3

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2003
•
VOLUME 29
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 2003 • VOLUME 29 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

July 2003

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa

Statistics Canada Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of *Survey Methodology* (Catalogue no. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $30 ($15 × 2 issues). Subscription order should be sent to Statistics Canada, Dissemination Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling 1 800 700-1033, by fax 1 800 889-9734 or by E-mail: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 29, Number 1, June 2003

CONTENTS

Waksberg Invited Paper Series

Special section on Small Area Methods

Regular Papers

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 1-2
Statistics Canada

In This Issue

This issue of *Survey Methodology* contains the third in an annual invited paper series in honour of Joseph Waksberg. A brief description of the series and a short biography of Joseph Waksberg were given in the June 2001 issue of the journal. I would like to thank the members of the Award Selection Committee, Chris Skinner (Chair), David Binder, Paul Biemer and Mike Brick for having chosen Tim Holt, who has had a very distinguished career in both academia and in official statistics, as the author of the third paper in the Waksberg Invited Paper Series.

In his paper entitled "Methodological Issues in the Development and Use of Statistical Indicators for International Comparisons", Holt first describes the wide range of national statistical indicators suggested by various United Nations committees to monitor and compare development in such areas as demography, health, economics and employment, and he considers how these can be prioritized for implementation. He then discusses the need for sound statistical infrastructure in each country, and the importance of base population estimates, administrative sources of data, and good meta-data for indicators that are produced. Holt goes on to discuss several methodological issues related to the implementation of such indicators, and interpretation of international comparisons.

The next six papers in this issue form a special section on small area estimation. The first three papers present general methodology, while the last three discuss small area estimation methods in more specific contexts.

Meeden presents a new Bayesian approach to small area estimation. Instead of using the usual Bayesian approach that implicitly links one area to another area, Meeden instead uses a noninformative or objective Bayesian approach. It applies a Polya posterior idea to obtain model-based estimates of small area parameters, all without introducing a model or a prior explicitly. One advantage of this approach is that population parameters other than means can be estimated with sensible estimates of their precision.

You, Rao and Gambino approach the problem of estimating unemployment in small domains by using an extension of the well-known Fay-Herriot model by borrowing strength across both areas and time. The authors use the structure of the Canadian Labour Force Survey to produce some interesting variations on this model. They use the short period – 6 months – that rotation groups are in the sample to produce efficient Hierarchical Bayes estimates which neatly avoids the seasonality problem common to designs with longer time periods. The result of this method is large reduction of the coefficient of variation especially in the smaller areas.

In their paper, Lehtonen, Särndal and Veijanen examine the effect of model choice for different types of estimators of domain totals. They point out that earlier literature on small domain estimation has not emphasized enough the distinction between the types of estimators and the model choice. They show analytically and empirically that model improvement has different effects on different estimator types. One of their main results is that, under some conditions, model improvement leads to a larger decrease in mean squared error in smaller domains for the generalized regression estimator. The opposite holds for the synthetic estimator. Also, model improvement is in general more beneficial to the synthetic estimator than to the generalized regression estimator since the former can have a large bias.

Chung, Lee and Kim consider small area estimation using the Korean Economically Active Population Survey. They compare synthetic estimation, a composite estimator that combines the synthetic and direct estimators, and a hierarchical Bayes estimation method based on multi-level modelling. They describe the estimators and the model selection for the hierarchical Bayesian approach. They find that all of these approaches improve significantly over direct estimates for unplanned small areas; however, the composite estimator was best overall.

Di Consiglio, Falorsi, Falorsi and Russo empirically compare several small area estimators using data from the Italian Labour Force Survey to estimate numbers of employed, unemployed, and persons looking for jobs within Local Labour Market Areas. Auxiliary data and target parameters are based on census data. Comparisons are done both conditionally on realized sample sizes within a small area and unconditionally. Several types of small area estimator – expansion, post-stratified ratio, synthetic, composite, sample size dependent, and empirical best linear unbiased predictors – are compared. They conclude that the best estimators overall are a composite estimator and a sample size dependent estimator.

In the final paper of the special section, Harter, Macaluso and Wolter present a case-study of small domain estimation techniques to estimate employment at the county/industry division level using data from the U.S. Current Employment Statistics program and lagged administrative data on employment. They discuss such issues as the availability, quality and choice of auxiliary data, problems in micro-matching of survey and administrative data, and regular monitoring of the entire process in order to build in the quality needed to support small area estimation.

The paper by de Waal deals with the error localization problem: the identification of erroneous fields in erroneous data. A well known method to solve this problem in numerical data is based on vertex generation, in particular the Chernikova algorithm. De Waal extends this approach to identify errors in a mix of categorical and numerical data. The paper shows that many results for numerical data also hold true for a mix of categorical and numerical data. This paper provides a nice readable introduction to Error Localization and its implementation.

Haziza and Rao discuss the problem of unweighted imputation for missing survey data. They show that unweighted imputation, unlike weighted imputation, generally leads to biased estimators under the design-based approach (*i.e.*, uniform response). They propose a bias-adjusted estimator which is simple to obtain and has the desirable property that it is approximately unbiased under both the design-based and the model-based approaches. They also derive linearization variance estimators for the proposed estimators. A simulation shows the good performance of the bias-adjusted estimator, especially when the correlation between the variable of interest and the inclusion probability is high.

The paper by Johnson and Deely develops optimal and approximately optimal fixed cost sampling allocations for simultaneous estimation in multiple independent Poisson processes based on the Bayes risk and the Bayes estimator under two different loss functions. The results from this approach are straightforward, interesting and are connected to the classical stratified random sampling allocations. Techniques for finding "representative" conjugate priors, under more general hierarchical models for allocation purposes are also presented.

In the last paper of this issue, Tracey, Singh and Arnab investigate calibrating to the second order moment of a auxiliary variable, when available, to improve the efficiency of estimators. They show that this new estimator can be more efficient than the combined regression estimator in stratified sampling and provide a variance estimator for the new estimator. Finally, they extend the method to double sampling and conclude with some limited simulation results.

Finally, we note that a paper from the December 2002 issue of this journal has just won an award. The paper by Balgobin Nandram, Geunshik Han, and Jai Won Choi, entitled "A Hierarchical Bayesian Nonignorable Nonresponse Model for Multinomial Data from Small Areas", has received the Statistical Science Award as the best paper of the year in applied statistics, awarded by the Statistical Awards Ceremony Committee of the Centers for Disease Control and Prevention and the Agency for Toxic Substances and Disease Registry. Congratulations to Drs. Nandram, Han and Choi!

M.P. Singh

Waksberg Invited Paper Series

Survey Methodology has established an annual invited paper series in honor of Joseph Waksberg, who has made many important contributions to survey methodology. Each year, a prominent survey researcher will be chosen to author a paper that will review the development and current state of a significant topic in the field of survey methodology. The author receives a cash award, made possible through a grant from Westat in recognition of Joe Waksberg's contributions during his many years of association with Westat. The grant is administered financially and managed by the *American Statistical Association*. The author of the paper is selected by a four-person committee appointed by *Survey Methodology* and the *American Statistical Association*.

The author of the Waksberg paper is announced at the annual Joint Statistical Meeting during the American Statistical Association Presidential Address and Awards session. In this session, recipients of awards such as Section, Chapter, Continuing Education-Excellence and other co-sponsored awards are congratulated. In particular, the Waksberg Award for outstanding contributions in the theory and practice of survey methodology is highlighted. Finally, the winner of the Waksberg award appears in the Awards program booklet.

Previous Waksberg Award Winners:

Gad Nathan (2001)
Wayne A. Fuller (2002)

Nominations:

Nominations of individuals to be considered as authors or suggestions for topics should be sent to the chair of the committee, J. Michael Brick at Westat, 1650 Research Boulevard, Rockville MD, U.S.A. 20850-3129 by e-mail at brickml@westat.com or by fax (301)294-2034. Nominations and suggestions for topics must be received by December 5, 2003.

2003 WAKSBERG INVITED PAPER

Author: Tim Holt

Tim Holt began his career as a survey methodologist at Statistics Canada from where he moved to the University of Southampton where he is currently professor. He has published a number of papers in academic journals. He has also been Director of the Office for National Statistics and Head of the UK Government Statistical Service. More recently he has undertaken various consultancies including for the United Nations, European Union, International Labour Office and International Monetary Fund.

MEMBERS OF THE WASKBERG PAPER SELECTION COMMITTEE (2002-2003)

J. Michael Brick (Chair), *Westat, Inc.*
David R. Bellhouse, *University of Western, Ontario*
Paul Biemer, *Research Triangle Institut, U.S.A.*
Gordon Brackstone, *Statistics Canada, Ontario*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 5-17
Statistics Canada

Methodological Issues in the Development and Use of Statistical Indicators for International Comparisons

DAVID HOLT[1]

ABSTRACT

International comparability of Official Statistics is important for domestic uses within any country. But international comparability matters also for the international uses of statistics; in particular the development and monitoring of global policies and assessing economic and social development throughout the world. Additionally statistics are used by international agencies and bilateral technical assistance programmes to monitor the impact of technical assistance.

The first part of this paper describes how statistical indicators are used by the United Nations and other agencies. The framework of statistical indicators for these purposes is described and some issues concerning the choice and quality of these indicators are identified.

In the past there has been considerable methodological research in support of Official Statistics particularly by the strongest National Statistical Offices and some academics. This has established the basic methodologies for Official Statistics and has led to considerable developments and quality improvements over time. Much has been achieved. However the focus has, to an extent, been on national uses of Official Statistics. These developments have, of course, benefited the international uses, and some specific developments have also occurred. There is however a need to foster more methodological development on the international requirements. In the second part of this paper a number of examples illustrate this need.

KEY WORDS: Official Statistics; Statistical Indicators; International Comparisons.

1. INTRODUCTION

Official Statistics matter in national life. They are used to develop and monitor public policies, allocate resources, support public administration and decisions made by businesses. They are used too by citizens as a window on the work of government and to monitor its performance.

As important are the international uses of Official Statistics. They are used by national governments to monitor the country's performance against comparators; to ensure that economic competitiveness is maintained or enhanced; to monitor economic and social developments in other countries and the outcome of alternative economic or social policies that other states may adopt. Increasingly in some regions they are used for national participation in international decision-making and resource allocation. For these purposes internationally comparable statistics are needed. They are required too by international agencies to monitor national performance and to make comparisons. The World Bank, IMF and bilateral funding agencies depend heavily on Official Statistics to monitor the impact of policies and technical assistance programmes.

Increasingly statistics and statistical indicators are being used to set and monitor global policies. For example a review of UN Summits and major conferences during the 1990's identified over 280 statistical indicators needed to monitor UN policies made through conference decisions.

Hence the need for internationally comparable statistics has never been greater. This paper has two purposes:

- To describe the current need for internationally comparable statistical indicators for UN and related agency purposes, and
- To suggest that despite the huge investment in methodological research and development to support national statistical needs, there has not been as much emphasis on methodological research supporting the international uses. Some examples will illustrate this.

2. UN STATISTICAL INDICATORS

What is an indicator?

The term "statistical indicator" has come into use particularly in relation to monitoring global policies. One might try to establish what characterizes a "statistical indicator" and what distinguishes it from the range of statistics published daily by National Statistical Offices. There are indicators, such as the Human Development Index, that are artificial constructs that combine disparate measures (GDP per capita, life expectancy at birth, literacy and educational attainment) into a single composite index number. Such indicators are not a statistical estimate of any single population characteristic and are intended only as a very broad and general measure. But most statistical indicators used by the UN, for example, are not of this kind. Rather they are simply statistical estimates of population characteristics (*e.g.* fertility rate, life expectation at birth,

[1] David Holt, Department of Social Statistics, University of Southampton United Kingdom. Former Director of the Office for National Statistics and Head of the UK Government Statistical Service.

GDP per capita). Each of these characteristics can be precisely defined even though the concept may be complex and the measurement difficult. Such statistics are important for both national and international purposes.

Since the statistical indicators are everyday statistics one may question the need for a different terminology. The reason is not based on the statistical properties but may reflect the way that the indicators are used. Indicators are meant to be high level (usually outcome) measures that are perceived to be related to some aspect of economic or social well-being. For example a low life expectancy at birth in a country is an indication of unsatisfactory living prospects and of health problems in particular. But two countries with similar life expectancies may have very different health situations and the policies needed to address these may be quite different. The statistic used as an indicator points at a problem but one would require much greater understanding of age-specific mortality rates, causes of death, the quality and range of health services and possible differences between sections of the population to formulate a policy response. That policy may be based on improved medical provision, preventative public health or social policies, greater education for those at risk or a combination of all of these. The statistical indicator is a high level monitoring instrument but policy development and monitoring require a much wider and richer statistical picture.

The fact that the indicator is used as a general measure of economic or social well-being does not imply that the methodology and sources used to measure it need not be tightly specified. The requirement is to get comparability both between countries and within a country at different points in time. Loosely specified sources and methods can give rise to inconsistencies that would invalidate the monitoring required. Indeed one of the problems of indicator use is that small changes that have no statistical or substantive significance but cause the ranks of countries to change are given far too much prominence particularly by national policy makers and commentators.

UN Statistical Indicators

In the last decade or so United Nations summits and major conferences (averaging almost two per year) have covered a wide range of economic and social issues. These meetings have resulted in declarations related to future goals and targets that have been endorsed by member states and are intended to improve the well-being of the world's population. Goals and targets call for a commitment to monitor progress towards them and, consequently indicators have been identified in relation to each goal. The intention is to monitor and report on these so that progress towards the declared goals and targets can be measured. The Millennium Development Goals, for example, subscribed to by 164 Heads of State or their representatives have resulted in 8 goals, 18 targets and 48 statistical indicators that will be monitored over the coming decades. For example there are two indicators for Goal 1, Target 2:

GOAL 1: ERADICATE EXTREME POVERTY AND HUNGER	
Target 2: Halve, between 1990 and 2015, the proportion of people who suffer from hunger	1. Prevalence of underweight children under-five years of age
	2. Proportion of population below minimum level of dietary energy consumption

In total over 280 indicators had been identified from UN Summits and major conferences in the last 10 years.

This process has gone on with too little co-ordination between officials concerned with the separate UN summits and major conferences in terms of the number and choice of indicators to be monitored. The result is a plethora of indicators of different levels of importance in policy terms. The meetings have varied considerably in terms of the number of resulting indicators (ranging from a handful or less to as many as 70 being identified from a single UN conference). Also there is potential for confusion among users because of an apparent inconsistency and lack of coherence among the indicators.

The UN conferences have adopted markedly different approaches to identifying the need for indicators. In most areas the number of indicators is relatively small and these focus on outcomes. In other areas the indicators are detailed and seek to measure many different facets of policy and service delivery. For Health for example the death rate for a specific disease may be required. Additionally the required indicators may include the disease prevalence rate, the inoculation rate, the proportion of cases treated under a specified treatment regime, public health preventative measures and public understanding of the causes of the disease.

The cumulative effect of indicators added at each conference has resulted in a large demand for statistical information from each member state: a demand that has to be set alongside the demands for statistical information for national policy purposes. For countries with less well-developed statistical infrastructure this total demand can be disproportionate to the resources available to meet it. Indeed some have a concern that the whole global indicator movement has gained too much momentum and the pressure from the UN and international agencies is distorting national priorities and reducing the provision of statistics to support public policy and sound public administration in some developing countries.

Attempts have been made to distil core sets of indicators that might be afforded higher priority. The United Nations Statistical Commission (UNSC) identified the Minimum National Data Set (MNDS: 15 indicators). The OECD Development Assistance Committee – in co-operation with the UN, World Bank and IMF – identified the International Development Goals (IDG: 21 indicators). This set drew heavily on international summits up to 1995. The United Nations Development Group identified indicators to support Common Country Assessment again based on an analysis of the requirements of UN summits (UNDAF-CCA: 57 indicators). Similarly the need to promote and assess sustainable development has resulted in an additional set (CSD: 57 indicators). There is also Basic Social Services for All (BSSA: 12 indicators). Most recently the UN has espoused the Millennium Declaration Goals and associated indicators (MDG). These sets have some common components and some differences as one might expect. Even these attempts illustrate the vagaries of the political process. For example the fact that the IDG indicators were repackaged and replaced within 5 years by the MDG indicators suggests a lack of constancy and political purpose.

In 2002 the UN Statistical Commission (UNSC 2002) adopted proposals to create a framework containing three levels of priority. The 123 most important indicators are allocated to 7 Domains:

- Demography,
- Health and Nutrition,
- Environment and Energy,
- Economics and Poverty,
- Employment and Labour,
- Education, and
- Other Social Indicators.

The Domains represent major divisions of policy responsibility that are commonly reflected by separate Ministries in many countries. Additionally important cross-cutting policy areas such as Poverty, Child Welfare or Gender that are distributed across these Domains are taken into account. Sub-Domains are identified within each Domain as being relatively self-contained policy areas. Indicators are allocated to the three priority tiers:

- First tier priority indicators reflected the need to monitor policies of the highest global and national importance. They represent the indicators that, no matter how limited the statistical capacity available, countries and international agencies would find essential for top-level monitoring of policy effectiveness. There are 2-6 tier 1 indicators per Domain.
- Tier 2 priority indicators mainly covered different policy objectives (different subdomains) from those covered by the highest priority indicators. These policy objectives should be of sufficient importance to merit a tier 2 priority indicator. Not all subdomains would necessarily do so. There are 0-13 tier 2 indicators per Domain with most Domains having much less than 13.
- Tier 3 priority indicators supported policy needs that are, albeit important, either subsidiary or judged to be less important than others. There are 2-8 tier 3 indicators per Domain.

The Criteria for Allocating Priorities to Indicators

Allocating priority must be grounded in the policy need but involves balancing a number of criteria surrounding the relevance to policy, the technical properties and current availability (or the feasibility, resource and statistical capacity implications of achieving an acceptable measure in a high proportion of countries). While one may aspire to the situation in which an indicator fully satisfies all of the criteria, in practice this will not be the case. The extent to which the indicator meets the criteria needs to be considered and a judgement made about whether any shortcomings are of such overriding concern as to disqualify a particular indicator from use.

A large number of criteria may be identified but the most important are:

Policy Relevance

- Indicators must be relevant to the policy requirement.
- Indicators should measure the real policy objective (or provide a proxy measure that is adequate for policy monitoring).
- Indicators should normally have global policy relevance.
- Indicators should be straightforward to interpret: changes over time in any direction should not be ambiguous in relation to the policy interpretation and significant differences between countries should be meaningful in terms of the policy goal.

Technical Properties

- Technical properties of the indicator should be adequate for the purpose, recognising that change over time is often more important than the level of the indicator.
- Indicators that fail to cover the target population fully should have sufficient coverage to ensure that the indicator values are unlikely to mislead policy users (*i.e.*, the potential bias as a measure of the true policy objective should be small).
- If possible, where indicators are difficult to measure for countries with less well-developed statistical capacity, simplified alternatives should be provided for use until the statistical capacity can support the more demanding measure.
- Indicators should be robust to institutional and cultural differences between countries and over time.
- Indicators should exhibit change over time at a rate that would support policy monitoring.
- Indicators should be produced with sufficient frequency and timeliness to support policy monitoring.
- Indicators should conform to international standards if these exist.

In a number of cases the application of these criteria to create the proposed framework revealed examples where the policy objective suggests allocation to a particular tier, but the inherent conceptual or statistical weaknesses of the proposed indicator and/or measurement problems cause the indicator to be allocated to a lower tier.

The numbers for each tier reflect the fact that the indicators are not intended to substitute for the mass of detailed statistical outputs from national statistical systems that support users' needs. They are intended as high level indicators for monitoring purposes.

3. GENERAL ISSUES

The process described in the previous section identified a number of general issues most of which have a technical dimension.

Choice of Indicators and Targets

There are two facets: first the precise form and definition of the indicator needs to be decided together with a methodology for measuring it. In practice, both national and international policy makers are inclined to express their goals directly in terms of a statistical indicator without particular concern for the definitional and measurement issues. Too often an indicator is identified with too little thought. The reality is that identifying statistical indicators for monitoring purposes should be neither a pure policy nor a pure statistical issue. The basic expression of the policy goal must drive the monitoring requirement but turning that expression into a statistical indicator that will be relevant, reliable and acceptable to the various stakeholders is a statistical function. The tension between the policy view of what is needed and the statistical view of what is feasible and technically sound needs resolution.

The second facet is the choice of a target. These are chosen in relation to the indicator (for example to halve the death rate due to a particular disease by a stated year). There are two views about such targets. One is that they should be based on policy analysis and set to reflect what effective policies might be expected to achieve. In this view it is unlikely that the same target is achievable or demanding enough in every country. The second view is that the targets are simply something to aspire to and not based on any reasoned analysis. In this view target setting is entirely a political process for binding countries into a political commitment.

From a statistical perspective the danger of aspirational targets is that they will not be met (or sometimes even approached) and the process of statistical monitoring itself may fall into disrepute as a result. There is also a threat to statistical integrity if the political pressure to show progress against an unrealistically set target is too strong.

Whichever view prevails targets that are framed in terms of improvements from a given base year do require that indicator values are available at that point. Given the lack of statistical capacity in many developing countries this is problematic and for a number of the Millennium Declaration Goals for example the global statistical picture for the baseline year from which progress will be measured is seriously inadequate.

Statistical Capacity

The ability to produce consistent, reliable statistical information requires a sustained statistical capacity. This requirement is not a one-off capability but implies the ability to produce statistics on a regular basis and with the timeliness needed.

In particular a sound statistical infrastructure is essential. By this is meant:

- Underpinning systems to create and maintain sampling frames for business and household surveys.
- A critical mass of ongoing statistical activities: survey design, data collection and analysis in order to nurture the basic professional skills.
- The technical and methodological capacity to maintain and develop systems in accordance with international standards as these are developed over time.
- A developed analytic capacity.
- Adequate statistical frameworks and IT infrastructure.
- Good management to make the most use of the resources that are available.
- All of the above embedded within a wider legal and administrative structure that recognises the importance of good statistical information and the need to sustain the conditions in which it can be produced with high professionalism and integrity, consistent with the *UN Fundamental Principles of Official Statistics*.

Without this core capacity and the ongoing resources to support it, neither the statistical needs of the country nor those of the international community will be reliably served. In many countries adequate ongoing financial support is a key issue. Where this core capacity is fragile the sporadic provision of additional funds from international or bilateral funding agencies to satisfy a particular statistical need will be much less effective and is no substitute for developing what one might term "statistical sustainability".

In this regard, statistical indicators need to be viewed as the end product of often complex statistical infrastructures that are essential if the indicators are to be produced with adequate quality. Too much emphasis has been placed on the indicators and too little on the statistical sources and infrastructure that underpin these.

Indicators as Rates and Ratios

International comparisons require that statistics be put on a basis that is immediately comparable and for this reason almost all of the indicators are presented as rates, proportions or in per capita terms. This places population estimates as a cornerstone of most of the statistical indicators. These depend on periodic Censuses to provide benchmarks and on systems of vital registration or other sources to permit inter-censal population estimates. Different statistical indicators call for population estimates for various age-sex groups as the appropriate denominator.

A particular difficulty is that the numerator of such indicators and the population denominator are often provided from different sources within a country and may be inconsistent. Hence the rates, when calculated, may not be recognised within the lead policy Ministries and can be challenged by them leading to a loss of confidence in the statistics. Population estimates from the National Statistical Institute, a policy Ministry and the UN Population Division may all differ. In extreme cases different population denominators may be used for different policy areas. This is clearly unsatisfactory and when it occurs implies a systemic problem of consistency and quality assurance and a lack of statistical co-ordination within a country.

For economic measures indicators are often expressed as per capita measures (in which case the comments above apply) or as ratios of expenditure (*e.g.* for health or education) in relation to GDP. Complex measures such as GDP require an extensive framework of business surveys, administrative sources and underpinning infrastructure if the statistics are to be of adequate quality.

The pervasive use of GDP and of population estimates in this way underlines the importance of the quality of these estimates if other indicators are to be sufficiently reliable. Both require a strong statistical capacity and infrastructure if they are to be regularly produced.

Inadequate Administrative Sources

There are a large number of indicators that are derived from administrative systems in countries where these are well established (*e.g.* mortality rates by cause, fertility rates, gross and net enrolment rates in education and many health indicators concerned with health services and provision). For some kinds of information often relating to public services (*e.g.* numbers of teachers, doctors or nurses and qualifications) the only realistic sources of information are administrative and where these are inadequate they need to be strengthened. For other measures a household survey may be an alternative although there can be conceptual and measurement differences between information obtained from administrative and survey sources.

Nonetheless, in countries where the administrative systems are inadequate survey based measures are widely used in which both the numerator and denominator of the indicator may be derived consistently from survey estimates. In this case a special survey devoted to one particular area of interest (*e.g.* health and fertility history) can provide a wide range of statistics. This is a viable possibility (at a cost) particularly when countries want a more comprehensive picture of a situation.

However, ad hoc surveys cannot provide the ongoing information needed to track important indicators. To ensure that critical information will be available on an ongoing basis it is necessary to invest resources into the statistical infrastructure so that surveys can be repeated regularly.

In general, even when they purport to measure the same thing, both administrative sources and surveys have strengths and weaknesses. The administrative source is often large and provides the opportunity to provide regional or local figures. However the concept contained is often not ideal for the statistical purpose. Also the source may not cover the whole population or may suffer from various inadequacies. Surveys can often measure the concept required but sample sizes are often small and there may be differences between the surveyed population and that intended because of inadequate sampling frames, response problems and measurement error.

The real methodological challenge is not to decide that one source is preferred to the other but to use all of the information available to produce the highest quality estimates possible. This will often require strong methodological effort if the statistics are to command confidence. However these data reconciliation problems often occur in countries where the methodological expertise is not strong.

Measuring Levels

There are some topics particularly concerning environmental indicators where the very idea of a measure of level may be very difficult to frame. It is often not the absolute level of the indicator so much as the trend over time within each country that is the key focus of policy.

For example there is no real meaning in measuring the average toxicity in Canada's coastal waters. One would need to define coastal water precisely and the sampling methods to achieve a representative sample of coastal water together with appropriate methods of statistical inference. In particular there would be a methodological question as to whether the sample should be weighted to represent the distribution of coastal water or that of the adjacent coastal population. In practice samples taken on a consistent basis from the same locations on repeated occasions will not provide a measure of toxicity level but will, under some strong assumptions, allow trends to be monitored. However development (such as new towns and industrial sites) will lead to new sources of toxicity over time and the location of sample sites may need to be reviewed to reflect this. At the same time data analysis will be needed to avoid the measured trends exhibiting discontinuities. The development of sample designs and methods of inference for populations of people and businesses has been one of the great achievements of Official Statistics. But there are some substantial unresolved methodological issues in designing and analysing samples of physical populations to an equivalent rigorous standard. The methods applied generally in Official Statistics may offer some contribution.

Meta-Data

This is essential if users are to understand any particular issues affecting the statistical indicator values for any country. Good meta-data (such as is required by the IM's SDDS and GDDS) is a general requirement but there are specific situations when countries should ensure that specific meta-data is provided.

- When national priorities result in an indicator which is not fully comparable with those produced by other countries. Failure to provide informative meta-data will fail those users who seek to use the indicator for comparative purposes.
- Where national standards or targets are adopted (for example in setting a national poverty standard) the basis of this measure needs to be available to users.
- Population forecasts (and inter-censal estimates in countries where vital registration systems are unreliable or unusable) will depend crucially on the data sources and assumptions made about age-specific fertility rates for example. A clear specification of the underpinning assumptions is essential to users.

Distributional Measures

A number of indicators call for separate analyses by sex and as a general rule if the data source can support it then this should be routinely provided. The same applies to analysis by subgroups (*e.g.* region, age-group, ethnic or social classifications). There is a broader issue about providing indicators that measure inequality and distribution within each country. There are a rather small number of indicators that focus on distributional issues (*e.g.* share of consumption by lowest quintile of population) but the large majority of indicators are based on national averages. This is a significant deficiency in the existing indicator list. Much deprivation and inequality in the world will be masked by indicators based on national averages. Analyses by subgroups (*e.g.* by gender, region, age group, income groups, ethnic or social classifications) where feasible would illuminate this issue much more. Similarly, additional measures of inequality, such as the ratio of consumption by the highest 20% of households to the lowest 20% have much to commend them.

4. SOME SELECTED METHODOLOGICAL ISSUES

In the second part of this paper a small number of methodological issues are discussed specifically in the context of international comparisons.

4.1 The Methodological Paradigm

In general the paradigm adopted by Official Statisticians to ensure comparability is based on several components:

- Conceptual clarity of the item to be measured.
- Precise definitions of relevant terms that can be applied in practice.
- And precisely defined classification systems.
- A clear specification of the target population to which the estimates apply.
- Development of appropriate sources and methods, even questionnaires, to obtain the data and compile it into the estimates required.
- Often, international standards, manuals and descriptions of best practice to cover all or most of the above.

The basic assumption is that if the measuring instrument and related methodology can be defined precisely enough then it can be applied independently in different countries and the resulting statistics will be internationally comparable. Hence: control the measurement process and the outputs will be comparable.

This approach generally yields relatively comparable statistics but not absolutely so and not all of the time.

4.2 Literacy

It is, of course, well known that the translation of some measures from one language and culture to another is fraught with difficulty and measuring functional literacy is an example. In any one country one can test comprehension of a text that is grounded in everyday experiences and the requirements of daily life. But the task of transferring this into another language and culture and getting a precisely comparable measure of functional illiteracy is very difficult. Even when great effort has been made to achieve this (*e.g.* the International Adult Literacy Study 1999) it may be that only approximate comparability can be achieved especially if the same measures are used over time so that within country changes may be monitored. In practice literacy measures for almost all countries are much cruder; for example a self-assessed respons'e to a Census question such as "Can the person read a letter? This approach may provide a broad estimate of the number of people who can read to a certain level in some circumstances but is unlikely to provide comparability either between countries or within a country over time. Large changes in the level of literacy within a country may provide evidence of real change but small changes may simply reflect the unreliability of the measure. In order to monitor literacy levels for global policy emphasis is placed on 15-24 year olds since these reflect the flow of newcomers to the adult pool and improvements in educational access and attainment will show larger changes to literacy levels for this group than for the adult population as a whole. Hence the inherent weaknesses in the measure may, to an extent, be mitigated by focussing on a group for which large change may be expected. Such an approach will, however, miss the effect of adult literacy programmes.

4.3 Interactions Between the State and the Citizen

International comparability is made more difficult whenever we seek to measure something that is affected by the interaction between the state and the citizen because the way in which the state provides for particular services may differ from country to country. In these cases precisely the same measuring instrument applied in different countries may give different results. Consider for example the case of housing provision for low income families. In some countries this is provided free or for very low rent. In others the rental cost is at market levels but families get state benefits to allow the payments to be covered. Hence the mechanisms by which the state interacts with the individual will affect important economic measures. As a consequence the international comparability of statistics collected and compiled under precisely the same conceptual framework can be impaired. In some circumstances money flows are imputed to reduce the discrepancies but this is impractical if the provision of cheap housing is very widespread.

Similar issues can arise for medical provision. In some countries medical services are provided by the state, free at the point of access and directly funded from taxation. In others the system is funded through a system of medical insurance which may have elements of both state and personal contributions. When medical services are provided to an individual the real flow of payments may vary. The medical practitioner may make a direct claim (to the state or to a medical insurance fund) for the services provided. In other systems the individual may be the formal claimant but with the payment made directly to the medical practitioner. Or the individual may be required to pay the costs and to claim these back from the state or insurance fund as a payment back to the individual. To some extent these arrangements may be regarded as alternative ways of achieving the same end: a state-facilitated system to ensure that individuals have good access to medical services. In practice money flows are imputed to eliminate most of the institutional differences.

A third example is the estimation of tax revenue which has a direct impact on the estimation of public expenditure and government deficit. Under SNA93 this assessment is made on an accruals basis and in the year in question will be based on the tax assessments made to individuals and businesses. In countries that use well-established self assessment methods and a high level of tax collection through employers the difference between the estimate of tax to be collected and that which is subsequently achieved in the following years may be very small (there will be companies that cease to function and default on the tax liability and people who may die without leaving an estate sufficient to cover the tax due). In other countries with different forms of tax assessment and recovery practice there may be much larger differences between the tax assessed and that which is eventually recovered. Where a shortfall occurs this will in due course be written off against the financial account. But this write-off will have no impact on the estimates of public expenditure and government deficit. Hence a system that "optimistically" estimates the level of taxation that will finally be recovered will result in a lower estimate of government deficit that will never be corrected when the tax shortfall is written off against the financial account. Given the importance attached by international bodies to levels of public expenditure and government deficit a lack of comparability in these key measures matters. In this third example there is no universally agreed method to eliminate the differences although in the European Union specific (non-SNA) rules have been introduced in the debt and deficit manual to eliminate the discrepancies.

4.4 Comparing Economic Measures – Purchasing Power Parities (PPP)

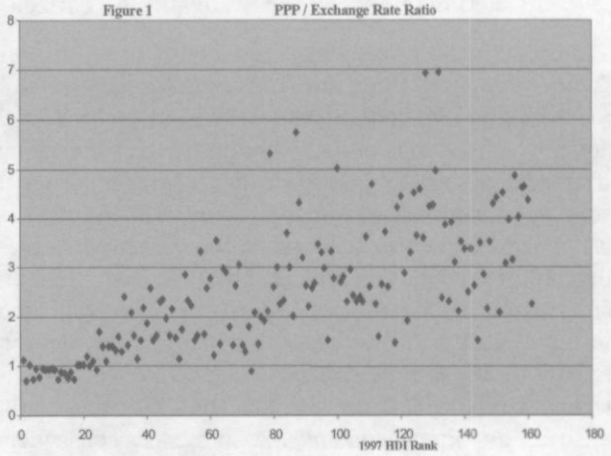
For comparative purposes economic measures (*e.g.* GDP, per capita income or expenditure on Health or Education, living standards) that are measured in national currencies must be converted to a common unit of measurement.

The point at issue is whether conversion from national currencies to a common unit (say US$) should be made using the comparative exchange rate values of different currencies, or should be made on the basis of equalizing the purchasing power of the currency. This is an important issue that can have a profound effect on international comparisons. For example in 1999 the Human Development Report (HDR) claimed that "the gap in per capita income (GNP) between the countries with the richest fifth of the world's population and those with the poorest fifth widened from 30:1 in 1960, to 60: 1 in 1970, to 74:1 in 1995." These statistics are based on exchange rate conversion and yet the corresponding PPP ratios are about 12:1 in 1960, 18:1 in 1990 and 16:1 in 1997. Not only are the ratios much smaller but the clear upward trend presented in the HDR figures is not apparent in the measure expressed in PPP.

The exchange rate conversion values of any currency are determined by the international financial markets and reflect the market forces in those institutions. Indeed, in the modern world, exchange rates are little affected by international trade and the exchange of goods and services in world markets. The second approach uses Purchasing Power Parities (PPP) to reflect domestic prices on an internationally comparable basis. The value of national income or economic output in any country is equated to others on this basis. In this approach, the PPPs provide an international valuation of what the local currency will buy within the country (United Nations 1992).

Figure 1 shows a plot of the ratio of exchange rate conversion to PPP conversion for most of the countries of the world. The x-axis is the 1997 Human Development Index (HDI) rank of each country. The most industrialized countries occupy the lowest 20 places at the left of the graph and the further right one goes the lower the level of development of the country as measured by the HDI. For the industrialised countries the ratio of PPP to US$ exchange rate conversion factors is fairly close to 1. However, for less developed countries the ratio is greater – in many cases much greater – than 1. The upwards slope of the plot shows that the ratio of PPP to US$ exchange rate conversion is generally larger, the lower the HDI rank. For the least developed countries the ratio can be as much as 4 or more. Hence, because the ratio is close to 1, a comparison of economic measures between the United States and a major European country, for example, would be fairly similar using either exchange rates or PPP conversions. However, a similar comparison between the United States, or any of the most industrialized countries, and a least developed country would be very different. In such a case, the conversion of per capita income, for example, using PPP conversion could be as much as 4-6 times larger than the conversion using exchange rates (an exchange rate measure of GDP per capita of $1,000 would be $4,000-6,000 in PPP terms). Hence, the choice of conversion factor has a significant effect across the developed/developing spectrum.

Figure 1 PPP / Exchange Rate Ratio



Source UNSC (2001)

Powerful reasons exist for using PPP conversion rather than US$ exchange rate conversions for real economic (rather than purely financial) phenomena such as standard of living comparisons (as reflected by per capita GDP) and, by extension, for comparisons of economic output (GDP) and national income (GDP or per capita GDP). (UNSC 1998).

Table 2

International Comparisons: Ratios of Per Capita Measures of Output or Use of Goods and Services

Comparison	Daily per capita suply of calories, 1996	Daily per capita supply of fat, Total (grams), 1996 (a)	Daily per capita supply of protein, Total (grams), 1996 (a)	GDP Index	TVs, per 1,000 people, 1996	Carbon dioxide emissions per capita (metric tons), 1996	Com'l energy use (oil equiv'nt) per capita (kgm), 1996	Per capita electricity consum'n, 1996	Main telephone lines, per 1,000 people, 1996	Intern'l tourism departs per 1,000 people, 1996	Personal computers per capita	Real GDP per capita (PPP$), 1997	Per capita GDP (US$) 1997
Japan/China	1.0	1.3	1.3	1.6	2.8	3.3	4.5	9.1	10.9	32.6	42.7	7.7	45.9
Sing/Indonesia				1.6	1.6	16.3	11.7	18.8	24.4	111.0	45.2	8.2	26.8
Korea/Vietnan	1.3	2.3	1.5	1.7	1.8	0.8	8.0	23.1	26.9		39.9	8.3	29.8
Mexico/ Nicaragua	1.3	1.8	1.6	1.5	1.1	5.3	2.9	3.9	3.7	1.6		4.2	10.0
SA/Mozbique	1.6	2.4	2.1	2.2	41.0	69.0	5.2	58.9	33.3		47.1	10.0	21.9
SA/C African Rep	1.5	1.2	1.6	1.7	24.6	69.0		125.4	33.3	0.8		5.5	11.3
Brazil/Ecuador	1.1	0.8	1.3	1.1	2.0	0.8	1.4	2.6	1.3		4.7	1.3	3.0
T&T/Haiti	1.5	2.4	1.5	1.7	63.6	86.0	22.9	40.2	21.0	11.5		5.4	12.6
Sey'lls/Sri Lanka	1.1	1.5	1.5	1.4	2.3	5.8		7.2	14.0	98.0		3.3	6.1
Sey'lls/India	1.0	1.6	1.3	1.6	3.0	2.1		3.0	13.1		10.3	4.9	12.7
Kuwait/Jordan	1.1	1.2	1.4	1.6		10.1	7.9	14.2	3.9		3.4	7.3	15.6
Lebanon/Jordan	1.2	1.4	1.2	1.2		1.8	1.1	1.7	2.5	19.0		1.7	4.2
Egypt/Ethiopia	1.8	2.6	1.5	2.1	31.5		2.2	36.4	16.7	41.9		6.0	10.6
Maur's/Madag'r	1.5	2.6	1.7	2.1		15.0		25.3	54.0			10.0	16.7

The approach due to Castles (2000) is illustrated using a range of bilateral comparisons of countries from the same region in Table 2. The ratio of per capita consumption of various items in each pair of countries is presented, together with ratios of the per capita GDP for the two countries based on PPP conversion and exchange rate conversion. A general pattern may be discerned. For items such as food consumption, which are price inelastic, the bilateral comparisons are relatively close to 1, with countries with higher per capita GDP having somewhat higher consumption. The ratios are much larger for items (*e.g.* televisions or personal computers) that depend on disposable income and are much more price elastic. In general, the PPP comparison for any pair of countries falls within this pattern, having a larger value than the ratios for food consumption but smaller than those for the technological items. This is what one would expect. The exchange rate comparisons, however, are generally much larger and often lie outside the range of consumption even for items such as PCs and televisions.

The PPP measure seems more consistent with the other measures and more relevant for the purposes intended.

There are, of course applications for which exchange rates are appropriate, such as the expression of a country's international debt relative to its GDP.

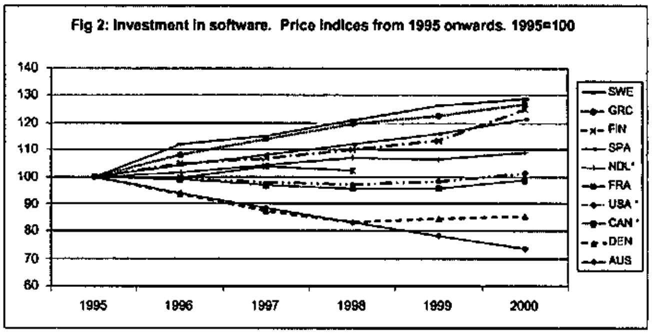
4.5 Price Indexes for International Market Prices

For some goods and services (particularly in Information and Communication Technologies – ICT) the rapid rate of technological change has made it much harder to estimate price changes by normal methods. The element of quality change in simultaneous price and product changes is significant and National Statistical Offices have responded to this by greater use of hedonic regression methods to adjust for quality changes. Even when these methods are applied independently by different countries there can still be large differences in the price deflators that are arrived at and yet, to a large extent the goods and services are traded in an active international market. Similarly it is possible for producers within the same NSO who compile national import and export price indexes to use different price deflators for the same type of goods and services.

These differences matter: within a country they can lead to significant impacts on key statistics such as the balance of trade and fixed capital formation. Between countries they distort the levels of ICT investment being made and the productivity analyses aimed at measuring the impact of ICT investment on growth and economic performance.

Wyckoff (1995) observed that, in the case of computer price indices in OECD countries, large differences in the prices were more likely to reflect methodological differences than real price differences between countries. Lequiller (2001) found significant country differences in the attribution of software expenditure between fixed capital formation and intermediate consumption. The question is whether these differences are due to methodological differences and hence distort international comparisons.

If we consider the case of computer software, for example, Figure 2 illustrates the range of price indexes applied to software by a range of countries. The differences in national estimates of the price indices are dramatic and will have a significant effect on the international comparability of statistics that depend on the price indices.



Source Edwards, Comisari and Johnson (2002) citing Ahmad (2002)

In a separate analysis Colecchia and Schreyer (2001) collate estimates of average annual percentage growth in software investment (1990-95) for a range of OECD countries. These estimates depend on nationally estimated price indices. They also recalculate the values using an internationally harmonised price index. The results are given in Figure 3. The latter raises the mean growth from 6.3 to 8.2. More significantly in terms of international comparisons it lowers the standard deviation from 4.8 to 2.9 making the national estimates more similar.

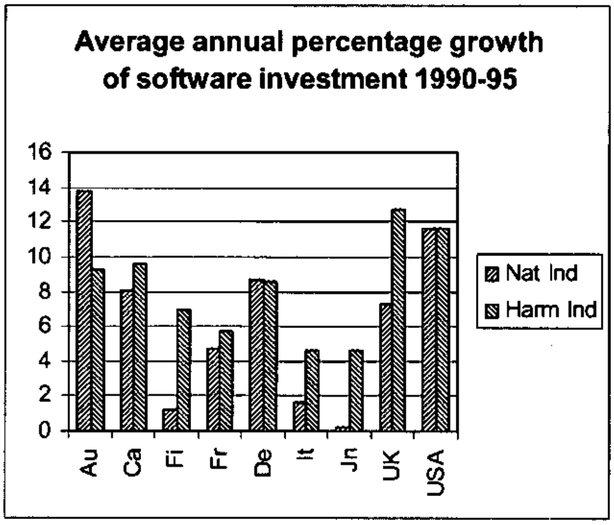


Figure 3

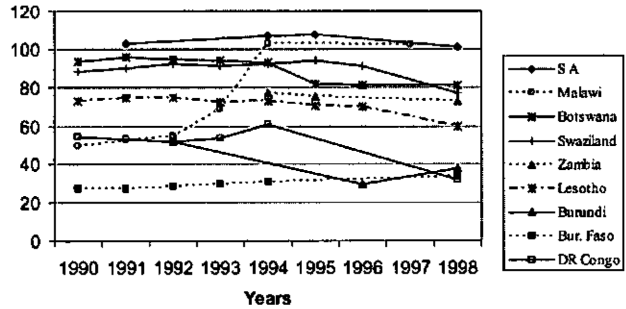
This is an example where broadly comparable procedures applied independently in different countries give such different measures of something that ought to be much the same in all countries that one must question the international comparability of the economic statistics that depend on the measure. In this case the methodological paradigm breaks down.

The basic question is whether it is appropriate for each country to independently apply somewhat similar methods to matters such as price indices for goods and services that have a strong international market. Alternatively it could be argued that to improve international comparability countries should cede an element of national statistical sovereignty by using internationally estimated indices. The issues are what methodology should be applied; to what data (presumably collected on a collaborative basis from a range of countries) and what are the consequential issues for economic analyses of national data. Using coherent estimates of price indices for import and export prices would also need to be considered.

4.6 Imputation and Aggregation

For the purposes of monitoring international policies it is not enough to measure statistical indicators at the national level. Most of the statistical series comprise rates, ratios or proportions. The country level measures need to be aggregated to provide measures at the regional and global level. This requirement generates a number of methodological problems that need further investigation and development.

Fig 4: Primary Enrolment Rates: African Countries



Source: UN Millennium Indicator Database

Figure 4 contains the primary level Net enrolment rates for some African Countries. Although Education statistics have been chosen here these illustrate a number of features that are common to a wide range of statistical indicators and countries:

- The series are incomplete with missing values in some years for all countries and the level of completeness varies from one country to another. Indeed there can be countries with only one figure in the recent past or, for some series, with none at all.
- The last figures available are for 1998.
- The data show different trends with participation rates increasing in some countries and decreasing in others.
- Some countries exhibit sudden changes in the participation rate from one year to the next (*e.g.* Botswana Malawi). Countries may exhibit erratic series (*e.g.* Rwanda).

The objective for inference

The objective for an aggregate statistic at the Regional or Global level needs to be clear. For a regional rate for Africa for example one might naturally assume that the objective is to estimate $Y_{R,T}$ the rate for the region R at time T:

$$\begin{aligned} Y_{R,T} &= \sum_{j\in R} Y_{j,T}\, w_{j,T} \Big/ \sum_{j\in R} w_{j,T}, \\ &= \sum_{j\in R} Y_{j,T}\, \mu_{j,T}. \end{aligned} \tag{1}$$

$Y_{j,T}$ is the corresponding rate for country j and year T and $\mu_{j,T} = w_{j,T}/\sum_{j\in R} w_{j,T}$.

In equation (1) the natural value for $w_{j,T}$ is the population size for the relevant age group in country j at year T. Thus for the enrolment rate data presented above the national enrolment rates would be aggregated to produce the regional (or global) rate. Corresponding estimates of change Δ_{T_1,T_2} between years T_1 and T_2 may be similarly defined at the national, regional or global level. For example:

$$\Delta_{R,T_1,T_2} = Y_{R,T_2} - Y_{R,T_1}. \tag{2}$$

$$= \sum_{j\in R} Y_{j,T_2} w_{j,T_2} \Big/ \sum_{j\in R} w_{j,T_2} - \sum_{j\in R} Y_{j,T_1} w_{j,T_1} \Big/ \sum_{j\in R} w_{j,T_1}. \tag{3}$$

Similarly annualized change $\Delta_{T_1,T_2}/(T_2 - T_1)$ may be defined.

The regional statistics defined by equations (1) and (2) will be dominated by the national rates (and changes) for the larger countries. In a region that contains China or India for example smaller countries may have national rates that are quite different and these will have little impact on the regional rate. The same is true for estimates of change. Similarly the variance of the regional statistics will tend to be dominated by the largest countries because of the impact of the weights squared $\mu^2_{j,T}$. For the regional estimate of the level for example:

$$V(Y_{R,T}) = \sum_{j\in R} V(Y_{j,T})\,\mu^2_{j,T}. \tag{4}$$

An alternative emphasis may be required if the global target is meant to impose a commitment on each country regardless of size. Here the country might be regarded as the unit of analysis (rather than the person as is implicit in the aggregate statistics described above). One possibility could be to define $Y_{R,T}$ and Δ_{T_1,T_2} by setting $w_{j,T}$ equal for all countries so that all countries contributed equally to the summary statistic. Clearly there are other alternatives such as summarizing the countries performance at the regional level by reporting on the number of countries that exceed/fall short of the population weighted regional statistic by a threshold amount together with the range of country levels (or changes) observed.

Constant or time-dependant weights for estimates of change

Many of the statistical series call for national statistics that incorporate a changing population structure over time. For example the proportion of people below a poverty threshold will be changing because of changes to household income (expenditure, or consumption) but also because the population size itself is changing. Indeed over a period of 15-20 years fertility rates in many developing countries imply very significant population growth. Hence the denominator implicit in $Y_{j,T}$ at different years T will properly reflect this change. When producing the aggregate measure $Y_{R,T}$ it is natural to use the population weights $w_{j,T}$ and hence the relative proportions $\mu_{j,T}$ relating to year T. It is less obvious whether the weights $w_{j,T}$ (and hence $\mu_{j,T}$) used to produce the aggregate measure of change Δ_{R,T_1,T_2} for a region or the whole world should change with time. The measure of change Δ_{R,T_1,T_2} may be decomposed as follows:

$$\Delta_{R,T_1,T_2} = \sum_{j\in R} (Y_{j,T_2} - Y_{j,T_1})(\lambda\,\mu_{j,T_2} + (I - \lambda)\,\mu_{j,T_1}) + \sum_{j\in R} (\mu_{j,T_2} - \mu_{j,T_1})((\lambda)\,Y_{j,T_1} + (1 - \lambda)\,Y_{j,T_2}). \quad (5)$$

The measure of change is thus a composite measure involving both the change in Y over time and the change in weights. Since the weights $\mu_{j,T}$ simply provide the linear combination of the country measures it is arguable that these should be held constant between T_1 and T_2 so that the second term in equation (5) is made zero. The first term in equation (5) is arguably a better measure as an index of change since it represents a linear combination of the country changes.

The same rationale may be applied when the national measures are economic and measured in the local currency and these have to be converted to a common basis using PPP conversion for example. It may be argued that a constant value of PPP conversion should be applied to all values in local currency whatever the time period to which they apply.

There still remains the choice of λ and values of 0 or 1 would use the weights for one of the reference years only.

Of course a measure of change based on the first term in equation (5) implies that this is not arithmetically equal to the difference between the measures of level in the two years.

Missing Values

Most of the statistical series used for monitoring global policy have gaps of various kinds. For some series most countries are represented with data for most years. For other series data may be available for a smaller subset of UN member states, but with perhaps only one or two data points for some of the countries represented and these related to different years for different countries. If inference is required for year T and this is missing then the question of imputation method arises. Figure 4 illustrates the situation for the primary education enrolment rates.

Much research and development on missing data in Official Statistics has focussed on the raw micro-data and causes such as non-response. In calculating aggregate measures for statistical indicators it is the national statistic for an entire country in a particular year that is unavailable. Common assumptions such as that data are missing at random are inapplicable in this case. In general the lack of completeness of statistical series for each country may often be related to the statistical capacity of the country to produce the range of statistics required. This in turn is often related to the level of development generally and to some extent the size of the country since the per capita statistical effort required is generally greater for small countries. This has two general consequences:

- If we consider the regional estimate $Y_{R,T} = \sum_{j\in R} Y_{j,T}\,\mu_{j,T}$ and only a small proportion of country values are missing (and if these relate to countries with small weights $\mu_{j,T}$) then the regional estimate will be relatively robust to any reasonable imputed value for the missing values. Moreover the weights associated with the imputed values and measures such as $\mu_{j,T}^2 / \sum_{j\in R} \mu_{j,T}^2$ will provide diagnostic information about the extent to which the regional (or global) estimate may be dependent on imputed values.
- It must be recognised however that if many of the statistical series are related to economic and social development and if countries with missing data are generally low in statistical capacity (and by extension development generally) then this is a case of informative non-response. Hence the term "reasonable imputed value" in the previous bullet point needs to take account of this.

In general there are three levels of information that might potentially assist with imputation for the missing values in a time series. These are (a) values for other years in the same series for the same country, (b) associated series from the same country that may convey information about the series in question and (c) statistical series from other countries that might be considered "similar" in some sense to the country for which the value needs to be imputed.

In addition the range and sophistication of available methods may vary greatly (see for example Chatfield 1996). The objective is to predict the value $Y_{j,T}$ of the trend at time T for the series in question. The length of the time series available are generally short and the series are non-stationary. Since the series are annual, for many of them seasonal effects may be less important. A simple method may involve naïve trend fitting (and then prediction for the missing values) using least squares fitting on the values available in the series. More sophisticated methods will remove the trend to arrive at a stationary series and then apply various approaches such as weighted averages and modelling the correlation structure of the series to arrive at predicted values for the missing values of the stationary time series. These are then combined with the initially removed trend estimates to yield predicted values for the missing values in the original series.

This general problem could benefit from some substantial methodological investigation perhaps taking account of some of the following:

- The ultimate objective is not to model the series, nor even to predict the missing value for use as an inference at the country level. The imputed value will be combined with observed values from other countries to produce the aggregate measure which is the ultimate objective.
- The time series available are often short.
- So long as the statistical series is not too noisy the highest quality predictive information will likely come from the values for other years in the same series and the same country. For many situations, since the objective is to predict the trend level at $t = T$, this may imply that simple trend estimation methods such as regression using year as an explanatory variable for the series in question may be adequate. For example:

$$Y_{j,t} = \alpha_j + \beta_j\, t + \varepsilon_{j,t}, \qquad (6)$$

where $V(\varepsilon_{j,T}) = \sigma^2$.

The use of data from the same series if it is reasonably stable will ensure that the informative non-response issue is taken into account since the parameter estimates α_j and β_j will relate to the specific country and will be estimated from data from that source.

Consider as an illustration equation (6) written in matrix form for a series of length $k(t = 1, ..., k)$ and where imputation is required for $t = k+1$. Prediction for a missing value at the end of the series is likely to be less reliable but is also likely to be realistic since it will occur in practice.

$$Y = X\beta + \Sigma, \qquad V(\Sigma) = \sigma^2 I \qquad (7)$$

$$V(\hat{Y}_{k+1}) = ((1, k+1)(X^T X)^{-1}(1, k+1)^T + 1)\sigma^2$$

$$= \left[\frac{2(2k+1)}{k(k-1)} + 1\right]\sigma^2. \qquad (8)$$

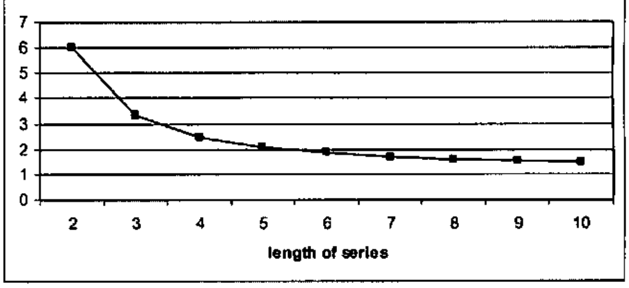


Fig 5: Variance for OLS prediction of missing value at t=k+1 for series of length k, = 1

Figure 5 shows the relative variance for predicting Y_{k+1} the missing value at $t = k+1$ for a series of length k under OLS assumptions. For a series of infinite length the variance will be 1. The point of interest is how quickly the variance drops for a series of 5 or 6 points and how relatively gradually further variance reduction occurs. The OLS assumptions may be replaced by some more general covariance structure such as $\mathrm{Corr}(\varepsilon_{j,t}, \varepsilon_{j,t+r}) = \rho'$. For small and moderate values of $\rho(\rho \le 0.5)$ the variance based on GLS estimates is very similar to Figure 5.

Of course the assumptions above are unrealistic since most time series prediction methods would take account of the correlation structure for recent periods by using exponential weighting of the most recent observations to predict the residual associated with $t = k+1$. However depending on the extent to which recent observations are correlated with $t = k+1$, estimating the parameters α_j and β_j from a very short series will, to some extent, automatically take account of the positive correlation of the residuals at recent periods. If this is so then the decay shown in Figure 5 may be a rough approximation to the impact of the length of the series used.

Clearly a more extensive study of the impact of simple and more sophisticated methods for imputing the missing value would be of considerable benefit.

- For such methods there will be a trade off between variance and bias related to the length of the series used. A relatively short part of the series where the local linearity of the prediction model is more likely to approximate reality may yield a less biased estimate of β_j due to model misspecification but provide parameter estimates of α_j and β_j with higher variance and hence a more variable predicted value.
- Alternative methods that take account of the correlation structure of the time series can be considered and the extent to which these provide a significant improvement in the quality of the prediction would be of interest. One needs to keep in mind that the ultimate objective is to generate the regional (or global) summary measure and the imputed value may have relatively little impact on this in terms of variance.

- When there are no values at all for the statistical series in a particular country or when the series is erratic and/or has very few values for other years the situation is much more difficult. There may be greater added benefit in using the information contained in other series from the same country and from other countries. One might conjecture that a regional effect, such as a drought for example, may affect other countries in a region and that the statistical series may display similar characteristics even if the series themselves are at different levels. The potential for borrowing strength from other time series in the country of interest and others needs to be explored. Hierarchical model-based methods developed for small area estimation could be investigated in this case although the total volume of data even in a region with 30-50 countries will not be large. Also the question of establishing which countries might be suitable sources of information in any situation may require both expert judgement of the similarities and dissimilarities between countries as well as formal statistical diagnostics. If the available series are short then identifying and fitting suitable models will be a challenge.

When we consider the estimation of change between two years T_1 and T_2 the same issues surrounding missing values and imputation emerge. As for the regional estimates of level, countries with small relative weights $\mu_{j,T}$ are unlikely to have a significant impact on the regional estimate of change. However under current international practice it is quite common for the regional estimates of $Y_{R,T}$ to be based on whatever national statistics are available for year T and hence for differences between two years to be based on different sets of countries. This is clearly unsatisfactory and will cause the estimate of change to be biased. Imputation for missing values is needed and the statistical properties of the resulting estimates of change need to be explored. The question of separate or joint imputation for missing values from the same series may also be considered.

5. CONCLUSIONS

A description of the use and importance of statistical indicators and the framework in which they are produced is provided.

It is suggested that there has been less focus on methodological development for statistics used for international comparisons than there has been on statistics used for national domestic purposes. A number of examples have been provided illustrating the need for additional methodological work in this field.

REFERENCES

AHMAD, N. (2002). Proposals on measuring software transactions. Paper presented to the 22-23 April meeting of the OECD Software Task Force, Paris.

CASTLES, I. (2000). Comments on use of PPP and exchange rate conversion in 1999 Human Development Report. Correspondence to Friends of the Chair Review Group.

CHATFIELD, C. (1996). *The Analysis of Time Series (5th edition).* London: Chapman and Hall.

COLECCHIA, A., and SCHREYER, P. (2001). ICT Investment and Economic Growth in the 1990's: Is the USA a Unique Case. OECD

EDWARDS, R., COMISARI, P. and JOHNSON, T. (2002). Beyond 1993: The system of national accounts and the new economy. Proceedings at IAOS Conference on Official Statistics and the New Economy, London.

UNITED NATIONS (1992). *Handbook of the International comparisons programme.* United Nations, New York.

UNSC (1998). *Evaluation of the International Comparison Project.* UN Statistical Commission, New York.

UNSC (2001). An assessment of the criticisms made of the human development report, 1999. UN Statistical Commission, New York.

UNSC (2002). An assessment of the statistical indicators derived from United Nations summit meetings. UN Statistical Commission, New York.

WYCKOFF, ANDREW W. (1995). The impact of computer prices on international comparisons of labour productivity. *Economics of Innovation and New Technology.* 3, 2, 277-293.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 19-24
Statistics Canada

A Noninformative Bayesian Approach to Small Area Estimation

GLEN MEEDEN[1]

ABSTRACT

In small area estimation one uses data from similar domains to estimate the mean in a particular small area. This borrowing of strength is justified by assuming a model which relates the small area means. Here we suggest a noninformative or objective Bayesian approach to small area estimation. Using this approach one can estimate population parameters other than means and find sensible estimates of their precision.

AMS 1991 subject classifications Primary 62D05; secondary 62C10.

KEY WORDS: Sample survey; Small area estimation; Polya posterior and noninformative Bayes.

1. INTRODUCTION

In the standard approach to small area estimation the parameters of interest, the small area means, are assumed to be related through some type of linear model. Drawing on linear model theory one can derive estimators which "borrow strength" by using data from related areas to estimate the mean of interest. Finding a good estimate of the precision of the estimator is often difficult however. Good recent summaries of the literature can be found in Rao (1999) and Ghosh and Rao (1994).

The Bayesian approach to statistical inference summarizes information concerning a parameter through its posterior distribution, which depends on a model and prior distribution and is conditional on the observed data. In finite population sampling the unknown parameter is just the entire population and the likelihood function for the model comes from the sampling design. A Bayesian must specify a prior distribution over all possible values of the population. Once the sample is observed the posterior is just the conditional distribution of the unobserved units given the the values of the observed units computed under the prior distribution for the population. For most designs this posterior does not depend on the design probability used to select the actual sample. The Bayesian approach to finite population sampling was very elegantly described in the writings of D. Basu. For further discussion see his collection of essays in Ghosh (1988).

Assume that given the sample one can simulate values for all the unobserved units from the posterior to generate a "complete copy"of the population. Then given the simulated and observed values one can compute the value of the population mean, $N^{-1}\sum_{i=1}^{N} y_i$, for this simulated copy of the entire population. By generating many independent simulated copies of the population and in each case finding the mean of the simulated population and then taking the average of these simulated means one has an estimate of the unknown population mean. This process computes approximately the Bayes estimate of the population mean under squared error loss for the given prior. More generally by simulating many such full copies of the population one can compute, approximately, the corresponding Bayes point or interval estimates for many population parameters. The problem then is to find a sensible Bayesian model which utilizes the type of prior information available for the small area problem at hand.

The Polya posterior is a noninformative Bayesian approach to finite population sampling which uses little or no prior information about the population. It is appropriate when a classical survey sampler would be willing to use simple random sampling as their sampling design. In Nelson and Meeden (1998) the authors considered several scenarios where it was assumed that information about the population quantiles of the auxiliary variable was known a priori. They demonstrated that an appropriately constrained Polya posterior, *i.e.*, one that used the prior knowledge about the quantiles of *x*, yielded sensible frequentist results. Here we will see that this approach can be useful for a variant of small area estimation problems.

We will consider a population that is partitioned into a domain D, of interest, and its complement D'. Also we suppose that it is partitioned into K areas, say $A_1, \ldots, A_K$. Let y be the characteristic of interest and x be an auxiliary variable. Suppose, using a random sample from the entire population, for some k we wish to estimate $\mu_{D,k}(y)$, the mean of y for the all units that belong to the small area $D \cap A_k$. Often the number of sampled units that belong to $D \cap A_k$ is quite small and using just these observations can lead to an imprecise estimator. As an example where this could arise imagine D is a region of a state which is broken up into counties. Each county in D is then paired with a similar county that is outside of D. Hence the kth county and its twin form the kth area and the collection of "twin" counties forms D'. Then a random sample is taken from $D \cup D'$ and one wishes to to estimate the means of the counties, or small areas, making up D.

[1] Glen Meeden, School of Statistics, University of Minnesota, Minneapolis, MN 55455. E-mail: glen@stat.umn.edu.

In order to improve on this naive estimator one needs to make some additional assumptions. Here we will assume that for each unit in the sample we learn both its y and x values. For units belonging to A_k we make two assumptions which formalize the idea that the small areas, $A_k \cap D$ and $A_k \cap D'$, are similar. First we assume that the small area means of the auxiliary variable, $\mu_{D,k}(x)$ and $\mu_{D',k}(x)$, although unknown are not too different. Secondly we assume that for units belonging to A_k the distribution of y_i depends only on its x_i value and not on its membership in D or D'. Finally we assume that $\mu_D(x)$, the mean of x for all the units that belong to D, is known. Note that we do not assume that $\mu_{D,k}(x)$ and $\mu_{D',k}(x)$ are known which is often the case in small area estimation.

Here we will demonstrate that when our assumptions are true a modification of the Polya posterior yields good point and interval estimators of $\mu_{D,k}(y)$ and of the the median of y in the small area $D \cap A_k$. In section two we will briefly review facts about the Polya posterior and in section three discuss simulating from a constrained version of it. In section four we present some simulation results that indicate how it could work in practice. Section five contains some concluding remarks.

2. THE POLYA POSTERIOR

Consider a finite population consisting of N units labeled 1, 2, ..., N. The labels are assumed to be known and to contain no information. For each unit i let y_i, a real number, be the unknown value of some characteristic of interest. The unknown state of nature, $y = (y_1, ..., y_N)$, is assumed to belong to some subset of N-dimensional Euclidean space, $\Re^N$. A sample s is a subset of $\{1, 2, ..., N\}$. We will let $n(s)$ denote the number of elements in s. A sample point consists of the set of observed labels s along with the corresponding values for the characteristic of interest. If $s = \{i_1, ..., i_{n(s)}\}$ then such a sample point can be denoted by (s, y_s).

Given the data the Polya posterior is a predictive joint distribution for the unobserved units in the population conditioned on the values in the sample. Given a data point (s, y_s) we now show how to generate a set of possible values for the unobserved units from this distribution. We consider an urn that contains $n(s)$ balls, where ball one is given the value $y_{s_{i_1}}$, ball two the value $y_{s_{i_2}}$ and so on. We begin by choosing a ball at random from the urn and assigning its value to the unobserved unit in the population with the smallest label. This ball and an additional ball with the same value are then returned to the urn. Another ball is chosen at random from the urn and we assign its value to the unobserved unit in the population with the second smallest label. This second ball and another with the same value are then returned to the urn. This process is continued until all $N - n(s)$ unobserved units are assigned a value. Once this is done we have generated one realization of the complete population from the Polya posterior distribution. This simulated, completed copy contains the $n(s)$ observed values along with the $N - n(s)$ simulated values for the unobserved members of the population. Hence by simple Polya sampling we have a predictive distribution for the unobserved given the observed.

One can verify that under this predicted distribution the expected value of the population mean is just the sample mean and it's posterior variance is approximately the frequentist variance of the sample mean under simple random sampling when $n(s) \geq 25$. Hence inference for the population mean under the Polya posterior agrees with standard methods. Although the design probabilities play no formal role in the inference based on the Polya posterior for it to be appropriate in the judgment of the survey sampler the values for the characteristic of interest for the observed and unobserved units need to be roughly exchangeable. This is usually the case when simple random sampling is used to select the sample.

It has been shown for a variety of decision problems that procedures based on the Polya posterior are admissible because they are stepwise Bayes. (See Ghosh and Meeden 1997). In these stepwise Bayes arguments a finite sequence of disjoint subsets of the parameter space is selected, where the order is important. A different prior distribution is defined on each of the subsets. First the Bayes procedure is found for each sample point that receives positive probability under the first prior. Next the Bayes procedure is found for each sample point which receives positive probability under the second prior and which was not considered under the first prior. Then the third prior is considered and so on. For a particular sample point the value of the stepwise Bayes estimate is the value for the Bayes procedure for that sample point for the Bayes procedure identified in the step at which the sample point was considered. It is the stepwise Bayes nature of the Polya posterior that explains its somewhat paradoxical properties. Given a sample it behaves just like a proper Bayesian posterior but the collection of possible posteriors that arise from all possible samples comes from a family of priors not from a single prior. From the Bayesian point of view it is appropriate when one's prior beliefs about the population is that the units are roughly exchange but nothing more about them is known. The stepwise Bayesian nature of the Polya posterior also helps to explain why it yields 0.95 Bayesian credible intervals that in most cases behave approximately like 95% confidence intervals. For more details and discussion on the theoretical properties of the Polya posterior see Ghosh and Meeden (1997). The Polya posterior is related to the Bayesian bootstrap of Rubin (1981). See also Lo (1988).

3. SIMULATION FROM THE POLYA POSTERIOR

The interval estimate of the population mean and point and interval estimates for other population quantities under the Polya posterior usually cannot be found explicitly. One must use simulation to find these values approximately. This is done by simulating many independent completed copies of the entire population and calculating the value of the parameter of interest for each copy. One may do this in a straightforward manner but often a well known approximation also works well. For simplicity assume the sample values y_s are all distinct and that the sampling fraction $n(s)/N$ is small. For $j = 1, ..., n(s)$ let λ_j be the proportion of units in a complete simulated copy of the entire population which take on the value y_{i_j}. Then under the Polya posterior $\lambda = (\lambda_1, ..., \lambda_{n(s)})$ has approximately a Dirichlet distribution with a parameter vector of all ones, *i.e.*, it is uniform on the $n(s) - 1$ dimensional simplex where $\sum_{j=1}^{n(s)} \lambda_j = 1$.

We now assume that there is an auxiliary characteristic associated with each element in the population. For unit i let x_i be the value of this auxiliary characteristic. The vector of these values for the auxiliary characteristic is denoted by x. The values of x are unknown but we assume their population mean is known. This is a common situation and either the regression estimator or the ratio estimator is often used in such cases. Let x_s denote the x values of the observed units in the sample. Now the Polya posterior can be adapted to use this additional information in the following way. When creating a simulated copy of the entire population using the values $\{(y_i, x_i) : i \in s\}$ one only uses completed copies whose simulated population mean of x is equal to the known mean of x.

Simulating from a constrained Polya posterior is more difficult than simulating from the unconstrained Polya. Let μ_x^* denote the known population mean of x. Suppose s is a sample such that x_s contains values smaller and larger than μ_x^*. When this is the case an approximate solution to the problem of generating simulated copies from the Polya posterior distribution which satisfies the mean constraint is available. For $j = 1, ..., n(s)$ let λ_j be the proportion of units in the simulated copy of the population which have the value (y_{i_j}, x_{i_j}). (Note the x_s need not be distinct.) If we ignore the constraint for a moment then, as we observed earlier, simulation from the Polya posterior is approximately equivalent to assuming a uniform distribution for $\lambda = (\lambda_1, ..., \lambda_{n(s)})$ on the $n(s) - 1$ dimensional simplex where $\sum_{j=1}^{n(s)} \lambda_j = 1$. In order to satisfy the mean constraint we must select λ's at random from the set which is the intersection of the hyperplane $\sum_{j=1}^{n(s)} \lambda_j x_{i_j} = \mu_x^*$ with the simplex for λ. In general one cannot generate independent random samples from this distribution. One may, however, use the Metropolis-Hasting algorithm to generate dependent simulated copies of the population from a convergent Markov chain. For more details on this algorithm see Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and Hastings (1970).

Using the approximate solution based on the Dirichlet distribution allows one to finesse a bothersome technical problem which has no practical significance. That is given the sample it is often impossible to get simulated copies of the population which satisfy the mean constraint exactly. For example suppose $N = 5$ and our sample of size three yielded x values of 0, 0 and 10. Now if we know $\mu_x = 4.5$ then under the Polya posterior it is impossible to generate simulated copies of the population since the only possible values for an x value of an unobserved unit is 0 or 10. This implies that given this sample under the Polya posterior the only possible values of μ_x are 2, 4 and 6. In general even if we have generated a λ which satisfies the constraint the $\lambda_i N$'s need not be integers and hence their need not be an actual copy of the population corresponding to λ. But in real problems this should not matter very much. For one thing the mean constraint will usually only be known approximately. Furthermore for larger sample sizes the approximate nature of the simulated copies is just not important.

Recently Nelson and Meeden (1998) and Meeden and Nelson (2001) have considered a variety of problems where a constrained Polya posterior is applicable. When the population mean of x is known Meeden and Nelson (2001) presented simulations that demonstrated that the point and interval estimators of the constrained Polya posterior were nearly identical those of the regression estimator. Hence just as the regression estimator does, when estimating the population mean of y the constrained Polya posterior utilizes the information contained in knowing the population mean of x.

4. A SMALL AREA PROBLEM

Consider again the small area estimation problem described in the introduction. A population is partitioned in two different ways. The first partitions the population into a domain of interest, D, and its complement D'. The second partitions it into K areas $A_1, ..., A_k$ where for each k we assume that the small areas $A_k \cap D$ and $A_k \cap D'$ are nonempty. Figure 1 gives a graphical representation of the population. A random sample is taken from the whole population and we wish to estimate $\mu_{D,k}(y)$, the mean of y for all the units belong to the small area $A_k \cap D$. For such problems one often assumes that for the auxiliary variable x all the means $\mu_{D,k}(x)$ and $\mu_{D',k}(x)$ are known. Here we make the weaker assumptions that $\mu_{D,k}(x)$ and $\mu_{D',k}(x)$ are unknown but not too different and that $\mu_D(x)$, the mean of x for all the units belonging to D, is known. We also assume that for units belonging to $A_k \cap D$ and $A_k \cap D'$ the distribution of y_i depends only on x_i and does not depend on whether it belongs to D or D'. In terms of Figure 1 we are assuming that the mean of x for all the units in the population which belong to the first column is known and that within each row the distribution of the units across the the two columns is roughly the same. As we will soon see this is enough to produce estimators of $\mu_{D,k}(y)$ which improve on the naive estimator.

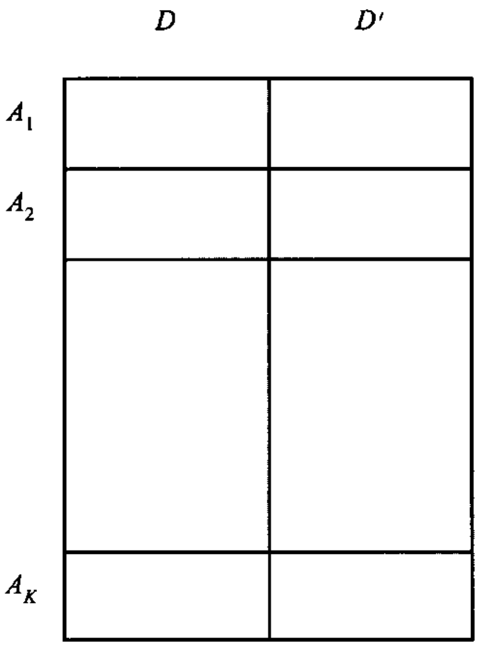


Figure 1. A population partitioned into a domain and its complement along with a second partition of K small areas.

Before explaining how this is done we need a bit more notation. Let $N_{D,k}$ be the number of units in the population that belong to $D \cap A_k$. We assume that the $N_{D,k}$'s are known. For unit i let $t_i = (1,k)$ if $i \in D \cap A_k$ and $t_i = (0,k)$ if $i \in D' \cap A_k$. Then given a sample s we must use $\{(y_i, x_i, t_i) : i \in s\}$ to estimate $\mu_{D,k}(y)$. The constrained Polya posterior is now constructed in two stages. In the first stage, using the members of the sample that fall into D and their (x_i, t_i) values, we create a completed copy of D which satisfies the known mean constraint $\mu_D(x)$. In the second stage we first find for the simulated copy of D the mean of the x values for all the units belonging to $D \cap A_k$. (Remember that this set contains both observed and simulated values.) Let $\hat{\mu}_{D,k}(x)$ denote this mean. Next using the observed sample values from $D \cap A_k$ and $D' \cap A_k$ we create a completed copy of $D \cap A_k$ which satisfies the mean constraint $\hat{\mu}_{D,k}(x)$. By repeating this two staged process many times one can construct simulated copies of $D \cap A_k$ which use the similarity of units within the small areas $A_k \cap D$ and $A_k \cap D'$ and the information from knowing $\mu_D(x)$.

To see how this approach could work in practice we present simulation results for some constructed populations. In all the cases $K = 2$ so there are just two areas and in Figure 1 there are just four cells or four small areas. The populations will be constructed so that there are 250 units in each of the four cells. For each cell we first generate 250 values for the auxiliary variable x by taking a random sample from a gamma distribution with some shape parameter and scale parameter one. Next within each area conditioned on the x values the y values are independent observations from normal distributions where the mean of $y_i \mid x_i$ depends on x_i and where the the variance of $y_i \mid x_i$ may be constant or in some cases depends on x_i.

In the first population, pop1, the shape parameter of the gamma distribution was four in both $A_1 \cap D$ and $A_1 \cap D'$ and was six in $A_2 \cap D$ and $A_2 \cap D'$. For units in A_1 $y_i \mid x_i$ was normal with mean $25 + 2x_i$ and variance 100. For units in A_2 $y_i \mid x_i$ is normal with mean $25 + 3x_i$ and variance 25.

Note that pop1 was generated under a model which is consistent with the assumptions underlying the constrained Polya posterior described above. In fact our method should work very well for pop1. This is because for each k the average values of the auxiliary variable in $A_k \cap D$ and $A_k \cap D'$ will be approximately equal. This is not necessary for our approach to work but if it does not work in this example then it is hard to imagine that it could work in practice. In two of the remaining populations for each k we will take the shape parameters generating the values of x in $A_k \cap D$ and $A_k \cap D'$ to be different. This is a more realistic assumption. We will also let the mean of $y_i \mid x_i$ be a nonlinear function of x_i and let the variance of $y_i \mid x_i$ depend on x_i. In all cases the form of the distribution of $y_i \mid x_i$ will be the same across $A_k \cap D$ and $A_k \cap D'$ for each k. This is the most crucial assumption. If this is not satisfied approximately then our method cannot work.

In the second population, pop2, the shape parameters of the gamma distributions were eight in $A_1 \cap D$, ten in $A_1 \cap D'$, six in $A_2 \cap D$ and four in $A_2 \cap D'$. For units in A_1 $y_i \mid x_i$ was normal with mean $25 + 2x_i$ and variance $9x_i$. For units in A_2 $y_i \mid x_i$ was normal with mean $25 + 3x_i$ and variance $4x_i$.

In the third population, pop3, the shape parameters of the gamma distributions were eight in $A_1 \cap D$ and $A_1 \cap D'$, and six in $A_2 \cap D$ and $A_2 \cap D'$. For units in A_1 $y_i \mid x_i$ was normal with mean $25 + 0.5(x_i - 8)^2$ and variance $9x_i$. For units in A_2 $y_i \mid x_i$ was normal with mean $25 + |x_i - 6|$ and variance $4x_i$.

In the fourth population, pop4, the shape parameters of the gamma distributions were four in $A_1 \cap D$, six in $A_1 \cap D'$, six in $A_2 \cap D$ and eight in $A_2 \cap D'$. For units in A_1 $y_i \mid x_i$ was normal with mean $25 + 0.5(x_i - 4)^2$ and variance $9x_i$. For units in A_2 $y_i \mid x_i$ was normal with mean $25 + |x_i - 6|$ and variance $4x_i$.

In the fifth population, pop5, the shape parameters for the gamma distributions were the same as those in pop2. For units in A_1 $y_i \mid x_i$ was normal with mean $25 + 0.5(x_i - 9)^2$ and variance $9x_i$. For units in A_2 $y_i \mid x_i$ was normal with mean $25 + |x_i - 5|^{1.5}$ and variance $4x_i$.

For each of these five populations we took 500 random samples of size 80. For each sample we calculated the usual point estimates and 95% confidence intervals for $\mu_{D,1(y)}$ and $\mu_{D,2(y)}$ using just the observations that fell into the small areas. We also found approximately the point estimate and 0.95 credible interval for the constrained Polya posterior. The results are given in Table 1. In each case the constrained Polya posterior estimates were computed using 500 simulated copies of the small area. Then our point estimate is just the average of these 500 computed values and our 0.95 credible interval ranges from the 0.025 quantile to the 0.975 quantile of this set.

We see that the constrained Polya posterior yields significantly better point estimators in every case but one, $\mu_{D,2(y)}$ of pop5. Its intervals are also considerable shorter than the usual. There is some evidence that their frequency of coverage is a bit less than the usual approximate 95% normal theory intervals. In particular this is true for the small area $A_2 \cap D$ in the fifth population.

The results in Table 1 are for the small area means. In Table 2 we give similar results for the small area medians. We compared our estimates to the sample median of the set of the sampled observations that fell into the small area and the usual confidence interval for the median due to Woodruff (1952). Compared to the usual estimators the performance of the constrained Polya posterior estimators for the small area medians is even better than it was for the small area means. In every case its point estimators are better than the sample median. Its interval estimators are always shorter than Woodruff's and for most cases their frequency of coverage seems to be quite close to the nominal 0.95.

Table 1

The average value and the average absolute error for the usual naive small area estimator and the constrained Polya posterior estimator (cstpp) for the small area means. Also given are the length and relative frequency of coverage for their nominal 0.95 intervals for 500 random samples of size 80 from five different populations

Pop	Small Area	Method	Ave value	Ave aberr	Ave lenght	Freq. of coverage
pop1	$A_1 \cap D$	usual	33.11	1.84	9.10	0.936
		cstpp	33.20	1.30	6.37	0.934
	$A_2 \cap D$	usual	43.03	1.47	7.78	0.946
		cstpp	43.13	1.03	5.15	0.940
pop2	$A_1 \cap D$	usual	40.39	1.79	8.69	0.932
		cstpp	40.29	1.20	5.62	0.944
	$A_2 \cap D$	usual	42.13	1.48	7.50	0.944
		cstpp	41.97	1.16	5.16	0.912
pop3	$A_1 \cap D$	usual	28.57	1.97	9.85	0.936
		cstpp	28.90	1.47	6.66	0.898
	$A_2 \cap D$	usual	26.71	1.01	5.08	0.940
		cstpp	26.83	0.70	3.24	0.930
pop4	$A_1 \cap D$	usual	27.73	1.27	6.57	0.960
		cstpp	27.64	0.81	4.09	0.940
	$A_2 \cap D$	usual	27.03	0.97	5.33	0.952
		cstpp	27.03	0.65	3.32	0.934
pop5	$A_1 \cap D$	usual	29.25	1.74	9.31	0.942
		cstpp	29.30	1.26	6.16	0.930
	$A_2 \cap D$	usual	27.73	1.08	5.85	0.954
		cstpp	28.82	1.28	4.40	0.850

Table 2

The average value and the average absolute error for the usual naive small area estimator and the constrained Polya posterior estimator for the small area medians. Also given are the length and relative frequency of coverage for their nominal 0.95 intervals for 500 random samples of size 80 from five different populations

Pop	Small Area	Method	Ave value	Ave aberr	Ave lenght	Freq. of coverage
pop1	$A_1 \cap D$	usual	33.88	2.01	11.48	0.944
		cstpp	33.25	1.44	7.81	0.930
	$A_2 \cap D$	usual	42.84	1.72	9.94	0.950
		cstpp	42.42	1.35	6.92	0.944
pop2	$A_1 \cap D$	usual	38.94	1.82	9.81	0.940
		cstpp	38.53	1.41	7.47	0.936
	$A_2 \cap D$	usual	40.99	1.77	8.75	0.970
		cstpp	40.33	1.38	6.36	0.914
pop3	$A_1 \cap D$	usual	27.64	1.73	9.52	0.952
		cstpp	27.73	1.24	6.46	0.958
	$A_2 \cap D$	usual	27.03	1.15	6.26	0.954
		cstpp	26.59	0.70	3.76	0.938
pop4	$A_1 \cap D$	usual	27.14	1.27	7.00	0.962
		cstpp	27.05	0.95	5.37	0.966
	$A_2 \cap D$	usual	26.84	1.07	5.99	0.960
		cstpp	26.81	0.78	4.32	0.954
pop5	$A_1 \cap D$	usual	29.10	2.06	11.01	0.956
		cstpp	28.89	1.51	8.28	0.944
	$A_2 \cap D$	usual	27.03	1.14	5.98	0.952
		cstpp	27.87	0.97	4.46	0.900

5. CONCLUDING REMARKS

Here we have presented a new method of "borrowing strength" when estimating parameters of a small area of a population. It makes weaker assumptions than those made by the usual approaches to such problems. It is an objective or noninformative Bayesian approach which uses no more prior information than is typically assumed by a frequentist. Simulations indicate that it should be applicable in a variety of situations and should work well especially for some of the problems which roughly satisfy the usual linear model type assumptions, often assumed in small area estimation. It has the advantage of not being restricted to estimating small area means but can estimate other parameters as well. Here we assumed that a certain mean of an auxiliary variable was known. This approach can be extended to when other parameters of an auxiliary variable are known, like the median. Also it should be possible to extend this method to situations where prior information is available for more than one auxiliary variable. In summary we believe that this is flexible approach which can yield point and interval estimators with good frequentist properties for a variety of problems.

ACKNOWLEDGEMENT

Research supported in part by NSF Grant DMS 9971331.

REFERENCES

GHOSH, J.K. (1988). *Statistical information and likelihood: A collection of critical essays by D. Basu.* Springer-Verlag, New-York.

GHOSH, M., and MEEDEN, G. (1997). *Bayesian Methods for finite Population Sampling.* Chapman and Hall, London.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science.* 9, 55-93.

HASTINGS, W.K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika.* 57, 97-109.

LO, A. (1988). Bayesian bootstrap for a finite population. *Annals of Statistics.* 16, 1684-1695.

MEEDEN, G., and NELSON, D. (2001). A noninformative Bayesian approach to problems in finite population sampling when information about an auxiliary variable is present. Technical Report.

METROPOLIS, N., ROSENBLUTH, A.W., ROSENBLUTH, M.N., TELLER, A.H. and TELLER, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics,* 21, 1087-1092.

NELSON, D., and MEEDEN, G. (1998). Using prior information about population quantiles in finite population sampling. *Sankhya A.* 60, 426-445.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology.* 25, 175-186.

RUBIN, D. (1981). The Bayesian bootstrap. *Annals of Statistics.* 9, 130-134.

WOODRUFF, R.S. (1952). Confidence intervals for medians and other position measures. *Journal of the American Statistical Association.* 47, 635-646.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 25-32
Statistics Canada

Model-Based Unemployment Rate Estimation for the Canadian Labour Force Survey: A Hierarchical Bayes Approach

YONG YOU, J.N.K. RAO and JACK GAMBINO[1]

ABSTRACT

The Canadian Labour Force Survey (LFS) produces monthly direct estimates of the unemployment rate at national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Census Agglomerations (CAs). However, for some sub-provincial areas, the direct estimates are not very reliable since the sample size in some areas is quite small. In this paper, a cross-sectional and time-series model is used to borrow strength across areas and time periods to produce model-based unemployment rate estimates for CMAs and CAs. This model is a generalization of a widely used cross-sectional model in small area estimation and includes a random walk or AR(1) model for the random time component. Monthly Employment Insurance (EI) beneficiary data at the CMA or CA level are used as auxiliary covariates in the model. A hierarchical Bayes (HB) approach is employed and the Gibbs sampler is used to generate samples from the joint posterior distribution. Rao-Blackwellized estimators are obtained for the posterior means and posterior variances of the CMA/CA-level unemployment rates. The HB method smooths the survey estimates and leads to substantial reduction in standard errors. Bayesian model fitting is also investigated based on posterior predictive distributions.

KEY WORDS: Gibbs sampling; Hierarchical Bayes; Labour Force Survey; Small area estimation; Unemployment rate.

1. INTRODUCTION

The unemployment rate is generally viewed as a key indicator of economic performance. In Canada, although provincial and national estimates get the most media attention, subprovincial estimates of the unemployment rate are also very important. They are used by the Employment Insurance (EI) program to determine the rules used to administer the program. In addition, the unemployment rates for Census Metropolitan Areas (CMAs, *i.e.*, cities with population more than 100,000) and Census Agglomerations (CAs, *i.e.*, other urban centres) receive close scrutiny at local levels. However, many CAs do not have a large enough sample to produce adequate direct estimates. Our objective in this paper is to obtain model-based estimators that lead to improvement over the direct estimator which is based solely on the sample falling in a given CMA or CA in a given month. For convenience, since CMAs are also CAs, we will refer to both CMAs and CAs as CAs.

In Canada, unemployment rates are produced by the Labour Force Survey (LFS). The LFS is a monthly survey of 53,000 households selected using a stratified, multistage design. Each month, one-sixth of the sample is replaced. Thus five-sixths of the sample is common between two consecutive months. This sample overlap induces correlations which can be exploited to produce better estimates by any method which borrows strength across time. For a detailed description of the LFS design, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998).

Traditional small area estimators borrow strength either from similar small areas or from the same area across time, but not both. In recent years, several approaches to borrowing strength simultaneously across both space and time have been developed. Estimators based on the approach developed by Rao and Yu (1994), such as those in Ghosh, Nangia and Kim (1996), Datta, Lahiri, Maiti and Lu (1999) and in this paper, successfully exploit the two dimensions simultaneously to produce improved estimates with desirable properties for small areas. Datta *et al.* (1999) applied their model to long time series ($T = 48$ months) data across small areas from the U.S. Current Population Survey. In this paper, we apply a similar model to the Canadian LFS. Unlike Datta *et al.* (1999), we have used short time series data across small areas. Therefore, our model does not contain seasonal parameters. This reduces substantially the number of parameters that need to be estimated; details on modelling and analysis are given in section 2 and section 4. Despite this simplification, we obtain both an adequate model fit and large reductions in the coefficients of variation (CVs) of the small area estimators of the unemployment rate. The CV reduction is due in part to our approach to computing covariance matrices, which uses smoothed CVs and lag correlations to obtain smoothed estimates of the sampling covariance matrices of the direct LFS estimators.

In section 2, we present the model, which borrows strength across small areas and time periods. In section 3, the model is placed in a hierarchical Bayes (HB) framework. The use of Gibbs sampling to generate samples from the joint posterior distribution is described and the corresponding HB estimators are obtained. The HB method is applied to the LFS data in section 4 to produce

[1] Yong You, Jack Gambino, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

unemployment rates for CAs. Specifically, subsections 4.2 and 4.3 present model selection and model fit analysis. Subsection 4.4 presents model-based estimates for the small area (CA) unemployment rates and the CV comparisons. Finally some concluding remarks are given in section 5.

2. CROSS-SECTIONAL AND TIME SERIES MODELS

Let y_{it} denote the direct LFS estimate of θ_{it}, the true unemployment rate of the i-th CA (small area) at time t, for $i = 1, ..., m$, $t = 1, ..., T$, where m is the total number of CAs and T is the (current) time of interest. We assume that

$$y_{it} = \theta_{it} + e_{it}, \; i = 1, ..., m, \; t = 1, ..., T, \qquad (1)$$

where e_{it}'s are sampling errors. Let $y_i = (y_{i1}, ..., y_{iT})'$, $\theta_i = (\theta_{i1}, ..., \theta_{iT})'$, and $e_i = (e_{i1}, ..., e_{iT})'$. Then e_i is a vector of sampling errors for the i-th CA. In the LFS design, the CAs are treated as strata. Thus the sampling vectors e_i are uncorrelated between areas (CAs). Because of the LFS sample rotation pattern, there is substantial sample overlap over short time periods within each area. As a result, the correlation between e_{it} and $e_{is} (t \neq s)$ has to be taken into account. It is customary to assume that e_i follows a multivariate normal distribution with mean vector 0 and covariance matrix Σ_i, *i.e.*, $e_i \sim N(0, \Sigma_i)$. Using (1), we have

$$y_i \sim N(\theta_i, \Sigma_i), \; i = 1, ..., m. \qquad (2)$$

Thus y_i is design-unbiased for θ_i. The variance- covariance matrix Σ_i in model (2) is assumed to be known. The assumption of normality and known Σ_i in model (2) is the customary practice in model-based small area estimation (see, for example, Fay and Herriot 1979; Ghosh and Rao 1994; Datta *et al.* 1999; Rao 1999). In this paper, we follow the customary approach and treat Σ_i as known. Specification of Σ_i may not be easy in practice. We use a smoothed estimator of Σ_i in the model, and then treat it as the true Σ_i. More details on constructing a smoothed estimator of Σ_i in the context of the LFS are given in section 4. Pfeffermann, Feder and Signorelli (1998) proposed a simple method of estimating the autocorrelations of sampling errors for rotating-panel designs, such as the Canadian LFS. It would be useful to study the feasibility of this approach in our context.

To borrow strength across small areas and time periods, we model the true unemployment rate θ_{it} by a linear regression model with random effects through auxiliary variables x_{it}. We assume that

$$\theta_{it} = x_{it}' \beta + v_i + u_{it}, \; i = 1, ..., m, t = 1, ..., T, \qquad (3)$$

where $x_{it} = (x_{it1}, ..., x_{itp})'$ is the vector of area level auxiliary data for the i-th CA at time t; β is a vector of regression parameters of length p; v_i is a random area effect with $v_i \sim$ iid $N(0, \sigma_v^2)$; u_{it} is a random time component. For a given area i, Datta *et al.* (1999) assumed that u_{it} follows a random walk process over time period $t = 1, ..., T$, that is,

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \; i = 1, ..., m, t = 2, ..., T, \qquad (4)$$

where $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Then $\text{cov}(u_{it}, u_{is}) = \min(t, s)\, \sigma_\varepsilon^2$. Also $\{v_i\}$, $\{\varepsilon_{it}\}$ and $\{e_i\}$ are assumed to be mutually independent. The regression parameter β and the variance components σ_v^2 and σ_ε^2 are unknown in the model. Rao and Yu (1994) used a stationary autoregressive model, AR(1), for u_{it}, that is, $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, and $|\rho| < 1$. Datta *et al.* (1999) included month and year effects as seasonal effects for θ_{it} in (3) using a long time series ($T = 48$ months) in their analysis. In our modelling, we intend to study the effects of borrowing strength across areas and over time using short time series data instead of long time series data. In particular, based on the Canadian LFS design's six-month rotation cycle, we used only 6 months of data for smoothing; see section 4 for details. Thus the linking model (3) is simpler than Datta *et al.* (1999)'s model. This simplification is likely to reduce the instability in the smoothed covariance matrix Σ_i.

Arranging the data $\{y_{it}\}$ as a vector $y = (y_1', ..., y_m')'$ with $y_i = (y_{i1}, ..., y_{iT})'$, we can write models (2), (3) and (4) in matrix form as

$$y_i = X_i \beta + 1_T v_i + u_i + e_i, \; i = 1, ..., m, \qquad (5)$$

where $X_i' = (x_{i1}, ..., x_{iT})$, $u_i' = (u_{i1}, ..., u_{iT})$, and 1_T is a $T \times 1$ vector of 1's. Model (5) is a special case of a general linear mixed effects model. It also extends the well-known Fay-Herriot model (Fay and Herriot 1979) by borrowing strength across both areas and time.

For comparison, we also considered the Fay-Herriot model for the time points $t = 1, ..., T$ in our data analysis. The model at time point t is given as

$$y_{it} = \theta_{it} + e_{it}, \; i = 1, ..., m, \qquad (6)$$

and

$$\theta_{it} = x_{it}' \beta_t + v_{it}, \; i = 1, ..., m, \qquad (7)$$

where the sampling errors $e_{it} \sim$ ind $N(0, \sigma_{it}^2)$ and the area random effects $v_{it} \sim$ iid $N(0, \sigma_{vt}^2)$ for each time point t and independent of $v_{it'}$, $t' \neq t$. The sampling variances σ_{it}^2 are assumed to be known (smoothed estimates) and σ_{vt}^2 is unknown. The Fay-Herriot model combines cross-sectional information at each t for estimating θ_{it}, but does not borrow strength over the past time periods.

We are interested in obtaining a model-based estimator of θ_{it}, in particular, for the current time $t = T$. Datta, Lahiri and Maiti (2002) and You (1999) obtained two-stage estimators for θ_{iT} and MSE approximations for the estimators through the empirical best linear unbiased prediction (EBLUP) approach. In this paper, we study both AR(1) and random walk models on u_{it}'s, under a complete HB approach using the Gibbs sampling method.

3. HIERARCHICAL BAYES ANALYSIS

In this section, we apply the hierarchical Bayes approach to the cross-sectional and time series model given by (2), (3) and (4) and the Fay-Herriot model given by (6) and (7). Estimates of the posterior mean and posterior variance of the small area means, θ_{iT}, are obtained using the Gibbs sampling method.

3.1 The Hierarchical Bayes Model

We now present the cross-sectional and time series model in a hierarchical Bayes framework as follows:

- Conditional on the parameters $\theta_i = (\theta_{i1}, ..., \theta_{iT})'$, $[y_i \mid \theta_i] \sim \text{ind } N(\theta_i, \Sigma_i)$;
- Conditional on the parameters β, u_{it} and σ_v^2, $[\theta_{it} \mid \beta, u_{it}, \sigma_v^2] \sim \text{ind } N(x_{it}'\beta + \rho u_{it}, \sigma_v^2)$;
- Conditional on the parameters $u_{i,t-1}$ and σ_ε^2, $[u_{it} \mid u_{i,t-1}, \sigma_\varepsilon^2] \sim \text{ind } N(\rho u_{i,t-1}, \sigma_\varepsilon^2)$;

Marginally β, σ_v^2 and σ_ε^2 are mutually independent with priors given as $\beta \propto 1$, $\sigma_v^2 \sim IG(a_1, b_1)$, and $\sigma_\varepsilon^2 \sim IG(a_2, b_2)$, where IG denotes an inverted gamma distribution and a_1, b_1, a_2, b_2 are known positive constants and usually set to be very small to reflect our vague knowledge about σ_v^2 and σ_ε^2. For the random walk model, we take $\rho = 1$ and for the AR(1) model, $|\rho| < 1$ and ρ is assumed to be known.

We are interested in estimating θ_i, and in particular, the current unemployment rate θ_{iT}. In the HB analysis, θ_{iT} is estimated by its posterior mean $E(\theta_{iT} \mid y)$ and the uncertainty associated with the estimator is measured by the posterior variance $V(\theta_{iT} \mid y)$. We use Gibbs sampling (Gelfand and Smith 1990; Gelman and Rubin 1992) to obtain the posterior mean and the posterior variance of θ_{iT}.

Similarly, the Fay-Herriot model (6)-(7) can be expressed as:

- Conditional on the parameters θ_{it}, $[y_{it} \mid \theta_{it}] \sim \text{ind } N(\theta_{it}, \sigma_{it}^2)$;
- Conditional on the parameters β_t, and σ_v^2, $[\theta_{it} \mid \beta_t, \sigma_{vt}^2] \sim \text{ind } N(x_{it}'\beta_t, \sigma_{vt}^2)$;

Marginally β_t and σ_{vt}^2 are mutually independent with priors given as $\beta_t \propto 1$, $\sigma_{v_t}^2 \sim IG(a_t, b_t)$.

3.2 Gibbs Sampling Method

The Gibbs sampling method is an iterative Markov chain Monte Carlo sampling method to simulate samples from a joint distribution of random variables by sampling from low dimensional densities and to make inferences about the joint and marginal distributions (Gelfand and Smith 1990). The most prominent application is for inference within a Bayesian framework. In Bayesian inference one is interested in the posterior distribution of the parameters. Assume that $y_i|\theta$ has conditional density $f(y_i \mid \theta)$ for $i = 1, ..., n$ and that the prior information about $\theta = (\theta_1, ..., \theta_k)'$ is summarized by a prior density $\pi(\theta)$. Let $\pi(\theta \mid y)$denote the posterior density of θ given the data $y = (y_1, ..., y_n)'$. It may be difficult to sample from $\pi(\theta \mid y)$ directly in practice due to the high dimensional integration with respect to θ. However, one can use the Gibbs sampler to construct a Markov chain $\{\theta^{(g)} = (\theta_1^{(g)}, ..., \theta_k^{(g)})'\}$ with $\pi(\theta \mid y)$ as the limiting distribution. For illustration, let $\theta = (\theta_1, \theta_2)'$. Starting with an initial set of values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$, we generate $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ by sampling $\theta_1^{(g)}$ from $\pi(\theta_1 \mid \theta_2^{(g-1)}, y)$ and $\theta_2^{(g)}$ from $\pi(\theta_2 \mid (\theta_1^{(g)}, y)$. Under certain regularity conditions, $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ converges in distribution to $\pi(\theta \mid y)$ as $g \to \infty$. Marginal inferences about $\pi(\theta_i \mid y)$ can be based on the marginal samples $\{\theta_i^{(g+k)};\ k = 1, 2, ...\}$ for large g.

For the hierarchical Bayes models in section 3.1, to implement the Gibbs sampler we need to generate samples from the full conditional distributions of the parameters β, σ_v^2 and σ_ε^2, u_{it} and θ_i. These conditional distributions are given in Appendix A.1. All the full conditional distributions in the Appendix are standard normal or inverted gamma distributions that can be easily sampled.

3.3 Posterior Estimation

To implement Gibbs sampling, we follow the recommendation of Gelman and Rubin (1992) and independently run $L (L > 2)$ parallel chains, each of length $2d$. The first d iterations of each chain are deleted. After d iterations, all the subsequent iterates are retained for calculating the posterior means and posterior variances, as well as for monitoring the convergence of the Gibbs sampler. The convergence monitoring is discussed in section 4.

We use the Rao-Blackwellization approach to obtain estimators for the posterior mean and the posterior variance of interest. The Rao-Blackwellization can substantially reduce the simulation errors compared to naive estimates based on the simulated samples (Gelfand and Smith 1991; You and Rao 2000). For the cross-sectional and time series model, the Rao-Blackwellized estimates of $E(\theta_i \mid y)$ and $V(\theta_i \mid y)$ are obtained as

$$\hat{E}(\theta_i \mid y) = \sum_{l=1}^{L} \sum_{k=d+1}^{2d}$$

$$[(\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1}) \times (\Sigma_i^{-1} y_i + \sigma_v^{-2(lk)} (X_i \beta^{(lk)} + u_i^{(lk)}))] / (Ld)$$

and

$$\hat{V}(\theta_i|y) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d} (\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1})/(Ld)$$
$$+\sum_{l=1}^{L}\sum_{k=d+1}^{2d}$$
$$[(\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1})^{-1} \times (\Sigma_i^{-1} y_{it} + \sigma_v^{-2(lk)}(X_i\beta^{(lk)} + u_i^{(lk)}))]$$
$$\times [(\Sigma_i^{-1} y_i + \sigma_v^{-2(lk)}(X_i\beta^{(lk)} + u_i^{(lk)}))'$$
$$\times (\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1})^{-1}]'/(Ld)$$
$$-\Bigg[\sum_{l=1}^{L}\sum_{k=d+1}^{2d} (\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1})^{-1}$$
$$\times\left(\Sigma_i^{-1} y_i + \sigma_v^{-2(lk)}\left(X_i\beta^{(lk)} + u_i^{(lk)}\right)\right)\Bigg]$$
$$\times\Bigg[\sum_{l=1}^{L}\sum_{k=d+1}^{2d} (\sigma_v^{-2(lk)} I_T + \Sigma_i^{-1})^{-1}$$
$$\times\left(\Sigma_i^{-1} y_i + \sigma_v^{-2(lk)}(X_i\beta^{(lk)} + u_i^{(lk)})\right)\Bigg]'/(Ld)^2,$$

where $\{\beta^{(lk)}, \sigma_v^{2(lk)}, u_i^{(lk)}; k = d+1, ..., 2d, l = 1, ..., L\}$ are the samples generated from the Gibbs sampler and I_T is the identity matrix of order T. Thus by using Gibbs sampling, we can estimate the current time small area mean θ_{iT} and the small area means θ_{it} for the past time periods $t = 1, ..., T-1$ simultaneously for each area. The posterior covariance matrix estimate $\hat{V}(\theta_i|y)$ also provides an estimate of the posterior covariance of θ_{it} and θ_{is} for $t \neq s = 1, ..., T$.

Under the Fay-Herriot model, letting $y_T = (y_{1T}, ..., y_{mT})'$ denote the current time cross-sectional data and using the conditional distributions given in Appendix A.2, we can similarly obtain the Rao-Blackwellized estimators of $E(\theta_{iT}|y_T)$ and $V(\theta_{iT}|y_T)$:

$$\hat{E}(\theta_{iT}|y_T) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d}$$
$$[(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)} x'_{iT}\beta_T^{(lk)}]/(Ld)$$

and

$$\hat{V}(\theta_{iT}|y_T) = \sum_{l=1}^{L}\sum_{k=d+1}^{2d} [\sigma_{iT}^2(1 - r_{iT}^{(lk)})]/(Ld)$$
$$+\sum_{l=1}^{L}\sum_{k=d+1}^{2d} [(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)} x'_{iT}\beta_T^{(lk)}]^2/(Ld)$$
$$-\{\sum_{l=1}^{L}\sum_{k=d+1}^{2d} [(1 - r_{iT}^{(lk)})y_{iT} + r_{iT}^{(lk)} x'_{iT}\beta_T^{(lk)}]\}^2/(Ld)^2,$$

where $r_{iT}^{(lk)} = \sigma_{iT}^2/(\sigma_{iT}^2 + \sigma_v^{2(lk)})$. Note that $E(\theta_{iT}|y_T)$ and $V(\theta_{iT}|y_T)$ use only the cross-sectional data at $t = T$. As a result, $E(\theta_{iT}|y_T)$ will be less efficient than the HB estimator $E(\theta_{iT}|y_T)$ based on all the data; see section 4.

4. APPLICATION TO THE LFS

4.1 Data Description and Implementation

We used the 1999 LFS unemployment estimates, y_{it}, in our HB analysis. There are 64 CAs across Canada. Employment Insurance (EI) beneficiary rates were used as auxiliary data, x_{it}, in the model. But the EI beneficiary data were available for only 62 CAs. So we included only those $m = 62$ CAs in the model. Within each CA, we considered six consecutive monthly estimates y_{it} from January 1999 to June 1999, so that $T = 6$ and the parameter of interest θ_{iT} is the true unemployment rate for area i in June, 1999. The reason that we only used six months of data is that the LFS sample rotation is based on a six-month cycle. Each month, one sixth of the LFS sample is replaced. Thus after six months, the correlation between estimates is very weak. The one-month lag correlation coefficient is about 0.48, and the lag correlation coefficients decrease as the lag increases. Figure 1 shows the estimated (smoothed) lag correlation coefficients for the LFS unemployment rate estimates. It is clear that after 6 months the lag correlation coefficients are all below 0.1.

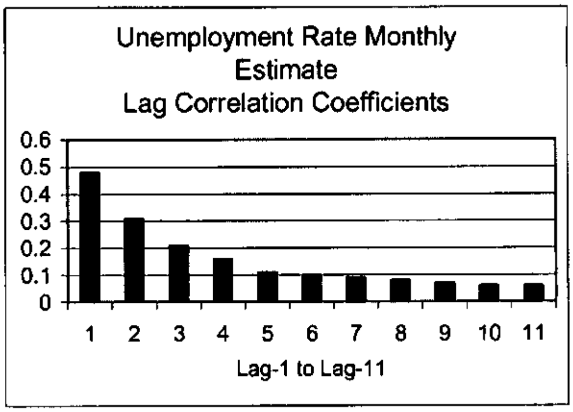


Figure 1. LFS unemployment rate lag correlation coefficients

To obtain a smoothed estimate of the sampling covariance matrix Σ_i used in the model, we first computed the average coefficient of variation (CV) for each CA over time (12 months in this study), denoted as $\overline{CV}_i$, and the average lag correlation coefficients over time and all CAs. By using these smoothed CVs and lag correlation coefficients, we obtained a smoothed estimate $\hat{\Sigma}_i$ with diagonal elements $\hat{\sigma}_{itt} = (\overline{CV}_i)^2 y_{it}^2$ and off-diagonal elements equal to $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|}(\hat{\sigma}_{itt}\hat{\sigma}_{iss})^{1/2}$ and treated $\hat{\Sigma}_i$ as the true Σ_i, where $\bar{\rho}_{|t-s|}$ is the average lag correlation coefficient of lag $|t-s|$. Our study found that using the smoothed estimate of Σ_i in the model can significantly improve the estimates in terms of CV reduction.

To implement the Gibbs sampling, we considered L=10 parallel runs, each of length $2d = 2{,}000$. The first $d = 1{,}000$ "burn-in" iterations were deleted. To monitor the convergence of the Gibbs sampler, for the parameters of interest $\theta_{iT}(i = 1, ..., m)$, we followed the method of

Gelman and Rubin (1992) involving the following steps: For each θ_{iT}, let $\theta_{iT}^{(lk)}$ denote the k-th simulated value in the l-th chain, $k = d+1, ..., 2d, l = 1, ..., L$. In the first step, compute the overall mean

$$\bar{\theta}_{iT} = \sum_{l=1}^{L} \sum_{k=d+1}^{2d} \theta_{iT}^{(lk)} / (Ld)$$

and the within sequence mean

$$\bar{\theta}_{iT}^{(l)} = \sum_{k=d+1}^{2d} \theta_{iT}^{(lk)} / d, \; l = 1, ..., L.$$

Then compute B_{iT}/d, the variance between the L sequence means as $B_{iT}/d = \sum_{l=1}^{L} (\bar{\theta}_{iT} - \bar{\theta}_{iT}^{(l)})^2 / (L-1)$. In the second step, calculate W_{iT}, the average of the L within sequence variances, s_{iTl}^2, each based on (d-1) degrees of freedom; that is, $W_{iT} = \sum_{l=1}^{L} s_{iTl}^2 / L$. In the third step, calculate $s_{iT}^2 = (d-1)\, W_{iT}/d + B_{iT}/d$ and $V_{iT} = s_{iT}^2 + B_{iT}/(Ld)$. In the last step, find the potential scale reduction factors $\hat{R}_{iT} = V_{iT}/W_{iT}$ $(i = 1, ..., m)$. If these potential scale reduction factors are near 1 for all of the scalar estimands θ_{iT} of interest, then this suggests that the desired convergence is achieved by the Gibbs sampler. In our study, the Gibbs sampler converged very well in terms of the values of $\hat{R}_{iT}$.

4.2 Model Selection

In this section, we compare the proposed model with the Rao and Yu (1994) AR(1) time component model. A number of methods for model comparison in a Bayesian framework have been developed, and several are implemented in the well-known BUGS program (see Spiegelhalter, Thomas, Best and Gilks 1996). In practice, when there is more than one model of interest, Bayesian model selection or model choice can be made on the basis of a Bayes factor, which is difficulty to calculate directly. Alternative strategies for model selection involve the predictive likelihood and predictive log-likelihood. In particular, Dempster (1974) suggested examining the posterior distribution of the log-likelihood of the observed data. The quantities of the posterior distribution of the log-likelihood may be obtained from the predictive posterior distribution of the deviance, $-2 \log f(y \mid \theta)$. The posterior deviance is straightforward to estimate using the Gibbs sampling output since it is the expectation of $-2 \log f(y \mid \theta)$ over the posterior $\pi(\theta \mid y)$. For non-hierarchical models, the minimum feasible value of $-2 \log f(y \mid \theta)$ is the traditional deviance statistic. For hierarchical models, the minimum of the deviance is likely to be very poorly estimated by the sample minimum, and the mean is a more reasonable measure (Karim and Zeger 1992; Gilks, Wang, Yvonnet and Coursagt 1993). For the AR(1) time component model, we considered two choices of ρ: $\rho = 0.75$ and $\rho = 0.5$. We calculated the log-likelihood at each iteration of the Gibbs sampler. Then we obtained the mean of the predictive posterior deviance: 1311.5 for the proposed model, 1372.8 for the AR(1) with $\rho = 0.5$ and 1358.3 for the AR(1) with $\rho = 0.75$. Thus, the deviance measure suggests that the random walk model on u_{it}'s provides a slightly better fit to the data than the AR(1) model.

For model comparison, we also computed the divergence measure of Laud and Ibrahim (1995) based on the posterior predictive distribution. Let θ^* represent a draw from the posterior distribution of θ given y, and let y^* represent a draw from $f(y \mid \theta^*)$. Then, marginally y^* is a sample from the posterior predictive distribution $f(y \mid y_{\text{obs}})$, where y_{obs} represents the observed data. The expected divergence measure of Laud and Ibrahim (1995) is given by $d(y^*, y_{\text{obs}}) = E(k^{-1} \| y^* - y_{\text{obs}} \|^2 \mid y_{\text{obs}})$, where k is the dimension of y_{obs}. Between two models, we prefer a model that yields a smaller value of this measure. As in Datta, Day and Maiti (1998) and Datta *et al.* (1999), we approximated the divergence measure $d(y^*, y_{\text{obs}})$ by using the simulated samples from the posterior predictive distribution. Using the Gibbs sampling output, we obtained a divergence measure of 13.36 for the proposed model, 14.62 for the AR(1) with $\rho = 0.5$ and 14.52 for the AR(1) with $\rho = 0.75$. Thus the divergence measure also suggests a slightly better fit of the random walk model compared to the AR(1) model.

It should be mentioned that the posterior deviance and the divergence measure are intended for comparing two or more alternative models. After selecting a model, we need to check if the selected model fits the data, which we turn to next.

4.3 Test of Model Fit

To check the overall fit of the proposed model, we used the method of posterior predictive p values (Meng 1994; Gelman, Carlin, Stern and Rubin 1995). In this approach, simulated values of a suitable discrepancy measure are generated from the posterior predictive distribution and then compared to the corresponding measure for the observed data. More precisely, let $T(y, \theta)$ be a discrepancy measure depending on the data y and the parameter θ. The posterior predictive p value is defined as

$$p = \text{prob}(T(y^*, \theta) > T(y_{\text{obs}}, \theta) \mid y_{\text{obs}}),$$

where y^* is a sample from the posterior predictive distribution $f(y \mid y_{\text{obs}})$. Note that the probability is with respect to the posterior distribution of θ given the observed data. This is a natural extension of the usual p value in a Bayesian context. If a model fits the observed data, then the two values of the discrepancy measure are similar. In other words, if the given model adequately fits the observed data, then $T(y_{\text{obs}}, \theta)$, should be near the central part of the histogram of the $T(y^*, \theta)$ values if y^* is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive p value is expected to be near 0.5 if the model adequately fits the data. Extreme p values (near 0 or 1) suggest poor fit. The p value is self-contained in the sense that it is computed without regard to an alternative model.

Computing the p value is relatively easy using the simulated values of θ^* from the Gibbs sampler. For each simulated value θ^*, we can simulate y^* from the model and compute $T(y^*,\theta^*)$ and $T(y_{\text{obs}},\theta^*)$. Then the p value is approximated by the proportion of times $T(y^*,\theta^*)$ exceeds $T(y_{\text{obs}},\theta^*)$. For the cross-sectional and time series model, the discrepancy measure used for overall fit is given by $d(y,\theta) = \sum_{i=1}^{m}(y_i-\theta_i)'\Sigma_i^{-1}(y_i-\theta_i)$. Datta *et al.* (1999) used the same discrepancy measure. We computed the p value by combining the simulated θ^* and y^* from all 10 parallel runs. We obtained a p value equal to 0.615. Thus we have no indication of lack of overall model fit for the random walk time series and cross-sectional model.

For the Fay-Herriot model that uses only the current cross-sectional data, an approximate discrepancy measure is given by

$$d_{\text{FH}}(y_T,\theta_T) = \sum_{i=1}^{m}(y_{iT}-\theta_{iT})^2/\sigma_{iT}^2,$$

where $\theta_T = (\theta_{1T},\ldots,\theta_{mT})'$. In this case, the estimated p value is about 0.587, indicating a good fit of the Fay-Herriot model for the current cross-sectional data only. However, the associated HB estimates are substantially less efficient compared to the HB estimates based on the proposed cross-sectional and time series model that borrows strength across regions and over time simultaneously; see Figures 3 and 4.

A limitation of the posterior predictive p value is that it makes "double use" of the observed data, y_{obs}, first to generate samples from $f(y|y_{\text{obs}})$ and then to compute the p value. This double use of the data can induce unnatural behaviour, as demonstrated by Bayarri and Berger (2000). To avoid double use of the data, Bayarri and Berger (2000) proposed two alternative p-measures, named the partial posterior predictive p value and the conditional predictive p value. These measures, however, seem to be more difficult to implement than the posterior predictive p value, especially for a complex model like the time series and cross-sectional small area model.

4.4 Estimation

We now obtain the posterior estimates of the unemployment rates under the random walk time series and cross-sectional model given by (3) and (4). We used the Rao-Blackwellized estimators, given in section 3.3, to obtain estimates for the posterior mean and the posterior variance of θ_{iT}. We denote these estimates by HB1. To study the impact of using a smoothed estimate of the sampling covariance matrix Σ_i, we also used the direct survey estimate of Σ_i in the model. We denote the estimates obtained in this case by HB2. To study the effect of borrowing strength over time, we also obtained the HB estimates of θ_{iT} under the Fay-Herriot model based only on the current cross-sectional data, denoted by HB3. Figure 2 displays the LFS direct estimates and the three HB estimates of the June 1999 unemployment rates for the 62 CAs across Canada. The 62 CAs appear in the order of population size with the smallest CA (Dawson Creek, BC, population is 10,107) on the left and the largest CA (Toronto, Ont., population is 3,746,123) on the right. For the point estimates, the Fay-Herriot model (HB3) tends to shrink the estimates towards the average of the unemployment rates, which leads to estimates that are too smooth in general. HB2 has more variation and tends to have more extreme values than HB1, since HB2 uses the direct estimates of Σ_i subject to sampling errors. HB1 leads to moderate smoothing of the direct LFS estimates. For the CAs with large population sizes and therefore large sample sizes, the direct estimates and the HB estimates are very close to each other; for smaller CAs, the direct and HB estimates differ substantially for some regions.

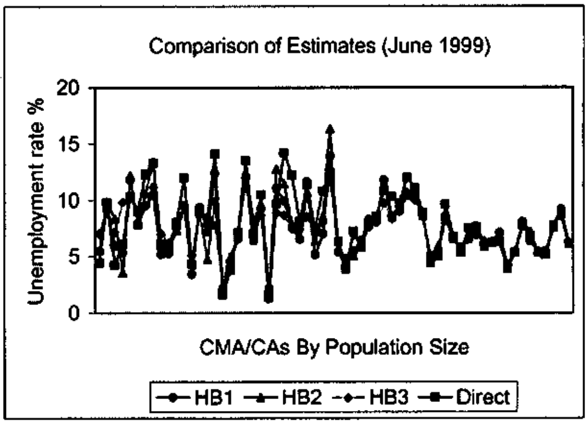


Figure 2. Comparison of direct and HB estimates

Figure 3 displays the coefficients of variation (CV) of the estimates. The CV of the HB estimate is taken as the ratio of the square root of the posterior variance and the posterior mean. It is clear from Figure 3 that the direct estimate has the largest CV and HB1 has the smallest CV. HB1 has smaller CV than HB2 for all CAs, and HB2 has smaller CV than HB3 for all CAs except two relatively small CAs. The efficiency gain of the HB estimates is obvious, particularly for the CAs with smaller population sizes.

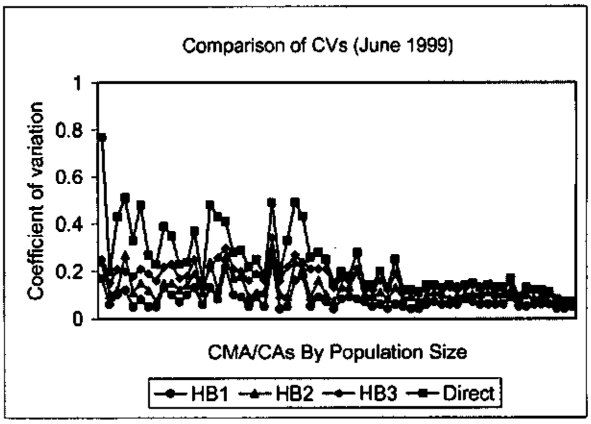


Figure 3. Comparison of CVs

Figure 4 shows the percent CV reduction over the direct survey estimates for HB1, HB2 and HB3. The percent CV reduction is defined as the difference of the LFS CV and the HB CV relative to the LFS CV and is expressed as a percentage. It is clear that HB1 achieves the largest CV reduction and that HB3 has the smallest reduction. The average percent reduction in CVs over the direct LFS estimates for the Fay-Herriot model (HB3) is 21%, for HB2 is 40%, and for HB1 is 62%. Also the CV reduction for smaller CAs is more significant than for larger CAs. As population size increases, the CV reduction tends to decrease.

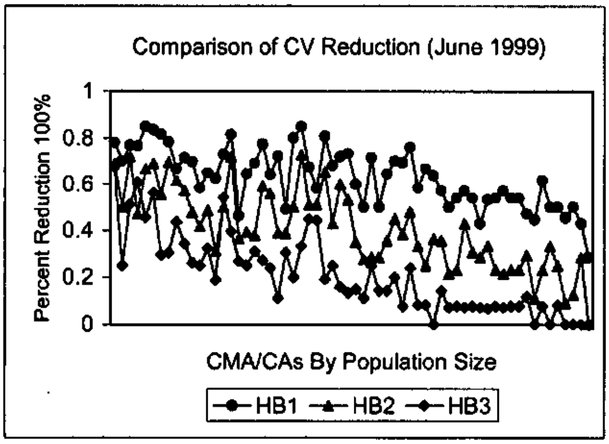


Figure 4. Comparison of CV reduction

In summary, we conclude the following: (1) The model-based HB estimates improve the direct LFS estimates. In particular, the cross-sectional and random walk time series model (HB1) improves the LFS estimates considerably in terms of CV reduction. (2) The cross- sectional and random walk time series model is more effective than the Fay-Herriot model. (3) Use of smoothed estimate of the sampling variance-covariance matrix Σ_i is very effective.

5. CONCLUDING REMARKS

In this paper we have presented a hierarchical Bayes cross-sectional and time series model to obtain model-based estimates of unemployment rates for CAs across Canada using LFS data. The model borrows strength across areas and over time periods simultaneously. Our analysis has shown that this model with a random walk process on the random time series components fits the data quite well. The hierarchical Bayes estimates, based on this model, improve the direct survey estimates significantly in terms of CV, especially for CAs with small population. However, these CVs are based on the assumption that the sampling variance covariance matrices Σ_i in the model are known. As a result, the uncertainty associated with the estimation of Σ_i is ignored.

We also considered the well-known Fay-Herriot model that combines cross-sectional information only, using the data at a specific time point, for example, at the current time of interest T. We found that the CVs under the Fay-Herriot model lie between the CVs for the direct and the model-based approach presented here. The cross-sectional and time series model is uniformly superior to the Fay-Herriot model in terms of CV reduction. This is expected since our model extends the Fay-Herriot model by borrowing strength over time as well as across space.

In our application to the LFS, we used simple smoothed estimates of the sampling variance-covariance matrices Σ_i and then treated them as the true Σ_i. We plan to study the sensitivity of the HB estimates of small area parameters θ_{iT} and the associated CVs to different methods of smoothing the Σ_i. In particular, it may be more realistic to use smoothed estimates of the form $\tilde{\sigma}_{itt} = (\overline{CV}_i)^2 \theta_{it}^2$ and $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|} (\tilde{\sigma}_{itt} \tilde{\sigma}_{iss})^{1/2}$ instead of the simple smoothed estimates we have used. However, it is more difficult to implement the HB method in this case since $\tilde{\sigma}_{itt}$ and $\tilde{\sigma}_{its}$ depend on the unknown parameters θ_{it}.

In this paper, we used a linear mixed linking model (3) for the parameters θ_{it}, which matches with the sampling model (1). Recently, You and Rao (2002) developed unmatched sampling and linking models for cross-sectional data, where the linking model is a non-linear mixed model, unlike the sampling model (1). You, Chen and Gambino (2002) extended this method to cross-sectional and time series data, using a log-linear linking model for θ_{it}.

ACKNOWLEDGEMENTS

The authors would like to thank two referees and the Editor for their helpful comments and suggestions. This work was partially supported by a research grant from the Natural Sciences and Engineering Research Council of Canada to J.N.K. Rao.

APPENDIX

A.1. Let $X = (X_1', ..., X_m')$, $\theta = (\theta_1', ..., \theta_m')$, $u = (u_1', ..., u_m')'$, with $\theta_i' = (\theta_{i1}, ..., \theta_{iT})$, $u_i' = (u_{i1}, ..., u_{iT})$. In the following, we list the full conditional distributions for the cross-sectional and time series model. For the proposed model (random walk time component), $\rho = 1$; for the alternative AR(1) time component model, $|\rho| < 1$.

- $\beta \mid y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim N((X'X)^{-1}(\theta - u), \sigma_v^2 (X'X)^{-1})$;
- $\sigma_v^2 \mid y, \beta, \sigma_\varepsilon^2, u, \theta \sim IG(a_1 + mT/2, b_1 + \sum_{i=1}^m \sum_{t=1}^T$ $(\theta_{it} - x_{it}' \beta - u_{it})^2/2)$;
- $\sigma_\varepsilon^2 \mid y, \beta, \sigma_v^2, u, \theta \sim IG(a_1 + m(T-1)/2, b_2 + \sum_{i=1}^m \sum_{t=2}^T$ $(u_{it} - \rho u_{i,t-1})^2/2)$;
- For $i = 1, ..., m$,

$$u_{i1} \mid y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta$$

$$\sim N((\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_\varepsilon^2})^{-1} (\frac{\theta_{i1} - x_{i1}' \beta}{\sigma_v^2} + \frac{\rho u_{i2}}{\sigma_\varepsilon^2}), (\frac{1}{\sigma_v^2} + \frac{\rho^2}{\sigma_\varepsilon^2})^{-1});$$

- For $i = 1, ..., m$, and $2 \leq t \leq T - 1$,

$$u_{i1}\,|\,y,\beta,\sigma_v^2,\sigma_\varepsilon^2,u_{i,t-1},u_{i,t+1},\theta$$
$$\sim N((\frac{1}{\sigma_v^2}+\frac{1+\rho^2}{\sigma_\varepsilon^2})^{-1}(\frac{\theta_{i1}-x_{i1}'\beta}{\sigma_v^2}+\frac{\rho u_{i,t-1}+\rho u_{i,t+1}}{\sigma_\varepsilon^2}),$$
$$(\frac{1}{\sigma_v^2}+\frac{1+\rho^2}{\sigma_\varepsilon^2})^{-1});$$

- For $i = 1, \dots, m$,

$$u_{i1}\,|\,y,\beta,\sigma_v^2,\sigma_\varepsilon^2,u_{i,T-1},\theta$$
$$\sim N((\frac{1}{\sigma_v^2}+\frac{1}{\sigma_\varepsilon^2})^{-1}(\frac{\theta_{iT}-x_{iT}'\beta}{\sigma_v^2}+\frac{\rho u_{i,T-1}}{\sigma_\varepsilon^2}),$$
$$(\frac{1}{\sigma_v^2}+\frac{1}{\sigma_\varepsilon^2})^{-1});$$

- For $i = 1, \dots, m$,

$$\theta_i\,|\,y,\beta,\sigma_v^2,\sigma_\varepsilon^2,u\sim N((\sigma_v^2 I_T+\Sigma_i^{-1})^{-1}$$
$$\times(\Sigma_i^{-1}y_i+\sigma_v^{-2}(X_i\beta+u_i)),(\sigma_\varepsilon^{-2}I_T+\Sigma_i^{-1})^{-1}).$$

A.2. Let $y_t=(y_{1t},\dots,y_{mt})'$, $X_t'=(x_{1t},\dots,x_{mt})$, $\theta_t'=(\theta_{1t},\dots,\theta_{mt})'$, $t=1,\dots,T$, we list the full conditional distributions for the Fay-Herriot model at time point t as follows:

- $\beta_t\,|\,y_t,\sigma_{vt}^2,\theta_t\sim N((X_t'X_t)^{-1}X_t'\theta_t,\sigma_{vt}^2(X_t'X_t)^{-1};$
- $\sigma_{vt}^2|y_t,\beta_t,\sigma_\varepsilon^2,u,\theta\sim IG(a_1+m/2,b_1+\sum_{i=1}^m(\theta_{it}-x_{it}'\beta_t)^2/2);$
- For $i = 1, \dots, m$,

$$\theta_{it}\,|\,y_t,\beta_t,\sigma_{vt}^2\sim N((1-r_{it})y_{it}+r_{it}x_{i1}'\beta_t,\sigma_{it}^2(1-r_{it})),$$
where $r_{it}=\sigma_{it}^2/(\sigma_{it}^2+\sigma_{vt}^2)$.

REFERENCES

BAYARRI, M.J., and BERGER, J.O. (2000). *P* values for composite null models. *Journal of the American Statistical Association*. 95, 1127-1142.

DATTA, G.S., DAY, B. and MAITI, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhya*. 60, 344-362.

DATTA, G.S., LAHIRI, P., MAITI, T. and LU, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*. 94, 1074-1082.

DATTA, G.S., LAHIRI, P. and MAITI, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*. 102, 83-97.

DEMPSTER, A.P. (1974). The direct use of likelihood for significance testing (with discussion). In *Proceedings of Conference on Foundational Questions in Statistical Inference* (Eds. O. Barndorff-Nielsen, P. Blaeslid and G. Schou). Dept. of Theoretical Statistics, University of Aarhus, Denmark. 335-354.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of Income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 74, 269-277.

GAMBINO, J.G., SINGH, M.P., DUFOUR, J., KENNEDY, B. and LINDEYER, J. (1998). *Methodology of the Canadian Labour Force Survey*. Statistics Canada, Catalogue No. 71-526.

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

GELFAND, A.E., and SMITH, A.F.M. (1991). Gibbs sampling for marginal posterior expections. *Communications In Statistics - Theory and Methods*. 20, 1747-1766.

GELMAN, A., CARLIN, J., STERN, H. and RUBIN, D. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.

GELMAN, A., and RUBIN, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*. 7, 457-472.

GHOSH, M., NANGIA, N. and KIM, D.H. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association*. 91, 1423-1431.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: An appraisal (with discussion). *Statistical Science*. 9, 55-93.

GILKS, W.R., WANG, C.C., YVONNET, B. and COURSAGT, P. (1993). Random-effects models for longitudinal data using Gibbs sampling. *Biometrics*. 49, 441-453.

KARIM, M.R., and ZEGER, S.L. (1992). Generalized linear models with random effects: Salamander mating revisited. *Biometrics*. 48, 631-644.

LAUD, P., and IBRAHIM, J. (1995). Predictive model selection. *Journal of Royal Statistical Society, Series B*. 57, 247-262.

MENG, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*. 22, 1142-1160.

PFEFFERMANN, D., FEDER, M. and SIGNORELLI, D. (1998). Estimation of autocorrelations of survey errors with application to trend estimation in small areas. *Journal of Business and Economic Statistics*. 16, 339-348.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*. 25, 175-186.

RAO, J.N.K., and YU, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*. 22, 511-528.

SPIEGELHALTER, D., THOMAS, A., BEST, N. and GILKS, W. (1996). BUGS 0.6: Bayesian inference Using Gibbs Sampling Manual. Available at http://www.mrc-bsu.cam.ac.uk/bugs.

YOU, Y. (1999). *Hierarchical Bayes and Related Methods for Model-Based Small Area Estimation*. Unpublished Ph.D. Thesis, School of Mathematics and Statistics, Carleton University, Ottawa, Canada.

YOU. Y., CHEN, E. and GAMBINO, J. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics* [CD-ROM], Alexandria, VA: American Statistical Association.

YOU, Y., and RAO, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*. 26, 173-181.

YOU, Y., and RAO, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*. 30, 3-15.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 33-44
Statistics Canada

The Effect of Model Choice in Estimation for Domains, Including Small Domains

RISTO LEHTONEN, CARL-ERIK SÄRNDAL and ARI VEIJANEN[1]

ABSTRACT

In this paper we examine the effect of model choice on different types of estimators for totals of domains, including small domains (small areas), for a sampled finite population. The paper asks: How do different estimator types compare for a common underlying model statement? We argue that estimator type (Synthetic, GREG, Composite, EBLUP, hierarchical Bayes, and so on) is one important aspect of domain estimation, and that the choice of the model, including its parameters and effects, is a second aspect, conceptually different from the first. Earlier work has not always kept this distinction clear. For a given estimator type, one can derive different estimators, depending on the choice of model. A number of estimator types have been proposed in the recent literature, but there is relatively little of an impartial comparison between them. In this paper we discuss three types: Synthetic, GREG, and, to a limited extent, Composite. We show that model improvement (the transition from a weaker to a stronger model) has very different effects on the different estimator types. We also show that the difference in accuracy between the different estimator types depends on the choice of model. For a well-specified model the difference in accuracy between Synthetic and GREG is negligible, but it can be substantial if the model is misspecified. Synthetic then tends to be highly inaccurate. We rely partly on theoretical results (for simple random sampling only), partly on empirical results. The empirical results are based on simulations with repeated samples drawn from two finite populations, one artificially constructed, the other constructed from real data from the Finnish Labour Force Survey.

KEY WORDS: Survey sampling; Generalized regression estimator; Synthetic estimator; Composite estimator; Multi-level models; Small areas; Small domains.

1. BACKGROUND

Most surveys require that estimates be made not only for the entire population under study but also for a number of sub-populations, called *domains* or *domains of interest.* Estevao and Särndal (1999) give a general outline of estimation for domains from a design-based perspective, with the use of auxiliary information. The sampling design is general, and so is the vector of auxiliary variables. The framework is also called model-assisted. Several national statistical agencies have in recent years constructed software that routinely handles domain estimation within the design-based, model-assisted framework. Examples of such software include CLAN97 by Statistics Sweden and GES by Statistics Canada. In a typical survey, some domains of interest are large enough, and the auxiliary information strong enough, so that the design-based estimators will be sufficiently accurate. But other domains may be so small (contain so few sampled units) that the design-based estimates will be too erratic. The statistical agency may then decide to suppress the publication of statistics for such domains.

Model-dependent estimates are less volatile, but an unattractive feature is their unknown bias, which can be substantial. The model-dependent synthetic estimator has occupied a prominent place in research on small area estimation from around 1970 and on, see for example, National Center for Health Statistics (1968), National Research Council (1980). Different estimators built on nested error regression models (Fuller and Battese 1973), random regression coefficients models (Dempster, Rubin and Tsutakawa 1981) and simple random effects models (Fay and Herriot 1979) provide examples of early propositions for alternatives to the synthetic estimator. Various composite estimators, constructed as weighted combinations of a model dependent estimator and a design-based estimator, were also proposed in the literature (for example Holt, Smith and Tomberlin 1979).

It was in connection with the synthetic estimator that the term "borrowing strength" began to be widely used. Today this term is invoked in virtually every one of the many published articles on small area estimation. Together, these articles now provide a rich source of possibilities for small area estimation, a majority of them model dependent. They draw on a variety of established statistical arguments and principles, such as generalized linear mixed models, composite estimation, empirical Bayes estimation, hierarchical Bayes, and so on.

Borrowing strength (or information) via modeling is a recurring theme in recent literature on small area estimation (for example Ghosh and Rao 1994; Pfeffermann 1999; Rao 1999). Borrowing strength is generally understood to mean that the estimator in use depends on data on the variable of interest, denoted y, from "related areas" or more generally from a larger area, in an effort to improve the accuracy for the small area. The resulting estimator is called *indirect,* in contrast to the one that uses y-data strictly from the domain itself, in which case it is called *direct.*

[1] Risto Lehtonen, University of Jyväskylä, Department of Mathematics and Statistics, P.O. Box 35 (MaD), FIN-40014 U. Jyväskylä, Finland; Carl-Erik Särndal, 2115 Embrook #44, Ottawa, Ontario, K1B 4J5; Ari Veijanen, Statistics Finland, P.O. Box 4 V, FIN-00022 Statistics Finland, Finland.

Underlying models and their features is another prominent theme in recent literature (for example Ghosh, Natarajan, Stroud and Carlin 1998; Marker 1999; Moura and Holt 1999; Prasad and Rao 1999; Feder, Nathan and Pfeffermann 2000). Small area estimates, and domain estimates more generally, are intrinsically linked to the idea of modeling. Holt and Rao (1995) hint that the use of y-information from other areas, although in a sense "necessary", should not be carried to an extreme. Instead there should be "specific allowance for local variation" through a model formulation that includes area-specific effects. This raises a certain ambiguity: borrowing strength from other areas is desirable, even necessary, but only within limits. It is unclear what these limits should be.

There is an extensive recent literature on small area estimation from a Bayesian point of view, including empirical Bayes and hierarchical Bayes techniques (for example Datta, Lahiri, Maiti and Lu 1999; Ghosh and Natarajan 1999; You and Rao 2000). Some recent publications relate frequentist and Bayesian approaches in small area estimation (for example Singh, Stukel and Pfeffermann 1998). Rao (2003) provides a good overview of current literature on model-based small area estimation.

The discussion in recent literature of domain estimation, including small area estimation, revolves around three crucial concepts: (i) borrowing strength; (ii) the type of (implicit or explicit) model, (iii) the parameters or effects admitted in the model statement, that is, whether they should be area specific or defined at some higher level of aggregation such as a set of "similar areas". We agree that these concepts are central and we use them in this paper.

Our starting point for the paper is summarized by (i) to (iii) as follows: (i) a number of different estimator types have been proposed for domain estimation and small area estimation: Synthetic estimator, Generalized Regression (GREG) estimator, Composite estimator, Empirical Best Linear Unbiased Predictor (EBLUP), empirical Bayes (EB) estimator, hierarchical Bayes estimator and so forth; (ii) for every estimator type, different estimators result from the choice of model; (iii) to borrow or not to borrow strength becomes an issue for some of the model choices. Attempts at borrowing strength takes place when the estimation of the parameters and effects in the model requires the use of y-values for units outside the domain itself.

2. STATEMENT OF OBJECTIVES

An objective in this paper is to examine domain estimation through a separation of two ideas: estimator type on the one hand, the choice of the underlying model on the other. We get a two-dimensional arrangement of possible estimators: By estimator type, by model choice. This distinction has not been emphasized enough in earlier literature.

We study the effect of model choice, and of model improvement, on selected estimator types: the Generalized Regression (GREG) estimator (which is design-based), the Synthetic (SYN) estimator (which is model dependent) and the Composite estimator with Empirical Best Linear Unbiased Predictor EBLUP as a special case (which also is model dependent). By construction, each type has its own particular features. For example the GREG estimator type is constructed to be design unbiased, the model dependent ones usually are not. The GREG estimator's variance, although of order n^{-1}, can be very large for a small domain if the "effective sample size" is small; GREG is a "strongly design consistent" estimator in that its relative bias (bias divided by standard deviation) tends to zero as $n^{-½}$. The SYN estimator is usually design biased; its bias does not approach zero with increasing sample size; its variance is usually smaller than that of GREG. The EBLUP is design consistent (although not strongly design consistent in the manner of GREG); is design biased for any fixed finite sample size; its variance ordinarily falls between that of GREG and that of SYN.

The chosen model specifies a hypothetical relationship between the variable of interest, y, and the vector of predictor variables, $\mathbf{x}$, and makes assumptions about its perhaps complex error structure. For every specified model, we can derive one GREG estimator, one SYN estimator, one composite estimator, by observing the respective construction principles. An "improved model" will influence all of GREG, SYN and composite, usually so that the MSE decreases. In other words, if Model A is better than Model B, the SYN estimator for Model A is usually better than the SYN estimator for Model B. The same is usually the case for GREG.

Model choice has two aspects: (i) the mathematical form, or the type, of the model, and (ii) the specification of the parameters and effects in the model. For a given variable of interest, some models are more appropriate than others. Model improvement can result either from a more appropriate model type, or from a better parametization, or both. We can distinguish linear models and nonlinear models. Logistic models are a special case of the latter. For a binary or polytomous variable of interest y, a (multinomial) logistic model type is arguably an improvement on a linear model type, because the fitted values under the former will necessary fall in the unit interval, which is not always true for a linear model. Lehtonen and Veijanen (1998) introduced the logistic GREG estimator and studied it in the context of the Finnish Labour Force survey. Another example is when a Bayesian model formulation is preferred to other forms.

The second aspect of model choice is the specification of the parameters and effects in the model. Some of these may be defined at the fully aggregated population level, others at the level of the domain (area specific parameters), yet others at some intermediate level (for a set of "related areas"). Using a multi-level model type, we can introduce stochastic effects that recognize domain differences, as in Goldstein (1995) for the SYN estimator and by Lehtonen and Veijanen (1999) for the GREG estimator. They found improved accuracy in small domains, compared to the GREG estimator based on a model with fixed effects at the population level. Generally, model improvement occurs when more parameters or effects are added to the model, as for example when it is formulated to include area specific effects reflecting local variation.

We show in this paper (i) that model improvement will generally, for any estimator type considered here, be accompanied by a decrease in MSE; (ii) that the effect on the MSE of model improvement is very different for different estimator types; (iii) that for a well-specified model, there are negligible differences only in the accuracy (the MSE) achieved by the estimator types under study, but under model failure the differences can be substantial. We emphasize that a comparison of estimators of different types should only take place under "similar conditions". That is, the model choice must be the same for all alternatives considered. An estimator is shown to be better than another estimator only if the MSE of the former is smaller than that of the latter, for one and the same model choice. (It is difficult to establish that one estimator type is uniformly better than another, that is, better under all model choices.)

Table 1 shows the estimators to be discussed, in a two-way arrangement by estimator type and by model choice. This table also shows our notation for the estimators to be considered. There are six SYN type estimators and six GREG type estimators in the table. Each of the six rows corresponds to a different model choice. A population model (P-model; rows 1 and 2) is one whose only parameters are fixed effects defined at the population level; it contains no domain specific parameters. A domain model (D-model) is one having at least some of its parameters or effects defined at the domain level. These are fixed effects for rows 3 and 4, or mixed with random effects for rows 5 and 6. "Linear" and "logistic" refer to the mathematical form. In this paper we discuss all estimators in Table 1 except the two in the last row.

Table 1
Schematic presentation of the SYN and GREG estimators by model choice and estimator type

			Estimator type	
	Model choice		Model-dependent synthetic	Model-assisted generalized regression
Fixed-effects models	Population models	Linear	SYN-P	GREG-P
		Logistic	LSYN-P	LGREG-P
	Domain models	Linear	SYN-D	GREG-D
		Logistic	LSYN-D	LGREG-D
Mixed models including fixed and random effects	Domain models	Linear	MSYN-D	MGREG-D
		Logistic	MLSYN-D	MLGREG-D

In addition to the SYN and GREG estimator types listed in Table 1, we can consider composite estimators of the type $\hat{\gamma}_d \text{GREG} + (1 - \hat{\gamma}_d)\text{SYN}$, being appropriately weighted combinations of the corresponding GREG and SYN estimators. In this paper we examine one estimator of this type, the EBLUP estimator.

The paper is organized as follows: Section 3 introduces three types of estimators for a domain total. In section 4, we describe the models used in the construction of these estimators. In section 5 we derive analytically the effect of model improvement, in a simple case. (Only simple cases can be treated analytically, because the formulas quickly attain a high degree of complexity, depending on the sampling design and other factors.) Section 6 is devoted to Monte Carlo simulations for two finite populations, illustrating the effect of model improvement on the three selected estimator types. Summary and discussion is given in section 7.

3. ESTIMATORS OF DOMAIN TOTALS

The finite population is denoted $U = \{1, 2, ..., k, ..., N\}$. A probability sample s is drawn from U by a given sampling design such that unit k is given the inclusion probability π_k. The sampling weight of unit k is then $a_k = 1/\pi_k$. Denote by y the variable of interest and by y_k its value for unit k. We consider a set of mutually exhaustive and exhaustive domains $U_1, ..., U_d, ..., U_D$. The target parameters are the set of domain totals, $Y_d = \sum_{U_d} y_k$, $d = 1, ..., D$.

Auxiliary information is essential for building accurate domain estimators, and increasingly so when domains of interest get smaller. Let $\mathbf{x}$ be the auxiliary vector of dimension $J \geq 1$ with a known value $\mathbf{x}_k$ for every unit $k \in U$. In a survey on individuals, $\mathbf{x}_k$ may specify known data about person k, such as age class, sex and other continuous or qualitative variable values. We assume that the vector value $\mathbf{x}_k$ and domain membership is known and specified in the frame for every $k \in U$. (For some estimators, it suffices to know the *total* of $\mathbf{x}_k$ for each domain of interest.)

The estimators we consider are constructed as follows: The first step is to estimate the designated model, using the sample data $\{(y_k, \mathbf{x}_k); k \in s\}$. Next, using the estimated parameter values, the vector value $\mathbf{x}_k$ and the domain membership of k, we compute the predicted value $\hat{y}_k$ for every $k \in U$, which is possible under our assumptions because $\mathbf{x}_k$ is known for every $k \in U$. The predictions, $\{\hat{y}_k; k \in U\}$, and the observations, $\{y_k; k \in s\}$, provide the material for the estimator types considered here.

Consider a fixed-effects model specification, linear or nonlinear, such that $\text{E}_m(y_k) = f(\mathbf{x}_k; \beta)$, for a given function $f(\cdot; \beta)$, where β is an unknown parameter vector requiring estimation, and E_m refers to the expectation under the model. The model fit yields the estimate $\hat{\beta}$. The supply of predicted values $\hat{y}_k = f(\mathbf{x}_k; \hat{\beta})$ is computed for $k \in U$. Similarly, for a linear mixed model involving random effects in addition to the fixed effects, the model specification is $\text{E}_m(y_k \mid \mathbf{u}_d) = \mathbf{x}_k'(\beta + \mathbf{u}_d)$ where $\mathbf{u}_d$ is a vector of random effects defined at the domain level. Using the estimated parameters, predicted values $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$ are computed for all $k \in U$. The models used in this paper are described in more detail in section 4. In more general terms, the models for the construction of GREG and SYN type estimators of domain totals are often members of the family of generalized linear mixed models (for example McCullogh and Searle 2001).

The predictions $\{\hat{y}_k; k \in U\}$ differ from one model specification to another. For a given model specification, the estimator of the domain total $Y_d = \sum_{U_d} y_k$ has the following structure for the three estimator types (Synthetic, Generalized Regression, Composite) to be studied:

$$\hat{Y}_{d\text{SYN}} = \sum_{U_d} \hat{y}_k \tag{3.1}$$

$$\hat{Y}_{d\text{GREG}} = \sum_{U_d} \hat{y}_k + \sum_{s_d} a_k(y_k - \hat{y}_k) \tag{3.2}$$

$$\hat{Y}_{d\text{COMP}} = \sum_{U_d} \hat{y}_k + \hat{\gamma}_d \sum_{s_d} a_k(y_k - \hat{y}_k) \tag{3.3}$$

where $a_k = 1/\pi_k, s_d = s \cap U_d$ is the part of the full sample s that falls in U_d, and $d = 1, ..., D$. $\hat{Y}_{d\text{SYN}}$ relies heavily on the truth of the model, and is usually biased. On the other hand, $\hat{Y}_{d\text{GREG}}$ has a second term that protects against model misspecification. The domain-specific weight $\hat{\gamma}_d$ in $\hat{Y}_{d\text{COMP}}$ is appropriately constructed to meet certain optimality properties, as explained in section 6. The weight $\hat{\gamma}_d$ approaches unity for increasingly large domain sample sizes, so that $\hat{Y}_{d\text{COMP}}$ approaches $\hat{Y}_{d\text{GREG}}$. At the other extreme, when $\hat{\gamma}_d$ is near zero, $\hat{Y}_{d\text{COMP}}$ is close to $\hat{Y}_{d\text{SYN}}$. We note that for a given model specification, (3.2) and (3.3) reduce to (3.1) for a domain d with no sample elements in s_d.

4. MODELS

4.1 Fixed-Effects Linear Models

Let $\mathbf{x}_k = (1, x_{1k}, ..., x_{jk}, ..., x_{Jk})'$ be a $(J+1)$-dimensional vector containing the values of $J \geq 1$ predictor variables $x_j, j = 1, ..., J$. This vector is used to create the predicted values $\hat{y}_k$ in the estimators (3.1), (3.2) and (3.3).

The estimators SYN-P and GREG-P build on the model specification (called the P-model)

$$\text{E}_m(y_k) = \mathbf{x}_k'\beta \tag{4.1}$$

for $k \in U$, where $\beta = (\beta_0, \beta_1, ...\beta_J)'$ is a vector of fixed effects defined for the whole population. If y-data were observed for the whole population, we could compute the generalized least squares (GLS) estimator of β given by

$$\mathbf{B} = \left(\sum_U \mathbf{x}_k\mathbf{x}_k'/c_k\right)^{-1} \sum_U \mathbf{x}_k y_k / c_k \tag{4.2}$$

where the c_k are specified positive weights. With no significant loss of generality we specify these to be of the form $c_k = \lambda'\mathbf{x}_k$ for $k \in U$, where the $(J+1)$-vector λ does not depend on k. Because (4.2) cannot be computed, the fit is carried out in practice on the observed sample data, yielding

$$\hat{\mathbf{B}} = \left(\sum_s a_k\mathbf{x}_k\mathbf{x}_k'/c_k\right)^{-1} \sum_s a_k\mathbf{x}_k y_k / c_k \tag{4.3}$$

where $a_k = 1/\pi_k$ is the sampling weight of unit k. The resulting predicted values are $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}$. They can be computed for all $k \in U$.

The estimators SYN-D and GREG-D are built with the same predictor vector $\mathbf{x}_k$, but with an improved model specification (called the D-model) allowing a fixed-effects vector β_d separately for every domain, so that

$$\text{E}_m(y_k) = \mathbf{x}_k'\beta_d \tag{4.4}$$

for $k \in U_d, d = 1, ..., D$, or equivalently,

$$\text{E}_m(y_k) = \sum_{d=1}^{D} \delta_{dk}\mathbf{x}_k'\beta_d \tag{4.5}$$

for $k \in U$, where δ_{dk} is the domain indicator of unit k, defined by $\delta_{dk} = 1$ for all $k \in U_d$, and $\delta_{dk} = 0$ for all $k \notin U_d, d = 1, ..., D$. If the model (4.3) could be fitted to data for the whole population, the GLS estimator of β_d would be

$$\mathbf{B}_d = \left(\sum_{U_d} \mathbf{x}_k\mathbf{x}_k'/c_k\right)^{-1} \sum_{U_d} \mathbf{x}_k y_k / c_k. \tag{4.6}$$

In practice, the fit must be based on the observed sample data, leading to

$$\hat{\mathbf{B}}_d = \left(\sum_{s_d} a_k\mathbf{x}_k\mathbf{x}_k'/c_k\right)^{-1} \sum_{s_d} a_k\mathbf{x}_k y_k / c_k. \tag{4.7}$$

The resulting predicted values are given by $\hat{y}_k = \mathbf{x}_k'\hat{\mathbf{B}}_d$ for $k \in U_d; d = 1, ..., D$. Because of the specification $c_k = \lambda'\mathbf{x}_k$, we have $\sum_{s_d} a_k(y_k - \hat{y}_k) = 0$. Consequently, SYN-D and GREG-D are identical, that is, $\hat{Y}_{d\text{SYN}-D} = \hat{Y}_{d\text{GREG}-D}$ for every sample s.

The transition from GREG-P to GREG-D, and from SYN-P to SYN-D, affects the MSE in a way to be analyzed in section 5. SYN-P and GREG-P will be examined empirically in section 6.

4.2 Linear Mixed Models

The estimators MSYN-D and MGREG-D build on a two-level linear model (called the D-model) involving fixed as well as random effects recognizing domain differences,

$$\begin{aligned}\text{E}_m(y_k \mid \mathbf{u}_d) &= \beta_0 + u_{0d} \\ &\quad + (\beta_1 + u_{1d})x_{1k} \\ &\quad + ... + (\beta_J + u_{Jd})x_{Jk} \\ &= \mathbf{x}_k'(\beta + \mathbf{u}_d)\end{aligned} \tag{4.8}$$

for $k \in U_d, d = 1, ..., D$. Each coefficient is the sum of a fixed component and a domain specific random component: $\beta_0 + u_{0d}$ for the intercept and $\beta_j + u_{jd}, j = 1, ..., J$ for the slopes. The components of $\mathbf{u}_d = (u_{0d}, u_{1d}, ..., u_{Jd})'$

represent deviations from the coefficients of the fixed-effects part of the model,

$$\mathrm{E}_m(y_k) = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_J x_{Jk} = \mathbf{x}_k' \beta, \qquad (4.9)$$

which agrees with (4.1). More generally, we can have that only some of the coefficients in (4.8) are treated as random, so that, for some j, $u_{jd} = 0$ for every d. One of the simplest special cases of (4.8), commonly used in practice, is the one that includes a domain-specific random intercept u_{0d} as the only random term, as in one of the models used in section 6. Another model used in section 6 is the special case of (4.8) for $J = 1$, with a random slope u_{1d} and a random intercept u_{0d}.

We insert the resulting fitted y-values, $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$, into (3.1) to obtain the two-level MSYN-D estimator. Inserting the fitted values, $\hat{y}_k = \mathbf{x}_k'(\hat{\beta} + \hat{\mathbf{u}}_d)$, into (3.2), we obtain the two-level MGREG-D estimator, introduced by Lehtonen and Veijanen (1999). MSYN-D and MGREG-D will be examined empirically in section 6.

For the simulations reported in section 6, we fitted the two-level model (4.8) by the iterative least squares fitting (IGLS) algorithm of Goldstein (1995). Random effects were estimated by equation (2.2.2) in Goldstein (1995). This algorithm appeals to an assumption that the random effects follow a joint normal distribution $N(\mathbf{0}, \Omega)$. Note however that this assumption of normality is in no way necessary to obtain favorable properties for the resulting MGREG-D estimator. It is nearly unbiased regardless of any such assumption. The fitting of a multi-level model is more demanding than the fitting of a linear fixed-effects model, since estimation of the covariance matrix Ω is required.

4.3 Logistic Models

The estimators LSYN-P and LGREG-P build on a multinomial logistic P-model. Assume an m-class polytomous response defined by the class variables y_i with value $y_{ik} = 1$ if k belongs to class i and $y_{ik} = 0$ otherwise, $i = 1, \ldots, m$, and modeled by

$$E_m(y_{ik}) = \frac{\exp(\mathbf{x}_k' \beta_i)}{\sum_{r=1}^{m} \exp(\mathbf{x}_k' \beta_r)} \qquad (4.10)$$

for $k \in U$, where $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ and $\beta_i = (\beta_{i0}, \beta_{i1}, \ldots \beta_{iJ})'$ are vectors of fixed effects defined for whole population. To avoid identifiability problems, we set $\beta_1 = 0$. The LSYN-P and LGREG-P estimators of the population frequency of class i in domain d, $Y_{id} = \sum_{U_d} y_{ik}$, are defined by (3.1) and (3.2), respectively, if we replace y_k and $\hat{y}_k$ by y_{ik} and $\hat{y}_{ik} = \exp(\mathbf{x}_k' \hat{\beta}_i)/(1 + \sum_{r=2}^{m} \exp(\mathbf{x}_k' \hat{\beta}_r))$, where β_i is the estimate of β_i obtained from the fit of (4.10).

LGREG-P was introduced and studied in Lehtonen and Veijanen (1998). LSYN-P and LGREG-P will be examined empirically in section 6, where $\hat{\beta}_i$ is derived as a pseudo-maximum likelihood estimator incorporating the sampling weights.

5. ANALYTIC EXAMINATION OF THE EFFECT OF MODEL IMPROVEMENT

In this section we analyze the transition from GREG-P to GREG-D, and from SYN-P to SYN-D in the case of Simple Random Sampling. For both estimator types, GREG and SYN, we find that the accuracy is improved when the model changes from the weaker P-model (4.1) (with fixed effects at the level of the whole population) to the stronger D-model (4.5) (admitting fixed effects at the domain level). Intuitively, this is to be expected. What is of interest here is the pattern of improvement. It is very different for the two types.

Our objective is to measure the effect of model improvement on $\hat{Y}_d$, where $\hat{Y}_d$ denotes either $\hat{Y}_{d\mathrm{GREG}}$ or $\hat{Y}_{d\mathrm{SYN}}$. For this purpose, we use the relative improvement in MSE,

$$\mathrm{RELIMP}(\hat{Y}_d) = (\mathrm{MSE}_{dP} - \mathrm{MSE}_{dD}) / \mathrm{MSE}_{dD} \qquad (5.1)$$

where MSE_{dP} and MSE_{dD} denote the MSE of $\hat{Y}_d$ under the P-model and under the D-model, respectively. Both MSE_{dP} and MSE_{dD} depend on the sampling design and on the composition of the $\mathbf{x}_k$-vector. The improvement factor (5.1) is in general a complex formula. It lends itself to easy analytic interpretation only in simple cases. Therefore, we examine here the case of Simple Random Sampling Without Replacement (SRS). For other designs and model formulations, empirical studies are necessary. One such study is reported in section 6.

We use the improvement factor (5.1) to measure the effect of changing from the P-model (4.1) (the weaker model) to the D-model (4.5) (the stronger model). The Technical Appendix gives the necessary expressions for bias and MSE of GREG and SYN estimators in the case of an SRS sample of size n from U. The size, n_d, of the sample from the domain U_d is random with expected value $nP_d = nN_d/N$. For GREG, we use (A.5) in Technical Appendix, and the two different forms of E_k presented there, to arrive at

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{GREG}}) = \frac{S^2_{E_P U_d}}{S^2_{E_d U_d}} - 1 + (1 - P_d)\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}}$$

$$\approx (1 - P_d)\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}} \qquad (5.2)$$

where $S^2_{E_d U_d} = (1/(N_d - 1)) \sum_{U_d} E^2_{dk}$ and $S^2_{E_P U_d} = (1/(N_d - 1)) \sum_{U_d} \{E_{Pk} - \bar{E}_{PU_d}\}^2$ with $\bar{E}_{PU_d} = \sum_{U_d} E_{Pk} / N_d$.

(Note that $E_{dU_d} = \sum_{U_d} E_{dk}/N_d = 0$). Similarly, for SYN, we use (A.6) in Technical Appendix, and the two different expressions for E_k presented there, to arrive at

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{SYN}}) = \frac{S^2_{(R_d E_p)U}}{P_d S^2_{E_d U_d}} - 1 + \frac{nP_d}{1-f}\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}}$$

$$\approx \frac{nP_d}{1-f}\frac{\bar{E}^2_{PU_d}}{S^2_{E_d U_d}} \qquad (5.3)$$

where $S^2_{(R_d E_p)U} = (1/(N-1)) \sum_U (R_{dk} E_{Pk})^2$. The approximation in (5.3) is a result of keeping only the term proportional to the total sample size n. By comparison, the other terms are negligible. The approximation in (5.3) is adequate in many cases, although the deleted part is not always insignificant. Comparing the improvement factors (5.2) and (5.3), we note:

(i) **Improvement factor as a function of the bias.** Comparing (5.2) and (5.3), we see that the improvement of SYN is large compared to that of GREG. The main reason is that SYN is handicapped, under the P-model, by an often considerable squared bias term. As the model improves, this handicap is greatly reduced. At the same time the variance term may increase moderately, so that, somewhat paradoxically, SYN becomes more volatile when the model is improved. For GREG, some improvement occurs when the model improves, as a result of a somewhat reduced variance. The improvement is small, compared to the dramatic improvement of SYN.

(ii) **Improvement factor as a function of domain size.** Suppose that $\bar{E}^2_{PU_d}/S^2_{E_d U_d}$ is constant for all domains. Then, the presence of the relative domain size P_d in (5.3) shows that $\hat{Y}_{d\mathrm{SYN}}$ improves more in larger domains than in small domains (where the need for accuracy improvement is relatively greater). For $\hat{Y}_{d\mathrm{GREG}}$, the pattern is more natural in that the improvement is more pronounced for the smaller domains, due to the factor $(1-P_d)$ in (5.2). But if $\bar{E}^2_{PU_d}/S^2_{E_d U_d}$ varies considerably between domains, these conclusions would be modified.

To throw further light on the generally complex improvement factors (5.2) and (5.3), consider the simple specification $\mathbf{x}_k = 1 = c_k$ for all k. Then $\hat{Y}_{d\mathrm{SYN}-P} = N_d \bar{y}_s$, $\hat{Y}_{d\mathrm{GREG}-P} = N_d \bar{y}_{s_d} - (1/f)(n_d - nP_d)\bar{y}_s$ with $f = n/N$ and $\hat{Y}_{d\mathrm{SYN}-D} = \hat{Y}_{d\mathrm{GREG}-D} = N_d \bar{y}_{s_d}$. (Overbar denotes the arithmetic mean over the set defined by the subscript.) Using $(N_d - 1)/(N-1) \approx N_d/N$, we get

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{GREG}}) \simeq (1-P_d)\frac{\left(\bar{y}_{U_d} - \bar{y}_U\right)^2}{S^2_{yU_d}} \qquad (5.4)$$

$$\mathrm{RELIMP}(\hat{Y}_{d\mathrm{SYN}}) \simeq P_d \frac{S^2_{yU}}{S^2_{yU_d}} - 1 + \frac{nP_d}{1-f}\frac{(\bar{y}_{U_d} - \bar{y}_U)^2}{S^2_{yU_d}}$$

$$\simeq \frac{nP_d}{1-f}\frac{(\bar{y}_{U_d} - \bar{y}_U)^2}{S^2_{yU_d}} \qquad (5.5)$$

where S^2_{yU} and $S^2_{yU_d}$ are the variances of y_k over U and U_d, respectively. The patterns are now very clear. The term $(\bar{y}_{U_d} - \bar{y}_U)^2/S^2_{yU_d}$ is present in both expressions. For SYN, we see from (5.5) that the improvement factor is proportional to the whole sample size n, hence it can be very large. For GREG, the improvement (5.4) is very small by comparison. If $(\bar{y}_{U_d} - \bar{y}_U)^2/S^2_{yU_d}$ is constant over all domains, GREG is improved more in smaller domains than in larger ones. The opposite holds for SYN.

The results in this section are limited by the complexity of the analytic expressions. Nevertheless they set the pattern for more general situations now to be studied by empirical examination. As the model improves, we can expect SYN to undergo a very large improvement, in terms of reduced MSE, compared to GREG.

6. EMPIRICAL EXAMINATION OF THE EFFECT OF MODEL IMPROVEMENT BY MONTE CARLO EXPERIMENTS

6.1 Experiments and Monte Carlo Summary Measures

The data for Experiment 1, presented in section 6.2, was generated entirely from a specified model, so it has no basis in any real data. For the 100 domains of this data set we compared the SYN estimator type (3.1) and the GREG estimator type (3.2) under different choices of model for a continuous variable of interest. We fitted a fixed-effects linear model (which created SYN-P and GREG-P estimators) and compared the results with those obtained from the fitting of a two-level linear model (which created MSYN-D and MGREG-D estimators).

In constructing the population for Experiment 2, presented in section 6.3, we took real data on ILO unemployment from Finland's Labour Force Survey (LFS) as a starting point for creating a larger artificial population with 84 regional domains. There, the variable of interest is binary (unemployed or not). We fitted, in addition to a fixed-effects linear model (which created SYN-P and GREG-P estimators) and a two-level linear model (which created MSYN-D and MGREG-D estimators), a fixed-effects binomial logistic model (which created LSYN-P and LGREG-P estimators). For this experiment we also constructed a composite estimator (3.3) as a weighted combination of GREG and SYN estimators, creating a COMP-D estimator.

In Experiments 1 and 2, by using estimates $\hat{Y}_d(s_v)$ from repeated samples $s_v; v = 1, 2, \ldots, K$, we computed for each domain $d = 1, \ldots, D$ the following Monte Carlo summary measures of bias, accuracy and relative improvement in MSE. We use two measures of accuracy, the relative root mean squared error (RRMSE) and the median absolute relative error (MdARE). For Experiment 1, where the response variable is continuous, these two measures give the same message about the accuracy. But for Experiment 2, where the response variable is binary, there is sometimes a difference in the conclusions drawn from the two measures.

(i) Absolute relative bias (ARB), defined as the ratio of the absolute value of bias to the true value:

$$\left| \frac{1}{K} \sum_{v=1}^{K} \hat{Y}_d(s_v) - Y_d \right| \Big/ Y_d. \tag{6.1}$$

(ii) Relative root mean squared error (RRMSE), defined as the ratio of the root MSE to the true value:

$$\sqrt{\frac{1}{K} \sum_{v=1}^{K} (\hat{Y}_d(s_v) - Y_d)^2} \Big/ Y_d. \tag{6.2}$$

(iii) Median absolute relative error (MdARE), defined as follows. For each simulated sample s_v; $v = 1, 2, \ldots, K$, the absolute relative error is calculated and a median is taken over the K samples in the simulation:

$$\underset{\text{over } v = 1, \ldots, K}{\text{Median}} \left\{ \left| \hat{Y}_d(s_v) - Y_d \right| \Big/ Y_d \right\}. \tag{6.3}$$

(iv) RELIMP, the relative improvement in MSE, defined in the manner of (5.1).

6.2 Experiment 1: Data Generated from a Model

Monte Carlo design

We used the two-level D-model (4.8) with $J = 1$ to generate an artificial population of one million elements distributed on 100 domains. The elements were randomly allocated to a set of 100 domains with probabilities proportional to $\exp(p_d)$ where p_d follows a uniform distribution in $(-3,3)$. In the generation of values for the x-variable and y-variable in the dth domain, $d = 1, \ldots, 100$, we operated in the following way. First, the values of the x-variable were obtained as independent realizations of $N(\mu_d, \sigma_d^2)$, where the domain-specific parameters (μ_d, σ_d^2) had first been generated from a bi-variate uniform distribution over $(5,15) \times (15,35)$. Then, the response variable values y_k were generated as

$$y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d}) x_k + \varepsilon_k \tag{6.4}$$

with $\beta_0 = 10$ and $\beta_1 = 0.6$. In (6.4), the values of ε_k are independent realizations of $N(0, 1)$, and the random effects u_{0d} and u_{1d} were realized from a bivariate normal distribution with $u_{0d} \sim N(0, 4)$, $u_{1d} \sim N(0, 0.01)$, $d = 1, \ldots, 100$. We report results for two values of the correlation of the random effects: (a) $\mathrm{Corr}(u_{0d}, u_{1d}) = 0$, and (b) $\mathrm{Corr}(u_{0d}, u_{1d}) = -0.5$. One case of a positive correlation, 0.5, was also studied but the results were similar with those in the zero correlation case and are thus omitted.

We examined four estimators: MSYN-D and MGREG-D based on the two-level D-model (4.8), $y_k = \beta_0 + u_{0d} + (\beta_1 + u_{1d}) x_{1k} + \varepsilon_k$, and SYN-P and GREG-P based on the fixed-effects P-model (4.9), that is, $y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$. Both sets of SYN and GREG estimators were calculated in the zero correlation and negative correlation cases. The conditions are thus ideal for MSYN-D and MGREG-D in the sense that the population follows exactly the model that lies behind these two estimators.

From the generated population we drew $K = 1{,}000$ samples, each of size $n = 10{,}000$, with Simple Random Sampling Without Replacement (SRS). For each estimator and for each domain, we computed the Monte Carlo summary measures of bias, accuracy and relative improvement in MSE in the manner described in (6.1), (6.2), (6.3) and (5.1). The Monte Carlo measures were then averaged with respect to a classification of the domains into Small (25 domains with average domain sample size <10), Medium-sized (50 domains with average domain sample size 10 ≥ and <50), and Large (25 domains with average domain sample size ≥ 50).

Results

The results for the cases of zero correlation (a) and negative correlation (b) are given in Tables 2 and 3. In both cases, SYN-P has a large bias (measured by the average ARB) for all the three domain size categories (Table 2). The bias is slightly larger in the zero correlation case. The bias in SYN-P is considerably reduced by MSYN-D, but is still significant in small domains. In the smallest domains, the estimated residuals (the estimates of the random effects) were biased towards zero, which created some bias in the estimates. The accuracy (measured by the average RRMSE and the average MdARE) of MSYN-D (based on the "ideal model") is much better than that of SYN-P (which is based on a population model). Accuracy gains are larger for the zero correlation case, and gains are substantial especially in larger domains. This result is in line with our theoretical results in section 5.

GREG-P and MGREG-D are essentially unbiased, confirming theory. Out of these two, accuracy is clearly better for MGREG-D, especially in small domains. In larger domains, accuracy gains are much smaller for the GREG estimator type than for the SYN estimator type. Bias and accuracy of GREG estimators are quite similar in both zero correlation and negative correlation cases.

Table 2
Average absolute relative biais (ARB) (%), average relative root mean squared error (RRMSE) (%) and average median absolute relative error (MdARE) (%) of total estimators in small, medium-sized and large domains of a synthetic population with (a) random slope and intercept independent or (b) random slope and intercept negatively correlated

	Average ARB (%)			Average RRMSE (%)			Average MdARE (%)		
	Expected domain size in sample			Expected domain size in sample			Expected domain size in sample		
	Small (1-9)	Medium (10-49)	Large (50+)	Small (1-9)	Medium (10-49)	Large (50+)	Small (1-9)	Medium (10-49)	Large (50+)
(a) Zero correlation									
Model-dependent SYN estimators									
SYN-P	10.29	12.37	10.54	10.3	12.4	10.6	10.3	12.4	10.5
MSYN-D	1.32	0.09	0.01	4.7	1.1	0.4	2.6	0.7	0.2
Model-assisted GREG estimators									
GREG-P	0.21	0.06	0.01	7.5	2.5	0.8	5.0	1.7	0.5
MGREG-D	0.83	0.03	0.01	4.8	1.1	0.4	2.7	0.7	0.2
(b) Negative correlation (-0.5)									
Model-dependent SYN estimators									
SYN-P	7.92	9.51	8.26	7.9	9.5	8.3	7.9	9.5	8.3
MSYN-D	1.20	0.09	0.01	4.2	1.1	0.4	2.5	0.7	0.2
Model-assisted GREG estimators									
GREG-P	0.18	0.05	0.01	6.4	2.1	0.6	4.2	1.4	0.4
MGREG-D	0.67	0.02	0.01	4.4	1.1	0.4	2.6	0.7	0.2

As the theoretical discussion in section 5 has also suggested, the effect on the SYN estimator type of model improvement depends strongly on the size of the domain. This is confirmed here: The D-model leads to a considerable MSE improvement (measured by the average RELIMP) for SYN. The improvement is striking for the large domains (Table 3). By contrast, the effect on the GREG estimator type of model improvement is modest, by comparison, and essentially independent of the domain size, as also suggested by the theoretical results.

Table 3
Average relative improvement in MSE (%) of total estimators in small, medium-sized and large domains of a synthetic population with (a) random slope and intercept independent or (b) random slope and intercept negatively correlated

	Average relative improvement in MSE (%)		
	Expected domain size in sample		
	Small (1-9)	Medium (10-49)	Large (50+)
(a) Zero correlation			
MSYN-D versus SYN-P	8.3	332.5	1278.3
MGREG-D versus GREG-P	1.9	6.0	3.7
(b) Negative correlation (-0.5)			
MSYN-D versus SYN-P	5.1	197.0	734.7
MGREG-D versus GREG-P	1.3	3.6	2.3

The reason for an improved behavior of SYN and GREG estimators is that a two-level (or more generally, a multi-level) model, because of the presence of domain parameters, produces fitted values $\hat{y}_k$ that are on the average closer to the (unobserved) y_k than those obtained by fitting simply the fixed part of the model. In addition, since MSYN-D takes domain differences into account, it is expected to be less biased than the SYN-P estimator based on the fixed part of the two-level model. Still, we find that the MSYN-D estimator has a significant bias, particularly in the smallest domains, for which the estimated random effects tend to be biased towards zero, which pulls the fitted values in the direction of those of the fixed part of the model. MSYN-D and MGREG-D estimators do not differ considerably in their accuracy, even in small domains.

6.3 Experiment 2: Data Adapted from Finland's Labour Force Survey

Monte Carlo design

The empirical data for our Experiment 2 came from the Finnish Labour Force Survey (LFS), conducted monthly by Statistics Finland. Details on the design and the estimation procedure of the LFS are described in Djerf (1997). In this experiment, we estimate the number of unemployed in 84 administrative regions of Finland, based on the NUTS4 classification (European Union's Nomenclature of Territorial Units for Statistics).

To emulate the sampling design of the Finnish LFS, in a fairly realistic manner, we generated a large artificial population by expanding a one-quarter sample data set of the Finnish LFS. The original data set of 32,564 individuals contained 29,024 respondents. The respondents were replicated by Simple Random Sampling With Replacement until we had reached a total of 3 million records approximating the size of the labour force in Finland.

The variable of interest, y, was a binary variable describing whether a person was unemployed of not. In LFS, the definition of unemployment is based on the ILO (International Labour Organisation) concept. Our population data included four auxiliary variables available from administrative registers (and used by Statistics Finland in their LFS): age, sex, region (NUTS2 level regional unit) and the job-seeker indicator, which is a dichotomous indicator showing whether or not a person is registered as an unemployed job-seeker in the administrative records of Finland's Ministry of Labour. Indicator variables were used for 6 age-by-sex classes (3 age groups, 2 sexes). These register-based data were merged with the survey data at the micro level by using personal identification numbers, which are unique in both data sources.

We examined seven estimators. Three model choices were used. First, we constructed the estimators (3.1) and (3.2), based on the linear fixed-effects P-model (4.9) incorporating the main effects for variables age, sex, region and the job-seeker indicator. The model also incorporates the two-variable interaction of age with the job-seeker indicator. The variables and terms in the model were selected in an exploratory data analysis. The resulting domain total estimators are SYN-P and GREG-P.

Secondly, we constructed the estimators (3.1) and (3.2) based on a binomial logistic model (4.10) involving the same model structure as the P-models for SYN-P and GREG-P. The resulting estimators are LSYN-P and LGREG-P.

Thirdly, we constructed the estimators (3.1) and (3.2) based on the two-level D-model (4.8) again involving the same structure in the fixed part as the previous models. The random component of the model, recognizing domain differences, consisted of random intercepts at the domain (NUTS4) level. The resulting estimators are MSYN-D and MGREG-D. For this model choice, we also constructed the composite estimator (3.3). The resulting estimator is denoted by COMP-D. The weight $\hat{\gamma}_d$ in COMP-D was computed as $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2/n_d)$, where $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ are sample based estimates for unknown parameters in the model's error structure (Ghosh and Rao 1994). The COMP-D estimator is perhaps best described as a pseudo EBLUP (Prasad and Rao 1999), by the fact that the residuals $y_k - \hat{y}_k$ are sample weighted. (A more conventional EBLUP uses unweighted residuals.)

We carried out four independent Monte Carlo experiments. In each experiment, we drew from the generated LFS population $K = 1{,}000$ samples, each of size $n = 12{,}000$ individuals, with SRS. We generated non-response in each sample using a model for the non-response. We modeled the non-response by a logistic model incorporating the same auxiliary variables as the LGREG-P model. The non-response probabilities were estimated from each sample, and the sampling weights were adjusted accordingly. For each estimator and for each domain, we computed the Monte Carlo summary measures defined in section 6.1. These measures were then averaged with respect to a classification of the domains into Small (32 domains with average domain sample size < 60) and Large (52 domains with average domain sample size ≥ 60). We finally averaged these figures over the four experiments.

Results

Table 4 shows the results for the seven estimators. In this experiment based on a real population, the results are far less dramatic than in Experiment 1. For all the models, the model-dependent SYN estimators SYN-P, LSYN-P and MSYN-D had a substantial bias. The bias was smallest, even though still substantial, for the multilevel-model based estimator MSYN-D. The bias continued to be large even in the large domains. Large bias might be due to the poor fit of the models, even if we used the best models available, and because the inclusion of random effects in the models was quite limited (only a random intercept term was included at the domain level). Accuracy in model-dependent estimators was best again for MSYN-D. As shown in Table 5, there was a slight positive effect of model improvement in MSE.

Table 4
Average absolute relative bias (ARB) (%), average relative root mean squared error (RRMSE) (%) and average median absolute relative error (MdARE) (%) of estimators of the number of ILO unemployed in small and large domains (LFS data)

	Average ARB (%)		Average RRMSE (%)		Average MdARE (%)	
	Expected domain size in sample		Expected domain size in sample		Expected domain size in sample	
	Small (1-59)	Large (60+)	Small (1-59)	Large (60+)	Small (1-59)	Large (60+)
Model-dependent SYN estimators						
SYN-P	36.5	14.2	37.6	16.3	36.6	14.9
LSYN-P	36.4	14.1	37.3	16.2	36.5	14.8
MSYN-D	27.3	9.1	31.8	15.9	29.0	12.1
Model-assisted GREG estimators						
GREG-P	1.2	0.6	46.7	24.0	30.6	16.0
LGREG-P	1.2	0.6	46.8	24.0	30.7	16.0
MGREG-D	1.2	0.6	46.4	24.0	30.6	16.0
Composite estimators						
COMP-D	26.9	8.8	31.8	16.0	28.9	12.1

In model-assisted GREG estimators, the differences in bias and accuracy were small between the multilevel-model assisted MGREG-D estimator and the GREG-P and LGREG-P estimators assisted by population-level fixed effects models. The fixed-effects linear and logistic models yielded quite similar results, but the multilevel model improved the results slightly, as shown in Table 5.

Table 5
Average relative improvement in MSE (%) of estimators of the number of ILO unemployed in small and large domains (LFS data)

Average relative improvement in MSE (%)		
	Expected domain size in sample	
	Small (1-59)	Large (60+)
MSYN-D versus SYN-P	32.4	1.3
MGREG-D versus GREG-P	0.4	0.2

As measured by the average MdARE, the difference in accuracy between MSYN-D and MGREG-D is small in small domains.

The composite estimates COMP-D were close to the synthetic estimates because the estimated variance of the random intercept was, in most cases, quite small.

7. SUMMARY AND DISCUSSION

In the introduction we made a point that, in our opinion, has not been emphasized in earlier literature on domain estimation, namely that the concept "model choice" must be distinguished from the concept "estimator type" when estimation methods are compared. To one and the same choice of model (same mathematical form, same specification of parameters or effects in the model) corresponds one estimator for each of the traditional estimator types discussed in the literature, Synthetic, Generalized Regression, Composite, EBLUP and so on. A first consequence of this is that one cannot make a fair comparison of estimators of different types unless all share the same model choice. Secondly, a change of model, say from a weaker to a stronger model, may have quite different impact on different estimator types. It is this second aspect that is highlighted in this paper.

We have studied the impact of model improvement especially for the Synthetic (SYN) type and Generalized Regression (GREG) type estimators, and found that the impact is very different, and the impact depends heavily of the size of the domain concerned, that is, of the number of sampled units in a domain. Especially in larger domains, the impact of model improvement is very large for SYN type estimators, and modest only for GREG type estimators. The progression is such that a SYN type estimator goes from being highly inaccurate estimator for a weaker model to a much improved estimator for a stronger model. In other words, SYN is highly dependent on the strength of the model. This is not the case for a GREG type estimator. It is slightly more accurate for the stronger model while maintaining a high accuracy for both kinds of models. Its improvement factor is modest compared to a SYN type estimator. We have not carried out our analysis in detail for other estimator types. This is an objective for future research.

The possibilities for efficient estimation for domains and small areas depend on the available statistical infrastructure. As evidenced in many recent papers on small area estimation, one must often start from a set of premises, where the data for model fitting are available not at a unit level, but at some aggregated level (this situation is typical for example in the United Kingdom and in the United States). The background for the methods described in this paper is typical in statistical infrastructures where a good supply of administrative registers exists, with data at the unit level (this holds for example the Scandinavian countries). In such an infrastructure it is often possible to use unit keys, such as personal identification numbers, to merge two or more administrative files at the micro level in building the vector of auxiliary variables. Also, domain membership is often specified for all units in the target population, as assumed in this paper. We can also assume that the collected survey data file can be merged with the auxiliary data file using the unit keys. The situation described above is increasingly found in many countries, for example in several member states of the European Union, where an increasing emphasis is being put on the use of administrative registers for purposes of statistics production.

TECHNICAL APPENDIX

This technical appendix includes the derivation of bias and MSE approximations for GREG and SYN estimators needed for the examination of the effect of model improvement in the case of Simple Random Sampling presented in section 5.

To measure how the accuracy $\hat{Y}_{d\text{GREG}}$ and $\hat{Y}_{d\text{SYN}}$ changes as the model progresses from (4.1) to (4.5), we need to evaluate the variance of each estimator, as well as the bias of $\hat{Y}_{d\text{SYN}}$. By contrast, $\hat{Y}_{d\text{GREG}}$ is nearly unbiased. An obstacle in the analysis of $\hat{Y}_{d\text{GREG}}$ and $\hat{Y}_{d\text{SYN}}$ is their nonlinear form. Therefore we work with the corresponding linearized forms, for which we can easily obtain the bias and the variance. The results are then used to approximate the corresponding characteristics of $\hat{Y}_{d\text{GREG}}$ and $\hat{Y}_{d\text{SYN}}$. Taylor linearization is a standard technique for these types of estimators, as illustrated, for example, in Särndal, Swensson and Wretman (1992), Chapter 6.

Consider first the GREG estimators, GREG-P and GREG-D. Let $\hat{Y}_{d\text{GREG}}$ denote either of those two. With linear approximation, the estimation error (the estimator's deviation from the target parameter Y_d) is

$$\hat{Y}_{d\text{GREG}} - Y_d \approx \sum\nolimits_s a_k \delta_{dk} E_k - \sum\nolimits_U \delta_{dk} E_k \qquad \text{(A.1)}$$

where E_k is the population fit residual for k. The difference between GREG-P and GREG-D lies in the residuals E_k. For GREG-P, they are $E_k = E_{Pk}$, where $E_{Pk} = y_k - \mathbf{x}_k' \mathbf{B}_P$

for $k \in U$, with $\mathbf{B}_P$ given by (4.2). For GREG-D, they are $E_k = E_{dk}$, with $E_{dk} = y_k - \mathbf{x}_k' \mathbf{B}_d$ for $k \in U_d, d = 1, ..., D$, with $\mathbf{B}_d$ given by (4.6).

In (A.1), $\sum_s a_k \delta_{dk} E_k$ is the Horvitz-Thompson (HT) estimator for the variable $\delta_{dk} E_k$. Using basic results for the HT estimator we get $E(\hat{Y}_{d\text{GREG}}) - Y_d \approx 0$, that is, $\hat{Y}_{d\text{GREG}}$ is nearly unbiased. It is easy to state the variance for a general design. We need it here for the special case of Simple Random Sampling Without Replacement (SRS). The MSE of $\hat{Y}_{d\text{GREG}}$ equals the variance of $\hat{Y}_{d\text{GREG}}$, to the order of approximation used here.

Next, consider the SYN estimators, SYN-P and SYN-D. Let $\hat{Y}_{d\text{SYN}}$ denote either of those two. After linearization, the estimation error is approximated as

$$\hat{Y}_{d\text{SYN}} - Y_d \approx \sum_s a_k r_{dk} E_k - \sum_U \delta_{dk} E_k \qquad \text{(A.2)}$$

where $E_k = E_{dk}, r_{dk} = \delta_{dk}$ for SYN-D, and $E_k = E_{Pk}, r_{dk} = R_{dk}$ for SYN-P, with

$$R_{dk} = \left(\sum_{U_d} \mathbf{x}_k\right)\left(\sum_U \frac{\mathbf{x}_k \mathbf{x}_k'}{c_k}\right)^{-1} \frac{\mathbf{x}_k}{c_k}.$$

The term $\sum_s a_k r_{dk} E_k$ in (A.2) is the HT estimator for the variable $r_{dk} E_k$. The quantities R_{dk} vary around a central value at or near the relative domain size, $P_d = N_d / N$. The mean $(1/N) \sum_U R_{dk}$ equals P_d if $\mathbf{x}_k$ contains the constant "1" for every k. From (4.2) we get

$$E(\hat{Y}_{d\text{SYN}}) - Y_d \approx -\sum_{U_d} E_k. \qquad \text{(A.3)}$$

The right hand side of (A.3) is zero for SYN-D, which is therefore nearly unbiased, but is different from zero for SYN-P, which is therefore biased.

For the fixed-effects linear model formulations in section 4.1, we now examine the relative improvement factor (5.1) under SRS with a sampling fraction equal to $f = n/N$.

Consider first the two GREG estimators. We get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{GREG}}) &\approx V_T(\hat{Y}_{d\text{GREG}}) \\ &= N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_U \\ &\qquad \left\{\delta_{dk} E_k - \frac{1}{N}\left(\sum_U \delta_{dk} E_k\right)\right\}^2 \qquad \text{(A.4)} \end{aligned}$$

where the index T indicates the approximations derived via the linearized $\hat{Y}_{d\text{GREG}}$, and $E_k = E_{Pk}$ for the P-model and $E_k = E_{dk}$ for the D-model. Developing the square in (A.4) and using $(N_d - 1)/N_d \approx 1$ and $(N_d - 1)/(N - 1) \approx N_d / N$ we get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{GREG}}) &\approx V_T(\hat{Y}_{d\text{GREG}}) \\ &= N_d^2 \frac{1-f}{n_{d0}} \left\{S_{EU_d}^2 + (1 - P_d) \bar{E}_{U_d}^2\right\} \qquad \text{(A.5)} \end{aligned}$$

where $n_{d0} = n P_d = n(N_d / N)$ is the expected size of the domain portion of the sample, $s_d = s \cap U_d$, and $S_{EU_d}^2 = (1/(N_d - 1)) \sum_{U_d} \{E_k - \bar{E}_{U_d}\}^2$ with $\bar{E}_{U_d} = (1/(N_d)) \sum_{U_d} E_k$. If n_{d0} is small, $\hat{Y}_{d\text{GREG}}$ has a poor precision (a high variance), except if the model fits extremely well so that the residual E_k is small for all units in the domain. For GREG-D, $E_{U_d} = 0$, so the second term within curly brackets disappears.

Next, consider the two SYN estimators. We get

$$\begin{aligned} \text{MSE}_T(\hat{Y}_{d\text{SYN}}) &= N^2 \frac{1-f}{n} \frac{1}{N-1} \sum_U (r_{dk} E_k)^2 \\ &\quad + N_d^2 \bar{E}_{U_d}^2 \qquad \text{(A.6)} \end{aligned}$$

where r_{dk} and E_k are as specified in (A.2). The first term in (A.6) is the variance; the second is the squared bias obtained from (A.3). The variance term is often very small because the sample size in the denominator is that of the entire sample, not the perhaps much smaller size of the domain part of the sample. The squared bias term is zero for SYN-D, but non-zero, perhaps large, and not tending to zero for SYN-P.

REFERENCES

DATTA, G.S., LAHIRI, P., MAITI, T. and LU, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*. 94, 1074-1082.

DEMPSTER, A.P,. RUBIN, D.B. and TSUTAKAWA, R.K. (1981). Estimation in covariance component models. *Journal of the American Statistical Association*. 76, 341-353.

DJERF, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*. 13, 29-39.

ESTEVAO, V.M., and SÄRNDAL, C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*. 25, 213-221.

FAY, R.E., and HERRIOT, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 74, 269-277.

FEDER, M., NATHAN, G. and PFEFFERMANN, D. (2000). Multilevel modelling of complex survey longitudinal data with time varying random effects. *Survey Methodology*. 26, 53-65.

FULLER, W.A., and BATTESE, G.E. (1973). Transformations for linear models with nested error structure. *Journal of the American Statistical Association*. 68, 626-632.

GHOSH, M., and RAO, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science*. 9, 55-93.

GHOSH, M., NATARAJAN, K., STROUD, T.W.F. and CARLIN, B. (1998). Generalized linear models for small area estimation. *Journal of the American Statistical Association*. 93, 273-282.

GHOSH, M., and NATARAJAN, K. (1999). Small area estimation: a Bayesian perspective. In Ghosh, S. (ed.). *Multivariate Analysis, Design of Experiments, and Survey Sampling*. New York: Marcel Dekker. 69-92.

GOLDSTEIN, H. (1995). *Multilevel Statistical Models*. 2nd edition. London: Arnold; New York: John Wiley & Sons, Inc.

HOLT, D., and RAO, J.N.K. (1995). Topic 3: Small area estimation. *Bulletin of the International Statistical Institute* 50th session. 56 (book 4), 1648-1650.

HOLT, D., SMITH, T.M.F. and TOMBERLIN, T.J. (1979). A model-based approach to estimation for small subgroups of population. *Journal of the American Statistical Association*. 74, 405-410.

LEHTONEN, R., and VEIJANEN, A. (1998). Logistic generalized regression estimators. *Survey Methodology*. 24, 51-55.

LEHTONEN, R., and VEIJANEN, A. (1999). Domain estimation with logistic generalized regression and related estimators. Proceedings, IASS Satellite Conference on Small Area Estimation, Riga, August 1999. Riga: Latvian Council of Science, 121-128.

MCCULLOGH, C.E., and SEARLE, S.R. (2001). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons, Inc.

MARKER, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*. 15, 1-24.

MOURA, F.A.S., and HOLT, D. (1999). Small area estimation using multilevel models. *Survey Methodology*. 25, 73-80.

NATIONAL CENTER FOR HEATH STATISTICS (1968). Synthetic State Estimates of Disability. PHS publication no. 1959. Washington, DC: Public Health Service, US Government Printing Office.

NATIONAL RESEARCH COUNCIL (1980). Panel on Small-Area Estimates of Population and Income. Estimating Population and Income of Small Areas. Washington, DC: National Academy Press.

PFEFFERMANN, D. (1999). Small area estimation – big developments. Proceedings, IASS Satellite Conference on Small Area Estimation, Riga, August 1999. Riga: Latvian Council of Science, 129-145.

PRASAD, N.G.N., and RAO, J.N.K. (1999). On robust small area estimation using a simple random effects model. *Survey Methodology*. 25, 67-72.

RAO, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*. 25, 175-186.

RAO, J.N.K. (2003). *Small Area Estimation*. Hoboken, New York: John Wiley & Sons, Inc.

SÄRNDAL, C.-E., SWENSSON, B. and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

SINGH, A.C., STUKEL, D.M. and PFEFFERMANN, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society, B*. 60, 377-396.

YOU, Y., and RAO, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*. 26, 173-181.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 45-52
Statistics Canada

Adjustment of Unemployment Estimates Based on Small Area Estimation in Korea

YEON SOO CHUNG, KAY-O LEE and BYUNG CHUN KIM[1]

ABSTRACT

The Korean Economically Active Population Survey (EAPS) has been conducted in order to produce unemployment statistics for Metropolitan Cities and Provincial levels, which are large areas. Large areas have been designated as planned domains, and local self-government areas (LSGA's) as unplanned domains in the EAPS. In this study, we suggest small area estimation methods to adjust for the unemployment statistics of LSGA's within large areas estimated directly from current EAPS data. We suggest synthetic and composite estimators under the Korean EAPS system, and for the model-based estimator we put forward the Hierarchical Bayes (HB) estimator from the general multi-level model. The HB estimator we use here has been introduced by You and Rao (2000). The mean square errors of the synthetic and composite estimates are derived by the Jackknife method from the EAPS data, and are used as a measure of accuracy for the small area estimates. Gibbs sampling is used to obtain the HB estimates and their posterior variances, and we use these posterior variances as a measure of precision for small area estimates. The total unemployment figures of the 10 LSGA's within the ChoongBuk Province produced by the December 2000 EAPS data have been estimated using the small area estimation methods suggested in this study. The reliability of small area estimates is evaluated by the relative standard errors or the relative root mean square errors of these estimates. We suggest here that under the current Korean EAPS system, the composite estimates are more reliable than other small area estimates.

KEY WORDS: Synthetic estimator; Composite estimator; Hierarchical Bayes; Multi-level model; Jackknife mean square error; Gibbs sampling.

1. INTRODUCTION

Sample surveys are a more cost-effective way of obtaining information than complete enumerations or censuses for most purposes. The surveys are usually designed to ensure that reliable estimates of totals and means for the population, pre-specified domains of interest, or major subpopulations can be derived from the survey data. There are also many situations in which it is desirable to derive reliable estimates for additional domains of interest, especially geographical areas or subpopulations, from existing survey data.

The Korean National Statistical Office conducts the Economically Active Population Survey (EAPS) in 30,000 sample households every month. The characteristics of the economically active for 16 large areas (7 Metropolitan Cities, 9 Provinces) of the country are based on these monthly EAPS results. The EAPS is a large city or provincial level survey. Many small cities in a large area would prefer to obtain the unemployment figures for individual cities without conducting their own survey, and the most cost-effective way would be to turn to the EAPS data. However, small cities belonging to a large area are unplanned regions in the EAPS and sample sizes for these small cities are typically too small due to the size of small cities. Therefore, if we estimate the unemployment statistics of small areas from the EAPS framework based on large areas, we may be unable to obtain an estimate with adequate precision since the sample size in specific small areas may not be large enough. The direct estimates for specific small areas from the EAPS cannot be sufficiently reliable in this situation. It is hence necessary to "borrow strength" from related areas to obtain more reliable estimates for a given small area. An example of such would be to gather separately published administrative records of related small areas. We define related areas as those areas with similar economic and demographic characteristics as the small area we wish to estimate. Our aim is to adjust the direct estimates derived from the National Statistical Office of Korea through design-based and model-based indirect estimators, and hence secure reliable estimates.

This paper focuses on discussion of the Hierarchical Bayes (HB) estimator using multi-level models, and the composite estimator that takes the weighted average of the direct estimator drawn from the Korean National Statistical Office and the synthetic estimator designed under the Korean EAPS system. The general multi-level model framework for small area estimation has been suggested in Moura and Holt (1999), and the HB estimation method using this multi-level model has been applied in more detail in You and Rao (2000). We use here the HB estimation method as in You and Rao (2000). Detailed accounts of synthetic and composite estimation are given by Ghosh and Rao (1994), Singh, Gambino and Mantel (1994) and

[1] Yeon Soo Chung, Department of Computer Science and Statistics, Korea Air Force Academy, Chungwon, Korea, e-mail: yschung@afa.ac.kr; Kay-O Lee, Gallup Koyed and Chungbuk National University, Seoul, Korea, e-mail: kolee@gallup.co.kr; Byung Chun Kim, Graduate School of Management, KAIST, Seoul, Korea, e-mail: bckim@kaist.ac.kr.

Marker (1999). Other references can be found in P.D. Falorsi, S. Falorsi and Russo (1994), and Chattopadhyay, Lahiri, Larsen and Reimnitz (1999). Falorsi *et al.* (1994) produce level estimates for unplanned small area territorial domains from the Italian Labor Force Sample Survey whereas Chattopadhyay *et al.* (1999) give a composite estimation of drug prevalence for sub-state areas to improve on the traditional design-based estimators. It is noted that both studies use supplementary information from the original survey data. For example, Chattopadhyay *et al.* (1999) uses additional information that relates various groups, counties and planning regions to one another.

In order to "borrow strength", we divide the EAPS data into two homogenous sub-regional groups (Cities and Counties), and each sub-regional group is classified into four categories of sex (male, female) and age (15-34, 35 and over). The unemployment characteristics of each category in the given small area are used as supplementary information for small area estimation. We also use the census of 2000 and the Resident Registration Population of 2000 as auxiliary information to calculate the small area estimates.

The contents of this paper are as follows. The Korean EAPS is described briefly in section 2. Section 3 gives the direct estimator drawn from the Korean National Statistical Office. Section 4 introduces design-based and model-based indirect estimators. We suggest synthetic and composite estimators under the current EAPS system, and the mean square errors of these estimates are derived using the Jackknife method. For the model-based indirect estimator we apply the HB multi-level model in estimating small areas. Section 5 illustrates the methodology, studies model selection and presents results employing the EAPS data. Finally, some closing comments are made in section 6.

2. ECONOMICALLY ACTIVE POPULATION SURVEY

The Korean National Statistical Office conducts the Economically Active Population Survey (EAPS) on a monthly basis. The characteristics of the economically active (such as employment and unemployment figures) are obtained from the EAPS. The EAPS provides monthly information on the employment trend, which plays an important role in policy making and evaluation for the 7 Metropolitan Cities and 9 Provinces. The interviewees of the EAPS are persons aged 15 and over residing in sample enumeration districts. The survey is conducted during the week just after the reference period, which is the week containing the 15th day of the month. The EAPS is conducted by visiting and interviewing each household.

The sample households for the Korean EAPS are selected from the sampled population using stratified two-stage sampling. The sampled population consists of 22,000 enumeration districts that are ten percent of the 1995 census. According to the classification of major administration regions, the country is divided into 16 large areas; there are 7 Metropolitan Cities and 9 Provinces, and the population is divided into 25 strata; 7 metropolitan strata, and 18 provincial strata consisting of 9 urban strata and 9 rural strata. The number of enumeration districts, which are primary sampling units (PSUs), selected in the 25 strata is computed using a preassigned relative standard error. Then PSUs are systematically selected with a probability proportional to their measure of size within each stratum. Each sampled PSU is divided into the same number of segments as the measure of size of each PSU, each segment containing 8 households on average. Within each PSU, 3 contiguous segments, secondary sampling units (SSUs), are randomly selected, and all households in each selected segment are surveyed. The sample is self-weighting in each stratum while the sampling rates are different from stratum to stratum. The selected sample households are surveyed repeatedly for 5 years without rotating.

The planned domains of the survey design are the 16 large areas (7 Metropolitan Cities and 9 Provinces), and local self-government areas (LSGAs) within those large areas are unplanned sub-regional domains. The sample size for the current EAPS is approximately 1,200 PSUs, and 30,000 households. The purpose of this study is to estimate unemployment statistics of the LSGAs from the EAPS.

3. DIRECT ESTIMATION

The direct estimator $\hat{Y}_{i\cdot}$ representing the total unemployment figure for small area i, based on data from the EAPS, is as follows:

$$\hat{Y}_{i\cdot} = \sum_{s=1}^{2} {}_s\hat{Y}_{i\cdot} = \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_s\hat{Y}_{ih} = \sum_{s=1}^{2}\sum_{h=1}^{n_i} {}_sM_i \; {}_sY_{ih} \qquad (3.1)$$

for $i = 1, 2, ..., I$, $s = 1, 2$ and $h = 1, 2, ..., n_i$, where s is an index of sex (male or female), n_i denotes the number of sample enumeration districts for small area i from the EAPS, and ${}_sY_{ih}$ is the number of unemployed persons by sex for the hth sample enumeration district within small area i from the EAPS. The multiplier ${}_sM_i = {}_s\hat{X}_{i\cdot} / {}_sX_{i\cdot}$ is calculated under the condition that $Y_{i\cdot}$ is an approximately unbiased estimator, where ${}_s\hat{X}_{i\cdot}$ is the estimate of the resident population in small area i, and ${}_sX_{i\cdot}$ is the sample survey resident population derived from the EAPS. The variance of $\hat{Y}_{i\cdot}$ in the ith small area is estimated using a linearization – based variance estimator.

4. INDIRECT SMALL AREA ESTIMATION

4.1 Synthetic Estimation

For the ith small area belonging to a large area, the direct estimator $\hat{Y}_{i\cdot}$ does not provide adequate precision because sample sizes in specific small areas are not large enough. The synthetic estimator $\hat{Y}_{i\cdot}^{S}$ is a design-based indirect estimator that borrows strength from related areas through implicit modeling of supplementary data along with the survey data. Suppose that there are I small areas in a large area. We then divide each large area into several homogenous sub-regional groups, in which $I = \sum_{l=1}^{L} I_l$. Each sub-regional group including I_l small areas is classified into J sex-age categories. It is assumed that each small area belongs to one of several sub-regional groups and we obtain auxiliary information from the sub-regional group. The synthetic estimator has a low variance since it is based on a larger sample, but it suffers from bias should the assumption of homogenous sub-regional groups not hold.

The following notations are used: N_i, for the number of enumeration districts in small area i; n_i, for the number of sample enumeration districts allocated to the ith small area; ${}_jP_{i,2000}^{C}$, for resident population derived from the census of 2000 in cell (i, j); ${}_jP_{i,2000}^{R}$, for Resident Registration Population of 2000 in cell (i, j); ${}_jP_{i,\text{month}}^{R}$, for Resident Registration Population at survey month in cell (i, j); ${}_j\hat{X}_i$, for the direct estimate of resident population in cell (i, j); ${}_jY_{ih}$, for the number of the unemployed in the hth sample enumeration district in cell (i, j).

We consider the estimation of the total unemployed $Y_{i\cdot}$ for all units belonging to small area i. A synthetic estimator for small area i within the sub-regional group including I_1 small areas is given by

$$\hat{Y}_{i\cdot}^{S} = \sum_{j=1}^{J} \frac{{}_j\hat{P}_{i\cdot}}{{}_j\hat{X}_{\cdot}} = {}_s\hat{Y}_{\text{dir}},\ i = 1, 2, \ldots, I_1, \tag{4.1}$$

where

$${}_j\hat{P}_i = \frac{{}_jP_{i,2000}^{C}\ {}_jP_{i,\text{month}}^{R}}{{}_jP_{i,2000}^{R}},$$

$${}_j\hat{X}_{\cdot} = \sum_{i=1}^{I_1} {}_j\hat{X}_i,$$

$${}_j\hat{Y}_{\text{dir}} = \sum_{i=1}^{I_1} \sum_{h=1}^{n_i} {}_jM_i\ {}_jY_{ih},$$

in which ${}_j\hat{P}_i$ denotes the estimate of resident population obtained from administrative sources for the j^{th} sex-age category (cell) in small area i, ${}_j\hat{X}_{\cdot}$ denotes the estimate of resident population of the j^{th} sex-age category, ${}_j\hat{Y}_{\text{dir}}$ denotes the direct estimate of the total unemployed of the j^{th} sex-age category in the EAPS, and the multiplier ${}_jM_i$ is expressed by ${}_jM_i = {}_j\hat{X}_i / {}_jX_{i\cdot}$. Note that ${}_j\hat{Y}_{\text{dir}}$ represent approximately unbiased estimates of ${}_jY_{\cdot\cdot} = \sum_{i=1}^{I_1} \sum_{h=1}^{N_i} {}_jY_{ih}$.

As a measure of accuracy for the synthetic estimator $\hat{Y}_{i\cdot}^{S}$, it is customary to take

$$\text{MSE}(\hat{Y}_{i\cdot}^{S}) = \text{Var}(\hat{Y}_{i\cdot}^{S}) + \left[\text{Bias}(\hat{Y}_{i\cdot}^{S})\right]^2. \tag{4.2}$$

In (4.2), the variance of $\hat{Y}_{i\cdot}^{S}$ is readily estimated, but it is more difficult to estimate the bias of $\hat{Y}_{i\cdot}^{S}$. Under the assumption $\text{Cov}(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S}) = 0$, where $\hat{Y}_{i\cdot}$ is a direct estimator of $Y_{i\cdot}$, the estimator of MSE of $\hat{Y}_{i\cdot}^{S}$ is given by

$$\text{mse}(\hat{Y}_{i\cdot}^{S}) \approx (\hat{Y}_{i\cdot}^{S} - \hat{Y}_{i\cdot})^2 - \widehat{\text{Var}}(\hat{Y}_{i\cdot}). \tag{4.3}$$

Note that mse $(\hat{Y}_{i\cdot}^{S})$ in (4.3) is approximately an unbiased estimator, but is potentially unstable should the number of sample enumeration districts not be large enough. Another measure would be to take the average of these MSE estimators over small areas. This average MSE estimator is expected to be stable, but it is not an area-specific measure of accuracy (Ghosh and Rao 1994).

The Jackknife method is an alternative method that can provide a more accurate area-specific measure. For small area i, the estimator for the mean square error of the estimate of the total unemployed is given as follows:

$$\text{mse}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = \widehat{\text{Var}}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) + \left[\widehat{\text{Bias}}_{\text{JN}}(\hat{Y}_{i\cdot}^{S})\right]^2, \tag{4.4}$$

where

$$\widehat{\text{Var}}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = \frac{n_i - 1}{n_i} \sum_{h=1}^{n_i} \left[\hat{Y}_{i\cdot}^{S}(h) - \frac{1}{n_i} \sum_{l=1}^{n_i} \hat{Y}_{i\cdot}^{S}(l)\right]^2,$$

$$\widehat{\text{Bias}}_{\text{JN}}(\hat{Y}_{i\cdot}^{S}) = (n_i - 1)\left[\frac{1}{n_i} \sum_{h=1}^{n_i} \hat{Y}_{i\cdot}^{S}(h) - \hat{Y}_{i\cdot}^{S}\right].$$

Here, $\hat{Y}_{i\cdot}^{S}(h)$ denotes the estimate of $Y_{i\cdot}$ obtained when district h is removed from the sample.

4.2 Composite Estimation

For small area i, the direct estimator $\hat{Y}_{i\cdot}$ derived from the EAPS does not provide adequate precision because sample sizes in specific small areas are seldom large enough. Also, the synthetic estimator $\hat{Y}_{i\cdot}^{S}$ that borrows strength from related small areas may be biased. A natural way to balance the synthetic estimator $\hat{Y}_{i\cdot}^{S}$ against the instability of the direct estimator $\hat{Y}_{i\cdot}$ is to take a weighted average of the two estimators. The following composite estimator $\hat{Y}_{i\cdot}^{C}$ can be considered to gain adequate precision for small area estimates:

$$\hat{Y}_{i\cdot}^{C} = \omega_i \hat{Y}_{i\cdot} + (1 - \omega_i)\,\hat{Y}_{i\cdot}^{S},\ i = 1, 2, \ldots, I_1, \tag{4.5}$$

where ω_i is the weight having a value between 0 and 1.

Under the assumption of $\text{Cov}(\hat{Y}_{i\cdot}, \hat{Y}_{i\cdot}^{S}) = 0$, the optimal weight $\omega_{i(\text{opt})}$ that minimizes the $\text{MSE}(\hat{Y}_{i\cdot}^{C})$ with respect to ω_i can be approximated by

$$\omega_{i(\text{opt})} = \frac{\text{MSE}(\hat{Y}_{i\cdot}^{S})}{\text{MSE}(\hat{Y}_{i\cdot}^{S}) + \text{Var}(\hat{Y}_{i\cdot})}. \tag{4.6}$$

The optimal weight $\omega_{i(\text{opt})}$ in (4.6) may be estimated by substituting the Jackknife estimator $\text{mse}_{\text{JN}}(\hat{Y}_{i\cdot}^{S})$ given in

(4.4) for $\mathrm{MSE}(\hat{Y}_{i\cdot}^{S})$, and replacing $\mathrm{Var}(\hat{Y}_{i\cdot})$ by $\widehat{\mathrm{Var}}(\hat{Y}_{i\cdot})$, the linearization-based estimator typically used by the National Statistical Office of Korea. The estimated weight $\hat{\omega}_{i(\mathrm{opt})}$ is then given by

$$\hat{\omega}_{i(\mathrm{opt})} = \frac{\mathrm{mse}_{\mathrm{JN}}(\hat{Y}_{i\cdot}^{S})}{\mathrm{mse}_{\mathrm{JN}}(\hat{Y}_{i\cdot}^{S}) + \widehat{\mathrm{Var}}\,(\hat{Y}_{i\cdot})}. \tag{4.7}$$

Using the estimated weight given in (4.7), we can obtain the composite estimator of the total unemployed as follows:

$$\hat{Y}_{i\cdot}^{C} = \hat{\omega}_{i(\mathrm{opt})}\hat{Y}_{i\cdot} + (1 - \hat{\omega}_{i(\mathrm{opt})})\,\hat{Y}_{i\cdot}^{S}\ ,\ i = 1, 2, ..., I_1. \tag{4.8}$$

The Jackknife method was used to obtain area-specific measures of accuracy.

4.3 Hierarchical Bayes Estimation Using Multi-level Models

Suppose that there are I small areas. We consider the following multi-level model that integrates variations within and between the small areas in a single model:

$$Y_{ik} = x_{ik}^{T}\beta_i + e_{ik}, \beta_i = Z_i\gamma + \nu_i, i = 1, 2, ..., I;$$
$$k = 1, 2, ..., K, \tag{4.9}$$

where y_{ik} are the direct estimates associated with the kth month in the ith small area, which may be adjusted through the model (4.9) with the auxiliary variables $x_{ik} = (x_{i1k}, x_{i2k}, ..., x_{ipk})^T$ selected from the EAPS, census and administrative records; β_i is a $p \times 1$ vector of regression coefficients; Z_i is a $p \times q$ design matrix; γ is a $q \times 1$ vector of fixed coefficients; and $\nu_i = (\nu_{i1}, \nu_{i2}, ..., \nu_{ip})^T$ is a $p \times 1$ vector of random effects for the ith small area.

The ν_i's are assumed to have a joint distribution $\nu_i \overset{\text{ind}}{\sim} N_p\ (0, \Phi)$ with an unknown variance covariance matrix Φ and the e_{ik}'s are assumed to be independent random error variables with $E\ (e_{ik}) = 0$ and $\mathrm{Var}(e_{ik}) = \sigma_i^2$. ν_i and e_{ik} are also assumed to be independent.

To obtain HB estimates for each small area and posterior variances of estimates obtained from (4.9), we apply You and Rao's (2000), HB multi-level model framework as follows:

Model 1: HB model with equal error variances.

(i) $[y_{ik}\,|\,\beta_i, \sigma_e^2] \overset{\text{ind}}{\sim} N(x_{ik}^{T}\beta_i, \sigma_e^2)$,

$$i = 1, 2, ..., I; k = 1, 2, ..., K, \tag{4.10}$$

(ii) $$[\beta_i\,|\,\gamma, \Phi] \overset{\text{ind}}{\sim} N_p(Z_i\ \gamma, \Phi), \tag{4.11}$$

(iii) Marginal prior distributions are as follows: $\gamma \sim N_q(0, D), \tau_e \sim G\,(a, b)$, and $\Omega \sim W_p(\alpha, R)$, where $\tau_e = \sigma_e^{-2}, \Omega = \Phi^{-1}$ and D, a, b, α and R are known and $G(a, b)$ denotes a gamma distribution with its density given by $f(x) = [b^{a}/\Gamma(a)]x^{a-1}e^{-bx}$ $(a > 0, b > 0, x \geq 0)$. $W_p\ (\alpha, R)$ denotes a Wishart distribution.

Model 2: HB model with unequal error variances

(i) $[y_{ik}\,|\,\beta_i, \sigma_e^2] \overset{\text{ind}}{\sim} N(x_{ik}^{T}\beta_i, \sigma_e^2)$,

$$i = 1, 2, ..., I; k = 1, 2, ..., K, \tag{4.12}$$

(ii) $$[\beta_i\,|\,\gamma, \Phi] \overset{\text{ind}}{\sim} N_p(Z_i\ \gamma, \Phi), \tag{4.13}$$

(iii) Marginal prior distributions are as follows: $\gamma \sim N_q\ (0, D)$, $\tau_i \overset{\text{ind}}{\sim} G(a_i, b_i)$, and $\Omega \sim W_p\ (\alpha, R)$, where $\tau_i = \sigma_i^{-2}$, $\Omega = \Phi^{-1}$, and D, a_i, b_i, α and R are known.

We can use the Gibbs sampler to obtain the posterior estimates of $\mu_{ik} = x_{ik}^{T}\beta_i$ for the kth month in the ith small area using the posterior distribution of β_i given $y = (\{y_{ik}\}, i = 1, 2, ..., I;\ k = 1, 2, ..., K)$. Its implementation requires generating samples from full conditional posterior distributions. The necessary full conditional posterior distributions under Model 1 are given by:

For $i = 1, 2, ..., I,\ k = 1, 2, ..., K$,

(i) $$[\beta_i\,|\,y, \gamma, \Omega, \tau_e] \overset{\text{ind}}{\sim} N_p((\tau_e \textstyle\sum_k x_{ik}x_{ik}^{T} + \Omega)^{-1}$$
$$(\tau_e \textstyle\sum_k y_{ik}x_{ik} + \Omega Z_i\gamma), (\tau_e \textstyle\sum_k x_{ik}x_{ik}^{T} + \Omega)^{-1}),$$

(ii) $$[\gamma\,|\,y, \beta, \Omega, \tau_e] \sim N_p((\textstyle\sum_i Z_i^{T}\Omega Z_i + D^{-1})$$
$$(\textstyle\sum_i Z_i^{T}\Omega\beta_i), (\textstyle\sum_i Z_i^{T}\Omega Z_i + D^{-1})^{-1}),$$

(iii) $$[\Omega|y, \beta, \gamma, \tau_e] \sim W_p\left(\alpha + I, R + \frac{1}{2}\textstyle\sum_i (\beta_i - Z_i\gamma)(\beta_i - Z_i\gamma)^{T}\right),$$

(iv) $$[\tau_e|y, \beta, \gamma, \Omega] \sim G\left(a + \frac{IK}{2}, b + \frac{1}{2}\textstyle\sum_i \sum_k (y_{ik} - x_{ik}^{T}\beta_i)^2\right).$$

Using initial values $\gamma^{(0)}, \Omega^{(0)}$ and $\tau_e^{(0)}$, we can generate samples iteratively based on (i)-(iv). The M Gibbs samples $\{\beta_i^{(m)}, \gamma^{(m)}, \Omega^{(m)}, \tau_e^{(m)};\ m = 1, 2, ..., M\}$ after implementing a "burn-in" period are assumed to be iterative samples from the joint posterior distribution of β_i, γ, Ω and τ_e. The posterior estimates of β_i can be calculated using the M iterative samples $\{\beta_i^{(m)};\ m = 1, 2, ..., M\}$.

The posterior mean of μ_{ik} and posterior variance of estimates can be obtained by implementing Markov chain Monte Carlo (MCMC) integration techniques from M Gibbs samples. It should be noted that should the Gibbs samples of the parameters be produced using the WinBUGS program (Spiegelhalter, Thomas and Best 2000), the need to derive the full conditional posterior distributions for the parameters mentioned above ceases to exist. This is due to the fact that the Gibbs samples would be produced by the full conditional posterior distributions of the parameters (inherent in the process of running the program), provided that the applicable model, priors and the initial values of the parameters are given to the WinBUGS program. The full conditional distributions for Gibbs sampling under Model 2 are similar to the above Model 1.

5. DATA ANALYSIS

5.1 Description of the Data and HB Model Fitted

Before we continue, we highlight the point that direct, synthetic, composite and HB estimates were all derived using the EAPS data of December 2000. However, the HB estimates were derived using additional EAPS data of May and July 2000 for model fitting.

The large area ChoongBuk Province in Korea consists of 10 local self-government areas (LSGAs), which are small areas. The number of sample enumeration districts of the ChoongBuk Province allocated in the EAPS is 63, and the number of sample households is 1,512. Under the EAPS, the planned domains are large areas such as the ChoongBuk Province, and hence small areas such as the LSGAs fall under the category of unplanned domains. This leads to the concern that should the estimates of the total unemployed of the LSGAs be derived using only the sample enumeration districts allocated under the LSGAs, the standard errors will become large. To address this problem, we have used data of neighboring small areas with similar economic and demographic characteristics as the areas considered here as complementary information for small area estimation. We have first divided the large area of ChoongBuk Province into two sub-regional groups with similar economic and demographic characteristics. The two sub-regional groups mentioned above are Cities and Counties. We next divided each sub-regional group into four categories of sex (male, female) by age (15-34, 35, and over). The unemployment and economically active population (EAP) estimates for each of the categories of each sub-regional group were derived from the EAPS data.

Using the above estimates and the estimated resident population for each of the four categories of LSGAs produced monthly by the Korean National Statistical Office as supplementary data, we have estimated the synthetic and composite estimates for the unplanned domains (10 LSGAs) within the ChoongBuk Province based on the EAPS data of December 2000.

Let the direct estimate for the *k*th month in small area *i* be y_{ik}. The direct estimates derived from the EAPS data of May, July and December 2000 were used as dependent variates in HB multi-level models. The additional auxiliary variates for the *k*th month in small area *i* are as follows:

$$x_{ik} = (x_{i1k}, x_{i2k}, x_{i3k}, x_{i4k})^T, \; i = 1,2,\ldots,I; \; k = 1,2,3$$

$$= \left(\left(\frac{{}_1\hat{P}_i}{{}_1\hat{X}_.} \, {}_1\hat{Y}_{\text{dir}} \right)_k, \left(\frac{{}_2\hat{P}_i}{{}_2\hat{X}_.} \, {}_2\hat{Y}_{\text{dir}} \right)_k, \left(\frac{{}_3\hat{P}_i}{{}_3\hat{X}_.} \, {}_3\hat{Y}_{\text{dir}} \right)_k, \left(\frac{{}_4\hat{P}_i}{{}_4\hat{X}_.} \, {}_4\hat{Y}_{\text{dir}} \right)_k \right)^T .$$

The element of x_{ik}, $({}_j\hat{P}_i / {}_j\hat{X}_.)\, {}_j\hat{Y}_{\text{dir}}$ ($j = 1, 2, 3, 4$), denotes the estimate of the total unemployed of the j^{th} sex-age category in small area *i*, which is given in (4.1). We tried to adjust the direct estimates, y_{ik}, through the HB multi-level model with auxiliary variates, x_{ik}. The random regression coefficient vector $\beta_i = (\beta_{i1}, \beta_{i2}, \beta_{i3}, \beta_{i4})^T$ of the i^{th} small area in (4.9) was assumed to have the following structure:

$$\beta_{i1} = \gamma_{10} + \nu_{i1}; \; \beta_{i2} = \gamma_{20} + \nu_{i2}; \beta_{i3} = \gamma_{30} + \nu_{i3}; \; \beta_{i4} = \gamma_{40} + \nu_{i4},$$

where the fixed regression parameter vector $\gamma = (\gamma_{10}, \gamma_{20}, \gamma_{30}, \gamma_{40})^T$ is an unknown value, and the random effect vector $\nu_i = (\nu_{i1}, \nu_{i2}, \nu_{i3}, \nu_{i4})^T$ of the *i*th small area follows $N_4\ (0, \Phi)$.

Using the vague proper priors for γ, τ and Ω determined by setting $D = \text{diag}\ (10^4, 10^4, 10^4, 10^4), \alpha = 4, a = b = a_i = b_i = 0.001$ and R with diagonal elements of 1 and off-diagonal elements of 0.001, we generated 6,000 Gibbs samples iteratively. Using the 3,000 samples after the "burn-in" period (3,001-6,000), the posterior means of unemployed persons of the *i*th small area and the posterior variances of the estimates were calculated. The data analysis was conducted using the WinBUGS program.

5.2 Model Selection

We considered model checking and comparison using MCMC methods under the two assumed HB multi-level model frameworks. First, we examined the posterior means of standardized residuals,

$$\text{resid}_{ik} = \frac{y_{ik} - E(y_{ik})}{\sqrt{\text{Var}(y_{ik})}}, \; i = 1, 2, \ldots, 10; \; k = 1, 2, 3,$$

which are directly computable in WinBUGS. Here y_{ik} are the direct estimates obtained from the data of the EAPS, and $E(y_{ik})$ and $\text{Var}(y_{ik})$ are obtained from the predictive distribution of y_{ik}. Figure 1 and Figure 2 give their normal Q-Q plots, both revealing a high degree of agreement with normality.

To make a comparison between the assumed HB multi-level models, we calculated a negative cross-validatory log-likelihood, $-\sum_{i,k} \log f(y_{ik} | y_{(ik)})$, and a posterior mean deviance, $-2 \sum_{i,k} \log f(y_{ik} | \theta)$, for each model. The two measures are also computable using the WinBUGS program. $y_{(ik)}$ denotes all data except y_{ik} and θ represents the parameters of the predictive distribution of y_{ik}. Table 1 gives the results for the HB multi-level model checks based on a 3,000 iteration BUGS run. Model 2 yielded a negative cross-validatory log-likelihood of 121.52 and a posterior mean deviance of 243.05, both of which are smaller than the corresponding Model 1 values. For our data, Model 2 provides a better fit than Model 1.

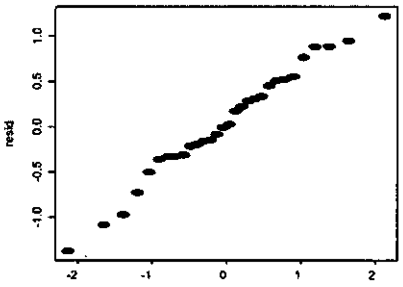


Figure 1. Normal Q-Q Plot (Model 1)

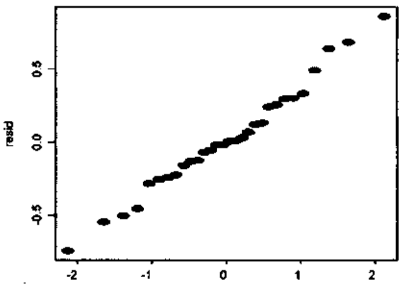


Figure 2. Normal Q-Q Plot (Model 2)

Table 1
Relative Comparison Between HB Multi-level Models

HB Model	Negative Cross-Validatory Log-likelihood	Deviance
Model 1	188.67	377.30
Model 2	121.52	243.05

In order to study how the direct estimates y_{ik} support the HB multi-level models, we employed conditional predictive ordinate (CPO) values (You and Rao 2000, page 178). The CPO values under Model 1 are calculated by

$$\widehat{\text{CPO}}_{ik}^{\text{HB}} = \frac{1}{\frac{1}{M}\sum_{m=1}^{M}\frac{1}{f(y_{ik}\,|\,\beta_i^{(m)},\sigma_e^{2\,(m)})}}$$

for $i = 1,2,...,10, k = 1,2,3,$ where $f(y_{ik}|\beta_i,\sigma_e^2)$ are the conditional normal densities given by (4.10). For model 2, the CPO values are calculated with $\sigma_i^{2(m)}$. Using the Gibbs sampler, we can calculate the CPO values for all points (see Gelfand (1995) for a more detailed discussion). Figure 3 gives a CPO comparison plot for the two assumed HB multi-level models.

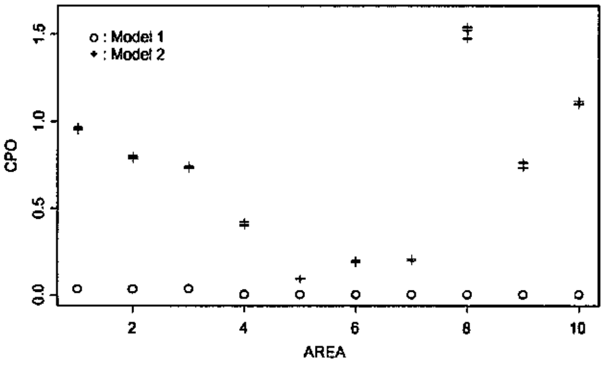


Figure 3. CPO comparison plot

Model 2 proves to be the better of the two HB multi-level models, since its CPO values are significantly larger in every small area than those for Model 1. Therefore, we conclude that Model 2 with unequal error variances is a good model for our data.

5.3 Estimation Results

Table 2 shows the estimates of the total unemployed of the 10 LSGAs within the ChoongBuk Province under the EAPS data of December 2000. The estimated standard errors of the direct and HB estimates are provided together with the Jackknife root mean square errors of the synthetic and composite estimates.

In general the direct estimates prove to be highly unstable. Studying the Jackknife root mean square errors of the estimates of the total unemployed in the LSGAs, we find that in comparison to the direct estimates, synthetic and composite estimates are more stable. Although the estimated standard errors of the HB estimates are clearly smaller than those of the direct estimates over all the LSGAs, they turn out to be highly variable in certain LSGAs such as areas 3, 4, and 5. Overall, the composite estimates are more stable than other estimates for our data.

In order to evaluate the reliability of the direct and HB estimates of each LSGA, the relative standard errors of these estimates were obtained. Similarly, the reliability of synthetic and composite estimates was evaluated by the relative bias values and the relative root mean square errors of these estimates. Denoting $\hat{Y}_{i\cdot}^{*}$ as the estimator of the total unemployed in the i^{th} small area, its relative bias (RB), relative standard error (RSE) and relative root mean square error (RRMSE) are given by the following respectively:

$$\text{RB}(\hat{Y}_{i\cdot}^{*}) = \frac{\widehat{\text{Bias}}(\hat{Y}_{i\cdot}^{*})}{\hat{Y}_{i\cdot}^{*}} \times 100,$$

$$\text{RSE}(\hat{Y}_{i\cdot}^{*}) = \frac{\sqrt{\widehat{\text{Var}}(\hat{Y}_{i\cdot}^{*})}}{\hat{Y}_{i\cdot}^{*}} \times 100,$$

$$\text{RRMSE}(\hat{Y}_{i\cdot}^{*}) = \frac{\sqrt{\text{mse}(\hat{Y}_{i\cdot}^{*})}}{\hat{Y}_{i\cdot}^{*}} \times 100.$$

Under the condition that $\hat{Y}_{i.}^*$ is an unbiased estimator, the RSE and the RRMSE of $\hat{Y}_{i.}^*$ are identical.

Table 3 shows the RB, RSE and RRMSE values of the estimates of the total unemployed of the 10 LSGAs within the ChoongBuk Province.

When comparing the bias values of synthetic and composite estimates, the average relative bias value of the composite estimates (*Av*.RB=10.26%) is somewhat smaller than that of the synthetic estimates (*Av*.RB=12.24%). However, both the synthetic and composite estimators show large values of bias in most small areas with the exception of two areas (areas 3 and 10).

We evaluate the reliability of these estimates based on the RSE (or RRMSE) values of small area estimates. It should be noted that since the direct estimates shown in Table 3 are unbiased, the RSE and RRMSE values of these direct estimates are identical. The National Statistical Office of Korea expects an approximate maximum RSE (or RRMSE) limit of 25% as the standard for reliability of small area estimates. With the exception of area 1, the RSE values of direct estimates do not satisfy this criterion for reliability. It follows that under the current EAPS system, direct estimates are unreliable. In contrast, both the RRMSE values of synthetic and composite estimates and the RSE values of the HB estimates were much smaller than the RSE(=RRMSE) values of the direct estimates in all LSGAs considered.

Table 2

Estimates of the Total Unemployed for Ten Local Self-Government Area (LSGA) in ChoongBuk (December, 2000)

Area No.	Direct		Synthetic		Composite		Hierarchical Bayes (Model 2)		n_i
	$\hat{Y}_{i.}$	Est.se	$\hat{Y}_{i.}^S$	$\sqrt{\text{mse}_{JN}}$	$\hat{Y}_{i.}^C$	$\sqrt{\text{mse}_{JN}}$	$\hat{\mu}_{ik}^{HB}$	Est.se	
1	8,517	1,733	7,969	580	8,023	493	8,514	358	22
2	3,949	1,445	2,823	725	3,050	607	3,773	474	11
3	365	390	1,830	110	1,723	101	399	152	4
4	503	373	612	234	581	196	440	106	2
5	781	676	1,164	169	1,140	158	567	261	3
6	1,275	577	1,230	282	1,238	233	1,138	270	3
7	1,032	646	1,459	295	1,384	252	1,035	117	5
8	1,795	893	1,825	346	1,821	306	1,790	69	6
9	1,023	602	2,888	574	2,000	270	970	200	5
10	512	384	872	94	851	92	511	63	2

Table 3

Relative Standard Errors (RSE) of Direct and HB Estimates for Ten Local Self-Government Areas (LSGA). Relative Bias (RB) Values and Relative Root Mean Square Errors (RRMSE) of Synthetic and Composite Estimates for Ten LSGAs (December, 2000)

Unit %

Area No.	Direct	Synthetic		Composite		Hierarchical Bayes (Model 2)
	RSE_i	RB_i	RRMSE_i	RB_i	RRMSE_i	RSE_i
1	20.35	6.92	7.27	5.99	6.15	4.20
2	36.59	23.77	25.69	18.39	19.91	12.56
3	106.91	-2.95	5.99	-2.87	5.89	37.97
4	74.15	16.26	38.30	14.37	33.73	24.00
5	86.58	-7.04	14.51	-6.67	13.84	45.94
6	45.23	17.56	22.90	14.43	18.80	23.69
7	62.56	14.86	20.25	13.29	18.21	11.28
8	49.77	15.25	18.97	13.49	16.78	3.87
9	58.83	15.01	19.88	10.20	13.50	20.65
10	74.93	-2.75	10.79	-2.82	10.79	12.29
Av.RB		12.24		10.26		
Av.RSE	61.59					19.65
Av.RRMSE			18.46		15.73	

Av.RB = average absolute relative bias over all LSGAs.
Av.RSE = average relative standard error over all LSGAs.
Av.RRMSE = average relative root mean square error over all LSGAs.

It has been noted that both composite and synthetic estimators produced reliable estimates for all the LSGAs, and also that the estimates were similar to each other. However, we stress that the composite estimator showed higher gains in efficiency against the synthetic estimator in all the LSGAs. Despite being efficient and reliable in eight of the LSGAs (areas 1, 2, 4, 6, 8, 9, and 10), the HB estimates fall below the criterion of reliability in the other two LSGAs (areas 3 and 5).

The RRMSE values of the composite estimates are on average 70.66% smaller than the RSE(=RRMSE) values of the direct estimates, with this figure ranging from 45.59% (area 2) to 94.49% (area 3). In comparing RSE values of the direct and HB estimates, HB estimates are on average 69.44% smaller than the direct estimates, with this figure ranging from 46.94% (area 5) to 92.22% (area 8). It is notable that $RSE_3 = 37.97\%$ and $RSE_5 = 45.94\%$ in HB estimation, which reflects not only that there are large variations within areas 3 and 5, but also possible variations of the estimates within each area for different months. For such areas as 3 and 5, it is suggested that additional sample enumeration districts should be allocated to reduce the standard errors of the estimates. Thus we come to the conclusion that under the current EAPS system, the composite estimator were more stable and reliable than the other estimators, and while the model-based HB estimator can be efficient in most areas, it has a major shortcoming in that it is highly variable in some areas

6. CONCLUSION

The Korean EAPS is a nation-wide sample survey, and the only official source producing monthly employment and unemployment figures. The monthly-published data includes the unemployment rate, employment rate, the economically active rate and also the demographic characteristics of the productive population. However, the EAPS design focuses on figures for large areas such as Metropolitan Cities and Provincial levels, and hence is a less than suitable source on its own for obtaining unemployment figures of unplanned sub-regional domains such as the LSGAs, especially since these areas are increasingly attracting interest. We have suggested the design-based indirect estimators (synthetic and composite estimators) and HB multi-level model estimators for deriving unemployment figures for the LSGAs within large areas, using only the EAPS data and the official figures of the Korean National Statistical Office (supplementary administrative information). The Jackknife mean square errors of the synthetic and composite estimates were introduced as measures of accuracy for the small area estimates. The posterior variances of the HB estimates were also used as measures of precision for the small area estimates.

The results using the EAPS data show that the small area estimators (synthetic, composite and HB multi-level model estimators) were much more effective in comparison to results obtained using the direct estimator, and moreover most of these estimates had significantly lower standard errors (or root mean square errors) than that of the direct estimates. In terms of gains in efficiency, the composite estimator performed much better than other estimators.

The Korean EAPS is conducted every month, in addition to which an overall review and redesign of the survey is carried out every five years. In constructing a new survey, a general review of population stratification, sample allocation and clustering is being considered so that the reliability of small area level estimates can be strengthened. Studies to estimate other relevant domains such as sex, age and education in addition to the existing sub-regional domains within large areas are under consideration, based on the new survey design.

REFERENCES

CHATTOPADHYAY, M., LAHIRI, P., LARSEN, M. and REIMNITZ, J. (1999). Composite estimation of drug prevalences for sub-state areas. *Survey Methodology*. 25, 81-86.

FALORSI, P.D., FALORSI, S. and RUSSO, A. (1994). Empirical comparison of small area estimation methods for the Italian Labour Force Survey. *Survey Methodology*. 20, 171-176.

GELFAND, A.E. (1995). Model determination using sampling-based methods. In *Markov Chain Monte Carlo in Practice* (Eds. W.R. Gilks, S. Richardson and D.J. Spiegelhalter). London: Chapman and Hall. 145-161.

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

MARKER, D.A. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*. 15, 1-24.

MOURA, F., and HOLT, D. (1999). Small area estimation using multi-level models. *Survey Methodology*. 25, 73-80.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*. 20, 3-22.

SPIEGELHALTER, D., THOMAS, A. and BEST, N. (2000). *WinBUGS Version 1.3 User Manual*. MRC Biostatistics.

YOU, Y., and RAO, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*. 26, 173-181.

Conditional and Unconditional Analysis of Some Small Area Estimators in Complex Sampling

LOREDANA DI CONSIGLIO, PIERO DEMETRIO FALORSI, STEFANO FALORSI and ALDO RUSSO[1]

ABSTRACT

This work deals with the unconditional and conditional properties of some well known small area estimators: expansion, post-stratified ratio, synthetic, composite, sample size dependent and the empirical best linear unbiased predictor. As it is commonly used in household surveys conducted by the National Statistics Institute of Italy, a two-stage sampling design is considered. An evaluation is carried out through a simulation based on 1991 Italian Census data. The small areas considered are the Local Labour Market Areas, which are unplanned domains that cut across the boundaries of the design strata.

KEY WORDS: Relative conditional bias; Relative root conditional MSE; Conditional coverage rate.

1. INTRODUCTION

Sampling theorists prefer to plan the sampling strategy on the basis of the unconditional sample space U_u, *i.e.*, the set of all possible samples (*unconditional approach*). However, after data collection, the reliability of an estimate obtained by means of an estimator $\tilde{Y}$, can be evaluated either unconditionally or conditionally; *i.e.*, the evaluation can be assessed on the conditional sample space U_C (*conditional approach*), where U_C is the set of samples with some specific properties.

The use of conditional arguments in sampling has been studied by Holt and Smith (1979) and Royall and Cumberland (1985). The use of the conditional approach for small area estimation has been studied by Rao (1985) and Särndal and Hidiroglou (1989). These papers consider the case of simple random sampling. In the context of small area estimation, the conditional and unconditional properties of some estimators for a two-stage sampling design with stratification of the primary sampling units have been studied in Russo and Falorsi (1993), Russo and Falorsi (1996), Falorsi and Russo (1999) and Falorsi, Falorsi and Russo (2000).

This paper considers a two-stage sampling design with stratification of the Primary Sampling Units (PSUs). This kind of design is generally used in household surveys conducted by the National Statistics Institute, *e.g.*, the Labour Force Survey (LFS). The aim of this work is to evaluate, on the basis of a simulation study, the conditional and unconditional properties of some important small area estimators.

The principal aspects of our investigation are:

- the simulation study is based on a sample design with strata, cluster delineation and sample size similar to those used in the LFS;
- the small areas considered are the Local Labour Market Areas (LLMAs), which are unplanned domains that cut across the boundaries of the design strata;
- the conditional analysis is developed using a sample space U_C, as reference set, consisting of all the possible samples containing a fixed number of PSUs belonging to the LLMA;
- the estimators examined are expansion, post-stratified ratio, synthetic, composite, sample size dependent and empirical best linear unbiased predictor. For a review see Ghosh and Rao (1994), Singh, Gambino and Mantel (1994), Pfeffermann (1999) and Rao (1999).

In section 2 the sampling design, the parameters of interest and the current estimator used by the LFS are described. Section 3 illustrates the small area estimators examined in the present work. In section 4 the empirical results of the simulation study are shown. Section 5 contains a short summary with suggestions for extension of the analysis.

2. DESCRIPTION OF THE LFS SAMPLING STRATEGY

2.1 Sample Design

The LFS is a quarterly sample of about 72,000 households designed to produce estimates of the labour force status of the population at national and regional levels. The survey in each quarter is based on a composite design. Within a given province (administrative area inside the region) the municipalities are divided into two area

[1] Loredana Di Consiglio, Piero Demetrio Falorsi and Stefano Falorsi, Istituto Nazionale di Statistica, Via Cesare Balbo, 16 - 00184 Roma, ITALY; Aldo Russo Università di Roma TRE Via C. Segre, 2-00142 Roma, ITALY.

types: the Self-Representing Area (SRA) – consisting of the larger municipalities – and the Non Self-Representing Area (NSRA) – consisting of the smaller ones.

In the SRA a stratified cluster sampling design is applied. Each municipality is a single stratum and the PSUs are the households selected by means of systematic sampling. All members of each sampled household are interviewed.

In the NSRA the sample is based on a stratified two-stage sample design. The PSUs are the municipalities, while the Secondary Sampling Units (SSUs) are the households. The PSUs are divided into strata of the same magnitude in terms of population size. Two sample PSUs are selected from each stratum without replacement and with probability proportional to the PSU's population size. The SSUs are selected by means of systematic sampling in each PSU. All members of each sample household are interviewed.

2.2 Notation and Parameter of Interest

For simplicity's sake we will introduce notation only for the two-stage sampling design of the NSRA. Note that the derivation of the quantities and expressions for the SRA case is a special case of NSRA.

With reference to the generic geographical region we introduce the following subscripts: $p\,(p = 1, ..., L)$ for province; $h\,(h = 1, ..., H_p)$ for stratum; i for municipality; j for household; $a\,(a = 1, ..., A)$ for age-sex group. A quantity associated to stratum h, municipality i, and household j will be briefly referred to as a quantity in hij; a quantity associated to stratum h and municipality i will be referred to as a quantity in hi. The following notation is also used: N_h for the number of municipalities in h; P_h for the number of persons in h; n_h for the number of sample municipalities in h; M_{hi} for the number of households in hi; P_{hi} for the number of persons in hi; m_{hi} for the number of sample households in hi; P_{ahij} for the number of persons in group a belonging to hij and P_{hij} for the number of persons in hij.

Further let

$$Y = \sum_{a=1}^{A} \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

be the total of the characteristic y for the regional population, where Y_{ahij} denotes the total of the characteristic of interest y for the P_{ahij} persons in group a in household hij.

2.3 Estimator of Y

An estimate of total Y is obtained by means of a post-stratified ratio estimator expressed by

$$\hat{Y}^R = \sum_{a=1}^{A} \frac{\hat{Y}_a^E}{\hat{P}_a^E} P_a \tag{1}$$

where

$$\hat{Y}_a^E = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ahij}$$

and (2)

$$\hat{P}_a^E = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} K_{hij} P_{ahij}$$

represent unbiased estimators of

$$Y_a = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

and

$$P_a = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} P_{ahij}.$$

The symbol K_{hij}, that denotes the *basic weight*, is expressed by (Cochran 1977)

$$K_{hij} = \frac{P_h}{n_h P_{hi}} \frac{M_{hi}}{m_{hi}}.$$

Note that for the SRA

$$n_h = 1 \text{ and } P_{hi} = P_h, \text{ so } K_{hij} = \frac{M_{hi}}{m_{hi}}.$$

3. SMALL AREA ESTIMATORS

We now consider the problem of estimating the total of a y variable for units belonging to a small area. Let $d\,(d = 1, ..., D)$ be the generic small area of a given geographical region. Since the LLMAs may cut across provinces, the total of interest in small area d is defined by

$$Y_d = \sum_{a=1}^{A} Y_{da} \tag{3}$$

with

$$Y_{da} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} Y_{ahij}$$

where L_d denotes the provinces including part of the small area d, H_{dp} are the strata of province p intersecting the small area d and N_{dh} denotes the municipalities of stratum h belonging to small area d.

The choice of an estimation method basically depends on available information. In Italy the accessible information at the small area level is currently very poor: only total persons in age-sex groups can be obtained at the municipality level; this is why all the small area estimators considered here will be based on this information only. In the simulation work we have considered the following *direct estimators*:

(i) the *expansion estimator*

$$\hat{Y}_d^E = \sum_{a=1}^{A} \hat{Y}_{da}^E \tag{4}$$

where

$$\hat{Y}_{da}^E = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ahij}$$

is the expansion estimator of Y_{da} and n_{dh} is the number of sampled municipalities of stratum h belonging to LLMA d;

(ii)the *post-stratified ratio estimator*

$$\hat{Y}_d^R = \sum_{a=1}^{A} \frac{\hat{Y}_{da}^E}{\hat{P}_{da}^E} P_{da} \tag{5}$$

in which

$$\hat{Y}_{da}^E = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij} Y_{ahij}\ ,\ \hat{P}_{da}^E = \sum_{l=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{n_{dh}} \sum_{j=1}^{m_{hi}} K_{hij},$$

$$P_{da} = \sum_{p=1}^{L_d} \sum_{h=1}^{H_{dp}} \sum_{i=1}^{N_{dh}} \sum_{j=1}^{M_{hi}} P_{ahij}.$$

In the simulation work reported here we have considered the following *design-based indirect estimators*:

(iii) the *synthetic estimator*

$$\hat{Y}_d^S = \sum_{a=1}^{A} \frac{\hat{Y}_a^E}{\hat{P}_a^E} P_{da} \tag{6}$$

in which $\hat{Y}_a^E$ and $\hat{P}_a^E$ are expressed by formulas (2). The estimator (6) is based on the underlying assumption that, for each post-stratum a, the small area mean equals the mean at the regional level;

(iv) the *composite estimator*, considered in two alternative forms

$$\hat{Y}_d^{C1} = \alpha_d \hat{Y}_d^R + (1 - \alpha_d) \hat{Y}_d^S \tag{7}$$

$$\hat{Y}_d^{C2} = \alpha \hat{Y}_d^R + (1 - \alpha) \hat{Y}_d^S \tag{8}$$

where $\alpha_d (0 \leq \alpha_d \leq 1)$ is a specific small area weight while $\alpha (0 \leq \alpha \leq 1)$ is a common weight for all the LLMAs of the region. The methods used to calculate weights α_d and α will be described in subsection 4.1. Both of the composite estimators equal by definition the synthetic estimator when the sample size in the small area equals zero;

(v) the *sample size dependent estimator* (SSD), expressed by

$$\hat{Y}_d^{\mathrm{SD}} = w_d \hat{Y}_d^R + (1 - w_d) \hat{Y}_d^S \tag{9}$$

where

$$w_d = \begin{cases} 1 & \text{if } \hat{P}_d^E \geq \lambda P_d \\ \hat{P}_d^E / (\lambda P_d) & \text{otherwise} \end{cases}$$

where λ is a given constant, $\hat{P}_d^E = \sum_{a=1}^{A} \hat{P}_{da}^E$ and $P_d = \sum_{a=1}^{A} P_{da}$.

The estimator (9) is based on the result that the performance of the post-stratified ratio estimator depends on the proportion of the sample falling in the small area. If the proportion of the sample within the small area is reasonably large then the estimator (9) equals the post-stratified ratio estimator. Otherwise it becomes a composite estimator with increasing weight $(1 - w_d)$ on the synthetic estimator, as the size of the sample in the small area decreases.

Finally, in the framework of *model-based indirect predictors*, we consider:

(vi) the *empirical best linear unbiased predictor (EBLUP)*

$$\hat{Y}_d^{EP} = \gamma_d \hat{Y}_d^R + (1 - \gamma_d) x_d' \tilde{\beta} \tag{10}$$

where

$$\tilde{\beta} = \left[\sum_{d=1}^{D} x_d x_d' / (\tilde{\sigma}_v^2 + \psi_d) \right]^{-1} \left[\sum_{d=1}^{D} x_d \hat{Y}_d^R / (\tilde{\sigma}_v^2 + \psi_d) \right],$$

$$\gamma_d = \tilde{\sigma}_v^2 / (\tilde{\sigma}_v^2 + \psi_d) \tag{11}$$

that is based on the well-known area level linear mixed model of Fay and Herriot (1979):

$$\hat{Y}_d^R = x_d' \beta + v_d + e_d \tag{12}$$

in which: β is the vector of regression parameters, x_d is a vector of area-specific auxiliary data, v_d are uncorrelated random area effects with mean zero and variance σ_v^2, e_d are independent sampling errors with mean zero and known variance ψ_d, $\tilde{\beta}$ is the weighted least squares estimator of β with weights $(\sigma_v^2 + \psi_d)^{-1}$ and $\tilde{\sigma}_v^2$ is suitable estimator of σ_v^2. In this work we utilise an asymptotically consistent estimator of σ_v^2 that can be obtained iteratively by alternating weighted least squares estimation for β with the solution of

$$\frac{\sum_{d=1}^{D} (\hat{Y}_d^R - x_d' \beta)^2}{\sigma_v^2 + \psi_d} = D - k$$

for σ_v^2, where k is the number of elements of vector x_d, corresponding to the number of auxiliary variables in the model (12). The previous description is based on the assumption that the variances ψ_d are known; in practice these variances are seldom known. In the present study we have considered two different methods (see subsection 4.1) for evaluating sampling variances. From these two methods we obtain two alternative empirical best linear unbiased predictors, $\hat{Y}_d^{EP1}$ and $\hat{Y}_d^{EP2}$.

4. EMPIRICAL STUDY

4.1 Simulation of the LFS Sample Design

In order to illustrate the conditional and unconditional properties of the estimators discussed in the preceding section, we carried out a simulation study involving repeated draws of a sample design with strata and cluster delineation and sample size similar to those used in LFS. The study can be summarised as follows:

- the information referring to the auxiliary variables and the totals of interest $Y_d (d = 1, ..., D)$ are taken from the 1991 General Population Census of Italy;
- the variables of interest are Employed, Unemployed and persons searching for their first job;
- the auxiliary variables for the post-stratification of the members of the sampling households are sex and age;
- the small areas of interest are the 27 LLMAs of the Lazio region;
- for the Monte Carlo simulation R = 2,000 two- stage LFS samples were selected for each one of the five provinces of the Lazio region;
- the number of sex-age classes considered in the construction of the synthetic estimators equals 28; the age groups are 0-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, more than 74;
- the SSD estimator has been evaluated with different values for the parameter λ ($\lambda = 2/3, \lambda = 1.5$ and $\lambda = 2$); the best performance in terms of mean square error has been obtained for $\lambda = 2$, so in this work only the results for SSD with $\lambda = 2$ are reported;
- for the empirical best linear unbiased predictors $\hat{Y}_d^{EP1}$ and $\hat{Y}_d^{EP2}$ we have removed from the analysis the LLMA of Rome. In fact the LLMA of Rome is very big in terms of population and we have verified that it has too much influence in the model. The model has been fitted separately for two groups of small areas (see section 5.1 for the definition of groups). The following covariates have been chosen:

  1) in the model for Employed and Unemployed, the province (administrative area contained in region) and the number of persons in age groups 14-35 and 35-65 by sex;
  2) in the model for persons searching for their first job, the province and the number of persons in age groups 14-25 and 25-35 by sex.

  The reduction of the number of classes with respect to the synthetic case was necessary because the number of small areas in this study is not large enough;
- the weights of composite estimator $\hat{Y}_d^{C1}$ correspond to the optimal weights given by the ratio of the MSE of the synthetic estimator over the sum of the variance of the direct estimator and the MSE of synthetic estimator (Schaible 1978). These quantities were actually evaluated on the 1991 census data;
- the unique regional weight of composite estimator $\hat{Y}_d^{C2}$ is the estimated optimal one for the average MSE of the composite estimators of all areas (Purcell and Kish 1979) given by

$$\alpha = 1 - \frac{\sum_{d=1}^{D} \widehat{\text{var}}\left(\hat{Y}_d^R\right)}{\sum_{d=1}^{D} \left(\hat{Y}_d^S - \hat{Y}_d^R\right)^2}.$$

  The resulting estimator is sample dependent. We have not pursued this method for small area specific weights due to the high variability of each area MSE and variance estimation. A smoothed model has been used to improve the stability of the evaluation of variances: the variance for the SRAs is obtained applying standard formulas for variance estimation on the linearized variables. For the NSRAs the variance is obtained applying a common design effect evaluated at the regional level to the simple random sampling variance estimate;
- in the predictor $\hat{Y}_d^{EP1}$, the sampling error variance ψ_d has been evaluated using census data; for predictor $\hat{Y}_d^{EP2}$ we have considered the alternative case in which ψ_d has to be evaluated through sample data: a regression model based on twelve simulated LFS samples was fitted and then the value of ψ_d predicted through the model.

4.2 Performance Measures

4.2.1 Overall Unconditional Measures

The following unconditional performance measures were calculated to assess the bias and the MSE of the estimators over the 2,000 replications and over all the D small areas:

- Percentage Average Absolute Relative Bias (AARB);
- Percentage Average Relative Root Mean Square Error (ARRMSE), expressed respectively by formulas

$$\text{AARB}(\hat{Y}^T) = \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{R}\sum_{r=1}^{R}\left[\frac{\hat{Y}_d^T(r) - Y_d}{Y_d}\right]\right| 100$$

$$\text{ARRMSE}(\hat{Y}^T) = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\frac{1}{R}\left(\sum_{r=1}^{R}\left[\frac{\hat{Y}_d^T(r) - Y_d}{Y_d}\right]^2\right)}\,100$$

in which $\hat{Y}_d^T(r)$ indicates the value of the generic small area estimator T (described in section 3) obtained in the r-th of the R=2,000 samples.

The same measures were also considered averaging only on subsets of small areas, with D replaced by the cardinality of the subset. For the definition of the subsets see section 5.1.

4.2.2 Conditional Measures

For each small area d, the 2,000 repeated samples were distributed over the different values of the realised number n_d of sampled municipalities belonging to small area d. For each value of n_d and for each small area d, the conditional performance measures were computed over that subset of the 2,000 samples for which the small area sample PSU count was exactly n_d.

The following conditional performance measures were considered:

- Percentage Relative Conditional Bias (RCB);
- Percentage Relative Root Conditional MSE (RCMSE);
- Conditional Coverage Rate (CCR).

These measures were calculated in the following way:

$$\text{RCB}(\hat{Y}_d^T) = \frac{1}{R_d}\sum_{r=1}^{R_d}\left[\frac{\hat{Y}_d^T(r) - Y_d}{Y_d}\right]100$$

$$\text{RRCMSE}(\hat{Y}_d^T) = \sqrt{\frac{1}{R_d}\sum_{r=1}^{R_d}\left[\frac{\hat{Y}_d^T(r) - Y_d}{Y_d}\right]^2}\,100$$

$$\text{CCR}(\hat{Y}_d^T) = \left(\frac{1}{R_d}\sum_{r=1}^{R_d} I(r)\right)100$$

in which R_d indicates the number of samples for which the PSU sample count in the small area d equals the fixed number n_d; $I(r) = 1$ if the r-th confidence interval based on $\hat{Y}_d^T(r)$ contains the true value Y_d and $I(r) = 0$ otherwise. The nominal value equals 95% and the confidence interval is the normal confidence interval where we have used as evaluation of variance the value resulting from the 2,000 replications.

5. ANALYSIS OF THE RESULTS

5.1 Unconditional Analysis

The LLMAs analysed in the simulation with their characteristics in terms of population, number of municipalities and number of LFS strata intersected are reported in Table 1. The small areas have been grouped on the basis of the ranking of the proportion of LLMA's population over the total regional population. The percent proportion of the first group ranges from 0.12% to 1.73%; the group is composed of 19 LLMAs. The percent proportion of the second group ranges from 1.9% to 5.05%; the group is composed of 7 LLMAs. The third group consists of the largest LLMA representing a percent proportion equal to 64%. The LLMAs are divided into these three groups because we expect the MSE to be larger for those LLMAs with smaller sample size.

Table 1
Local Labour Market Area (LLMA), Population, Percent Population, Number of Municipalities and Number of LFS Strata Intersected by the LLMA

LLMA	Population	Population%	Number Municipalities	Number Strata
398	6,005	0.12	5	2
396	7,364	0.14	3	2
391	8,901	0.17	4	2
407	11,392	0.22	2	2
393	12,500	0.24	3	3
414	12,656	0.25	4	2
406	13,051	0.25	3	2
395	16,012	0.31	5	3
390	19,823	0.39	8	3
411	23,226	0.45	5	2
394	30,193	0.59	6	3
408	45,274	0.88	5	2
392	51,789	1.01	13	5
416	59,512	1.16	10	4
402	71,906	1.40	15	5
401	72,080	1.40	34	8
400	72,235	1.41	4	3
412	78,249	1.52	5	3
409	88,984	1.73	7	4
399	97,680	1.90	42	5
405	114,361	2.23	3	2
397	133,303	2.60	18	5
413	146,133	2.85	41	5
410	170,945	3.33	6	4
404	198,010	3.86	16	8
415	259,382	5.05	35	7
403	3,314,237	64.54	65	13

In Table 2 we present the values of the unconditional performance measures AARB and ARRMSE for one of the three LFS characteristics studied: the number of Unemployed. This variable has been chosen since it is one of the most important characteristic produced by the LFS.

Table 2

Percentage Average Absolute Relative Bias and Percentage Average Root Relative Mean Square Error of the Estimators of Unemployed

Estimator	AARB	ARRMSE
Expansion	2.67	96.07
Post stratified ratio	26.20	58.29
Synthetic	18.10	19.40
Composite C1	15.52	17.34
Composite C2	8.94	31.48
SSD	10.14	29.84
EBLUP EB1*	13.36	66.57
EBLUP EB2*	12.98	74.88

*The averages for the EBLUPs do not include LLMA=403

Table 3 reports the same measures for each of the three previously defined groups of LLMAs.

From the analysis of the results in Tables 2 and 3, the following conclusions emerge:

- with the exclusion of the direct estimator, the bias of composite estimator $\hat{Y}^{C2}$ is almost always the smallest, or among the smaller ones, and it is very close to the bias of the SSD estimator;
- composite estimator $\hat{Y}^{C1}$ is almost always the best in terms of ARRMSE; its performance is similar to that of the synthetic estimator when taking account of the overall measure. This is due to the fact that the optimal weights are close to zero on many of the small areas considered in the simulation (note that many small areas have a percentage population under 2%). This can be confirmed by examining the results for Group 1 where the similarity of the two estimators is evident;
- the overall bias of the post-stratified ratio estimator is very high; this can be explained by the very high bias of the estimator for the areas belonging to Group 1, where the typical sample size is small;
- the model used for the empirical best linear unbiased predictors does not seem adequate, likely because we are far from the hypothesis of unbiasedness for the direct component (post-stratified ratio estimator) and due to the choice of the auxiliary variables; this is true in particular for the variable unemployment reported in Tables 3 and 4; it is important to note that these predictors have not been considered for Group 3 since this group includes only LLMA = 403 (Rome);
- comparing the SSD estimator and the composite estimator $\hat{Y}^{C2}$, both combining a direct component with a synthetic component with sample weights, the SSD estimator seems preferable since the performance of the two estimators is very close but SSD is superior in terms of computational simplicity. Since in actual surveys the optimal weights are not known, the present analysis suggests using the SSD estimator; a drawback is that a specific study has to be carried out for the choice of the parameter λ.

Table 3

Percentage Average Absolute Relative Bias and Percentage Average Root Relative Means Square Error of the Estimators of Unemployed by Group of Local Labour Market Areas

Estimator	AARB	ARRMSE	AARB	ARRMSE	AARB	ARRMSE
	Group 1		Group 2		Group 3	
Expansion	3.52	123.30	0.71	35.01	0.11	6.19
Post-stratified ratio	36.94	72.07	0.77	28.43	0.08	5.68
Synthetic	17.06	18.24	22.68	24.28	5.84	7.33
Composite C1	16.52	17.85	14.71	17.66	2.19	5.50
Composite C2	9.95	35.59	6.89	23.86	3.98	6.68
SSD	10.11	34.77	11.27	19.89	2.99	5.70
EBLUP EB1	13.84	80.14	12.06	29.75	*	*
EBLUP EB2	14.44	91.89	9.02	28.74	*	*

5.2 Conditional Analysis

For the conditional measures we limit ourselves to the presentation of the results for the following four LLMAs: Bagnoregio (code number = 391) and Civita Castellana (code number = 392) in the small group, Cassino (code number = 413) in the medium group, and Rome (code number = 403) for the large group. The frequency distributions over the 2,000 replications of the PSUs' counts in each selected area are very different as a consequence of the LLMAs' different sizes.

Recall that we could not consider EBLUPs for LLMA 403 since it is the only one in GROUP 3.

In Table 4 the results of the study areas are reported for the variable number of Unemployed.

The following points arise:

- the post-stratified ratio estimator usually has conditional bias near zero when the sample size, n_d, takes an inner value of its frequency distribution;
- the post-stratified ratio estimator usually shows better conditional performance, in terms of conditional bias and of RRCMSE, than the expansion estimator;

Table 4
Percentage Relative Conditional Bias and Percentage Relative Root Conditional MSE of the Estimators Conditioned on the Number of Sampled Municipalities for given LLMAs

Number of sampled Municipalities	Proportion of simulations %	Expansion	Post stratified Ratio	Synthetic	Composite C1	Composite C2	Sample Size Dependent	EBLUP EB1	EBLUP EB2
LLMA= 391									
Percentage Relative Conditional Bias									
0	72.43	-100.00	-100.00	28.69	28.69	28.69	28.69	26.76	39.68
1	25.29	208.30	-4.28	28.39	28.21	7.28	-4.28	-2.35	-4.21
2	2.24	527.40	0.66	29.81	29.65	9.22	0.66	1.21	0.97
3	0.05	637.88	-16.53	24.68	24.45	1.41	-16.53	85.54	-12.22
LLMA=391									
Percentage Relative Root Conditional MSE									
0	72.43	100.00	100.00	29.33	29.33	29.33	29.33	141.08	163.51
1	25.29	281.18	68.54	29.03	28.85	48.20	68.54	70.44	66.58
2	2.24	588.02	45.51	30.23	30.07	33.49	45.51	84.47	47.33
3	0.05	637.88	16.53	24.68	24.45	1.41	16.53	85.54	12.22
LLMA=392									
Percentage Relative Conditional Bias									
0	8.79	-100.00	-100.00	10.26	10.26	10.26	10.26	-6.80	-7.45
1	27.32	-48.19	1.51	9.94	9.85	5.31	8.19	-5.31	0.80
2	34.03	-2.01	-3.07	10.30	10.15	1.71	3.85	-5.88	-4.18
3	20.57	43.54	-2.95	10.22	10.08	1.01	0.50	-3.84	-3.64
4	7.65	108.22	4.05	10.88	10.81	6.58	4.34	-6.07	0.98
5	1.39	159.44	3.80	13.33	13.22	6.04	3.80	-6.21	1.06
6	0.25	169.30	-13.82	10.14	9.87	-5.39	-13.82	-14.92	-13.08
LLMA=392									
Percentage Relative Root Conditional MSE									
0	8.79	100.00	100.00	11.47	11.47	11.47	11.47	40.38	43.91
1	27.32	60.11	74.67	11.24	11.19	58.36	21.25	33.99	65.82
2	34.03	48.50	48.03	11.50	11.37	34.87	24.39	28.57	41.19
3	20.57	70.07	38.01	11.54	11.41	27.09	27.52	28.41	32.86
4	7.65	129.85	35.12	11.92	11.87	24.80	33.96	29.47	30.96
5	1.39	171.38	26.29	14.09	13.97	18.80	26.29	26.84	24.56
6	0.25	173.01	20.23	11.07	10.84	15.04	20.23	19.06	17.92
LLMA=413									
Percentage Relative Conditional Bias									
0	0.05	-100.00	-100.00	2.47	2.47	2.47	2.47	-100.00	-100.00
1	1.29	-74.42	8.04	5.60	5.63	8.36	6.08	-9.08	4.88
2	7.40	-49.73	0.92	4.56	4.52	2.72	3.68	-16.72	-2.08
3	21.31	-26.46	0.93	5.06	5.01	2.37	3.55	-15.33	-1.86
4	28.96	-4.60	-1.01	5.11	5.04	1.14	2.26	-17.29	-3.83
5	25.48	19.41	-0.31	4.92	4.86	1.72	1.78	-16.93	-3.18
6	11.43	42.48	0.14	4.64	4.58	1.91	1.45	-16.70	-2.81
7	3.68	66.82	0.86	5.04	4.99	1.77	1.58	-15.11	-1.91
8	0.40	59.75	-14.54	4.74	4.51	-7.72	-13.37	-28.08	-17.24
LLMA=413									
Percentage Relative Root Conditional MSE									
0	0.05	100.00	100.00	2.47	2.47	2.47	2.47	100.00	100.00
1	1.29	76.71	77.02	8.00	8.14	66.06	13.49	72.44	75.79
2	7.40	54.07	46.04	6.69	6.69	36.83	12.75	46.44	45.42
3	21.31	36.86	36.07	7.11	7.09	27.54	14.15	39.35	35.63
4	28.96	32.02	32.26	7.28	7.24	23.93	16.22	36.92	32.12
5	25.48	38.45	27.51	6.97	6.94	20.05	16.99	33.58	27.43
6	11.43	53.61	22.02	6.52	6.47	16.06	15.95	29.75	21.94
7	3.68	77.79	24.58	6.83	6.79	17.86	20.53	28.44	23.81
8	0.40	65.42	18.76	8.19	7.96	12.71	17.42	34.65	21.61
LLMA=403									
Percentage Relative Conditional Bias									
8	0.15	-5.20	3.17	-3.96	0.56	-1.82	-0.71	*	*
9	0.20	-2.87	3.38	-2.10	1.37	-0.67	0.43	*	*
10	1.59	-4.66	-0.15	-5.82	-2.23	-3.45	-3.15	*	*
11	4.82	-2.98	0.36	-6.13	-2.02	-3.53	-3.04	*	*
12	11.38	-2.41	-0.03	-5.98	-2.21	-3.91	-3.11	*	*
13	20.32	-1.52	-0.30	-6.15	-2.44	-4.16	-3.30	*	*
14	23.40	-0.15	-0.10	-5.84	-2.20	-4.05	-3.00	*	*
15	18.68	1.01	-0.07	-5.51	-2.06	-3.93	-2.79	*	*
16	12.67	2.51	0.20	-5.64	-1.94	-3.85	-2.69	*	*
17	4.42	3.73	0.25	-5.49	-1.85	-3.66	-2.55	*	*
18	1.84	1.86	-2.55	-7.20	-4.25	-6.32	-4.80	*	*
19	0.55	6.28	0.71	-4.70	-1.28	-3.24	-1.88	*	*

Table 4 (continued)
Percentage Relative Conditional Bias and Percentage Relative Root Conditional MSE of the Estimators Conditioned on the Number of Sampled Municipalities for given LLMAs

Number of sampled Municipalities	Proportion of simulations %	Expansion	Post stratified Ratio	Synthetic	Composite C1	Composite C2	Sample Size Dependent	EBLUP EB1	EBLUP EB2
				LLMA=403					
				Percentage Relative Root Conditional MSE					
8	0.15	6.79	5.24	5.52	4.05	4.46	4.04	*	*
9	0.20	5.64	6.06	4.79	4.81	4.69	4.49	*	*
10	1.59	7.81	6.19	6.86	5.49	6.26	5.54	*	*
11	4.82	6.54	5.75	7.51	5.41	6.54	5.66	*	*
12	11.38	6.14	5.56	7.34	5.37	6.61	5.62	*	*
13	20.32	6.34	6.01	7.72	5.86	7.12	6.09	*	*
14	23.40	5.83	5.62	7.23	5.43	6.58	5.63	*	*
15	18.68	5.98	5.58	7.10	5.42	6.51	5.59	*	*
16	12.67	6.02	5.20	7.10	5.07	6.31	5.28	*	*
17	4.42	7.33	5.82	7.16	5.53	6.72	5.66	*	*
18	1.84	6.40	6.38	8.76	6.90	8.56	7.16	*	*
19	0.55	8.38	5.42	6.53	5.04	5.84	5.12	*	*

- synthetic estimators and the composite estimator $\hat{Y}_d^{C1}$ show the best performances in terms of RRCMSE for LLMAs 391, 392, 413 and 403, confirming what was observed in the unconditional analysis. The only relevant exception is for LLMA 403 for the variable Employed (not reported here) where the post-stratified ratio is the best. In fact the variances of the different estimators are very low for this small area so that the bias is decisive;
- in terms of RRCMSE neither $\hat{Y}_d^{C2}$ nor SSD seems to outperform the other.

We have not reported here the results for the conditional coverage rate (CCR), but we can summarize them as follows:

- the post stratified estimator, the composite estimator $\hat{Y}_d^{C2}$ and the SSD estimator have CCR close to the nominal value apart from extremes values of the PSU counts;
- the EBLUPs' CCRs are also close to the nominal value but we suspect this is due to their high variances;
- for the LLMA = 403 and the Employed variable, the CCR of all the estimators is far from the nominal value.

5.3 Conclusions

As we have already observed, the results for the EBLUP estimators are unsatisfactory; the model used is not adequate, likely because we are far from the hypothesis of unbiasedness for the direct component (post-stratified ratio estimator) in many cases and because of the choice of the auxiliary variables. One of the main points we intend to address in future work is the improvement of the explicit models for EBLUP.

The composite estimator $\hat{Y}_d^{C1}$ turns out to be the best in terms of ARRMSE and RRCMSE. If weights are thought to be stable they may be evaluated, for example, at a Census point and $\hat{Y}_d^{C1}$ applied. If sample dependent weights are to be used, then the SSD estimator seems preferable to the composite estimator $\hat{Y}_d^{C2}$ because of its computational simplicity, even if some *ad hoc* study may be necessary for the choice of the parameter λ, since the two estimators' unconditional and conditional properties do not differ greatly. In any case, some improvements can be gained for the composite and SSD estimators through use of a better synthetic estimator, in terms of the number and the choice of post-strata, or in terms of a better choice of the auxiliary variables as observed for the EBLUP.

In this work we have examined conditional and unconditional properties of some common estimators; our interest in the future will be to examine also the empirical properties from the conditional point of view of the conditional estimators proposed in the work by Falorsi and Russo (1999).

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

FALORSI, P.D., and RUSSO, A. (1996). A conditional analysis of some small area estimators in two-stage sampling. In *Proceedings of 1996 Annual Research Conference*, Bureau of the Census, Washington. 613- 637.

FALORSI, P.D., and RUSSO, A. (1999). A conditional analysis of some small area estimators in two-stage sampling. *Journal of Official Statistics*. 15, 4, 537-550.

FALORSI, P.D., FALORSI, S. and RUSSO, A. (2000). A conditional analysis of some small area estimators in sampling with two primary units selected in each stratum. *Statistics in Transitions, Journal of the Polish Statistical Association*. 4, 4, 565-585.

FAY, R.E., and HERRIOT, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 74, 366, 269-277.

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

HOLT, D., and SMITH, T.M.F. (1979). Post-stratification. *Journal of the Royal Statistical Society*. A, 142, 33-46.

PFEFFERMANN, D. (1999). Small area estimation – big developments. In *Proceedings of the IASS Satellite Conference Small Area Estimation*, Riga 1999. 129-145.

PURCELL, N.J., and KISH, L. (1979). Estimation for small domain. *Biometrics*. 35, 365-384.

RAO, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology*. 11, 15-31.

RAO, J.N.K. (1999). Some recent advances in model based small area estimation. *Survey Methodology*. 25, 175-186.

ROYALL, R.M., and CUMBERLAND, W.G. (1985). Conditional coverage properties of finite population confidence intervals. *Journal of the American Statistical Association*. 80, 355-359.

RUSSO, A., and FALORSI, P.D. (1993). Conditional and unconditional properties of small area estimators in two-stage sampling. In *Proceedings of the International Scientific Conference of the International Association of Survey Statistician*, Warsaw, 1992. 251-270.

SÄRNDAL, C.E., and HIDIROGLOU, M. A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*. 84, 266-275.

SCHAIBLE, W.L. (1978). Choosing weights for composite estimators for small area statistics. In *Proceedings of the American Statistical Association, Survey Research Section*. 741-746.

SINGH, M.P., GAMBINO, J. and MANTEL, H. J. (1994). Issues and strategies for small area data. *Survey Methodology*. 20, 3-22.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 63-70
Statistics Canada

Evaluating the Fundamentals of a Small Domain Estimator

RACHEL HARTER, MICHAEL MACALUSO and KIRK WOLTER[1]

ABSTRACT

The Illinois Department of Employment Security is using small domain estimation techniques to estimate employment at the county/industry division level. The estimator is a standard synthetic estimator, based on the ability to match Current Employment Statistics sample data to ES202 administrative records and an assumed model relationship between the two. This paper is a case study; it reviews the steps taken to evaluate the appropriateness of the model and the difficulties encountered in linking the two data sources.

KEY WORDS: Small domain; Employment; Labor market; Generalized regression model; Auxiliary data.

1. INTRODUCTION

The Current Employment Statistics (CES) program of the U.S. Bureau of Labor Statistics (BLS) is a federal-state cooperative survey of employers used for estimating employment, women workers, production workers, production worker hours, and production worker earnings on a monthly basis. The estimates are among America's leading economic indicators. The sample was designed to support estimates at the national, state, and large metropolitan statistical area (MSA) levels. CES is roughly comparable to Statistics Canada's monthly Survey of Employment, Payroll and Hours (SEPH).

The Illinois Department of Employment Security (IDES), and similar agencies in other states across the nation, participates with the BLS in the collection, tabulation, and publication of the CES data. The state agencies have considerable customer demand for employment estimates at smaller sub-state levels than the CES sample was intended to support. In particular, IDES needs monthly employment estimates at the county/industry division level, and it formed a partnership with the National Opinion Research Center (NORC) to find a solution to this small domain estimation problem.

In a prior paper (Harter, Wolter and Macaluso 1999), we discussed some simulations done to test various small domain estimators. In this paper, we focus on the practical aspects of finding suitable auxiliary data, determining an appropriate model, merging the data sources, and monitoring the estimation process.

2. EVALUATING AUXILIARY DATA

Purcell and Kish (1980), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994) provide excellent overviews of many small domain estimators. Most small domain estimators improve on direct sample-based estimators by (1) taking advantage of known auxiliary data, and (2) assuming and fitting a model relationship between the auxiliary data and the sample data. In this section we describe the auxiliary data for Illinois' small domain estimation problem and our evaluation of the data for this purpose.

The CES has a sister federal-state cooperative program – known as the Covered Employment and Wages (or ES202) program – in which employment and wage data are collected quarterly from all employers that participate in states' unemployment insurance programs. The employment figures from the ES202 are available approximately five months following the reference quarter. The ES202 records provide the sampling frame for the CES program. Furthermore, since the ES202 data are available for essentially all employers in the sampling frame, ES202 employment figures are considered "truth" for practical purposes.

CES monthly estimates are regularly benchmarked to ES202 figures. While they are revised several times as more complete information becomes available, the first release of CES data occurs on the first Friday of the month following the reference month. Although the ES202 employment figures lag behind the initial CES estimates by several months, ES202 employment is an obvious candidate for auxiliary data in our small domain estimation project.

A good auxiliary variable should be highly correlated with the estimation variable. In this case, ES202 employment is measuring the same concept as CES employment, except for minor scope and coverage differences, such as student workers at colleges and universities. Therefore, we expect ES202 employment and CES employment to be highly correlated.

Illinois data for a matched sample of employers from 1995 and 1996 shows that, indeed, ES202 employment and CES employment are highly correlated, regardless of the

[1] Rachel Harter, National Opinion Research Center, 55 East Monroe, Suite 4800, Chicago, IL 60603; Michael Macaluso, Illinois Department of Employment Security, Economic Information and Analysis, 401 South State Street, 7 North, Chicago, IL 60605; Kirk Wolter, Interdisciplinary Research Institute for Survey Science, 218 Snedecor Hall, Ames, IA 50010.

time lag between the two. Table 1 shows simple correlation coefficients for various industries and time lags. The correlations are slightly higher for shorter lags in growing industries, such as Finance, Insurance, and Real Estate, and for 12-month lags in seasonal industries, such as Construction. Nevertheless, we conclude from these statistics that any recent period of ES202 data is likely to serve successfully as auxiliary data for CES estimation.

Table 1
Mean Correlations of CES Employment with ES202 Employment*

Industry Division	ES202 lagged 12 months from CES	Most recent March ES202 available for CES month	Average monthly ES202 for most recent available quarter to CES month
Mining	0.951	0.965	0.980
Construction	0.936	0.909	0.909
Manufacturing	0.983	0.984	0.985
Transportation & Utilities	0.978	0.981	0.982
Trade	0.979	0.979	0.979
Finance, Insurance, & Real Estate	0.982	0.985	0.987
Services	0.975	0.966	0.966
Government Ownership	0.996	0.995	0.993

* Within 2-digit Standard Industrial Classification (SIC) codes, we computed correlations for pairs of CES and ES202 months with the lagged relationships shown. We averaged the correlations across reference months and across SICs within the industry divisions shown.

We reviewed the scope and coverage differences between CES and ES202 to determine where the use of ES202 data may require special attention. The student worker example cited above was one such difference. Railroad workers do not participate in state unemployment insurance programs, so this industry is one in which ES202 data are not likely to be helpful. We reviewed the processing schedules for both CES and ES202 to help us determine which period of ES202 data would be available for estimation on the CES schedule. We reviewed the edits applied in both programs to see where differences may affect outcomes. For both of these programs, many anomalies in the data are explained through the use of comment variables containing standard coded values for various business conditions. We reviewed these comment variables to see how special cases are handled. All of these background checks were necessary to identify potential pitfalls in using ES202 data as an auxiliary variable for the small domain estimation problem.

Finally, we needed some indication that CES and ES202 data could be successfully linked for individual employers. To examine this issue, we matched and plotted CES and ES202 data. See Figures 1-3 for examples of statewide plots by 2-digit SIC (Standard Industrial Classification). The plots immediately alert us to potential matching problems in individual cases (Points considerably off the straight line signify potential matching or data problems), but assure us that most observations can be successfully matched. We discuss this issue in greater detail in section 4.

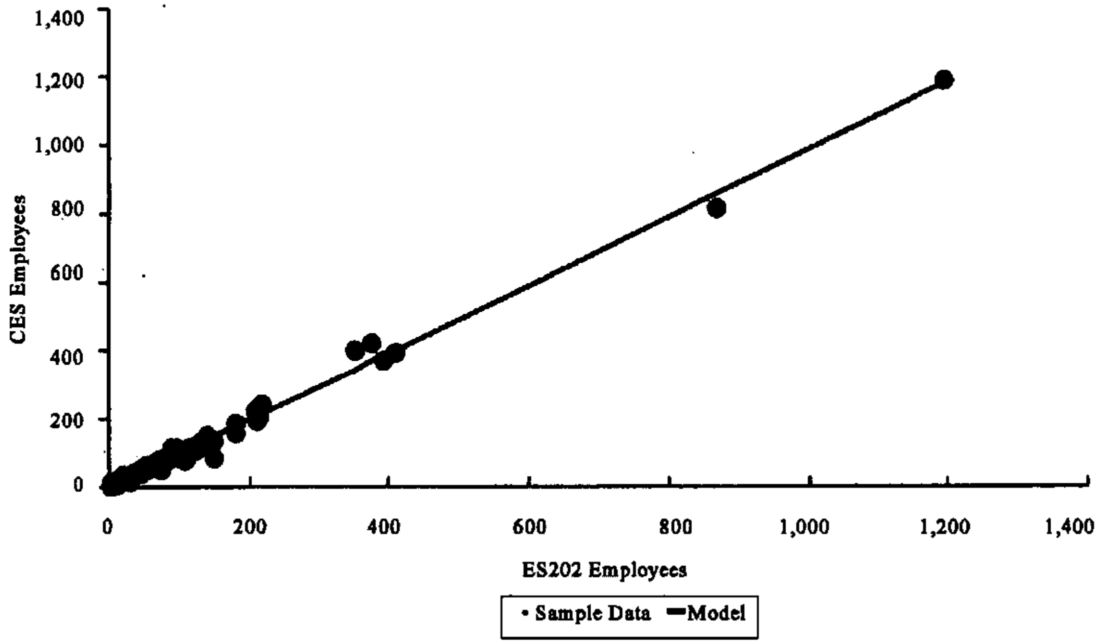


Figure 1. CES Versus ES202 Employment for a Sample of 103 Illinois Employers Classified in the Primary Metal Manufacturing Industry.

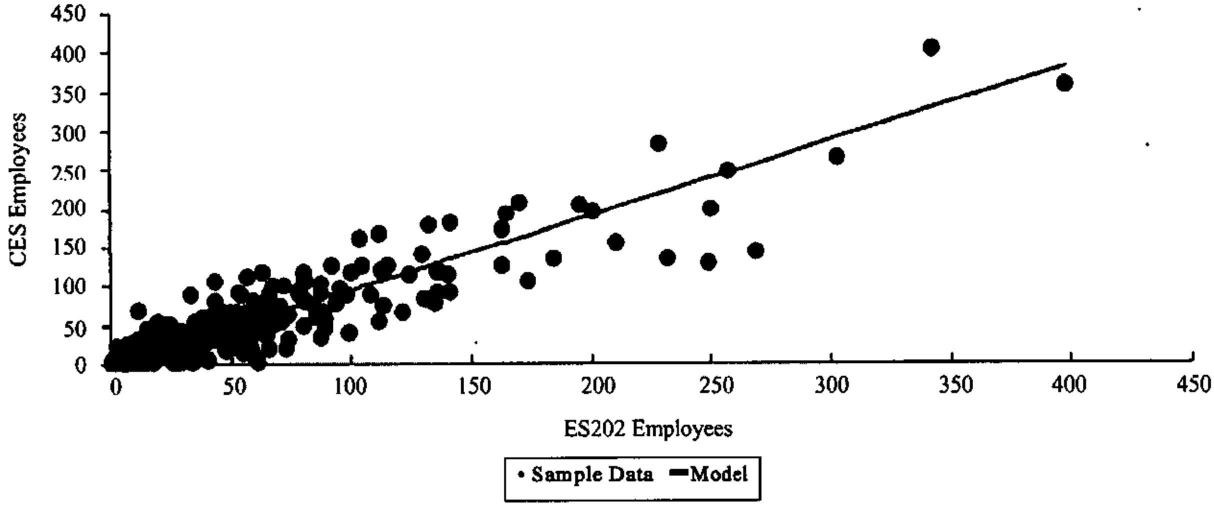


Figure 2. CES Versus ES202 Employment for a Sample of 701 Employers Classified in the Trade Contractors Industry

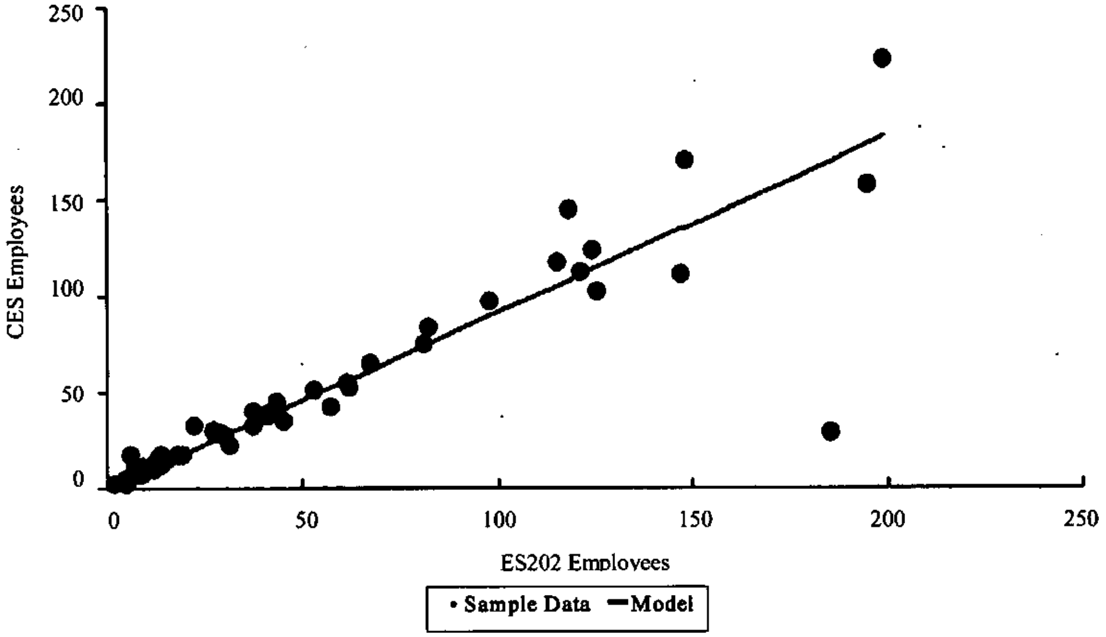


Figure 3. CES Versus ES202 Employment for a Sample of 50 Employers Classified in the Apparel Manufacturing Industry

3. EVALUATING THE MODEL

Since the CES and ES202 programs are both measuring employment, we expect the relationship between the two to be linear with intercept zero and slope close to one. The plots in Figures 1-3, and the many other similar plots we produced and reviewed, indicate that this is generally true. Industries with changes over time or differences in scope and coverage sometimes display slopes other than one. The plots also indicate variability in the linear relationships, and some industries exhibit more variability about the linear relationship than others. Generally, the residual variance about the line increases with employment.

The standard "ratio" model adequately describes most of our data. Let y_j be the current month CES employment for employer j, and let x_j be the ES202 employment for the same employer at some previous time period. Then the assumed model relationship is

$$y_j = x_j \beta + \varepsilon_j, \ \ \varepsilon_j \sim \mathrm{NID}(0, \sigma^2 x_j). \qquad (1)$$

The model parameter β can be estimated by generalized least squares, resulting in the ratio estimator $\hat{\beta} = \bar{y} / \bar{x}$, where $\bar{y}$ and $\bar{x}$ are the means of the observed current-month and auxiliary data, respectively (Sampling weights may or may not be employed in the analysis depending on many considerations beyond the scope of this article).

If Model (1) is true, then the best linear unbiased predictor of current month employment for sub-state domain D_{kl} (industry division k and geographic area l) is

$$\hat{Y}(D_{kl}) = \sum_{j \in U} x_j \hat{\beta}_{kl} \delta_j (D_{kl}), \qquad (2)$$

where $\delta_j(D_{kl})$ indicates whether unit j is in small domain D_{kl}; the summation is over all employers j within the state (or universe U); and $\hat{\beta}_{kl}$ is the ratio estimator within D_{kl}. With insufficient sample data to estimate the model parameters reliably at the small domain level, we instead estimate the parameters for model cell m (typically a 2-digit SIC at the state level), and apply the estimated model parameters to each of the small domains within the state. The resulting synthetic estimator is of the form

$$\hat{Y}(D_{kl}) = \sum_{m \in k} \sum_{j \in U_m} x_j \hat{\beta}_m \delta_j (D_{kl}), \qquad (3)$$

where the first summation is over all model cells that overlap with domain D_{kl} and the second summation is over all employers within the model cell. The estimator is a simple sum of predicted employment over all employers in the universe within the domain.

We tried an intercept in the model and verified that it was not significantly different from zero, in most cases. We tested that the slope was significantly different from zero. We plotted the residuals to verify that they were suitably well behaved. We checked the R^2 values to quickly assess the explanatory power of the model.

To illustrate this work, Table 2 gives summary statistics for models in Trade using January 1996 CES and January 1995 ES202 data. All of the R^2 values in Table 2 are quite high, ranging from .87 to .96. Only two of the intercepts are significantly different from zero. Except for Retail Trade, Apparel, where the intercept is significantly different from zero, all of the slopes are between .9 and 1.1.

The largest employers are selected into the sample with certainty. Because they are so influential and not necessarily typical, we decided to exclude them from the estimation of the model parameters.

We also tried Estimator (3) corresponding to large sub-state model cells. This approach loses sample size (and thus precision) relative to the statewide model cells, but presumably gains some greater ability to target local economic conditions (thus reducing bias, if any). Yet in comparing the resulting small domain estimates with "true" values in simulations, we found the estimators from statewide model cells to have the smaller mean squared errors.

Following the work of Battese, Harter, and Fuller (1988), we fit a components-of-variance model of the form

$$y_{ij} = x_{ij}\beta + v_i + \varepsilon_{ij}, \ v_i \sim \mathrm{NID}(0, \sigma_v^2), \ \varepsilon_{ij} \sim \mathrm{NID}(0, \sigma_e^2 x_{ij}) \quad (4)$$

and tested the homogeneity of the county-level variance components, v_i. While there was some indication of heterogeneity, the variability in the variance component estimates actually increased the mean squared errors of the small domain estimates in our simulations. We decided that the variance components approach was not superior to the simple synthetic estimator.

Table 2

Generalized Regression Models for CES All Employment on ES202 Year-Ago Employment: Trade Industries

Industries Defined by 2-Digit SIC Code	n	R^2	Intercept	Slope
Wholesale trade, durable goods	700	0.96	-0.061	1.015 **
Wholesale trade, nondurable goods	381	0.95	-0.032	0.978 **
Retail trade, building and garden supplies	189	0.96	0.420	0.918 **
Retail trade, general merchandise	42	0.95	-1.325	1.081 **
Retail trade, food stores	156	0.95	0.410	0.934 **
Retail trade, automobiles	379	0.97	0.130	0.971 **
Retail trade, apparel	112	0.90	1.320 **	0.750 **
Retail trade, furniture	110	0.95	0.242	0.931 **
Retail trade, eating & drinking establishments	460	0.89	0.382	0.968 **
Miscellaneous retail trade	332	0.87	0.810 **	0.915 **

* Significant at .05 level ** Significant at .01 level

We evaluated the synthetic estimator and other small domain estimators in a simulation study using Illinois data. The study included the simple unbiased estimator, the link relative estimator (Madow and Madow 1978, and West 1983,1984), raked estimators using CES estimates at higher aggregations as marginal totals, two variations of generalized regression estimators (Särndal and Hidiroglou 1989), and three variations of synthetic estimators. For some of the simulations, the data were restricted to cases for which the CES and ES202 data could be cleanly linked. We then drew repeated samples from this "universe" and tested the results against "truth." For later simulations, the data files included non-matches with rules for special handling based on likely causes of the mismatches. The handling of non-matches is described in the next section.

In the simulations, we used all the samples and the known truth to compute bias, relative bias, mean squared error, and relative mean squared error of estimated total employment and month-to-month change in employment. We also plotted the 5th, 50th, and 95th percentiles of the distribution of the estimators and examined the distributions in relation to the true values.

Results of the simulation study are reported in Harter *et al.* (1999). In general, we found that estimators that used ES202 as auxiliary data performed better than the direct sample-based estimator, the link relative estimator, and the raked estimators that used only sample data. The estimator that performed best overall was a variation of the synthetic estimator, derived from the prediction theory approach to survey sampling (Royall 1970, 1988, and Royall and Cumberland 1981a, 1981b). This estimator

$$\begin{aligned}\hat{Y}(D_{kl}) &= \sum_{m\in k}\sum_{j\in s_m} y_j\delta_j(D_{kl}) \\ &\quad+\sum_{m\in k}\sum_{j\notin s_m} x_j\hat{\beta}_m\delta_j(D_{kl}) \\ &= \sum_{m\in k}\sum_{j\in U_m} x_j\hat{\beta}_m\delta_j(D_{kl}) \\ &\quad+\sum_{m\in k}\sum_{j\in s_m}(y_j-x_j\hat{\beta}_m)\delta_j(D_{kl}) \qquad (5)\end{aligned}$$

is intuitively appealing to non-statisticians because the sample data are used directly for sample employers, while the model predictions are used only for nonsample employers. It is the synthetic estimator plus a sample-based correction for any lack of fit in the models.

4. MERGING THE DATA

The success of the small domain estimator depends, in part, on the ability to accurately match the CES and ES202 data. We can match CES and ES202 records by unemployment insurance number (UI) and establishment or reporting unit number (RU). When the CES reporter is an aggregate of establishments, such as a multi-site employer reporting all employees together without distinguishing individual work sites, the corresponding ES202 records must be aggregated to match. Figure 3 demonstrates an isolated instance of a bad aggregate match.

Plots of the kind presented in Figures 1-3 enabled us to identify many miscoded observations. For example, an aggregate reporter coded in the files as containing all the company's work sites, but that actually covers only a single work site, should have been coded as a single establishment. The process of checking outliers in all the plots was time-consuming, but resulted in major improvements in the micro data, which in turn improved the estimated model parameters.

Several situations make the match process problematic. First, the ES202 data contain employers that have gone out of business. Conversely, the CES data contain new employers that were not in existence at the time the ES202 data were collected, although difficulty in identifying new businesses in a timely fashion makes this scenario less common. Births and deaths of businesses, then, cause real mismatches in the data.

Second, nonresponse to either the CES or ES202 causes mismatches. Missing or delinquent reporters to the ES202 are usually imputed for a time. At present, imputation is not done for missing CES cases. A key difficulty with both programs is distinguishing nonresponse from a death.

Third, businesses often reorganize, merge, acquire other businesses, divest divisions, and so on. Any of these status changes can cause states to assign new unemployment insurance numbers. The predecessor businesses and successor businesses are treated as deaths and births. Alternatively, if a single predecessor can be linked to a single successor, their records could be joined to form one unified record. Unfortunately, the linkages are often not one-to-one. In many instances, predecessors are indistinguishable from deaths and delinquent CES reporters, and successors are indistinguishable from births and missing ES202 data.

For the initial implementation of our small domain estimator, we treat missing CES units as nonsample units; that is, we use their ES202 data and the model to predict their current month values. Since we cannot distinguish deaths and predecessors from missing CES data, we predict their current month employment using their ES202 data and the model. We use imputed ES202 data as real observations. Because it is relatively rare for a new business to appear in the CES sample data before it appears in the ES202, we treat CES records without ES202 counterparts as successor records. That is, in the small domain estimator, we treat them as nonmembers of the CES sample and predict their employment from the unmatched predecessor records in the ES202 file and the model. All of these decisions or judgments were based on IDES' experience.

Even if the UI and RU numbers match, the CES and ES202 records may differ in their industry or geographic codes due to differences in the programs' update cycles. Discrepancies might represent errors or legitimate changes. Originally, our thought was to use the CES codes in the small domain estimator, assuming CES codes were the more current. However, as the small domain estimator was being implemented, more and more of the CES data collection operations were being transferred from Illinois' control to central data collection centers operated by the BLS. IDES felt this loss of control could compromise the quality of the CES codes and thus they decided to use the ES202 codes instead. In actual production, we use these classification codes for all purposes, including definition of model cells, estimation of the slope parameters, and calculation of the small domain estimates.

Sometimes a well-matched sample unit experiences employment shifts that are not typical of the industry or the region as a whole. Both the CES and ES202 systems allow for comment codes in the data files so that anomalies and their reasons can be flagged. We developed an extensive set of rules for determining when a matched sample record may be used in the estimation of model parameters, and when this would be unwise. For example, a drop in employment due to weather or climate conditions, such as flooding along the Mississippi River, is a situation likely to be common to other businesses in the area. A record with a code for this type of anomaly should probably be included in the estimation of model parameters. A fire, on the other hand, is likely to affect one and only one business, and a drop in employment due to the fire could be very misleading if applied to nonsample businesses. In this case, the sample unit with the fire stands for itself, but it is not part of the calculation of the model parameters.

All the potential data problems and potential mismatches led us to modify the estimator slightly. The revised estimator is

$$\hat{Y}(D_{kl}) = \sum_{m \in k} \sum_{j \in s_m} y_j \delta_j(D_{kl}) + \sum_{m \in i} \sum_{j \notin s_m} x_j \hat{\beta}_m \delta_j(D_{kl}) + A_{kl}, \qquad (6)$$

where A_{kl} is an additive adjustment for known data deficiencies. This concession to practical realities was originally intended for situations such as the addition of railroad workers, where Illinois' CES manager obtains information on railroad employment from the Railroad Retirement Board because railroad workers are not covered by the state unemployment insurance program, and thus are missing from the ES202 data file. Clergy and summer youth workers are often added the same way. The CES manager and affiliated local economists scattered throughout the state have found the adjustment option useful for other known problems, such as employees that are reported at headquarters when they are really located around the state. Employees whose location is unknown are usually assigned to a nonspecific county "999" for inclusion in statewide estimates, but traditionally have been omitted from sub-state estimates. With the adjustment option, the CES manager can allocate the county 999 employment to individual counties in proportion to other employees in the same industry. Major births and deaths can be reflected in the estimates through the adjustments until the CES and ES202 files can catch up.

The danger of this adjustment capability is that it can be used to force small domain estimates to conform to the CES manager's or economists' judgments, rather than letting the data and models speak for themselves. The best possible model is useless if it is ignored or "fudged".

Despite the danger, Estimator (6) is the one that we have actually moved into production in Illinois. All matched respondent records contribute to the first term. All matched records not designated as atypical or certainty contribute to the estimated slope in the second term. The summation in the second term includes nonmatched ES202 cases and missing sample cases – all cases that are treated as nonsample cases that month. If we have a CES record that does not match anything in ES202, it is dropped altogether. At present, all data adjustments, A, are coordinated and approved through the CES manager, who operates under strict guidelines, including a requirement to maintain consistency with the CES estimates published by the BLS. Within the guidelines, the manager is granted discretion to determine when adjustments are in the best interest of the estimation process.

5. MONITORING THE PROCESS

It is preferable to discover and fix data problems prior to estimation rather than rely on the adjustment capability in estimation. Illinois has developed several tools for monitoring the data that feed the monthly estimation process. Many of these tools reside in Illinois' software that pre-processes and matches the data prior to estimation.

Matching proceeds as a by-product of CES' daily processing activities. The editing and registry maintenance of CES records involves review of ES202 records, which are available to CES staff through simple "point and click" tools. The CES staff designates a match between CES and ES202 records by a special code manually applied to the CES record and later read by the pre-processing software. Those CES records so indicated as matched are subsequently checked for ES202 congruence and uniqueness on the combination of UI, RU, industry, ownership type, county, and delinquency status. The clean matches are added to a *matched file*, which is available for further review through special diagnostic or exception reports. We developed and implemented an extensive set of rules for the staff to follow in resolving the messy matches – the one to many and many to one matches. The pre-processing software executes the rules and prints all cases of a certain type in a table for staff review. After applying all the rules and resolving the match statuses of the cases in the printed tables, we write remaining non-matching records to a separate *nonmatched file* for diagnostic reports and additional staff review.

From the matched file, we develop diagnostic or exception reports for CES staff. For instance, the pre-processing software generates a report of sample records whose CES and ES202 data differ more than one might expect. The basis for this exception report is a statistic derived from information theory. See Theil (1967), Strobel (1982), and Harter (1987). The statistic is computed for each sample observation as follows:

$$E_j = \frac{(y_j - x_j)^2}{(y_j + x_j)/2}. \tag{7}$$

It is a Taylor series approximation of a measure of entropy and under the null hypothesis has a χ^2 distribution with 1 df. The statistic provides a way of ranking data differences, and balancing absolute differences, dominated by larger employers, and relative differences, dominated by smaller employers. The CES manager can evaluate the cases with the largest values of E, identifying and correcting miscoded data prior to small domain estimation.

Other exception reports display duplicate CES records that were removed from the files. Duplicates are rare but can happen, for example, if two respondents from the same company each file CES reports. The exception reports display for review single establishment records in CES incorrectly matched to an aggregation in the ES202 that were dropped by the pre-processing software. Also displayed for review are unmatched CES records that could represent a successor or a birth employer. Other specialized diagnostics check the sums of ES202 records at county, MSA, and statewide levels for comparison with their respective CES counterparts.

After going through these exception reports and making changes where appropriate, CES staff may decide to rerun the pre-processing software using the newly updated data, if the production schedule permits.

The software that computes the small domain estimates has a final data check built in. The input data values and the estimated model parameters are checked against tables of "sanity values" for reasonableness. This is a gross check only, designed to signal when something very unexpected has occurred.

The estimation system produces tables of matched sample data and tables of nonsample data at the individual reporting unit level. The authorized users of the small domain estimation software – the CES manager and the affiliated local economists, among others – can review the micro data as well as the computed estimates. Based on their review, they can provide useful guidance regarding specification of the adjustment term A_{kt}.

The CES manager and local economists review the estimates themselves along with historical estimates to see whether the trends and seasonality in the observed time series are reasonable. For instance, Construction, Retail Trade, and Education Services all have strong seasonal patterns. Deviation from such patterns would suggest to the analyst that further review is needed. Manufacturing employment is thought to be trending downward over the long term, and there is a natural tendency to examine its time series in this context.

Finally, the CES manager and local economists summarize all of the labor market areas into one large entity. The larger employment numbers allow sharper delineation of seasonal and trend expectancies. They also allow for subsequent comparison with statewide estimates.

6. CONCLUSION

Many aspects of small domain estimation must be checked and rechecked in production on a monthly basis. The auxiliary variable must be investigated carefully with respect to its correlation with the survey variable and its reliability, compatibility, and availability. The record linkage process is challenging (but highly rewarding) and requires vigilance. The models and assumptions underlying the estimator must be checked and verified for reasonableness. The estimates themselves must be scrutinized regularly. Development of the small domain estimator forcefully shows that even with the most ideal auxiliary variable and a textbook model, practical issues can intrude and require that flexibility be built into the estimation process.

REFERENCES

BATTESE, G.E., HARTER, R.M. and FULLER, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*. 83, 28-36.

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

HARTER, R. (1987). Exception reporting: judging what is significant. *Nielsen Marketing Trends*. January. 20-23.

HARTER, R., WOLTER, K. and MACALUSO, M. (1999). Small domain estimation of employment using CES and ES202 data. In *Statistical Policy Working Paper 30, 1999 Federal Committee on Statistical Methodology Research Conference: Complete Proceedings, Part 1 of 2*. Washington DC: Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget.

MADOW, L., and MADOW, W. (1978). On link relative estimators. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 534-539.

PURCELL, N.J., and KISH, L. (1980). Postcensal estimates for local areas (or domains). *International Statistical Review*. 48, 3-18.

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

ROYALL, R.M. (1988). The prediction approach to sampling theory. In *Handbook of Statistics*, (Eds. P.R. Krishnaiah and C.R. Rao). New York: North Holland. 6, 399-413.

ROYALL, R.M., and CUMBERLAND, W.G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*. 76, 66-77.

ROYALL, R.M., and CUMBERLAND, W.G. (1981b). The finite population linear regression estimator and estimators of its variance – an empirical study. *Journal of the American Statistical Association*. 76, 924-930.

SÄRNDAL, C.-E., and HIDIROGLOU, M.A. (1989). Small domain estimation: A conditional analysis. *Journal of the American Statistical Association*. 84, 266-275.

SINGH, M.P., GAMBINO, J. and MANTEL, H.J. (1994). Issues and strategies for small area data (with discussions). *Survey Methodology*. 20, 3-22.

STROBEL, D. (1982). Determining outliers in multivariate surveys by decomposition of a measure of information. *Proceedings of Section on Business and Economic Statistics,* American Statistical Association.

THEIL, H. (1967). Economics and Information Theory. *Studies in Mathematical and Managerial Economics*, (Ed. H. Theil). Amsterdam: North Holland.

WEST, S. (1983). A comparison of different ratio and regression type estimators for the total of a finite population. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 388-393.

WEST, S. (1984). A comparison of estimators for the variance of regression-type estimators in a finite population. *Proceedings of the Section on Survey Research Methods*, American Statistical Association. 170-175.

Solving The Error Localization Problem by Means of Vertex Generation

TON DE WAAL[1]

ABSTRACT

To automate the data editing process the so-called error localization problem, *i.e.*, the problem of identifying the erroneous fields in an erroneous record, has to be solved. A paradigm for identifying errors automatically has been proposed by Fellegi and Holt in 1976. Over the years their paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights. A reliability weight of a variable is a non-negative number that expresses how reliable one considers the value of this variable to be. Given this paradigm the resulting mathematical problem has to be solved. In the present paper we examine how vertex generation methods can be used to solve this mathematical problem in mixed data, *i.e.*, a combination of categorical (discrete) and numerical (continuous) data. The main aim of this paper is not to present new results, but rather to combine the ideas of several other papers in order to give a "complete", self-contained description of the use of vertex generation methods to solve the error localization problem in mixed data. In our exposition we will focus on describing how methods for numerical data can be adapted to mixed data.

KEY WORDS: Chernikova's algorithm; Error localization; Fellegi-Holt paradigm; Fourier-Motzkin elimination; Mathematical programming; Mixed data editing; Statistical data editing; Vertex generation.

1. INTRODUCTION

An important problem that has to be solved in order to automate the data editing process is the so-called error localization problem, *i.e.*, the problem of identifying the erroneous fields in an erroneous record. Fellegi and Holt (1976) describe a paradigm for identifying errors in a record automatically. According to this paradigm the data of a record should be made to satisfy all edits by changing the values of the fewest possible number of variables. In due course the original Fellegi-Holt paradigm has been generalized to: the data of a record should be made to satisfy all edits by changing the values of the variables with the smallest possible sum of reliability weights. A reliability weight of a variable is a non-negative number that expresses how reliable one considers the value of this variable to be. A high reliability weight corresponds to a variable of which the values are considered trustworthy, a low reliability weight to a variable of which the values are considered not so trustworthy.

Describing a paradigm for identifying the erroneous fields in an erroneous record is only a first step towards solving the error localization problem, however. The second step consists of actually solving the resulting mathematical problem. This mathematical problem can be solved in several ways, see *e.g.* Fellegi and Holt (1976); De Waal and Quere (2003), and De Waal (2003). One of these ways is by generating vertices of a certain polyhedron. Unfortunately, the number of vertices of this polyhedron is often too high for this approach to be applicable in practice. Instead, one should therefore generate a suitable subset of the vertices only. There are a number of vertex generation algorithms that efficiently generate such a suitable subset of vertices of a polyhedron. An example of such a vertex generation algorithm is an algorithm proposed by Chernikova (1964, 1965). Probably most computer systems for automatic edit and imputation of numerical data are based on adapted versions of this algorithm. The best-known example of such a system is GEIS (Kovar and Whitridge 1990). Other examples are CherryPi (De Waal 1996), AGGIES (Todaro 1999), and a SAS program developed by the Central Statistical Office of Ireland (see Central Statistical Office 2000). The original algorithm of Chernikova is rather slow for solving the error localization problem. It has been accelerated by various modifications (see Rubin 1975 and 1977; Sande 1978; Schiopu-Kratina and Kovar 1989; Fillion and Schiopu-Kratina 1993).

Only the last three of these papers focus on the error localization problem. Sande (1978) discusses the error localization problems for numerical data, categorical data and mixed data. The discussion of the error localization problem in mixed data is very brief, however. Schiopu-Kratina and Kovar (1989) and Fillion and Schiopu-Kratina (1993) propose a number of improvements on Sande's method for solving the error localization problem for numerical data. They do not consider the error localization problems for categorical or mixed data.

In the present paper we examine how vertex generation methods can be used to solve the error localization problem in mixed data, *i.e.*, a combination of categorical (discrete) and numerical (continuous) data. The main aim of this paper is not to present new results, but rather to combine

[1] Ton de Waal, Statistics Netherlands, PO Box 40000, 2270 JM Voorburg, The Netherlands. E-mail: twal@cbs.nl.

the ideas of the above-mentioned papers in order to give a "complete", self-contained description of the use of vertex generation methods to solve the error localization problem in mixed data. We will especially describe how modifications to accelerate Chernikova's algorithm for numerical data can also be used for mixed data.

The remainder of the present paper is organized as follows. Section 2 gives a formal definition of the edits that we consider as well as a number of examples. Section 3 formulates the error localization problem as a mixed integer programming problem. Section 4 describes how the error localization problem can be solved by generating vertices of an appropriate polyhedron. We describe how Chernikova's algorithm can be used to generate these vertices in sections 5 and 6. In these sections we also describe modifications to the algorithm in order to improve its performance. Section 7 concludes the paper with a brief discussion. In the Appendix we give Rubin's description of Chernikova's algorithm. In this paper proofs are omitted for most results. The interested reader is referred to the literature for those proofs.

2. THE EDITS

2.1 Formal Definition of the Edits

We denote the categorical variables by $v_i\,(i = 1, ..., m)$ and the numerical variables by $x_i\,(i = 1, ..., n)$. For categorical data we denote the domain, *i.e.*, the set of possible values, of variable *i* by D_i. We assume that every edit $E^j\,(j = 1, ..., J)$ is written in the following form: edit E^j is satisfied by a record $(v_1, ..., v_m, x_1, ..., x_n)$ if and only if the following statement holds true:

$$\text{IF} \quad v_i \in F_i^j \quad \text{for} \quad i = 1, ..., m$$

$$\text{THEN} \quad (x_1, ..., x_n) \in \{\mathbf{x} \mid a_{1j}x_1 + ... + a_{nj}x_n + b_j \geq 0\}, \tag{2.1}$$

where $F_i^j \subset D_i\,(j = 1, ..., J)$. Numerical variables may attain negative values. For non-negative variables an edit of type (2.1) needs to be introduced in order to ensure non-negativity. A numerical equality can be expressed as two inequalities.

All edits have to be satisfied simultaneously. A record that satisfies all edits is called a consistent record. The condition after the IF-statement, *i.e.*, "$v_i \in F_i^j$ for all $i = 1, ..., m$", is called the IF-condition of edit $j\,(j = 1, ..., J)$. The condition after the THEN-statement is called the THEN-condition. If the IF-condition does not hold true, the edit is always satisfied, irrespective of the values of the numerical variables. If the set in the THEN-condition of (2.1) is the entire *n*-dimensional real vector space, then the edit is always satisfied and may be discarded. If the set in the THEN-condition of (2.1) is empty, then the edit is failed by any record for which the IF-condition holds.

In many practical cases, certain kinds of missing values are acceptable, *e.g.* when the corresponding questions are not applicable to a particular respondent. We assume that for categorical variables such acceptable missing values are coded by special values in their domains. Non-acceptable missing values of categorical variables are not coded. The optimization problem of section 3 will identify these missing values as being erroneous. We also assume that numerical THEN-conditions are only be triggered if none of the values of the variables involved may be missing. Hence, if – in a certain record – a THEN-condition involving a numerical variable of which the value is missing is triggered by the categorical values, then either the missing numerical value is erroneous or at least one of the categorical values.

2.2 Examples of Edits

Below we illustrate what kind of edits can be expressed in the form (2.1) by means of a number of examples.

1. *Turnover* – *Profit* ≥ 0. (2.2)

   This is an example of a numerical edit. For every combination of categorical values the edit should be satisfied. The edit can be formulated in our standard form as:

   $$\text{IF} \quad v_i \in D_i \quad \text{for all} \quad i = 1, ..., m$$

   $$\text{THEN } (\mathit{Profit}, \mathit{Turnover}) \in \{(\mathit{Profit}, \mathit{Turnover}) \mid \mathit{Turnover} - \mathit{Profit} \geq 0\}. \tag{2.3}$$

   In the remaining examples we will be slightly less formal with our notation. In particular, we will omit the terms "$v_i \in D_i$" from the edits.

2. IF (*Gender* = "*Male*") THEN (*Pregnant* = "*No*")(2.4)

   This is an example of a categorical edit. It can be formulated in our standard form as:

   IF (*Gender* = "*Male*") AND (*Pregnant* = "*Yes*")
   THEN ∅. (2.5)

3. IF (*Occupation* = "*Statistician*")
   THEN (*Income* ≥ 1,000 Euro). (2.6)

   This is a typical example of a mixed edit. Given certain values for the categorical variables, a certain numerical constraint has to be satisfied.

4. IF (*Occupation* = "*Statistician*")
   OR (*Education* = "*University*")
   THEN (*Income* ≥ 1,000 Euro). (2.7)

   This edit can be split into two edits given by (2.6) and

   IF (*Education* = "*University*")
   THEN (*Income* ≥ 1,000 Euro). (2.8)

5. IF (*Tax on Wages* > 0)
THEN (*Number of Employees* $\geq$ 1). (2.9)

Edit (2.9) is not in standard form (2.1), because the IF-condition involves a numerical variable. To handle such an edit, one can carry out a pre-processing step to introduce an additional categorical variable *TaxCond* with domain {"*False*", "*True*"}. Initially, *TaxCond* is given the value "*True*" if *Tax on Wages* > 0 in the unedited record, and the value "*False*" otherwise. The reliability weight *TaxCond* is set to zero. We can now replace (2.9) by the following three edits of type (2.1):

IF (*TaxCond* = "*False*")
THEN (*Tax on Wages* $\leq$ 0), (2.10)

IF (*TaxCond* = "*True*")
THEN (*Tax on Wages* $\geq \varepsilon$), (2.11)

IF (*TaxCond* = "*True*")
THEN (*Number of Employees* $\geq$ 1), (2.12)

where ε is a sufficiently small positive number.

3. THE ERROR LOCALIZATION PROBLEM AS A MIXED INTEGER PROGRAMMING PROBLEM

We assume that the values of the numerical variables are bounded. That is, we assume that for the i-th numerical variable ($i = 1, \ldots, n$) constants α_i and β_i exist such that

$$\alpha_i \leq x_i \leq \beta_i \tag{3.1}$$

for all consistent records. In practice, such values α_i and β_i always exist although they may be very large, because numerical variables that occur in data of statistical offices are bounded. The values of α_i and β_i may be negative. If the value of the i-th numerical variable is missing, we code this by assigning a value less than α_i or larger than β_i to x_i. Numerical variables for which the value should be missing, *e.g.* because the corresponding question was non-applicable, will nonetheless receive a value after the termination of the algorithm that is described in subsequent sections, but this value may subsequently be ignored.

For the i-th categorical variable, let $D_i = \{c_{ik}, k = 1, \ldots, g_i\}$ ($i = 1, \ldots, m$) be its domain. We introduce the binary variable γ_{ik}

$$\gamma_{ik} = \begin{cases} 1 & \text{if the value of categorical variable } i \text{ equals } c_{ik} \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

To the i-th categorical variable there corresponds a vector $(\gamma_{i1}, \ldots, \gamma_{ig_i})$ such that $\gamma_{ik} = 1$ if and only if the value of this categorical variable equals c_{ik}, otherwise $\gamma_{ik} = 0$. For each categorical variable i of a consistent record the relation

$$\sum_k \gamma_{ik} = 1 \tag{3.3}$$

has to hold, *i.e.*, exactly one categorical value should be filled in. The vector $(\gamma_{i1}, \ldots, \gamma_{ig_i})$ will also be denoted by $\boldsymbol{\gamma}_i$. If the value of the i-th categorical variable ($i = 1, \ldots, m$) is missing, we set all γ_{ik} equal to zero ($k = 1, \ldots, g_i$). In terms of the binary variables γ_{ik} an edit j given by (2.1) can be written as

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j \geq M\left(\sum_{i=1}^{m}\left(\sum_{c_{ik} \in F_i^j} \gamma_{ik} - 1\right)\right), \tag{3.4}$$

where a positive M is chosen so that $-M$ is less than the lowest possible value of $a_{1j}x_1 + \ldots + a_{nj}x_n + b_j$. If the IF-condition of (2.1) and condition (3.3) hold true, the right-hand side of (3.4) equals zero. Consequently, the THEN-condition of (2.1) has to hold true for the numerical variables. If the IF-condition of (2.1) does not hold true, by (3.2) the right-hand side of (3.4) equals a large negative value. Consequently, (3.4) holds true irrespective of the values of numerical variables.

If (2.1) is not satisfied by a record $(v_1^0, \ldots, v_m^0, x_1^0, \ldots, x_n^0)$, or equivalently if (3.4) is not satisfied by $(\boldsymbol{\gamma}_1^0, \ldots, \boldsymbol{\gamma}_m^0, x_1^0, \ldots, x_n^0)$, then we seek values $e_{ik}^P (k = 1, \ldots, g_i; i = 1, \ldots, m), e_{ik}^N (k = 1, \ldots, g_i; i = 1, \ldots, m)$, $z_i^P (i=1, \ldots, n)$ and $z_i^N (i=1, \ldots, n)$ that have to satisfy certain conditions mentioned below. The e_{ik}^P and the e_{ik}^N correspond to positive and negative changes, respectively, in the value of γ_{ik}^0. Likewise, the z_i^P and the z_i^N correspond to positive and negative changes, respectively, in the value of x_i^0. The vector $(e_{i1}^P, \ldots, e_{ig_i}^P)$ will also be denoted as $\mathbf{e}_i^P$ and the vector $(e_{i1}^N, \ldots, e_{ig_i}^N)$ as $\mathbf{e}_i^N$.

The objective function we consider in this paper is given by

$$\sum_{i=1}^{m} w_i^c \left(\sum_k e_{ik}^P\right) + \sum_{i=1}^{n} w_i^r \left(\delta(z_i^P) + \delta(z_i^N)\right), \tag{3.5}$$

where w_i^c is the reliability weight of the i-th categorical variable $(i=1, \ldots, m)$, w_i^r the reliability weight of the i-th real-valued variable $(i=1, \ldots, n)$, $\delta(x) = 1$ if $x \neq 0$ and $\delta(x) = 0$ otherwise. The objective function (3.5) is the sum of the reliability weights of the variables for which a new value must be imputed. Note that minimizing (3.5) is equivalent to minimizing

$$\sum_{i=1}^{m} w_i^c \left(\sum_k e_{ik}^N\right) + \sum_{i=1}^{n} w_i^r \left(\delta(z_i^P) + \delta(z_i^N)\right). \tag{3.6}$$

The objective function (3.6) is the sum of the reliability weights of the variables of which the original values must be modified. The value of the objective function (3.5) is equal to the value of the objective function (3.6) plus the sum of reliability weights of the categorical variables for which the original value was missing.

The objective function (3.5) is to be minimized subject to the following constraints:

$$e_{ik}^{P}, e_{ik}^{N} \in \{0,1\}, \qquad (i=1, ..., m) \qquad (3.7)$$

$$z_i^{P}, z_i^{N} \geq 0, \qquad (i=1, ..., n) \qquad (3.8)$$

$$e_{ik}^{P} + e_{ik}^{N} \leq 1 \qquad (i=1, ..., m) \qquad (3.9)$$

$$\sum_k e_{ik}^{P} \leq 1, \qquad (i=1, ..., m) \qquad (3.10)$$

$$e_{ik}^{N} = 0 \qquad \text{if } \gamma_{ik}^{0} = 0 \qquad (i=1, ..., m) \qquad (3.11)$$

$$\sum_k (\gamma_{ik}^{0} + e_{ik}^{P} - e_{ik}^{N}) = 1, \qquad (i=1, ..., m) \qquad (3.12)$$

$$\alpha_i \leq x_i^{0} + z_i^{P} - z_i^{N} \leq \beta_i \qquad (i=1, ..., n) \qquad (3.13)$$

and

$$\sum_{i=1}^{n} a_{ij}(x_i^{0} + z_i^{P} - z_i^{N}) + b_j \geq M\left(\sum_{i=1}^{m}\left(\sum_{c_{ik} \in F_i^j} (\gamma_{ik}^{0} + e_{ik}^{P} - e_{ik}^{N}) - 1\right)\right) \qquad (3.14)$$

for all edits $j = 1, ..., K$.

Relation (3.9) expresses that a negative correction and a positive one may not be applied to the same reported value of a categorical variable. Relation (3.10) expresses that at most one value may be imputed, *i.e.*, estimated and subsequently filled in, for a categorical variable, and relation (3.11) that a negative correction may not be applied to a categorical value that was not filled in. Relation (3.12) ensures that a value for each categorical variable is filled in, even if the original value was missing. Relation (3.13) states that the value of a numerical variable must be bounded by the appropriate constants. In particular, relation (3.13) also states that the value of a numerical variable may not be missing. Finally, relation (3.14) expresses that the modified record should satisfy all edits given by (2.1).

After solving this optimization problem the resulting, modified record is given by

$$(\gamma_1^{0} + \mathbf{e}_1^{\mathbf{P}} - \mathbf{e}_1^{\mathbf{N}}, ..., \gamma_m^{0} + \mathbf{e}_{\mathbf{m}}^{\mathbf{P}} - \mathbf{e}_{\mathbf{m}}^{\mathbf{N}}, x_1^{0} + z_1^{P} - z_1^{N}, ..., x_n^{0} + z_n^{0} - z_n^{N}).$$

This modified record is consistent, *i.e.*, satisfies all edits. A solution to the above mathematical problem corresponds to a solution to the error localization problem, which simply consists of a list of variables of which the values have to be changed without specifying their new values. There may be several optimal solutions to the error localization problem. Our aim is to find all these optimal solutions. Note that the above optimization problem is a translation of the generalized Fellegi-Holt paradigm in mathematical terms.

We end this section with two remarks. First, note that in practice only one e_{ik}^{N}-variable for each variable i is needed, namely for the index k for which $\gamma_{ik}^{0} = 1$. The other e_{ik}^{N} equal zero. In the present paper we use g_i binary e_{ik}^{N}-variables for each variable i to cover all possible cases. Second, note that in an optimal solution to the above optimization problem either $z_i^{P} = 0$ or $z_i^{N} = 0$, and that, similarly, in any feasible solution either $e_{ik}^{P} = 0$ or $e_{ik}^{N} = 0$ (or both).

4. VERTEX GENERATION METHODS AND ERROR LOCALIZATION FOR MIXED DATA

In this section we explain how vertex generation methods can be used to solve the error localization problem in mixed data. To this end we show that a minimum of (3.5) subject to (3.7) to (3.14) is attained in a vertex of a certain polyhedron P described by linear, non-integer constraints. Suppose a minimum of (3.5) subject to (3.7) to (3.14) is attained in a point given by:

1. $e_{ik}^{N} = 0$ for $(i, k) \in I_e^{N}$, $e_{ik}^{N} = 1$ otherwise,
2. $e_{ik}^{P} = 0$ for $(i, k) \in I_e^{P}$, $e_{ik}^{P} = 1$ otherwise,
3. $z_i^{P} = 0$ for $i \in I_z^{P}$, $z_i^{P} \neq 0$ otherwise,
4. $z_i^{N} = 0$ for $i \in I_z^{N}$, and $z_i^{N} \neq 0$ otherwise,

for certain index sets $I_e^{N}, I_e^{P}, I_z^{N}$ and I_z^{P}. We now consider the problem of minimizing the linear function given by

$$\sum_{(i,k) \in I_e^N} e_{ik}^{N} + \sum_{(i,k) \notin I_e^N} (1 - e_{ik}^{N}) + \sum_{(i,k) \in I_e^P} e_{ik}^{P} + \sum_{(i,k) \notin I_e^P} (1 - e_{ik}^{P}) + \sum_{i \in I_z^P} z_i^{P} + \sum_{i \in I_z^N} z_i^{N} \qquad (4.1)$$

subject to (3.8) to (3.14) and

$$0 \leq e_{ik}^{N}, e_{ik}^{P} \leq 1. \qquad (4.2)$$

Subject to (3.8) to (3.14) and (4.2), which together form our polyhedron P, the function (4.1) is non-negative. Moreover, its value equals zero only for the point given by 1 to 4 above. In other words, our selected minimum of (3.5) subject to (3.7) to (3.14) is also the minimum of (4.1) subject to (3.8) to (3.14) and (4.2).

It is well known that a linear function subject to a set of linear constraints attains its minimum, if such a minimum exists, in a vertex of the feasible polyhedron described by the set of linear constraints (see *e.g.* Chvátal 1983). So, the minimum of (4.1) subject to (3.8) to (3.14) and (4.2), zero, is attained in a vertex of the feasible polyhedron P described by (3.8) to (3.14) and (4.2). We conclude that the point given by 1 to 4 above, *i.e.*, an arbitrary optimum of (3.5) subject to (3.7) to (3.14), is a vertex of the polyhedron defined by (3.8) to (3.14) and (4.2).

The above observation implies that the minimum of (3.5) subject to (3.7) to (3.14) can be found by generating all vertices of the polyhedron given by (3.8) to (3.14) and (4.2). From these vertices we select the vertices that satisfy (3.7). From those latter vertices we subsequently select the vertices for which the value of the objective function (3.5) is minimal. These vertices correspond to the optimal solutions to the error localization problem.

5. CHERNIKOVA'S ALGORITHM AND THE ERROR LOCALIZATION PROBLEM

Chernikova's algorithm (Chernikova 1964 and 1965) was designed to generate the edges of a system of linear inequalities given by

$$\mathbf{Cx} \geq \mathbf{0} \tag{5.1}$$

and

$$\mathbf{x} \geq \mathbf{0}, \tag{5.2}$$

where $\mathbf{C}$ is a constant $n_r \times n_c$-matrix and $\mathbf{x}$ an n_c-dimensional vector of unknowns. The algorithm is described in the Appendix. It can be used to find the vertices of a system of linear inequalities because of the following lemma (see Rubin 1975 and 1977).

Lemma 5.1. *The vector* $\mathbf{x}^0$ *is a vertex of the system of linear inequalities*

$$\mathbf{Ax} \leq \mathbf{b} \tag{5.3}$$

and

$$\mathbf{x} \geq \mathbf{0} \tag{5.4}$$

if and only if $\{(\lambda \mathbf{x}^0 \mid \lambda)^T, \lambda \geq 0\}$ *is an edge of the cone described by*

$$(-\mathbf{A} \mid \mathbf{b}) \begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} \geq \mathbf{0} \tag{5.5}$$

and

$$\begin{pmatrix} \mathbf{x} \\ \xi \end{pmatrix} \geq \mathbf{0}. \tag{5.6}$$

Here A *is an* $n_r \times n_c$*-matrix,* b *an* n_r*-vector,* x *an* n_v*-vector, and* ξ *and* λ *scalar variables.*

For notational convenience we write

$$n_c = n_v + 1 \tag{5.7}$$

throughout this paper. The matrix in (5.5) is then an $n_r \times n_c$-matrix just like in (5.1), so we can use the same notation as in Rubin's formulation of Chernikova's algorithm.

If Chernikova's algorithm is used to determine the edges of (5.5) and (5.6), then after the termination of the algorithm the vertices of (5.3) and (5.4) correspond to those columns j of $\mathbf{L}^{n_r}$ (see Appendix) for which $l_{n_c,j}^{n_r} \neq 0$. The entries of such a vertex $\mathbf{x}'$ are given by

$$x_i' = l_{ij}^{n_r} / l_{n_c,j}^{n_r} \quad \text{for} \quad i = 1, \ldots, n_v. \tag{5.8}$$

Now, we explain how Chernikova's algorithm can be used to solve the error localization problem in mixed data. The set of constraints (3.8) to (3.14) and (4.2) can be written in the form (5.3) and (5.4). We can find the vertices of the polyhedron corresponding to this set of constraints by applying Chernikova's algorithm to (5.5) and (5.6). Vertices of the polyhedron defined by (3.8) to (3.14) and (4.2) are given by columns $y_{*s}^{n_r}$ for which $u_{is}^{n_r} \geq 0$ for all i and $l_{n_c,s}^{n_r} > 0$, where n_c is the number of rows of the final matrix $\mathbf{L}^{n_r}$ (see Appendix). In our case, n_c equals the total number of variables z_i^P, z_i^N, e_{ik}^P and e_{ik}^N plus one (corresponding to ξ in (5.5) and (5.6)), *i.e.*, $n_c = 2n + 2G + 1$, where $G = \sum_i g_i$. The values of the variables z_i^P, z_i^N, e_{ik}^P and e_{ik}^N in such a vertex are given by the corresponding values $l_{js}^{n_r} / l_{n_c,s}^{n_r}$.

Two technical problems must be overcome when Chernikova's algorithm is applied to solve the error localization problem for mixed data. First, the algorithm must be sufficiently fast. Second, the solution found must be feasible for the error localization problem for mixed data, *i.e.*, the values of the variables e_{ik}^P and e_{ik}^N must be either 0 or 1. Both problems can be overcome by removing certain "undesirable" columns from the current matrix $\mathbf{Y}^k$, *i.e.*, by deleting columns that cannot yield an optimal solution to the error localization problem. That such undesirable columns may indeed be removed from the current matrix $\mathbf{Y}^k$ is essentially demonstrated by Rubin (1975 and 1977). We state this result as Theorem 5.1.

Theorem 5.1. *Columns that cannot yield an optimal solution to the error localization problem because they contain too many non-zero entries may be removed from an intermediate matrix.*

To accelerate Chernikova's algorithm, we aim to limit the number of vertices that are generated as much as possible. Once we have found a (possibly suboptimal) solution to the error localization problem for which the objective value (3.5) equals η, say, we from then on look only for vertices corresponding to solutions with an objective value at most equal to η. A minor technical problem is that we cannot use the objective function (3.5) directly when applying Chernikova's algorithm, because the values of $e_{ik}^P, e_{ik}^N, z_i^N$ and z_i^P are not known during the execution of this algorithm. Therefore, we introduce a new objective function that associates a value to each column of the matrix $\mathbf{Y}^k$ (see Appendix). Assume that the first G entries of a column l_{*s}^k of $\mathbf{L}^k$ correspond to the e_{ik}^P-variables, the next G entries to the e_{ik}^N-variables, the next n entries to the z_i^P-variables, and the subsequent n entries to the z_i^N-variables. We define the following objective function

$$\sum_{i=1}^{m} w_i^c \left(\sum_{k=1}^{g_i} \delta(l_{t,s}^k) \right) + \sum_{i=1}^{n} w_i^r \times \left(\delta(l_{2G+i,s}^k) + \delta(l_{2G+n+i,s}^k) \right), \tag{5.9}$$

where $t = \sum_{l=1}^{i-1} g_l + r$ for each pair $\{i, r\}$ $(i = 1, ..., m; r = 1, ..., g_i)$. Differences between (3.5) and (5.9) are that for each e_{ik}^P or e_{ik}^N in (3.5) several variables $l_{t,s}^k$ occur in (5.9), and that the e_{ik}^P and e_{ik}^N attain values in $\{0,1\}$ whereas the $l_{t,s}^k$ can attain any value between zero and one. If column y_{*s}^k of $\mathbf{Y}^k$ corresponds to a solution to the error localization problem, then the value of the objective function (5.9) for y_{*s}^k equals the value of the objective function (3.5) for this solution. This implies that we can use the objective function (5.9) to update the value of η.

The computing time of Chernikova's algorithm can be further reduced by noting that in an optimal solution to the error localization problem either $z_i^P = 0$ or $z_i^N = 0$ (or both). This implies that in Step 7 of Chernikova's algorithm (see Appendix) columns y_{*s}^k and y_{*t}^k need not be combined if one of these columns corresponds to $z_i^P \neq 0$ and the other to $z_i^N \neq 0$. Theorem 5.1 implies that not combining such columns is allowed.

We now consider the problem of constructing a feasible solution to the error localization problem for mixed data. This problem can, of course, be solved by first generating vertices without taking into account that values of e_{ik}^P and e_{ik}^N must be either 0 or 1 and then selecting the best vertices that possess this property, but this is rather inefficient so we suggest a different approach. It suffices to ensure that for each variable $i (i = 1, ..., m)$ at most one e_{ik}^P differs from zero, and that the e_{ik}^N and e_{ik}^P equal either zero or one after the termination of the algorithm. We can ensure that for each i at most one e_{ik}^P differs from zero in the following way. If in Step 7 of Chernikova's algorithm the entry of y_{*s}^k corresponding to $e_{ik_1}^P$ differs from zero and the entry of y_{*t}^k corresponding to $e_{ik_2}^P$ $(k_2 \neq k_1)$ differs from zero as well, then columns y_{*s}^k and y_{*t}^k are not combined to generate a new column. We can also ensure that the e_{ik}^N equal either zero or one after the termination of the algorithm. For each i this is a problem only for the unique $e_{ik_0}^N$ for which $\gamma_{ik_0}^0 = 1$. We introduce variables $\tilde{e}_i$ that can attain values between zero and one. These variables have to satisfy

$$e_{ik_0}^N + \tilde{e}_i = 1. \tag{5.10}$$

Relation (5.10) is treated as a constraint for the values of the variables $e_{ik_0}^N$ and $\tilde{e}_i$. Because the value of $e_{ik_0}^N$ has to be either zero or one, we demand that either $e_{ik_0}^N = 0$ or $\tilde{e}_i = 0$. This can be ensured in the same manner as for the z_i^P and the z_i^N. Finally, we have to ensure that the e_{ik}^P equal either zero or one after the termination of the algorithm. This is automatically the case if for each i at most one e_{ik}^P differs from zero, at most one e_{ik}^N equals one and the remaining e_{ik}^N equal zero, because relation (3.12) has to hold true. We have already ensured that these conditions are satisfied, so all e_{ik}^P equal zero or one after the termination of the algorithm. With the adaptations described above Chernikova's algorithm can be applied to solve the error localization problem in mixed data. Theorem 5.1 again implies that these modifications are allowed.

6. ADAPTING CHERNIKOVA'S ALGORITHM TO THE ERROR LOCALIZATION PROBLEM

6.1 Advanced Adaptations

In this section we consider more advanced adaptations of Chernikova's algorithm in order to make the algorithm better suited for solving the error localization problem. Sande (1978) notes that when two columns in the initial matrix $\mathbf{Y}^0$ have exactly the same entries in the upper matrix $\mathbf{U}^0$, they will be treated exactly the same in the algorithm. The two columns are always combined with the same other columns, and never with each other. Keeping both columns in the matrix only makes the problem unnecessarily bigger. One of the columns may therefore be temporarily deleted. After the termination of the algorithm, the solutions to the error localization problem involving the temporarily deleted column can easily be generated.

A correction pattern associated with column y_{*s}^k in an intermediate matrix $\mathbf{Y}^k$, where $\mathbf{Y}^k$ can be split into an upper matrix $\mathbf{U}^k$ and lower matrix $\mathbf{L}^k$ with n_r and n_c rows respectively (see Appendix), is defined as the n_c-dimensional vector with entries $\delta(y_{js}^k)$ for $n_r < j \leq n_r + n_c$. For each z_i^P, z_i^N, e_{ik}^P, and e_{ik}^N a correction pattern contains an entry with value in $\{0,1\}$. Sande (1978) notes that Theorem 5.1 implies that once a vertex has been found, all columns with correction patterns with ones on the same places as in the correction pattern of this vertex can be removed.

The concept of correction patterns has been improved upon by Fillion and Schiopu-Kratina (1993), who note that it is not important how the value of a variable is changed, but only whether the value of a variable is changed or not. A *generalized correction pattern* associated with column y_{*s}^k in an intermediate matrix $\mathbf{Y}^k$ is defined as the $(m+n)$-dimensional vector of which the j-th entry equals 1 if and only if an entry corresponding to the j-th variable in column y_{*s}^k is different from 0, and 0 otherwise. Here m denotes the number of categorical variables and n the number of numerical variables. For each variable involved in the error localization problem, a generalized correction pattern contains an entry with value in $\{0, 1\}$. Again Theorem 5.1 implies that once a vertex has been found, all columns with generalized correction patterns with ones on the same places as in the generalized correction pattern of this vertex can be deleted.

Fillion and Schiopu-Kratina (1993) define a *failed row* as a row that contains at least one negative entry placed on a column of which the last entry is non-zero. They note that in order to solve the error localization problem we can already terminate Chernikova's algorithm as soon as all failed rows have been processed. This result is stated as Theorem 6.1.

Theorem 6.1. *If an intermediate matrix contains no failed rows, then all (generalized) patterns corresponding to vertices for which (5.9) is minimal have been found.*

The final adaptation of Fillion and Schiopu-Kratina (1993) to Chernikova's algorithm is a method to speed-up the algorithm in case of missing values. Suppose the error localization problem has to be solved for a record with missing values. For each numerical variable of which the value is missing we first fill in an arbitrary value, say zero. Next, only the entries corresponding to variables with non-missing values are taken into account when calculating the value of function (5.9) for a column. An optimal solution to the error localization problem is given by the variables corresponding to a determined optimal generalized correction pattern plus the variables with missing values. In this way, unnecessary generalized correction patterns according to which many variables with non-missing values should be changed are discarded earlier than in the standard algorithm.

6.2 Duffin's Rules

Chernikova's algorithm does not generate any redundant columns, *i.e.*, columns whose information is already contained in another column. Its problem is, however, that in order to achieve this the algorithm requires a considerable amount of computing time. This is for a substantial part caused by its Step 7 where a time- consuming check has to be performed to prevent the generation of redundant columns. Duffin (1974) demonstrates that this step can be split into two parts. In Duffin's version of the algorithm Step 7 consists of two parts:

- For each pair (s,t) for which $y_{rs}^{k} \times y_{rt}^{k} < 0$ we choose $\mu_1, \mu_2 > 0$ such that $\mu_1 y_{rs}^{k} + \mu_2 y_{rt}^{k} = 0$ and adjoin the column $\mu_1 y_{*s}^{k} + \mu_2 y_{*t}^{k}$ to $\mathbf{Y}^{\mathbf{k+1}}$.
- Delete (some of) the redundant columns of $\mathbf{Y}^{\mathbf{k+1}}$.

Duffin (1974) gives the following two rules to delete redundant columns of $\mathbf{Y}^{\mathbf{k+1}}$.

Refined elimination rule: When t rows have been processed, delete any columns that have been generated by combining $t+2$ or more original columns.

This first rule allows the generation of redundant columns, but is much faster to apply than Step 7 of Chernikova's algorithm. The second rule, the dominance rule, makes sure that no redundant columns are generated. A column y_{*u}^{k} is called dominated by another column y_{*v}^{k} if $y_{iv}^{k} = 0$ implies $y_{iu}^{k} = 0$.

Dominance rule: Delete any column y_{*u}^{k} in $\mathbf{Y}^{\mathbf{k}}$ that is dominated by some other column y_{*v}^{k}.

One could consider using the refined elimination rule during most iterations of Chernikova's algorithm and only resort to the dominance rule when the number of columns becomes too high to be handled efficiently. After all failed rows have been processed the dominance rule has to be applied to remove redundant columns from the final matrix $\mathbf{Y}^{\mathbf{k}}$. One may hope that this leads to an algorithm that is faster than Chernikova's algorithm, but this remains to be tested.

7. DISCUSSION

At Statistics Netherlands a prototype computer program based on the adapted version of Chernikova's algorithm described in sections 5 and 6.1 of the present paper has been developed. The possibly more efficient rules described in section 6.2 have not been implemented in this prototype program. For purely numerical data a production version of this program has been used for several years in the day-to-day routine at Statistics Netherlands in order to produce clean data for most of our structural business statistics.

For Statistics Netherlands improving the efficiency of the data editing process for economic, and hence mainly numerical, data is much more important than for social, and hence mainly categorical, data. In particular, edits of type 1 (see *e.g* (2.2)) mentioned in section 2.2 are the most important ones for us, followed by edits of type 5 (see *e.g.* (2.9)). Because improving the efficiency of data editing for numerical data is much more important to us than for social data, the developed prototype program has only been evaluated for purely numerical test data. For these numerical test data, the program has been compared to several other prototype programs, namely a program based on a standard mixed integer programming problem formulation (see *e.g.* De Waal 2003), a program based on cutting planes (see Garfinkel, Kunnathur and Liepins 1988; Ragsdale and McKeown 1996, and De Waal 2003), and a program based on a branch-and-bound algorithm (see *e.g.* De Waal and Quere 2003). Our evaluation results show that the computing speed of our program based on the adapted version of Chernikova's algorithm is acceptable in comparison to other algorithms (for details on our evaluation experiments we refer to De Waal 2003). They also show, however, that this program is out-performed by the program based on the branch-and-bound algorithm. Besides being faster than the adapted version of Chernikova's algorithm, the branch-and-bound algorithm is less complex, and hence easier to maintain.

Further improvements to the adapted version of Chernikova's algorithm may reduce its computing time. Examples of such potential improvements are: better selection criteria for the row to be processed, and better ways to handle missing values. However, these improvements would at the same time increase the complexity of the algorithm, thereby making it virtually impossible for software-engineers at Statistics Netherlands to maintain the program. For the above reasons, computing time for numerical data and complexity of the algorithm, we recently decided to switch to the branch-and-bound algorithm instead of the adapted version of Chernikova's algorithm for our production software. In our latest version of our production software, a version of the branch-and-bound algorithm suitable for a mix of categorical, continuous, and integer data has been implemented. We sincerely hope, however, that the present paper will inspire some readers to find further improvements to Chernikova's algorithm.

ACKNOWLEDGEMENTS

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author would like to thank two anonymous referees for their useful comments.

APPENDIX: CHERNIKOVA'S ALGORITHM

Rubin's formulation (Rubin 1975 and 1977) of Chernikova's algorithm is as follows:

1. Construct the $(n_r + n_c) \times n_c$-matrix $\mathbf{Y}^0 = \begin{pmatrix} \mathbf{U}^0 \\ \mathbf{L}^0 \end{pmatrix}$, where $\mathbf{U}^0 = \mathbf{C}$ and $\mathbf{L}^0 = \mathbf{I}_{n_c}$: the $n_c \times n_c$-identity matrix. The j-th column of $\mathbf{Y}^0$, y^0_{*j}, will also be denoted as
$$y^0_{*j} = \begin{pmatrix} u^0_{*j} \\ l^0_{*j} \end{pmatrix},$$
where u^0_{*j} and l^0_{*j} are the j-th columns of $\mathbf{U}^0$ and $\mathbf{L}^0$, respectively.
2. k:= 0
3. If any row of $\mathbf{U}^k$ has all components negative, $\mathbf{x} = \mathbf{0}$ is the only point satisfying (5.1) and (5.2), and the algorithm terminates.
4. If all the elements of $\mathbf{U}^k$ are non-negative, the columns of $\mathbf{L}^k$ are the edges of the cone described by (5.1) and (5.2), and the algorithm terminates.
5. If neither 3 nor 4 holds: choose a row of $\mathbf{U}^k$, say row r, with at least one negative entry.
6. Let $R = \{j \mid y^k_{rj} \geq 0\}$. Let v be the number of elements in R. Then the first v columns of the new matrix $\mathbf{Y}^{k+1}$ are all the columns y^k_{*j} of $\mathbf{Y}^k$ for $j \in R$.
7. Examine the matrix $\mathbf{Y}^k$.
   a. If $\mathbf{Y}^k$ has only two columns and $y^k_{r1} \times y^k_{r2} < 0$, then choose $\mu_1, \mu_2 > 0$ such that $\mu_1, y^k_{r1} + \mu_2 y^k_{r2} = 0$. Adjoin the column $\mu_1 y^k_{*1} + \mu_2 y^k_{*2}$ to $\mathbf{Y}^{k+1}$. Go to Step 9.
   b. If $\mathbf{Y}^k$ has more than two columns then let $S = \{(s,t) \mid y^k_{rs} \times y^k_{rt} < 0 \text{ and } t > s\}$, *i.e.*, let S be the set of all pairs of columns of $\mathbf{Y}^k$ whose elements in row r have opposite signs. Let I_0 be the index set of all non-negative rows of $\mathbf{Y}^k$ *i.e.*, all rows of $\mathbf{Y}^k$ with only non-negative entries. For each $(s, t) \in S$, find all $i \in I_0$ such that $y^k_{is} = y^k_{it} = 0$. Call this set $I_1\ (s, t)$.
      - If $I_1\ (s, t) = \varnothing$, then y^k_{*s} and y^k_{*t} do not contribute another column to the new matrix.
      - If $I_1\ (s, t) \neq \varnothing$, check to see if there is a v not equal to s or t such that $y^k_{iv} = 0$ for all $i \in I_1\ (s,t)$. If such a v exists, then y^k_{*s} and y^k_{*t} do not contribute a column to the new matrix. If no such v exists, then choose $\mu_1, \mu_2 > 0$ such that $\mu_1 y^k_{rs} + \mu_2 y^k_{rt} = 0$. Adjoin the column $\mu_1 y^k_{*s} + \alpha_2 y^k_{*t}$ to $\mathbf{Y}^{k+1}$.
8. When all pairs in S have been examined, and the additional columns (if any) have been added, we say that row r has been processed. We then define matrices $\mathbf{U}^{k+1}$ and $\mathbf{L}^{k+1}$ by $\mathbf{Y}^{k+1} = \begin{pmatrix} \mathbf{U}^{k+1} \\ \mathbf{L}^{k+1} \end{pmatrix}$, where $\mathbf{U}^{k+1}$ is a matrix with n_r rows and $\mathbf{L}^{k+1}$ a matrix with n_c rows. The j-th column of $\mathbf{Y}^{k+1}$, y^{k+1}_{*j}, will also be denoted as
$$y^{k+1}_{*j} = \begin{pmatrix} u^{k+1}_{*j} \\ l^{k+1}_{*j} \end{pmatrix},$$
where u^{k+1}_{*j} and l^{k+1}_{*j} are the j-th columns of $\mathbf{U}^{k+1}$ and $\mathbf{L}^{k+1}$, respectively.
9. $k := k + 1$, and go to Step 3.

Chernikova's algorithm can be modified in order to handle equalities more efficiently than treating them as two inequalities. Steps 3, 5 and 6 should be replaced by

3. If any row of $\mathbf{U}^k$ corresponding to an inequality or equality has all components negative or if any row of $\mathbf{U}^k$ corresponding to an equality has all components positive, $\mathbf{x} = 0$ is the only point satisfying (5.1) and (5.2), and the algorithm terminates.

5. If neither 3 nor 4 holds: choose a row of $\mathbf{U}^k$, say row r, with at least one negative entry if the row corresponds to an inequality, and with at least one non-zero entry if the row corresponds to an equality.

6. If row r corresponds to an inequality, then apply Step 6 of the standard algorithm. If row r corresponds to an equality then let $R = \{j \mid y^k_{rj} = 0\}$. Let v be the number of elements in R. Then the first v columns of the new matrix $\mathbf{Y}^{k+1}$ are all the columns y^k_{*j} of $\mathbf{Y}^k$ for $j \in R$.

In Step 5 of Chernikova's algorithm a failed row has to be chosen. Rubin (1975) proposes the following simple rule. Suppose a failed row has z entries equal to zero, p positive entries, and q negative ones. We then calculate for each failed row the value $N_{\max} = z + p + pq$ if the row corresponds to an inequality and the value $N_{\max} = z + pq$ if the row corresponds to an equality, and choose a failed row with the lowest value of $N_{\max}$.

REFERENCES

CENTRAL STATISTICAL OFFICE (2000). Editing and Calibration in Survey Processing. Report SMD-37, Ireland.

CHERNIKOVA, N.V. (1964). Algorithm for finding a general formula for the non-negative solutions of a system of linear equations. *USSR Computational Mathematics and Mathematical Physics*. 4, 151-158.

CHERNIKOVA, N.V. (1965). Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Mathematical Physics*. 5, 228-233.

CHVÁTAL, V. (1983). *Linear Programming*. W.H. Freeman and Company.

DE WAAL, T. (1996). CherryPi: A Computer Program for Automatic Edit and Imputation. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Voorburg.

DE WAAL, T. (2003). Processing of Erroneous and Unsafe Data. Ph.D. Thesis, Erasmus University Rotterdam.

DE WAAL, T. and QUERE, R. (2003). A fast and simple algorithm for automatic editing of mixed data. Paper submitted to *Journal of Official Statistics*.

DUFFIN, R.J. (1974). On Fourier's analysis of linear inequality systems. *Mathematical Programming Study*. North-Holland Publishing Company. 1, 71-97.

FELLEGI, I.P., and HOLT, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*. 71, 17-35.

FILLION, J.M., and SCHIOPU-KRATINA, I. (1993). On the Use of Chernikova's Algorithm for Error Localization. Report, Statistics Canada.

GARFINKEL, R.S., KUNNATHUR A.S. and LIEPINS, G.E. (1988). Error localization for erroneous data: continuous data, linear constraints. *SIAM Journal on Scientific and Statistical Computing*. 9, 922-931.

KOVAR, J., and WHITRIDGE, P. (1990). Generalized edit and imputation system. Overview and applications. *Revista Brasileira de Estadistica*. 51, 85-100.

RAGSDALE, C.T., and MCKEOWN, P.G. (1996). On solving the continuous data editing problem. *Computers & Operations Research*. 23, 263-273.

RUBIN, D.S. (1975). Vertex generation and cardinality constrained linear programs. *Operations Research*. 23, 555-565.

RUBIN, D.S. (1977). Vertex generation methods for problems with logical constraints. *Annals of Discrete Mathematics*. 1, 457-466.

SANDE, G. (1978). An Algorithm for the Fields to Impute Problems of Numerical and Coded Data. Report, Statistics Canada.

SCHIOPU-KRATINA, I., and KOVAR, J.G. (1989). Use of Chernikova's Algorithm in the Generalized Edit and Imputation System. Report, Statistics Canada.

TODARO, T.A. (1999). Overview and Evaluation of the AGGIES Automated Edit and Imputation System. Paper presented at the UN/ECE Work Session on Statistical Data Editing, Rome.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 81-90
Statistics Canada

Inference for Population Means Under Unweighted Imputation for Missing Survey Data

DAVID HAZIZA and J.N.K. RAO[1]

ABSTRACT

In the presence of item nonreponse, unweighted imputation methods are often used in practice but they generally lead to biased estimators under uniform response within imputation classes. Following Skinner and Rao (2002), we propose a bias-adjusted estimator of a population mean under unweighted ratio imputation and random hot-deck imputation and derive linearization variance estimators. A small simulation study is conducted to study the performance of the methods in terms of bias and mean square error. Relative bias and relative stability of the variance estimators are also studied.

KEY WORDS: Bias-adjusted estimator; Item nonresponse; Random hot-deck imputation; Ratio imputation.

1. INTRODUCTION

Item nonresponse occurs when a sampled unit fails to provide information on some variables of interest. Many surveys use imputation to handle item nonresponse but one should be aware of the difficulties when imputation is used. For example, the imputed values are commonly treated as if they are true values, and the variance estimates are computed using standard formulas. This can lead to serious underestimation of the true variance of the estimators when the proportion of missing values is not small. The relationships between variables may also be distorted.

Imputation methods can be classified into two broad classes: deterministic and stochastic. Deterministic methods include ratio or regression imputation and nearest neighbour imputation, using auxiliary variables observed on all the sampled units. For nearest neighbour imputation, a nonrespondent item is assigned the respondent item value of the "nearest" respondent, where "nearest" is usually defined in terms of a distance function based on the auxiliary variables. Stochastic methods include random hot-deck imputation where the value assigned for a missing response is randomly selected from the set of respondents within an imputation cell.

In the presence of item nonresponse, weighted or unweighted imputation may be used. Weighted (deterministic or stochastic) imputation uses the sampling weights induced by the sampling design to select donors. However, weighted imputation is not feasible in practice when the sampling weights are not available at the imputation stage. Note that unweighted and weighted imputation methods lead to identical results for self-weighting designs (*i.e.*, designs with equal weights). Also, unweighted imputation methods are appealing to users.

Unweighted imputation generally leads to biased estimators under uniform response within imputation classes. Following the approach of Skinner and Rao (2002), we propose bias-adjusted estimators of population means under unweighted imputation and derive linearization variance estimators.

Let θ be a finite population parameter and $\hat{\theta}_I$ be its estimator based on the observed and imputed data respectively. Using the traditional two-phase approach: population → complete sample → sample with nonrespondents, we have

$$E(\hat{\theta}_I) = E_p\left[E_r(\hat{\theta}_I)\right], \tag{1}$$

$$V(\hat{\theta}_I - \theta) = E_p\left[V_r(\hat{\theta}_I - \theta)\right] + V_p E_r\left[(\hat{\theta}_I - \theta)\right] \tag{2}$$

under deterministic imputation, where $E_r(.)$ and $V_r(.)$ denote respectively the expectation and the variance with respect to the response mechanism given the sample, and $E_p(.)$ and $V_p(.)$ denote respectively the expectation and the variance with respect to sampling under the given design. In the model-based approach (see section 2), we replace $E_r(.)$ and $V_r(.)$ by $\tilde{E}_m(.) = E_r E_m(.)$ and $\tilde{V}_m(.) = E_r V_m(.) + V_r E_m(.)$ respectively, where $E_m(.)$ and $V_m(.)$ denote respectively the expectation and the variance with respect to the imputation model.

Fay (1991) proposed a different approach obtained by reversing the order of sampling and response: population → census with nonrespondents → sample with nonrespondents. Fay's approach facilitates variance estimation, as explained below. Using this approach, we have

$$E(\hat{\theta}_I) = E_r\left[E_p(\hat{\theta}_I)\right], \tag{3}$$

and

$$V(\hat{\theta}_I - \theta) = E_r\left[V_p(\hat{\theta}_I - \theta)\right] + V_r\left[E_p(\hat{\theta}_I - \theta)\right], \tag{4}$$

[1] David Haziza, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6.

see Shao and Steel (1999). Note that the inner expectation and variance in (4) are with respect to sampling, conditional on the response. An estimator of the overall variance $V(\hat{\theta}_I - \theta)$ in (4) is given by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\hat{\theta}_I - \theta)$ conditional on the response indicators, and v_2 is an estimator of $V_r E_p(\hat{\theta}_I - \theta)$. The estimator v_1 does not depend on the response mechanism or the assumed model, and hence v_1 is valid under either the design-based approach or the model-based approach (see section 2).

In the case of stochastic imputation, $V_p(\hat{\theta}_I - \theta)$ in (4) may be written as

$$V_p(\hat{\theta}_I - \theta) = V_p\left[E_*(\hat{\theta}_I - \theta)\right] + E_p\left[V_*(\hat{\theta}_I - \theta)\right], \qquad (5)$$

where the inner expectation and variance, E_* and V_*, denote respectively the expectation and the variance with respect to the imputation scheme given the sample with respondents and nonrespondents. An estimator of $V_p(\hat{\theta}_I - \theta)$, denoted v_1^*, is then given by $v_1^* = v_1 + v_*$ where v_1 is an estimator of $V_p E_*(\hat{\theta}_I - \theta)$ and v_* an estimator of $E_p V_*(\hat{\theta}_I - \theta)$. Also, in the case of stochastic imputation we replace $E_p(.)$ by $E_p E_*(.)$ in (4) and the formula for v_2 is the same as in the case of deterministic imputation provided $E_*(\hat{\theta}_I)$ agrees with the imputed estimator for the deterministic case. Hence, an estimator of the overall variance $V(\hat{\theta}_I - \theta)$ is given by $v_t = v_1 + v_* + v_2$.

We set out our basic framework and assumptions in section 2. In section 3, we study both weighted and unweighted ratio imputation. We show that the imputed estimator under unweighted imputation is asymptotically biased, and propose a bias-adjusted estimator. The estimator under weighted imputation and the bias-adjusted estimator under unweighted imputation are shown to be robust in the sense of validity under both the design-based and model-based approaches. We also derive linearization variance estimators of the imputed estimators in section 3. We consider the case of random hot-deck imputation in section 4. A small simulation is conducted in section 5 to compare the performances of the imputed estimators in terms of bias and mean square error. Relative bias and relative stability of the variance estimators are also studied.

2. FRAMEWORK AND ASSUMPTIONS

Let P be a finite population of possibly unknown size N. The objective is to estimate the population mean $\bar{Y} = 1/N \sum_P y_i$ when imputation has been used to compensate for nonresponse. For brevity, $\sum_A$ will be used for $\sum_{i \in A}$, where $A \subseteq P$. Suppose a probability sample, s, of size n is selected according to a specified design $p(s)$ from P. Let s_r be the set of respondents of size r and let s_m be the set of nonrespondents of size m; $r + m = n$.

Imputation is often done by first dividing the population into J nonoverlapping imputation classes and then imputing sample nonrespondents within each imputation class using sample respondents within the same class as donors, independently across the J imputation classes. For simplicity, we assume that $J = 1$; the extension to $J > 1$ imputation classes is straightforward.

The usual imputed estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_I = \frac{1}{\sum_s w_i}\left[\sum_{s_r} w_i y_i + \sum_{s_m} w_i y_i^*\right], \qquad (6)$$

where w_i is the sampling (or design) weight attached to unit i and y_i^* denotes the value imputed for missing y_i. We use the Horvitz-Thompson weight $w_i = 1/\pi_i$, where π_i is the probability of including unit i in the sample.

We consider two approaches: (i) design-based and (ii) model-based. Under the design-based approach, we assume a uniform response mechanism within classes so that the following assumption holds:

Assumption DB: Within an imputation cell, the response probability for a given variable of interest is constant and the response statuses for different units are independent.

Under the model-based approach, the following assumption holds:

Assumption MB: Within an imputation cell the response mechanism is ignorable or unconfounded in the sense that the response status of a unit does not depend on the variable being imputed but may depend on covariates used for imputation. In this case, an imputation model is assumed.

The imputation classes are chosen to make the assumption DB or MB hold approximately. The response mechanism in assumption MB is much weaker than the uniform response in assumption DB, but inferences depend on the assumed imputation model. Under ratio imputation, the imputation model used is the "ratio model" given by

$$E_m(y_i) = \beta z_i,\ V_m(y_i) = \sigma^2 z_i, \mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j, \qquad (7)$$

where β and σ^2 are unknown parameters, z_i is an auxiliary variable available for all $i \in s$. Under random hot-deck imputation, the imputation model used is given by

$$E_m(y_i) = \mu,\ V_m(y_i) = \sigma^2, \mathrm{Cov}_m(y_i, y_j) = 0 \text{ if } i \neq j. \qquad (8)$$

3. RATIO IMPUTATION

In this section, we study the properties of the imputed estimator (6) under both weighted and unweighted ratio imputation. We also derive linearization variance estimators. We study point estimation in section 3.1 under weighted and unweighted ratio imputation, and corresponding variance estimation in section 3.2.

3.1 Estimation of a Mean

3.1.1 Weighted Ratio Imputation

Weighted ratio imputation uses $y_i^* = \hat{R}_r z_i$ for missing y_i, where $\hat{R}_r = \bar{y}_r / \bar{z}_r$ and $(\bar{y}_r, \bar{z}_r) = \sum_{s_r} w_i (y_i, z_i) / \sum_{s_r} w_i$ are the weighted means of respondents for variables y and z respectively. Using the y_i^*'s, the imputed estimator (6) reduces to

$$\bar{y}_{\mathrm{IR}} = \hat{R}_r \bar{z}, \tag{9}$$

where $\bar{z} = \sum_s w_i z_i / \sum_s w_i$. It is easy to verify that $\bar{y}_{\mathrm{IR}}$ is approximately unbiased for $\bar{Y}$ under both the design-based and the model-based approaches, (Särndal 1992). Hence $\bar{y}_{\mathrm{IR}}$ is robust in the sense of validity under both approaches.

3.1.2 Unweighted Ratio Imputation

Unweighted ratio imputation uses $y_i^* = \hat{R}_r^{\mathrm{un}} z_i$ for missing y_i, where $\hat{R}_r^{\mathrm{un}} = \bar{y}_r^{\mathrm{un}} / \bar{z}_r^{\mathrm{un}}$ and $(\bar{y}_r^{\mathrm{un}}, \bar{z}_r^{\mathrm{un}}) = \sum_{s_r} (y_i, z_i) / r$ are the unweighted means of respondents for variables y and z respectively. Using the y_i^*'s, the imputed estimator (6) reduces to

$$\bar{y}_{\mathrm{IR}} = \frac{1}{\sum\limits_s w_i} \left[\sum_{s_r} w_i y_i + \hat{R}_r^{\mathrm{un}} \sum_{s_m} w_i z_i \right], \tag{10}$$

where $\hat{R}_r^{\mathrm{un}} = \bar{y}_r^{\mathrm{un}} / \bar{z}_r^{\mathrm{un}}$. Under the ratio model (7) and assumption MB, the imputed estimator (10) is approximately unbiased for $\bar{Y}$, *i.e.*, $E_r E_p E_m (\bar{y}_{\mathrm{IR}}) \approx E_m(\bar{Y})$. However, it is biased under uniform response (assumption DB). We have $E_p E_r (\bar{y}_{\mathrm{IR}}) \approx p\bar{Y} + (1-p)\ \bar{Y}_\pi / \bar{Z}_\pi \bar{Z}$, where $(\bar{Y}_\pi, \bar{Z}_\pi) = \sum_P \pi_i (y_i, z_i) / \sum_P \pi_i$. Hence, the relative bias of $\bar{y}_{\mathrm{IR}}$, $\mathrm{RB}(\bar{y}_{\mathrm{IR}}) = (E_p E_r (\bar{y}_{\mathrm{IR}}) - \bar{Y}) / \bar{Y}$, is given by

$$\mathrm{RB}(\bar{y}_{\mathrm{IR}}) \approx (1-p) \left[\frac{\bar{Z}}{\bar{Z}_\pi} \frac{\bar{Y}_\pi}{\bar{Y}} - 1 \right] \tag{11}$$

$$= (1-p) \frac{\bar{Z}}{\bar{Z}_\pi} C_\pi \left[C_y \rho_{\pi y} - C_z \rho_{\pi z} \right], \tag{12}$$

where $\bar{Z} = 1/N \sum_P z_i$, $\rho_{\pi y}$ and $\rho_{\pi z}$ are the finite population correlation coefficients between the variables π and y and π and z respectively, C_π, C_z and C_y are respectively the coefficients of variation of π, z and y, and p is the probability of response to y. The bias is nonzero generally. It vanishes in the full response case (*i.e.*, $p = 1$) or if

$$C_\pi \left[C_y \rho_{\pi y} - C_z \rho_{\pi z} \right] = 0, \tag{13}$$

which is satisfied when $C_\pi = 0$ (the case when the design is self-weighting) or when

$$\frac{\rho_{\pi y}}{\rho_{\pi z}} = \frac{C_z}{C_y}. \tag{14}$$

We further explore the relative bias (11) for three cases. First, we consider unweighted mean imputation, $y_i^* = \bar{y}_r^{\mathrm{un}}$, which is a special case of unweighted ratio imputation with $z_i = 1$. Assume that a size variable x is available for all the units in the population and that the sample s is selected according to a probability proportional to size (PPS) sampling without replacement design, using x as the size, such that $\pi_i = n x_i / X$, where $X = \sum_P x_i$. For example, one may use the well-known Sampford method (Sampford 1967). Noting that $\rho_{\pi y} = \rho_{xy}$, $\bar{Z} / \bar{Z}_\pi = 1$ and $C_\pi = C_x$, the expression (12) for the relative bias may be written as

$$\mathrm{RB}(\bar{y}_{\mathrm{IR}}) \approx (1-p)\, C_x C_y \rho_{xy}. \tag{15}$$

Two particular cases of (15) are of interest. First, if x and y are uncorrelated, the bias of the imputed estimator vanishes. The case of weakly correlated x and y (*i.e.*, $\rho_{xy} \approx 0$) may occur in surveys with multiple characteristics y (Rao 1966). Second, if $y_i \propto x_i$, the relative bias (15) reduces to $(1-p) C_x^2$ which decreases with C_x. Note that, since $C_x = C_\pi$, the sampling design approaches a self-weighting design as C_x decreases.

Consider next the more general case of unweighted ratio imputation based on $z_i, i \in s$, and PPS sampling based on $x_i, i \in s$. In this case, the relative bias (11) is zero if and only if

$$\frac{\rho_{xy}}{\rho_{xz}} = \frac{C_y}{C_z},$$

provided $p < 1$ and $C_\pi \neq 0$. If $C_y = C_z$, then the relative bias (11) is zero if and only if $\rho_{xy} = \rho_{xz}$.

Finally, we consider the case of stratified random sampling. In this case, the population P is partitioned into H strata P_h with N_h sampling units in the h-th stratum; $P = \cup_{h=1}^H P_h$, $N = \sum_{h=1}^H N_h$. We then independently select a simple random sample without replacement s_h of size n_h from each stratum; $s = \cup_{h=1}^H s_h$ and $n = \sum_{h=1}^H n_h$. Two situations may occur in practice: (1) Imputation is done independently in each stratum (*i.e.*, the imputation classes coincide with the strata). In this case, under unweighted ratio imputation, the imputed estimator is approximately unbiased under uniform response within strata. (2) The imputation is done across strata. In this case, we note from (11) that the imputed estimator is approximately unbiased if and only if $n_h = n\,(N_h / N)$ (proportional allocation).

A bias-adjusted estimator of $\bar{Y}$ under unweighted ratio imputation is given by

$$\bar{y}_{\mathrm{IR}}^a = \hat{p}^{-1} \bar{y}_{\mathrm{IR}} + \left(1 - \hat{p}^{-1}\right) \frac{\bar{z}}{\bar{z}^{\mathrm{un}}} \bar{y}_{\mathrm{IR}}^{\mathrm{un}}, \tag{16}$$

where $\hat{p} = (\sum_{s_r} w_i / \sum_s w_i)$ is a consistent estimator of the response probability p, $\bar{z}^{\mathrm{un}} = 1/n \sum_s z_i$ and $\bar{y}_{\mathrm{IR}}^{\mathrm{un}}$ is the unweighted mean of the observed values y_i and the imputed values $y_i^* = \hat{R}_r^{\mathrm{un}} z_i$. This estimator may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\bar{y}_{\mathrm{IR}}) = p\bar{Y} + \left(1 - p\right) \frac{\bar{Y}_\pi}{\bar{Z}_\pi} \bar{Z},$$

for $\bar{Y}$ and replacing $E(\bar{y}_{\mathrm{IR}})$ by its estimator $\bar{y}_{\mathrm{IR}}$, $(\bar{Y}_\pi / \bar{Z}_\pi)\bar{Z}$ by its estimator

$$\hat{R}_r^{\text{un}}\bar{z} = \left(\frac{\bar{z}}{\bar{z}^{\text{un}}}\right)\bar{y}_{\text{IR}}^{\text{un}}, \tag{17}$$

and p^{-1} by its estimator $\hat{p}^{-1}$. Note that the estimator $\bar{z}$ of $\bar{Z}$ makes use of the full sample z-values, unlike $\bar{z}_r$. If $\bar{z}_r$ is used to estimate $\bar{Z}$, then the bias-adjusted estimator requires response identifiers, unlike (16).

We now show that the bias-adjusted estimator (16) is approximately unbiased under both the design-based and the model-based approaches. Hence, unlike the unadjusted estimator (10), the adjusted estimator is robust in the sense of validity under both approaches. First, noting that $\bar{y}_{\text{IR}}$ may be expressed as $\hat{p}\,\bar{y}_r + \hat{R}_r^{\text{un}}(\bar{z} - \hat{p}\bar{z}_r)$ and using (17), the bias-adjusted estimator (16) reduces to

$$\bar{y}_{\text{IR}}^{a} = \bar{y}_r + \hat{R}_r^{\text{un}}(\bar{z} - \bar{z}_r). \tag{18}$$

Comparing (9) and (18), we see that $\bar{y}_{\text{IR}}$ under weighted ratio imputation is not equal to the bias-adjusted estimator $\bar{y}_{\text{IR}}^{a}$ under unweighted ratio imputation, unless $z_i = 1$ for all i. In the latter case, both estimators reduce to $\bar{y}_r$. However, the form (16) for $\bar{y}_{\text{IR}}^{a}$ does not require response identifiers, provided $\hat{p}$ is available.

Since $E_m(\bar{y}_{\text{IR}}^{a}) = \beta\,\bar{z}$ and $E_m(\bar{Y}) = \beta\,\bar{Z}$ under the ratio model (7), we have $E_pE_m(\bar{y}_{\text{IR}}^{a} - \bar{Y}) \approx 0$; that is, the adjusted estimator is approximately unbiased under the model-based approach. On the other hand, since $E_pE_r(\bar{y}_r) \approx \bar{Y}$ and $E_r(\bar{z} - \bar{z}_r) \approx 0$ under uniform response, it follows that $E_pE_r(\bar{y}_{\text{IR}}^{a}) \approx \bar{Y}$ so that the adjusted estimator is approximately design-unbiased under uniform response.

We note several points here: (1) The survey analyst can easily implement the adjusted estimator $\bar{y}_{\text{IR}}^{a}$, given by (16), from the imputed data file without response identifiers, *i.e.*, $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$. Note that the response identifiers are not needed on the data file, but the response rate $\hat{p}$ should be available to the analyst, which we assume to be the case here. In the case of multiple imputation classes, response rates within classes and imputation class identifiers need to be provided with the file. (2) The bias-adjusted estimator coincides with the unadjusted estimator $\bar{y}_{\text{IR}}$, given by (10), under a self-weighting design $w_i = w$. (3) The adjusted estimator $\bar{y}_{\text{IR}}^{a}$ in (18) has the form of a regression estimator in two-phase sampling. (4) Under mean imputation, (18) reduces to the weighted mean of respondents $\bar{y}_r$, so the correction made to the unadjusted estimator eliminates the effect of using unweighted mean imputation.

Another approach to getting a bias-adjusted estimator, $\bar{y}_{\text{IR}}^{a}$, is to subtract an estimator, $b(\bar{y}_{\text{IR}})$, of the bias of $\bar{y}_{\text{IR}}$, from $\bar{y}_{\text{IR}}$, *i.e.*,

$$\bar{y}_{\text{IR}}^{a} = \bar{y}_{\text{IR}} - b(\bar{y}_{\text{IR}}). \tag{19}$$

It follows from (11) that an estimator of the bias of $\bar{y}_{\text{IR}}$ is given by

$$b^{(1)}(\bar{y}_{\text{IR}}) = (1 - \hat{p})\left(\hat{R}_r^{\text{un}}\,\bar{z} - \bar{y}_r\right). \tag{20}$$

But the resulting bias-adjusted estimator is not identical to (16), and it depends on response identifiers, unlike (16). On the other hand, if one uses

$$b^{(2)}(\bar{y}_{\text{IR}}) = (1 - \hat{p})\left(\hat{R}_r^{\text{un}}\,\bar{z}_r - \bar{y}_r\right), \tag{21}$$

it is easy to verify that the resulting bias-adjusted estimator is identical to (16).

3.2 Variance estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

3.2.1 Variance Estimation under Weighted Ratio Imputation

In this subsection, we obtain a linearization variance estimator of the imputed estimator (9) based on weighted ratio imputation, using the reverse approach of Fay (1991). First, express (9) as

$$\bar{y}_{\text{IR}} = \frac{\sum_s w_i a_i y_i}{\sum_s w_i a_i z_i}\,\bar{z},$$

where a_i is a response indicator to item y such that $a_i = 1$ if $i \in s_r$ and $a_i = 0$, otherwise. It follows from (4) that the variance $V(\bar{y}_{\text{IR}})$ of $\bar{y}_{\text{IR}}$ can be estimated by $v_t = v_1 + v_2$, where v_1 is an estimator of $V_p(\bar{y}_{\text{IR}} - \bar{Y})$ conditional on the a_i's, and v_2 is an estimator of $V_rE_p(\bar{y}_{\text{IR}} - \bar{Y})$. Denote the estimator of the variance of the estimated total $\hat{Y} = \sum_s w_i y_i$ based on the full sample as $v(y_i)$. Then, using the delta method, a linearization variance estimator, v_1, in the operator notation $v(.)$, is given by

$$v_1 = v\left(\hat{\xi}\right), \tag{22}$$

where the value of $\hat{\xi}$ for $i \in s$ is given by

$$\hat{\xi}_i = \frac{1}{\sum_s w_i}\left[\hat{\xi}_{1i} - \bar{y}_{\text{IR}}\right],$$

with

$$\hat{\xi}_{1i} = a_i y_i + (1 - a_i)\hat{R}_r z_i + \hat{c}\,a_i\left(y_i - \hat{R}_r z_i\right),$$

where

$$\hat{c} = \frac{\sum_s w_i(1 - a_i)z_i}{\sum_s w_i a_i z_i}.$$

Note that v_1 is valid regardless of the response mechanism and the imputation model. The derivation of (22) is given in Appendix A. Shao and Steel (1999) derived a linearization variance estimator of the imputed estimator $\hat{Y} = \sum_s w_i a_i y_i + \sum_s w_i(1 - a_i)\hat{R}_r z_i$ of the total Y. They first expressed $\hat{Y}$ as

$$\hat{Y} = \sum_s w_i\left[a_i y_i + (1 - a_i)R_a z_i\right] + \hat{c}\sum_s w_i a_i\left(y_i - R_a z_i\right),$$

where $R_a = Y_a/Z_a$ with $(Y_a, Z_a) = \sum_P a_i(y_i, z_i)$ and then replaced $\hat{c}$ by $\tilde{c} = \sum_P(1-a_i)z_i/\sum_P a_i z_i$ to get a linear approximation for $\hat{Y} \approx \sum_s w_i \eta_i$, where

$$\eta_i = a_i y_i + (1-a_i) R_a z_i + \tilde{c}\, a_i \left(y_i - R_a z_i\right).$$

Now replacing R_a by $\hat{R}_r$ and $\tilde{c}$ by $\hat{c}$ in the above expression for η_i we get $\hat{\eta}_i = a_i y_i + (1-a_i)\hat{R}_r z_i + \hat{c}\, a_i(y_i - \hat{R}_r z_i)$ which leads to the linearization variance estimator $v_1 = v(\hat{\eta})$. The delta method in Appendix A may be used to obtain this result in a straightforward manner.

Next, using the delta method,

$$V_r E_p(\bar{y}_{\mathrm{IR}} - \bar{Y}) \approx p(1-p)\left(\frac{Z}{E_r(Z_a)}\right)^2 \frac{S_e^2}{N}, \tag{23}$$

Under asswumption DB where $Z = \sum_P z_i$, and $S_e^2 = 1/N \sum_P (y_i - E_r(R_a) z_i)^2$. The component v_2 is then obtained by substituting estimators for the unknown quantities in (23). We obtain

$$v_2 = \hat{p}(1-\hat{p}) \left(\frac{\hat{Z}}{\hat{Z}_a}\right)^2 \frac{s_{er}^2}{\hat{N}}, \tag{24}$$

where $\hat{Z} = \sum_s w_i z_i$, $\hat{Z}_a = \sum_s w_i a_i z_i$, $\hat{N} = \sum_s w_i$ and

$$s_{er}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i \left(y_i - \hat{R}_r z_i\right)^2.$$

The sum of (22) and (24) gives v_t, the estimator of the overall variance of $\bar{y}_{\mathrm{IR}}$.

3.2.2 Variance Estimation under Unweighted Ratio Imputation

We now give a linearization estimator of variance of the imputed estimator (10) based on unweighted ratio imputation. Using the delta method, see Appendix A, we obtain

$$v_1 = v\left(\hat{\xi}\right), \tag{25}$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i} \left[\hat{\xi}_{1i} - \bar{y}_{\mathrm{IR}}\right],$$

with

$$\hat{\xi}_{1i} = a_i y_i + (1-a_i)\hat{R}_r^{\mathrm{un}} z_i + \hat{d}\, \frac{a_i}{w_i}\left(y_i - \hat{R}_r^{\mathrm{un}} z_i\right)$$

and $\hat{d} = \sum_s w_i(1-a_i)z_i/\sum_s a_i z_i$. The component v_2 is given by (B.2) in Appendix B.

3.2.3 Variance Estimation for the Bias-Adjusted Estimator

In this subsection, we give a linearization variance estimator of the bias-adjusted estimator (18). Using the delta method, we obtain

$$v_1 = v\left(\hat{\xi}\right), \tag{26}$$

where

$$\hat{\xi}_i = \frac{a_i}{\sum_s w_i a_i}\left[(y_i - \bar{y}_r) + \hat{R}_r^{\mathrm{un}}(z_i - \bar{z}_r)\right] + \frac{\hat{R}_r^{\mathrm{un}}}{\hat{N}}(z_i - \bar{z})$$

$$+ (\bar{z} - \bar{z}_r)\frac{1}{\sum_s a_i z_i}\frac{a_i}{w_i}\left(y_i - \hat{R}_r^{\mathrm{un}} z_i\right);$$

see Appendix A. The component v_2 is given by (C.2) in Appendix C.

4. RANDOM HOT-DECK IMPUTATION

In this section, we study the properties of the imputed estimator (6) under weighted and unweighted random hot-deck imputation. We also derive linearization variance estimators under uniform response.

4.1 Estimation of a Mean

In section 4.1 we study point estimators under weighted and unweighted random hot-deck imputation.

4.1.1 Weighted Random Hot-Deck Imputation

Under weighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with selection probabilities $w_j/\sum_{s_r} w_i$ and use $y_i^* = y_j, i \in s_m$. The imputed estimator, $\bar{y}_{\mathrm{IH}}$, is given by (6) with the above imputed values. It is approximately unbiased for the population mean $\bar{Y}$ under both the design-based and the model-based approaches. The latter uses the mean model (8).

4.1.2 Unweighted Random Hot-Deck Imputation

Under unweighted random hot-deck imputation, we select the donors $j \in s_r$ with replacement with equal probabilities $1/r$ and use $y_i^* = y_j, i \in s_m$. The imputed estimator, $\bar{y}_{\mathrm{IH}}$, is given by (6) with the above imputed values. It is approximately unbiased for $\bar{Y}$ under the mean model (8), but biased under uniform response. The bias of $\bar{y}_{\mathrm{IH}}$ is given by

$$\mathrm{B}(\bar{y}_{\mathrm{IH}}) = (1-p)(\bar{Y}_\pi - \bar{Y}). \tag{27}$$

A bias-adjusted estimator of $\bar{Y}$ under unweighted random hot-deck imputation is given by

$$\bar{y}_{\mathrm{IH}}^{a} = \hat{p}^{-1}\,\bar{y}_{\mathrm{IH}} + (1-\hat{p}^{-1})\,\bar{y}_{\mathrm{IH}}^{\mathrm{un}}, \tag{28}$$

where $\hat{p} = (\sum_{s_r} w_i/\sum_s w_i)$ is a consistent estimator of the response probability p and $\bar{y}_{\mathrm{IH}}^{\mathrm{un}}$ is the unweighted mean of the observed values y_i and the imputed values y_i^*. The estimator (28) may be derived from the method of moments, following Skinner and Rao (2002), by solving

$$E(\bar{y}_{\mathrm{IH}}) = p\,\bar{Y} + (1-p)\,\bar{Y}_\pi$$

for $\bar{Y}$ replacing by $E\,(\bar{y}_{\mathrm{IH}})$ its estimator $\bar{y}_{\mathrm{IH}}$, $\bar{Y}_\pi$ by its estimator $\bar{y}_{\mathrm{IH}}^{\mathrm{un}}$ and p^{-1} by its estimator $\hat{p}^{-1}$. The adjusted estimator is approximately unbiased for $\bar{Y}$ under both the design-based and the model-based approaches. As in section 3.1.2, note that the survey analyst can easily implement the adjusted estimator $\bar{y}_{\text{IR}}^{a}$ from the imputed data file without response identifiers, *i.e.*, $(w_i, \tilde{y}_i, z_i, i \in s)$, where $\tilde{y}_i = y_i$ if $i \in s_r$ and $\tilde{y}_i = y_i^*$ if $i \in s_m$, provided the response rate, $\hat{p}$, is available.

Note that the method of subtracting an estimator of the bias of $\bar{y}_{\text{I}}$ from $\bar{y}_{\text{I}}$, using (27), will lead to a bias-adjusted estimator that depends on response identifiers, unlike (28). It is not possible to obtain the bias-adjusted estimator (28) by this approach, unlike in the case of deterministic ratio imputation studied in subsection 3.1.2.

4.2 Variance Estimation

We study variance estimation under uniform response in this section. We assume that response identifiers are available with the variance estimation file. If imputation classes are used, their identifiers are also needed.

4.2.1 Variance Estimation under Weighted Random Hot-Deck Imputation

We now obtain a linearization variance estimator of the imputed estimator $\bar{y}_{\text{IH}}$ under weighted random hot-deck imputation. First, note that under weighted random hot-deck imputation, $E_*(\bar{y}_{\text{IH}}) = \bar{y}_r$. This is a particular case of (9) with $z_i = 1$ for all i. Hence, using (22), v_1 is given by

$$v_1 = v\left(\hat{\xi}\right), \tag{29}$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i} \left[\hat{\xi}_{1i} - \bar{y}_r\right],$$

$$\hat{\xi}_i = a_i y_i + (1 - a_i)\,\bar{y}_r + \hat{c}\, a_i \left(y_i - \bar{y}_r\right),$$

with $\hat{c} = \sum_s w_i (1 - a_i) / \sum_s w_i a_i$. Straightforward algebra shows that $\hat{\xi}_i$ simplifies to $\hat{\xi}_i = a_i (y_i - \bar{y}_r) / \sum_s w_i a_i$. Now, noting that $V_*(y_i^*) = (1/\sum_s w_i a_i) \sum_s w_i a_i (y_i - \bar{y}_r)^2 = s_{yr}^2$, we have

$$v_* = V_*(\bar{y}_{\text{IH}} - \bar{Y}) = \frac{\sum_s w_i^2 (1 - a_i)}{\left(\sum_s w_i\right)^2} s_{yr}^2. \tag{30}$$

As noted in section 1, v_2 is the same as for the deterministic case. Hence, under weighted random hot-deck imputation, v_2 is given by (24) with $z_i = 1$ for all i, which leads to

$$v_2 = \hat{p}(1 - \hat{p}) \left(\frac{\hat{N}}{\sum_s w_i a_i}\right)^2 \frac{s_{yr}^2}{\hat{N}}. \tag{31}$$

The sum of (29), (30) and (31) gives v_t, the estimator of overall variance.

4.2.2 Variance Estimation under Unweighted Random Hot-Deck Imputation

We now obtain a linearization estimator of variance of the imputed estimator (6) under unweighted random hot-deck imputation. First, note that $E_*(\bar{y}_{\text{IH}})$ reduces to (10) with $z_i = 1$ for all i. Hence, v_1 is given by

$$v_1 = v\left(\hat{\xi}\right), \tag{32}$$

where

$$\hat{\xi}_i = \frac{1}{\sum_s w_i} \left[\hat{\xi}_{1i} - E_*\left(\bar{y}_{\text{IH}}\right)\right],$$

$$\hat{\xi}_{1i} = a_i y_i + (1 - a_i)\,\bar{y}_r^{\text{un}} + \hat{d}\frac{a_i}{w_i}\left(y_i - \bar{y}_r^{\text{un}}\right),$$

with $\hat{d} = \sum_s w_i (1 - a_i) / \sum_s a_i$. Now, nothing that $V_*(y_i^*) = 1/r \sum_s a_i (y_i - \bar{y}_r^{\text{un}})^2 s_{yr}^{2\text{un}}$, we have

$$v_* = \frac{\sum_s w_i^2 (1 - a_i)}{\left(\sum_s w_i\right)^2} s_{yr}^{2\text{un}}. \tag{33}$$

As noted in section 1, v_2 is the same as for the deterministic case. Hence, under unweighted random hot-deck imputation, v_2 is given by (B.2) with $z_i = 1$ for all i. The sum of (32), (33) and (B.2) gives v_t.

4.2.3 Variance Estimation for the Bias-Adjusted Estimator

We now obtain a linearization variance estimator of the bias-adjusted estimator given by (28). First, note that, $E_*(\bar{y}_{\text{IH}}^{a})$ reduces to $\bar{y}_r$, the mean of the y-values respondent. Hence, v_1 is given by (29) and v_2 is given by (31). Now, noting that $V_*(y_i^*) = s_{yr}^{2\text{un}}$ and $\text{Cov}_*(y_i^*, y_j^*) = 0$ for $i \neq j$, one can show that $V_*(\bar{y}_{\text{IH}}^{a} - \bar{Y})$ is given by

$$v_* = \left[\frac{\hat{p}^{-2}}{\left(\sum_s w_i\right)^2} \sum_s w_i^2 (1 - a_i) - (1 - \hat{p}^{-1})^2 \left(\frac{r + n}{n^2}\right)\right] s_{yr}^{2\text{un}}. \tag{34}$$

The sum of (29), (31) and (34) gives v_t. Note that even though v_* given by (34) is expressed as the difference between two terms, it is always nonegative, as shown in Appendix D.

5. SIMULATION STUDY

As a complement to the theory, we present some results from a limited simulation study. We generated a population of $N = 800$ values (y_i, z_i) according to the ratio model $y = \beta z + \varepsilon$, where z and ε were generated from a normal distribution such that the correlation, ρ_{yz}, between y and z equaled 0.05, 0.30, 0.70 and 0.90. The objective is to estimate the population total $Y = \sum_p y_i$. We drew $R = 10{,}000$ PPS samples, each of size $n = 75$, according to Sampford's pps sampling method, using item z as the measure of size. Nonresponse to item y was then generated from each PPS sample according to a uniform response mechanism with a response rate of 0.7; item z was observed for all units in the sample. We used weighted and unweighted random hot-deck imputation to compensate for nonresponse to item y.

The estimator of the first component in the variance formula (4) was computed using the well known Sen-Yates-Grundy estimator. Let $v(\xi)$ denote the variance estimator of $\sum_s w_i \xi_i$. The Sen-Yates-Grundy estimator of variance is then given by

$$v(\xi) = \frac{1}{2} \sum_{i \in s} \sum_{j \in s} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{\xi_i}{\pi_i} - \frac{\xi_j}{\pi_j} \right)^2, \qquad (35)$$

where $\pi_{ij} = P(i \in s \text{ and } j \in s)$ is the joint probability of inclusion of units i and j in the sample. Sampford's method ensures $\pi_i \pi_j - \pi_{ij} \geq 0$ for all i, j so that the variance estimator in (35) is always nonnegative.

As a measure of the bias of an imputed estimator $\hat{Y}_I$ of Y, we used the bias ratio $B_r(\hat{Y}_I) = \text{Bias}(\hat{Y}_I) / \text{s.e.}(\hat{Y}_I)$, where $\text{s.e.}(\hat{Y}_I)$ denotes the standard error of $\hat{Y}_I$. To compare the efficiencies, we used the coefficient of variation of $\hat{Y}_I$, denoted $\text{CV}(\hat{Y}_I)$ and given by $\text{CV}(\hat{Y}_I) = (\sqrt{\text{MSE}}/Y)$. The variance estimators were compared in terms of their relative bias and CV. The relative bias of a variance estimator, v_t, is measured by $B_{\text{rel}}(v_t) = (E(v) - \text{MSE}(\hat{Y}_I))/\text{MSE}(\hat{Y}_I)$ and its CV is given by $\text{CV}(v_t) = \sqrt{\text{MSE}}(v_t)/\text{MSE}(\hat{Y}_I)$. Values of the above measures were calculated from the simulated PPS samples.

Table 1 reports the simulation results on the bias ratio (B_r) of the three imputed estimators of Y, denoted B_r (weighted), B_r (unweighted) and B_r (adjusted) and the CVs of the estimators, denoted CV (weighted), CV (unweighted) and CV (adjusted). It is clear from Table 1 that the bias ratio of the estimator under unweighted imputation is large ($\geq$ 30%) if $\rho_{xy} \geq 0.5$, while the bias ratios of the estimator under weighted imputation and the adjusted estimator are small ($\leq$4%) for all values of ρ_{xy}. Due to large bias, the CV of the unweighted estimator is larger than the CV of the weighted estimator if $\rho_{xy} \geq 0.5$ and also larger than the CV of the adjusted estimator if $\rho_{xy} \geq 0.7$, but the increase in CV is not large. Also, CV (weighted) is slightly smaller than CV (adjusted) for all values of ρ_{xy}.

Table 2 reports the relative bias (B_{rel}) and the CV ratios of the variance estimators. As expected, the variance estimator v_t (unweighted) leads to serious underestimation of MSE of the estimator for large ρ_{xy} (≥ 0.7), while the absolute relative bias of the variance estimators v_t (weighted) and v_t (adjusted) is small ($\leq$ 6%) for all values of ρ_{xy}. Turning to the CV ratios of the variance estimators, Table 2 shows that v_t (unweighted) has the smallest CV followed by v_t (weighted) and v_t (adjusted) for $\rho_{xy} \geq 0.3$.

Table 1
Bias Ratio (%) and CV (%) of the Imputed Estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
B_r(weighted)	-0.78	1.99	-0.79	0.40	3.27
B_r(unweighted)	1.82	18.60	30.50	49.20	64.20
B_r(adjusted)	-1.12	1.47	0.01	0.61	2.94
CV(weighted)	18.80	15.30	11.60	5.87	4.69
CV(unweighted)	18.00	15.20	12.50	6.83	5.93
CV(adjusted)	20.90	16.80	13.50	6.10	4.78

Table 2
Relative Bias (%) of the Variance Estimators and Comparisons of the CV ratios of the Variance Estimators

	$\rho_{xy} = 0.05$	$\rho_{xy} = 0.30$	$\rho_{xy} = 0.50$	$\rho_{xy} = 0.70$	$\rho_{xy} = 0.90$
$B_{\text{rel}}(v_t)$(weighted)	-2.43	-4.78	-4.28	3.96	-1.95
$B_{\text{rel}}(v_t)$(unweighted)	-1.03	-3.47	-11.80	-18.50	-29.30
$B_{\text{rel}}(v_t)$(adjusted)	-5.42	-1.06	-4.21	1.61	0.07
$\frac{\text{CV}(v_t)(\text{unweighted})}{\text{CV}(v_t)(\text{weighted})}$	1.016	0.984	0.931	0.875	0.781
$\frac{\text{CV}(v_t)(\text{unweighted})}{\text{CV}(v_t)(\text{adjusted})}$	1.032	0.829	0.701	0.819	0.692
$\frac{\text{CV}(v_t)(\text{weighted})}{\text{CV}(v_t)(\text{adjusted})}$	1.016	0.843	0.751	0.935	0.886

6. CONCLUDING REMARKS

Unweighted imputation methods are often used in practice to compensate for item nonresponse when the survey weights are not available at the imputation stage. Also, unweighted imputation is appealing to users even when the weights are available at the imputation stage. But it leads to biased estimators under uniform response within imputation classes. We have proposed bias-adjusted estimators under ratio imputation and random hot-deck imputation. These estimators can be implemented from the imputed data file, even if the imputation flags within classes are not given, provided estimates of response rates within classes are reported. We have shown that the bias-adjusted estimator performs better than the unadjusted estimator under unweighted imputation, and is robust in the sense of validity under both the frequentist and model-based approaches.

We have obtained linearization variance estimators for the bias-adjusted estimators. For variance estimation, imputation flags should be provided in the variance estimation file.

If the imputation flags are available in the data file and imputation is deterministic, the imputed values can be replaced by those under weighted imputation. For example, in the case of unweighted ratio imputation, $y_i^* = \bar{y}_r^{\text{un}} / \bar{z}_r^{\text{un}} z_i$, one could either multiply each imputed value by $\bar{z}_r^{\text{un}} / \bar{y}_r^{\text{un}} \times \bar{y}_r / \bar{z}_r$ to reproduce the values $\bar{y}_r / \bar{z}_r \; z_i$ under weighted ratio imputation, provided edits are not applied after imputation. Alternatively, one could reimpute values using the sampling weights w_i. In both cases, the adjusted estimator does not present advantages over the imputed estimator based on weighted imputation other than assuring that the imputed values in the data file are not changed.

In the case of random hot-deck imputation, however, the only way to implement weighted random hot-deck imputation is to reimpute using a weighted hot-deck scheme. We believe that analysts do not like to change the imputed values on the data file produced by the edit and imputation system.

The imputed estimator (10) can use poststratification (or calibration) weights, $\tilde{w}_i(s)$, based on known population auxiliary information, instead of design weights w_i. Note that the calibration weights, $\tilde{w}_i(s)$, depend on the whole sample s unlike the design weights w_i. If the calibration weights are used for ratio imputation, then we simply replace w_i by $\tilde{w}_i(s)$ in section 3.1.1 and the resulting linearization variance estimator, v_1, uses $\hat{\xi}$ in (22) with w_i changed to $\tilde{w}_i(s)$. However, $v(.)$ in (22) now refers to the linearization variance estimator of the full sample post-stratified estimator $\sum_s \tilde{w}_i(s) y_i$.

Under unweighted imputation, linearization variance estimation becomes more complex because the bias-adjusted estimator based on the calibration weights will involve both design weights and calibration weights. If the design weights, w_i, are available at the imputation stage but not the calibration weights, $\tilde{w}_i(s)$, the design weights can be used for imputation and the calibration weights for estimation. The resulting imputed estimator (6) based on calibration weights remains asymptotically unbiased under uniform response (within classes), but linearization variance estimation becomes more complex because both sets of weights are involved in the imputed estimator. We propose to study poststratification and some other extensions in a separate paper, and derive corresponding linearization variance estimators.

ACKNOWLEDGEMENT

The authors would like to thank the referee, the Associate Editor and Jae Kim of Hankuk University, Korea, for useful comments and constructive suggestions. David Haziza would also like to thank Jean-Francois Beaumont and Eric Rancourt of Statistics Canada for their encouragement and valuable discussions.

APPENDIX

A. Derivation of v_1

Suppose that an estimator $\hat{\theta}$ is expressed as

$$\hat{\theta} = \frac{1}{\hat{Y}_1}\left[\hat{Y}_2 + \frac{\hat{Y}_3}{\hat{Y}_4}\left(\hat{Y}_5 - \hat{Y}_6\right)\right] =: g(\hat{Y}), \qquad \text{(A.1)}$$

where $\hat{Y}_j = \sum_s w_i y_{ji}, j = 1, ..., 6$ and $\hat{\mathbf{Y}} = (\hat{Y}_1, ..., \hat{Y}_6)'$. Letting $\theta = g(\hat{\mathbf{Y}})$, $R_{34} = Y_3 / Y_4, \hat{Y}_j = Y_j(1 + \delta \hat{Y}_j)$ with $\delta \hat{Y}_j = (\hat{Y}_j - Y_j)/Y_j$ and $Y_j = E_p(\hat{Y}_j)$, we have

$$\hat{\theta} - \theta = \frac{1}{Y_1\left(1 + \delta \hat{Y}_1\right)}\left\{Y_2\left(1 + \delta \hat{Y}_2\right)\right.$$

$$\left. + R_{34}\frac{\left(1 + \delta\hat{Y}_3\right)}{\left(1 + \delta\hat{Y}_4\right)}\left[Y_5\left(1 + \delta\hat{Y}_5\right) - Y_6\left(1 + \delta\hat{Y}_6\right)\right]\right\} - \theta$$

$$\approx \frac{1}{Y_1}\left\{\left(\delta \hat{Y}_2 - \delta\hat{Y}_1\right) Y_2 + R_{34} Y_5\left(\delta \hat{Y}_3 - \delta\hat{Y}_4 + \delta\hat{Y}_5 - \delta\hat{Y}_1\right) - \right.$$

$$\left. R_{34} Y_6\left(\delta \hat{Y}_3 - \delta \hat{Y}_4 + \delta \hat{Y}_6 - \delta\hat{Y}_1\right)\right\}, \qquad \text{(A.2)}$$

neglecting higher order terms in $\delta \hat{Y}_j$'s. The expression (A.2) reduces to

$$\hat{\theta} - \theta \approx \frac{1}{Y_1}\left\{\hat{Y}_2 + R_{34}\left(\hat{Y}_5 - \hat{Y}_6\right) + \frac{Y_5 - Y_6}{Y_4}\left(\hat{Y}_3 - R_{34}\hat{Y}_4\right) - \theta\hat{Y}_1\right\}$$

$$= \sum_s w_i \xi_i,$$

where

$$\xi_i = \frac{1}{Y_1}\left(\xi_{1i} - \theta\right) \tag{A.3}$$

with

$$\xi_{1i} = y_{2i} + R_{34}\left(y_{5i} - y_{6i}\right) + \frac{Y_5 - Y_6}{Y_4}\left(y_{3i} - R_{34}\, y_{4i}\right).$$

Hence, the variance estimator of $\hat{\theta}$ from the delta method may be expressed as $v(\xi)$. Now, replacing unknown quantities in (A.3) by their estimators, we get

$$\text{estvar}\,(\hat{\theta}) = v\,(\hat{\xi}),$$

where

$$\hat{\xi}_i = \frac{1}{\hat{Y}_1}\left(\hat{\xi}_{1i} - \hat{\theta}\right)$$

with

$$\hat{\xi}_{1i} = y_{2i} + \hat{R}_{34}\left(y_{5i} - y_{6i}\right) + \frac{\hat{Y}_5 - \hat{Y}_6}{\hat{Y}_4}\left(y_{3i} - \hat{R}_{34}\, y_{4i}\right).$$

Note that the delta method avoids evaluation of partial derivatives of $g(\hat{\mathbf{Y}})$ with respect to its components Y_j, unlike the usual Taylor linearization method.

Letting $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_s w_i a_i y_i$, $\hat{Y}_4 = \hat{Y}_6 = \sum_s w_i a_i z_i$ and $\hat{Y}_5 = \sum_s w_i z_i$ in (A.1), we get the variance estimator (22) of $\bar{y}_{\text{IR}}$ based on weighted ratio imputation. Also, letting $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \sum_s w_i a_i y_i$, $\hat{Y}_3 = \sum_s w_i\, a_i (y_i / w_i)$, $\hat{Y}_4 = \sum_s w_i a_i (z_i / w_i)$, $\hat{Y}_5 = \sum_s w_i z_i$ and $\hat{Y}_6 = \sum_s w_i a_i z_i$ in (A.1), we get the variance estimator (25) of $\bar{y}_{\text{IR}}$ based on unweighted imputation. Finally, we note that the bias-adjusted estimator (16) written in the form (18) can be expressed as the sum of three components: $\bar{y}_r$, $\hat{R}_r^{\text{un}}\, \bar{z}$ and $-\hat{R}_r^{\text{un}}\, \bar{z}_r$. Each of these components is a special case of (A.1). Indeed, the component $\bar{y}_r$ is a special case of (A.1) with $\hat{Y}_1 = \sum_s w_i a_i$, $\hat{Y}_2 = \sum_s w_i a_i y_i$ with $\hat{Y}_5 = \hat{Y}_6$. The component $\hat{R}_r^{\text{un}}\, \bar{z}$ is a special case of (A.1) with $\hat{Y}_1 = \sum_s w_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_s w_i a_i (y_i / w_i)$, $\hat{Y}_4 = \hat{Y}_6 = \sum_s w_i a_i\, (z_i / w_i)$ and $\hat{Y}_5 = \sum_s w_i z_i$. The component $\hat{R}_r^{\text{un}}\, \bar{z}_r$ is a special case of (A.1) with $\hat{Y}_1 = \sum_s w_i a_i$, $\hat{Y}_2 = \hat{Y}_3 = \sum_s w_i a_i (y_i / w_i)$, $\hat{Y}_4 = \hat{Y}_6 = \sum_s w_i a_i (z_i / w_i)$ and $\hat{Y}_5 = \sum_s w_i a_i z_i$. We apply the delta method to each component separately to obtain $v_1 = v(\xi)$ given by (26).

B. Derivation of v_2 for the estimator $\bar{y}_{\text{IR}}$ under unweighted imputation

Using the delta method, it can be shown that $V_r E_p\, (\bar{y}_{\text{IR}}^{(1)} - \bar{Y})$ under unweighted ratio imputation is given by

$$V_r E_p (\bar{y}_{\text{IR}} - \bar{Y}) \approx p\left(1 - p\right) \frac{1}{N}$$

$$\times \left[S_{e(1)}^2 + \left(\frac{Z - E_r(Z_a)}{E_r(Z_{\pi a})} \right)^2 S_{e(2)}^2 + 2 \left(\frac{Z - E_r(Z_a)}{E_r(Z_{\pi a})} \right) S_{e(3)}^2 \right], \tag{B.1}$$

where

$$S_{e(1)}^2 = \frac{1}{N} \sum_P \left(y_i - E_r\, (R_{\pi a}) z_i\right)^2,$$

$$S_{e(2)}^2 = \frac{1}{N} \sum_P \pi_i^2 \left(y_i - E_r\, (R_{\pi a}) z_i\right)^2,$$

$$S_{e(3)}^2 = \frac{1}{N} \sum_P \pi_i \left(y_i - E_r\, (R_{\pi a}) z_i\right)^2,$$

with $R_{\pi a} = Y_{\pi a} / Z_{\pi a}$ and $(Y_{\pi a}, Z_{\pi a}) = \sum_P \pi_i a_i (y_i, z_i)$. The component v_2 is obtained by estimating unknown quantities in (B.1). It is given by

$$v_2 \approx \hat{p}\,(1 - \hat{p})$$

$$\times \frac{1}{\hat{N}} \left[s_{er(1)}^2 + \left(\frac{\hat{Z} - \hat{Z}_a}{\hat{Z}_{\pi a}} \right)^2 s_{er(2)}^2 + 2 \left(\frac{\hat{Z} - \hat{Z}_a}{\hat{Z}_{\pi a}} \right) s_{er\,(3)}^2 \right], \tag{B.2}$$

where

$$s_{er(1)}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i \left(y_i - \hat{R}_r^{\text{un}}\, z_i\right)^2,$$

$$s_{er(2)}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i^{-1} a_i \left(y_i - \hat{R}_r^{\text{un}}\, z_i\right)^2,$$

$$s_{er(3)}^2 = \frac{1}{\sum_s w_i a_i} \sum_s a_i \left(y_i - \hat{R}_r^{\text{un}}\, z_i\right)^2,$$

and $\hat{Z}_{\pi a} = \sum_s a_i z_i$.

C. Derivation of v_2 for the estimator $\bar{y}_{\text{IR}}^{a}$

Using the delta method, it can be shown that $V_r E_P (\bar{y}_{\text{IR}}^{a} - \bar{Y})$ for the bias-adjusted estimator is given by

$$V_r E_p\, (\bar{y}_{\text{IR}}^{a} - \bar{Y}) \approx p\,(1 - p) \frac{N}{E_r\left[\left(\sum_P a_i\right)^2\right]}$$

$$\times \Bigg\{ S_{ay}^2 + E_r(R_{\pi a}^2) S_{az}^2 - 2 E_r(R_{\pi a}) S_{ayz} + E_r\left[h \sum_P a_i\right]^2 S_{e\,(2)}^2$$

$$+ 2 E_r\left[h \sum_P a_i\right] \left(S_{\pi e y} - E_r(R_{\pi a}) S_{\pi e z}\right) \Bigg\}, \tag{C.1}$$

where

$$S_{ay}^2 = \frac{1}{N} \sum_P \left(y_i - E_r\,(\bar{Y}_a)\right)^2,$$

$$S_{az}^2 = \frac{1}{N} \sum_P \left(z_i - E_r(\bar{Z}_a)\right)^2,$$

$$S_{ayz} = \frac{1}{N} \sum_P \left(y_i - E_r(\bar{Y}_a)\right)\left(z_i - E_r(\bar{Z}_a)\right),$$

$$S_{e(2)}^2 = \frac{1}{N} \sum_P \pi_i^2 \left(y_i - E_r(R_{\pi a}) z_i\right)^2,$$

$$S_{\pi e y} = \frac{1}{N} \sum_P \pi_i \left(y_i - E_r\,(\bar{Y}_a)\right)\left(y_i - E_r(R_{\pi a}\,) z_i\right),$$

$$S_{\pi e z} = \frac{1}{N} \sum_P \pi_i \left(z_i - E_r(\bar{Z}_a)\right)\left(y_i - E_r(R_{\pi a}) z_i\right),$$

$(\bar{Y}_a, \bar{Z}_a) = \sum_P a_i (y_i, z_i) / \sum_P a_i$ and $h = (\bar{Z} - \bar{Z}_a)/Z_{\pi a}$. The component v_2 is obtained by estimating unknown quantities in (C.1). It is given by

$$v_2 \approx \hat{p}(1-\hat{p}) \frac{\hat{N}}{\sum_s w_i a_i} \times \left\{ s_{yr}^2 + \left(\hat{R}_r^{\text{un}}\right)^2 s_{zr}^2 - 2\hat{R}_r^{\text{un}} s_{yzr} + \left(\hat{h}\sum_s w_i a_i\right)^2 s_{er(2)}^2 + 2\left(\hat{h}\sum_s w_i a_i\right)\left(s_{eyr} - \hat{R}_r^{\text{un}} s_{ezr}\right)\right\}, \tag{C.2}$$

where

$$s_{yr}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i \left(y_i - \bar{y}_r\right)^2,$$

$$s_{zr}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i \left(z_i - \bar{z}_r\right)^2,$$

$$s_{yzr} = \frac{1}{\sum_s w_i a_i} \sum_s w_i a_i \left(y_i - \bar{y}_r\right)\left(z_i - \bar{z}_r\right),$$

$$s_{er(2)}^2 = \frac{1}{\sum_s w_i a_i} \sum_s w_i^{-1} a_i \left(y_i - \hat{R}_r^{\text{un}} z_i\right)^2,$$

$$s_{eyr} = \frac{1}{\sum_s w_i a_i} \sum_s a_i \left(y_i - \bar{y}_r\right)\left(y_i - \hat{R}_r^{\text{un}} z_i\right),$$

$$s_{ezr} = \frac{1}{\sum_s w_i a_i} \sum_s a_i \left(z_i - \bar{z}_r\right)\left(y_i - \hat{R}_r^{\text{un}} z_i\right),$$

and $\hat{h} = (\bar{z} - \bar{z}_r)/\sum_s a_i z_i$.

D. Nonegativity of $V_*(\bar{y}_{\text{IH}}^a - \bar{Y})$

We show that the variance formula in (34) is always nonnegative. First, note that this expression can be expressed as

$$V_*(\bar{y}_{\text{IH}}^a - \bar{Y}) = \frac{n^2 \sum_{s_m} w_i^2 - (r+n)\left(\sum_{s_m} w_i\right)^2}{n^2 \left(\sum_{s_r} w_i\right)^2} \geq 0$$

$$\Leftrightarrow n^2 \sum_{s_m} w_i^2 - (r+n)\left(\sum_{s_m} w_i\right)^2 \geq 0$$

$$\Leftrightarrow n^2 \sum_{s_m} w_i^2 - m(r+n)\frac{\left(\sum_{s_m} w_i\right)^2}{m} \geq 0.$$

On one hand, $n^2 \geq m(r+n) \Leftrightarrow n \geq m$ which is always true. On the other hand, using Cauchy-Schwarz inequality, it is easily seen that $\sum_{s_m} w_i^2 \geq (\sum_{s_m} w_i)^2/m$. The result follows.

REFERENCES

FAY, R.E. (1991). A Design-Based Perspective on Missing Data Variance. *Proceedings of the 1991 Annual Research Conference, US Bureau of the Census*. 429-440.

RAO, J.N.K. (1966). Alternative estimators in pps sampling for multiple characteristics. *Sankhyā*, Series A. 28, Part 1. 47-59.

SAMPFORD, M.R. (1967). On sampling without replacement with unequal probabilities of selection. *Biometrika*. 54, 499-513.

SHAO, J., and STEEL, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*. 94, 254-265.

SÄRNDAL, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*. 18, 241-252.

SKINNER, C.J., and RAO, J.N.K. (2002). Jackknife variance for multivariate statistics under hot deck imputation from common donors. *Journal of Statistical Planning and Inference*. 102, 149-167.

Minimum Risk, Fixed Cost Sampling Designs for Independent Poisson Processes

BRAD C. JOHNSON and JOHN DEELY[1]

ABSTRACT

Optimal and approximately optimal fixed cost Bayesian sampling designs are considered for simultaneous estimation in independent homogeneous Poisson processes. We develop general allocation formulae for a basic Poisson-Gamma model and compare these with more traditional allocation methods. We then discuss techniques for finding representative gamma priors under more general hierarchical models and show that, in many practical situations, these provide reasonable approximations to the hierarchical prior and Bayes risk. The methods developed are general enough to apply to a wide variety of models and are not limited to Poisson Processes.

KEY WORDS: Optimal sampling allocations; Poisson processes; Poisson-Gamma hierarchy.

1. INTRODUCTION

The topic of Bayesian survey sampling techniques is well represented in the literature. A number of articles focus on sampling from finite populations and most make use of normality or a "posterior linearity" property (*cf.* Godambe 1955; Ericson 1988; Ericson 1969; Scott and Smith 1971; Tiwari and Lahiri 1989). An excellent review of recent Bayesian methods for sampling finite populations is contained in (Ghosh and Meeden 1997) as well as some interesting new approaches. Lindley and Deely (1993) discuss optimal allocation in stratified sampling under a normal model when only partial information is available. In terms of Poisson models, Clevenson and Zidek (1975) discuss the simultaneous estimation of means in independent Poisson processes and Leite, Rodrigues and Milan (2000) discuss a Bayesian analysis when estimating the number of species in a population using a non-homogeneous Poisson process. Little work has been done on model specific sampling designs from a Bayesian perspective.

In the present paper we take a model based approach to develop optimal and approximately optimal fixed cost sampling allocations for simultaneous estimation in multiple independent Poisson processes. Section 2 introduces the model and some notation. Section 3 presents the general allocation problem and gives the minimum Bayes risk allocations when independent conjugate gamma priors are assumed for each process. Comparisons are made with classical stratified random sampling allocations. In section 4 we describe techniques for finding "representative" conjugate priors under more general hierarchical models thus allowing (at least approximately) optimal sampling allocations to be determined for this larger class of models. In many situations, these representative conjugate priors can be used to reduce the hierarchical model for the purposes of posterior analysis as well. A full numerical example is presented in section 5.

2. MODEL AND NOTATION

To avoid the necessity for subscripting, we first present the model and notation in terms of a single homogeneous Poisson process. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space, let $\{N(A): A \in \mathcal{F}\}$ be a homogeneous Poisson process on $(\Omega, \mathcal{F}, \nu)$ with unknown intensity $\theta \in \Theta = (0, \infty)$ and, for any $A \in \mathcal{F}$, let $X = (X, m) = (N(A), \nu(A))$ denote a complete sufficient statistic with realization $x = (x, m)$. Less formally, x is the realization of a Poisson count from a sample of "size" m. The p.m.f. of X is given by

$$f(x \mid \theta) = \frac{(m\theta)^x e^{-m\theta}}{\Gamma(x+1)} I_{\{0,1,2,\dots\}}(x), \qquad \theta \in (0,\infty). \quad (1)$$

We express our prior beliefs about the parameter θ by a conjugate gamma distribution with shape parameter α and scale parameter β, denoted $\text{Gamma}(\alpha, \beta)$, with density

$$\pi(\theta \mid \lambda) = \frac{\theta^{\alpha-1} e^{-\theta/\beta}}{\beta^\alpha \Gamma(\alpha)} I_{(0,\infty)}(\theta), \qquad \lambda = (\alpha, \beta) \in (0,\infty)^2. \quad (2)$$

We presently restrict our attention to the case when λ can be specified; the addition of hyper-priors on λ is considered in section 4.

For an arbitrary action a in the action space $A = \Theta$, we consider the loss functions

$$L_k(\theta, a) = \frac{(\theta - a)^2}{\theta^k}, \qquad k = 0,1. \quad (3)$$

L_0 is the ordinary squared error loss and L_1 is the relative squared error loss. For L_1 we require that $\alpha > 1$ which implies the gamma prior is unimodal.

[1] Brad C. Johnson, Department of Statistics, Purdue University, West Lafayette, IN 47907; e-mail: bradj@stat.purdue.edu; John Deely, Department of Statistics, Purdue University, West Lafayette, IN 47907.

Under the loss functions L_k the above model is extremely well understood. To simplify notation somewhat, let $\pi^\lambda = \pi(\theta \mid \lambda)$ and let δ_k^λ denote the Bayes procedure for π^λ under the loss function L_k. We recall that the posterior distribution of θ given x is

$$\theta \mid x \sim \text{Gamma}\left(\alpha + x, \frac{\beta}{m\beta + 1}\right).$$

The Bayes procedure for loss function L_k is given by

$$\delta_k^\lambda(x) = \frac{\beta(\alpha + x - k)}{m\beta + 1}, \qquad \alpha > k. \tag{4}$$

The posterior expected loss in using δ_k^λ under the loss function L_k is

$$\rho(\pi^\lambda, \delta_k^\lambda, L_k) = \left(\frac{\beta}{m\beta + 1}\right)^{2-k} (\alpha + x - k)^{1-k}, \qquad \alpha > k; \tag{5}$$

with Bayes risk

$$r(\pi^\lambda, \delta_k^\lambda, L_k) = \frac{\alpha^{1-k}\beta^{2-k}}{m\beta + 1}, \qquad \alpha > k. \tag{6}$$

It is interesting to note that under L_1, (4) and (5) imply that the Bayes procedure $\delta_1^\lambda(x)$ is the mode of the posterior and that $\rho(\pi^\lambda, \delta_1^\lambda, L_1)$ does not depend on the observed count x and hence is constant.

It is often more convenient, in terms of the elicitation process, to allow the shape parameter α of the gamma prior for θ to depend on the scale parameter β. In particular, the following alternate parameterizations are used throughout:

$$\theta \mid \lambda \sim \text{Gamma}(\mu/\beta, \beta), \quad \lambda = (\mu, \beta), \quad E(\theta \mid \lambda) = \mu; \tag{7}$$

$$\theta \mid \lambda \sim \text{Gamma}(\eta/\beta + 1, \beta), \ \lambda = (\eta, \beta), \ \text{Mode}(\theta \mid \lambda) = \eta. \tag{8}$$

Unless specified otherwise, results and formulae for these alternate parameterizations can be obtained by simply substituting the proper value for α. For λ as in (7) or (8) we substitute $\alpha = \mu/\beta$ or $\eta/\beta + 1$ respectively.

3. OPTIMAL ALLOCATION

We now discuss the allocation of sampling effort when $\{N_s(A) : A \in \mathcal{F}_s\}$, for $s = 1, \ldots, S$ are independent homogeneous Poisson processes on corresponding measure spaces $(\Omega_s, \mathcal{F}_s, \nu_s)$ with unknown intensities θ_s. The realization of a sample is now denoted $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S)$ where the $\boldsymbol{x}_s = (x_s, m_s)$ have the same meanings as $\boldsymbol{x} = (x, m)$ in section 2. For each process, $s = 1, \ldots, S$, we assume that

$$X_s \mid \theta_s \sim \text{Poisson}(m_s\theta_s);$$

$$\theta_s \mid \lambda_s \sim \text{Gamma}(\alpha_s, \beta_s), \quad \lambda_s = (\alpha_s, \beta_s).$$

Notice that we have not assumed that the θ_s are exchangeable so that prior information about one process is not influenced by the others.

Let $\boldsymbol{\delta}_k^\lambda = \boldsymbol{\delta}_k^\lambda(\boldsymbol{x}) = (\delta_k^{\lambda_1}(\boldsymbol{x}), \ldots, \delta_k^{\lambda_S}(\boldsymbol{x}_S))$ be the componentwise vector of Bayes procedures for estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)$ under the loss function L_k and let π^λ denote the overall prior specification. We assume that the overall loss for estimating some (possibly vector valued) function $g(\boldsymbol{\theta})$ with $g(\boldsymbol{\delta}_k^\lambda)$ can be expressed as

$$L_k(g(\boldsymbol{\theta}), g(\boldsymbol{\delta}_k^\lambda)) = \sum_{s=1}^{S} w_s L_k(\theta_s, \delta_k^{\lambda_s}(\boldsymbol{x}_s)), \tag{9}$$

where the w_s are known arbitrary non-negative weights. In particular this covers the case when we are interested in the simultaneous estimation of $\boldsymbol{W\theta}$ where $\boldsymbol{W} = (w_{js})$ is a $J \times S$ matrix and the loss structure is of the form

$$L_k(\boldsymbol{W\theta}, \boldsymbol{W\delta}_k^\lambda) = \sum_{j=1}^{J} \sum_{s=1}^{S} L_k(w_{js}\theta_s, w_{js}\delta_k^{\lambda_s})$$

$$= \sum_{s=1}^{S} \left(\sum_{j=1}^{J} w_{js}^{2-k}\right) L_k(\theta_s, \delta_k^{\lambda_s}). \tag{10}$$

The weights in (9) become $w_s = \sum_{j=1}^{J} w_{js}^{2-k}$ and, by the linearity of the expectation operator, the overall Bayes risk is given by

$$r(\pi^\lambda, \boldsymbol{W\delta}_k^\lambda, L_k) = \sum_s w_s r(\pi^{\lambda_s}, \delta_k^{\lambda_s}, L_k).$$

Let $\xi = (\xi_1, \ldots, \xi_S)$ denote the full specification where $\xi_s = (\alpha_s, \beta_s, w_s, c_s)$ denotes the specification for process s and c_s is the per unit sampling cost within that process. The general allocation problem involves finding an $\boldsymbol{m} = (m_1, \ldots, m_S)$ that minimizes the total risk $r(\pi^\lambda, g(\boldsymbol{\delta}^\lambda), L_k)$ of $g(\boldsymbol{\delta}^\lambda)$ subject to the constraint

$$C = \sum_{s=1}^{S} c_s m_s;$$

where C is the fixed total sampling budget. The proof of the following result is routine and deferred to the appendix.

Result 1. *Let $\xi = (\xi_1, \ldots, \xi_S)$ be given. The allocation $\boldsymbol{m} = (m_1, \ldots, m_S)$ that minimizes $r(\pi^\lambda, g(\boldsymbol{\delta}_0^\lambda), L_0)$ for fixed total cost C is*

$$m_s = \frac{\sqrt{w_s\alpha_s\beta_s/c_s}}{\sum_s \sqrt{w_s\alpha_s\beta_s c_s}} \left(C + \sum_s \frac{c_s}{\beta_s}\right) - \frac{1}{\beta_s}. \tag{11}$$

The allocation that minimizes $r(\pi^\lambda, g(\boldsymbol{\delta}_1^\lambda), L_1)$ is

$$m_s = \frac{\sqrt{w_s/c_s}}{\sum_s \sqrt{w_s c_s}} \left(C + \sum_s \frac{c_s}{\beta_s}\right) - \frac{1}{\beta_s}. \tag{12}$$

Equations (11) and (12) can result in one or more $m_s \le 0$ (*i.e.*, we take no samples from the offending processes) in which case we would remove these processes and reallocate C among the remaining processes. Of course, for the removed processes, our posterior mean and variance are equivalent to the prior mean and variance of θ_s.

We also comment that the allocation which minimizes $r(\pi^\lambda, \delta_1^\lambda, L_1)$ in (12) also minimizes $\rho(\pi^\lambda, \delta_1^\lambda, L_1)$ since this latter quantity is free of the observed counts x_s.

3.1 Comparisons with Traditional Frequentist Sampling Allocations

A special case of the above result is when the $\{N_s(A): A \in \mathcal{F}_s\}$ can be thought of as a stratification of a single non-homogeneous Poisson process $\{N(A): A \in \mathcal{F}\}$ and we are interested in estimating the overall population mean, say θ. To this end, let W_s denote the relative size of each Ω_s (which is assumed to be finite) and consider estimating the overall population mean $\theta = \mathbf{W\theta}$, where $\mathbf{W} = (W_1, ..., W_S)$, with the decision rule $\mathbf{W}\delta_0$. The weights in this case are $w_s = W_s^2$ and, substituting into (11), we obtain

$$m_s = \frac{W_s\sqrt{\alpha_s\beta_s/c_s}}{\sum_s W_s\sqrt{\alpha_s\beta_s c_s}}\left(C + \sum_s \frac{c_s}{\beta_s}\right) - \frac{1}{\beta_s}.$$

Letting $\beta_s \to \infty$ and $\alpha_s \to 0$ such that $\alpha_s\beta_s \to \mu_s$ simultaneously for each process is equivalent to letting $E(\theta_s) \to \mu_s$ and $\mathrm{Var}(\theta_s) \to \infty$ for all s and we obtain

$$m_s = \frac{CW_s\sqrt{\mu_s/c_s}}{\sum_s W_s\sqrt{\mu_s c_s}}. \tag{13}$$

This expression, up to the finite population correction factor, is the traditional frequentist allocation under the parametric model $X_s = (X_s, 1) \sim \mathrm{Poisson}(\mu_s)$ where μ_s represents our "best guess" for the mean (and hence variance) of X_s (*cf.* Cochran 1977). When $c_s = 1$ for all s, this becomes the Neyman allocation when the finite population correction factor is ignored.

Assuming that all of the μ_s are the same in (13) yields

$$m_s = \frac{CW_s/\sqrt{c_s}}{\sum_s W_s\sqrt{c_s}}; \tag{14}$$

and, when $c_s = 1$ for all s, we obtain the usual proportional allocation for fixed total sample size $C = N$.

In this sense, we see that the traditional frequentist solutions to the allocation problem can be obtained as the appropriate limit of Bayes solutions just as the traditional frequentist estimates can be obtained as a limit of Bayes procedures.

4. REPRESENTATIVE CONJUGATE PRIORS UNDER HIERARCHICAL MODELS

Up until now we have assumed that the λ_s were known. Returning to the notation of section 2 we now consider a more general hierarchical model

$$\begin{aligned} \boldsymbol{X} \mid \theta &\sim \mathrm{Poisson}(m\theta); \\ \theta \mid \lambda &\sim \mathrm{Gamma}(\alpha, \beta); \\ \lambda &\sim h(\lambda) \quad \lambda \in \mathcal{H}. \end{aligned} \tag{15}$$

We restrict our attention to choices of $h(\lambda)$ where the Bayes risk is finite and this precludes, among other things, the use of improper $h(\lambda)$. The unconditional prior for θ under this model can, at least in principle, be obtained as

$$\pi(\theta) = E^{h(\lambda)}\pi(\theta \mid \lambda).$$

In practice however, there is little to be gained since the resulting $\pi(\theta)$ will usually not be expressible in closed form. Indeed, it is usually the case that numeric integration and/or simulation is required to obtain the required posterior quantities and the Bayes risk.

We propose two methods for finding a "representative" *single conjugate prior* which, in most cases, can be substituted for $\pi(\theta)$ for the purposes of allocation. Indeed, for many practical situations, we find that these "representative" conjugate priors can replace the hierarchical model completely, greatly simplifying the posterior analysis.

We assume that it is relatively easy to simulate a sequence of random variables, $\{\lambda_n\}_{n=1}^N$, from $h\{\lambda\}$ and, as such, a sequence of random variables, $\{\theta_n\}_{n=1}^N$, can be obtained easily from $\pi(\theta)$ by taking $\theta_i \sim \pi(\theta \mid \lambda_i)$.

We now discuss the two techniques for finding the representative conjugate prior.

4.1 The Minimum L_∞ Conjugate Prior.

Let $F(\theta)$ and $F(\theta \mid \lambda)$ denote the distribution functions of $\pi(\theta)$ and $\pi(\theta \mid \lambda)$ respectively. The L_∞ conjugate prior, or L_∞-C prior, is defined to be the prior $\pi^\infty = \pi(\theta \mid \lambda^\infty)$ where λ^∞ is chosen such that

$$\|F(\theta) - F(\theta|\lambda^\infty)\|_\infty = \inf_{\lambda \in \mathcal{H}} \|F(\theta) - F(\theta|\lambda)\|_\infty.$$

That is, the L_∞-C prior is the prior $\pi(\theta \mid \lambda)$ which minimizes the L_∞ distance between $F(\theta)$ and $F(\theta \mid \lambda)$.

In order to estimate such a $\pi(\theta \mid \lambda^\infty)$ let $\{\theta_i\}_{i=1}^N$ be N simulated values from the unconditional prior $\pi(\theta)$; let $\theta_{i:N}$ denote the i^{th} ordered value of the $\{\theta_i\}$; and define the function

$$d_N(\lambda) = \max_i \left| F(\theta_{i:N}|\lambda) - \frac{i - .5}{N} \right|. \tag{16}$$

It is usually a routine matter to numerically find an (at least approximate) minimizing λ for (16) and our L_∞-C prior is $\pi(\theta \mid \lambda^\infty)$ where λ^∞ satisfies

$$d_N(\lambda^\infty) = \inf_{\lambda \in \mathcal{H}} d_N(\lambda).$$

Note that we are essentially minimizing the Kolmogorov-Smirnov statistic and the obvious appeal of estimating $\pi(\theta)$ in this manner is that $d_N(\lambda^\infty)$ can be directly interpreted as the estimated maximum difference of cumulative probabilities under $\pi(\theta)$ and $\pi(\theta \mid \lambda^{\infty})$. In the sequel, we will denote the Bayes procedure under the prior π^{∞} and loss function L_k as $\delta_k^{\infty}(x)$.

4.2 The ML Conjugate Prior

Let $\theta_1, ..., \theta_N$ be N simulated values from $\pi(\theta)$. The ML conjugate prior, or ML-C prior, when it exists, is defined to be the prior $\pi(\theta \mid \lambda^{\mathrm{ml}})$ where λ^{ml} satisfies

$$\pi(\theta \mid \lambda^{\mathrm{ml}}) = \sup_{\lambda \in \mathcal{H}} L(\lambda \mid \theta) = \sup_{\lambda \in \mathcal{H}} \prod_{i=1}^{N} \pi(\theta_i \mid \lambda);$$

or, equivalently,

$$\ln \pi(\theta \mid \lambda^{\mathrm{ml}}) = \sup_{\lambda \in \mathcal{H}} l(\lambda \mid \theta) = \sup_{\lambda \in \mathcal{H}} \sum_{i=1}^{N} \ln \pi(\theta_i \mid \lambda).$$

That is, λ^{ml} is the usual maximum likelihood estimator of λ if $\theta_1, ..., \theta_N$ were i.i.d. from $\pi(\theta \mid \lambda)$. Again, it is usually a simple matter to obtain λ^{ml} by numerical or simulation techniques. As in the L_{∞} method, we will let π^{ml} and δ_k^{ml} denote the estimated prior and the Bayes procedure under π^{ml} and loss function L_k.

Examples. The following four examples give an indication of how these procedures perform for a few different choices of $h(\lambda)$. In all of the examples we consider the general fist stage setup to be

$$X \mid \theta \sim \text{Poisson}(m\theta)$$
$$\theta \mid \lambda \sim \text{Gamma}(\eta/\beta + 1, \beta) \qquad \lambda = (\eta, \beta)$$

Furthermore, we assume η and β are independent so that $h(\lambda)$ may be written as $h_1(\eta)h_2(\beta)$. Adopting the notational conventions

$$Y \sim \text{Beta}_{(a,b)}(\zeta_1, \zeta_2) \Rightarrow f(y) \propto (y-a)^{\zeta_1 - 1}(b-y)^{\zeta_2 - 1} I_{(a,b)}(y);$$
$$Y \sim \text{InvGamma}\,(a, b) \rightarrow f(y) \propto y^{-(a+1)} e^{-1/yb} I_{(0,\infty)}(y);$$

the four examples considered are

Example	η	β
(a)	Uniform (4, 6)	$\text{Beta}_{(0.5,2)}(2, 5)$
(b)	Gamma (6.25, 0.8)	InvGamma (11, 1 / 30)
(c)	Uniform (2, 18)	Uniform (0.2, 0.5)
(d)	$\text{Beta}_{(3,15)}(2, 1)$	$\text{Beta}_{(0.1,0.5)}(1, 2)$

Table 1 gives the estimated λ^{∞} and λ^{ml} with $d_N(\lambda^{\infty})$ and, for comparison, $d_N(\lambda^{\mathrm{ml}})$ for each of these examples where all of the estimates are based on $N = 100{,}000$ simulated values from $\pi(\theta)$. In examples (a) and (b) both methods give very similar results and provide very good fits to $\pi(\theta)$ as indicated by the small values of d_N. Examples (c) and (d) were chosen to illustrate what happens when $\pi(\theta)$ deviates noticeably from a gamma distribution. Example (c) has a "plateau" distribution and example (d) is skewed in the wrong direction. As expected, the fits are less convincing in these examples. Figure 1 shows the simulated $\pi(\theta)$ along with π^{∞} and π^{ml} for each of these examples.

Table 1
Estimated λ^{∞} and λ^{ml} for examples (a) – (d)

Example	λ^{∞}	$d_N(\lambda^{\infty})$	λ^{ml}	$d_N(\lambda^{\mathrm{ml}})$
(a)	(4.94, 0.98)	0.003	(4.93, 1.00)	0.006
(b)	(4.42, 3.53)	0.003	(4.35, 3.63)	0.006
(c)	(7.72, 2.92)	0.043	(7.38, 2.93)	0.065
(d)	(10.44, 1.01)	0.040	(10.12, 1.10)	0.068

A more important consideration, for the purposes of the allocations discussed in section 3, is how well the Bayes risks are approximated under π^{∞} and π^{ml}. Table 2 gives the Bayes risk, $r(\pi, \delta_k^{\pi}, L_k)$ under the hierarchical model and the values for $r_k^*(\pi^{\infty})$ and $r_k^*(\pi^{\mathrm{ml}})$ where

$$r_k^*(\pi^{\bullet}) = \frac{r(\pi, \delta_k^{\pi}, L_k) - r(\pi^{\bullet}, \delta_k^{\bullet}, L_k)}{r(\pi, \delta_k^{\pi}, L_k)} \tag{17}$$

and where $\bullet = \infty$ or ml for each of the examples. The values $r(\pi, \delta_k^{\pi}, L_k)$ in this table were obtained by simulation and are subject to a certain amount of variation. Repeated simulations produced similar results. In examples (a) and (b) the correspondence between the Bayes risk under the full hierarchical model and the Bayes risk under the representative priors is very close, especially for the ML-C priors. In examples (c) and (d) the correspondence is still quite good considering these relatively small sample sizes. Overall, the ML-C prior appears to perform slightly better in the sense that the Bayes risks $r(\pi^{\mathrm{ml}}, \delta_k^{\mathrm{ml}}, L_k)$ tend to be closer to $r(\pi, \delta_k^{\pi}, L_k)$ with the exception of example (c) under the loss function L_1 where the L_{∞}-C prior is slightly better.

In examples (a) and (b) one may ask why a hierarchical model would be considered in the first place. The answer lies in the relative ease of eliciting absolute or probabilistic bounds on the hyper-parameters involved and taking $h(\lambda)$ to represent this uncertainty. The methods above can then, in many practical situations, be used to determine a representative single conjugate gamma prior for θ thus greatly simplifying the posterior analysis. The next section illustrates this with an example.

We also point out that it is relatively easy to construct examples where the methods described in this section will fail miserably at not only approximating π but also the Bayes risk. The method is best suited to cases where $h(\lambda)$ is chosen to represent uncertainty about λ. In situations when $h(\lambda)$ is being used to change the fundamental behavior of the first stage gamma prior (to create a bimodal prior for example) the representative priors π^{∞} and π^{ml} would normally not be used as a replacement for π in the posterior analysis but may still give suitable approximations of the Bayes risk for the purposes of allocation.

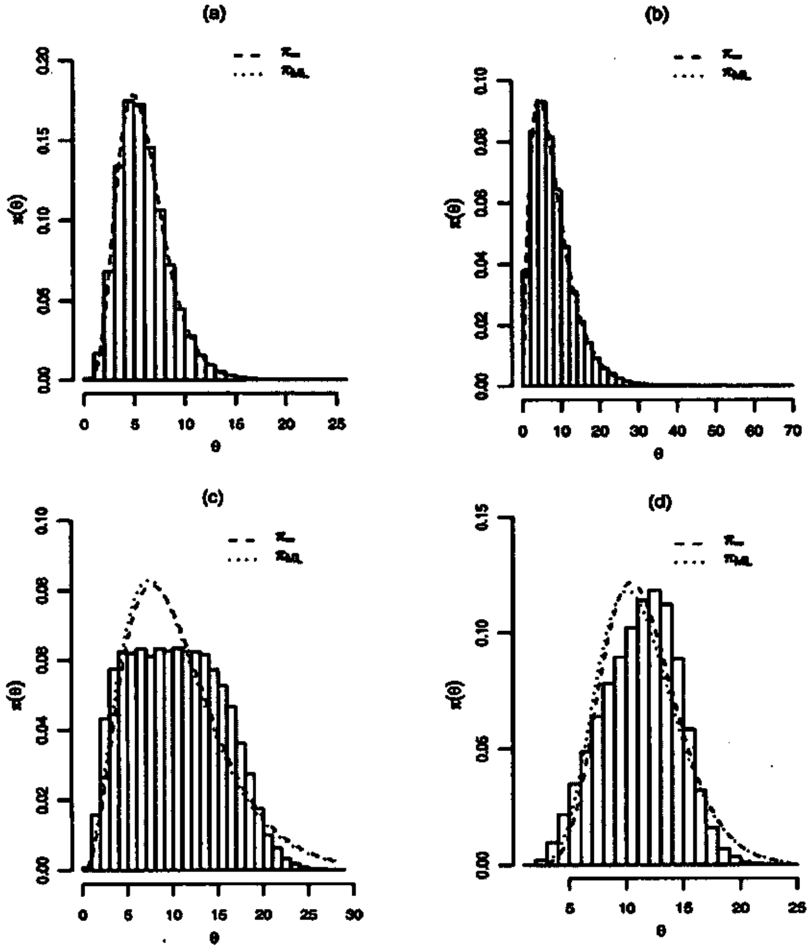


Figure 1. Simulated prior $\pi(\theta)$ (histogram) and the representative priors π^∞ and π^{ml} for examples (a) - (d).

Table 2
Bayes risks for examples (a) – (d)

	Example (a)					
	L_0			L_1		
m	$r(\pi, \delta^\pi)$	$r_0^*(\pi^\infty)$	$r_0^*(\pi^{ml})$	$r(\pi, \delta^\pi)$	$r_1^*(\pi^\infty)$	$r_1^*(\pi^{ml})$
1	2.997	-0.018	-0.006	0.500	-0.006	-0.002
5	0.986	-0.004	0.001	0.167	-0.003	0.000
10	0.540	-0.006	-0.002	0.091	-0.002	0.000
	Example (b)					
	L_0			L_1		
m	$r(\pi, \delta^\pi)$	$r_0^*(\pi^\infty)$	$r_0^*(\pi^{ml})$	$r(\pi, \delta^\pi)$	$r_1^*(\pi^\infty)$	$r_1^*(\pi^{ml})$
1	6.320	-0.021	-0.009	0.791	-0.015	-0.008
5	1.524	-0.014	-0.007	0.190	-0.003	-0.002
10	0.779	-0.008	-0.002	0.097	-0.002	-0.001
	Example (c)					
	L_0			L_1		
m	$r(\pi, \delta^\pi)$	$r_0^*(\pi^\infty)$	$r_0^*(\pi^{ml})$	$r(\pi, \delta^\pi)$	$r_1^*(\pi^\infty)$	$r_1^*(\pi^{ml})$
1	6.861	0.154	0.121	0.725	0.027	0.028
5	1.836	0.084	0.052	0.183	0.023	0.024
10	0.968	0.062	0.031	0.095	0.015	0.015
	Example (d)					
	L_0			L_1		
m	$r(\pi, \delta^\pi)$	$r_0^*(\pi^\infty)$	$r_0^*(\pi^{ml})$	$r(\pi, \delta^\pi)$	$r_1^*(\pi^\infty)$	$r_1^*(\pi^{ml})$
1	5.251	0.096	0.121	0.523	-0.040	0.002
5	1.778	0.075	0.068	0.165	0.013	0.027
10	0.986	0.057	0.043	0.090	0.013	0.021

5. NUMERIC EXAMPLE

We now present a numerical example based on data in Lindley and Deely (1993). The data consists of traffic counts between the hours of 7 a.m. and 6 p.m. over a 341 day period (3,751 hours) for a particular street in Auckland, New Zealand. The hours are stratified into $M_1 = 2{,}673$ weekday hours and $M_2 = 1{,}078$ weekend hours and we assume that the number of vehicles per hour can be modeled by two independent Poisson processes. For the purposes of this example we assume a total budget of \$1,500 is to be allocated and that per hour sampling costs are $c_1 = \$10$ and $c_2 = \$5$ for weekdays and weekends respectively. The relative strata sizes in this case are $W_1 = 0.71261$ and $W_2 = 0.28739$ for weekdays and weekends respectively.

The prior belief is that the weekend traffic rate is 40 vehicles per hour and that weekend traffic accounts for 5% of the total weekly traffic which yields a weekday traffic rate of 304 vehicles per hour. Suppose also that, for weekday traffic, we have elicited that the number of vehicles per hour will rarely exceed 400 and that, for weekend days, the number of vehicles per hour will rarely exceed 60, that is, say

$$\Pr(X_1 \le 400) \approx .95 \quad \text{and} \quad \Pr(X_2 \le 600) \approx 0.95.$$

Making use of the fact that the marginal distribution of x_s given λ_s is a "number of failure" negative binomial distribution of "size" $\alpha = \eta/\beta + 1$ and success probability $1/(m\beta + 1)$ we find that, when $\eta_1 = 304$ and $\eta_2 = 40$, the β_s's that come closest to satisfying these elicited probabilities are $\beta_1 = 7.51$ and $\beta_2 = 1.74$ respectively.

We now assume that the modes of the traffic rates for weekdays and weekends are equally likely to be within approximately 10% of the elicited traffic rates of 304 and 40 respectively and take

$$\eta_1 \sim \text{Uniform}(274, 334) \quad \text{and} \quad \eta_2 \sim \text{Uniform}(36, 44).$$

To represent our uncertainty about the β_s we take

$$\beta_1 \sim \text{InvGamma}(11, 0.0136)$$

and

$$\beta_2 \sim \text{InvGamma}(14.25, 0.043);$$

which yields $E(\beta_1) = 7.5$ with $\Pr(4 < \beta_1 \le 13.4) \approx 0.95$ and $E(\beta_2) = 1.75$ with $\Pr(1.03 \le \beta_2 \le 2.97) \approx 0.95$.

Using the ML-C technique in section 4 with $N = 100{,}000$, the specifications for weekday ($s = 1$) and weekend ($s = 2$) hourly traffic rates along with the values $d_N(\pi^{\text{ml}})$ are

s	c_s	W_s	η_s^{ml}	β_s^{ml}	$d_N(\pi^{\text{ml}})$
1	10	0.71261	302.98	8.303	0.0055
2	5	0.28739	39.876	1.889	0.0060

For the remainder of this section we will dispense with the superscript "ml" and simply refer the prior specification as π^λ and let $\boldsymbol{\delta}^\lambda(\boldsymbol{x}) = (\delta_1^\lambda, \ldots, \delta_S^\lambda)$ denote the component-wise vector of Bayes procedures for estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_S)$ under the prior specification π^λ and loss function L_0.

We consider three different allocations based on estimating $\mathbf{W}_1\boldsymbol{\theta}$, $\mathbf{W}_2\boldsymbol{\theta}$ and $\mathbf{W}\boldsymbol{\theta}$ where

$$\mathbf{W}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \ \mathbf{W}_2 = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \ \text{and} \ W = \begin{bmatrix} \mathbf{W}_1 \\ \hline \mathbf{W}_2 \end{bmatrix}.$$

With $\mathbf{W}_1$ we are primarily interested in estimating the weekday and weekend traffic rates θ_1 and θ_2 individually; with $\mathbf{W}_2$ we are only interested in estimating the overall traffic rate $\theta = W_1\theta_1 + W_2\theta_2$; and, with $\mathbf{W}$, we are interested in estimating all of these. In the sequel, we will refer to the allocations as $\boldsymbol{m}(\mathbf{W}_1)$, $\boldsymbol{m}(\mathbf{W}_2)$ and $\boldsymbol{m}(\mathbf{W})$ respectively.

Table 3 gives the allocations and corresponding weights w_s for these examples based on (11) and table 4 shows the Bayes risks in estimating θ_1, θ_2, $\mathbf{W}_1\boldsymbol{\theta}$, $\mathbf{W}_2\boldsymbol{\theta}$ and $\mathbf{W}\boldsymbol{\theta}$ under these 3 allocations. While allocation $\boldsymbol{m}(\mathbf{W}_2)$ is optimal for estimating the overall traffic rate θ, it results in large increases in the Bayes risks when estimates for the weekday and weekend traffic rates are also desired – the Bayes risk for estimating θ_2 under $\boldsymbol{m}(\mathbf{W}_2)$ is almost double compared to the Bayes risk under $\boldsymbol{m}(\mathbf{W})$.

Table 3

Weights and allocations for $\mathbf{W}_1$, $\mathbf{W}_2$ and $\mathbf{W}$.

	$\boldsymbol{m}(\mathbf{W}_1)$		$\boldsymbol{m}(\mathbf{W}_2)$		$\boldsymbol{m}(\mathbf{W})$	
s	w_s	m_s	w_s	m_s	w_s	m_s
1	1	119.33	0.5078	136.04	1.5078	123.20
2	1	61.35	0.0826	27.92	1.0826	53.60

Table 4

Bayes risks under allocations $\boldsymbol{m}(\mathbf{W}_1)$, $\boldsymbol{m}(\mathbf{W}_1)$ and $\boldsymbol{m}(\mathbf{W}_1)$.

	Estimand				
$\boldsymbol{m}$	θ_1	θ_2	$\mathbf{W}_1\boldsymbol{\theta}$	$\mathbf{W}_2\boldsymbol{\theta}$	$\mathbf{W}\boldsymbol{\theta}$
$\boldsymbol{m}(\mathbf{W}_1)$	2.61	0.68	3.28	1.38	4.66
$\boldsymbol{m}(\mathbf{W}_2)$	2.29	1.47	3.75	1.28	5.04
$\boldsymbol{m}(\mathbf{W})$	2.52	0.77	3.29	1.35	4.64

6. CONCLUDING COMMENTS

The techniques employed in the present paper are general enough to apply to a wide variety of Bayesian models. Optimal allocation equations for other Bayesian models in which the prior beliefs can, at least approximately, be modeled by conjugate priors are usually easy to obtain. The idea of finding "representative" conjugate priors, as discussed in section 4, is also applicable to a wide variety of hierarchical models with first stage conjugate priors. Areas of additional research in this area include allocations under loss functions other that L_0 and L_1 as well as more complicated cost functions.

7. ACKNOWLEDGEMENTS

The authors would like to thank Dennis Lindley for his comments on an earlier version of this paper. We would also like to thank the referees for their valuable input and comments. Partially supported by National Science Foundation grant NFS/DMS 9983601.

APPENDIX A. PROOF OF RESULT 1

Proof of Result 1. Introducing the Lagrange multiplier λ, we wish to minimize, for loss function L_k,

$$\sum_{s=1}^{S} \frac{w_s \alpha_s^{1-k} \beta_s^{2-k}}{m_s \beta_s + 1} + \lambda\left(\sum_s m_s c_s - C\right).$$

Differentiating with respect to m_s, setting equal to 0 and solving for m_s yields

$$m_s = \frac{\sqrt{w_s \alpha_s^{1-k} \beta_s^{1-k}/c_s}}{\sqrt{\lambda}} - \frac{1}{\beta_s}.$$

Therefore, to minimize the risk for fixed cost, we take

$$C = \sum_s m_s c_s = \sum_{s=1}^{S} \frac{\sqrt{w_s \alpha_s^{1-k} \beta_s^{1-k} c_s}}{\sqrt{\lambda}} - \sum_{s=1}^{S} \frac{c_s}{\beta_s}$$

or

$$\sqrt{\lambda} = \frac{\Sigma_s \sqrt{w_s \alpha_s^{1-k} \beta_s^{1-k} c_s}}{C + \Sigma_s c_s/\beta_s}.$$

Substituting this back into the equation for m_s yields

$$m_s = \frac{\sqrt{w_s \alpha_s^{1-k} \beta_s^{1-k}/c_s}}{\Sigma_s \sqrt{w_s \alpha_s^{1-k} \beta_s^{1-k} c_s}}\left(C + \sum_s \frac{c_s}{\beta_s}\right) - \frac{1}{\beta_s}.$$

Taking $k = 0$ or 1 gives the desired result.

REFERENCES

CLEVENSON, M.L., and ZIDEK, J.V. (1975). Simultaneous estimation of the means in independent Poisson laws. *Journal of the American Statistical Association.* 70(351), 698-705.

COCHRAN, W.G. (1977). *Sampling Techniques* (Third ed.). New York: John Wiley & Sons, Inc.

ERICSON, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations. In *New Developments in Survey Sampling,* (Eds. N.L. Johnson and H. Smith Jr.). New York: John Wiley & Sons, Inc. 326-357.

ERICSON, W.A. (1988). Bayesian Inference in Finite Populations. In *Handbook of Statistics,* (Eds. P. Krishnaiah and C. Rao). Amsterdam: Elsevier Science. 6, 213-246.

GHOSH, M., and MEEDEN, G. (1997). *Bayesian Methods for Finite Population Sampling.* Monographs on Statistics and Applied Probability. New York: Chapman & Hall. 79.

GODAMBE, V.P. (1955). A unified theory for sampling from finite populations. *Journal of the Royal Statistical Society Series B.* 17, 267-278.

LEITE, J.G., RODRIGUES, J. and MILAN, L.A. (2000). A Bayesian analysis for estimating the number of species in a population using a nonhomogenoeous Poisson process. *Statistics & Probability Letters.* 48, 153-161.

LINDLEY, D.V., and DEELY, J.J. (1993). Optimal allocation in stratified sampling with partial information. *Test.* 2(1), 147-160.

SCOTT, A., and SMITH, T.M.F. (1971). Bayes estimates for subclasses in stratified sampling. *Journal of the American Statistical Association.* 66(336), 834-836.

TIWARI, R.C., and LAHIRI, P. (1989). On robust emperical Bayes analysis of means and variances from stratified samples. *Communication in Statistics, Part A - Theory and Methods.* 18(3), 913-926.

Survey Methodology, June 2003
Vol. 29, No. 1, pp. 99-104
Statistics Canada

Note on Calibration in Stratified and Double Sampling

D.S. TRACY, SARJINDER SINGH and RAGHUNATH ARNAB[1]

ABSTRACT

In the present investigation, new calibration equations making use of second order moments of the auxiliary character are introduced for estimating the population mean in stratified simple random sampling. Ways for estimating the variance of the proposed estimator are suggested, as well. The resultant new estimator can be more efficient than the combined regression estimator is in stratified sampling. The idea has been extended to double sampling in a stratified population and some simulation results studied.

KEY WORDS: Calibration; Stratified Sampling; Double Sampling.

1. INTRODUCTION

Calibration estimation (Deville and Särndal 1992) has been much studied and practitioners have already offered many useful approaches (*e.g.*, Dupont 1995, Hidiroglou and Särndal 1998, Sitter and Wu 2002). Still more seems to remain to be done, as the use of this powerful technique expands further among practitioners.

This paper offers a modest extension of calibration estimation in the stratified and double sampling settings. We begin in this introduction by describing a new calibration estimator for the conventional stratified sample setting. Section 2 derives the variance of the proposed new estimator, followed by the derivation of a variance estimator. Section 3 extends these results to the important special case of double sampling. To explore the performance characteristics of the new estimator, some simulation results are presented in section 4 which concludes this brief note.

1.1 Standard Stratified Sampling Estimator

Suppose we have a population of N units that is first subdivided into L homogeneous subgroups called strata, such that the h-th stratum consists of N_h units, where $h = 1,2,...,L$ and $\sum_{h=1}^{L} N_h = N$. Suppose further that a sample of size n_h is drawn by Simple Random Sampling Without Replacement (SRSWOR) from the h-th population stratum such that $\sum_{h=1}^{L} n_h = n$, the required sample size. Finally, suppose the value of the i-th unit of the study variable selected from the h-th stratum is denoted by y_{hi}, where $i = 1,2,...,n_h$ and $W_h = N_h/N$ is the known proportion of population units falling in the h-th stratum.

In this standard set up (Cochran 1977), it can be shown that an unbiased estimator of population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1.1}$$

where $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$ denotes the h-th stratum sample mean. Under SRSWOR sampling, the variance of the estimator $\bar{y}_{st}$ is given by

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) S_{hy}^2 \tag{1.2}$$

where $S_{hy}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$ denotes the h-th stratum population variance, $\bar{Y}_h = N_h^{-1} \sum_{i=1}^{N_h} Y_{hi}$ denotes the h-th stratum population mean and $f_h = n_h / N_h$.

1.2 Proposed New Calibration Estimator

Let $X_{hi}, i = 1,2,...,N_h; h = 1,2,...,L$ denote the value of the i-th unit of the auxiliary variable in the h-th stratum about which information may be known at the unit level or at the stratum level. Consider a new alternative (calibration) estimator for stratified sampling of the form

$$\bar{y}_{st}(\text{new}) = \sum_{h=1}^{L} \Omega_h \bar{y}_h \tag{1.3}$$

where the weights Ω_h are chosen such that the chi-square distance function

$$\sum_{h=1}^{L} \frac{(\Omega_h - W_h)^2}{W_h Q_h} \tag{1.4}$$

where Q_h denotes suitable weights to form different forms of estimators such as combined ratio and combined regression type estimators, is minimized subject to the following two calibration constraints

$$\sum_{h=1}^{L} \Omega_h \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h \tag{1.5}$$

and

[1] D.S. Tracy, Department of Mathematics and Statistics, University of Windsor, Windsor, Ontario, N9B 3P4, Canada; Sarjinder Singh, Department of Statistics, St. Cloud State University, 720 Fourth Avenue South, St. Cloud, MN 56301-4498, U.S.A. E-mail: sarjinder@yahoo.com and Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa. E-mail: arnab@pixie.udw.ac.za.

$$\sum_{h=1}^{L} \Omega_h s_{hx}^2 = \sum_{h=1}^{L} W_h S_{hx}^2, \qquad (1.6)$$

where x_{hi}, $i = 1,2,...,n_h$; $h = 1,2,...,L$ denotes the value of sampled i-th unit from the h-th stratum such that $\bar{x}_h = n_h^{-1}\sum_{i=1}^{n_h} x_{hi}$ denotes the h-th stratum sample mean estimator of the known h-th stratum population mean $\bar{X}_h = N_h^{-1}\sum_{i=1}^{N_h} X_{hi}$, and $s_{hx}^2 = (n_h - 1)^{-1}\sum_{i=1}^{n_h}(x_{hi} - \bar{x}_h)^2$ denotes the h-th stratum sample variance estimator of the known h-th stratum population variance $S_{hx}^2 = (N_h - 1)^{-1}\sum_{i=1}^{N_h}(X_{hi} - \bar{X}_h)^2$ of the auxiliary variable.

Now it can be shown that minimization of (1.4) subject to (1.5) and (1.6) leads to new calibrated weights given by

$$\Omega_h = W_h + \left\{W_h Q_h \bar{x}_h\left[\sum_{h=1}^{L} W_h(\bar{X}_h - \bar{x}_h)\sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \sum_{h=1}^{L} W_h (S_{hx}^2 - s_{hx}^2)\sum_{h=1}^{L} W_h Q_h s_{hx}^2\right]\right\} \Big/ \left\{\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2\right)^2\right\}$$

$$+ \left\{W_h Q_h s_{hx}^2\left[\sum_{h=1}^{L} W_h(S_{hx}^2 - s_{hx}^2)\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 - \sum_{h=1}^{L} W_h (\bar{X}_h - \bar{x}_h)\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2\right]\right\} \Big/ \left\{\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2\right)^2\right\}. \qquad (1.7)$$

On substituting (1.7) in (1.3), we get

$$\bar{y}_{st}(\text{new}) = \sum_{h=1}^{L} W_h\left[\bar{y}_h + \hat{\beta}_1(\bar{X}_h - \bar{x}_h) + \hat{\beta}_2\left(S_{hx}^2 - s_{hx}^2\right)\right] \qquad (1.8)$$

where

$$\hat{\beta}_1 = \left\{\sum_{h=1}^{L} W_h Q_h \bar{x}_h \bar{y}_h\left[\sum_{h=1}^{L} W_h(\bar{X}_h - \bar{x}_h)\sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \sum_{h=1}^{L} W_h\left(S_{hx}^2 - s_{hx}^2\right)\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2\right]\right\} \Big/ \left\{\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2\right)^2\right\}$$

and

$$\hat{\beta}_2 = \left\{\sum_{h=1}^{L} W_h Q_h s_{hx}^2 \bar{y}_h\left[\sum_{h=1}^{L} W_h\left(S_{hx}^2 - s_{hx}^2\right)\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 - \sum_{h=1}^{L} W_h(\bar{X}_h - \bar{x}_h)\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2\right]\right\} \Big/ \left\{\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h Q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^2\right)^2\right\}.$$

Since the ratio $\Omega_h / W_h \rightarrow 1$ in probability, as the sample size in each stratum tends to infinity, the proposed estimator of the population mean is consistent.

Note that we are calibrating the estimates of the sample mean and the sample variance from each stratum, instead of each value of x_i, to the corresponding population parameters. Further note that if the population variance for each stratum is unknown, but the population means $\bar{X}_h$, $h = 1,2,...,L$ are known (or $\bar{X}$ is known), then it is advised to use only the single constraint (1.5).

2. VARIANCE AND VARIANCE ESTIMATION

While the new estimator $\bar{y}_{st}(\text{new})$ has been shown above to have acceptable asymptotic properties, what about the variance of the estimator and how does one go about estimating the variance? These questions are addressed in this section. We begin by looking (in subsection 2.1) at the variance of $\bar{y}_{st}(\text{new})$ and then go on to show how that variance can be estimated (in subsection 2.2).

2.1 Variance of New Estimator

The variance of the estimator $\bar{y}_{st}(\text{new})$ is given by

$$V(\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right) S_{hy}^2\left\{1 - \lambda_{h11}^2 - \frac{(\lambda_{h11}\lambda_{h03} - \lambda_{h12})^2}{\lambda_{h04} - 1 - \lambda_{h03}^2}\right\} \qquad (2.1)$$

where $\lambda_{hrs} = \mu_{hrs} / \mu_{h20}^{r/2}\mu_{02}^{s/2}$ and $\mu_{hrs} = (N_h - 1)^{-1}\sum_{i=1}^{N_h}(Y_{hi} - \bar{Y}_h)^r (X_{hi} - \bar{X}_h)^s$.

The expression (2.1) shows that the proposed estimator is always at least as efficient as the combined regression estimator in stratified sampling defined as

$$\bar{y}_{st}(c) = \sum_{h=1}^{L} W_h\left[\bar{y}_h + \hat{\beta}(\bar{X}_h - \bar{x}_h)\right] \qquad (2.2)$$

with variance

$$V(\bar{y}_{st}(c)) = \sum_{h=1}^{L} W_h^2\left(\frac{1-f_h}{n_h}\right) S_{hy}^2\left\{1 - \lambda_{h11}^2\right\}. \qquad (2.3)$$

The variance $V\,(\bar{y}_{st}(\text{new}))$ can be written as

$$V(\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \epsilon_{hi}^2 \qquad (2.4)$$

where

$$\epsilon_{hi} = (Y_{hi} - \bar{Y}_h) - \beta_1 (X_{hi} - \bar{X}_h) - \beta_2 \left\{ (X_{hi} - \bar{X}_h)^2 - \sigma_{hx}^2 \right\} (2.5)$$

with $\sigma_{hx}^2 = N_h^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$.

2.2 Estimation of the Variance

An estimator for estimating the variance $V(\bar{y}_{st}(\text{new}))$ is given by

$$\hat{V}_0 (\bar{y}_{st}(\text{new})) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h} \right) \frac{1}{n_h - 3} \sum_{i=1}^{n_h} e_{hi}^2 \qquad (2.6)$$

where

$$e_{hi} = (y_{hi} - \bar{y}_h) - \hat{\beta}_1 (x_{hi} - \bar{x}_h) - \hat{\beta}_2 \left\{ (x_{hi} - \bar{x}_h)^2 - s_{hx}^{*2} \right\} \qquad (2.7)$$

with $s_{hx}^{2*} = n_h^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ being the maximum likelihood estimator of σ_{hx}^2.

We also consider a calibrated estimator of the variance defined as

$$\hat{V}_1 (\bar{y}_{st}(\text{new}))_1 = \sum_{h=1}^{L} \Omega_h^2 \left(\frac{1-f_h}{n_h} \right) \frac{1}{n_h - 3} \sum_{i=1}^{n_h} e_{hi}^2. \qquad (2.8)$$

The estimator proposed by Wu (1985) is a special case of this estimator.

3. DOUBLE SAMPLING

In this section we extend our stratified sampling results to the stratified double sampling case. In particular, suppose the population of N units consists of L strata such that the h-th stratum consists of N_h units and $\sum_{h=1}^{L} N_h = N$. From the h-th stratum of N_h units, draw a preliminary large sample of m_h units by SRSWOR sampling and measure the auxiliary character x_{hi} only. Select a sub-sample of n_h units from the given preliminary large sample of m_h units by SRSWOR sampling and measure both the study variable y_{hi} and auxiliary variable x_{hi}. Let $\bar{x}_h^* = m_h^{-1} \sum_{i=1}^{m_h} x_{hi}$ and $s_{hx}^{*2} = (m_h - 1)^{-1} \sum_{i=1}^{m_h} (x_{hi} - \bar{x}_h^*)^2$ denote the first phase sample mean and variance. Also let $\bar{x}_h = n_h^{-1} \sum_{i=1}^{n_h} x_{hi}$, $s_{hx}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ and $\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi}$, $s_{hy}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ denote the second phase sample mean and variances for the auxiliary and study characters, respectively. We are considering an estimator of the population mean in stratified double sampling as

$$\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h^* \bar{y}_h \qquad (3.1)$$

where W_h^* are the calibrated weights such that the chi-square distance

$$\sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h} \qquad (3.2)$$

where Q_h are predefined weights used to obtain to different types of estimators, is minimized subject to the constraints

$$\sum_{h=1}^{L} W_h^* \bar{x}_h = \sum_{h=1}^{L} W_h \bar{x}_h^* \qquad (3.3)$$

and

$$\sum_{h=1}^{L} W_h^* s_{hx}^2 = \sum_{h=1}^{L} W_h s_{hx}^{*2} \qquad (3.4)$$

where $W_h = N_h / N$ are known stratum weights. We then get the calibrated weights, for stratified double sampling, as

$$W_h^* = W_h + \left\{ W_h Q_h \bar{x}_h \left[\sum_{h=1}^{L} W_h \left(\bar{x}_h^* - \bar{x}_h \right) \sum_{h=1}^{L} W_h Q_h \, s_{hx}^4 - \sum_{h=1}^{L} W_h \left(s_{hx}^{*2} - s_{hx}^2 \right) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right] \right\} \Bigg/ \left\{ \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right) \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^4 \right) - \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right)^2 \right\}$$
$$+ \left\{ W_h Q_h s_{hx}^2 \left[\sum_{h=1}^{L} W_h \left(s_{hx}^{*2} - s_{hx}^2 \right) \sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 - \sum_{h=1}^{L} W_h (\bar{x}_h^* - \bar{x}_h) \sum_{h=1}^{L} W_h Q_h \bar{x}_h s_{hx}^2 \right] \right\} \Bigg/ \left\{ \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right) \left(\sum_{h=1}^{L} W_h Q_h s_{hx}^4 \right) - \left(\sum_{h=1}^{L} W_h Q_{hx_h}^{-} s_{hx}^2 \right)^2 \right\}. \qquad (3.5)$$

Substitution of (3.5) in (3.1) leads to a new estimator of the population mean in stratified random sampling. Thus a calibrated estimator of the population mean in stratified double sampling is given by

$$\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}_1^* \left[\sum_{h=1}^{L} W_h \left(\bar{x}_h - \bar{x}_h^* \right) \right] + \hat{\beta}_2^* \left[\sum_{h=1}^{L} W_h \left(s_{hx}^2 - s_{hx}^{*2} \right) \right] \qquad (3.6)$$

where $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ have their usual meanings. It is to be noted that the estimator (3.6) makes the use of the estimated first phase variance of the auxiliary character while estimating the population mean. Thus the estimator (3.6) is different than the usual separate regression type estimator available in the literature.

Since the ratio $W_h^* / W_h \to 1$ in probability, as the second-phase sample size in each stratum tends to infinity, the proposed estimator is a consistent estimator of the population mean. The conditional variance of the stratified double sampling estimator, $\bar{y}_{st}(d) = \sum_{h=1}^{L} W_h^* \bar{y}_h$, is

$$V\left[\bar{y}_{st}(d) \mid W_h^*\right] = \sum_{h=1}^{L} W_h^{*2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{hy}^2 \qquad (3.7)$$

where $S_{hy}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$.

A conditionally unbiased estimator of $V[\bar{y}_{st}(d) \mid W_h^*]$ is

$$\hat{V}\left[\bar{y}_{st}(d) \mid W_h^*\right] = \sum_{h=1}^{L} W_h^{*2}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) s_{hy}^2 \qquad (3.8)$$

where $s_{hy}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$.

It may be noted that in the proposed strategy, there is no need to go for higher order calibration for estimating the variance, because the calibrated weights W_h^* already make use of the estimated first phase variance of the auxiliary character. The minimum variance of the stratified double sampling estimator $\bar{y}_{st}(d)$, to the first order of approximation, is given by

$$V(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h}\right) S_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h}\right) S_{hy}^2 \left\{1 - \lambda_{h11}^2 - \frac{(\lambda_{h11}\lambda_{h03} - \lambda_{h12})^2}{\lambda_{h04} - 1 - \lambda_{h03}^2}\right\}\right]. \qquad (3.9)$$

The variance of the stratified double sampling estimator $\bar{y}_{st}(d)$ can also be written as

$$V(\bar{y}_{st}(d)) \approx \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h}\right) S_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h}\right) \frac{1}{N_h} \sum_{i=1}^{N_h} \epsilon_{hi}^2\right] \qquad (3.10)$$

where

$$\epsilon_{hi} = (Y_{hi} - \bar{Y}_h) - \beta_1 (X_{hi} - \bar{X}_h) - \beta_2 \left\{(X_{hi} - \bar{X}_h)^2 - \sigma_{hx}^2\right\}. \qquad (3.11)$$

An estimator of variance $V(\bar{y}_{st}(d))$ is given by

$$\hat{V}(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^2 \left[\left(\frac{1}{m_h} - \frac{1}{N_h}\right) s_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h}\right) \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}^2\right] \qquad (3.12)$$

where $e_{hi} = (y_{hi} - \bar{y}_h) - \hat{\beta}_1 (x_{hi} - \bar{x}_h) - \hat{\beta}_2 \{(x_{hi} - \bar{x}_h)^2 - s_{hx}^{*2}\}$ denotes the estimate of the residual term and $s_{hx}^{*2} = n_h^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ denotes the maximum likelihood estimator of σ_{hx}^2.

We suggest here a new estimator of the variance in stratified double sampling as

$$\hat{V}(\bar{y}_{st}(d)) = \sum_{h=1}^{L} W_h^{*2} \left[\left(\frac{1}{m_h} - \frac{1}{N_h}\right) s_{hy}^2 + \left(\frac{1}{n_h} - \frac{1}{m_h}\right) \frac{1}{n_h - 1} \sum_{i=1}^{n_h} e_{hi}^2\right]. \qquad (3.13)$$

Clearly $\lim_{m_h \to N_h} \hat{V}(\bar{y}_{st}(d)) = \hat{V}(\bar{y}_{st}(\text{new}))$ because $\lim_{m_h \to N_h} W_h^* \to \Omega_h$. Note that in two-phase sampling, an estimate of population parameter of the auxiliary character based on first-phase sample information (large sample) will always be better than the corresponding estimate based on only second-phase sample information. One can refer to Hidiroglou and Särndal (1998) to see that calibration to an estimate of such an unknown quantity works well.

4. EARLY SIMULATION RESULTS AND SOME CONCLUSIONS

To begin our study of the operating performance of the proposed estimator with respect to the usual combined regression estimator in stratified sampling, we performed a few simulation experiments. These are described below and then some overall observations are made to conclude the paper.

4.1 Simulation Results

The following procedure for doing the simulation experiment was adopted. We assumed that the population consists of three strata and within each stratum the population followed the distributions shown in Table 1.

In each stratum different transformations on x_{hi}^* and y_{hi}^* were made by examining all possible combinations of the correlation coefficients $\rho_h = 0.5, 0.7$ and 0.9 and sample sizes $n_h = 5$, 10, and 15. The quantities $S_{1x} = 4.5$, $S_{2x} = 6.2$, $S_{3x} = 8.4$ and $S_{hy} = 4.8$ were fixed in each stratum.

We generated 50,000 populations each of size 75 units and having 25 units in each stratum. From each stratum, SRSWOR samples were drawn and an average of the empirical mean squared error of the combined regression estimator was computed as:

$$\text{MSE}(\bar{y}_{st}(c)) = \frac{1}{50{,}000} \sum_{j=1}^{50{,}000} \left[\left(\sum_{h=1}^{3} W_h\left(\bar{y}_h + \hat{\beta}(\bar{X}_h - \bar{x}_h)\right)\right)_j - \bar{Y}\right]^2 \qquad (4.1)$$

Table 1
Characteristics of the Population

Population	Stratum 1	Stratum 2	Stratum 3
	$y_{1i} = 15 + \sqrt{(1-\rho_1^2)} y_{1i}^* + \rho_1 \frac{S_{1x}}{S_{1y}} x_{1i}^*$ $y_{1i} = 50 + x_{1i}^*$	$y_{2i} = 100 + \sqrt{(1-\rho_2^2)} y_{2i}^* + \rho_2 \frac{S_{2x}}{S_{2y}} x_{2i}^*$ $y_{2i} = 150 + x_{2i}^*$	$y_{3i} = 200 + \sqrt{(1-\rho_3^2)} y_{3i}^* + \rho_3 \frac{S_{3x}}{S_{3y}} x_{3i}^*$ $y_{3i} = 100 + x_{3i}^*$
1	$f(z_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} z_{hi}^{*\alpha_h - 1} e^{-z_{hi}^*}, \alpha_h = 0.3;$ for $z_{hi}^* = x_{hi}^*;$ $\alpha_h = 1.5$ for $z_{hi}^* = y_{hi}^*; h = 1,2,3$		
2	$f(y_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} y_{hi}^{*\alpha_h - 1} e^{-y_{hi}^*}, \alpha_h = 0.3; f(x_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_{hi}^{*2}}{2}}; h = 1,2,3$		
3	$f(x_{hi}^*) = \frac{1}{\Gamma_{\alpha_h}} x_{hi}^{*\alpha_h - 1} e^{-x_{hi}^*}, \alpha_h = 0.3; f(y_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y_{hi}^{*2}}{2}}; h = 1,2,3$		
4	$f(z_{hi}^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_{hi}^{*2}}{2}}$ for $z_{hi}^* = x_{hi}^*, z_{hi}^* = y_{hi}^*; h = 1,2,3$		

where $\bar{Y} = \frac{25\times 15 + 25\times\ 100 + 25\times 200}{75} = 100.5.$

Similarly the empirical mean squared error of the proposed estimator is given by

$$\text{MSE}(\bar{y}_{st}(p)) = \frac{1}{50{,}000} \sum_{j=1}^{50{,}000}$$

$$\left[\left(\sum_{h=1}^{3} W_h \left(\bar{y}_h + \hat{\beta}_1 (\bar{X}_h - \bar{x}_h) + \hat{\beta}_2 (S_{hx}^2 - s_{hx}^2) \right) \right)_j - \bar{Y} \right]^2. \tag{4.2}$$

The percent relative efficiency of the proposed estimator with respect to combined regression estimator is given by

$$\text{RE} = \frac{\text{MSE}(\bar{y}_{st}(c))}{\text{MSE}(\bar{y}_{st}(p))} \times 100. \tag{4.3}$$

The results so obtained demonstrated a modest improvement over all combinations studied for all four populations. The range of improvements was about 4.46% to 13.08% with the median being 5.19%.

Several empirical studies were also carried out similar in structure to those presented above. In particular we were able to illustrate the extent to which our approach was more efficient than that considered by Singh, Horn and Yu (1998) in stratified sampling. Quite similar results were observed for the double sampling setting. Using the simulation program with $m_h = 20, h = 1,2,3,$ with the same four populations as described earlier, the median improvement was observed as 3.17%, 7.20%, 5.28%, and 3.12%, respectively.

4.2 Some Overall Observations

We are comfortable that our new calibration estimator will perform well in many settings. Our simulation results demonstrate this in several special cases. As with other calibration estimators, however, there has been an appeal at various points to asymptotic results. Such appeals raise concerns in small samples. For example in section 3 we stated that the ratio $W_h^* / W_h - 1$ in probability. This allowed us to conclude that our new double sampling estimator was asymptotically unbiased. We recommend that such appeals be checked before our estimator is used in an application, possibly by employing simulation studies similar to those in this paper but for situations like those that are to be sampled in the practitioner's particular setting.

ACKNOWLEDGEMENTS

The authors' are thankful to the Associate Editor, the Assistant Editor and the two anonymous referees for their fruitful comments on the original version this manuscript. We are sorry to inform you that Professor Tracy passed away on 27/12/1998.

REFERENCES

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

DEVILLE, J.C., and SÄRNDAL, C.-E.(1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 87, 376-382.

DUPONT, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*. 21, 125-135.

HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*. 24, 11-20.

SINGH, S., HORN, S. and YU, F. (1998). Estimation of variance of the general regression estimator: Higher level calibration approach. *Survey Methodology*. 24, 41-50.

SITTER, R.R., and WU, C. (2002). Efficient estimation of quadratic finite population functions. *Journal of the American Statistical Association*. 97, 535-543

WU, C.F.J. (1985). Variance estimation for combined ratio and combined regression estimators. *Journal of the Royal Statistical Society, B*. 47, 147-154.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 18, No. 4, 2002

Volume 19, No. 1, 2003

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 30, No. 4, December/décembre 2002

The Canadian Journal of Statistics **La Revue Canadienne de Statistique**

CONTENTS TABLE DES MATIÈRES

Volume 31, No. 1, March/mars 2003

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

1.1 Manuscripts should be typed on white bond paper of standard size (8½ × 11 inch), one side only, entirely double spaced with margins of at least 1½ inches on all sides.
1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
1.4 Acknowledgements should appear at the end of the text.
1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

3.1 Avoid footnotes, abbreviations, and acronyms.
3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as "exp(·)" and "log(·)", etc.
3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
3.4 Write fractions in the text using a solidus.
3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.