American Journal of Epidemiology*, 117, 292-304.

LaRue , A., Bank, L., Jarvik , L., Hetland, M. (1979), "Health in old age: how do physician ratings and self-ratings compare?" *Journal of Genontology*, 14, 687-691.

Linn, B.S., Linn, M.W. (1980), "Objective and self-assessed health in the old and very old," *Social Science and Medicine*, 14A, 311-315.

Linn, M.W., Hunter, K.I., Linn, B.S. (1980), "Self-assessed health, impairment and disability in Anglo, black and Cuban elderly," *Medical Care*, 18, 282-288.

Maddox, G.L., Douglas, E. (1973), "Self-assessment of health: a longitudinal study of elderly subjects," *Journal of Health and Social Behavior*, 14, 87-93.

Mossey, J.M., Shapiro, E. (1982), "Self-rated health: a predictor of mortality among the elderly," *American Journal of Public Health*;72:800-808.

Nagi, S.Q. (1976), "An epidemiology of disability among adults in the United States," *Milbank Quarterly*, 439-467.

Nova Scotia Royal Commission on Health Care (1989), *Towards a new strategy*. Halifax: Government of Nova Scotia, p. 130-132.

Ontario Ministry of Health (1992*), Ontario Health Survey 1990: User's guide, Documentation*. (Vol. 1). Toronto: Author, p. 25.

Ontario Ministry of Treasury and Economics (1996), *Public Accounts of Ontario 1995-96*. Toronto: Author.

Premier's Commission on Future Health Care for Albertans (1989), *The rainbow report: Our vision for health*. (Vol. II). Edmonton: Government of Alberta, pp. 124-125.

Saskatchewan Commission on Directions in Health (1990), *Future directions for health care in Saskatchewan*. Regina: Government of Saskatchewan, pp. 39-41.

Saskatchewan Health (1994), *Introduction of needs-based allocation of resources to Saskatchewan District Health Boards for 1994-95*. Regina: Saskatchewan Health.

Statistics Canada (1998), *1996/97 National Population Health Survey Microdata File Documentation*. Ottawa: Minister of Industry.

Tissue, T. (1972), "Another look at self-rated health in the elderly," *Journal of Gerontology*, 7, 91-94.

Townsend, P., Davidson, N. (1982), *Inequalities in health: The Black report*. Harmondsworth: Penguin.

Waldron, I. (1986), "Why do women live longer than men?" *The sociology of health and illness*. (2nd ed.) ed. P. Conrad, K. Dern. New York: St. Martin's Press.

Wilkins, R., Adams, O. (1983), *The healthfulness of life*. Montreal: Institute for Research in Public Policy.

Wyme, S., Berkman, L. (1986), "Social class, susceptibility and sickness," *The sociology of health and illness*. (2nd ed.) ed. P. Conrad, K. Dern. New York: St. Martin's Press.

SESSION VII

ROUND TABLE

"COMBINING INFORMATION FROM DIFFERENT SOURCES: HAVE THE STATISTICAL AGENCIES GONE FAR ENOUGH?"

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

ROUND TABLE

COMBINING INFORMATION FROM DIFFERENT SOURCES: HAVE THE STATISTICAL AGENCIES GONE FAR ENOUGH?

Black, C., Manitoba Centre for Health Policy and Evaluation (Canada)

Deville, J.-C., CREST (France)

Péladeau, P., Institut de recherche clinique de Montréal (Canada

Venne, M., Le Devoir, Newspaper (Canada)

SESSION VIII

METHODOLOGICAL ISSUES: ESTIMATION

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

ESTIMATION USING THE GENERALISED WEIGHT SHARE METHOD: THE CASE OF RECORD LINKAGE

Pierre Lavallée and Pierre Caron[1]

ABSTRACT

To augment the amount of available information, data from different sources are increasingly being combined. These databases are often combined using record linkage methods. When there is no unique identifier, a probabilistic linkage is used. In that case, a record on a first file is associated with a probability that is linked to a record on a second file, and then a decision is taken on whether a possible link is a true link or not. This usually requires a non-negligible amount of manual resolution. It might then be legitimate to evaluate if manual resolution can be reduced or even eliminated. This issue is addressed in this paper where one tries to produce an estimate of a total (or a mean) of one population, when using a sample selected from another population linked somehow to the first population. In other words, having two populations linked through probabilistic record linkage, we try to avoid any decision concerning the validity of links and still be able to produce an unbiased estimate for a total of one of the two populations. To achieve this goal, we suggest the use of the Generalised Weight Share Method (GWSM) described by Lavallée (1995).

The paper will first provide a brief overview of record linkage. Secondly, the GWSM will be described. Thirdly, the GWSM will be adapted to provide three different methods where no decision on the validity of the links has to be made. These methods will be: (1) use all possible links; (2) use all possible links above a threshold; and (3) choose the links randomly using Bernoulli trials. For each of the methods, an estimator of a total will be presented together with a proof of design unbiasedness and a variance formula. Finally, to compare the three proposed methods to the classical one where some decision is taken about the links, some simulation results will be presented where it will be seen that using all possible links can improve the precision of the estimates.

KEYWORDS: Generalised Weight Share Method, Record Linkage, Estimation, Clusters.

1. INTRODUCTION

To augment the amount of available information, data from different sources are increasingly being combined. These databases are often combined using record linkage methods. When the files involved have a unique identifier that can be used, the linkage is done directly using the identifier as a matching key. When there is no unique identifier, a probabilistic linkage is used. In that case, a record on the first file is linked to a record on the second file with a certain probability, and then a decision is taken on whether this link is a true link or not. Note that this process usually requires a certain amount of manual resolution.

Manual resolution usually requires a large of amount of resources with respect to time and employees. It might then be legitimate to see if it would be possible to evaluate if manual resolution can be reduced or even eliminated. This issue will be addressed in this paper where one tries to produce an estimate of a total (or a mean) of one population, when using a sample selected from another population linked somewhat to the first population. In other words, having two populations linked through record linkage, we will try to avoid any decision concerning the validity of links, but still be able to produce an unbiased estimate for a total of one of the two populations.

The problem that is considered here is to estimate the total of a characteristic of a population that is naturally divided into clusters. Assuming that the sample is obtained by the selection of units within clusters, if at least one unit of a cluster is selected, then the whole cluster will be interviewed. This usually leads to cost reductions

1 Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 CANADA, plavall@statcan.ca and caropie@statcan.ca

as well as the possibility of producing estimates on the characteristics of both the clusters and the units.

The present paper will show that avoiding deciding on the validity of the links can be achieved using the Generalised Weight Share Method (GWSM) that has been described by Lavallée (1995). This method is an extension of the Weight Share Method presented by Ernst (1989). Although this last method has been developed in the context of longitudinal household surveys, it was shown that the Weight Share Method can be generalised to situations where a population of interest is sampled through the use of a frame which refers to a different population, but linked somehow to the first one.

2. RECORD LINKAGE

The concepts of record linkage were introduced by Newcome *et al.* (1959) and formalised in the mathematical model of Fellegi and Sunter (1969). As described by Barlett *et al.* (1993), *record linkage* is the process of bringing together two or more separately recorded pieces of information pertaining to the same unit (individual or business). Record linkage is sometimes also called *exact matching*, in contrast to *statistical matching*. This last process attempts to link files that have few units in common. Linkages are then based on similar characteristics rather than unique identifying information. In the present paper, we will discuss more the context of record linkage. However, the developed theory could also be used for statistical matching.

Suppose that we have two files A and B containing characteristics related to two populations U^A and U^B, respectively. The two populations are somehow related to each other. The purpose of record linkage is to link the records of the two files A and B. If the records contain unique identifiers, then the matching process is trivial. Unfortunately, often a unique identifier is not available and then the linkage process needs to use some probabilistic approach to decide whether two records of the two files are linked together or not. With this linkage process, the likelihood of a correct match is computed and, based on the magnitude of this likelihood, it is decided whether we have a link or not.

Formally, we consider the product space $\boldsymbol{AXB}$ from the two files A and B. Let j indicates a record (or unit) from file A (or population U^A) and k a record (or unit) from file B (or population U^B). For each pair (j,k) of $\boldsymbol{AXB}$, we compute a linkage weight θ_{jk} reflecting the degree to which the pair (j,k) is likely to be a true link. The higher the linkage weight θ_{jk} is, the more likely the pair (j,k) is a true link. The linkage weight θ_{jk} is commonly based on the ratios of the conditional probabilities of having a match μ and an unmatch $\overline{\mu}$, given the result of the outcome of comparison C_q of the characteristic q of the records j from A and k from B, $q=1,...,Q$.

Once a linkage weight θ_{jk} has been computed for each pair (j,k) of $\boldsymbol{AXB}$, we need to decide whether the linkage weight is sufficiently large to consider the pair (j,k) a link. This is typically done using a decision rule. With the approach of Fellegi and Sunter (1969), we use an upper threshold θ_{High} and a lower threshold θ_{Low} to which each linkage weight θ_{jk} is compared. The decision is made as follows:

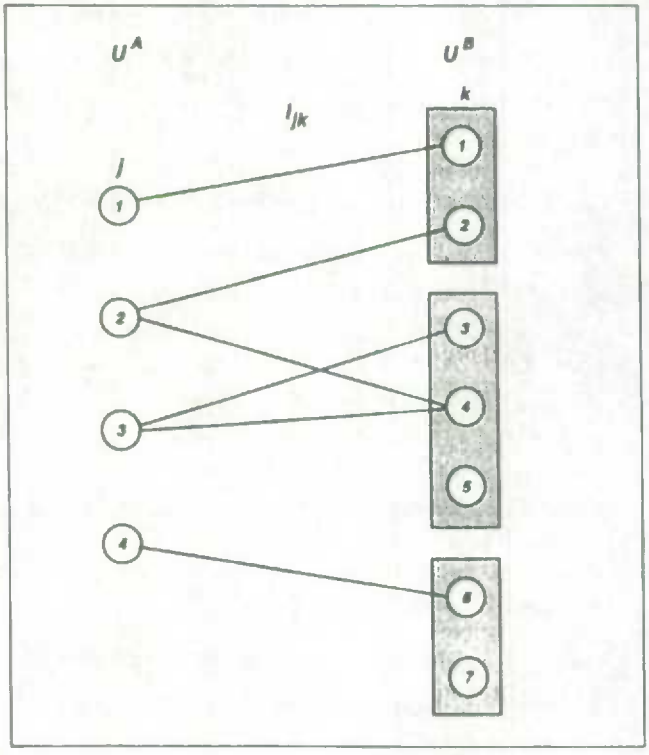
$$D(j,k) = \begin{cases} \text{link} & \text{if } \theta_{jk} \geq \theta_{High} \\ \text{can be a link} & \text{if } \theta_{Low} < \theta_{jk} < \theta_{High} \\ \text{nonlink} & \text{if } \theta_{jk} \leq \theta_{Low} \end{cases} \qquad (2.1)$$

The lower and upper thresholds θ_{Low} and θ_{High} are determined by *a priori* error bounds based on false links and false nonlinks. When applying decision rule (2.1), some clerical decisions will be needed for those linkage weights falling between the lower and upper thresholds. This is generally done by looking at the data, and also by using auxiliary information. By being automated and also by working on a probabilistic basis, some errors could be introduced in the record linkage process. This has been discussed in several papers, namely Barlett

et al. (1993), Belin (1993) and Winkler (1995).

The application of decision rule (2.1) leads to the definition of an indicator variable $l_{jk} = 1$ if the pair *(j,k)* is considered to be a link, 0 otherwise. As for the decisions that need to be taken for those linkage weights falling between the lower and upper thresholds, some manual intervention can be needed to decide on the validity of the links. Note that decision rule (2.1) does not prevent to have many-to-one or one-to-many links. This can be represented by Figure 1.

Figure 1.



3. THE GENERALISED WEIGHT SHARE METHOD

The GWSM is described in Lavallée (1995). It is issued from the Weight Share Method described by Ernst (1989) in the context of longitudinal household surveys. The GWSM can be viewed as a generalisation of *Network Sampling* and also of *Adaptive Cluster Sampling*. These two sampling methods are described in Thompson (1992), and Thompson and Seber (1996).

Suppose that a sample s^A of m^A units is selected from the population U^A of M^A units using some sampling design. Let π_j^A be the selection probability of unit *j*. We assume $\pi_j^A > 0$ for all $j \in U^A$.

Let the population U^B contain M^B units. This population is divided into *N* clusters where cluster *i* contains M_i^B units. From population U^B, we are interested in estimating the total $Y^B = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} y_{ik}$ for some characteristic *y*.

An important constraint that is imposed in the measurement (or interviewing) process is to consider all units within the same cluster. That is, if a unit is selected in the sample, then every units of the cluster containing the selected unit will be interviewed. This constraint is one that often arises in surveys for two reasons: cost reductions and the need for producing estimates on clusters. As an example, for social surveys, there is normally a small marginal cost for interviewing all persons within the household. On the other hand, household estimates are often of interest with respect to poverty measures, for example.

With the GWSM, we make the following assumptions:

1) There exists a link between each unit *j* of population U^A and at least one unit *k* of cluster *i* of population U^B, i.e. $L_j^A = \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik} \geq 1$ for all $j \in U^A$.
2) Each cluster *i* of U^B has at least one link with a unit *j* of U^A.
3) There can be zero, one or more links for a unit *k* of cluster *i* population U^B, i.e. it is possible to have $L_{ik} = \sum_{j \in U^A} l_{j,ik} = 0$ or $L_{ik} = \sum_{j \in U^A} l_{j,ik} > 1$ for some $k \in U^B$.

These situations are illustrated in Figure 1. We will see in Section 4 that in the context of record linkage, some of these assumptions might not be satisfied.

By using the GWSM, we want to assign an estimation weight w_{ik} to each unit k of an interviewed cluster i. To estimate the total Y^B belonging to population U^B, one can then use the estimator

$$\hat{Y} = \sum_{i=1}^{n} \sum_{k=1}^{M_i^B} w_{ik}\, y_{ik} \tag{3.1}$$

where n is the number of interviewed clusters and w_{ik} is the weight attached to unit k of cluster i. With the GWSM, the estimation process uses the sample s^A together with the links existing between U^A and U^B to estimate the total Y^B. The links are in fact used as a bridge to go from population U^A to population U^B, and vice versa.

The GWSM allocates to each sampled unit a final weight established from an average of weights calculated within each cluster i entering into $\hat{Y}$. An *initial weight* that corresponds to the inverse of the selection probability is first obtained for unit k of cluster i of $\hat{Y}$ having a non-zero link with a unit $j \in s^A$. An initial weight of zero is assigned to units not having a link. The *final weight* is obtained by calculating the ratio of the sum of the initial weights for the cluster over the total number of links for that cluster. This final weight is finally assigned to all units within the cluster. Note that the fact of allocating the same estimation weight to all units has the considerable advantage of ensuring consistency of estimates for units and clusters.

Formally, each unit k of cluster i entering into $\hat{Y}$ is assigned an initial weight w'_{ik} as :

$$w'_{ik} = \sum_{j=1}^{M^A} l_{j,ik}\, \frac{t_j}{\pi_j^A} \tag{3.2}$$

where $t_j = 1$ if $j \in s^A$ and 0 otherwise. Note that a unit k having no link with any unit j of U^A has automatically an initial weight of zero. The final weight w_i is given by

$$w_i = \frac{\sum_{k=1}^{M_i^B} w'_{ik}}{\sum_{k=1}^{M_i^B} L_{ik}} \tag{3.3}$$

where $L_{ik} = \sum_{j=1}^{M^A} l_{j,ik}$. The quantity L_{ik} represents the number of links between the units of U^A and the unit k of cluster i of population U^B. The quantity $L_i = \sum_{k=1}^{M_i^B} L_{ik}$ then corresponds to the total number of links present in cluster i. Finally, we assign $w_{ik} = w_i$ for all $k \in i$ and uses equation (3.1) to estimate the total Y^B.

Now, let $z_{ik} = Y_i / L_i$ for all $k \in i$. As shown in Lavallée (1995), $\hat{Y}$ can also be written as

$$\hat{Y} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} \sum_{i=1}^{N} \sum_{k=1}^{M_i^B} l_{j,ik}\, z_{ik} = \sum_{j=1}^{M^A} \frac{t_j}{\pi_j^A} Z_j \tag{3.4}$$

Using this last expression, it can easily be shown that the GWSM is design unbiased. The variance of $\hat{Y}$ is directly given by $Var(\hat{Y}) = \sum_{j=1}^{M^A} \sum_{j'=1}^{M^A} \frac{(\pi_{jj'}^A - \pi_j^A \pi_{j'}^A)}{\pi_j^A \pi_{j'}^A} Z_j Z_{j'}$ where $\pi_{jj'}^A$ is the joint probability of selecting units j and j' (See Särndal, Swensson and Wretman (1992) for the calculation of $\pi_{jj'}^A$ under various sampling designs).

4. THE GWSM AND RECORD LINKAGE

With record linkage, the links $l_{j,ik}$ have been established between files A and B, or population U^A and population U^B, using a probabilistic process. As mentioned before, record linkage uses a decision rule D such as (2.1) to decide whether there is a link or not between unit j from file A and unit k from file B. Once the links are established, we have seen that it is then possible to estimate the total Y^B from population U^B using a sample obtained from population U^A. One could then ask if it is necessary to make such a decision. That is, is it necessary to establish whether there is positively a link for the given pair (j,k), or not? Would it be easier to simply use the linkage weights θ_{jk} (without using any decision rule) to estimate the total Y from U^B using a sample from U^A? If this were the case, it is easy to see that reducing the amount of clerical intervention required in the record linkage process could save time and resources. In the present section, we will see that the GWSM can be used to answer the previous question. Three methods will be considered.

Method 1: Use all possible links with their respective linkage weights

When using all possible links with the GWSM, one wants to give more importance to links that have large linkage weights θ than those that have small linkage weights. We no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not between unit j from U^A and unit k of cluster i from U^B. Instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process. By doing so, we do not need any decision to be taken to establish whether there is a link or not between two units.

By definition, for each pair (j,ik) of $A \times B$, the linkage weight $\theta_{j,ik}$ reflects the degree to which the pair (j,ik) is likely to be a true link. This linkage weight can then directly replace the indicator variable l in equations (3.2) and (3.3) that define the estimation weight obtained through the GWSM. We then get the estimation weight w_{ik}^{RL}.

The assumptions of Section 3 concerning the GWSM still apply. For instance, the existence of a link between each unit j of population U^A and at least one unit k of population U^B is translated into the need of having a non-zero linkage weight $\theta_{j,ik}$ between each unit j of U^A and at least one unit k of cluster i of U^B. The assumption that each cluster i of U^B must have at least one link with a unit j of U^A translates into the need of having for each cluster i of U^B at least one non-zero linkage weight $\theta_{j,ik}$ with a unit j of U^A. Finally, there can be a zero linkage weight $\theta_{j,ik}$ for a unit k of cluster i of population U^B. In theory, the record linkage process does not insure that these constraints are satisfied. This is because the decision rule (2.1) does not prevent to have many-to-one or one-to-many links, or no link at all. For example, it might turn out that for a cluster i of U^B, there is no non-zero linkage weight $\theta_{j,ik}$ with any unit j of U^A. In that case, the estimation weight (4.2) underestimate the total Y^B. To solve this problem, one practical solution is to collapse two clusters in order to get at least one non-zero linkage weight $\theta_{j,ik}$ for cluster i. Unfortunately, this solution might require some manual intervention, which we try to avoid. A better solution is to force to have a link by choosing one link at random within the cluster.

Method 2: Use all possible links above a given threshold

Using all possible links with the GWSM as in Method 1 might require the manipulation of large files of size $M^A \times M^B$. This is because it might turn out that most of the records between files A and B have non-zero linkage weights θ. In practice, even if this happens, we can expect that most of these linkage weights will be relatively small or negligible to the extent that, although non-zero, the links are very unlikely to be true links.

In that case, it might be useful to only consider the links with a linkage weight θ above a given threshold θ_{High}.

For this method, we again no longer use the indicator variable $l_{j,ik}$ identifying whether there is a link or not , but instead, we use the linkage weight $\theta_{j,ik}$ obtained in the first steps of the record linkage process and above the threshold θ_{High}. The linkage weights below the threshold are considered as zeros. We therefore define the linkage weight: $\theta^T_{j,ik} = \theta_{j,ik}$ if $\theta_{j,ik} \geq \theta_{High}$, 0 otherwise. The estimation weight w^{*RLT}_{ik} is then directly obtained by replacing the indicator variable l in equations (3.2) and (3.3) by $\theta^T_{j,ik}$.

The number of zero linkage weights θ^T will be greater than or equal to the number of zero linkage weights θ used for Method 1. Therefore, the assumption of having a non-zero linkage weight $\theta^T_{j,ik}$ between each unit j of U^A and at least one unit k of U^B might be more difficult to satisfy. The assumption that each cluster i of U^B must have at least one non-zero linkage weight $\theta^T_{j,ik}$ with a unit j of U^A can also possibly not be satisfied. In that case, the estimation weight (4.7) underestimate the total Y^B. To solve this problem, one solution is to force the selection of the link with the largest linkage weights θ. This will lead to accepting links with weights θ^T below the threshold. If there is still no link, then choose one link at random within the cluster is a possible solution.

Method 3: Choose the links by random selection

In order to avoid taking a decision on whether there is a link or not between unit j from U^A and unit k of cluster i from U^B, one can decide to simply choose the links at random from the set of possible links. For this, it is reasonable to choose the links with probabilities proportional to the linkage weights θ. This can be achieved by Bernoulli trials where, for each pair *(j,ik)*, we decide on accepting a link of not by generating a random number $u_{j,ik}$ that is compared to the linkage weight $\theta_{j,ik}$.

The first step before performing the Bernoulli trials is to rescale the linkage weights in order to restrict them to the [0,1] interval. This can be done by dividing each linkage weight $\theta_{j,ik}$ by the maximum possible value θ_{Max}. Although in most practical situations, the value θ_{Max} exists, it is not the case in general. When this is not possible, one can then use a transformation such as the inverse logit function $f(x) = e^x/(1+e^x)$ to force the adjusted linkage weights $\tilde{\theta}$ to be in the [0,1] interval. The chosen function should have the desirable property that the adjusted linkage weights $\tilde{\theta}$ sum to the expected total number of links L in ***AXB***, i.e. $\sum_{j=1}^{M^A}\sum_{i=1}^{N}\sum_{k=1}^{M_i^B}\tilde{\theta}_{j,ik} = L$.

Once the adjusted linkage weights $\tilde{\theta}_{j,ik}$ have been obtained, for each pair *(j,ik)*, we generate a random number $u_{j,ik} \sim U(0,1)$. Then, we set the indicator variable $\tilde{l}_{j,ik}$ to 1 if $u_{j,ik} \leq \tilde{\theta}_{j,ik}$, and 0 otherwise. This process provides a set of links similar to the ones used in the original version of the GWSM, with the exception that now the links have been determined randomly instead of through a decision process comparable to (2.1). The estimation weight w^{*RLT}_{ik} is then directly obtained by replacing the indicator variable l in equations (3.2) and (3.3) by $\tilde{l}_{j,ik}$.

By conditioning on the accepted links $\tilde{l}$, it can be shown that estimator (4.11) is conditionally design unbiased and hence, unconditionally design unbiased. Note that by conditioning on $\tilde{l}$, the estimator (4.11) is then equivalent to (3.1). To get the variance of $\hat{\tilde{Y}}$, again conditional arguments need to be used.

With the present method, by randomly selecting the links, it is very likely that one or more of the three assumptions of Section 3 will not be satisfied. For example, for a given unit j of population U^A, there might not be a link with any unit k of population U^B. Again, in practice, we can overcome this problem by forcing a link for all unit j of population U^A by choosing the link that has the highest linkage weight $\theta_{j,ik}$. The constraint that each cluster i of U^B must have has at least one link with a unit j of U^A can also be not satisfied. Again, we can force to have a link by choosing the one with the highest linkage weight $\theta_{j,ik}$. If there is still no link, it is possible to choose one link at random within the cluster. It should be noted that this solution preserves the design unbiasedness of the GWSM.

5. SIMULATION STUDY

A simulation study has been performed to evaluate the proposed methods against the classical approach (Fellegi-Sunter) where the decision rule (2.1) is used to determine the links. This study was made by comparing the design variance obtained for the estimation of a total Y^B using four different methods: (1) use all links; (2) use all links above a threshold; (3) choose links randomly using Bernoulli trials; (4) Fellegi-Sunter. Given that all four methods yield design unbiased estimates of the total Y^B, the quantity of interest for comparing the various methods was the standard error of the estimate, or simply the coefficient of variation (i.e., the ratio of the square root of the variance to the expected value).

For the record linkage step, data from the 1996 Farm Register (population U^A) was linked to the 1996 Unincorporated Revenue Canada Tax File (population U^B). The Farm Register is essentially a list of all records collected during the 1991 Census of Agriculture with all the updates that have occurred since 1991. It contains a farm operator identifier together with some socio-demographic variables related to the farm operators. The 1996 Unincorporated Revenue Canada Tax File contains data on tax filers declaring at least one farming income. It contains a household identifier, a tax filer identifier, and also socio-demographic variables related to the tax files.. For the purpose of the simulations, the province of New Brunswick was considered. For this province, the Farm Register contains 4,930 farm operators while the Tax File contains 5,155 tax filers.

The linkage process used for the simulations was a match using five variables. It was performed using the statement MERGE in SAS®. All records on both files were compared to one another in order to see if a potential match had occurred. The record linkage was performed using the following five key variables common to both sources: (1) first name (modified using NYSIIS); (2) last name (modified using NYSIIS); (3) birth date; (4) street address; (5) postal code. The first name and last name variables were modified using the NYSIIS system. This basically changes the name in phonetic expressions, which in turn increases the chance of finding matches by reducing the probability that a good match is rejected because of a spelling mistake or a typo.

Records that matched on all 5 variables received the highest linkage weight (θ=60). Records that matched on only a subset of at least 2 of the 5 variables received a lower linkage weight (as low as θ=2). Records that did not match on any combination of key variables were not considered as possible links, which is equivalent as having a linkage weight of zero. A total number of 13,787 possible links were found.

Two different thresholds were used for the simulations: $\theta_{High} = \theta_{Low}$ =15 and $\theta_{High} = \theta_{Low}$ =30. The upper and lower thresholds, θ_{High} and θ_{Low}, were set to be the same to avoid the grey area where some manual intervention is needed when applying the decision rule (2.1) of Fellegi-Sunter.

For the simulations, we assumed that the sample from U^A (i.e. the Farm Register) would be selected using Simple Random Sampling Without Replacement (SRSWOR), without any stratification. We also considered two sampling fractions: 30% and 70%. The quantity of interest Y to be estimated is the Total Farming Income. It was possible for us to calculate the theoretical variance for these estimates for various sampling fractions. We could also estimate this variance by simulations (i.e. performing a Monte-Carlo study). Both approaches were used. For the simulations, 500 simple random samples were selected for each method for two different sampling fractions (30% and 70%). The two thresholds (15 and 30) were also used to better understand the properties of the given estimators. The results of the study are presented in Figures 2.1 and 2.2.

By looking at the above figures, it can be seen that in all cases, Method 1 provided the smallest design variances for the estimation of the Total Farming Income. Therefore, using all possible links leads to greatest precision. This results is important since it indicates that, in addition to saving resources by eliminating manual interventions needed to decide on the links, we also gain in the precision of the estimates. This conclusion also seems to hold regardless of the sampling fraction, the threshold and the province.

As previously mentioned, the number of links to handle using Method 1 might be quite large. Therefore, a compromise method such as Method 2 (use all links above a threshold) or Method 3 (choose links randomly using Bernoulli trials) might be appealing. The precision of Method 2 seems to be comparable to Method 4 (Fellegi-Sunter). Using Method 2 can then be chosen because, unlike Method 4, it does not involve any decision rule to decide on the validity of the links.

Figure 2.1. CVs with $\theta_{High} = \theta_{Low}$ =15.

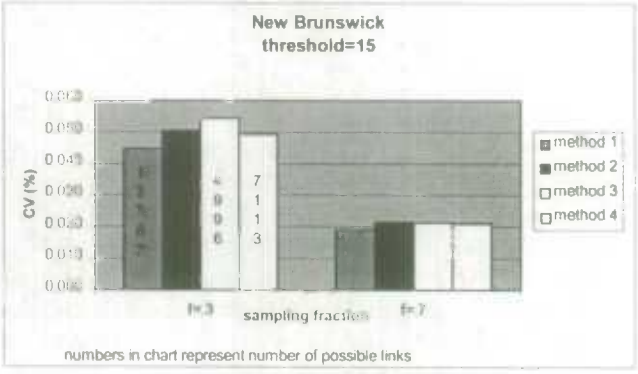
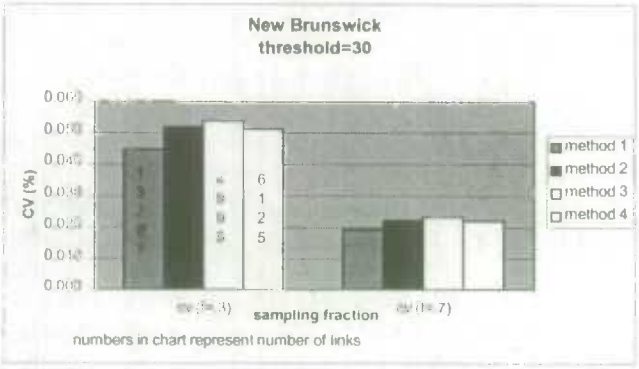


Figure 2.2. CVs with $\theta_{High} = \theta_{Low}$ =30.



Method 3 turned out to have the largest variance. Therefore, choosing the links randomly using Bernoulli trials does not seem to help with respect to precision. However, Method 3 is the one that used the least number of links. If this is of concern, this method can turn out to be appealing in some situations.

REFERENCES

Barlett, S. , Krewski, D., Wang, Y., Zielinski, J.M. (1993). Evaluation of Error Rates in Large Scale Computerized Record Linkage Studies. *Survey Methodology*, Vol. 19, No. 1, pp. 3-12.

Belin, T.R. (1993). Evaluation of Sources of Variation in Record Linkage through a Factorial Experiment. *Survey Methodology*, Vol. 19, No. 1, pp. 13-29.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (Kasprzyk, D., Duncan, G., Kalton, G., Singh, M.P., Editors), John Wiley and Sons, New York, pp. 135-159.

Fellegi, I.P., Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, Vol. 64, pp. 1183-1210.

Lavallée, P. (1995). Cross-sectional Weighting of Longitudinal Surveys of Individuals and Households Using the Weight Share Method, *Survey Methodology*, Vol. 21, No. 1, pp. 25-32.

Newcome, H.B., Kennedy, J.M., Axford, S.J., James, A.P. (1959), Automatic Linkage of Vital Records. *Science*, Vol. 130, pp. 954-959.

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

Thompson, S.K. (1992), *Sampling*. John Wiley and Sons, New York.

Thompson, S.K., Seber, G.A. (1996), *Adaptive Sampling*. John Wiley and Sons, New York.

Winkler, W.E. (1995). Matching and Record Linkage. In *Business Survey Methods* (Cox, Binder, Chinnappa, Colledge and Kott, Editors), John Wiley and Sons, New York, pp. 355-384.

Proceedings of the Statistics Canada Symposium 99
Combining Data from Different Sources
May 99

DUAL SYSTEM ESTIMATION AND THE 2001 CENSUS COVERAGE SURVEYS OF THE UK

James Brown[1], Ian Diamond[1], Ray Chambers[1], and Lisa Buckner[2]

ABSTRACT

The application of dual system estimation (DSE) to matched Census / Post Enumeration Survey (PES) data in order to measure net undercount is well understood (Hogan, 1993). However, this approach has so far not been used to measure net undercount in the UK. The 2001 PES in the UK will use this methodology. This paper presents the general approach to design and estimation for this PES (the 2001 Census Coverage Survey). The estimation combines DSE with standard ratio and regression estimation. A simulation study using census data from the 1991 Census of England and Wales demonstrates that the ratio model is in general more robust than the regression model.

KEY WORDS: Dual System Estimation, Ratio Estimation, Regression Estimation, Census Underenumeration

1. INTRODUCTION

An increasing problem for many countries is what to do when the census, the gold standard for population counts and their distribution at very small areas, suffers from underenumeration. The United Kingdom (UK) has faced this problem for a number of censuses with net underenumeration measured by a post enumeration survey (PES). In 1991, though, the net underenumeration suggested by the PES was rather less than that suggested, at the national level, by demographic estimates. As a result it was not possible to allocate underenumeration below the national level using the PES and a demographic strategy was developed (Diamond, 1993 and Simpson *et al*, 1997). To ensure that this does not occur in 2001 a major research project has been undertaken so that, in 2001, the Office for National Statistics (ONS) will be in a position to **estimate** and **adjust** accurately census outputs for net underenumeration. The ultimate aim is to adjust the actual census database and create a 'One Number Census' (ONC) so that all tabulations reflect the estimated underenumeration and all figures reported by the census are consistent. However, before this can be done, estimates of the population at sub-national administrative areas, the level at which most resource allocation takes place, are required. It is this estimation problem which is the focus of this paper. The overall ONC research is described in Brown *et al* (1999).

1.1 Dual System Estimation

A standard method for estimating underenumeration is Dual System Estimation. This was the approach used by the US Census Bureau following both the 1980 and 1990 US Censuses (Hogan, 1993). Shortly after the census a PES is used to obtain an independent re-count of the population in a sample of areas. Dual system estimation combines these two counts to estimate the true population, allowing for people missed by both the census and the PES, in the PES sample areas. Although the method is theoretically straightforward, in practice it has some problems.

[1] Department of Social Statistics, University of Southampton, Southampton, Hants SO17 1BJ, UK
[2] Census Division, Room 4200W, Office for National Statistics, Titchfield, Fareham, Hants PO15 5RR, UK

a) The DSE assumes that in the target population the matched PES and census counts follow a multinomial distribution. That is, the probabilities of being counted by either or both the PES and the census are **homogeneous** across the target population. This is unlikely for most populations.

b) Unbiased estimation requires statistical **independence** between the census count and the PES count. This is impossible to guarantee.

c) It is necessary to **match** the two data sources to determine whether individuals on the lists were counted once or twice. Errors in matching become biases in the dual system estimator (DSE).

In the 1990 Census the US Census Bureau tackled problem a) by splitting the population up into post strata based on factors (e.g. race) which were thought to affect an individual's probability of being counted, a method originally proposed by Sekar and Deming (1949). Problem b) is typically handled by operational procedures that ensure the operational independence of the census and the PES. Problem c) is essentially unavoidable but it is absolutely essential to ensure that errors due to matching are minimised.

Generalisation of the DSE counts from the sampled PES areas to the whole population can be carried out using a variety of survey estimation methods. In this paper DSE methodology is combined with ratio and regression estimation to achieve this aim. A series of estimators are proposed and the robustness of the ratio and regression models evaluated using a simulation study.

2. POPULATION ESTIMATION USING THE 2001 CENSUS COVERAGE SURVEY

2.1 Design of the Census Coverage Survey (CCS)

In the UK in 2001 the PES will concentrate only on coverage (as opposed to 1991 when both quality and coverage were addressed) using a short questionnaire and large sample size. This will be known as the Census Coverage Survey (CCS). The aim of the CCS is to estimate population totals by age and sex for groups of local administrative areas, called design areas, with total expected populations of around 0.5 million. It will be a two-stage sample design based on a set of auxiliary variables, the 1991 Census counts, for each age-sex group by enumeration district within a design group.

At the first stage a stratified random sample of enumeration districts (EDs) will be drawn within a design area. An ED is an area containing about 200 households and represents the workload for one census enumerator. Within a design area the enumeration districts are stratified by a hard to count (HtC) index with categories d = 1 to D. The index is based on the social, economic, and demographic characteristics associated with people who were considered hard to count in the 1991 Census. At the design stage its role is to ensure that all types of EDs are sampled. Further size stratification of EDs, within each level of the HtC index, based on the 1991 Census counts improves efficiency by reducing within stratum variance.

The second stage of the CCS design will consist of the random selection of a fixed number of small areas called postcodes (containing on average fifteen households) within each selected enumeration district. All households in the selected CCS postcodes will then be independently enumerated using a short personal interview that will ascertain the household structure at the time of the census.

2.2 Models for Population Estimation Using the CCS

After the CCS there will be two population counts for each postcode in the CCS sample, the census count and the CCS count. One approach would be to assume that the CCS count is equal to the true population count in the sampled postcodes and that, therefore, there is no underenumeration in the CCS. However, it is

more sensible to assume that there will be underenumeration in both census and CCS and hence for each sampled postcode both counts are subject to non-response. Under the assumptions already specified the DSE can be used to estimate the true population counts, Y_{aid}, for age-sex group a in postcode i in HtC stratum d. The problem is then how to estimate the overall population total in the design area, T_a, for age-sex group a using this information.

2.2.1 Ratio Model for Population Estimation

The simplest approach to make such population estimates is to assume that information is only available for the census and the CCS in the sample areas. In this situation Alho (1994) proposes an adaptation of the Horvitz-Thompson estimator for T_a. However, census counts are available for all postcodes and can be used as auxiliary information to improve this estimator. The simplest way to introduce these auxiliary data is to assume that the true count is proportional to the census count. For estimation this leads to the classical ratio model for each age-sex group. Dropping the age-sex group indicator a, and representing the census count in postcode i of HtC stratum d by X_{id}, this model can be written as

$$\begin{aligned} &E\{Y_{id} \mid X_{id}\} = R_d X_{id} \\ &Var\{Y_{id} \mid X_{id}\} = \sigma_d^2 X_{id} \\ &Cov\{Y_{id}, Y_{jf} \mid X_{id}, X_{jf}\} = 0 \text{ for all } i \neq j \end{aligned} \tag{1}$$

where R_d and σ_d^2 are unknown parameters. Under the model in (1) it is straightforward to show (Royall, 1970) that the best linear unbiased predictor (BLUP) for the true population total T of an age-sex group is the stratified ratio estimator of this total given by

$$\hat{T}_{RAT} = \sum_{d=1}^{5} \hat{R}_d \sum_{i=1}^{N_d} X_{id} \tag{2}$$

where N_d is the total number of postcodes in HtC stratum d and $\hat{R}_d$ is an estimate of the population ratio of true to census counts. Strictly speaking the assumption in model (1) of zero covariance between postcodes counts is violated as the design of the CCS has the postcodes clustered. However, this is not a problem for estimation of the population total, as (2) remains unbiased when this assumption is violated with only a small loss of efficiency (Scott and Holt, 1982). Typically $\hat{R}_d = \frac{\sum_{S_d} Y_{id}}{\sum_{S_d} X_{id}}$ where S_d represents the sampled postcodes in HtC index d. In practice, of course, the Y_{id} are unknown and replaced by their corresponding DSEs.

An alternative to this estimator is to compute one DSE across all the CCS sample postcodes within a HtC stratum and then 'ratio' this total up to a population estimate for that stratum by multiplying it by the ratio of the overall census count for the stratum to the census count for the CCS postcodes in the stratum. This is analogous to treating the HtC stratum as a post-stratum in the US Census context and applying the ratio estimator proposed by Alho (1994). One would expect this second approach to have a lower variance due to the larger counts contributing to the DSE but be subject to more bias due to heterogeneity of capture probabilities within each HtC stratum. Defining the HtC strata after the census can reduce this bias. However, it appears unlikely that all the necessary data for such a post-stratification will be available in time for such an exercise to be carried out after the 2001 UK Census.

2.2.2 Regression Model for Population Estimation

The model in (1) forces a strictly proportional relationship between the census and true counts. Such a relationship is unlikely to be the case where census counts are close to zero, as will be the situation if

estimation is carried out at the postcode level. Therefore, Brown *et al* (1999) suggested the use of a simple regression model to allow for the situation where the census counts for a particular postcode are close, but not equal to, zero. This model is given by

$$\begin{aligned} &E\{Y_{id} \mid X_{id}\} = \alpha_d + \beta_d X_{id} \\ &Var\{Y_{id} \mid X_{id}\} = \sigma_d^2 \\ &Cov\{Y_{id}, Y_{jf} \mid X_{id}, X_{jf}\} = 0 \text{ for all } i \neq j \end{aligned} \tag{3}$$

Under (3) it is straightforward to show (Royall, 1970) that the best linear unbiased predictor (BLUP) for the true population total T of an age-sex group is then the stratified regression estimator

$$\hat{T}_{REG} = \sum_{d=1}^{5} \sum_{i=1}^{N_d} (\hat{\alpha}_d + \hat{\beta}_d X_{id}) \tag{4}$$

where $\hat{\alpha}_d$ and $\hat{\beta}_d$ are the OLS estimates of α_d and β_d in (3). Like the ratio estimator (2), (4) is robust to the correlation of postcodes due to the sample design (Scott and Holt, 1982). Unfortunately, it is not robust to a large number of zero census / CCS counts, since the fitted regression line can then be significantly influenced by the large number of sample points at the origin.

3. SIMULATION STUDY OF POPULATION ESTIMATION

To assess the performance of the three estimators of the population total described in Section 2.2 when the CCS design described in Section 2.1 is applied to a population a simulation study was undertaken and is described in this section. Anonymised individual records for a local administrative area from the 1991 Census augmented by a HtC index are treated as a design area and used as the basis for the simulation.

3.1 Applying the CCS Design to the Simulation Population

As already stated it is expected that underenumeration in the 2001 Census will be higher in areas characterised by particular social, economic and demographic characteristics. For example, people in dwellings occupied by more than one household (multi-occupancy) have a relatively high probability of not being enumerated in a census. This heterogeneity is accounted by stratifying the enumeration districts (and hence the postcodes contained within them) by a 'Hard to Count' (HtC) index. The "prototype" HtC index used here (see Brown *et al*, 1999 for details) is based on a linear combination of variables including: language difficulty for heads of households in the enumeration district (ED), migration of young adults into the ED, households in multi-occupied buildings, and privately rented households. The distribution of the 930 enumeration districts in the simulation population across the categories of this HtC index is between 150 and 210. The CCS sample design was carried out using size stratified sampling based on the design described in Brown *et al* (1999), leading to a total of 35 EDs (and hence 175 postcodes) being selected.

3.2 Simulating a Census and its CCS

Each individual in the population was assigned a fixed probability of being counted in a census. These probabilities depend on individual characteristics and are based on research by the 'Estimating With Confidence Project' (Simpson *et al*, 1997) following the 1991 Census. In particular, there is considerable variation in the individual probabilities by age and sex. They also vary by HtC index; the census variable 'Primary Activity Last Week'; and there is also a small enumeration district effect. However, the probabilities are still heterogeneous even when all these factors are taken into account. Whole households are also assigned probabilities of being counted in the census. These are based on averaging the individual probabilities associated with the adults within the households. Household probabilities also vary according

to the tenure of the household and the household size. The household and individual probabilities remain fixed throughout the simulation study. Each individual and household is also assigned a factor that defines the differential nature of response in the CCS. These mirror the same pattern as the census probabilities but the differentials are less extreme. This extends the simulation study in Brown *et al* (1999) so that there is heterogeneity in both the census and the CCS for age-sex groups at the postcode level.

To generate a census and its corresponding CCS, independent Bernoulli trials are used to determine first whether the household is counted and second whether the individuals within a counted household are counted. There is also a check that converts a counted household to a missed household if all the adults in the household are missed. In these simulations the census and CCS outcome for households and individuals are independent. Coverage in the CCS is set at approximately 90 per cent for households with 98 per cent of individuals within those households being counted. For each census ten CCS postcode samples are selected. The estimators described in Section 2.2 are then applied to each age-sex group and population totals are calculated. The whole process is repeated for 100 independent censuses.

3.3 Population Estimation Results

For the simulation of 100 censuses the average census coverage is 94.90 per cent. This is rather less than 1991 where overall coverage was around 98 per cent and aims to assess the robustness of the procedure to increased probabilities of underenumeration. The three estimators being evaluated are:

1) The ratio estimator with the DSE at the postcode level (Postcode Ratio)
2) The ratio estimator with the DSE at the HtC index level (Index DSE)
3) The regression estimator with the DSE at the postcode level (Postcode Regression)

As this is a simulation relative root mean square errors (RRMSE) and the relative biases can be used to assess the performance of the estimators relative to each other (and the census) over the 1000 CCSs. For each estimator the RRMSE is defined as:

$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_j - \text{truth}\right)^2} \times 100 \qquad (5)$$

and can be considered as a measure of the total error due to bias and variance. Relative bias is defined as:

$$\text{Relative Bias} = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_j - \text{truth}\right) \times 100 \qquad (6)$$

Bias in an estimator is usually considered a poor feature, as it cannot be estimated from the sample. However, it can be better overall to adopt a slightly biased estimator if its total error is small.

TABLE 1

Performance of the three population estimators for the population total

	Postcode Ratio	Index DSE	Postcode Regression
Relative Bias (%)	0.27	0.26	0.41
RRMSE (%)	0.57	0.55	0.68

Table 1 summarises the results for the estimation of the total population by summing the individual age-sex estimates. There are not tremendous differences between the estimators. However, the two estimators based on the ratio model both have lower bias and lower total error. Looking at the total can hide problems with the estimation of the individual age-sex population totals. Across age-sex groups the three estimators are very similar in terms of RRMSE and are always better than the census with the exception of men aged 85 years and over. With respect to bias Figure 1 shows that the postcode regression estimator has a higher bias for the young age groups of both sexes. This is what gives the postcode regression estimator its high relative bias in Table 1. The two estimators based on the ratio model have a higher relative bias for the

oldest age groups. However, in terms of the population total these are small groups and so do not impact on the results in Table 1. It is also possible to plot the distribution of the individual errors from the 1000 CCSs for each age-sex group. This shows that while the estimators based on the ratio model have smaller inter-quartile ranges for the distributions of their errors they are slightly more prone to outliers.

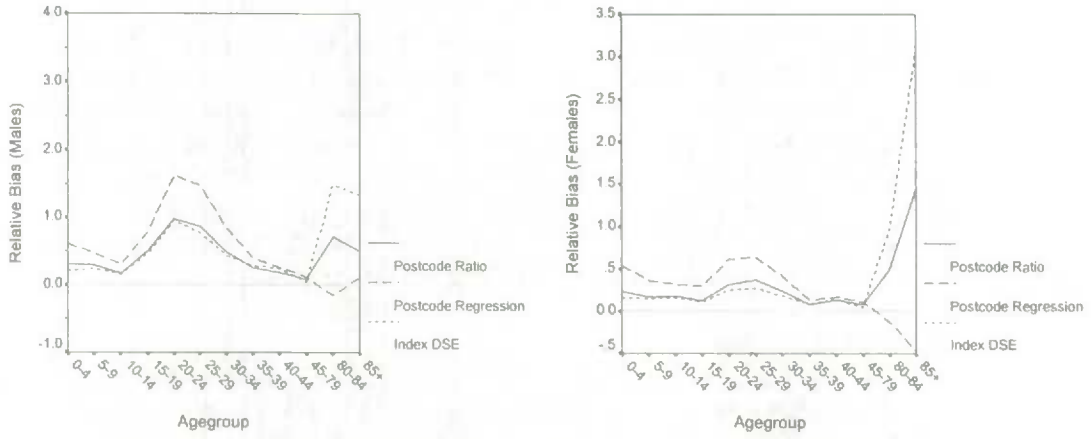


FIGURE 1: Relative Bias of the three population estimators by age and sex compared to the census.

Figure 1 and Table 1 suggest that, in general, the estimators based on the ratio model are better with the Index DSE estimator being 'best' overall. However, this particular estimator needs to be treated with care as it relies heavily on the HtC index defining homogeneous strata. In the simulation this is the case once age and sex are also controlled for. However, in 2001 this assumption will be much harder to defend given that many postcodes that will have changed in ten years. This will cause the DSE calculated at the HtC stratum level to be biased. For the postcode-based estimators this will not impact on the individual DSEs to cause bias but it will increase the variance as the relationship between the census and CCS counts within each stratum will not be as strong.

There are also problems with both the ratio and the regression model as the census count gets small. As stated in Section 2.2 the regression model will fit well when census counts are approaching zero and the CCS is finding extra people but it will not be robust to a large number of origin points. As the postcode is a very small geographic area the count for a particular age-sex group will often be zero. In the simulation about one third of the sampled postcodes counts are at the origin for most age-sex groups. The presence of the points at the origin tends to rotate the fitted line increasing the estimate of the slope leading to the positive bias. The origin points do not affect the ratio model as it is constrained to pass through the origin. However, postcodes where the census count is zero and the CCS is greater than zero do. These happen in a few postcodes for all the age-sex groups. However, their affect is most dramatic in the oldest age groups when there are only a few non-zero census counts and the observed counts are in general all close to zero. It is this that generates the positive bias for these estimators that can be seen in Figure 1.

4. ADDITIONAL CONSIDERATIONS

4.1 Matching Error and the DSE

One of the most challenging aspects of using the DSE is the matching of the census and CCS databases to determine whether individuals have been counted twice or only counted once. Any error at this stage feeds directly into the DSE and becomes bias, the direction of which depends on the nature of the matching error. Current research (Baxter, 1998) is developing an automated matching strategy that minimises the error from matching. It tackles the problem by first using exact computer matching, which is expected to handle

the majority of cases. The remaining individuals are then matched using probability-based methods. Those records that cannot be matched this way will be matched by hand.

4.2 Estimation of Overenumeration in the 2001 Census

All the work presented in this paper has assumed that overenumeration in the UK Censuses is minimal. In previous censuses no attempt has been made to adjust for overenumeration and it has been assumed that careful implementation of the census in the field minimises the risk of it occurring. It is very unlikely that a person or household will complete two forms for the same place. However, overenumeration will occur when a person is erroneously added to a form for a household. Research is being carried out to assess its importance in the UK context. Methods of identifying overenumeration and erroneous census counts using the CCS and the matching exercise are also being investigated.

4.3 Variance Estimation

This paper has concentrated on the performance of the estimators in a simulation study where the truth is known. In 2001 the truth will not be known and therefore the measures of performance will be estimated variances for the estimators. Results for the postcode level regression estimator in Brown *et al* (1999) using the 'Ultimate Cluster Variance Estimator' were very promising. This estimator gave good coverage for estimated confidence intervals in the situation of independence between the census and CCS. However, more work is needed to assess both its robustness and stability.

5. CONCLUSIONS

This paper has presented research that is being undertaken as part of a research project by the Office for National Statistics to estimate and adjust for underenumeration in the 2001 Censuses of the UK. The standard technique for estimating underenumeration, dual system estimation, has been combined with both ratio and regression estimation models. The simulation study, which extends Brown *et al* (1999) to include heterogeneity in the CCS as well as the census, shows that while estimators based on both models perform well there is a robustness issue. This particularly affects the estimator based on the regression model.

The work presented in this paper is ongoing. The next major steps will be the further development of robust estimation models for both the regression and ratio approaches. More work is also needed to more fully assess the effect of increasing heterogeneity on the estimator. Issues relating to overenumeration, and its estimation, along with variance estimation also need to be fully addressed to ensure that a robust and efficient estimation strategy is adopted in 2001.

REFERENCES

Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *J. Off. Statist.* **10**, 245 - 256.

Baxter, J. (1998) One Number Census Matching. *One Number Census Steering Committee Working Papers* **ONC(SC)98/14**. Available from the Office for National Statistics, Titchfield. (jennet.baxter@ons.gov.uk)

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999) A methodological strategy for a One Number Census. *J. R. Statist. Soc. A*, **162**, 247-267.

Diamond, I. D. (1993) Where and who are the 'missing million'? Measuring census of population undercount. In *Statistics Users' Council Conference Proceedings on Regional and Local Statistics, 16th November 1993*. Published by IMAC Research, Esher.

Hogan, H. (1993) The 1990 post-enumeration survey: operations and results. *J. Am. Statist. Ass.*, **88**, 1047-1060.

Royall, R. M. (1970) On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.

Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *J. Am. Statist. Ass.*, **77**, 848-854.

Sekar, C. C. and Deming W. E. (1949) On a method of estimating birth and death rates and the extent of registration. *J. Am. Statist. Ass.*, **44**, 101-115.

Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends*, **90**, 31-39.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

SIMULTANEOUS CALIBRATION OF SEVERAL SURVEYS

Jean-Claude Deville[1]

ABSTRACT

Often, the same information is gathered almost simultaneously for several different surveys. In France, this practice is institutionalized for household surveys that have a common set of demographic variables, i.e., employment, residence and income. These variables are important co-factors for the variables of interest in each survey, and if used carefully, can reinforce the estimates derived from each survey. Techniques for calibrating uncertain data can apply naturally in this context. This involves finding thebest unbiased estimator in common variables and calibrating each survey based on that estimator. The estimator thus obtained in each survey is always a linear estimator, the weightings of which can be easily explained and the variance can be obtained with no new problems, as can the variance estimate. To supplement the list of regression estimators, this technique can also be seen as a ridge-regression estimator, or as a Bayesian-regression estimator.

When data external to a questionnaire survey is available, composed of totals of auxiliary variables (some of these totals are often exact values) the estimate derived from the survey must be compared with the exact values. One fairly widespread use consists of modifying the extrapolation weights to determine those values. The theory of calibrating estimators consists of studying the statistical impacts of those practices. It shows that, using broad assumptions, calibration produces estimators that are significantly more precise than the Horvitz-Thompson estimator. The variance is very small for variables that are closely linked to calibration variables, and even nil for these latter variables.

However, this assumes that the auxiliary totals are known exactly, with no errors, which is not always the case. There are other possibilities, including:

- The auxiliary totals are derived from a survey considered to be more precise because of its size. At INSEE, for example, auxiliary data is often imported from the survey of employment (which uses 80,000 households) to reinforce the precision of small multi-annual surveys (called PCVs) that use only 5 to 8,000 households;
- there are measurement errors in the method for gathering auxiliary information;
- the auxiliary information is derived from experts' estimates (for example, national accountants) who guarantee a certain margin of error in their estimates.

In this paper, we acknowledge that the external data is derived from other surveys, even though, officially, the word "survey" can be replaced with the word "source," on the condition we are somewhat flexible regarding what is called a chance variable (one which casts doubt on the level of accuracy claimed by the expert, for example!)

First we will look at the case of a single external source where the total calibration variables X is estimated by $\hat{X}$ in the survey and by $\widetilde{X}$ in the external source. We will then look at a case involving results from several almost simultaneous surveys with common information. This is the case used by INSEE, where all population surveys must include a common set of questions on demography, residence, employment and income. Finally, we will look at what can be attempted when several surveys with information that partially overlaps.

[1] CREST- Survey Statistics Laboratory, Ker-Lann Campus, rue Blaise Pascal, 35170 BRUZ

1. CALIBRATING FROM AN UNCERTAIN SOURCE

We thus have two estimates, $\hat{X}$ and $\tilde{X}$ for X that we assume (at least for the second one) are unbiased. We assume that we also know (or can reliably estimate) the variances V and $\tilde{V}$ of these random vectors as well as the covariance C between $\hat{X}$ and $\tilde{X}$ Finally, we know (or estimate) the covariances (vectors) Γ and $\tilde{\Gamma}$ for $\hat{Y}$ with $\hat{X}$ et $\tilde{X}$. We apply $W = V + \tilde{V} - C - C' = Var(\hat{X} - \tilde{X})$ and $\Gamma_0 = \Gamma - \tilde{\Gamma}$.

We can then look for the best linear estimator of Y in the formula $\hat{Y} + B'(\hat{X} - \tilde{X})$ which leaves the estimate bias-free. The optimizing vector is obviously $B = W^{-1} \Gamma_0$. . The minimum variance (with respect to quadratic forms) thus obtained is

$$Var(\hat{Y}) - \Gamma_0' W^{-1} \Gamma_0 \qquad (1).$$

This estimator looks like a regression estimator and it may have analogous properties. Is it a linear estimator?

If Γ is estimated by $\sum_{k,\ell} \Delta_{k\ell} \frac{x_k}{\pi_k} \frac{y_\ell}{\pi_\ell} = \sum_\ell z_\ell \frac{y_\ell}{\pi_\ell}$, we can write:

$$\hat{Y}_0 = \sum_s \frac{y_k}{\pi_k} + (\tilde{X} - \hat{X})' W^{-1} (\sum_s z_\ell \frac{y_\ell}{\pi_\ell} - \tilde{\Gamma})$$
$$= \sum_s w_k y_k - (\tilde{X} - \hat{X})' W^{-1} \tilde{\Gamma}$$

The estimator is linear if $\tilde{\Gamma} = 0$; otherwise it is dependent on a quantity which is a function of $\hat{Y}$ which is hardly easy to manipulate.

The variance for this estimator can be estimated using the formula (1) above, or indeed we can write $Var(\hat{Y}_0) = Var(\hat{Y} - B'\hat{X}) + B'\tilde{\Gamma} + B'(\hat{W} - C')B$. The first term can be seen as the variance of the total of residues $y_k - B'x_k$, the others deduced from the data of the problem. Only the last additional term exists if the two sources of information are independent.

We can also see that, in the case of independent sources, the result is a ridge regression, because $B = (V + \tilde{V})^{-1} \Gamma = \left(Var(\tilde{X})^{-1} Cov(\hat{Y}_0, \tilde{X})\right)$.

The use of this type of regression to create an estimator (Chambers) is thus interpreted as a lack of confidence in the auxiliary information to which a variance is implicitly attributed.

Another interpretation is possible, particularly if the auxiliary information is based on statements by the experts who associate a level of precision with their prognosis for the value of X. The information is thus *a priori* information in the Bayesian sense of the term. We know that, in this case, the estimate of Y and the regression of Y on X results in a ridge regression (Deville,(1977).

2. USE OF THE BEST ESTIMATOR OF AUXILIARY INFORMATION

We will now introduce the best unbiased linear estimator X^* based on $\hat{X}$ and $\tilde{X}$. Let's review its construction.

We are looking for a square matrix A with the same dimensions as X liable to reduce to a minimum, in the ordered vector space of symmetric matrices, the variance $A\tilde{X} + (1-A)\hat{X}$, which results in: $A = (V - C)W^{-1}$ and $1 - A = (\tilde{V} - C')W^{-1}$ and thus:

$$X^* = (V - C)W^{-1}\tilde{X} + (\tilde{V} - C')W^{-1}\hat{X}.$$

We can thus express the optimal estimator of Y using X^* and $\hat{X}$. We can easily see that:

$$\hat{Y}_0 = \hat{Y} + \Gamma_0'(V - C)^{-1}(X^* - \hat{X})$$

Again, we have something that looks like a regression estimator. Also, if we choose y_k as linear combinations of x_k ($y_k = U' x_k$) then

$$\Gamma_0 = E(U'\hat{X})(\hat{X} - \tilde{X}) = U'(V - C)$$

such that $\left(U'\hat{X}\right)_0 = U'X^*$.

In other words the estimator is calibrated to X^*, the best unbiased estimator of X taking into account the information available.

In the event that the correlations between the two sources are nil, we obviously have a linear estimator, the weights of which are:

$$w_k = \frac{1}{\pi_k}\left(1 + \left(X^* - \hat{X}\right)'V^{-1}z_k\right) \qquad (2).$$

If the survey involves simple random sampling, these weights are those of the regression estimator.

3. PRACTICAL APPLICATION

This observation generates an approximate practical rule when using uncertain auxiliary information. It is not always a good idea to use values derived from a variance estimate in point estimators. We know that those values are generally estimated in a highly unstable fashion and do not, in practice, always lead to satisfactory results. This is why, in practice, we use the regression estimator (or various simple forms of calibrated estimators) rather than the optimal estimator recommended, for example, by Montanari (1987).

In any case, it seems natural to try to obtain the optimal estimator of X. Correlations between the two sources are often neglected. In the case of overlapping surveys, we can draw on those surveys by taking into account disjoint samples only, despite the artificial nature of that practice, particularly in the case of two-phase surveys. The first approximation is thus: $X^* = \tilde{V}W^{-1}\hat{X} + VW^{-1}\tilde{X}$.

If $\widetilde{V}$ is unknown (expert's estimate, external data for which the precision is known only approximately), then $\widetilde{V} \cong \alpha V$; a prudent approach would be to take α fairly broadly, thus taking a discrete value of the variance of the external source, and minimizing its confidence level, resulting in $X^* = \dfrac{\alpha \hat{X} + \widetilde{X}}{\alpha + 1}$. If the external information is derived from a survey of the same type as the survey we are interested in, with respective sample sizes of $\widetilde{n}$ et n , then $\alpha \cong n / \widetilde{n}$, which leads us to select $X^* = (n \hat{X} + \widetilde{n} \widetilde{X}) / (n + \widetilde{n})$. This choice seems sensible, and is justified by the fact that the optimums we are looking for are very 'flat' and can be achieved approximately without too much problem.

We then use a calibration estimate of X^*. If this is reduced to a regression estimate, we thus calculate $\hat{B} = T_s^{-1} \sum_s \dfrac{x_k \, y_k}{\pi_k}$, with $T_s = \sum_s (x_k \, x_k^{'}) / \pi_k$ to create the estimator:

$$\hat{Y}_0 = \hat{Y} + \hat{B}^{'}(X^* - \hat{X}) = \sum_s w_k y_k$$
$$\text{with } w_k = \frac{1}{\pi_k}\left(1 + \left(X^* - \hat{X}\right)^{'} T_s^{-1} x_k\right) \qquad (3)$$

The calculation of the variance (and its estimate) is fairly accessible. Generally, the optimal estimator chosen will have the formula $X^* = A \widetilde{X} + (1 - A) \hat{X}$, where A is chosen and then $\hat{Y}_0 = (\hat{Y} - \hat{B}^{'} A \hat{X}) + \hat{B}^{'} A \widetilde{X}$ for which we deduce the variance:

$$Var(\hat{Y}_0) = Var(\sum_s \frac{\breve{e}_k}{\pi_k}) + \hat{B}^{'} A \widetilde{V} A^{'} \hat{B}$$
$$\text{with } \breve{e}_k = y_k - \hat{B}^{'} A x_k$$

The first term is like a variance for a standard survey, and the second is a supplementary term due to the variance of the external information, the relevance of which is dependent on the degree of accuracy with which it is known.

Another slightly different approach is possible, using only the variance of X^* , and of $\hat{X}$, and the residuals of the regression linked to the estimator $e_k = y_k - \hat{B}^{'} x_k$ .By distributing the relationships (3), we see that $Var(\hat{Y}_0) = Var(\sum_s \dfrac{e_k}{\pi_k}) + \hat{B}' Var(X^*) \hat{B} + Cov(\hat{Y} - \hat{B}\hat{X}, \hat{B}X^*)$ .This last term is fairly easy to calculate and equals $\hat{B}^{'}(1 - A)(\Gamma - V\hat{B})$. It is nil if the regression is that which is linked to the optimal estimator according to Montanari(1988), particularly if it is that of the standard least squares regression in a simple random design. It can be assumed that in practice, this term will not be very important.

4. CASE OF MULTIPLE EXTERNAL INFORMATION

We assume now that we have multiple external information $\widetilde{X}_i$ estimating each unbiased X with a known variance $\widetilde{V}_i$. To simplify, although it is not subsequently essential, we will assume that these sources are independent of eac least squares other and independent of our survey. All the covariances that cross between sources are thus assumed to be nil. As before, we want to improve the estimator for the current survey by looking for an estimator of the form:

$\hat{Y}_B = \hat{Y} + \sum_{i=0}^{I} B_i' \widetilde{X}_i$ where by convention $\widetilde{X}_0 = \hat{X}$.

The condition of non-bias results in the constraint $\sum_{i=0}^{I} B_i = 0$. The variance for this estimator is:

$$Var(\hat{Y}_B) = Var(\hat{Y}) + B_0'\Gamma + \sum_{i=0}^{I} B_i'\hat{V}_i B_i$$

The solution to the problem of minimizing this quadratic form is expressed by:

$$B_0 = -(W - V_0^{-1})W^{-1}V_0^{-1}\Gamma$$
$$B_i = V_i^{-1}W^{-1}V_0^{-1}\Gamma$$
$$\text{with } W = \sum_{i=0}^{I} V_i^{-1}$$

It can be read and interpreted in different ways. Specifically, it is a very good idea to go back to the previous problem. We note that: $\sum_{i=1}^{I} B_i' \widetilde{X}_i = \Gamma' V_0^{-1} W^{-1} (\sum_{i=1}^{I} V_i^{-1} \widetilde{X}_i)$ and, in the parentheses, we will see the 'numerator' of the best unbiased linear estimator of X formed on the $\widetilde{X}_i$ (for i= 1 to I) i.e., $X^* = (W - V_0^{-1}) \sum_{i=1}^{I} V_i^{-1} \widetilde{X}_i$. Thus, the optimal estimator is correctly written as: $\hat{Y}_0 = \hat{Y} + B_0' (X_0^* - \hat{X})$. The variance of X_0^* is calculated immediately and equals $(W - V_0^{-1})^{-1}$, such that $\hat{Y}_0$ is the best unbiased estimator using the external information represented by X_0^*.

Thus, all previous results apply with no modification. In particular, $\hat{Y}_0$ is a linear estimator, whose weights are given by (2). Its variance is obtained using the same technique:

- Estimate the variance for the residual artificial variable $y_k - B_0' x_k$.
- Add the variance $B_0' (W - V_0^{-1}) B_0$ of $B_0' X_0^*$.

Finally, as above, we can easily see that $\hat{Y}_0$ is an estimator calibrated on $X^* = W^{-1}\left[\left(W - V_0^{-1}\right)\hat{X} + V_0^{-1} X_0^*\right] = W^{-1}\left(\sum_{i=0}^{I} V_i^{-1} \widetilde{X}_i\right)$, the variance of which is W^{-1}. In this case, we write $\hat{Y} = \widetilde{B}'(X^* - \hat{X})$ *with* $\widetilde{B} = V_0^{-1}\Gamma$, which is not dependent on the data from our

survey of interest. The estimator $\hat{Y}_0$, its variance and its variance estimate can now be calculated by referring just to external information X_0^* and Var(X^*) .

5. APPLICATION

Everything stated in paragraph 3 remains true in the case of calibration using several uncertain sources. If the sources I+1 are surveys, we can proceed symmetrically. We define a reasonable compromise by the linear combination of sources I+1 to obtain an 'almost best unbiased estimator' for X . Taking a mean weighted by sample sizes is a minimum possibility that can always be applied. We then apply to each survey a calibration estimator on the best compromise. The variance in estimators created for each survey can be estimated exactly, as stated in Section 3.

6. CONCLUSIONS AND EXTENSIONS

The ideas presented above are useful for numerous applications and easy to envision: calibration on structural surveys, administrative data containing measurement errors, or even on national accounts data. They can also be used when an institution is conducting several surveys almost simultaneously on different subjects, but the surveys contain the same descriptive population variables. We can extend the field of application to successive surveys involving a rapidly changing variable, while gathering structural data that varies much more slowly. Eventually, the addition of some temporal series techniques could again reinforce the reliability of estimates.

Finally, we can study more complex auxiliary information structures. It is not difficult to extend what was presented in the case where there are correlations among information sources. One apparently more complex case consists of the presence of different information within different sources. This can be formally resolved by acknowledging that there is only one vector of auxiliary variables, and some of its coordinates have nil variances (if they are known exactly) or infinite variances if they are absent from the source.

REFERENCES

Chambers, R,W.(1994) *Journal of Official Statistics*

Deville, J-C.(1979) Estimation linéaire et paramètres bornés, *Les cahiers du BURO, ISUP, Paris,* Vol 22, pp 137-173

Montanari,G.E.(1987).Post-sampling efficient prediction in large-scale surveys. *International Statistical Review,* Vol 55, pp 191-202

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

DIAGNOSTICS FOR COMPARISON AND COMBINED USE OF DIARY AND INTERVIEW DATA FROM THE U.S. CONSUMER EXPENDITURE SURVEY

J. L. Eltinge[1]

ABSTRACT

The U.S. Consumer Expenditure Survey uses two instruments, a diary and an in-person interview, to collect data on many categories of consumer expenditures. Consequently, it is important to use these data efficiently to estimate mean expenditures and related parameters. Three options are: (1) use only data from the diary source; (2) use only data from the interview source; and (3) use generalized least squares, or related methods, to combine the diary and interview data. Historically, the U.S. Bureau of Labor Statistics has focused on options (1) and (2) for estimation at the five or six-digit Universal Classification Code level. Evaluation and possible implementation of option (3) depends on several factors, including possible measurement biases in the diary and interview data; the empirical magnitude of these biases, relative to the standard errors of customary mean estimators; and the degree of homogeneity of these biases across strata and periods. This paper reviews some issues related to options (1) through (3); describes a relatively simple generalized least squares method for implementation of option (3); and discusses the need for diagnostics to evaluate the feasibility and relative efficiency of the generalized least squares method.

KEY WORDS: Efficiency; Generalized least squares; Mean squared error; Measurement bias; Measurement quality; Total survey design

1. INTRODUCTION

1.1 The U.S. Consumer Expenditure Survey

The U.S. Consumer Expenditure Survey (CE) is a large-scale household survey carried out for the U.S. Bureau of Labor Statistics (BLS). The principal stated goals of this survey are to: (a) produce mean and quantile estimates for specified expenditures by U.S. consumers; (b) obtain cost weight estimates used in computation of the U.S. Consumer Price Index; and (c) provide economic, social and policy researchers with population-based sample data on consumer expenditures, income and assets. The present paper will focus on issues related to goals (a) and (b). For some general background on the CE, see, e.g., U.S. Bureau of Labor Statistics (1997), Pearl (1977, 1979), Silberstein and Scott (1991), Tucker (1992), Hsen (1998a) and references cited therein.

1.2 Interview and Diary Data in the Consumer Expenditure Survey

CE data are collected from sample consumer units (CUs, which are roughly equivalent to households). Each selected sample CU is assigned to one of two data collection modes: interview or diary. A CU assigned to the interview group is asked to participate in five interviews spaced approximately three months apart. The first interview is used primarily for bounding purposes. In each of the second through fifth interviews, CUs are asked to report expenditures from the previous three months. For many variables, CUs are asked to report their expenditures *separately* for each of the preceding three *record months*, where each record month corresponds to a single calendar month. For example, in an interview conducted in April, a CU is asked to report expenditures separately for January, February and March.

[1] J.L. Eltinge, Department of Statistics, Texas A&M University and Office of Survey Methods Research, U.S. Bureau of Labor Statistics, PSB 4915, 2 Massachusetts Avenue NE, Washington, DC 20212 USA
Eltinge_J@bls.gov

A CU assigned to the diary group is asked to complete a weekly expenditure diary for each of two consecutive weeks. The diary instrument is intended to cover all expenditures incurred in the specified week, excluding some special items like food and lodging expenditures on overnight trips. Diary data are considered to be of special interest for small and frequently purchased items, for which one could not realistically anticipate accurate recall in an interview one or more months after the purchase. Conversely, interview data are of interest for purchases that one could reasonably expect to be recalled as long as three or four months later.

1.3 Possible Combination of Interview and Diary Data

In general, expenditure data are collected by the CE at relatively fine levels of aggregation known as the five- and six-digit Universal Classification Code (UCC) code levels. For analytic purposes (a) and (b) listed in Section 1.1, five or six-digit UCC-level mean or total expenditure estimates could potentially be based on one of three options:

(1) use only data from the diary source;

(2) use only data from the interview source; or

(3) use generalized least squares, or related methods, to combine the diary and interview data.

At present, the Bureau of Labor Statistics uses only options (1) and (2), with the specific data source selected in the following way. First, for many six-digit UCCs, data are collected through only the interview mode or only the diary mode, so the selection of a data source is straightforward. Second, for six-digit UCCs for which data are collected by both the diary and interview, the choice between diary and interview is determined through balanced consideration of several factors. These factors include both quantitative measures, e.g., reporting rates and mean reported levels; and qualitative considerations, e.g., cognitive properties of the interview or diary-based data collection, or related substantive issues associated with a given survey item. For some background on the selection of interview or diary sources for specific UCCs, see, e.g., Jacobs (1984, 1988), Shipp (1984), U.S. Bureau of Labor Statistics (1984), Brown (1986), Jacobs et al. (1989), Branch and Jayasuriya (1997) and references cited therein.

For calendar year 1994, the U.S. Bureau of Labor Statistics (1997, p. 164) reported 20,517 completed interviews and 10,884 completed diaries, where, e.g., completion of a diary in two consecutive weeks by a selected household is counted as two completed diaries. Although these sample sizes are relatively large, for many specific six-digit UCCs, nonzero expenditures in a given year are relatively rare. Consequently, single-source estimators of those six-digit UCCs can have sampling errors that are larger than desirable, which in turn can suggest consideration of option (3).

The remainder of this paper focuses on option (3). Section 2 considers limitations of the CE interview and diary data, with special emphasis on measurement error issues which could have a substantial effect on the comparison and possible combination of CE data sources. Section 3 develops some notation approximations for measurement bias terms. The approximations lead to a relatively simple generalized least squares method for combination of the CE data sources, and also lead to several related diagnostic issues. Section 4 briefly considers three extensions of the main issues considered here.

2. LIMITATIONS OF INTERVIEW AND DIARY DATA

The U.S. Bureau of the Census conducts fieldwork and initial data processing for the CE, and the BLS produces final point estimates, variance estimates and related data analyses. Sample consumer units are selected through a complex design; some details of this design are reported in U.S. Bureau of the Census (1999). Some design characteristics of special interest in the present work are as follows. First, standard analyses treat the data as arising from a stratified multistage design, with two primary sample units (PSUs) selected within each stratum. Second, subsampling within each selected PSU leads to selection of sample

consumer units, roughly equivalent to households. Third, within each PSU, some sample CUs are assigned to an interview group, and others are assigned to a diary group. The assignment of CUs to the interview or diary group is considered to be independent across PSUs.

For UCCs covered by both instruments, the interview and diary are intended to record the same expenditures in their respective specified reference periods. However, previous authors' empirical results indicate that data quality may vary substantially between and within the interview and diary sources. See, e.g., Silberstein and Scott (1991), Tucker (1992) and references cited therein. Of special concern are observed differences in reported means among the three record months covered in a given interview, with lower means reported for more distant months; and between the two diary weeks, with lower means reported for the second week.

3. COMBINATION OF INTERVIEW AND DIARY DATA

3.1 Bias Approximations and Generalized Least Squares Estimation

The bias issues summarized in Section 2 suggest that one view CE data as arising from five distinct methods: interview data for the most recent record month; interview data for the second most recent record month; interview data for the most distant record month; diary data from the first week of data collection; and diary data from the second week of data collection. Let $j = 1,...,5$ serve as an index to label these five methods. In addition, for a specified time period p, stratum h, and CU i, let x_{phi} equal the true expenditure; let x_{ph} equal the finite population mean of x_{phi} for period p and stratum h; and let x_p equal the mean of x_{phi} for period p. Also, let y_{phij} equal the hypothetical value that consumer unit i in stratum h would report if it were selected and measured using method j for period p; let y_{p_j} equal the corresponding finite population mean of the values y_{phij} for specified p and j; let $y_{p\_} = (y_{p\_1},...,y_{p\_5})'$; and define the five-dimensional vector $\hat{y}_{ph} = \left(\hat{y}_{ph1},...,\hat{y}_{ph5}\right)'$, where $\hat{y}_{phj}$ is a customary survey weighted point estimator of x_{ph} based only on data from source j.

Now define the vector of ratios $\beta_p = (\beta_{p1},...,\beta_{p5})' = (y_{p\_1},...,y_{p\_5})' / x_p$. Three features of β_p are of special interest here. First, β_p is itself a finite population quantity, and depends (through the finite population means y_{p_j}) on the measurement errors $y_{phij} - x_{phi}$ for consumer unit i in stratum h during period p. Second, for the P periods $p = 1,...,P$, define the averaged vector $\beta = (\beta_1 + ... + \beta_P) / P$ and define the differences $u_{ph} = y_{ph} - \beta x_{ph}$. Then in the decomposition,

$$\hat{y}_{ph} = y_{ph} + e_{ph} = \beta x_{ph} + u_{ph} + e_{ph} \qquad (3.1)$$

the vector β reflects systematic measurement bias effects, averaged across periods 1 through P; u_{ph} involves variability in the ratio vectors y_{ph} / x_{ph} across strata and across periods; and e_{ph} equals the sampling error in $\hat{y}_{ph}$, conditional on the finite population or error-contaminated reports y_{phij}.

Third, under identifying restrictions (e.g., the assumption that for one of the sources j the ratio β_{pj} equals one for all p and j) and additional regularity conditions, one may obtain consistent estimators of the remaining elements of β and of the covariance matrix of the combined errors $u_{ph} + e_{ph}$. Routine arguments then indicate that one may use the abovementioned parameter estimators and the vector estimators $\hat{y}_{ph}$ to compute a generalized least squares estimator $\tilde{x}_p$, say, of the true mean x_p for period p.

3.2 Diagnostics for Combined-Data Estimation

The potential benefits of $\tilde{x}_p$ in practical applications will depend on the extent to which the approximations described in Section 3.1 are consistent with observed data; and on the relative efficiency of $\tilde{x}_p$. To address these issues, one could consider several groups of diagnostic tools, including the following.

(i) Stability of the vectors β_p across the periods $p = 1,..., P$. This can be assessed through formal tests based on the nominal t statistics $(\hat{\beta}_{pj} - \hat{\beta}_j) / se(\hat{\beta}_{pj} - \hat{\beta}_j)$ where $\hat{\beta}_{pj}$ and $\hat{\beta}_j$ are point estimators of the j-th elements of β_p and β, respectively; and $se(\hat{\beta}_{pj} - \hat{\beta}_j)$ is a standard error estimator for the associated difference. In addition, plots of these nominal t statistics against the indices p can help one explore specific patterns of change in the β_p across time.

(ii) Other diagnostics for the adequacy of the approximation (3.1) include a t test for the presence of a nonzero intercept in (3.1); tests for inclusion of other regressors in (3.1); tests for the homogeneity of the variances of e and u across periods; and tests for the stability of the weighting factors produced by the generalized least squares method.

(iii) Efficiency (quantified by estimated mean squared error) of $\tilde{x}_p$, relative to simpler estimators. These competing simpler estimators may include single-source means $\hat{y}_{p\_j}$; the average of the three interview means, $(\hat{y}_{p\_1} + \hat{y}_{p\_2} + \hat{y}_{p\_3}) / 3$; and the average of the two diary means, $(\hat{y}_{p\_4} + \hat{y}_{p\_5}) / 2$. In keeping with standard approaches to total survey design (e.g., Andersen et al., 1979; and Linacre and Trewin, 1993), interpretation of the efficiency gains, if any, achieved by $\tilde{x}_p$ should also account for cost issues, e.g., the additional computational and administrative burden involved in implementation of the generalized least squares method.

4. DISCUSSION

In closing, we note three related points. First, the present discussion has restricted attention to combination of CE data across sources within a specified time period. However, one could also consider combination of data across time, either using a single data source or multiple data sources. This currently is done, e.g., with composite estimation for the U.S. Current Population Survey. See, e.g., Cantwell and Ernst (1992), Lent et al. (1996) and references cited therein.

Second, this paper has considered only the use of data from currently implemented forms of the U.S. Consumer Expenditure Survey. One could consider modification of current CE data collection methods, e.g., through the direct use of cash register receipts or scanner data as discussed in Raines (1996), Hsen (1998b) and Dashen et al. (1998, 1999). Also, in principle, one could use U.S. consumer expenditure data from other sources, e.g., the Personal Consumption Expenditure (PCE) estimates from the National Income and Product Accounts (NIPA) produced by the U.S. Department of Commerce. For some background on the Personal Consumption Expenditure data and comparisons with CE data, see, e.g., Gieseman (1987), Branch and Cage (1989), Branch (1994) and references cited therein. In general, PCE estimates are reported at relatively high levels of aggregation, and with some publication delay following the reference year of interest. Thus, the PCE data may be of use primarily in assessing bias of CE-based estimators. This in turn may provide some additional insight into the identifying restrictions considered in Section 2. In addition, if PCE data were available for the relevant periods p, one could consider expansion of the least squares estimators of Section 2 to incorporate both CE and PCE data. The resulting application would require some indication of the variances of the PCE estimates, and of the correlation of the PCE estimates with the CE data. However, any use of PCE data, either for comparison with CE estimates, or for

combination with CE data, should be viewed with some caution due to issues involving the comparability of definitions of expenditure classes and of underlying populations.

Finally, recall from Section 1.1 that CE data are used in the construction of cost weights for the U.S. Consumer Price Index. In some cases, one can view price indices as measures of the shift in an expenditure distribution between two periods, conditional on a given distribution of item counts of purchased goods and services. For such cases, two important issues would be quantification of the following.

(a) The extent to which univariate price indices can be supplemented by broader measures of distributional change, e.g., offset functions or shift functions, as discussed in Doksum and Sievers (1976) and Oja (1981).

(b) The extent to which the combination of interview and diary data, or selection of a single data source, will affect the operating characteristics of an estimated offset function from (a).

ACKNOWLEDGEMENTS

The author thanks Stephen H. Cohen, Chris Cope, Monica Dashen, Cathryn S. Dippo, Alan H. Dorfman, Thesia Garner, Mary McCarthy, Bill Mockovak, Adriana Silberstein, David Swanson and Wolf Weber for many helpful comments on the U.S. Consumer Expenditure Survey; and Adriana Silberstein for suggesting separate analysis of the three interview record months and two diary weeks. The views expressed here are those of the author and do not necessarily represent the policies of the U.S. Bureau of Labor Statistics.

REFERENCES

Andersen, R., Kasper, J., Frankel, M.R. et al. (1979). *Total Survey Error*. San Francisco: Jossey-Bass.

Branch, E.R. (1994). The consumer expenditure survey: A comparative analysis. *Monthly Labor Review*, December, 1994, 47-55.

Branch, R. and Cage, R. (1989). Data comparison: Quarterly results from the consumer expenditure survey and personal consumption expenditures 1984 through 1987. Manuscript, Division of Consumer Expenditure Surveys, U.S. Bureau of Labor Statistics, April 28, 1989.

Branch, R. and Jayasuriya, B. (1997). Consumer expenditure interview and diary data selection: A new method. Manuscript, Division of Consumer Expenditure Surveys, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, August 8, 1997.

Brown, G. (1986). Analysis of 82/84 data for integration source selection. U.S. Bureau of Labor Statistics memorandum to S. Shipp, September 3, 1986.

Cantwell, P.J. and Ernst, L.R. (1992). New developments in composite estimation for the Current Population Survey. *Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys*, 121-130.

Dashen, M., Davis, J., Rogers, J. and Silberstein, A. (1998). Report on the use of scanners to collect diary survey data. Manuscript, Office of Survey Methods Research, U.S. Bureau of Labor Statistics, September 25, 1998.

Dashen, M., Davis, J., Rogers, J. and Silberstein, A. (1999). Team recommendations on using receipts in the CE diary. Manuscript, Office of Survey Methods Research, U.S. Bureau of Labor Statistics, February 26, 1999.

Doksum, K.A. and Sievers, G.L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika* **63**, 421-434.

Gieseman, R. W. (1987). The consumer expenditure survey: Quality control by comparative analysis. *Monthly Labor Review,* March, 1987, 8-14.

Hsen, P. (1998a). Specifications for weighting the 1998 diary survey using the calibration method. U.S. Bureau of Labor Statistics memorandum to S. Groves, September 28, 1998.

Hsen, P. (1998b). Analysis results from the field representative receipts survey. U.S. Bureau of Labor Statistics memorandum to M. McCarthy, October 30, 1998.

Jacobs, C.A. (1984). Criteria for expenditure data source selection. U.S. Bureau of Labor Statistics memorandum to E. Jacobs, June 18, 1984.

Jacobs, C.A. (1988). CE integrated publication 1984/86 – source selection for apparel estimates. U.S. Bureau of Labor Statistics memorandum, September 27, 1988.

Jacobs, E.E., Mason, C. and Shipp, S. (1989). Item source selection differences between the CE and CPI. U.S. Bureau of Labor Statistics memorandum to K.V. Dalton, April 26, 1989.

Lent, J., Miller, S.M. and Cantwell, P.J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 130-139.

Linacre, S. and Trewin, D. (1983). Total survey design – application to a collection of the construction industry. *Journal of Official Statistics* **9**, 611-621.

Oja, H (1981). On location, scale, skewness and kurtosis of univariate distributions. *Scandinavian Journal of Statistics* **8**, 154-168.

Pearl, R. B. (1977). The 1972-1973 U.S. consumer expenditure survey: A preliminary evaluation. In *Proceedings of the Social Statistics Section, American Statistical Association*, 492-497.

Pearl, R.B. (1979). Re-evaluation of the 1972-73 U.S. Consumer Expenditure Survey: A further examination based on revised estimates of personal consumer expenditures. Technical Paper No. 46, U.S. Bureau of the Census. U.S. Government Printing Office stock number 003-024-01938-8.

Raines, M.D. (1996). Field representative (FR) receipts survey. U.S. Bureau of Labor Statistics CE office memorandum 96-36, September 3, 1996.

Shipp, S. (1984). Data source selection for the 1980/81 CE integrated publications. U.S. Bureau of Labor Statistics memorandum to E. Jacobs, September 26, 1984.

Silberstein, A. R. and Scott, S. (1991). Expenditure diary surveys and their associated errors. In *Measurement Errors in Surveys* (Paul P. Biemer, Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz and Seymour Sudman, eds.) 303-326. New York: Wiley.

Tucker, C. (1984). Discussion on the use of composite estimators for the publication of the integrated diary and interview suerveys. U.S. Bureau of Labor Statistics memorandum to A. Lippert, May 15, 1984.

Tucker, C. (1992). The estimation of instrument effects on data quality in the Consumer Expenditure Diary Survey. *Journal of Official Statistics* **8**, 41-61.

U.S. Bureau of the Census (1999). Summary of the 1990 redesign for the consumer expenditure surveys and the rent and property tax survey. Documentation memorandum from the CE Section, Victimization and Expenditures Branch, Demographic Statistical Methods Division, U.S. Bureau of the Census, January 13, 1999.

U.S. Bureau of Labor Statistics (1984). Initial results concerning integration of the diary and interview surveys by composite estimation. BLS internal memorandum, July 23, 1984.

U.S. Bureau of Labor Statistics (1997). Chapter 16: Consumer expenditures and income. *BLS Handbook of Methods*. U.S. Department of Labor, Bureau of Labor Statistics Bulletin 2490, April, 1997. Washington, DC: U.S. Government Printing Office.

SESSION IX

APPLICATIONS

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

COMBINING DATA SOURCES: AIR POLLUTION AND ASTHMA CONSULTATIONS IN 59 GENERAL PRACTICES THROUGHOUT ENGLAND AND WALES – A CASE STUDY

J Charlton[1], S Stevenson[2], B Armstrong[2], T Fletcher[2], P Wilkinson[2]

ABSTRACT

The geographical and temporal relationship between outdoor air pollution and asthma was examined by linking together data from multiple sources. These included the administrative records of 59 general practices widely dispersed across England and Wales for half a million patients and all their consultations for asthma, supplemented by a socio-economic interview survey. Postcode enabled linkage with: (i) computed local road density, (ii) emission estimates of sulphur dioxide and nitrogen dioxides, (iii) measured/interpolated concentration of black smoke, sulphur dioxide, nitrogen dioxide and other pollutants at practice level. Parallel Poisson time series analysis took into account between-practice variations to examine daily correlations in practices close to air quality monitoring stations. Preliminary analyses show small and generally non-significant geographical associations between consultation rates and pollution markers. The methodological issues relevant to combining such data, and the interpretation of these results will be discussed.

KEY WORDS: Linkage; environment; health; methodology.

1. INTRODUCTION

1.1 Description of the Problem

Recently there has been burgeoning public and scientific interest in the relationship between the environment and health, in particular cardio-respiratory disease. Evidence mainly from daily time-series and panel studies suggests that air pollution can exacerbate asthmatic symptoms although its effects, e.g. on hospital admissions (Wordley 1997; Sunyer 1997), emergency room visits (Jorgensen 1996)), and decreased lung function/ increased respiratory symptoms (Zmirou 1997). The epidemiological evidence in relation to *chronic* effects is more limited, and currently insufficient to conclude that air pollution contributes to the development of asthma in previously healthy subjects (Anderson 1997; Guidotti 1997). Certainly air pollution does not appear to be a key determinant of the large international differences in asthma prevalence (ISAAC 1998). However studies of variation in asthma prevalence in relation to air pollution are often difficult to interpret because of potential bias and confounding when comparing different population groups, coupled with the difficulty of deriving comparable measures of pollution exposure. The evidence of a positive association between asthma and proximity to traffic pollution has been contradictory. There is, moreover, little evidence that asthma prevalence differs between urban and rural areas, or that it correlates with long-term changes in pollution levels (COMEAP 1995, Anderson 1997, Dockery 1989). This paper describes an analysis of spatial and temporal variations in asthma consultations using a large sample obtained by linking together different datasets. In particular the methodological issues involved in such exercises are discussed.

1.2 Methodological issues for environmental exposure studies

It is important to recognise that time-series and prevalence studies address fundamentally different questions. Though analytically complex, time series are attractive because the same population is compared day to day. Uncertainties of comparing different populations are thereby avoided. However

[1] Office for National Statistics, London, (UK)
[2] London School of Hygiene and Tropical Medicine (UK)

their focus is very specific and limited to short-term effects. Standard analytical techniques entail removing all long-term variation in disease frequency, leaving only the daily fluctuations from the seasonal baseline for comparison with pollution concentrations. Consequently the evidence of time-series relates only to short-term effects, i.e. only to exacerbation of symptoms rather than induction of disease.

From a public health perspective, however, it is the relationship between chronic exposure and new disease onset that is of greatest interest. As McMichael (1998) and colleagues have pointed out, chronic health effects may have a different pathophysiological basis from the acute, and the health impacts of long-term exposure may not reliably be estimated from an 'annualisation' of the risk relationships derived from daily time-series. Their quantification requires comparison of the prevalence/long-term incidence of disease in populations exposed to differing levels of outdoor pollution.

Ideally, such comparisons should be based on cohort studies with individual-level data on confounding factors and long-term follow-up. However, studies of this kind are expensive and difficult to set up, and interest has grown in the potential offered by small area analysis of routine data. In recent years, the sophistication of small area studies has been enhanced by developments in computer technology Vine (1997), Briggs (1995), Coggan (1995) and the improved geographical referencing of health datasets (for example through use of the postcode of residence).

Experience with these designs remains limited, however, and there are various methodological issues which affect their interpretation or limit their use. The more important are:

(1) **Exposure misclassification**. Categorising exposure on the basis of location may not accurately reflect personal dose (Coggan 1995) and so may lead to bias and loss of statistical power. The factors that contribute to misclassification include:

(i) use of over-simplistic exposure models (e.g. ones which assume unrealistic dispersion patterns, or that fail to take account of micro-variations in ambient exposure levels)
(ii) unmeasured exposure sources, including sources in the home or at work
(iii) population movement through the local environment such as in commuting to work
(iv) long term population migration (especially important when studying hazards with long latency)

(2) **Small relative risks**. Exposures to environmental pollutants and the health risks associated with them are often small and hence difficult to distinguish from underlying disease variation. Statistical power and precision can often be improved by increasing the size of the study population and indeed, in some studies based on very large populations, relative risk are estimated with narrow confidence intervals. But if the relative risk is small, the possibility of bias as an alternative explanation is difficult to discount.

(3) **Confounding**. Population health is influenced by a wide range of social, behavioural and health care factors few of which are recorded in routine data. In studies of respiratory disease, for example, data are seldom available on variation in smoking prevalence.

(4) **Data quality**. Integrity and consistency of data are pre-requisite to the sound interpretation of any geographical study. Even where comparisons are made over large populations or in relation to extended emissions sources, variations in case ascertainment may still produce bias. Routine data cannot be checked with the same rigour as data collected for the purpose, so uncertainties about their completeness, accuracy, and spatial consistency often remain. Propensity scores (Rubin 1997) may help in such analyses.

(5) **Sensitivity**. For various reasons, available measures of health impact may not be sensitive to changes in health status. This may be the case if the health outcome is rare, delayed in onset, likely to develop only with extreme exposure or with pre-disposing susceptibility, or if it shows marked spatial variation due to unmeasured risk factors.

(6) **Confidentiality issues.** Certain data that could be used for environmental studies are collected subject to confidentiality undertakings, e.g. for the MSGP4 data described here the data collectors had to guarantee that it would be impossible to identify any patient or participating practice. Such undertakings are met by using "masking procedures" (Armstrong 1999), which, although necessary for ethical considerations, do

place some constraints on the research.

However, geographical studies also have many advantages. The description of the spatial distribution of exposures, populations and disease risks is useful in itself, sometimes revealing aspects of their inter-relationships that remain hidden in other sorts of analysis. Exposure assessment, though indirect, can usually be done with comparative ease and applied over wide areas. The regional or national coverage of many routine datasets provides the opportunity to study large populations. This has benefits in terms of power and statistical precision, enables study areas of contrasting exposure to be selected. Investigation of multiple sites offers possibilities for independent testing of hypotheses generated in a subset of sites.

The study presented in this report is based on data from the fourth Morbidity Survey in General Practice (McCormick 1995) which recorded consultations in 60 general practices across England and Wales between September 1991 to August 1992. It has a number of advantages for studying air pollution effects. First, it covers a large population base with a broad geographical distribution in both urban and rural settings, and wide contrasts of exposure. Secondly, data based on GP consultations will pick up milder cases than hospital admission data and may thus be a more sensitive marker of pollution effects. Thirdly, the data include the patients full postcode of residence, which allows analysis with very fine geographical resolution. Finally, the availability of individual-level survey-derived variables such as socio-economic and smoking status means that important potential confounding effects may be controlled for. At the same time, the date of consultations makes it possible to study the relationship between short-term fluctuations in exposure and general practice consultations, and hence to compare time-series results with those of geographical analysis.

2. METHODS

2.1 Asthma consultation data

Analyses were based on data from the Fourth Morbidity Survey in General Practice (MSGP4). Data from one of the 60 participating practices was considered potentially unreliable and was excluded from further analysis. MSGP4 records information on the characteristics of all patients registered at each participating practice, as well as on all consultations during the study year. Variables made available for analysis included: date of consultation, diagnosis, referral (if any), who contacted, days in study, distance of residence from practice, ethnicity, housing tenure, marital status, social class, occupational code, smoking habit, and economic position. Consultations were coded as first ever for the specified diagnosis (`first ever'), first consultation for a new episode of disease (`new'), and 'on-going'. The analyses presented here were based on first ever and new consultations only. Analyses were carried out for both sexes and all ages combined, and separately for the 0-4, 5-14, 15-64, 65+ years age-groups. Two sets of analysis were undertaken: spatial comparisons and daily time series, both of which entailed integrating MSGP4 data with data on air pollutants, meteorology and aero-allergens from national monitoring.

2.2 Geographical exposure data

Markers of pollution exposure

Geographical comparisons were based on the following data:

(1) Annual averages of the daily mean pollution concentrations of black smoke, SO_2 and ozone used in the time-series analysis (described below)

(2) 5km x 5km raster grid of interpolated 6-month mean NO_2 concentrations modelled from the results of a 1991 diffusion tube survey and population data.

(3) Data from the National Emissions Inventory, generated from data on fixed and mobile emissions sources, and available as 10 x 10 km raster grid for nitrogen oxides (NOx), SO_2, and CO.

(4) Road density (length of road/km^2) of motorways, primary roads, A-roads, B-roads and unclassified roads, computed as a 0.25 x 0.25 km raster from the AA digitised national road network.

Area socio-demographic characteristics

In addition to the data on the registered population derived from the MSGP4 data, the following area-based characteristics were obtained and used to classify populations by the areas in which they live:

(1) The Carstairs (1991) deprivation index - a composite of overcrowding, social class, unemployment and car access computed at enumeration district level from 1991 census data.

(2) Population density (persons/kilometre^2), calculated at ward level by dividing the total number of residents (1991 census data) by the ward area (computed from digitised ED-line™ ward boundaries).

Integration of spatial datasets, and generation of road density maps, was carried out within the Geographical Information System ARC-INFO v 7.1. The road density grid, initially generated at 250 x 250 m resolution, was spatially smoothed to obtain densities averaged over circles of 500m, 1km, 5km & 10km radius of each grid square. This was done to provide measures of road density (separate for each class of road) over the wider area within which someone might live and work. Similarly, ward population densities were also smoothed by taking (i) the average population density over all wards sharing a boundary with the index ward (`adjacency smoothing') and (ii) averaging all wards whose centroids lay within 5 km of the index ward centroid (`5 km smoothing'). These smoothed measures of population density distinguish built up areas in small towns from those in major conurbations, and hence were useful for examining urban-rural differentials.

2.3 Time series data

The time series analysis was an analysis of 59 parallel time series (one for each practice), though reliable pollution and aero-allergen data in particular were not available for all sites. The following data sources were used.

Pollutants

Daily pollution data were derived from monitoring sites in the National Environmental Technology Centre co-ordinated by AEA Technology. There were two broad categories:

(i) Non-automatic black smoke and sulphur dioxide (SO_2) sites providing daily average smoke (smoke stain method) and SO_2 (acidemetric bubbler), and

(ii) Automatic monitoring stations providing daily average, 15-minute maximum and hours above EC directive limits for nitrogen dioxide (NO_2), sulphur dioxide, carbon monoxide (CO) and ozone (O_3); particle (PM_{10}) concentrations were also obtained, but no site had a full year of PM_{10} data, and no analyses of PM_{10} are presented in this report.

Selection of monitoring sites, and decisions of how representative their data would be of pollutant concentrations in the study areas, were made by AEA Technology. These judgements took account of the location and type of each monitoring station. Practices for which the best available pollution data were considered unrepresentative were excluded from analyses.

Meteorological data

Meteorological stations were selected site by site by collaborators from the Meteorological Office. Satisfactory matches were made for all practices. The following variables were calculated from hourly data: minimum daily temperature; maximum daily temperature; minimum daytime (7.00 -19.00 GMT) temperature, rainfall; thunder indicator; daily mean speed and direction of wind; maximum heat loss in $kJ.m^{-2}.hr^{-1}$; maximum vapour loss in $g.m^{-2}$; maximum rate of cooling; maximum rate of vapour pressure reduction; and Monin-Obukov length (a derived measure of air mixing relevant to dispersion of pollutants).

Pollen data

Daily pollen counts for 1991-92 were obtained from monitoring stations co-ordinated by the Pollen Research Unit: counts of grass pollen in each case, and a breakdown of other taxa for five sites. Only a small part of the study population was covered by both pollen and pollution data. Because of the difficulty of integrating pollen data with data on consultations, only very limited analyses of aero-allergen effects has so far been possible. Further work is now continuing to explore methods of extrapolating the available pollen data more widely; but the results of this work are not yet complete and are not reported here.

2.4 Statistical analyses

The relationship between first and new consultation rates and explanatory factors was examined by tabulation and regression methods. Associations between air pollution markers and consultations are shown as consultation odds ratios pertaining to an increase in exposure from the fifth to the 95th centile of the distribution of the relevant marker. For geographical comparisons, the proportion of the registered population who consulted for asthma in the study period was computed. Analysis was confined to patients registered for at least 75% (275 days) of the study year. Initial analyses were based on logistic models using the Huber/White sandwich estimator of variance, specifying practice-level clustering to allow for dependence of errors within practices (Huber 1967). Subsequent analyses were based on Generalised Estimation Equations with dependence assumed between consultations of patients from the same practice.

Time-series analyses were restricted to the areas for which pollution measurements are available. The analysis broadly followed the methods applied to analyses of air pollution and health events described by Schwartz(1996). The number of new and first ever consultations was calculated by practice and day and analysed using log-linear models with Poisson error. The main confounding factors incorporated into the analysis were practice, season, meteorological conditions (primarily minimum and maximum daily temperature), upper respiratory tract infections, influenza, and aero-allergens. Initial analyses of aero-allergen data, available for only a small number of sites, did not show an important association with consultation rates, and to avoid excluding most practices, pollen was not further included in regression models of pollutant effects. Season was controlled for by using a practice-specific weighted 29-day moving average of asthma consultations (14 days either side of the index day) to compute the expected number of consultations. The weights used followed a pyramidal function, i.e. they declined linearly away from the index day, reaching zero at 15 days either side of it (Burnett 1994). Minimum and maximum temperature effects were each modelled as three-piece linear splines with cut points at -5° and 9°C (min) and 4° and 26°C (max). Other meteorological variables were used in more detailed analyses, but only minimum and maximum temperature, the main meteorological determinants of variation in asthma consultations, were used in models presented in this report. In summary, we investigated whether days of high pollution were associated with days of higher rates of consultation than those of the previous and subsequent two weeks, after allowing for day of the week, public holidays, minimum and maximum temperature, and frequency of consultations for upper respiratory tract infection.

Table 1. Summary of practice characteristics and daily data.

	No. of	Mean	Centile	
PRACTICE AVERAGES	Practices	(s.d.)	5th	95th
Age	59	37.8 (2.89)	31.7	39.4
%male	59	49.2 (2.0)	48.4	51.6
Pop. Density /km^2	59	1294 (1152)	89	4157
% manual households	59	51.9 (14.1)	31.8	62.1
% current smokers	59	28.5 (6.6)	19.4	31.6
Percent consulting in year for				
Asthma	59	2.4 (1.0)	0.8	4.2
URTI	59	25.0 (5.5)	16.4	28.0
Annual av. Pollutant (ppm)				
Black smoke	42	12.5 (4.6)	6.2	20.3
Sulphur dioxide	42	31.4 (12.6)	14.4	47.5
Ozone	24	21.5 (5.8)	12	30
Carbon monoxide	11	0.8 (0.31)	0.5	1.4
Nitrogen dioxide	59	13.6 (5.3)	4.9	21.3
Emissions(tonnes/yr)				
Nitrogen oxides	59	3627 (3451)	243	11520
Sulphur dioxide	59	9386 (9103)	717	34000
Carbon monoxide	59	1709 (2860)	64	7646
DAILY AVERAGES	No. of days			
Asthma consultations	21960	0.68 (1.12)	0	3
URTI consultations	21960	8.4 (8.5)	0	25
Meteorological factors				
Minimum temperature	21960	6.5 (5.1)	-2.2	14.1
Maximum temperature	21960	13.9 (6.2)	4.5	24.0
Evaporative loss	21960	27.4 (1.3)	24.9	29.2
Pollutants				
Black smoke	13362	12.9 (16.0)	1	40
Sulphur dioxide	13309	31.9 (23.1)	6	74
Nitrogen dioxide	3190	27.5 (13.7)	11	50
Ozone	6320	22.1 (12.1)	4	42

3. RESULTS

3.1 Geographical analysis

The analyses are still in progress and only preliminary results are reported here. Tabulations of the proportion of registered population who consulted for asthma (new and first ever consultations) during the study period are shown in Table 1. Results of the early regression analyses are shown in Table 2. Despite wide variation in estimated pollutant concentrations and pollutant emissions none of the markers analysed showed evidence of association with asthma consultations after adjustment for age (18 five-year age-groups) and sex, though confidence intervals were wide. In fact, all point estimates were below 1.0. Additional individual-level adjustment for social class of head of household left odds ratios unchanged or reduced them. The width of the confidence intervals reflected the fact that pollutant concentrations were assigned on the basis of practice-level data, with emission categories based of fairly large-scale geographical data that came close to being practice-level. Thus, information on the association of pollution and emissions with consultations came from between-practice comparisons, which were blunted by unexplained differences between practices. Road density was classified by the enumeration district of residence (linked through the postcode) and yielded estimates with narrow confidence intervals. However, the results again provided no evidence of positive association, the point estimate odds ratio for a 5th to 95th percentile increase in road density being 1.00 with or without adjustment for socio-economic deprivation. Further analyses are being undertaken to investigate the influence of between-practice effects on these findings. They include analyses based on road and population density at various levels of spatial smoothing.

Table 2. Proportion of registered population consulting for asthma in the study year: odds ratios (95% confidence intervals) for a 5th to 95th increase in exposure marker

Exposure marker	Practices	Observations	ODDS RATIOS	
			Adjusted for age & sex	Adjusted for age, sex and social class of head of household
POLLUTANT CONCENTRATIONS				
Black smoke	42	337361	0.99 (0.68 - 1.43)	0.99 (0.68 - 1.44)
Sulphur dioxide	42	337361	0.92 (0.66 - 1.28)	0.84 (0.61 - 1.17)
Nitrogen dioxide	59	445341	0.92 (0.60 - 1.41)	0.88 (0.59 - 1.32)
Ozone	24	171787	0.78 (0.46 - 1.31)	0.76 (0.47 - 1.23)
POLLUTANT EMISSIONS				
Oxides of nitrogen	59	445341	0.98 (0.66 - 1.45)	0.95 (0.64 - 1.40)
Carbon monoxide	59	445341	0.97 (0.94 - 1.01)	0.97 (0.94 - 1.01)
Sulphur dioxide	59	445341	0.99 (0.69 - 1.42)	0.96 (0.68 - 1.37)
ROAD DENSITY	59	445363	1.00 (0.99 - 1.01)	1.00 (0.99 - 1.01)

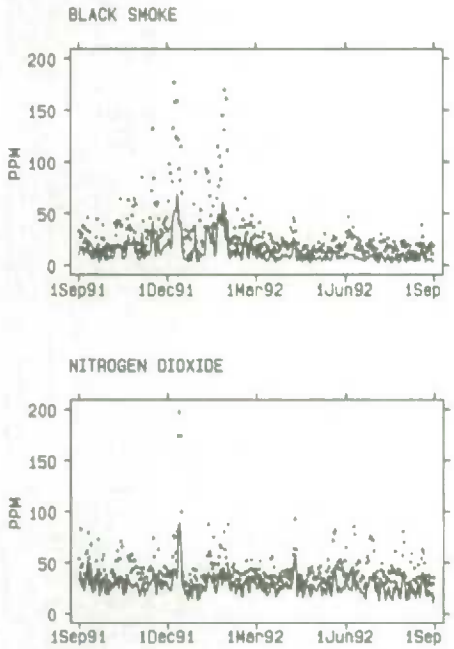


Figure 1: daily outdoor concentrations: mean (black line), 5th (top dot) and 95th (bottom dot) centile of daily mean values across practices

3.2 Daily time-series

Air pollution concentrations varied appreciably between practices and throughout the year (Figure 1). Black smoke and sulphur dioxide levels were highest in winter months. A large peak in nitrogen dioxide concentrations also

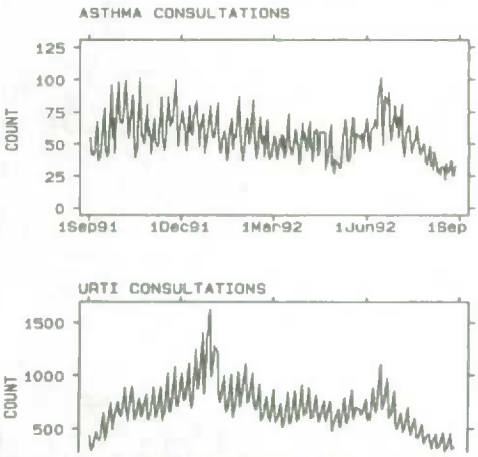


Figure 2: Daily frequency of consultations for asthma

accompanied the first episode. Ozone showed clearer seasonal fluctuation in mean concentrations, the higher baseline and peaks being observed during summer months. Meteorological parameters followed a broadly sinusoidal variation with local fluctuations in minimum and maximum daily temperature as well as in the maximum daytime rate of heat loss and evaporative loss.

Asthma consultations showed a rise in September, remained generally high in late autumn to early winter (October - December), declined steadily until May, but then showed a surprisingly sharp peak in June before declining rapidly again (Figure 2). Upper respiratory tract infections showed a steadier and clear rise to a peak in December and January before declining steadily, with a small rise in June. Of the specific determinants of variations in asthma consultations, day of the week and public holidays were among the most important - consultations were most frequent on a Monday, and less than 20% as frequent on weekend days. The association with minimum and maximum daily temperatures showed roughly U-shaped curves: an increase in consultations at low temperatures and some increase at high maximum temperatures, but fairly constant over the middle part of the temperature distributions. The consultation rate also showed a small positive gradient with vapour pressure loss and maximum rate of drying, as it did with the maximum rate of heat loss and, somewhat more strongly, with the maximum rate of temperature decline (cooling) during daytime hours. Asthma consultations showed an association with upper respiratory tract infections in the preceding week (averaged across practices).

After considering these relationships with asthma consultation ratios and the correlations between them, the models chosen to examine pollution effects included the following covariates: day of the week, indicator of public holidays, three-piece splines for minimum temperature and maximum temperature, and practice-specific count of upper respiratory tract (URTI) consultations in the preceding week. Seasonal and other broad-scale variation in asthma consultations was controlled for by using the practice-specific 29-day moving average of asthma consultations.

Table 3 shows that, unadjusted for these factors, black smoke, sulphur dioxide and nitrogen dioxide each showed a significant positive gradient with asthma consultations and ozone a negative association. However, the statistical significance of these associations was lost after adjusting for the day of week, public holidays, minimum and maximum temperature - point estimates were moved close to 1.0. After additional adjustment for URTIs and using a 29-day moving average, none of the pollutants showed even suggestive evidence of association with new and first ever consultations either for asthma or respiratory disease. The effect on pollutant rate ratios of adding covariates sequentially is shown in Table 3. Plotting the period of the large winter peaks in black smoke revealed no perceptible change in asthma consultations following the pollution peaks. Only two associations with black smoke and sulphur dioxide were nominally significant for asthma: black smoke at ages 0-4 years and lag of 2 days; and sulphur dioxide at ages 65+, no lag. Given the frequency of significance tests, not all of which were independent, this is no more than might be expected by chance. Overall, as many point estimates were below 1.0 as above it, and there is little in the pattern of results to suggest clinically important associations. The strongest, but still very weak, evidence is that for the older age-group and infants. But taken as a whole (although analyses are not yet complete), the results do not provide evidence of association between pollution levels and daily consultations for asthma.

Table 3. Daily time series results: first and new consultation ratios for asthma (95% confidence intervals) for 5th to 95th percentile increase in pollutant concentration

	Consultation ratios adjusting for:				
	Unadjusted	Day of week & public holidays	Variables of previous column + min & max temperature	Variables of previous column + 29-day moving average	Variables of previous column + URTIs
Black smoke	1.11 (1.05-1.18)	1.04 (0.99-1.09)	1.01 (0.94-1.08)	0.95 (0.90-1.00)	0.94 (0.89-0.99)
Sulphur dioxide	1.18 (1.09-1.29)	1.06 (1.00-1.14)	1.06 (0.99-1.14)	1.01 (0.96-1.07)	1.01 (0.96-1.06)
Nitrogen dioxide	1.28 (1.09-1.51)	1.05 (0.97-1.14)	0.97 (0.90-1.11)	0.99 (0.92-1.06)	0.99 (0.92-1.06)
Ozone	0.73 (0.64-0.84)	0.91 (0.79-1.04)	0.94 (0.83-1.07)	1.03 (0.94-1.12)	1.02 (0.94-1.11)

4. DISCUSSION

The spatial analyses were limited by the method of exposure classification, which entailed classifying individuals by rather broad and indirect markers, often, as in the case of pollution concentrations, at practice level. This was reflected in the width of confidence intervals. However, the fact that all the main comparisons yielded point estimates close to or below 1.0, even for road density, does not suggest that a real association is being obscured by lack of precision.
Practice-level effects may arise with such factors as thresholds for contact with health services, diagnostic categorisation, data recording etc, and these may therefore obscure pollution effects which were also largely based on practice to practice comparisons. On the other hand, a strength of this design was that it drew on data from practices with contrasting exposures. One of the potential problems of studies based in single towns or cities is the uncertainty about exposure contrast and exposure misclassification. Even though there may be appreciable micro-variation in pollutants at street level, exposure misclassification is likely to arise from the imprecision of defining pollution concentrations at particular locations, compounded by population movements between different environments, such as in commuting to work. By taking practices in very different areas, not only are the contrasts in average ambient levels maximised, but the effect of population movement is less important, as the exposure categorisation is based on broader scale averaging. There is thus a trade-off: at practice level, contrasts are maximised and the effects of population movements diminished, but the associations may be obscured by other practice level factors; within practice comparisons are not affected by such factors, but contrasts in pollution are generally smaller and population movement more problematic. Further analyses are being undertaken to examine how the scale of spatial analysis affects the findings of the study.

The time series spanned only a single year and covered a comparatively small (for time-series) base population of around 400,000 people. That number was further limited when analysing the effects of specific pollutants by the lack of reliable air quality data for all practice locations. None the less, the findings in relation to specific determinants of daily variation in asthma consultations was consistent with expected patterns and risk estimates were made with reasonable precision.

The lack of association is compatible with a weak association with air pollution (though our results provide little direct evidence of this) and would certainly be consistent with other published evidence that suggests that pollution effects on asthma are at most very small. Other time-varying factors had clear associations with asthma consultations: there were broad seasonal and day-of-the-week associations and expected relationships with temperature and upper respiratory tract infections. Unadjusted for such confounding factors, black smoke, sulphur dioxide and nitrogen dioxide concentrations all showed modest but statistically significant, positive associations with the asthma consultations, but adjustment for these factors removed all statistical significance and reduced point estimates to very close to 1.0.

The use of the 29-day moving average was probably conservative in dealing with seasonal effects: the pyramidal weighting function gives three-quarters of its weight to days in the week either side of the index day, so that all but very short-term associations may in part be removed by the moving average. While focusing on such short term associations may be appropriate for very acute effects (e.g. pollution-related deaths in those with end-stage respiratory disease) this is not so clear in the case of mild asthma symptoms - patients may not experience such acute effects that they need to seek medical advice immediately, but there might nonetheless still be an increase in consultations over the span of several days to a week. Such a diffuse swell in consultations over several days may not be clearly apparent in the sort of analysis presented here. However, it is worth noting that even without inclusion of the 29-day moving average, the associations between asthma consultations and pollutant levels was at best very weak and most of the variation across the year is explained by other environmental factors.

It can also be argued that inclusion of URTIs as a covariate will remove any effect of air pollution on asthma that operates through increased susceptibility to respiratory infection; and perhaps there may also be some diagnostic confusion between URTIs and exacerbation of asthma symptoms. But URTIs clearly exacerbate asthma and so it was important to examine its influence when quantifying pollution effects.

The age-group analysis provides only the weakest suggestion that pollution effects, if they exist at all, are

strongest in the elderly. But interactions with such factors as age, socio-economic status and average ambient pollution levels are currently being explored, as are methodological issues relating to the scale of analysis in spatial comparisons, and the time-averaging of temporal associations, and cumulative effects.

Overall the lack of associations in both the geographical and time series analyses argues against air pollution being an important determinant of asthma morbidity, though small effects cannot be excluded. The findings broadly support the balance of scientific evidence that outdoor air pollution is not a major contributor to induction of asthma, and that it has no more than a fairly small effect on exacerbation of symptoms in some individuals

REFERENCES

Anderson HR. (1997). Air pollution and trends in asthma. *CIBA Foundation symposia*, 206:190-207

Armstrong MP, Rushton G, Zimmerman DL (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18, 497-525.

Briggs, D.J., Elliott, P. (1995) The use of geographical information systems in studies on environment and health. *World Health Statist Qurt*, 48, 85-94

Burnett R, Krewski D.(1994). Air pollution effects on hospital admission rates: a random effects modelling approach. *Canadian J Stats*, 22(4):441-58

Carstairs V, Morris R. (1991). *Deprivation and health in Scotland*. Aberdeen University Press, Aberdeen.

Coggan D. (1995) Assessment of exposure to environmental pollutants. Occup Environ Med,52:562-64

COMEAP (Committee on the Medical Effects of Air Pollutants). (1995). *Asthma and outdoor air pollution*. Department of Health. London, HMSO.

Dockery DW, Speizer FE, Stram DO, et al. (1989). Effects of inhalable particles on respiratory health of children. Am Rev Respir Dis,139:587-94

Guidotti TL.(1997). Ambient air quality and asthma: a northern perspective. *J Invest Allergology Clin* Immunol,7(1):7-13)

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221-233

ISAAC (International Study of Asthma and Allergies in Childhood) Steering Committee. (1998). Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema. *Lancet* 1998; 351:1225-32

Jorgensen B, Lundbye-Christensen S, Song X-K, Sun L. (1996) A longitudinal study of emergency room visits and air pollution for Prince George, British Columbia. *Stats in Medicine*, 15(7-9):823-836

McCormick A, Fleming D, Charlton J (1995). *Morbidity Statistics from General Practice, Fourth national study 1991-92*. London, OPCS

McMichael AJ, Anderson HR, Brunekreef B, Cohen AJ. (1998). Inappropriate use of daily mortality analyses to estimate longer-term mortality effects of air pollution. *Internatl J Epidemiol*, 27:450-53

Rubin DB (1997). Estimating causal effects from large data sets using propensity scores. *Ann Intern Med*

15;127:(8 Pt 2): 757-63.

Sunyer J, Spix C, Quenel P, Ponce de Leon A, Ponka A, Baramandzadeh T, et al. (1997). Urban air pollution and emergency admissions for asthma in four European cities: the AHEA Project. *Thorax*, 52:760-65

Schwartz J, Spix C, Touloumi G, Bacharova L, Barumamdzadeh T et al. (1996). Methodological issues in studies of air pollution and daily counts of deaths or hospital admissions. *J Epidemiol Commun Health*, 50(suppl 1):S3-11

Vine, M.F., Degnan, D., Hanchette, C. (1997) Geographic Information Systems: their use in environmental epidemiologic research. *Environ Health Perspec* 105, 598-605

Wordley J Walters S Ayres JG (1997) Short term variations in hospital admissions and mortality and particulate air pollution. *Occup Environ Med*, 54(2):108-116

Zmirou D, Balducci F, Dechenaux J, Piras A, Benoit-Guyod F, Filippi J-L. (1997). Meta-analysis and dose-response functions of respiratory effects of air pollution. *Revue d Epidemiologie et de Sante Publique*, 45(4):293-304

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

A METHOD OF GENERATING A SAMPLE OF ARTIFICIAL DATA FROM SEVERAL EXISTING DATA TABLES : APPLICATION BASED ON THE RESIDENTIAL ELECTRIC POWER MARKET

Christian Derquenne[1]

ABSTRACT

The artificial sample was generated in two steps. The first step, based on a master panel, was a Multiple Correspondence Analysis (MCA) carried out on basic variables. Then, "dummy" individuals were generated randomly using the distribution of each "significant" factor in the analysis. Finally, for each individual, a value was generated for each basic variable most closely linked to one of the previous factors. This method ensured that sets of variables were drawn independently. The second step consisted in grafting some otherdata bases, based on certain property requirements. A variable was generated to be added on the basis of its estimated distribution, using a generalized linear model for common variables and those already added. The same procedure was then used to graft the other samples. This method was applied to the generation of an artificial sample taken from two surveys. The artificial sample that was generated was validated using sample comparison testing. The results were positive, demonstrating the feasibility of this method.

KEYWORDS: Statistical Data Fusion, Survey Data, Multiple Correspondence Analysis, Calibration, Generalized Linear Models.

1. BACKGROUND

A major strategy used by Electricité de France has been to develop electric power consumption among residential clients. This has required a knowledge of the relationships between various stakeholders of France's residential electric power market (clients, builders, competitors, etc.). Since the overall information is not available in a single data base including the same customers, a market simulation project has been undertaken. The first step has been to generate a sample of artificial individuals, drawn from several independent data bases. Each individual corresponds to different characteristics (socio-demographic, behaviour, use of equipment, etc.). The author provides a method of generating a sample of artificial individuals [Derquenne, 1998] involving two main steps based on different areas of statistics: sampling, data analysis and generalized linear models.

2. A METHOD OF GENERATING A SAMPLE OF ARTIFICIAL INDIVIDUALS

Two sets of survey samples were available (*primary sample* $\mathcal{I}$ and *secondary samples* $\mathcal{Y}^1,...,\mathcal{Y}^K$, respectively). The primary sample included the variables of the sample design taken from the survey design: X_{MP} = {gender; age; size of town; occupation} and the measured variables X_M = {dwelling and household characteristics; main heating source of the dwelling; etc.}. The secondary samples had some variables common to X_{MP}, i.e. $Y_{MP}^{(k)}$, and other variables (common and non-common to X_M), i.e. $Y_M^{(k)}$. The principle used to generate the sample of artificial individuals involved two main steps:

(i) Generating the *first artificial sample* based on primary sample $\mathcal{I}$,

---

[1] Christian Derquenne - EDF - Direction des Etudes et Recherches - 1, av. du Général de Gaulle - 92141 CLAMART cedex - FRANCE.

(ii) Statistically grafting a secondary sample y^1 onto the first artificial sample to obtain the *second artificial sample*.

Step (ii) was repeated to provide the *final artificial sample* by progressively grafting other secondary samples $y^2,...,y^K$.

2.1 Generating the first artificial sample

Let $X_{MP} = \left(X_{MP(1)},\ldots,X_{MP(Q)}\right)$ denote the sampling variables (all qualitative), and let $X_M = \left(X_{M(1)},\ldots,X_{M(R)}\right)$ denote the measured variables (nominal, ordinal or discretized). This step consists in applying a Multiple Correspondence Analysis (MCA) [Benzecri et al., 1979] to X_{MP} as active variables (used to generate the principal components) and to X_M as supplementary variables. The MCA is used to provide reduced dimension as well as new, uncorrelated variables (principal components).

Let $Z = \left(Z_1,\ldots,Z_T\right)$ denote the principal components derived by MCA. We then select a subset Z^* of Z corresponding to "significant" eigenvalues l_t, where $Z^* = \left(Z_1,\ldots,Z_{T^*}\right)$, with T^* = Number of eigenvalues greater than $1/MP(Q)$. We then generate groups of variables which best correlate with the principal components. To do so, we calculate the correlation ratio η^2 between sampling variables and "significant" principal components, and we establish the maximum of η^2 such that:

$$\eta^2\left(X_{MP(q)},Z_u\right) = \max_{t=1,T^*} \eta^2\left(X_{MP(q)},Z_t\right) \tag{1}$$

then $X_{MP}^{(t)} = \left\{X_{MP(q)}\ ;\ \eta^2\left(X_{MP(q)},Z_u\right),\forall t,t=u\right\}$ = {group of variables correlated with the principal component $Z_t \in Z^*$}. We follow the same procedure to obtain supplementary variable groups: $X_M^{(t)}$. Then, artificial individuals can be drawn in two steps.

As a first step, each principal component of Z^* is discretized into k_t intervals so as to generate a paving space such that:

$$k_t = \left[(n_d)^{1/T^*} \times \lambda_t \Big/ \prod_{t=1}^{T^*} \lambda_t\right] \text{ where } (n_d = n/s,\ s \text{ fixed}) \tag{2}$$

and $\tilde{K} = \prod_{t=1}^{T^*} k_t \le n_d$ is the number of windows $\left(w_l\right)$ in the paving space. Let $f_l = n_l/n$ denote the distribution observed in that space, where $n_l = \sum_{i=1}^{n} 1_{(i\in w_l)}$ $\left(l = 1,\tilde{K}\right)$. Thus N artificial individuals are drawn from this distribution, where N is the size of the sample generated (normally greater than that of the primary sample). These artificial individuals are called "dummy individuals": $\tilde{z}_{\tilde{i}} = \left(\tilde{z}_{\tilde{i}(1)},\ldots,\tilde{z}_{\tilde{i}(T^*)}\right)$, for $\tilde{i} = 1,N$.

As a second step, since each $X^{(l)}_{MP(q)}$ and $X^{(l)}_{M(r)}$ also involves a distribution observed in each window of the space, N artificial individuals are drawn, knowing $\tilde{z}_i$. These new artificial individuals are called "first replicates": $\tilde{x}_{i} = \left(\tilde{x}_{i(1)}, \ldots, \tilde{x}_{i(MP(Q)+M(R))}\right)$.

2.2 Statistical graft based on secondary samples

The first sample is chosen within the set of secondary samples $y^1,...,y^K$. This choice is normally based on the number of sampling variables common to the primary sample and the variables of the secondary sample deemed important for the study. Let $Y^{(1)}_{MP} = \left\{Y^{(1)}_{MP(1)}; \ldots; Y^{(1)}_{MP(Q_1)}\right\}$ denote the sampling variables common to the primary sample, and let $Y^{(1)}_{M} = \left\{Y^{(1)}_{M(1)}; \ldots; Y^{(1)}_{M(R_1)}\right\}$ denote the other measured variables and G_1 the number of variables to be grafted onto the first artificial sample (generally, these variables are not common to the primary sample). The second artificial sample is generated in two steps.

As a first step, we adjust the secondary sample in relation to the sample design of the primary sample, using a marginal calibration method [Deming and Stephan, 1940].

As a second step, we graft the G_1 variables (one by one) onto the first artificial sample. For example, let $y^{(1)}_{M(j)}$ $\left(j = 1, G_1\right)$ denote an ordinal variable taken from a multinomial distribution, and let $Y^{(1)}_{C}$ denote a set of variables common to the primary sample and the secondary sample. In terms of the Generalized Linear Models [Mc Cullagh and Nelder, 1990], $y^{(1)}_{M(j)}$ is the variable to be explained, $Y^{(1)}_{C}$ are the candidate explanatory variables, and the link function is cumulative logit. Thus G_1 models are generated on $y^{(1)}_{M(j)}$, knowing $Y^{(1)}_{C}$. The first grafted variable to be explained corresponds to the variable which provides the best fit with the data (e.g. using the likelihood ratio test). Let $y^{(1)}_{M(j^*)}$ denote this variable, and $\hat{y}^{(1)}_{M(j^*)}$ its estimate. Then N artificial individuals $\tilde{y}^{(1)}_{M(j^*)}$ are drawn from the estimated distribution of $y^{(1)}_{M(j^*)}$, the characteristics of the significant "explanatory" variables in the first artificial sample being known. There then remain G_1-1 for which G_1-1 models are generated on $y^{(1)}_{M(j)}$, with $Y^{(1)}_{C}$ and $y^{(1)}_{M(j^*)}$ being known.

The previous procedure for selecting the $y^{(1)}_{M(j)}$ and the drawing of N artificial individuals are again applied so as to generate a third artificial sample. The generation of other artificial samples follows the same procedure until the final artificial sample is obtained.

3. STATISTICAL VALIDATION OF THE GENERATED ARTIFICIAL SAMPLE

The statistical validation of the generated artificial sample (the final artificial sample) involves six test levels of increasing complexity depending on the nature of the variables.

(i) *Univariate marginal*: Comparing the marginal distributions for an existing survey and the generated sample.

(ii) *Univariate category*: Comparing the category representation percentages for an existing survey and the generated sample (test and confidence interval using a binomial distribution or a normal distribution approximation).

(iii) *Comparing correlations*:
(a) Two generated variables and two observed variables taken from the same survey (e.g. age and occupation).
(b) Two generated variables (common and non-common to the primary sample) and two observed variables (common and non-common to the primary sample), e.g. the level of satisfaction with respect to cost and the age of the client taken from two different surveys.

(iv) *Crossed categories*: Comparing the crossed category percentages for an existing survey and the generated sample (test and confidence interval as in (ii)).

(v) *Characterization*: Comparing characterizations of $y^{(1)}_{M(J^*)}$ and $\tilde{y}^{(1)}_{M(J^*)}$ with respect to common variables for a survey and the generated sample (univariate: variable by "explanatory" variable).

(vi) *Modelling*: Comparing models applied to $y^{(1)}_{M(J^*)}$ and $\tilde{y}^{(1)}_{M(J^*)}$ with respect to common variables for a survey and the generated sample (multivariate: several "explanatory" variables).

4. APPLICATION AND RESULTS

This method was applied to the generation of a sample of 10,000 artificial individuals for variables that were common and non-common to two surveys. The primary sample was linked to the 1990 CREDOC survey on the living conditions and aspirations of the French, whereas the secondary sample was taken from a 1990 SOFRES survey on home heating. The following table shows the variables of the sample design and the measured variables.

Table 1: List of the grafted variables under study

	CREDOC 1990 (2,000 persons)	***SOFRES 1990 (8,000 clients)***
Sampling variables	- gender x age - occupation - size of town	- age - occupation - size of town
Measured variables	- dwelling characteristics - household characteristics - principal heat source - opinion regarding nuclear power stations - opinion regarding the environment	- dwelling characteristics - household characteristics - principal heat source in the dwelling - level of satisfaction with respect to heat source (cost, safety, thermal comfort, etc.)

The statistical validation was carried out with a significance level of 5% for the tests and a level of confidence of 95% for the confidence intervals. The following table shows satisfactory results. As was to be expected, there was a decrease in the success of the statistical validation as the complexity of the statistical tests increased. This was due in large part to the fact that the grafting procedure was carried out variable by variable, which is not the "best" possible solution (see the discussion in § 5.).

Table 2: Results of the statistical validation

Level of complexity	*Number of successful attempts/number of tests (or % of success)*
(i) *Univariate marginal*	21/21
(ii) *Univariate category*	95/97
(iii) *Correlation*: (a)	46/48
(b)	98/104
(iv) *Crossed categories*	95%
(v) *Characterization*	91%
(vi) *Modelling*	85%

5. ADVANTAGES, LIMITS, PROSPECTS AND APPLICATIONS

The proposed method of generating a sample of artificial individuals involves two main steps. The first consists in applying Multiple Correspondence Analysis to the primary sample, so as to obtain a lower-dimensional space of uncorrelated variables and generate a first sample. The second step consists in using Generalized Linear Models to estimate the distribution of variables to be grafted (secondary samples) and to generate them. Three major advantages of the method are the possibility of generating a sample of variable size (generally large), the positive results obtained, and the generation of artificial individuals using a priori knowledge. There are two limitations, however, i.e. the fact that the size of survey samples is generally small, and the complexity of the generation process increases with the number of grafted variables and secondary samples. Nevertheless, our current research is aimed at improving the generation process using a multivariate approach for the variables to be explained, in order to avoid grafting variable by variable. Moreover, our method will be generalized in terms of longitudinal data (panel data). In terms of applications, the second step of the method has been used to enhance uninformed data from an EDF client data base. The proposed method (first and second steps) should be applied to other segments of our clientele (hotels, offices, major industry, etc.).

REFERENCES

Benzecri, J.-P., et al. (1979). *L'analyse des données*, Tome 1, : *la taxinomie*, Tome 2 : *l'Analyse des correspondances*, 3e éd., Dunod, Paris.

Deming, W.E. and Stephan, F.F. (1940). On a least square adjustment of sampled frequency tables when the expected marginal totals are known, *Annals of Mathematical Statistics*, **11**, 427-444.

Derquenne, C. (1998). A Method to Generate a Sample of Artificial Individuals in Frame of a Simulator of Household Market of Electricity, Royal Statistical Society, *Proceeding of International Conference*.

Mc Cullagh, P., and Nelder, J.A. (1990). *Generalized Linear Models*, Monographs on Statistics and Applied Probability **37**, Chapman & Hall.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

USING META-ANALYSIS TO UNDERSTAND THE IMPACT OF TIME-OF-USE RATES

K. H. Tiedemann[1]

ABSTRACT

Electricity rates that vary by time-of-day have the potential to significantly increase economic efficiency in the energy market. A number of utilities have undertaken economic studies of time-of-use rate schemes for their residential customers. This paper uses meta-analysis to examine the impact of time-of-use rates on electricity demand pooling the results of thirty-eight separate programs. There are four key findings. First, very largpeak to off-peak price ratios are needed to significantly affect peak demand. Second, summer peak rates are relatively effective compared to winter peak rates. Third, permanent time-of-use rates are relatively effective compared to experimental ones. Fourth, demand charges rival ordinary time-of-use rates in terms of impact.

KEY WORDS: Meta-analysis; regression modeling; electricity pricing.

1. INTRODUCTION

Electricity has traditionally been sold as an undifferentiated commodity, with little attention paid to the time at which it was delivered. In some respects this seems peculiar, since electricity can not be economically stored in bulk but is perishable in the sense that it must be consumed as it is produced. This provides a considerable contrast to other perishable commodities where seasonal or even daily variations in price are often the rule.

More recently, however, increasing attention has been paid to the potential of time-varying electricity prices. With these time-varying electricity prices - generally referred to as time-of-use rates, electricity prices are higher during peak periods when marginal costs of generation and distribution are higher, and lower in off-peak periods when marginal costs of generation and distribution are lower. A considerable body of economic evidence suggests that moving prices closer to marginal costs will, in general, increase economic efficiency. Some standard references include Houthakker (1951), Williamson (1966) and Train and Mehrez (1995).

Time-of-use rates were introduced in France after World War Two, and over the past twenty years have made significant inroads into North America. A number of Canadian and American utilities have implemented time-of-use rates on either an experimental or a permanent basis. Generally these schemes have involved a limited number of experimental designs (say a standard rate and several experimental rates) so that each treatment has a reasonable number of cases for analysis. There is a substantial theoretical and empirical literature on these individual time-of-use rate schemes, with much of it focused on the estimation of own-price and cost-price elasticities of demand. Key references include two special issues of the Journal of Econometrics edited by Lawrence and Aigner (1979) and Aigner (1984).

Unlike the literature on individual time-of-use rate schemes, this paper uses meta-analysis to integrate main results of various studies. More specifically, it examines the effect of time-of-use rates and demand charges on residential peak energy demand, using the results of thirty-eight time-of-use schemes. An outline of the paper is as follows. The next section provides a brief description of meta-analysis and explains why it is

[1] Ken Tiedemann, BCHydro, 6911 Southpoint Drive (E13), Burnaby, BC, Canada,V3N 4X8

appropriate in the present context. This is followed by a description of the regression model used, a discussion of the variables employed and a description of the sample. The next section describes the estimation method used and provides the results of the regression analysis. The final section of the paper discusses the implications of the findings for electricity rate design.

2. METHODOLOGY

We noted above that the analysis of time-of-use rates based on individual utility data has two significant limitations. First, although some rate design variables are modified in any given rate scheme, others remain fixed. For example, most studies compare two or more alternative energy rates with a standard rate, but fewer studies include demand charges, which are based on maximum demand and designed to cover fixed as opposed to variable costs. For a single experiment, it is difficult to understand the effect of those variables which are fixed for that experiment. Second, a given utility experiment provides information on customer response for customers within the service territory of the utility, but this information may or may not be transferable to other utilities whose customers have different characteristics.

One way of dealing with these issues is to use multiple studies which provide the necessary variation in experimental treatment and customer characteristics. Meta-analysis is a means of integrating the effects of multiple studies using regression analysis. In social science applications, meta-analysis typically begins with the concept of effect size. In other words, it starts by asking by how much the recipients of a treatment have changed their behavior as a result of the treatment. Often the outcome variables from different experiments are diverse making direct comparisons difficult. The procedure in this case is to standardize the difference by dividing the difference by the pooled standard deviation (or sometimes by the standard deviation of the comparison group). With an effect size measured in this way, multiple regression analysis can then be used to investigate the determinants of the effect size.

Where suitable data for all experiments is available, a simpler procedure is often appropriate. Standardization is needed mainly because the scaling of outcomes varies across studies. But if all studies use the same outcome measure, the effect size is just the experimental outcome minus the comparison outcome. This is the approach used in this paper.

3. MODEL AND DATA

In this paper, the outcome variable is the change in relative peak load. Relative peak load is the ratio of the average hourly energy consumption during the peak period to the average hourly energy consumption for the whole day. Change in relative peak load is then the relative peak load of the experimental group minus the relative peak load of the comparison group. Only weekday consumption periods are included in the calculation, because electricity demand and consumption are usually high during the weekdays, so that this is the time when utilities wish to modify consumption patterns of their customers.

Instead of relative peak load, a number of alternative outcome measures could have been utilized. These alternative outcome measures include: (1) change in total electricity consumption; (2) change in load by hour of day; (3) change in customer maximum kilowatt demand; and (4) change in load at system peak. For this study, change in relative peak load appears to be the best measure since it captures the essential nature of the issue and all four alternatives have logical or practical limitations. First, change in peak load is of more interest for system planning than change in total energy consumption, since peak load drives investment decisions on capacity requirements because of utilities legal obligation to serve. Second, although change in utility load (i.e. change in the whole hourly system load shape) is of considerable interest, very few utilities have (or at least have published) this information. Third, change in maximum customer kilowatt demand provides little insight into system behavior because individual maximum

demand occurs at different times on different days for different customers. Fourth, change in load at system peak is subject to relatively high random variation compared to a broader measure such as change in relative peak load.

Review of previous research on time-of-use rates listed above suggest that several factors play a key role in determining the change in relative peak load. These include peak to off-peak price ratios, whether the rate applies to summer peak or winter peak, whether the rate is permanent or experimental, and whether or not there are demand charges. Definition of the variables used in the regression analysis are given in Table 1.

Table 1. Definition of Variables and Sample Characteristics

Variable	Definition
CRPL	Change in relative peak load = change in the ratio of average energy use per hour during peak to average energy use for whole day (mean = -.16, standard deviation = .14)
POP	POP = peak/off-peak price ratio (mean = 5.03; standard deviation = 4.88)
SUM	SUM = 1 if summer peak, 0 if winter peak (mean = .61, standard deviation = .50)
PERM	PERM = 1 if time-of-use rate is permanent, 0 if time-of-use rate is experimental (mean =.11, standard deviation = .31)
DEM	DEM =1 if the rate has a demand charge, 0 if no demand charge (mean = .34, standard deviation = .48)

The first independent variable is the ratio of peak to off-peak prices. Like the dependent variable, the ratio of peak to off-peak prices is a pure number and has no units. The higher the ratio of peak to off-peak prices, the greater should be the reduction in relative peak. In other words, the expected sign on this regression coefficient is negative.

The second independent variable is the dummy variable for summer peak, rather than winter peak. Summer response should be greater than winter response for two reasons: first, system loads and individual loads have greater coincidence in summer so that summer peak residential load is more likely to coincide with the peak period for a peak rate in the summer; second, residential air conditioner loads are the largest moveable residential load and are much larger in summer. The expected sign on this regression coefficient is negative.

The third independent variable is the dummy variable for permanent rather than experimental time-of-use rates. With permanent time-of-use rates, customers should have a greater incentive to make changes which facilitates their shifting electricity loads. The expected sign on this regression coefficient is negative.

The fourth independent variable is the dummy variable for the presence of a demand charge. A demand charge should reduce peak period demand since it is based on maximum demand which tends to be very coincident with system peak. The expected sign of this regression coefficient is negative.

4. ESTIMATION METHOD AND RESULTS

We estimated several regression models using ordinary least squares regressions. However, inspection of residuals suggested the presence of heteroscedasticity so that the usual estimates of standard errors and associated t-statistics may not be appropriate. For this reason, White's heteroscedasticity-corrected covariance matrix was used to calculate the standard errors and t-statistics. This doesn't affect the estimated regression coefficients, just their standard errors.

The regression models are shown in Table 2. The t-statistics are shown below the estimated regression coefficients. With the exception of the constant terms, all regression coefficients are significant at the five percent level or better.

Table 2. Determinants of Change in Relative Peak Load
(T-ratios for coefficients in parentheses)

Variable	Model 1 (full sample)	Model 2 (full sample)	Model 3 (summer peak)	Model 4 (winter peak)
Constant	-.0138	.0164	-.0376	.0119
	(-.582)	(.707)	(-1.104)	(.551)
POP	-.0153	-.0143	-.0145	-.0141
	(-6.947)	(-6.325)	(-5.097)	(-4.556)
SUM	--	-.0578	--	--
		(-1.832)		
PERM	-.2612	-.2656	-.2578	-.2733
	(-8.959)	(-10.248)	(-6.370)	(-8.711)
DEM	-.1206	-.1206	-.1309	-.1915
	(-2.699)	(-2.807)	(-1.960)	(-2.186)
Sample size	38	38	23	15
Adj. R-squared	.46	.49	.37	.54
F statistic	11.50	9.79	5.36	6.56

Model 1 uses the full sample while ignoring the distinction between summer peak and winter peak. All coefficients have the expected signs, i.e. a higher peak to off-peak price ratio, permanent rather than experimental time-of-use rates, and demand charges all significantly reduce relative peak load. Overall equation fit is quite good for a cross section model.

Model 2 again uses the full sample but adds a dummy variable for summer peak as a regressor. Again, all coefficients have the expected sign, with the summer peak dummy in particular having the expected negative coefficient. The magnitudes of the coefficients are essentially unchanged from Model 1. There is a small increase in explanatory power of the regression.

Given the statistical significance of the summer peak variable, it seems worthwhile to split the sample into summer peak and winter peak schemes. Model 3 examines summer peak observations only. Coefficients again have the expected signs. Explanatory power drops somewhat.

Model 4 examines winter peak observations only. Once again coefficients have the expected signs. Explanatory power is good. However the number of observations here is quite small.

5. SUMMARY AND CONCLUSIONS

This paper has examined the role of time-of-use rates for residential electricity customers. Time-of-use rates are an attractive means of improving the efficiency with which generation and distribution assets are employed. This can be an effective means of increasing economic efficiency while reducing costs to customers. Using meta-analysis, the results of thirty-eight individual utility schemes are examined in the paper.

The regression analysis suggest several key implications for electricity rate design. First, the impact of time-of-use rates on relative peak load is fairly small in absolute terms. Large peak to off-peak rates are needed to significantly influence relative peak load. Second, summer peak rates are comparatively more effective than winter peak rates. This may reflect the ease with which space cooling loads can be shifted.

Third, permanent time-of-use rates are more effective than experimental ones. This suggest that introducing time-of-use rate to customers with an assurance that they will be in place for a substantial length of time is helpful. Fourth, demand charges rival conventional time-of-use rates in terms of their impact on relative peak load. Since residential demand charges are not yet widely used, this suggests an additional means of influencing residential peak demand.

REFERENCES

Aigner, D., ed. (1985). Welfare Econometrics of Peak Load Pricing for Electricity. *Journal of Econometrics*. 26(1).

Houthakker, H. (1951). Electric Tariffs in Theory and Practice. *Economic Journal*, 61(1), 1-25.

Lawrence, A. and Aigner, D. (1979). Modeling and Forecasting Time-of-day and Seasonal Electricity Demands, *Journal of Econometrics*, 9(1/2).

Train, K. and Mehrez, G. (1995). The Impacts of Optional Time-of-use Prices, *Energy and Buildings*, 22(3), 267-278.

Williamson, O. (1966). Peak-load Pricing and Optimal Capacity under Indivisibility Constraints, American *Economic Review*, 56(4), 810-827.

Proceedings of Statistics Canada Symposium 99
Combining Data from Different Sources
May 1999

META-ANALYSIS OF POPULATION DYNAMICS DATA: HIERARCHICAL MODELLING TO REDUCE UNCERTAINTY

Nicholas J. Barrowman[1] and Ransom A. Myers[2]

ABSTRACT

We use data on 14 populations of coho salmon to estimate critical parameters that are vital for management of fish populations. Parameter estimates from individual data sets are inefficient and can be highly biased, and we investigate methods to overcome these problems. Combination of data sets using nonlinear mixedeffects models provides more useful results, however questions of influence and robustness are raised. For comparison, robust estimates are obtained. Model-robustness is also explored using a family of alternative functional forms. Our results allow ready calculation of the limits of exploitation and may help to prevent extinction of fish stocks. Similar methods can be applied in other contexts where parameter estimation is part of a larger decision-making process.

KEY WORDS: mixed effects models; population dynamics; spawner-recruitment; meta-analysis.

1. INTRODUCTION

In this paper we discuss the analysis of fish population dynamics data, but the problem is nearly identical to problems in a wide range of fields including pharmacokinetics, dose-response, and bioassay. We view the statistical methodology as part of a process, the endpoint of which is a scientific or management decision. For example, the output of our component may be used as input to a risk assessment. A key element is the estimation of variability.

Traditionally, most fisheries scientists have focused their attention on data sets for individual fish populations of interest, which often exhibit great variability, may be contaminated by unreliable observations, and often span a small number of years. Borrowing the words of Thomas Hobbes, such data sets may be characterized as "nasty, brutish, and short". Statistical inferences based on this approach are inefficient and may be biased. Dangerous management decisions may result. Only by combining data from many studies can these problems be overcome.

To illustrate our methodology, we use data on 14 populations of coho salmon (*Oncorhynchus kisutch*), one of the most widespread of Pacific salmon (Hunter, 1959 and additional information from Pacific Biological Station, Nanimo, B.C. Salmon Archive, BL/2/5). Adult coho spawn in streams and rivers. About 1.5 years later, their offspring — known as smolts at this life stage — migrate to sea. Roughly another 1.5 years later, the survivors return to spawn. We study the population dynamics of the coho using standardized units based on painstaking observations from counting fences erected on the rivers. Let S represent the quantity of spawners, measured as the number of spawning females per kilometre of river, and let R represent the quantity of "recruits", measured as the number of female smolts per kilometre of river. (The river length is a crude way to measure habitat size and allows comparison among different coho populations.) Since egg production is proportional to the quantity of spawners, the ratio R/S is an index of juvenile survival, and it is often helpful to consider the data on this scale. An example of a coho salmon "spawner-recruitment" data set is given in Figure 1.

[1] Nicholas J. Barrowman, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 3J5
[2] Ransom A. Myers, Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada, B3H 4J1

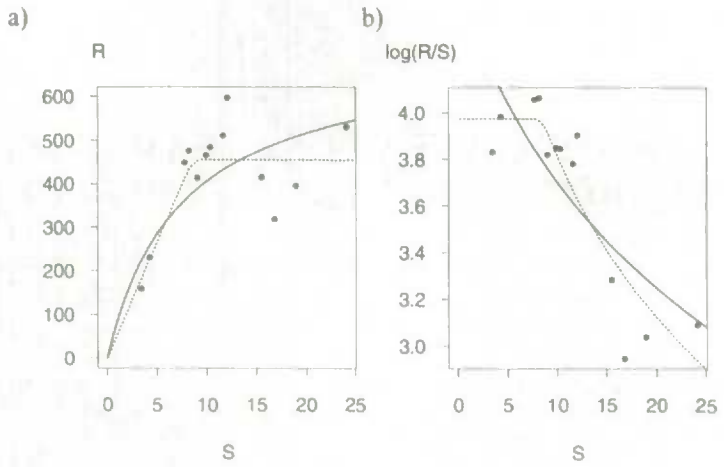


FIGURE 1. Coho salmon spawner-recruitment data set for Deer Creek, Oregon shown on two different scales: (a) recruitment, R, versus spawners, S, and (b) log survival versus spawners. Two maximum likelihood model fits, assuming lognormal recruitment, are shown in each panel: a Michaelis-Menten (solid curve) and a generalized hockey stick (dotted). These models are explained in more detail subsequently in the paper.

One simple, parametric spawner-recruitment model is

$$R = \frac{\alpha S}{1 + S/K}, \tag{1}$$

proposed by Beverton and Holt (1957), and also known as the Michaelis-Menten curve in chemistry. It is easily shown that K is the half-saturation point for the curve and that α is its initial slope, representing the average reproductive rate at low abundance. In this paper, our interest is primarily in α because it is critical in the analysis of extinction and recovery rates, and in determining sustainable levels of fishing. For the coho salmon data, α has units of for female smolts per spawning female, which are identical for all of the data sets. However, estimates of α based on individual data sets are rarely precise, and frequently biased. While the model fit for Deer Creek (Fig. 2a) appears reasonable, the model fit for Bingham Creek (Fig. 2b) gives an effectively infinite slope, which is clearly not reasonable biologically.

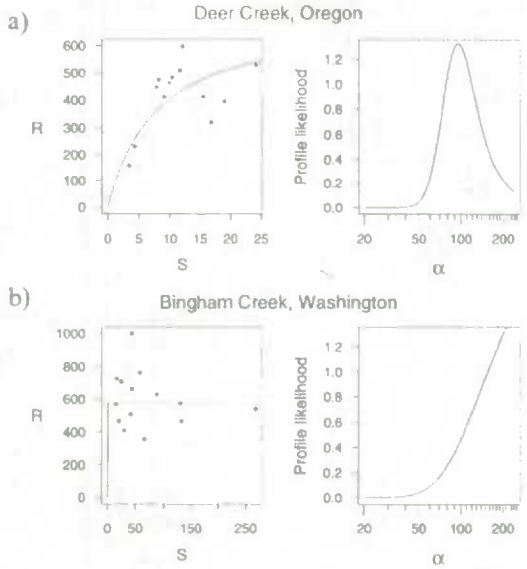


FIGURE 2. Coho salmon spawner-recruitment data sets for (a) Deer Creek, Oregon and (b) Bingham Creek, Washington, with corresponding profile likelihoods for the Michaelis-Mentenα parameter, assuming lognormal recruitment.

2. NONLINEAR MIXED EFFECTS MODELS

The individual estimates discussed in section 1 are clearly inadequate. But how can information be shared between data sets? Suppose we have M populations with data sets of the form (S_{ij}, R_{ij}) for $j = 1, \ldots, n_i$. Letting $y_{ij} = \log(R_{ij}/S_{ij})$, the log Michaelis-Menten model with additive error $\varepsilon_{ij} \sim N(0,\sigma^2)$ is

$$y_{ij} = \log \alpha_i - \log(1 + S_{ij} / K_i) + \varepsilon_{ij}.$$

(Notice that the intercept in this model is log α_i). One way to share information across data sets is by modelling the parameters α_i and K_i as random effects. For this to be valid, the parameters must be measured on the same scale from data set to data set. This is why the data were standardized by river length, as discussed in section 1. Since α_i and K_i must both be positive, we model their logarithms as being jointly normal. The result is a nonlinear mixed effects model to which the maximum-likelihood methods of Lindstrom and Bates (1990) can be applied. For the 14 coho salmon populations such a model gives much more reasonable results than the individual estimates (Fig. 3).

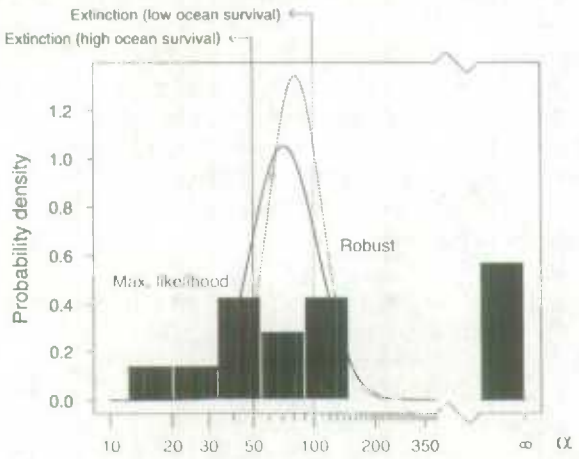


FIGURE 3. Estimates of α for the 14 coho-salmon populations with a superimposed probability density curve from a nonlinear mixed model fit (solid curve labelled "Max. likelihood") and a robust nonlinear mixed model fit (dotted curve labelled "Robust"; explained in section 3). In 4 cases, the estimates of α were effectively infinite. The Bingham Creek population shown in Figure 2 is one example. The vertical lines show cutoff values of α below which populations would be expected to tend towards extinction, assuming 10% ("high") and 5% ("low") ocean survival followed by 20% survival from fishing.

Empirical Bayes fits obtained from the mixed model analysis show a characteristic "shrinkage toward the mean". For example, implausibly high estimates of α, such as that for Bingham Creek, Washington, are replaced by much more reasonable estimates (Figure 4). In some cases, e.g. Hunt's Creek, BC, a single data point seems to have undue influence on the individual fit; this is somewhat improved in the mixed model fit.

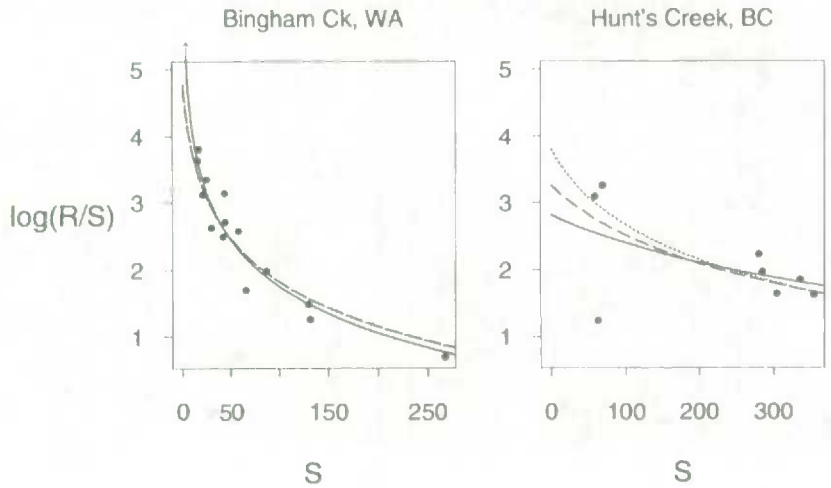


FIGURE 4. Coho salmon log survival versus spawner abundance for Bingham Creek, Washington and Hunt's Creek, British Columbia with fitted transformed Michaelis-Menten curves from individual fits (solid curve), mixed model fits (dashed curve), and robust mixed model fits (dotted curve; explained in section 3). For Bingham Creek, the estimated value of $\log\alpha$, (the y-intercept) from the individual fit is essentially infinite (represented by a vertical arrow at the top of the plot) whereas the mixed model estimate and its robust counterpart coincide at approximately 4.8.

3. INFLUENCE AND ROBUSTNESS

Any meta-analysis raises questions of influence and robustness. For example, is a single study driving the results? A simple way to assess this is a leave-one-dataset-out approach. For the coho salmon data, the parameter estimates seem to be reasonably stable, although one data set, Needle Branch Creek, Oregon, seems to be quite influential.

We might also wonder whether errors in the data might be strongly affecting the estimates. A related question is whether one of the data sets is somehow different from the others, and really ought not to be used in a combined analysis. This is partly a non-statistical issue, depending on the expertise of fisheries scientists (or other subject area specialists). We might hope, however, to identify influential or outlying observations or data sets as part of the data analysis process. Another approach is to develop methods that are robust to the presence of influential or outlying observations or data sets. Many standard statistical methods assume that unknown quantities are normally distributed. This assumption makes methods "fragile" in the sense that a small proportion of outlying values can drastically change estimates. Robust methods generally trade off the efficiency of the standard methods in order to buy robustness. They often provide diagnostic information as well.

Welsh and Richardson (1997) review methods for robust estimation of linear mixed effects models. Such methods generally involve replacing the sum of squares in the log likelihood by a less rapidly increasing function. These methods can be extended to the alternating two-step algorithm of Lindstrom and Bates (1990) for nonlinear mixed effects models. For the coho salmon data, the estimated distribution of α obtained using the robust procedure has a higher mean and smaller variance than that obtained using maximum likelihood (Fig. 3). For some populations, e.g., Hunt's Creek, the fit obtained using the robust procedure is quite different from the maximum-likelihood fit due to downweighting of outliers (Fig. 4). Some caution is advisable. With its lower estimate of the mean of α, the robust fit is less precautionary, and its lower estimate of the variance of α may not be trustworthy.

4. MODEL ROBUSTNESS

A different kind of robustness issue also arises: how confident are we that the assumed functional form relating spawner abundance and recruitment is correct? The Michaelis-Menten model assumes that survival increases monotonically with decreasing spawner abundance, an assumption that is questionable for territorial species such as coho salmon. A piecewise-linear model known as the "hockey stick" (Fig. 5a) may be more appropriate. However, the sharp bend in the hockey-stick model may not be realistic and smoother "generalized hockey sticks" (Fig. 5a) have also been proposed (Barrowman and Myers, submitted).

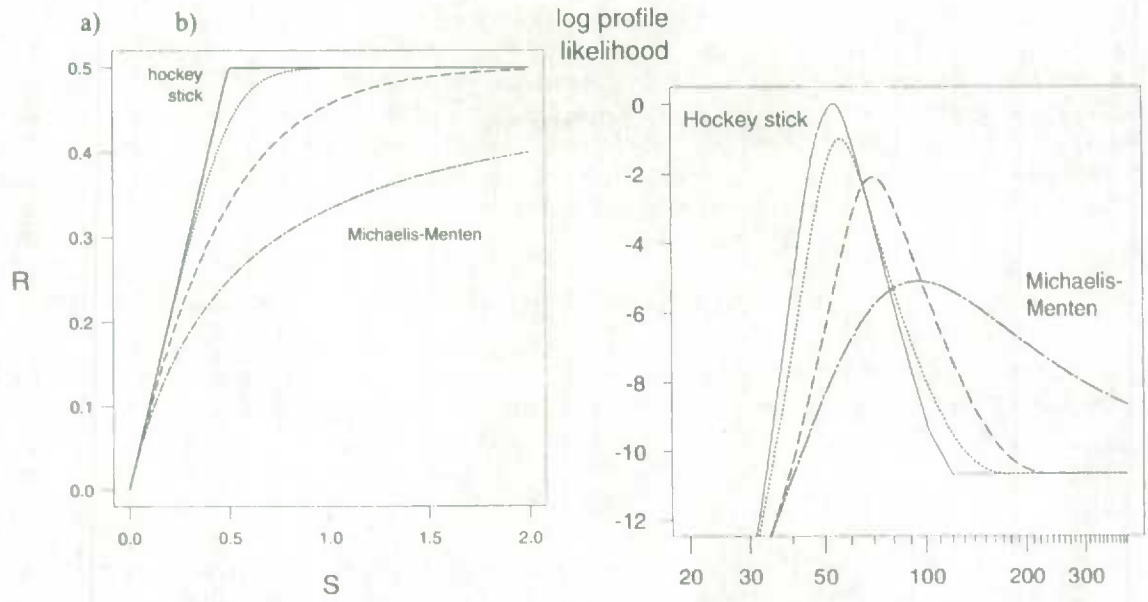


FIGURE 5. Generalized hockey-stick models compared to the Michaelis-Menten: (a) recruitment as a function of spawner abundance for generalized hockey-stick models with varying smoothness (solid, dotted, and dashed curves) and for a Michaelis-Menten (dot-dash curve); (b) log profile likelihoods for α from the corresponding models shown in (a). A difference in maximized log profile likelihood of approximately 2 is significant at the 95% level. In comparisons between generalized hockey-stick models this is equivalent to a likelihood ratio test; in comparisons between a generalized hockey-stick model and the Michaelis-Menten, this is equivalent to using Akaike's information criterion.

Profile log likelihoods for individual populations can be used to obtain confidence intervals for model parameters and to compare the fits of competing models. For the coho salmon data, the hockey-stick model and its generalizations often fit better than the Michaelis-Menten (Fig. 5b). The generalized hockey-stick can also be used in nonlinear mixed effects models. For the coho data, the result is an estimated distribution of α with lower mean and higher variance than that obtained using the Michaelis-Menten (Fig. 6).

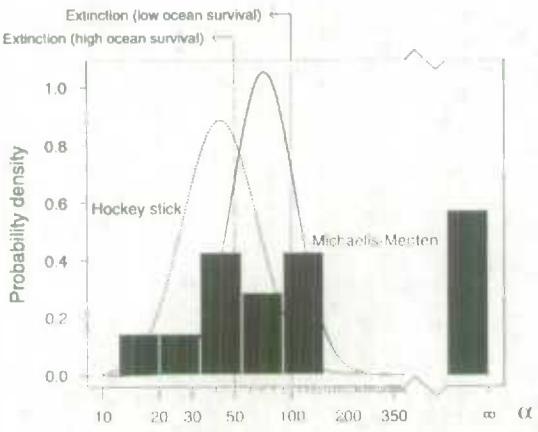


FIGURE 6. Estimates of α for the 14 coho-salmon populations with a superimposed probability density curve from a Michaelis-Menten mixed model fit (solid curve) and a hockey-stick mixed model fit (dotted curve). For more details see caption for Fig. 3.

5. CONCLUSIONS

Research into these methods is ongoing. Robust estimation of nonlinear mixed effects models requires further study. In particular we intend to examine robust estimation for mixed models using the generalized hockey-stick models. Our results suggest that choice of model can be critical and a robust procedure for model choice is desirable. Ultimately, a decision-theoretic approach may be needed, in order to address the role of estimation explicitly in a larger decision-making process.

ACKNOWLEDGMENTS

This work was supported by the Killam Foundation. We thank Chris Field for guidance with robust methods and David Hamilton for helpful suggestions regarding the hockey-stick models.

REFERENCES

Beverton, R.J.H., and Holt, S.J. (1959). A review of the lifespans and mortality rates of fish in nature, and their relation to growth and other physiological characteristics. *in* Wolstenholme, G.E.W., and O'Connor, M. eds. *CIBA Foundation Colloquia on Ageing 5.* 142-174.

Hunter, J. G. (1959). Survival of pink and chum salmon in a coastal stream. *Canadian Journal of Fisheries and Aquatic Sciences*, 16, 835-886.

Lindstrom, M.J., and Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46, 673-687.

Myers, R.A., Bowen, K.G., and Barrowman, N.J. (1999). The maximum reproductive rate of fish at low population sizes. *Canadian Journal of Fisheries and Aquatic Sciences*. In press.

Welsh, A.H., and Richardson, A.M. (1997). Approaches to the robust estimation of mixed models. *in* Maddala, G.S., and Rao, C.R. eds. *Handbook of Statistics, Vol.15.* Elsevier Science B.V.

AUTHOR LIST
IN ALPHABETIC ORDER

KEY WORD LIST

This list is based on the *key words* provided by the authors in their abstracts. Note that some papers are not referred to in this list because the authors did not provide any key word.