Special Issue, Volume 21, Number 4, August 2020 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

STATISTICAL DATA INTEGRATION – Special Issue

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, e-mail: m.cierpial-wolan@stat.gov.pl
Secretary
Patryk Barszcz, e-mail: p.barszcz@stat.gov.pl, phone number + 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, e-mail: r.litkowiec@stat.gov.pl

Address for correspondence
Statistics Poland, al. Niepodległości 208, 00-925 Warsaw, Poland, Tel./fax:00 48 22 — 825 03 95

CONTENTS

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. II

From the Editor

The Editors and Editorial Board of the Statistics in Transition new series (SiTns) have great pleasure in presenting this special issue on statistical data integration to our readers. We are very grateful for the efforts taken by all those who contributed to the production of this special issue that made its publication possible. We believe that this volume represents not only the state-of-the-art in the relevant topic areas, but that it will also help to identify new research avenues for study in the years to come.

Behind such an ambitious and demanding endeavor, there is always a key role to be played by an intellectual and organizational leader. Practically, we owe this product personally to Professor Partha Lahiri, who kindly accepted an invitation by SiTns Editorial Board member Graham Kalton and me to act as Editor-in-Chief of this special issue. We are very grateful to Malay Ghosh, another long-term member of the SiTns' Editorial Board, for initially putting forward the idea of a special issue on statistical data integration under Partha Lahiri's leadership. This special issue would not have been possible without Partha Lahiri's guidance and intellectual leadership, supported by a team of leading international experts who generously accepted his invitation to serve as Guest co-Editors.

This special issue is the third in the series of SiTns special issues. The two previous special issues were: (1) a two-volume special issue on small area estimation that was published jointly with Survey Methodology, and that arose out of a conference held in Poznan, with Ray Chambers, Malay Ghosh, Graham Kalton, and Risto Lehtonen serving as Guest co-Editors; and (2) a special issue on subjective well-being in survey research, co-edited by Graham Kalton and Christopher MacKie.

The focus of this special issue is broader than those of the previous ones because the subject-matter of statistical data integration encompasses a wide range of analytic objectives and of statistical techniques. It can be well argued that data integration is the dominant innovation in national statistical offices. If so, the efforts of everyone involved in the preparation of this volume would be duly appreciated. Let us believe that most of our readers share this view.

Last but not least, I would like to express my appreciation to the work of our Editorial Office members for their work done in parallel with the preparation of the regular SiTns release.

Wlodzimierz Okrasa,

Editor

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. III–VI

Preface

The demand for statistics on a range of socio-economic, agricultural, health, transportation, and other topics is steadily increasing at a time when government agencies are desperately looking for ways to reduce costs to meet fixed budgetary requirements. A single data source may not be able to provide all the data required for estimating the statistics needed for many applications in survey and official statistics. However, information complied through different data linkage or integration techniques may be a good option for addressing a specific research question or for multi-purpose uses. For example, information from multiple data sources can be extracted for producing statistics of desired precision at a granular level, for a multivariate analysis when a single data source does not contain all variables of interest, for reducing different kinds of nonsampling errors in probability samples or self-selection biases in nonprobability samples, and other emerging problems.

The greater accessibility of administrative and Big Data and advances in technology are now providing new opportunities for researchers to solve a wide range of problems that would not be possible using a single data source. However, these databases are often unstructured and are available in disparate forms, making data linkages quite challenging. Moreover, new issues of statistical disclosure avoidance arise naturally when combining data from various sources. There is, therefore, a growing need to develop innovative statistical data integration tools to link such complex multiple data sets. In the US federal statistical system, the need to innovate has been emphasized in the following report: National Academies of Sciences, Engineering, and Medicine. (2017), *Innovations in Federal Statistics: Combining Data Sources While Protecting Privacy*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24652.

The idea of organizing an international week-long workshop on statistical data integration arose in 2017. I joined Dr. Sanjay Chaudhuri, a faculty member at the National University of Singapore (NUS), Dr. Danny Pfeffermann, National Statistician of Israel, and Dr. Pedro Silva of the Instituto Brasileiro de Geografia e Estatística (IBGE), Brazil, and former President of the International Statistical Institute, to organize this international workshop. Eventually, with generous funding from the Institute for Mathematical Sciences at the National University of Singapore, the workshop was held on the NUS campus during August 5–8, 2019. The World Statistics Congress Satellite meeting on Current Trends in Survey Statistics took place at the same venue in the following week, August 13–16, 2019. We had great success with

participants and speakers from more than 18 countries in these two meetings, at which a number of papers on statistical data integration were presented.

A few months before the two Singapore events, in February of 2019, I had a fruitful lunch meeting in the Washington DC area with Professor Wlodzimierz Okrasa, Editor-in-Chief, and Dr. Graham Kalton, a member of the Editorial Board, of the Statistics in Transition (SiT) New Series. During that meeting they invited me to edit a special issue for the journal. We discussed a few options for the focus of the special issue. Our discussions led to the idea of focusing on statistical data integration, in view of the current importance of the topic, and the value of disseminating the findings from current research. We felt the issue would be timely, given the emphasize on this topic in the two Singapore workshops that were to be held later that year. We agreed that anyone, including the participants of the two Singapore meetings, could submit papers for possible publication in the special issue, and all papers would go through a thorough review process.

Out of the nineteen papers submitted for possible publication in this special issue, we finally accepted ten papers, after they went through a referring and revision process. In addition, this special issue features an invited discussion paper on a selective review of small area estimation by Professor Malay Ghosh, which is based on his 2019 Morris Hansen lecture delivered in Washington DC on October 30, 2019. We are pleased to have seven experts, including Professor J. N. K. Rao and Dr. Julie Gershunskaya – the two invited discussants of Professor Ghosh's Morris Hansen lecture – as discussants of Professor Ghosh's paper.

For over 75 years, survey statisticians have been using information from multiple data sources in solving a wide range of problems. One early example of combining surveys can be traced back to a 1943 Sankhya paper (www.jstor.org/stable/25047787) by Mrs. Chameli Bose. Mrs Bose developed the regression estimation for double sampling used by Professor P.C. Mahalanobis in 1940–41 to estimate the yield of cinchona bark in the Government Cinchona Plantation at Mungpoo, Bengal, India. Over the years, we have witnessed tremendous progress in such research topics as small area estimation, probabilistic record linkage, combining multiple surveys, multiple frame estimation, microsimulation, poststratification, all of which incorporate multiple data sources and can be brought under the broader umbrella of statistical data integration or data linkages. In a 2020 Sankhya B paper (doi 10.1007/s13571-020-00227-w), Professor J. N. K. Rao provides an excellent review of a selected subtopics of statistical data integration.

It is difficult to cover all interesting statistical data integration topics in a single issue of SiT. But we are happy that the invited discussion review paper plus the ten contributed papers published in this special issue collectively cover a broad spectrum of topics in statistical data integration. The papers can be broadly classified into the following subtopics: 1) small area estimation, 2) advances in probabilistic record

linkage and analysis of linked data, 3) statistical methods for longitudinal data, multiple-frame, and data fusion, and 4) synthetic data for microsimulations, disclosure avoidance and multi-purpose inferences.

Professor Ghosh's paper, along with the discussions, provide an excellent review of some topics in small area estimation and they should prove to be a valuable reference for those working on small area estimation. In addition, this issue features two more papers on small area estimation by (i) Cai, Rao, Dumitrescu, and Chatrchi, and (ii) Neves, Silva, and Moura that address variable selection and modeling to capture uncertainties of sampling errors of survey estimates, respectively. These are indeed important and yet understudied problems in small area estimation.

This special issue includes two papers that advance knowledge on probabilistic record linkage. Consiglio and Tuoto investigate potential advantages of using probabilistic record linkage in small area estimation. Bera and Chatterjee discuss a problem of probabilistic record linkage on high-dimensional data. This is a novel approach to the probabilistic record linkage methodology that can be applied in absence of any common matching field among the data sets.

The three papers by (i) Saegusa, (ii) Zhang, Pyne, and Kedem, and (iii) Bonnery, Cheng, and Lahiri investigate potential benefits of using nonparametric and semi-parametric methods to combine information from multiple data sources. The nature of the available multiple data sources differs between the three papers. Saegusa develops a nonparametric method to construct confidence bands for a distribution function using multiple overlapping data sources – this is an advancement in the multiple-frame theory. To overcome a relatively small sample of interest, Zhang et al. propose a semi-parametric data fusion technique for combining multiple spatial data sources using variable tilts functions obtained by model selection. Bonnery et al. carefully devise a complex simulation study, using the U.S. Current Population Survey (CPS) rotating panel survey data, to evaluate different possible estimators of levels and changes in the context of labor force estimation.

The three papers by (i) Bugard, Dieckmann, Krause, Münnich, Neufang, and Schmaus, (ii) Alam, Dostie, Drechsler, and Vilhuber, and (iii) Lahiri, and Suntornchost demonstrate how the synthetic data approach can be useful for solving seemingly unrelated problems. Bugard et al. discuss microsimulations that are used for evidence-based policy. Using a general framework for official statistics, they use synthetic data created from multiple data sets to approximate a realistic universe. The synthetic data discussed in the Alam et al. paper relates to statistical data disclosure. The authors consider a feasibility study to understand if the synthesis method for longitudinal business data used in a US project can be effectively applied to two other longitudinal business projects, in Canada and Germany. In the context of poverty estimation for small geographic areas, Lahiri and Suntornchost point out the inappropriateness of using point estimates for all inferential purposes. Using a Bayesian approach,

they demonstrate how synthetic data can be created for multipurpose inferences in small area estimation problems.

I would like to thank Professor Wlodzimierz Okrasa and Dr. Graham Kalton for encouraging me to take a lead on this project. I appreciate all the help I received from Professor Okrasa and his editorial staff. Thanks are also due to the anonymous referees who offered many constructive suggestions to improve the quality of the original submissions. Last but not the least, I would like to thank my distinguished guest co-editors Drs. Jean-Francois Beaumont, Sanjay Chaudhuri, Jörg Drechsler, Michael Larsen, and Marcin Szymkowiak for their diligent editorial work. Without their enormous help, we would not have this high quality special issue.

Partha Lahiri,

Guest Editor of Statistics in Transition new series

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. VII

Submission information for Authors

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and Statistics Poland, on a quarterly basis (during 1993–2006 it was issued twice and since 2006 three times a year). Also, it has extended its scope of interest beyond its originally primary focus on statistical issues pertinent to transition from centrally planned to a market-oriented economy through embracing questions related to systemic transformations of and within the national statistical systems, world-wide.

The SiT-*ns* seeks contributors that address the full range of problems involved in data production, data dissemination and utilization, providing international community of statisticians and users – including researchers, teachers, policy makers and the general public – with a platform for exchange of ideas and for sharing best practices in all areas of the development of statistics.

Accordingly, articles dealing with any topics of statistics and its advancement – as either a scientific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

Demonstration of the role played by statistical research and data in economic growth and social progress (both locally and globally), including better-informed decisions and greater participation of citizens, are of particular interest.

Each paper submitted by prospective authors are peer reviewed by internationally recognized experts, who are guided in their decisions about the publication by criteria of originality and overall quality, including its content and form, and of potential interest to readers (esp. professionals).

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

It is assumed, that the submitted manuscript has not been published previously and that it is not under review elsewhere. It should include an abstract (of not more than 1600 characters, including spaces). Inquiries concerning the submitted manuscript, its current status etc., should be directed to the Editor by email, address above, or w.okrasa@stat.gov.pl.

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. IX–X

Editorial Policy

The broad objective of *Statistics in Transition new series* is to advance the statistical and associated methods used primarily by statistical agencies and other research institutions. To meet that objective, the journal encompasses a wide range of topics in statistical design and analysis, including survey methodology and survey sampling, census methodology, statistical uses of administrative data sources, estimation methods, economic and demographic studies, and novel methods of analysis of socio-economic and population data. With its focus on innovative methods that address practical problems, the journal favours papers that report new methods accompanied by real-life applications. Authoritative review papers on important problems faced by statisticians in agencies and academia also fall within the journal's scope.

***

ABSTRACTING AND INDEXING DATABASES

Statistics in Transition new series is currently covered in:

Databases indexing the journal:

- BASE – Bielefeld Academic Search Engine
- CEEOL – Central and Eastern European Online Library
- CEJSH (The Central European Journal of Social Sciences and Humanities)
- CNKI Scholar (China National Knowledge Infrastructure)
- CNPIEC – cnpLINKer
- CORE
- Current Index to Statistics
- Dimensions
- DOAJ (Directory of Open Access Journals)
- EconPapers
- EconStore
- Electronic Journals Library
- Elsevier – Scopus
- ERIH PLUS (European Reference Index for the Humanities and Social Sciences)
- Genamics JournalSeek
- Google Scholar
- Index Copernicus
- J-Gate
- JournalGuide
- JournalTOCs
- Keepers Registry
- MIAR
- Microsoft Academic
- OpenAIRE
- ProQuest – Summon
- Publons
- QOAM (Quality Open Access Market)
- ReadCube
- RePec
- SCImago Journal & Country Rank
- Ulrichsweb & Ulrich's Periodicals Directory
- WanFang Data
- WorldCat (OCLC)
- Zenodo.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 1–22, DOI 10.21307/stattrans-2020-022
Received – 31.01.2020; accepted – 30.06.2020

Small area estimation: its evolution in five decades

Malay Ghosh[1]

ABSTRACT

The paper is an attempt to trace some of the early developments of small area estimation. The basic papers such as the ones by Fay and Herriott (1979) and Battese, Harter and Fuller (1988) and their follow-ups are discussed in some details. Some of the current topics are also discussed.

Key words: template, article, journal.

1. Prologue

Small area estimation is witnessing phenomenal growth in recent years. The vastness of the area makes it near impossible to cover each and every emerging topic. The review articles of Ghosh and Rao (1994), Pfeffermann (2002, 2013) and the classic text of Rao (2003) captured the contemporary research of that time very successfully. But the literature continued growing at a very rapid pace. The more recent treatise of Rao and Molina (2015) picked up many of the later developments. But then there came many other challenging issues, particularly with the advent of "big data", which started moving the small area estimation machine faster and faster. It seems real difficult to cope up with this super-fast development.

In this article, I take a very modest view towards the subject. I have tried to trace the early history of the subject up to some of the current research with which I am familiar. It is needless to say that the topics not covered in this article far outnumber those that are covered. Keeping in mind this limitation, I will make a feeble attempt to trace the evolution of small area estimation in the past five decades.

2. Introduction

The first and foremost question that one may ask is "what is small area estimation"? Small area estimation is any of several statistical techniques involving estimation of parameters in small 'sub-populations' of interest included in a larger 'survey'. The term 'small area' in this context generally refers to a small geographical area such as a county, census tract or a school district. It can also refer to a 'small domain' cross-classified by

[1]Department of Statistics, University of Florida, Gainesville, FL, USA. E-mail: ghoshm@stat.ufl.edu. ORCID: https://orcid.org/0000-0002-8776-7713.

several demographic characteristics, such as age, sex, ethnicity, etc. I want to emphasize that it is not just the area, but the 'smallness' of the targeted population within an area that constitutes the basis for small area estimation. For example, if a survey is targeted towards a population of interest with prescribed accuracy, the sample size in a particular subpopulation may not be adequate to generate similar accuracy. This is because if a survey is conducted with sample size determined to attain prescribed accuracy in a large area, one may not have the resources available to conduct a second survey to achieve similar accuracy for smaller areas.

A domain (area) specific estimator is 'direct' if it is based only on the domain-specific sample data. A domain is regarded as 'small' if domain-specific sample size is not large enough to produce estimates of desired precision. Domain sample size often increases with population size of the domain, but that need not always be the case. This requires use of 'additional' data, be it either administrative data not used in the original survey, or data from other related areas. The resulting estimates are called 'indirect' estimates that 'borrow strength' for the variable of interest from related areas and/or time periods to increase the 'effective' sample size. This is usually done through the use of models, mostly 'explicit', or at least 'implicit' that links the related areas and/or time periods.

Historically, small area statistics have long been used, albeit without the name "small area" attached to it. For example, such statistics existed in eleventh century England and seventeenth century Canada based on either census or on administrative records. Demographers have long been using a variety of indirect methods for small area estimation of population and other characteristics of interest in postcensal years. I may point out here that the eminent role of administrative records for small area estimation cannot but be underscored even today. A very comprehensive review article in this regard is due to Erciulescu, Franco and Lahiri (2020).

In recent years, the demand for small area statistics has greatly increased worldwide. The need is felt for formulating policies and programs, in the allocation of government funds and in regional planning. For instance, legislative acts by national governments have created a need for small area statistics. A good example is SAIPE (Small Area Income and Poverty Estimation) mandated by the US Legislature. Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on local socio-economic conditions. Small area estimation is of particular interest for the transition economics in central and eastern European countries and the former Soviet Union countries. In the 1990's these countries have moved away from centralized decision making. As a result, sample surveys are now used to produce estimates for large areas as well as small areas.

3. Examples

Before tracing this early history, let me cite a few examples that illustrate the ever increasing current day importance of small area estimation. One important ongoing small

area estimation problem at the U.S. Bureau of the Census is the small area income and poverty estimation (SAIPE) project. This is a result of a Bill passed by the US House of Representatives requiring the Secretary of Commerce to produce and publish at least every two years beginning in 1996, current data related to the incidence of poverty in the United States. Specifically, the legislation states that "to the extent feasible", the secretary shall produce estimates of poverty for states, counties and local jurisdictions of government and school districts. For school districts, estimates are to be made of the number of poor children aged 5-17 years. It also specifies production of state and county estimates of the number of poor persons aged 65 and over.

These small area statistics are used by a broad range of customers including policy makers at the state and local levels as well as the private sector. This includes allocation of Federal and state funds. Earlier the decennial census was the only source of income distribution and poverty data for households, families and persons for such small geographic areas. Use of the recent decennial census data pertaining to the economic situation is unreliable especially as one moves further away from the census year. The first SAIPE estimates were issued in 1995 for states, 1997 for counties and 1999 for school districts. The SAIPE state and county estimates include median household income number of poor people, poor children under age 5 (for states only), poor children aged 5-17, and poor people under age 18. Also starting 1999, estimates of the number of poor school-aged children are provided for the 14,000 school districts in the US (Bell, Basel and Maples, 2016).

Another example is the Federal-State Co-Operative Program (FSCP). It started in 1967. The goal was to provide high-quality consistent series of post-censal county population estimates with comparability from area to area. In addition to the county estimates, several members of FSCP now produce subcounty estimates as well. Also, the US Census Bureau used to provide the Treasury Department with Per Capita Income (PCI) estimates and other statistics for state and local governments receiving funds under the general revenue sharing program. Treasury Department used these statistics to determine allocations to local governments within the different states by dividing the corresponding state allocations. The total allocation by the Treasury Dept. was $675 billion in 2017.

United States Department of Agriculture (USDA) has long been interested in prediction of areas under corn and soybeans. Battese, Harter and Fuller (JASA, 1988) considered the problem of predicting areas under corn and soybeans for 12 counties in North-Central Iowa based on the 1978 June enumerative survey data as well as Landsat Satellite Data. The USDA statistical reporting Service field staff determined the area of corn and soybeans in 37 sample segments of 12 counties in North Central Iowa by interviewing farm operators. In conjunction with LANDSAT readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels in the 12 counties.

There are many more examples. An important current day example is small area "poverty mapping" initiated by Elbers, Lanjouw and Lanjouw (2003). This was extended as well as substantially refined by Molina and Rao (2010) and many others.

4. Synthetic Estimation

An estimator is called 'Synthetic' if a direct estimator for a large area covering a small area is used as an indirect estimator for that area. The terminology was first used by the U.S. National Center for Health Statistics. These estimators are based on a strong underlying assumption is that the small area bears the same characteristic for the large area.

For example, if $y_1, \cdots, y_m$ are the direct estimates of average income for m areas with population sizes $N_1, \cdots, N_m$, we may use the overall estimate $\bar{y}_s = \sum_{j=1}^{m} N_j y_j / N$ for a particular area, say, i ,where $N = \sum_{j=1}^{m} N_j$. The idea is that this synthetic estimator has less mean squared error (MSE) compared to the direct estimator y_i if the bias $\bar{y}_s - y_i$ is not too strong. On the other hand, a heavily biased estimator can affect the MSE as well.

One of the early use of synthetic estimation appears in Hansen, Hurwitz and Madow (1953, pp 483-486). They applied synthetic regression estimation in the context of radio listening. The objective was to estimate the median number of radio stations heard during the day in each of more than 500 counties in the US. The direct estimate y_i of the true (unknown) median M_i was obtained from a radio listening survey based on personal interviews for 85 county areas. The selection was made by first stratifying the population county areas into 85 strata based on geographical region and available radio service type. Then one county was selected from each stratum with probability proportional to the estimated number of families in the counties. A subsample of area segments was selected from each of the sampled county areas and families within the selected area segments were interviewed.

In addition to the direct estimates, an estimate x_i of M_i, obtained from a mail survey was used as a single covariate in the linear regression of y_i on x_i. The mail survey was first conducted by sampling 1,000 families from each county area and mailing questionnaires. The x_i were biased due to nonresponse (about 20% response rate) and incomplete coverage, but were anticipated to have high correlation with the M_i. Indeed, it turned out that $\text{Corr}(y_i, x_i) = .70$. For nonsampled counties, regression synthetic estimates were $\hat{M}_i = .52 + .74 x_i$.

Another example of Synthetic Estimation is due to Gonzalez and Hoza (JASA, 1978, pp 7-15). Their objective was to develop intercensal estimates of various population characteristics for small areas. They discussed synthetic estimates of unemployment where the larger area is a geographic division and the small area is a county.

Specifically, let p_{ij} denote the proportion of labor force in county i that corresponds to cell j $(j = 1, \cdots, G)$. Let u_j denote the corresponding unemployment rate for cell j

based on the geographic division where county i belongs. Then, the synthetic estimate of the unemployment rate for county i is given by $u_i^* = \sum_{j=1}^{G} p_{ij} u_j$. These authors also suggested synthetic regression estimate for unemployment rates.

While direct estimators suffer from large variances and coefficients of variation for small areas, synthetic estimators suffer from bias, which often can be very severe. This led to the development of composite estimators, which are weighted averages of direct and synthetic estimators. The motivation is to balance the design bias of synthetic estimators and the large variability of direct estimators in a small area.

Let y_{ij} denote the characteristic of interest for the jth unit in the ith area; $j = 1, \cdots, N_i; i = 1, \cdots, m$. Let x_{ij} denote some auxiliary characteristic for the jth unit in the ith local area. Note that the population means are $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$ and $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$. We denote the sampled observations as y_{ij}, $j = 1, \cdots, n_i$ with corresponding auxiliary variables x_{ij}, $j = 1, \cdots, n_i$. Let $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$. $\bar{x}_i$ is obtained from the sample. In addition, one needs to know $\bar{X}_i$, the population average of auxiliary variables.

A Direct Estimator (Ratio Estimator) of $\bar{Y}_i$ is $\bar{y}_i^R = (\bar{y}_i/\bar{x}_i)\bar{X}_i$. The corresponding Ratio Synthetic Estimator of $\bar{Y}_i$ is $(\bar{y}_s/\bar{x}_s)\bar{X}_i$, where $\bar{y}_s = \sum_{i=1}^{m} N_i \bar{y}_i / \sum_{i=1}^{m} N_i$ and $\bar{x}_s = \sum_{i=1}^{m} N_i \bar{x}_i / \sum_{i=1}^{m} N_i$. A Composite Estimator of $\bar{Y}_i$ is

$$(n_i/N_i)\bar{y}_i + (1 - n_i/N_i)(\bar{y}_s/\bar{x}_s)\bar{X}_i',$$

where $\bar{X}_i' = (N_i - n_i)^{-1} \sum_{j=n_i+1}^{N_i} x_{ij}/(N_i - n_i)$. Note $N_i\bar{X}_i = n_i\bar{x}_i + (N_i - n_i)\bar{X}_i'$. All one needs to know is the population average $\bar{X}_i$ in addition to the already known sample average $\bar{x}_i$ to find $\bar{X}_i'$. Several other weights in forming a linear combination of direct and synthetic estimators have also been proposed in the literature.

The Composite Estimator proposed in the previous paragraph can be given a model-based justification as well. Consider the model $y_{ij} \overset{\text{ind}}{\sim} (bx_{ij}, \sigma^2 x_{ij})$. Best linear unbised estimator of b is obtained by minimizing $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - bx_{ij})^2/x_{ij}$. The solution is $\hat{b} = \bar{y}_s/\bar{x}_s$. Now estimate $\bar{Y}_i = (\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} y_{ij})/N_i$ by $\sum_{j=1}^{n_i} y_{ij}/N_i + \hat{b}\sum_{j=n_i+1}^{N_i} x_{ij}/N_i$. This simplifies to the expression given in the previous paragraph. Holt, Smith and Tomberlin (1979) provided more general model-based estimators of this type.

5. Model-Based Small Area Estimation

Small area models link explicitly the sampling model with random area specific effects. The latter accounts for between area variation beyond that is explained by auxiliary variables. We classify small area models into two broad types. First, the "area level" models that relate small area direct estimators to area-specific covariates. Such models are necessary if unit (or element) level data are not available. Second, the "unit level" models that relate the unit values of a study variable to unit-specific covariates. Indirect

estimators based on small area models will be called "model-based estimators".

The model-based approach to small area estimation offers several advantages. First, "optimal" estimators can be derived under the assumed model. Second, area specific measures of variability can be associated with each estimator unlike global measures (averaged over small areas) often used with traditional indirect estimators. Third, models can be validated from the sample data. Fourth, one can entertain a variety of models depending on the nature of the response variables and the complexity of data structures. Fifth, the use of models permits optimal prediction for areas with no samples, areas where prediction is of utmost importance.

In spite of the above advantages, there should be a cautionary note regarding potential model failure. We will address this issue to a certain extent in Section 7 when we discuss benchmarking. Another important issue that has emerged in recent years, is design-based evaluation of small area predictors. In particular, design-based mean squared errors (MSE's) is of great appeal to practitioners and users of small area predictors, because of their long-standing familiarity with the latter. Two recent articles addressing this issue are Pfeffermann and Ben-Hur (2018) and Lahiri and Pramanik (2019).

The classic small area model is due to Fay and Herriot (JASA, 1979) with Sampling Model: $y_i = \theta_i + e_i,\ i = 1,\ldots,m$ and Linking Model: $\theta_i = x_i^T b + u_i,\ i = 1,\ldots,m$. The target is estimation of the $\theta_i,\ i = 1,\ldots,m$. It is assumed that e_i are independent $(0, D_i)$, where the D_i are known and the u_i are iid $(0, A)$, where A is unknown. The assumption of known D_i can be put to question because they are, in fact, sample estimates. But the assumption is needed to avoid nonidentifiablity in the absence of microdata. This is evident when one writes $y_i = x_i^T b + u_i + e_i$. In the presence of microdata, it is possible to estimate the D_i as well. An example appears in Ghosh, Myung and Moura (2018).

A few notations are needed to describe the Fay-Herriot procedure. Let $y = (y_1,\ldots,y_m)^T$; $\theta = (\theta_1,\ldots,\theta_m)^T$: $e = (e_1,\ldots,e_m)^T$; $u = (u_1,\ldots,u_m)^T$; $X^T = (x_1,\ldots,x_m)$; $b = (b_1,\ldots,b_p)^T$. We assume X has rank $p(< m)$. In vector notations, we write $y = \theta + e$ and $\theta = Xb + u$.

For known A, the best linear unbiased predictor (BLUP) of θ_i is $(1 - B_i)y_i + B_i x_i^T \tilde{b}$ where $\tilde{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y,\ V = \text{Diag}(D_1 + A, \cdots, D_m + A)$ and $B_i = D_i/(A + D_i)$. The BLUP is also the best unbiased predictor under assumed normality of y and θ.

It is possible to give an alternative Bayesian formulation of the Fay-Herriott model. Let $y_i|\theta_i \overset{\text{ind}}{\sim} \text{N}(\theta_i, D_i)$; $\theta_i|b \overset{ind}{\sim} N(x_i^T b, A)$. Then the Bayes estimator of θ_i is $(1 - B_i)y_i + B_i x_i^T b$, where $B_i = D_i/(A + D_i)$. If instead we put a uniform(R^p) prior for b, the Bayes estimator of θ_i is the same as its BLUP. Thus, there is a duality between the BLUP and the Bayes estimator.

However, in practice, A is unknown. A hierarchical prior joint for both b and A is $\pi(b,A)=1$. (Morris, 1983, JASA). Otherwise, estimate A to get the resulting empirical Bayes or empirical BLUP. We now describe the latter.

There are several methods for estimation of A. Fay and Herriot (1979) suggested solving iteratively the two equations (i) $\tilde{b}=(X^TV^{-1}X)^{-1}X^TV^{-1}y$ and (ii) $\sum_{i=1}^m(y_i-x_i^T\tilde{b})^2=m-p$. The motivation for (i) comes from the fact that $\tilde{b}$ is the best linear unbiased estimator (BLUE) of b when A is known. The second is a method of moments equation noting that the expectation of the left hand side equals $m-p$.

The Fay-Herriot method does not provide an explicit expression for A. Prasad and Rao (1990, JASA) suggested instead a unweighted least squares approach, which provides an exact expression for A. Specifically, they proposed the estimator $\hat{b}_L=(X^TX)^{-1}X^Ty$. Then $E||y-X\hat{b}_L||^2=(m-p)A+\sum_{i=1}^m D_i(1-r_i),\ r_i=x_i^T(X^TX)^{-1}x_i,\ i=1,\cdots,m$. This leads to $\hat{A}_L=\max\left(0,\frac{||y-X\hat{b}_L||^2-\sum_{i=1}^m D_i(1-r_i)}{m-p}\right)$ and accordingly $\hat{B}_i^L=D_i/(\hat{A}_L+D_i)$. The corresponding estimator of θ is $\hat{\theta}_i^{EB}=(1-\hat{B}_i^L)y_i+\hat{B}_i^Lx_i^T\tilde{b}(\hat{A}_L)$, where

$$\tilde{b}(\hat{A}_L)=[X^TV^{-1}(\hat{A}_L)X]^{-1}X^TV^{-1}(\hat{A}_L)y.$$

Prasad and Rao also found an approximation to the mean squared eror (Bayes risk) of their EBLUP or EB estimators. Under the subjective prior $\theta_i\overset{\text{ind}}{\sim}\mathrm{N}(x_i^Tb,A)$, the Bayes estimator of θ_i is $\hat{\theta}_i^B=(1-B_i)y_i+B_ix_i^Tb$, $B_i=D_i/(A+D_i)$. Also, write $\tilde{\theta}_i^{EB}(A)=(1-B_i)y_i+B_ix_i^T\tilde{b}(A)$. Then $E(\hat{\theta}_i^{EB}-\theta_i)^2=E(\hat{\theta}_i^B-\theta_i)^2+E(\tilde{\theta}_i^{EB}(A)-\hat{\theta}_i^B)^2+E(\hat{\theta}_i^{EB}-\tilde{\theta}_i^{EB}(A))^2$. The cross-product terms vanish due to their method of estimation of A, by a result of Kackar and Harville (1984). The first term is the Bayes risk if both b and A were known. The second term is the additional uncertainty due to estimation of b when A is known. The third term accounts for further uncertainty due to estimation of A.

One can get exact expressions $E(\theta_i-\hat{\theta}_i^B)^2=D_i(1-B_i)=g_{1i}(A)$, say and $E(\hat{\theta}_i^{EB}(A)-\hat{\theta}_i^B)^2=B_i^2x_i^T(X^TV^{-1}X)^{-1}x_i=g_{2i}(A)$, say. However, the third term, $E(\hat{\theta}_i^{EB}-\hat{\theta}_i^{EB}(A))^2$ needs an approximation. An approximate expression correct up to $O(m^{-1})$, i.e. the remainder term is of $o(m^{-1})$, as given in Prasad and Rao, is $2B_i^2(D_i+A)^{-1}A^2\sum_{i=1}^m(1-B_i)^2/m^2=g_{3i}(A)$, say. Further, an estimator of this MSE correct up to $O(m^{-1})$ is $g_{1i}(\hat{A})+g_{2i}(\hat{A})+2g_{3i}(\hat{A})$. This approximation is justified by noticing $E[g_{1i}(\hat{A})]=g_{1i}(A)-g_{3i}(A)+o(m^{-1})$.

A well-known example where this method has been applied is estimation of median income of four-person families for the 50 states and the District of Columbia in the United States. The U.S. Department of Health and Human Services (HHS) has a direct need for such data at the state level in formulating its energy assistance program for low-income families. The basic source of data is the annual demographic supplement to the March sample of the Current Population Survey (CPS), which provides the median income of

four-person families for the preceding year. Direct use of CPS estimates is usually undesirable because of large CV's associated with them. More reliable results are obtained these days by using empirical and hierarchical Bayesian methods.

Here sample estimates of the state medians for the current year (c) as obtained from the Current Population Survey (CPS) were used as dependent variables. Adjusted census median (c) defined as the base year (the recent most decennial census) census median (b) times the ratio of the BEA PCI (per capita income as provided by the Bureau of Economic Analysis of the United States Bureau of the Census) in year (c) to year (b) was used as an independent variable. Following the suggestion of Fay (1987), Datta, Ghosh, Nangia and Natarajan (1996) used the census median from the recent most decennial census as a second independent variable. The resulting estimates were compared against a different regression model employed earlier by the US Census Bureau.

The comparison was based on four criteria recommended by the panel on small area estimates of population and income set up by the US committee on National Statistics. In the following, we use e_i as a generic notation for the ith small area estimate, and $e_{i.TR}$ the "truth", i.e. the figure available from the recent most decennial census. The panel recommended the following four criteria for comparison.
Average Relative Absolute Bias $= (51)^{-1}\sum_{i=1}^{51}|e_i - e_{i,TR}|/e_{i,TR}$.
Average Squared Relative Bias $= (51)^{-1}\sum_{i=1}^{51}(e_i - e_{i,TR})^2/e_{i,TR}^2$.
Average Absolute Bias $= (51)^{-1}\sum_{i=1}^{51}|e_i - e_{i,TR}|$.
Average Squared Deviation $= (51)^{-1}\sum_{i=1}^{51}(e_i - e_{i,TR})^2$.

Table 1 compares the Sample Median, the Bureau Estimate and the Empirical BLUP according to the four criteria as mentioned above.

Table 1. Average Relative Absolute Bias, Average Squared Relative Bias, Average Absolute Bias and Average Squared Deviation (in 100,000) of the Estimates.

	Bureau Estimate	Sample Median	EB
Aver. rel. bias	0.325	0.498	0.204
Aver. sq. rel bias	0.002	0.003	0.001
Aver. abs. bias	722.8	1090.4	450.6
Aver. sq. dev.	8.36	16.31	3.34

There are other options for estimation of A. One due to Datta and Lahiri (2000) uses the MLE or the residual MLE (RMLE). With this estimator, g_{3i}^{DL} is approximated by $2D_i^2(A+D_i)^{-3}[\sum_{i=1}^{m}(A+D_i)^{-2}]^{-1}$, while g_{1i} and g_{2i} remain unchanged. Finally, Datta, Rao and Smith (2005), went back to the original Fay-Herriot method of estimation of A, and obtained $g_{3i}^{DRS} = 2D_i^2(A+D_i)^{-3}m[\sum_{i=1}^{m}(A+D_i)^{-2}]^{-1}$.

The string of inequalities

$$m^{-1}\sum_{i=1}^{m}(A+D_i)^2 \geq [m^{-1}\sum_{i=1}^{m}(A+D_i)]^2 \geq m^2[\sum_{i=1}^{m}(A+D_i)^{-1}]^2$$

leads to $g_{3i}^{PR} \geq g_{3i}^{DRS}$. Another elementary inequality $\sum_{i=1}^{m}(A+D_i)^{-2} \geq m^{-1}[\sum_{i=1}^{m}(A+D_i)^{-1}]^2$ leads to $g_{3i}^{DRS} \geq g_{3i}^{DL}$. All three expressions for g_{3i} equal when $D_1 = \ldots = D_m$. It is also pointed out in Datta, Rao and Smith that while both Prasad-Rao and REML estimators of A lead to the same MSE estimator correct up to $O(m^{-1})$, a further adjustment to this estimator is needed when one uses either the the ML or the Fay-Herriot estimator of A. The simulation study undertaken in Datta, Rao and Smith also suggests that the ML, REML and Fay-Herriot methods of estimation of A perform quite similarly in regards to the MSE of the small area estimators, but the Prasad-Rao approach usually leads to a bigger MSE. However, they all perform far superior to the MSE's of the direct estimators.

Over the years, other approaches to MSE estimation have appeared, some quite appealing as well as elegant. The two most prominent ones appear to be the ones due to Jackknife and Bootstrap. Jiang and Lahiri (2001), Jiang, Lahiri and Wan (2002), Chen and Lahiri (2002), Das, Jiang and Rao (2004) all considered Jackknife estimation of the MSE that avoid the detailed Taylor series expansion of the MSE. A detailed discussion paper covering many aspects of related methods appears in Jiang and Lahiri (2006). Pfeffermann and Tiller (2005), Butar and Lahiri (2003) considered bootstrap estimation of the MSE. More recently, Yoshimori and Lahiri (2014) considered adjusted likelihood estimation of A. Booth and Hobert (1998) introduced a conditional approach for estimating the MSE. In a different vein, Lahiri and Rao (1995) dispensed with the normality assumption of the random effects, assuming instead its eighth moment in the Fay-Herriot model.

Pfeffermann and Correa (2012) proposed an approach which they showed to perform much better than the "classical" jackknife and bootstrap methods. Pfeffermann and Ben-Hur (2018) used a similar approach for estimating the design-based MSE of model-based predictors.

Small area estimation problems have also been considered for the general exponential family model. Suppose $y_i|\theta_i$ are independent with $f(y_i|\theta_i) = \exp[y_i\theta_i - \psi(\theta_i) + h(y_i)]$, $i = 1,\ldots,m$. An example is the Bernoulli (p_i) where $\theta_i = \text{logit}(p_i) = \log(p_i/(1-p_i))$ and Poisson(λ_i) where $\theta_i = \log(\lambda_i)$. One models the θ_i as independent N$(x_i^T b, A)$ and proceeds. Alternately, use beta priors for the p_i and gamma priors for the λ_i.

The two options are to estimate the prior parameters either using an empirical Bayes approach or alternately using a hierarchical Bayes approach assigning distributions to the prior parameters. The latter was taken by Ghosh et al. (1998) in a general framework. Other work is due to Raghunathan (1993) and Malec et al. (1997). A method for MSE estimation in such contexts appears in Jiang and Lahiri (2001).

Jiang, Nguyen and Rao (2011) evaluated the performance of a BLUP or EBLUP using only the sampling model $y_i \overset{\text{ind}}{\sim} (\theta_i, D_i)$. Recall $B_i = D_i/(A+D_i)$. Then

$$E[\{(1-B_i)y_i + B_i x_i^T b - \theta_i\}^2 | \theta_i] = (1-B_i)^2 D_i + B_i^2 (\theta_i - x_i^T b)^2.$$

Noting that $E[(y_i - x_i^T b)^2 | \theta_i] = D_i + (\theta_i - x_i^T b)^2$, an unbiased estimator of the above MSE is $(1-B_i)^2 D_i - B_i^2 D_i + B_i^2 (y_i - x_i^T b)^2$. When one minimizes the above with respect to b and A, then the resulting estimators of of b and A are referred to as observed best predictive estimators. The corresponding estimators of the θ_i are referred to as the "observed best predictors". These authors suggested Fay-Herriot or Prasad-Rao method for estimation of b and A.

6. Model Based Small Area Estimation: Unit Specific Models

Unit Specific Models are those where observations are available for the sampled units in the local areas. In addition, unit-specific auxiliary information is available for these sampled units, and possibly for the non-sampled units as well.

To be specific, consider m local areas where the ith local area has N_i units with a sample of size n_i. We denote the sampled observations by $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, m$. Consider the model

$$y_{ij} = x_{ij}^T b + u_i + e_{ij}, j = 1, \ldots N_i, i = 1, \ldots, m.$$

The u_i's and e_{ij}'s are mutually independent with the u_i iid $(0, \sigma_u^2)$, and the e_{ij} independent $(0, \sigma^2 \psi_{ij})$.

The above nested error regression model was considered by Battese, Harter and Fuller (BHF, 1988), where y_{ij} is the area devoted to corn or soybean for the jth segment in the ith county; $x_{ij} = (1, x_{ij1}, x_{ij2})^T$, where x_{ij1} denotes the no. of pixels classified as corn for the jth segment in the ith county and x_{ij2} denotes the no. of pixels classified as soybean for the jth segment in the ith county; $b = (b_0, b_1, b_2)^T$ is the vector of regression coefficients. BHF took $\psi_{ij} = 1$. The primary goal of BHF was to estimate the $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, the population average of area under corn or soybean for the 12 areas in North Central Iowa, N_i denoting the population size in area i.

A second example appears in Ghosh and Rao (1994). Here y_{ij} denotes wages and salaries paid by the jth business firm in the ith census division in Canada and $x_{ij} = (1, x_{ij})^T$, where x_{ij} is the gross business income of the jth business firm in the ith census division. In this application, $\psi_{ij} = x_{ij}$ was found more appropriate than the usual model involving homoscedasticity.

I consider in some detail the BHF model. Their ultimate goal was to estimate the population means $\bar{Y}_i = (N_i)^{-1} \sum_{j=1}^{N_i} y_{ij}$, In matrix notation, we write $y_i = (y_{i1}, \ldots, y_{in_i})^T$,

$X_i = (x_{i1}, \ldots, x_{in_i})^T$, $e_i = (e_{i1}, \ldots, e_{in_i})^T$, $i = 1, \ldots, m$. Thus, the model is rewritten as

$$y_i = X_i b + u_i 1_{n_i} + e_i, i = 1, \ldots, m.$$

Clearly, $E(y_i) = X_i b$ and $V_i = V(y_i) = \sigma_e^2 I_{n_i} + \sigma_u^2 J_{n_i}$, where J_{n_i} denote the matrix with all elements equal to 1. Write $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ and $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$. The target is estimation of $\bar{X}_i^T b + u_i 1_{n_i}$, where $\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}$, $i = 1, \ldots, m$.

For known σ_u^2 and σ_e^2, the BLUP of $\bar{x}_i^T b + u_i 1_{n_i}$ is $(1 - B_i) y_i + B_i \bar{x}_i^T \tilde{b}$, where $B_i = (\sigma_e^2/n_i)/(\sigma_e^2/n_i + \sigma_u^2)$ and $\tilde{b} = (\sum_{i=1}^m X_i^T V_i^{-1} X_i)^{-1} (\sum_{i=1}^m X_i^T V_i^{-1} y_i)$. Hence, the BLUP of $\bar{X}_i^T b + u_i 1_{n_i}$ is $[(1 - B_i)[\bar{y}_i + (\bar{X}_i - \bar{x}_i)^T \tilde{b}] + B_i \bar{X}_i^T \tilde{b}$.

BHF used method of moment estimation to get unbiased estimators of unknown σ_u^2 and σ_e^2. The EBLUP of $\bar{X}_i^T b + u_i$ is now found by substituting these estimates of σ_u^2 and σ_e^2 in the BLUP formula. Estimation of σ_e^2 is based on the moment identity

$$E[\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i - (x_{ij} - \bar{x}_i)^T \tilde{b}]^2 = (n - m - p_1),$$

where p_1 is the number of non-zero x deviations. The second moment identity is given by

$$E[\sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - x_{ij})^T \hat{b})^2] = (n - p)\sigma_e^2 + \sigma_u^2 [m - \sum_{i=1}^{m} n_i^2 \bar{x}_i^T (X^T X)^{-1} \bar{x}_i],$$

where $\hat{b} = (X^T X)^{-1} X^T y$, $y = (y_1^T, \ldots, y_m^T)^T$. If this results in a negative estimator of σ_u^2, they set the estimator equal to zero.

Of course, the method of moments estimators can be replaced by maximum likelihood, REML or other estimators as discussed in the previous section. Alternately, one can adopt a hierarchical Bayesian approach as taken in Datta and Ghosh (1991). First, it may be noted that if the variance components σ_e^2 and σ_u^2 were known, a uniform prior on b leads to a HB estimator of $\bar{X}_i^T b + u_i$, which equals its BLUP. Another interesting observation is that the BLUP of $\bar{X}_i^T b + u_i$ depends only on the variance ratio $\sigma_u^2/\sigma_e^2 = \lambda$, say. Rather than assigning priors separately for σ_u^2 and σ_e^2, it suffices to assign a prior to λ. This is what was proposed in Datta and Ghosh (1991), who assigned a Gamma prior to λ. The Bayesian approach of Datta and Ghosh (1991) did also acccommodate the possibility of multiple random effects.

7. Benchmarking

The model-based small area estimates, when aggregated, may not equal the corresponding estimated for the larger area. On the other hand, the direct estimate for a larger area, for example, a national level estimate, is quite reliable. Moreover, matching the latter may be a good idea, for instance to maintain consistency in publication, and very

often for protection against model failure. The latter may not always be achieved, for example in time series models, as pointed out by Wang, Fuller and Qu (2008).

Specifically, suppose θ_i is the ith area mean and $\theta_T = \sum_{i=1}^m w_i\theta_i$ is the overall mean, where w_j may be the known proportion of units in the jth area. The direct estimate for θ_T is $\sum_{i=1}^m w_i\hat{\theta}_i$. Also, let $\tilde{\theta}_i$ denote an estimator of θ_i based on a certain model. Then $\sum_{i=1}^m w_i\tilde{\theta}_i$ is typically not equal to $\sum_{i=1}^m w_i\hat{\theta}_i$

In order to address this, people have suggested (i) ratio adjusted estimators

$$\hat{\theta}_i^{RA} = \hat{\theta}_i^G(\sum_{j=1}^m w_j\hat{\theta}_j)/(\sum_{j=1}^m w_j\hat{\theta}_j^G)$$

and (ii) difference adjusted estimator $\hat{\theta}_i^{DA} = \hat{\theta}_i^G + \sum_{j=1}^m w_j\hat{\theta}_j - \sum_{j=1}^m w_j\hat{\theta}_j^G$, where $\hat{\theta}_j^G$ is some generic model-based estimator of θ_j.

One criticism against such adjustments is that a common adjustment is used for all small areas regardless of their precision. Wang, Fuller and Qu (2008) proposed instead minimizing $\sum_{j=1}^m \phi_j E(e_j - \theta_j)^2$ for some specified weights $\phi_j(>0)$ subject to the constraint $\sum_{j=1}^m w_j e_j = \hat{\theta}_T$. The resulting estimator of θ_i is

$$\hat{\theta}_i^{WFQ} = \hat{\theta}_i^{BLUP} + \lambda_i(\sum_{j=1}^m w_j\hat{\theta}_j - \sum_{j=1}^m w_j\hat{\theta}_j^{BLUP}),$$

where $\lambda_i = w_i\phi_i^{-1}/(\sum_{j=1}^m w_j^2\phi_j^{-1})$.

Datta, Ghosh, Steorts and Maples (2011) took instead a general Bayesian approach and minimized $\sum_{j=1}^m \phi_j[E(e_j - \theta_j)^2|data]$ subject to $\sum_{j=1}^m w_j e_j = \hat{\theta}_T$ and obtained the estimator $\hat{\theta}_i^{AB} = \hat{\theta}_i^B + \lambda_i(\sum_{j=1}^m w_j\hat{\theta}_j - \sum_{j=1}^m w_j\hat{\theta}_j^B)$, with the same λ_i. This development is similar in spirit to those of Louis (1984) and Ghosh (1992) who proposed constrained Bayes and empirical Bayes estimators to prevent overshrinking. The approach of Datta, Ghosh, Steorts and Maples extends readily to multiple benchmarking constraints. In a frequentist context. Bell, Datta and Ghosh (2013) extended the work of Wang, Fuller and Qu (2008) to multiple benchmarking constraints.

There are situations also when one needs two-stage benchmarking. A current example is the cash rent estimates of the Natural Agricultural Statistics Service (NASS), where one needs the dual control of matching the aggregate of county level cash rent estimates to the corresponding agricultural district (comprising of several counties) level estimates, and the aggregate of the agricultural district level estimates to the final state level estimate. Berg, Cecere and Ghosh (2014) adopted an approach of Ghosh and Steorts (2013) to address the NASS problem.

Second order unbiased MSE estimators are not typically available for ratio adjusted benchmarked estimators. In contrast, second order unbiased MSE estimators are available for difference adjusted benchmarked estimators, namely, $\hat{\theta}_i^{DB} = \hat{\theta}_i^{EB} + (\sum_{j=1}^m w_j \hat{\theta}_j - \sum_{j=1}^m w_j \hat{\theta}_j^{EB})$. Steorts and Ghosh (2013) have shown that $\text{MSE}(\hat{\theta}_i^{DB}) = \text{MSE}(\hat{\theta}_i^{EB}) + g_4(A) + o(m^{-1})$, where $\text{MSE}(\hat{\theta}_i^{EB})$ is the same as the one given in Prasad and Rao (1990), and

$$g_4(A) = \sum_{i=1}^m w_i^2 B_i^2 (D_i + A) - \sum_{i=1}^m \sum_{j=1}^m w_i w_j B_i B_j x_i^T (X^T V^{-1} x_j).$$

We may recall that $B_i = D_i/(A+D_i)$, $X^T = (x_1, \ldots, x_m)$ and $V = \text{Diag}(A+D_1, \ldots, A+D_m)$ in the Fay-Herriot model. A second order unbiased estimator of the benchmarked EB estimator is thus $g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}) + g_{4i}(\hat{A})$.

There are two available approaches for self benchmarking that do not require any adjustment to the EBLUP estimators. The first, proposed in You and Rao (2002) for the Fay-Herriot model replaces the estimator $\hat{b}$ in the EBLUP by an estimator which depends both on $\hat{b}$ as well as the weights w_i. This changes the MSE calculation. Recall the Prasad-Rao MSE of the EBLUP given by $\text{MSE}(\hat{\theta}_i^{EB}) = g_{1i} + g_{2i} + g_{3i}$, where $g_{1i} = D_i(1-B_i)$, $g_{2i} = B_i^2 x_i^T (X^T V^{-1} X)^{-1} x_i$ and $g_{3i} = 2D_i^2 (A+D_i)^{-3} m^{-2} \{\sum_{j=1}^m (A+D_j)^2\}$. For the Benchmarked EBLUP, g_{2i} changes.

The second approach is by Wang, Fuller and Qu (2008) and it uses an augmented model with new covariates (x_i, w_i, D_i). This second approach was extended by Bell, Datta and Ghosh (2013) to accommodate multiple benchmarking constraints.

8. Fixed versus Random Area Effects

A different but equally pertinent issue has recently surfaced in the small area literature. This concerns the need for random effects in all areas, or whether even fixed effects models would be adequate for certain areas. Datta, Hall and Mandal (DHM, 2011) were the first to address this problem. They suggested essentially a preliminary test-based approach, testing the null hypothesis that the common random effect variance was zero. Then they used a fixed or a random effects model for small area estimation based on acceptance or rejection of the null hypothesis. This amounted to use of synthetic or regression estimates of all small area means upon acceptance of the null hypothesis, and composite estimates which are weighted averages of direct and regression estimators otherwise. Further research in this area is due to Molina, Rao and Datta (2015).

The DHM procedure works well when the number of small areas is moderately large, but not necessarily when the number of small areas is very large. In such situations, the null hypothesis of no random effects is very likely to be rejected. This is primarily due to a

few large residuals causing significant departure of direct estimates from the regression estimates. To rectify this, Datta and Mandal (2015) proposed a Bayesian approach with "spike and slab" priors. Their approach amounts to taking $\delta_i u_i$ instead of u_i for random effects where the δ_i and the u_i are independent with δ_i iid Bernoulli(γ) and u_i iid N$(0, \sigma_u^2)$.

In contrast to the spike and slab priors of Datta and Mandal (2015), Tang, Ghosh, Ha and Sedransk (2018) considered a different class of priors that meets the same objective. as the spike and slab priors, but uses instead absolutely continuous priors. These priors allow different variance components for different small areas, in contrast to the priors of Datta and Mandal, who considered prior variances to be either zero or else common across all small areas. This seems to be particularly useful when the number of small areas is very large, for example, when one considers more than 3000 counties of the US, where one expects a wide variation in the county effects. The proposed class of priors, is usually referred to as "global-local shrinkage priors" (Carvalho, Polson and Scott (2010); Polson and Scott (2010)).

The global-local priors, essentially scale mixtures of normals, are intended to capture potential "sparsity", which means lack of significant contribution by many of the random effects, by assigning large probabilities to random effects close to zero, but also identifying random effects which differ significantly from zero. This is achieved by employing two levels of parameters to express prior variances of random effects. The first, the "local shrinkage parameters", acts at individual levels, while the other, the "global shrinkage parameter" is common for all random effects. This is in contrast to Fay and Herriot (1979) who considered only one global parameter. These priors also differ from those of Datta and Mandal (2015), where the variance of random effects is either zero or common across all small areas.

Symbolically, the random effects u_i have independent N$(0, \lambda_i^2 A)$ priors. While the global parameter A tries to cause an overall shrinking effect, the local shrinkage parameters λ_i^2 are useful in controlling the degree of shrinkage at the local level. If the mixing density corresponding to local shrinkage parameters is appropriately heavy-tailed, the large random effects are almost left unshrunk. The class of "global-local" shrinkage priors includes the three parameter beta normal (TPBN) priors (Armagon, Clyde and Dunson, 2011) and Generalized Double Pareto priors (Armagon, Dunson and Lee, 2012). TPBN includes the now famous horseshoe (HS) priors (Scott and Berger, 2010) and the normal-exponential-gamma priors (Griffin and Brown, 2005).

As an example, consider estimation of 5-year (2007–2011) county-level overall poverty ratios in the US. There are 3,141 counties in the data set. The covariates are foodstamp participation rates. The map given in Figure 1 gives the poverty ratios for all the counties of US. Some salient findings from these calculations are given below.

(i) Estimated poverty ratios are between 3.3% (Borden County, TX) and 47.9% (Shannon County, SD). The median is 14.7%.

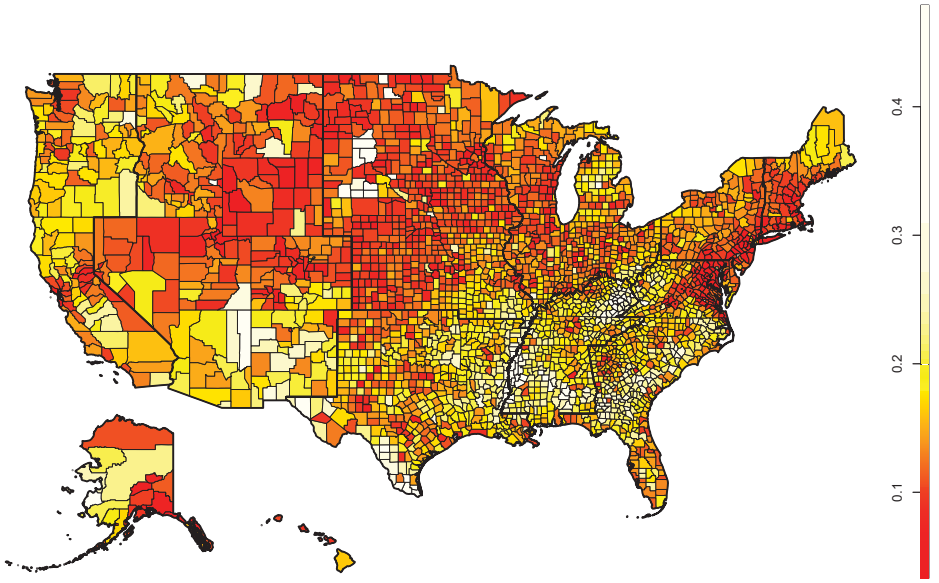


Figure 1: Map of posterior means of θ's.

(ii) In Mississippi, Georgia, Alabama and New Mexico, 55%+ counties have poverty rates $>$ the third quartile (18.9%).
(iii) In New Hampshire, Connecticut, Rhode Island, Wyoming, Hawaii and New Jersey, 70%+ counties have poverty rates $<$ the first quartile (11.1%).
(iv) Examples of counties with high poverty ratios are Shannon, SD; Holmes, MS; East Carroll, LA; Owsley, KY; Sioux, IA.
(v) Examples of counties with large random effects are Madison, ID; Whitman, WA; Harrisonburg, VA; Clarke, GA; Brazos, TX.

Dr. Pfeffermann suggested splitting the counties, whenever possible, into a few smaller groups, and then use the same global-local priors for estimating the random effects separately for the different groups. From a pragmatic point of view, this may sometimes be necessary for faster implementation. It seems though that the MCMC implementation even for such a large number of counties was quite easy since all the conditionals were standard disributions, and samples could be generated easily from these distributions at each iteration.

9. Variable Transformation

Often the normality assumption can be justified only after transformation of the original data. Then one performs the analysis based on the transformed data, but transform back properly to the original scale to arrive at the final predictors. One common example is transformation of skewed positive data, for example, income data where log transfor-

mation gets a closer normal approximation. Slud and Maiti (2006) and Ghosh and Kubokawa (2015) took this approach, providing final results for the back-transformed original data.

For example, consider a multiplicative model $y_i = \phi_i \eta_i$ with $z_i = \log(y_i)$, $\theta_i = \log(\phi_i)$ and $e_i = \log(\eta_i)$. Consider the Fay-Herriott (1979) model (i) $z_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, D_i)$ and (ii) $\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(x_i^T\beta, A)$. θ_i has the $\mathrm{N}(\hat{\theta}_i^B, D_i(1-B_i))$ posterior with $\hat{\theta}_i^B = (1-B_i)z_i + B_i x_i^T\beta$, $B_i = D_i/(A+D_i)$. Now $E(\phi_i|z_i) = E[\exp(\theta_i)|z_i] = \exp[\hat{\theta}_i^B + (1/2)D_i(1-B_i)]$.

Another interesting example is the variance stabilizing transformation. For example, suppose $y_i \overset{\text{ind}}{\sim} \mathrm{Bin}(n_i, p_i)$. The arcsine transformation is given by $p_i = \sin^{-1}(2p_i - 1)$. The back transformation is $p_i = (1/2)[1 + \sin(\theta_i)]$.

A third example is the Poisson model for count data. There $y_i \overset{\text{ind}}{\sim} \mathrm{Poisson}(\lambda_i)$. Then one models $z_i = y_i^{1/2}$ as independent $\mathrm{N}(\theta_i, 1/4)$ where where $\theta_i = \lambda_i^{1/2}$. An added advantage in the last two examples is that the assumption of known sampling variance, which is really untrue, can be avoided.

10. Final Remarks

As acknowledged earlier, the present article leaves out a large number of useful current day topics in small area estimation. I list below a few such topics which are not covered at all here. But there are are many more. People interested in one or more of the topics listed below and beyond should consult the book of Rao and Molina (2015) for their detailed coverage of small area estimation and an excellent set of references for these topics.

- Design consistency of small area estimators.
- Time series models.
- Spatial and space-time models.
- Variable Selection.
- Measurement errors in the covariates.
- Poverty counts for small areas.
- Empirical Bayes confidence intervals.
- Robust small area estimation.
- Misspecification of linking models.

- Informative sampling.
- Constrained small area estimation.
- Record Linkage.
- Disease Mapping.
- Etc, Etc., Etc.

Acknowledgements

I am indebted to Danny Pfeffermann for his line by line reading of the manuscript and making many helpful suggestions, which improved an earlier version of the paper. Partha Lahiri read the original and the revised versions of this paper very carefully, and caught many typos. A comment by J.N.K. Rao was helpful. The present article is based on the Morris Hansen Lecture delivered by Malay Ghosh before the Washington Statistical Society on October 30, 2019. The author gratefully acknowledges the Hansen Lecture Committee for their selection.

REFERENCES

ARMAGAN, A., CLYDE, M., and DUNSON, D. B., (2013). Generalized double pareto shrinkage. *Statistica Sinica*, 23, pp. 119–143.

ARMAGAN, A., DUNSON, D. B., LEE, J., and BAJWA, W. U., (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100, pp. 1011–1018.

BATTESE, G. E., HARTER, R. M., and FULLER, W. A., (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, pp. 28–36.

BELL, W. R., DATTA, G, S., and GHOSH, M., (2013). Benchmarking small area estimators. *Biometrika*, 100, pp. 189–202.

BELL, W. R., BASEL, W. W., and MAPLES, J. J., (2016). An overview of U.S. Census Bureau's Small Area Income and Poverty Estimation Program. In *Analysis of Poverty Data by Small Area Estimation*. Ed. M. Pratesi. Wiley, UK, pp. 349–378.

BERG, E., CECERE, W., and GHOSH, M., (2014). Small area estimation of county level farmland cash rental rates. *Journal of Survey Statistics and Methodology*, 2, pp. 1–37. Bivariate hierarchical Bayesian model for estimating cropland cash rental rates at the county level. *Survey Methodology*, in press.

BOOTH, J. G., HOBERT, J., (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93, pp. 262–272.

BUTAR, F. B., LAHIRI, P., (2003). On measures of uncertainty of empirical Bayes small area estimators. *Journal of Statistical Planning and Inference*, 112, pp. 63–76.

CARVALHO, C. M., POLSON, N. G., SCOTT, J. G., (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97, pp. 465–480.

CHEN, S., LAHIRI, P., (2003).A comparison of different MPSE estimators of EBLUP for the Fay-Herriott model. In *Proceedings of the Section on Survey Research Methods*. Washington, D.C. American Statistical Association, pp. 903–911.

DAS, K., JIANG, J., RAO, J. N. K., ((2004). Mean squared error of empirical predictor. *Annals of Statistics*, 32, pp. 818–840.

DATTA, G. S., GHOSH, M., (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics*, 19, pp. 1748–1770.

DATTA, G., GHOSH, M., NANGIA, N., and NATARAJAN, K., (1996). Estimation of median income of four-person families: a Bayesian approach. In *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*. Eds. D. Berry, K. Chaloner and J. Geweke. North Holland, pp. 129–140.

DATTA, G. S., LAHIRI. P., (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, pp. 613–627.

DATTA, G. S., RAO, J. N. K., and SMITH, D. D., (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92, pp. 183–196.

DATTA, G. S., GHOSH, M., STEORTS, R., and MAPLES, J. J., (2011). Bayesian benchmarking with applications to small area estimation. *TEST*, 20, pp. 574–588.

DATTA, G. S., HALL, P., and MANDAL, A., (2011). Model selection and testing for the presence of small area effects and application to area level data. *Journal of the American Statistical Association*, 106, pp. 362–374.

DATTA, G. S., MANDAL, A., (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110, pp. 1735–1744.

ELBERS, C., LANJOUW, J. O., and LANJOUW, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71. pp. 355–364.

ERCIULESCU, A. L., FRANCO, C., and LAHIRI, P., (2020). Use of administrative records in small area estimation. To appear in *Administrative Records for Survey Methodology*. Eds. P. Chun and M. Larson. Wiley, New York.

FAY, R. E., (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*. Eds. R. Platek, J.N.K. Rao, C-E Sarndal and M.P. Singh. Wiley New York, pp. 91–102.

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

GHOSH, M., (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, pp. 533–540.

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

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

GHOSH, M., STEORTS, R., (2013). Two-stage Bayesian benchmarking as applied to small area estimation. *TEST*, 22, pp. 670–687.

GHOSH, M., KUBOKAWA, T., and KAWAKUBO, Y., (2015). Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation. *Biomerika*, 102, pp. 647–659.

GHOSH, M., MYUNG, J., and MOURA, F. A. S., (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44, pp. 101–115.

GONZALEZ, M. E., HOZA, C., (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73, pp. 7–15.

GRIFFIN, J. E., BROWN, P. J., (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5, pp. 171–188.

HANSEN, M. H., HURWITZ, W. N., and MADOW, W. G., (1953). *Sample Survey Methods and Theory*. Wiley, New York.

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

JIANG, J., LAHIRI, P., (2001). Empirical best prediction of small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53, pp. 217–243.

JIANG, J., LAHIRI, P., and WAN, S-M., (2002). A unified jackknife theory. *The Annals of Statistics*, 30, pp. 1782–1810.

JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation (with discussion). *TEST*, 15, pp. 1–96.

JIANG, J., NGUYEN, T., and RAO, J. S., (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, 106, pp. 732–745.

KACKAR, R. N., HARVILLE, D. A., (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, pp. 853–862.

LAHIRI, P., RAO, J. N. K., (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, pp. 758–766.

LAHIRI, P., PRAMANIK, S., (2019). Evaluation of synthetic small area estimators using design-based methods. *Austrian Journal of Statistics*, 48, pp. 43–57.

LOUIS, T. A., (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, pp. 393–398.

MALEC, D., DAVIS, W. W., and CAO, X., (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics and Medicine*, 18, pp. 3189–3200.

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, pp. 369–385.

MOLINA, I., RAO, J. N. K., and DATTA, G. S., (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects. *Survey Methodology*.

PFEFFERMANN, D., TILLER, R. B., (2005). Bootstrap approximation of prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26, pp. 893–916.

MORRIS, C. N., (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78, pp. 47–55.

POLSON, N. G., SCOTT, J. G., (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9, pp. 501–538.

PFEFFERMANN, D., (2002). Small area estimation: new developments and direction. *International Statistical Review*, 70, pp. 125–143.

PFEFFERMANN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28, pp. 40–68.

PRASAD, N. G. N., RAO, J. N. K., (1990). The estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 85, pp. 163–171.

RAGHUNATHAN, T. E., (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, 88, pp. 1444–1448.

RAO, J. N. K., (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2, pp. 145–169.

RAO, J. N. K., (2006). Inferential issues in small area estimation: some new developments. *Statistics in Transition*, 7, pp. 523–526.

RAO, J. N. K., Molina, I., (2015). *Small Area Estimation*, 2nd Edition. Wiley, New Jersey.

SCOTT, J. G., BERGER, J. O., (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38, pp. 2587–2619.

SLUD, E. V., MAITI, T., (2006). Mean squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society, B*, 68, pp. 239–257.

TANG, X., GHOSH, M., Ha, N-S., and SEDRANSK, J., (2018). Modeling random effects using global-local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, 113, pp. 1476–1489.

WANG, J., FULLER, W. A., and QU, Y., (2008). Small area estimation under restriction. *Survey Methodology*, 34, pp. 29–36.

YOSHIMORI, M., LAHIRI, P., (2014). A new adjusted maximum likelihood method for the Fay-Herriott small area model. *Journal of Multivariate Analysis*, 124, pp. 281–294.

YOU, Y., RAO, J. N. K., and HIDIROGLOU, M. A., (2013). On the performance of self-benchmarked small area estimators under the Fay-Herriott area level model. *Survey Methodology*, 39, pp. 217–229.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 23–29, DOI 10.21307/stattrans-2020-023
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Julie Gershunskaya[1]

1. Introduction

I would like to begin by congratulating Professor Ghosh for his many contributions to small area estimation, both as an original researcher and effective communicator of complex ideas. The current paper provides a lucid overview of the history and developments in small area estimation (SAE) and offers a synopsis of some of the most recent innovations. As is well illustrated in the paper, the development of the field is driven by real-world demands and problems emerging in actual applications. Let us ponder on this practical side of the SAE methodology that, by offering a set of tools and concepts, provides an *engineering framework* for present day official statistics.

From the very beginning of large-scale sample surveys in the official statistics, there was the realization that the survey practice should be based on both theoretical developments and clear practical strategy. Morris Hansen (1987) applied the term "*total survey design*" to describe the fusion of theory and operational planning, a paradigm used from the early days of sampling surveys at the U.S. Bureau of Census. In a similar spirit, P. C. Mahalanobis (1946) characterized the whole complex of activities involved in the managing of large-scale sample surveys in the Indian Statistical Institute by calling it "*statistical engineering*".

Traditionally, a great deal of theory, experimentation, and practical considerations are focused on the design stage of sample surveys. Yet, no matter how well the survey is designed, there is a growing demand in extracting ever more information from already collected data. Even more, in many present day surveys, the required "unplanned" domains number in thousands. In such an environment, the production of small domain estimates becomes a substantial part of a large-scale enterprise. Developments in the SAE field address the demands by providing survey practitioners with necessary gear, whereas an applied statistician acts as *engineer* that employs a variety of available tools and creates an appropriate operational plan.

[1] U. S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, DC 20212, USA.
E-mail: gershunskaya.julie@bls.gov. ORCID: https://orcid.org/0000-0002-0096-186X.

2. Model building considerations

To illustrate some aspects of the planning and model development for estimation in small domains, I will describe, in broad strokes, considerations involved in the model choice for the U.S. Bureau of Labor Statistics' Current Employment Statistics (CES) survey. The specific context that affects approaches to small domain modeling in CES includes:

- the tight production timeline, where estimates are produced monthly within only a few weeks after the data collection;
- the demand for estimates over a large number of small areas. Monthly estimates are published for about 10 thousands domains defined by intersections of detailed industry and geography. Of those, roughly 40 percent of domains have sufficient sample, so that direct sample-based estimates are deemed reliable for the use in publication; the other domains may have only a handful of sample units and require modeling;
- the dynamic and heterogeneous nature of the population of business establishments, a feature that could generally manifest itself – thus affecting the model fit – in two ways: 1. in the form of a frequent appearance of sample-influential observations or; 2. as irregularities in the signal for groups of domains.

Because of the above characteristics of the CES survey process, essential requirements for any model considered in CES are (i) computational scalability, (ii) flexibility of modeling assumptions, and (iii) robustness to model outliers. To demonstrate how the above aspects are taken into account, we examine three models.

Our baseline model M0 is the classical Fay-Herriot area level model. In the Bayesian formulation, using the notation of Professor Ghosh's paper, the *sampling model* for domain $i = 1..., m$ is

$$y_i \mid \theta_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \tag{1}$$

and the *linking model* is

$$\theta_i \mid \mathbf{b} \overset{ind}{\sim} N\left(x_i^T \mathbf{b}, A\right). \tag{2}$$

The parsimonious structure and the ease of implementation of the FH model make it particularly appealing under the tight CES production schedule. The posterior mean in the form of the weighted average of direct sample based and synthetic estimators has clear intuitive interpretation, thus facilitating communication of the reasoning to a wider, less quantitatively oriented, community.

However, the dynamic nature of the population of business establishments affects the FH model fit and reduces the attractiveness of the model in two important respects:

1) On the one hand, sampling model (1) is not robust to extreme y_i values. Noisy direct estimates y_i could result from the appearance of influential observations in the sample data. In the ideal world, the additional variability induced by noisy sample data would be reflected in larger values of respective variances $D_i's$, that are assumed to be known. If that would be the case, larger $D_i's$ would lessen the influence of noisy $y_i's$ on the model fit. In practice, however, true variances are not known, and the usual method is to plug in values based on a generalized variance function (GVF). Such plug-in $D_i's$ may not properly reflect the amount of noise in respective $y_i's$.
2) On the other hand, the linking model (2) normality assumption may fail, for example, when groups of domains form clusters or when some domains deviate from the linearity assumption $x_i^T\mathbf{b}$. This is especially likely to happen when a large number of domains is included in the same model.

In model M1, we address the concern regarding the non-robustness of sampling model (1). Here, sample-based estimates $\hat{D}_i$ of variances D_i are treated as data and modeled jointly with $y_i's$. The joint modeling approach was considered by Arora and Lahiri (1997), You and Chapman (2006), Dass et al. (2012), Liu et al. (2014), among others. Model M1 is related to the model proposed by Maiti et al. (2014) who used the EM algorithm for estimation of the model parameters within the empirical Bayes paradigm. The Bayesian extension of the model was developed by Sugasawa et al. (2017). Assume in domain i, $i=1...,m$, the following model M1 holds for pair $\left(y_i,\hat{D}_i\right)$:

$$y_i \mid \theta_i, D_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \qquad \theta_i \mid \mathbf{b}, A \overset{ind}{\sim} N\left(x_i^T\mathbf{b}, A\right), \tag{3}$$

$$\hat{D}_i \mid D_i \overset{ind}{\sim} G\left(\frac{n_i-1}{2}, \frac{n_i-1}{2D_i}\right), \qquad D_i \mid \gamma \overset{ind}{\sim} IG\left(a_i, c_i\gamma\right), \tag{4}$$

where (3) is the usual FH model for the point estimate and (4) describes a companion model for observed variance $\hat{D}_i$ (here, direct sample-based estimates of variances are termed "observed variances" in the model input context); $G(\cdot)$ and $IG(\cdot)$ denote the gamma and inverse gamma distributions, respectively; γ is an unknown parameter; a_i and c_i are positive known constants, Sugasawa et al. (2017) suggested the choice of $a_i=2$ and $c_i=n_i^{-1}$, n_i is the number of respondents in domain i.

Although model M1 mitigates the effect caused by noisy direct sample estimates, it still ignores the problem of possible deviations from the normality assumption in linking model (2). When there is a large number of domains, we can more fully explore the underlying structure and relax the assumption of linking model (2) by replacing the normality with a finite mixture of normal distributions. Model M2, proposed by Gershunskaya and Savitsky (2020), is given by (5) and (6):

$$y_i \mid \theta_i, D_i \overset{ind}{\sim} N\left(\theta_i, D_i\right), \qquad \theta_i \mid \boldsymbol{\pi}, \mathbf{b}_0, \mathbf{b}, A \overset{ind}{\sim} \sum_{k=1}^{K} \pi_k N\left(b_{0k} + \tilde{\mathbf{x}}_i^T \mathbf{b}, A\right), \qquad (5)$$

$$\hat{D}_i \mid D_i \overset{ind}{\sim} G\left(\frac{sn_i}{2}, \frac{sn_i}{2D_i}\right), \qquad D_i \mid \boldsymbol{\gamma}, \boldsymbol{\pi} \overset{ind}{\sim} \sum_{k=1}^{K} \pi_k IG\left(2, \exp\left(z_i^T \boldsymbol{\gamma}_k\right)\right). \qquad (6)$$

In this model, we assume the existence of K latent clusters having cluster-specific intercepts b_{0k}, $k = 1,\ldots,K$, and common variance A; in addition, we relax the inverse gamma assumption of (4) by specifying a mixture of the inverse gamma distributions with the cluster-specific coefficient vectors $\boldsymbol{\gamma}_k$; z_i is a vector of covariates for the variance model for area i; s is a model parameter that regulates the shape and scale of the gamma distribution, it depends on the quality of variance estimates.

The Stan modeling language and the Variational Bayes algorithm within Stan proved to be effective in fitting the above models.

3. Model selection and evaluation plan

Due to the tight CES production schedule, a *production* model has to be chosen in advance, before a statistician obtains the actual data. Models for CES are pre-selected and pre-evaluated based on a comparison to historical employment series derived from the universe of data that is available from an administrative source, known as the Quarterly Census of Employment and Wages (QCEW) program. These data become available to BLS on a quarterly basis with the time lag of 6 to 9 months after the reference date and are considered a "gold standard" for CES. After an evaluation based on several years of data, that include periods of economic growths and downturns, the best model from a set of candidates would be accepted for the use in production.

Thus, the availability of a "gold standard" defines the CES strategy for the model development and evaluation. This approach differs from the usual model selection and checking methods used in statistics, yet it is common for government agencies.

4. Real-time analysis protocol

The quality of the production model is regularly re-assessed based on newly available data from QCEW. This kind of evaluation can be performed only post hoc, several months after the publication of CES estimates. While the "gold standard" based approach of model selection and evaluation works well overall and provides reassurance and the perception of objectivity of the chosen model, the following question remains: Suppose a particular model (say, model M2) is accepted for the production based on its historical performance; however, what if in a given month during the production such history-based best model would fit poorly for some of the domains? To diagnose possible problems in the real production time, analysts have to be equipped with formal tests and graphical tools allowing the efficient detection of potential problems, and with the guidelines for ways to proceed whenever problems arise.

One example of a tool for the routine diagnostics of outlying cases is given by the model-based domain screening procedure proposed by Gershunskaya and Savitsky (2020). The idea for this procedure is to flag the domains whose direct estimates $y_i's$ have low probability of following the *posterior predictive distribution* obtained based on the model. The list of "suspect" domains is sent to analysts for checking; analysts review the list and decide if the reason for a given extreme direct estimate is one of the following: (i) the deficiency of the domain sample or (ii) a failure of modeling assumptions. In general, if the domain sample size is small, the outlyingness of the direct sample estimate would likely be attributed to the deficiency of the sample; in such a case, analysts would decide to rely on the model estimate for this domain. For domains with larger samples, the direct estimates may be deemed more reliable than the model-based estimates. In addition, to these general considerations, analysts would also have the ability to check the responses in the suspect domains to determine if there are any erroneous reports overlooked at the editing stage. Such reports would have to be corrected or removed from the sample. Analysts may also possess the knowledge of additional facts that may guide their decision, such as, information about the economic events not reflected in the modeling assumptions or, conversely, in the available sample.

5. Summary

The growing demand for estimates in "unplanned" domains instigated development of the SAE methods. Theoretical advances in SAE over past five decades, along with the proliferation of powerful computers and software, invited even more, ever increasing demand in estimates for small areas. Contemporary small area estimation becomes a *large-scale* undertaking. The present day *statistical engineers*

require development of tools – as well as philosophy and guidelines – for the *quality control* in the *production environment* to help ensure estimates in small domains are reliable and impartial.

Acknowledgement

The views expressed here are those of the discussant and do not necessarily constitute the policies of the U.S. Bureau of Labor Statistics.

REFERENCES

ARORA, V., LAHIRI, P., (1997). On the superiority of the Bayesian methods over the BLUP in small area estimation problems. *Statistica Sinica* 7, pp. 1053–1063.

BUREAU OF LABOR STATISTICS, (2004). Employment, Hours, and Earnings from the Establishment Survey, BLS Handbook of Methods, Washington, DC: US Department of Labor.

DASS, S. C., MAITI, T., REN, H., SINHA, S., (2012). Confidence interval estimation of small area parameters shrinking both means and variances. *Survey Methodology*, 38, pp. 173–187.

GERSHUNSKAYA, J., SAVITSKY, T. D., (2020) Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, Vol. 8, Issue 2, pp. 181–205, https://doi.org/10.1093/jssam/smz004.

HANSEN, M. H., (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, Vol. 2, No. 2, pp. 180–190.

LIU, B., LAHIRI, P., KALTON, G., (2014). Hierarchical Bayes modelling of survey-weighted small area proportions. *Survey Methodology*, Vol. 40, No. 1, pp. 1–13.

MAHALANOBIS, P. C., (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, pp. 325–378.

MAITI, T., H. REN, A. SINHA, (2014). Prediction Error of Small Area Predictors Shrinking Both Means and Variances. *Scandinavian Journal of Statistics*, 41, pp. 775–790.

STAN DEVELOPMENT TEAM, (2017). Stan modeling Language User's Guide and Reference Manual, Version 2.17.0 [Computer Software Manual], available at http://mc-stan.org/. Accessed February 28, 2019.

SUGASAWA, S., TAMAE, H., KUBOKAWA, T., (2017). Bayesian Estimators for Small Area Models Shrinking Both Means and Variances. *Scandinavian Journal of Statistics*, 44, pp. 150–167.

YOU, Y., CHAPMAN, B., (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology,* 32, pp. 97–103.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 30–34, DOI 10.21307/stattrans-2020-024
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Ying Han[1]

1. Introduction

I would like to thank Prof. Ghosh for his significant contributions to small area estimation, not only for his phenomenal research, but also for the talents that he cultivated and brought into this field. It is my great honor to be an invited discussant of Prof. Ghosh's paper "Small Area Estimation: Its Evolution in Five Decades".

In the paper, Prof. Ghosh presents a nice overview of the history and development of small area estimation. He clearly explains the reason why small area estimation techniques are important in providing accurate estimates for small regions or domains, illustrates the increasing importance of small area estimation through examples in different fields, introduces different small area estimates developed from area-level and unit-level models, etc. He traces back to the starting point of small area estimation, demonstrates its development, and shows us its bright future.

The basic idea of small area estimation is to increase the effective sample size by borrowing strengths from variable of interest from other related areas. This is primarily done by linking related small areas using auxiliary information related to the variable of interest. The auxiliary information often comes from administrative records. So, the availability of good administrative records is of great importance to small area estimation. As Prof. Ghosh said in the paper, "the eminent role of administrative records for small area estimation cannot but be underscored even today."

The unit-level small area estimation models require the joint observations on the variable of interest y and the auxiliary variables x for the sampled units in small areas. If administrative records are used, we need to know which administrative record represents the same population unit as one in the survey data. Consider the case where the data comes from two separate files: one survey data set containing the observations on y and an administrative data set containing the observations on x. If a unique and error-free identifier exists in both files, the two files can be linked without any errors and a merged dataset with joint observations on y and x is obtained. Under this data layout, a huge literature on small area estimation is available. We refer reader to Rao and Molina (2015), Jiang and Lahiri (2006), and Pfeffermann (2013).

[1]Gallup.Inc, USA. E-mail: ying_han@gallup.com. ORCID: https://orcid.org/0000-0003-0082-5654.

Most of the time, however, such identifier is not available in either the survey data set or the administrative data set. In this case, the administrative records can rarely be used for unit-level small area estimation model. This limits the application of small area estimation. Record linkage, a data integration technique, is a potential approach to link the files even when a unique and error-free identifier is not available. The application of record linkage extends the application of small area estimation to the case when administrative records cannot be linked to the survey data by using unique identifiers. This is one of the most emerging topics that was not covered in Prof. Ghosh overview paper. In this discussion, I would like to provide a brief description on this topic.

2. Probabilistic Record Linkage

Record linkage, or exact matching, is a technique to identify records for the same entity (e.g., person, household, etc.) that are from two or more files when a unique, error-free identifier (such as Social Security Number) is missing. The first theoretical framework for record linkage was developed by Fellegi and Sunter (1969). A linked dataset, created by record linkage, is of great interest to analysts interested in certain specialized multivariate analysis, which would be otherwise either impossible or difficult without advanced statistical expertise as variables are stored in different files.

However, the linked dataset is subject to linkage errors. If one simply ignores the linkage errors, analysis of the linked data could yield misleading results in a scientific study. Neter et al. (1965) demonstrated that a relatively small amount of linkage errors could lead to substantial bias in estimating a regression relationship. Therefore, the importance of accounting for linkage errors in statistical analysis cannot be overemphasized. In the past couple of decades, researchers have been focused on how to correct the bias caused by linkage errors when fitting linear regression model on linked data. Chambers (2009), Kim and Chambers (2012), Samart and Chambers (2014) tackled the problem from the second analyst point of view, assuming that they can only get access to the linked data and limited information is available about the linkage process. In contrast, Lahiri and Larsen (2005) solved the problem from the primary analyst point of view by taking advantage of the summary information generated during the record linkage process. But there is little literature on the how to apply small area estimation on the linked data generated through record linkage process.

The importance of integrating probabilistic record linkage in small area estimation was highlighted in the SAE International Statistical Institute Satellite Meeting held in Paris during July 10-12 , 2017. In his keynote address at the meeting, Professor Partha Lahiri introduce the concept of merging survey data with administrative records together through record linkage technique to obtain an enhanced dataset for small area estimation. It can cut down the cost in data collection by preventing the need to collect new survey data with all necessary information. Later, I worked with Professor Lahiri in proposing a unified way for performing small area estimation using data from multiple

files. A brief description of the methodology is provided in the next section. Readers interested in the details are referred to Lahiri (2017), Han (2018), and Han and Lahiri (2019).

3. Small area estimation within linked data

We are interested in predicting an area-specific parameter, which can be expressed as a function of fixed effects and random effects related to the conditional distribution of y given x. For simplicity, we restrict our research to the case where the observations on y and x come from two files, rather than more than two files (e.g., one survey dataset and multiple administrative data sets). Suppose the observations on y (x) are available for a sample S_y (S_x) and are recorded in file F_y (F_x). The matching status between any record in F_y and any record in F_x is unknown. We assume that (1) there is no duplicate in either F_y or F_x, (2) $S_y \subset S_x$, and (3) the records in both files can be partitioned into small areas without error.

We propose a general integrated model to propagate the uncertainty of the linkage process in the later estimation step under the assumption of data availability described above. The model is developed from a primary analyst point of view. The primary analyst can get access to the original two files, which contains both the separate observations on y and x and the values of matching fields (a set of variables for record linkage). The proposed model is built directly on the data values from the original two files (rather than on data in the linked dataset) and is based on the actual record linkage method that is used (rather than making a strong assumption on the linkage process afterwards). The general proposed integrated model includes three important components: a unit-level small area estimation model, a linkage error model, and a two-class mixture model on comparison vectors. The unit-level model is used to characterize the relationship between y and x in the target population. The linkage error model is used to characterize the randomness of the linkage process. It is developed by exploiting the relationship between X^* (the unobserved x values corresponding to the observed y values in F_y) and X (the observed x values in F_x). It is the key to the general integrated model, serving as a connector between the unit-level small area model and the record linkage model. The two-class mixture model is used to estimate the probability of a record pair being a match given the observed data and designate all record pairs into links and non-links.

Under the general integrated model, we provide a general methodology for obtaining an empirical best prediction (EBP) estimator of an area-specific mixed parameter. The unified jackknife resampling method proposed by Jiang et al. (2002) and its alternative proposed by Lohr and Rao (2009) can be used to estimate the mean squared error of the empirical best prediction estimator. The jackknife methods proposed by Jiang et al. (2002) and Lohr and Rao (2009) require closed-form expressions for the mean squared error (MSE) and conditional mean squared error (CMSE) of the best prediction estimator (BP), respectively. So, the choice of the jackknife methods depends on whether a

closed-form expression for MSE or CMSE is available.

Application of the general methodology is not limited to the mutual independence of measurements. It can be applied to measurements that are correlated within small areas but independent across small areas. Unit-level models such as general linear model with correlated sampling errors within small areas, general linear mixed model with nested errors can all be considered. To illustrate our general methodology, we consider the situation where the unit-level small area model of the general integrated model is set to be the general linear mixed model with block diagonal covariance structure. The Best Prediction (BP) estimator for the mixed parameter is derived under the general integrated model. The conditional mean squared error (CMSE) of its corresponding Best Prediction (BP) Estimator can be expressed in a closed form, making it possible to estimate its mean squared error using the jackknife method provided by Lohr and Rao (2009).

As a special example, we consider the estimation of small area means when a nested error linear model is used. We provide two methods for estimating the unknown parameters: the Maximum Likelihood (ML) method and the Pseudo Maximum Likelihood (PML) method. We also discuss the use of numerical algorithms in approximating the maximum likelihood estimates (MLE), including Newton-Raphson method and Fish scoring algorithm, and further propose a quasi-scoring algorithm in order to reduce the computational burden.

4. Summary

Due to the increasing demand of small area estimation in different fields and the accessibility of administrative records, it is of great interest for researchers and analysts to use probabilistic record linkage in extracting additional information from administrative records as additional auxiliary variable in unit-level small area models. It is an example of the more recent topics in small area estimation that are not covered by Prof. Ghosh in his overview paper. As Prof. Ghosh said, "the vastness of the area makes it near possible to cover each and every emerging topic". That means, small area estimation is still under its rapid development driven by its high demand, and it is a field full of vitality.

REFERENCES

CHAMBERS, R., (2009). Regression analysis of probability-linked data. *Statisphere*, 4.

FELLEGI, I., SUNTER, A., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, pp. 1183–1210.

HAN, Y., (2018). Statistical inference using data from multiple files combined through record linkage. PhD thesis, University of Maryland.

HAN, Y., LAHIRI, P., (2019). Statistical analysis with linked data. *Journal of the American Statistical Association*, 87, pp. S139–S157.

JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation. *Test*, 15(1), pp. 1–96.

JIANG, J., LAHIRI, P., WAN, S. W., (2002). A unified jackknife theory for empirical best prediction with m-estimation. *Annals of Statistics*, 30(6), pp. 1782–1810.

KIM, J., CHAMBERS, R., (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56(9), pp. 2756–2770.

LAHIRI, P., (2017). Small area estimation with linked data. Keynote address at the ISI Satellite Meeting on Small Area Estimation, Paris, France, July, pp. 10–12.

LAHIRI, P., LARSEN, M., (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469), pp. 222–230.

LOHR, S. L., RAO, K., (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Boimetrika*, 96(2), pp. 457–468.

NETER, J., MAYNES, E., RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response error. *Journal of the American Statistical Association*, 60(312), pp. 1005–1027.

PFEFFERMAN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28(1), pp. 40–68.

RAO, J., MOLINA, I., (2015). Small Area Estimation. Wiley, second edition.

SAMART, K., CHAMBERS, R., (2014). Linear regression with nested errors using probability-linked data. *Australian and New Zealand Journal of Statistics*, 56(1), pp. 27–46.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 35–39, DOI 10.21307/stattrans-2020-025
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Yan Li[1]

Prof. Ghosh leads us step gradually into the realm of small area estimation (SAE) through the evolution of SAE for the past five decades, introducing various SAE methods of synthetic estimators, composite estimators, and model-based estimators for small area parameters, mean squared error approximations, adjustment methods of benchmarking and transformation, etc. The paper broadens and deepens our understanding of different perspectives of the SAE and provides a few illustrative real-life applications. It is a great review paper for general audience, especially for our graduate students in survey statistics and related areas, who wish to have a snapshot of the SAE research.

Prof. Ghosh focuses his review on the inferential aspects of the two celebrated small area models ----- the Fay-Herriot (FH) area model and the unit level nested error regression (NER) model. In the implementation of these models, variable selection plays a vital role and my discussion centers around this topic, which complements Professor Ghosh's paper.

There is a vast literature on variable selection, a subtopic of model selection. We refer to the Institute of Mathematical Statistics Monograph edited by Lahiri (2001) for different approaches and issues in model selection and the book by Jiang and Nguyen (2015) for model selection methodology especially designed for mixed models. Variable selection methods for general linear mixed model can be, of course, applied to select variables for the FH and NER models as they are special cases of the general linear mixed model. Many data analysts not familiar with mixed models, however, use software meant for linear regression models to select variables. This approach may result in loss of efficiency in variable selection. Lahiri and Suntornchost (2015) and Li and Lahiri (2019) proposed simple adjustment methods so that the data users can select reasonable models by calculating their favorite variable selection criteria, such as AIC, BIC, Mallow's C_p, and adjusted R^2, which are developed for standard linear regression model assuming independent identically distributed (*iid*) errors. The goal of the two

[1] Joint Program in Survey Methodology and Department of Epidemiology and Biostatistics, University of Maryland, College Park, USA. E-mail: yli6@umd.edu. ORCID: https://orcid.org/0000-0001-8241-7464.

papers is to propose adjustment methods, instead of advocating a specific variable selection method. Cai et al. (2020), with the same goal, creatively combined the two variable selection methods (Lahiri and Suntornchost, 2015 and Li and Lahiri, 2019) and proposed a variable selection method for another popular two-fold subarea model.

The above-mentioned three methods consider commonly used variable selection criteria under a standard regression model with *iid* errors, including

1) Adjusted R^2: adjRsq $= 1 - \frac{MSE_k}{MST}$,
2) Mallows C_P: $C_P = \frac{SSE_k}{MSE_K} + 2k - n$,
3) AIC: $AIC = 2k + n \cdot \log\left(\frac{SSE_k}{n}\right)$, and
4) BIC: $BIC = k \cdot \log(n) + n \cdot \log\left(\frac{SSE_k}{n}\right)$,

where

$$MSE_k = \frac{SSE_k}{n-k} \text{ with}$$
$$SSE_k = y^T[I - X_k(X_k^T X_k)^{-1}X_k^T]y, \text{ and}$$
$$MST = \frac{SST}{n-1} \text{ with}$$
$$SST = y^T[I - n^{-1}11^T]y.$$

Note that $y = (y_1, \dots, y_n)$ is a vector of observations on the dependent variable; X_k is a $n \times (1+k)$ design matrix with columns of one's and k auxiliary variables, corresponding to the intercept and k unknown parameters; $SSE_k(MSE_k)$ is the SSE (MSE) based on the standard regression model for $k = 1, \dots, K$. Here K is the total number of auxiliary variables considered in model selection and n is the sample size. When $k = K$, $MSE_K = \frac{SSE_K}{n-K}$ is the MSE based on the full model with all K auxiliary variables. As noted, these variable selection criteria can be expressed as a smooth function of MSE_k and MST.

Next, adjustments proposed for the three small area models are briefly discussed before above variable selection criteria designed for standard regression model can be used.

1. Consider the Fay-Herriot area model given by:

$$y_i = \theta_i + e_i \text{ and } \theta_i = x_i^T\beta + v_i, \quad (1)$$

where θ_i is the unobserved true mean for small area i; y_i is the survey-weighted estimate of θ_i; $v_i$is the random effect for small area i; v_i's and e_i's are independent with $v_i \sim N(0, A)$ and $e_i \sim N(0, D_i)$ $i = 1, \dots, m$. Let $\epsilon_i = v_i + e_i$, and its variance is $A + D_i$. The vector $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$ is a vector of length k+1 of unknown parameters.

Lahiri and Suntornchost (2015) proposed a simple adjustment to the standard variable selection methods by replacing MSE_k and MST in above variable selection criteria by

$$\widehat{MSE}_k = MSE_k - \bar{D}_w$$

and

$$\widehat{MST} = MST - \bar{D},$$

where $\bar{D}_w = \frac{\sum_{i=1}^{m}(1-h_{ii})D_i}{m-k}$, $h_{ii} = x_i^T(X^TX)^{-1}x_i$, and $\bar{D} = m^{-1}\sum_{i=1}^{m} D_i$. The new variable selection criteria track the corresponding true variable selection criteria much better than naïve methods. Lahiri and Suntornchost (2015) also proposed a transformation method and a truncation method to prevent negative values of $\widehat{MSE}_k$ and $\widehat{MST}$. As noted, the Lahiri-Suntornchost method can be implemented using two simple steps: 1) adjusting MSE_k and *MST*, and 2) computing the variable selection criteria of users' choice under the standard regression model with adjusted $\widehat{MSE}_k$ and $\widehat{MST}$.

2. Consider a unit level nested error regression model given by:

$$y_{ij} = x_{ij}^T\beta + v_i + e_{ij} \tag{2}$$

for unit $j = 1, \ldots, n_i$ in area $i = 1, \ldots, m$, where n_i is the sample size for small area *i* and the total sample size $n = \sum_{i=1}^{m} n_i$. In Model (2), we assume the area effect v_i ~ iid $N(0, \sigma_v^2)$ is independent of e_{ij} ~ iid $N(0, \sigma_e^2)$. Define $\sigma^2 = \sigma_e^2 + \sigma_v^2$. The outcome in unit *j* of area *i* is denoted by y_{ij}, and $x_{ij} = \left(1, x_{ij1}, x_{ij2}, \ldots, x_{ijk}\right)$ is a vector of length *k*+1 with the values of the covariates $x_1, x_2, \ldots x_k$ for unit *j* in area *i*. In order to make the observations independent and at the same time to avoid the estimation of the intra-cluster correlation, Li and Lahiri (2019) specified P_i to be an $(n_i - 1) \times n_i$ matrix such that $\begin{pmatrix} n_i^{-\frac{1}{2}} 1_{n_i}^T \\ P_i \end{pmatrix}$ is orthogonal for $i = 1,2, \ldots, m$, and transformed the data by

$$\begin{gathered} y_i^{LL} = P_i y_i, \\ x_i^{LL} = P_i x_i, \text{and} \\ u_i^{LL} = P_i u_i. \end{gathered}$$

The transformed model can then be written as:

$$y_i^{LL} = x_i^{LL}\beta + u_i^{LL} \text{ for } i = 1, 2, \ldots, m, \tag{3}$$

where the vector of the error term in area *i* follows $u_i^{LL} \sim N(0, \sigma^2(1-\rho)I_{n_i-1})$ with I_{n_i-1} a $(n_i - 1) \times (n_i - 1)$ identity matrix. The MSE_k and MST estimated from Model (3) can then be plugged into the various variable selection criteria, from which users can pick their favorite to select model variables. Same as the Lahiri-Suntornchost

method, the Li-Lahiri (LL) method is implemented with two steps, but with a different first step: estimating MSE_k and MST by fitting the LL-transformed data to Model (3): a standard regression model with *iid* error.

3. Consider two-fold subarea model given by:

$$y_{ij} = \theta_{ij} + e_{ij} \text{ and } \theta_{ij} = x_{ij}^T\beta + v_i + \gamma_{ij}. \quad (4)$$

Compared to the unit-level nested error regression model (2), an additional error term γ_{ij}~iid $N(0, \sigma_\gamma^2)$ is assumed and independent of v_i or e_{ij}. Cai et al. (2020) first employed the LL data transformation to construct a new linking model for θ_{ij}, given by

$$\theta_i^{LL} = x_i^{LL}\beta + u_i^{LL}, \quad (5)$$

which is similar to Model (3) but with unobserved response θ_i^{LL}. The Lahiri-Suntornchost method are then employed to adjust the MSE_k and MST in estimating the information criteria under Model (5) with MSE_k and MST estimated by replacing the unobserved response θ_i^{LL} by y_i^{LL}, the LL-transformed observed response.

All the three papers aim at making simple adjustments to the regression packages available to data users, and their objective is not to decide on the best possible regression model selection criterion, but to suggest ways to adjust the MSE_k and MST before employing a data user's favorite model selection criterion. Given the conceptual and computational simplicity of the methods and wide availability of software packages for the standard regression model, these adjustments are likely to be adopted by users. To carry out variable selection under an assumed model (Fay-Herriot area model, nested error regression model, or two-fold subarea model), users can choose one of the above information criteria and estimate its values for a set of submodels under consideration with adjusted MSE and MST. The submodel with the smallest estimated information criterion value is selected as the final model.

Prof. Ghosh discussed various inferential aspects, including MSE approximations, under the FH and NER models, assuming the underlying model is true. In practice, variable selection is often conducted to select the optimal model so that inferential accuracy can be improved conditional on the selected model. An important follow-up question is how we can incorporate this additional uncertainty introduced by model selection into the MSE approximation at the inferential stage.

REFERENCES

CAI, S., RAO, J. N. K., DUMITRESCU, L., CHATRCHI, G., (2020). Effective transformation-based variable selection under two-fold subarea models in small area estimation. Statistics in Transition (*to appear*).

HAN, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. Statistical Methodology, 11, pp. 53–67.

JIANG, J., THUAN, N., (2015). The Fence Methods. World Scientific Publishing Co. Pte. Ltd., Singapore.

LAHIRI, P. ed., (2001). Model Selection. Beachwood, OH: Lecture Notes–Monograph Series, Institute of Mathematical Statistics.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. Sankhya B, 77(2), pp. 312–320.

LI, Y., LAHIRI, P., (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. Sankhya B, 81(2), 302–371.

MEZA, J. L., LAHIRI, P., (2005). A note on the PC statistic under the nested error regression model. Survey Methodology 31, pp. 105–109.

RAO, J. N. K., MOLINA, I., (2015). Small Area Estimation, 2nd Edition. Hoboken: Wiley.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 40–44, DOI 10.21307/stattrans-2020-026
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Isabel Molina[1]

Extending on poverty mapping methods

The paper gives a nice overview of small area estimation, putting emphasis on important applications that have led to notable methodological contributions to the field. I would like to extend further on one of the important applications of unit level models that is mentioned in the paper, which is the estimation of poverty or inequality indicators in small areas. The characteristic of this application that makes it particular is that many of these indicators are defined as much more complex functions of the values of the target variable in the area units than simple means or totals.

The traditional method used by the Word Bank, due to Elbers, Lanjouw and Lanjouw (2003 – ELL), was designed to estimate general small area indicators (and perhaps several of them together), defined in terms of a welfare measure for the area units (i.e. households) with a single unit level model for the welfare variable. The model is traditionally a nested error model similar to that of Battesse et al. (1988), for the log of the welfare variable in the population units. This model is fit to the survey data, and the resulting model parameter estimates are then used to generate multiple censuses based on census auxiliary information. With each census, indicators are calculated for each area, and averages across the censuses are taken as ELL estimators. Similarly, variances across the indicators from the different censuses are taken as ELL noise measures of the estimators.

When estimating simple area means with a model for the welfare variable without transformation, the final averaging makes the area effect vanish (it has zero expectation), making ELL estimators essentially synthetic. In fact, ELL method seems to be inspired by the literature on multiple imputation rather than by the small area estimation literature.

[1] Department of Statistics, Universidad Carlos III de Madrid, Spain. E-mail: isabel.molina@uc3m.es. ORCID: https://orcid.org/0000-0002-4424-9540.

Molina and Rao (2010 - MR) proposed to consider empirical best/Bayes (EB) estimators of general small area indicators based on a similar nested error model as in ELL method. The only difference in the model was that, in the traditional applications of ELL method, the random effects were for the clusters of the sampling design (i.e. primary sampling units), which are generally nested in the small areas of interest (e.g., census tracks). In the EB procedure by MR, as in typical small area applications with unit level models, the random effects in the nested error model are for the areas of interest. Considering the clusters as the small areas of interest for more fair comparisons, MR showed substantial gains of EB estimators with respect to ELL ones in a (limited) simulation experiment. In fact, EB estimators are optimal in the sense of minimizing the mean squared error (MSE) under the assumed model and hence cannot be worse than ELL estimators under the same model assumptions. The main reason for the large gains in efficiency is that the EB estimator is theoretically (i.e., under completely known model) defined as the conditional expectation of the indicator given the survey welfares, whereas ELL estimator is theoretically defined as the unconditional expectation which does not make use of the precious information on the actual welfare variable, coming from the survey.

The MSE of the EB estimators in MR (2010) was estimated using the parametric bootstrap approach for finite populations of González-Manteiga et al. (2008), which can be computationally very intensive for large populations and very complex indicators. Molina, Nandram and Rao (2014) proposed a hierarchical Bayes (HB) alternative that avoids performing a bootstrap procedure for MSE estimation, since posterior variances are obtained directly from the predictive distribution of the indicators of interest. They use a reparameterization of the nested error model in terms of the intraclass correlation coefficient, which allows to draw directly from the posterior using the chain rule of probability, avoiding MCMC methods.

Ferretti and Molina (2011) introduced a fast EB approach for the case when the target area parameter is computationally very complex, such as when the indicators are based on pairwise comparisons or sorting area elements, or when the population is too large. Faster HB approaches can be implemented similarly.

Marhuenda et al. (2017) extended the EB procedure for estimation of general parameters to the twofold nested error model with area and (nested) subarea effects, considered in Stukel and Rao (1999) for the case of linear parameters. They obtained clear losses in efficiency when the random effects are specified for the subareas (e.g. clusters) but estimation is desired for areas, except for the case when the areas of interest are not sampled. In this case, they recommend the inclusion of both area and subarea random effects.

Another subtle difference between the traditional ELL approach and the EB method of MR lies in the fact that the original EB method requires to link the survey and census units, because the expectation defining the EB estimator is with respect to the distribution of the non-sample welfares given the sample ones. The Census EB estimator (Molina, 2019) is a slight variation of the original EB estimator based on the nested error model, which does not require linking the survey and census data sets, similarly as ELL procedure. Molina (2019) presents a slight variation of the parametric bootstrap procedure of González-Mateiga et al. (2008) for estimation of the MSE of the Census EB estimator that avoids linking the survey and census data sets.

The Word Bank revised their methodology in 2014 introducing a new bootstrap procedure intended to obtain EB predictors according to Van der Weide (2014), but this procedure is not leading to the original EB (or Census EB) predictors. They also incorporated heteroscedasticity and survey weights, to account for complex sampling designs. They include the survey weights in the estimates of the regression coefficients and variance components according to Huang and Hidiroglou (2003), and also in the predicted area effects following You and Rao (2002). Recently, Corral, Molina and Nguyen (2020) show that the resulting bootstrap procedure leads to substantially biased small area estimators. They also show that MSEs are not correctly estimated with this approach. This has lead to a very recent revision of the World Bank methodology and software, incorporating now the original Census EB estimators and the parametric bootstrap procedure of González-Manteiga et al. (2008), adapted for the case when the survey and census data cannot be linked. The new estimators account for heteroscedasticity and include also survey weights in the model parameter estimators and in the predicted area effects similarly as in Van der Weide (2014). The implemented estimators are the Census versions of the pseudo EB estimators of Guarrama, Molina and Rao (2018) designed to reduce the bias due to complex sampling designs, accounting for heteroscedasticity and using estimates of the variance components that include the survey weights as well.

In small area estimation of welfare-related indicators, another important issue is the transformation taken to the welfare variable in the model. Since welfare variables are most often severely right-skewed and may show heteroscedasticity, log transformation is customarily taken in the nested error model. For the special parameters of area means of the original variables, Molina and Martín (2018) studied the analytical EB predictors under the model with log transformation and obtained second-order correct MSE estimators.

In fact, the EB method of MR for the estimation of general indicators requires normality of area effects and unit level errors, so care should be taken with the transformation taken in order to achieve at least approximate normality. Popular

families of transformations are the power or Box-Cox families. The appropriate member of these families may be selected beyond log in the implemented function for EB method ebBHF() from the R package sae (Molina and Marhuenda, 2015). In fact, in the presence of very small values of the welfare variable, the log transformation shifts these small values towards minus infinity, which may produce now a thin yet long tail in the distribution. A simple way of avoiding such effect is just adding a shift to the welfare variable before taking log. A drawback is that selection of this shift, as well as selection of the Box-Cox or power transformation, needs to be based on the actual survey data. A different approach is to consider a skewed distribution for welfare. Diallo and Rao (2018) extended the EB procedure to the skew normal distribution and Graf, Martín and Molina (2019) considered the EB procedure under a generalized beta of the second kind (GB2). This distribution contains four parameters, one for each tail, offering a more flexible framework for modeling skewed data of different shapes.

Acknowledgement

This work was supported by the Spanish grants MTM2015-69638-R and MTM2015-64842-P from Ministerio de Economía y Competitividad.

REFERENCES

BATTESE, G. E., HARTER, R. M., 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, pp. 28–36.

CORRAL, P., MOLINA, I., NGUYEN, M., (2020). Pull your small area estimates up by the bootstraps. World Bank Policy Research Working Paper 9256.

DIALLO, M., RAO, J., (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45, pp. 1092–1116.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, pp. 355–364.

FERRETTI, C., MOLINA, I., (2012). Fast EB Method for Estimating Complex Poverty Indicators in Large Populations. *Journal of the Indian Society of Agricultural Statistics*, 66, pp. 105–120.

GONZÁLEZ -MANTEIGA, W., LOMBARDÍA,M.J.,MOLINA,I.,MORALES,D.,SANTAMARÍA,L., (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78, pp. 443–462.

GRAF, M., MARÍN, J. M., MOLINA, I., (2019). A generalized mixed model for skewed distributions applied to small area estimation. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 28, pp. 565–597.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis*, 121, pp. 20–40.

HUANG, R., HIDIROGLOU, M., (2003). Design consistent estimators for a mixed linear model on survey data. Proceedings of the Survey Research Methods Section, American Statistical Association (2003), pp. 1897–1904.

MARHUENDA, Y., MOLINA, I., MORALES, D., RAO, J. N. K., (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A*, 180, pp. 1111–1136.

MOLINA, I., (2019). Desagregación de datos enencuestas de hogares: metodologías de estimación en áreas pequeñas. Seriesdela Comisión Económicapara América Latinayel Caribe (CEPAL), Estudios Estadísticos LC/TS.2018/82/Rev.1, CEPAL.

MOLINA, I., MARHUENDA, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7, pp. 81–98.

MOLINA, I., NANDRAM, B., RAO, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8, pp. 852–885.

MOLINA, I., RAO, J. N. K., (2010). Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics*, 38, pp. 369–385.

STUKEL, D., RAO, J. N. K., (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference*, 78, pp. 131–147.

VAN DER WEIDE, R., (2014). Gls estimation and empirical Bayes prediction for linear mixed models with heteroskedasticity and sampling weights: a background study for the POVMAP project. World Bank Policy Research Working Paper 7028.

YOU, Y., RAO, J. N. K., (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, pp. 431–439.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 45–50, 10.21307/stattrans-2020-027
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

David Newhouse[1]

The overview paper by Dr. Malay Ghosh provides a valuable historical perspective on the development of the statistics of small area estimation, giving particular emphasis to important past contributions and recent developments. It is a testament to the phenomenal recent research activity in the field that such a comprehensive overview cannot fully do justice to several relevant topics. I will focus on my comments on, first, detailing practical aspects of small area estimation as it is typically applied by the World Bank for client National Statistics Offices. The second part will discuss how particular aspects of small area estimation as it is traditionally carried out may be altered by the increasing use of "big data", which as the review paper mentions has been driving a resurgence of interest in small area estimation in recent years.

Nearly all small area estimation conducted by the World Bank focuses on generating poverty maps by linking survey data with auxiliary census data, which enables policymakers to obtain estimates of poverty rates at more granular subnational areas than is possible with survey data alone. This method is applicable when the survey and census are conducted around the same time, and has been used to generate poverty maps in over 60 countries. It is typically not feasible, however, to link survey data with census data at the household level due to confidentiality restrictions. Therefore, analysts typically estimate a nested error household-level model in a household expenditure or income survey, and then use the estimated parameters to generate repeated simulations of household income or consumption, adjusted for household size, in the census. These simulations can then be used to generate estimates of the poverty rate and gap, and corresponding measures of uncertainty. Traditionally the World Bank has followed the method described in Elbers, Lanjouw, and Lanjouw (2003), otherwise known as ELL, but more recently, "Empirical Best" methods are increasingly being used (Van der Weide, 2014, Nguyen et al., 2018, Corral et al., 2020). Most models have traditionally specified the random effect at the survey cluster level, following ELL, but there is also

[1] Senior Economist, Poverty and Equity Global Practice, The World Bank, Washington DC, USA. E-mail: dnewhouse@worldbank.org. ORCID: https://orcid.org/0000-0003-4051-8130.

an ongoing shift towards specifying the random effect at the area level, as recommended by Marhuendra et al. (2018).

An important first step when using the traditional method is to identify variables that are common to the census and the household expenditure or income survey, and to verify that the questions are asked in the same way in both surveys. These are typically tested empirically by conducting a t test of means for common variables, although these tests should be interpreted with caution since the results depend in part on the size of the survey. Aggregate means of the variables at the target area level are usually considered as candidate variables and included in the model. This improves the accuracy of the estimates of both poverty rates and their confidence intervals by shrinking the variance of the estimated area effect (Elbers, Lanjouw, and Leite 2008).

The analyst, sometimes in consultation with the national statistics office, determines a model or a set of models to apply. Two important decisions are how many model specifications to estimate and how to select variables. Estimating separate models, for example for urban and rural areas or different subnational regions, can better account for heterogeneity in model coefficients and may be politically appealing. On the other hand, estimating too many distinct models can reduce efficiency. This trade-off is typically navigated based on manual inspection of model results in consultation with national statistics offices.

Model selection is also typically conducted manually, with guidance from automated procedures and model diagnostics such as R2, AIC and BIC. Traditionally, analysts have used stepwise regression to provide a starting point for investigating different models, but are now also employing variance inflation factor thresholds, and occasionally the LASSO, to help select an initial model. A rule of thumb outlined in Zhao (2006) is that the number of variables should be less than the square root of the number of observations. Models are then tweaked manually, in part to obtain national estimates that match survey direct estimates. Studies that follow good practice also examine diagnostics such as residual plots, higher moments of the residuals, and the proportion of variance explained by the area effect. Once the model is selected, the simulations are conducted using one of the three versions of the Stata SAE package. The latest version, which will be universally adopted in the coming months, improves on previous versions by implementing a parametric bootstrap approach to generate mean squared error estimates (Gonzalez-Manteiga et al., 2008, Marhuenda and Molina, 2015). In many cases, estimates are not benchmarked to the level at which the survey is considered representative, although they are in some cases to maintain consistency with published figures.

The resulting poverty estimates are typically published in either reports written jointly with the national statistics offices, or World Bank poverty assessments or

systematic country diagnostics. Most reports highlight subnational estimates of the poverty incidence and the number of poor, which are of greatest interest to policymakers. How these are in turn used in national planning and the allocation of resources varies greatly from country to country. One important application of small area estimates, however, is to inform assessments of the geographic targeting of social assistance programs and the rebalancing of program caseloads across target domains.

The traditional constraint that poverty maps can only be estimated when a new census is available is being challenged by the increasing availability of alternative sources of auxiliary data such as satellite and mobile phone data and administrative records. This offers the possibility to conduct small area poverty estimation each time a new household survey round is collected. In addition, it opens up the possibility of using each new survey to conduct small area estimation for a number of other important socioeconomic characteristics besides poverty, such as population density, labor market, educational outcomes, and health outcomes including disease mapping (Hay et al., 2009)

Several recent innovative studies have demonstrated that satellite imagery and mobile phone data can predict cross-sectional variation in key socioeconomic indicators remarkably well. Mobile phone data is strongly correlated with wealth and multidimensional poverty in a variety of developing country contexts (Steele et al., 2017, Pokhriyal and Jacques, 2017, Blumenstock, 2018). Geospatial data, meanwhile, are broadly predictive of spatial variation in measures of wealth and consumption (Jean et al., 2016, Engstrom et al., 2016, Watmough et al., 2017). Besides wealth and poverty, high-resolution imagery can also accurately predict agricultural yields (Jin et al., 2017, Lobell et al, 2019). Finally, geospatial data correlates very strongly with population density and can be used to estimate small area population and migration statistics from micro census or survey listing data (Wardrop et al., 2018, Engstrom et al., 2018).

Despite the impressive performance of these new sources of data in explaining cross-sectional variation in several socio-economic indicators, most existing research uses big data to generate purely synthetic predictions and has yet to utilize either Bayesian or empirical Bayesian methods to integrate survey data into the estimates[2]. It is also important to emphasize that, with the exception of Pokhriyal and Jacques (2017), these estimates have generally not yet been validated rigorously against census data. In addition, little attention has been paid to appropriately estimating uncertainty. This is unfortunate, because statistics offices typically adopt a minimum threshold of precision, which defines the lowest level of disaggregation for which survey statistics can be published. There is a strong argument that official estimates should adhere to the same standards for precision whether they are derived solely from sample survey data or draw on non-traditional data sources. It is therefore crucial to estimate

[2] Important exceptions are Pokhriyal and Jacques (2017) and Erciulescu et al (2018).

uncertainty accurately when combining survey data with novel forms of big data for official statistics.

The small area estimation methods detailed by Dr. Ghosh are the natural framework to consider how best to combine survey data with "big" auxiliary data. Empirical best models, in particular, are easier to explain and communicate than Bayesian methods, and have the additional advantage of not requiring the specification of a prior distribution. Since auxiliary data is typically available only at the sub-area level, it is natural to employ a sub-area empirical best model such as the one outlined in Torabi and Rao (2014). Unfortunately, as of now there is no well-documented software options for estimating sub-area models using empirical best methods. In the short run, sub-area level predictors can be used in household level models to conduct this estimation using existing software such as the SAE package in Stata or the SAE or EMDI packages in R. These models offer the advantage of continuity with existing census-based methods, since they use the same basic nested error structure employed in ELL and Molina and Rao (2010). In the medium term, there is an important agenda to develop software that estimates sub-area models that employ appropriate transformations and generate sound estimates of uncertainty, and to compare the performance of these with household-level models that rely exclusively on sub-area predictors.

Another important area for further research includes understanding which indicators, in both census data and in alternative "big data" data, are most effective in tracking local shocks. Currently, census-based poverty maps rely heavily on household size and educational attainment as explanatory variables, which do not change quickly in response to local economic shocks. Alternative indicators such as weather patterns, predicted crop yields, or new housing construction may better reflect local economic conditions. When applying traditional census-based small area estimation, it would also be useful to better understand the extent of bias caused by time lags between the survey and census data (Lange et al, 2019). This would inform the choice of whether to use older census data at the household level or more current auxiliary data at the sub-area level. Finally, it is critical to validate different methods of combining survey with big data at the sub-area level, to build confidence that the resulting estimates can be relied upon to guide high-stakes policy decisions.

REFERENCES

BLUMENSTOCK, JOSHUA, GABRIEL CADAMURO, ROBERT ON, (2015). Predicting poverty and wealth from mobile phone metadata. Science 350.6264: pp. 1073–1076.

CORRAL, PAUL, ISABEL MOLINA, MINH CONG NGUYEN, (2020). Pull your sae up by the bootstraps, mimeo.

ELBERS, CHRIS, JEAN O. LANJOUW, PETER LANJOUW, (2003). Micro–level estimation of poverty and inequality, Econometrica, 71.1: pp. 355–364.

ELBERS, CHRIS, PETER LANJOUW, PHILLIPPE GEORGE LEITE, (2008). Brazil within Brazil: Testing the poverty map methodology in Minas Gerais, The World Bank.

ENGSTROM, RYAN, JONATHAN HERSH, DAVID NEWHOUSE, (2017). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. The World Bank.

ENGSTROM, RYAN, DAVID NEWHOUSE, VIDHYA SOUNDARARAJAN, (2019a). Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka, The World Bank.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, M. J., MOLINA, I., MORALES, D., SANTAMARÍA, L., (2008). Bootstrap mean squared error of a small-area EBLUP. *Journal of Statistical Computation and Simulation*, 78(5), pp. 443–462.

HAY, SIMON I., et al., (2009). A world malaria map: Plasmodium falciparum endemicity in 2007, PLoS medicine 6.3.

JEAN, NEAL, et al., (2016). Combining satellite imagery and machine learning to predict poverty, Science 353.6301, pp. 790–794.

JIN, Z., AZZARI, G., BURKE, M., ASTON, S., LOBELL, D. B., (2017). Mapping smallholder yield heterogeneity at multiple scales in eastern Africa, Remote Sensing, 9.9.

LANGE, S., UTZ JOHANN PAPE, PETER PÜTZ, (2018). Small area estimation of poverty under structural change, The World Bank.

LOBELL, D. B., AZZARI, G., BURKE, M., GOURLAY, S., JIN, Z., KILIC, T., MURRAY, S., (2019). Eyes in the sky, boots on the ground: assessing satellite- and ground-based approaches to crop yield measurement and analysis. *American Journal of Agricultural Economics.*

MARHUENDA, Y., et al., (2017). Poverty mapping in small areas under a twofold nested error regression model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180.4, pp. 1111–1136.

MOLINA, I., J. N. K. RAO, (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38.3, pp. 369–385.

MOLINA, I., MARHUENDA, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), pp. 81–98.

NGUYEN, MINH, C., et al., (2017). *Small Area Estimation: An extended ELL approach*. mimeo.

POKHRIYAL, N., DAMIEN CHRISTOPHE J., (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, 114.46, E9783–E9792.

STEELE, JESSICA, E., et al., (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14.127, 20160690.

TORABI, M., RAO, J. N. K., (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, pp. 36–55.

VAN DER WEIDE, ROY, (2014). *GLS estimation and empirical Bayes prediction for linear mixed models with Heteroskedasticity and sampling weights: a background study for the POVMAP project*, The World Bank.

WARDROP, N. A., et al., (2018). Spatially disaggregated population estimates in the absence of national population and housing census data, Proceedings of the National Academy of Sciences, 115.14, pp. 3529–3537.

WATMOUGH, GARY, R., et al., (2019). *Socioecologically informed use of remote sensing data to predict rural household poverty*. Proceedings of the National Academy of Sciences, 116.4, pp. 1213–1218.

ZHAO, QINGHUA., (2006). User manual for POVMAP, World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ ManualPovMap. pdf.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 51–52, DOI 10.21307/stattrans-2020-028
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

Danny Pfeffermann[1]

This review article will help to promote further the "exponentially" expanding literature on small area estimation (SAE), which became one of the most researched and practiced topics in statistics in the last three decades. The areas are small, but the research and applications are huge. Malay Ghosh is undoubtedly one of the world leading experts in the theory and application of SAE, and his pioneering articles with his students and colleagues paved the way for new research and applications all over the world. No wonder that he is frequently invited to make keynote presentations in conferences and workshops, and from time to time to write review articles as this one.

I have sent Malay already a few remarks, leaving him the choice to include them in the text or just ignore them, which I shall not repeat here. (I was asked to send a short review anyway.) In the last section of the paper, Malay acknowledges that "the present article leaves out a large number of useful current day topics in small area estimation", referring the readers to look for them in the very comprehensive book of Rao and Molina (2015) and the extensive list of references therein. I shall therefore list a few topics which have been researched more recently (but need to be researched further), and topics that to the best of my knowledge have not been researched so far, but in my view should be investigated. (Unfortunately, due to my extensive administrative roles in the last 7 years, I no longer follow the SAE literature as I used in the past.)

1. SAE with unit observations in the presence of NMAR nonresponse. As well known, the response rate in surveys is steadily declining all over the world, and the nonresponse is often informative, implying inevitably the same problem in at least some of the areas. NMAR nonresponse need to be handled properly, irrespectively of the method of inference, whether design- or model-based; following the frequentist or the Bayesian approach.

[1] National Statistician and Head of CBS, Hebrew University of Jerusalem, Israel & Southampton Statistical Sciences Research Institute, UK. E-mail: msdanny@soton.ac.uk. ORCID: https://orcid.org/0000-0001-7573-2829.

2. Accounting for mode effects. Modern surveys leave the sampled units the choice whether to respond via the internet, by telephone or via a "face to face" interview. As well known, the responses obtained from the different modes are often different, either before different profiles of people respond with different modes, or because the answers depend on the mode chosen. Mode effects can bias the estimates, if not accounted for properly. This is a well-known problem in national surveys, which cannot be ignored in SAE either.
3. Accounting for measurement errors in the covariates in generalised linear mixed models (GLMM). Malay mentions the problem of measurement errors as one of the topics that he has not covered but from my knowledge, this topic has only been investigated (quite extensively) for linear models. Has someone investigated the problem in the context of GLMM?
4. Benchmarking with GLMM. Malay discusses in some detail the issue of benchmarking, citing several studies published in the literature under the frequentist and Bayesian approaches. However, almost all these studies consider linear models. A PhD student of mine just completed his dissertation in which he considers among other topics benchmarking when fitting GLMM, but his study is under the frequentist approach. Extensions under the Bayesian approach will be welcome.
5. Estimation of design-based MSE of model-dependent estimators. The use of models for SAE is often inevitable. Users, (not statisticians), don't care much how the area parameters are estimated, but they are familiar with the concept of design-based (randomization) MSE. The concept that the true target mean or other area characteristics are random makes little sense to them; they like to know how well the predictors estimate the true (finite) area value. Hence, the often need to estimate the design-based MSE. Some work in this direction has been published in recent years, but much more need to be done, depending on the form of the model-dependent predictors.

I follow Malay by acknowledging that the 5 topics listed above are only few drops in a big pool of problems that call for new or further investigation. However, I can see that my review is no longer "short", so let me finish by congratulating Malay for this thoughtful, inspiring review.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 53–58, DOI 10.21307/stattrans-2020-029
Received – 31.01.2020; accepted – 30.06.2020

Discussion of "Small Area Estimation: Its Evolution in Five Decades", by Malay Ghosh

J. N. K. Rao[1]

1. Introduction

It is my great pleasure to act as an invited discussant of this overview paper on small area estimation (SAE) by Malay Ghosh, based on his 28th Annual Morris Hansen Lecture held on October 30, 2019 in Washington, D.C. I was closely associated with the late Morris Hansen while we were both members of the Statistics Canada Methodology Advisory Committee for several years chaired by Hansen. I greatly benefited from his pioneering contributions to survey sampling theory and practice. Ghosh and I collaborated on a SAE review paper 26 years ago (Ghosh and Rao, 1994), which has received more than 1000 Google citations and partly stimulated much research on SAE over the past 25 years. The greatly increased demand for reliable small area statistics worldwide of course is the primary factor for the explosive growth in the SAE methodology. My joint paper with Ghosh stimulated me to write my 2003 Wiley book on SAE (Rao 2003). Because of the extensive developments in SAE after my 2003 book appeared, I wrote the second edition of my Wiley book in 2015 jointly with Isabel Molina (Rao and Molina 2015). Perhaps, my 2015 book is now obsolete given the rapid new developments in SAE theory and practice over the past 5 years!

Direct area-specific estimates are inadequate for SAE due to small domain or area sample sizes or even zero sample sizes in some small areas. It is therefore necessary to take advantage of the information in related areas through linking models to arrive at reliable model-dependent or indirect small area estimates. Hansen et al. (1983) demonstrated that model-dependent strategies can perform poorly for large samples even under small model misspecification, unlike design-based strategies leading to design-consistent estimators. On the other hand, Hansen et al. (1983) also note that the model-dependent strategies might enjoy substantial advantage in small samples if the model is appropriate and the sampling plan need not be probability based. The latter statement has implications to current focus on non-probability samples. Kalton (2018)

[1] School of Mathematics and Statistics, Carleton University, Canada. E-mail: jrao@math.carleton.ca. ORCID: https://orcid.org/ 0000-0003-1103-5500.

says "Opposition to using models has been overcome by the demand for small area estimates".

Ghosh provides a nice overview of methods for indirect estimation of small area means or totals over the past 50 years, starting with the use of synthetic estimation in the context of a radio listening survey (Hansen et al. 1953, pp. 483–486). In the early days, indirect estimates were based on simple implicit linking models (Rao and Molina, 2015, Chapter 3), but methods based on explicit linking models have taken over because of many advantages including the following: (a) model diagnostics to find suitable models can be implemented, (b) area-specific estimates of mean squared error (MSE) can be associated with each small area estimate, unlike the global measures of precision (averaged over small areas) often used with traditional synthetic estimates, and (c) "optimal" estimates of small area parameters under linear mixed and generalized linear mixed models can be obtained using empirical best unbiased prediction (EBLUP), empirical best (EB) or hierarchical Bayes (HB) methods. The HB method is currently popular because of its ability to handle complex models in an orderly manner and the availability of powerful computer programs to implement sophisticated HB methods. Ghosh has made significant contributions to the HB method for SAE. It is interesting to note that his first two papers on HB were jointly with his former students Partha Lahiri and Gauri Datta (Ghosh and Lahiri 1989 and Datta and Ghosh 1991). As we all know, both Lahiri and Datta have become leading researchers in SAE.

2. Basic area-level model

For simplicity, Ghosh focused his paper on the basic area level model (also called the Fay-Herriot model) in sections 5, 7 and 8 supplemented by a brief account of model based SAE under a basic unit level nested error linear regression model (also called the Battese-Harter-Fuller model) in Section 6. He presents the empirical best linear unbiased predictor (EBLUP) which avoids the normality assumption, using the moment estimator of the random effect variance A proposed by Prasad and Rao (1990). He also gives the estimator of the mean squared prediction error (MSPE) proposed by Prasad and Rao (PR), which is second-order unbiased for the MSPE, under normality assumption. He also mentions the work of Lahiri and Rao (1995), which proved the second-order unbiasedness of the PR MSE estimator without normality assumption on the random area effects in the model, provided the PR moment estimator of A is used. Fay and Herriot (1979) proposed a different moment estimator of A by solving two equations iteratively.

The moment estimators of A as well as the maximum likelihood (ML) and the residual ML (REML) estimators might produce zero estimates. In this case, the EBLUPs will give zero weight to the direct estimates in all areas, regardless of the efficiency of

the direct estimator in each area. On the other hand, survey practitioners often prefer to give always a strictly positive weight to direct estimators because they are based on the area-specific unit level data without a model assumption. For this situation, Li and Lahiri (2010) proposed an adjusted ML (AML) estimator that delivers a strictly positive estimator of A. Molina et al. (2015) proposed modifications of the AML estimator that use the AML estimator only when the REML estimator is zero or when the data does not provide enough evidence against the hypothesis. Their simulation study suggested that the EBLUPs based on the modified estimators of A lead to smaller average MSE than the A AML-based EBLUPs when A is small relative to the variance of the direct estimator. They also proposed an MSE estimator that performed well in terms of average absolute relative bias even when A is small relative to the variance of the direct estimator.

In my books I emphasized the need for external evaluations by comparing the small area estimates to corresponding gold standard values, say from the recent census, in terms of absolute relative error (ARE) averaged over groups of areas, where ARE for a specific area is equal to |est. – truth|/truth. Ghosh mentions an external evaluation in the context of estimating median income of four-person families for the 50 states and the District of Columbia in USA. His Table 1 shows that the EBLUP leads to significant reduction in ARE averaged over the areas relative to the corresponding direct estimate obtained from the Current Population Survey (CPS). Hidiroglou et al. (2019) report the results of a recent external evaluation on Canadian data. Here Census Areas (CAs) are small areas, direct estimates are unemployment rates from the Canadian Labor Force Survey (LFS) and Employment Insurance (EI) beneficiary rate is the area level covariate, which is an excellent predictor of unemployment rate. Direct estimates from a much larger National Household Survey (NHS) were treated as gold standard or true values. The external evaluation showed that for the 28 smallest areas ARE for the LFS estimates is 33.9% compared to 14.7% for the EBLUP. Statistics Canada is now embarked on a very active SAE program and the demand for reliable small area estimates has greatly increased.

EBLUP-type model dependent estimates are often deemed suitable by National Statistical Agencies to produce official statistics, after careful external evaluations as mentioned above. However, those agencies often prefer estimators of design mean squared error (DMSE) of the EBLUP rather than its estimator of model-based MSPE, similar to estimators of DMSE of the direct estimator, conditional on the small area parameters , see Pfeffermann and Ben-Hur (2019). Exact design-unbiased estimator of EBLUP can be obtained but it is highly unstable due to small sample size in the area and also it can take negative values often when the sampling variance of the direct estimator is large relate to the model variance of the random area effect (Datta et al., 2011). Recent research attempts to remedy the difficulty with the design unbiased

estimator. Rao et al. (2018) proposed a composite estimator of design MSE of EBLUP by taking a weighted combination of the design-unbiased MSE estimator and the model-based estimator of MSPE, using the same weights as those used in constructing the EBLUP as a weighted sum of the direct estimator and the synthetic estimator. It performed well in simulations in overcoming the instability associated with the design unbiased MSE estimator and reducing the probability of getting negative values. Pfeffermann and Ben-Hur (2019) proposed an alternative estimator of DMSE of EBLUP, based on a bootstrap method restricted to the distribution generated by the sampling design.

3. Some extensions

Ghosh mentions an extension of the basic FH model that allows different random effect variances for different small areas. In this case, he refers to the HB method of Tang et al. (2018) based on "global-local shrinkage priors", which can capture potential "sparsity" by assigning large probabilities to random area effects close to zero and at the same time identifying random area effects significantly different from zero. Ghosh mentions that such priors are particularly useful when the number of small areas is very large. I believe this extension is very useful and I expect to see further work on this topic.

Ghosh lists several important topics not covered in his review, including robust SAE, misspecification of linking models and estimation of complex area parameters such as poverty indicators. I will make few remarks on the latter topics.

An excellent review paper by Jiang and J. S. Rao (2020) covers robust SAE and model misspecification. They mention the work of Sinha and Rao (2009) on robust EBLUP (REBLUP) under unit level models that can provide protection against representative outliers in the unit errors and/or area effects. Dehnel and Wawrowski (2020) applied the REBLUP method to provide robust estimates of wages in small enterprises in Poland's districts. Jiang and J. S. Rao (2020) also mention their earlier work (Jiang et al. 2011) on misspecification of the linking model under the FH model.

Most of the past work on SAE focused on area means or totals under area level and unit level models. However, in recent years the estimation of complex small area parameters has received a lot of attention, such as small area poverty indicators that are extensively used for constructing poverty maps. We refer the reader to a review paper (Guadadarrama et al. 2014) and Rao and Molina (2015, Chapter 9) on estimating poverty indicators proposed by the World Bank: poverty rate, poverty gap and poverty severity. They studied empirical best or Bayes (EB) and HB methods and compared them to a method used by the World Bank, called ELL method.

There is also current interest in using estimates from big data or nonprobability samples as additional predictors or covariates in area level models. Rao (2020) mentions some recent applications of using big data as covariates.

4. Production of small area official statistics

Tzavidis et al. (2019) provide a framework for production of small area official statistics using model-dependent methods. Molina and Marhuenda (2015) developed an R package for SAE that was used in the book by Rao and Molina (2015).

REFERENCES

DEHNEL, G., WAWROWSKI, L., (2020). Robust estimation of wages in small enterprises: the application to Poland's districts, Statistics in Transition new series, 21, pp. 137–157.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2016). A comparison of small area estimation methods for poverty mapping, Statistics in Transition new series, 1, pp. 41–66.

GHOSH, M., LAHIRI, P., (1989). A hierarchical Bayes approach to small area estimation with auxiliary information, In: Proceedings of the Joint Indo- US Workshop on Bayesian Inference in Statistics and Econometrics.

HANSEN, M. H., MADOW, W. G., TEPPING, B. J., (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys, Journal of the American Statistical Association, 78, pp. 776–793.

HIDIROGLOU, M. A, BEAUMONT, J-F., YUNG, W., (2019). Development of a small area estimation system at Statistics Canada, Survey Methodology, 45, pp. 101–126.

JIANG, J., RAO, J. S., (2020). Robust small area estimation: An overview, Annual Reviews, 7, pp. 337–360.

KALTON, G., (2019). Developments in survey research over the past 60 years: A personal perspective, International Statistical Review, 87, pp. S10–S30.

LI, H., LAHIRI, P., (2010). An adjusted maximum likelihood method for solving small area estimation problems, Journal of Multivariate Analysis, 101, pp. 882–892.

MOLINA, I., MARHUENDA, Y., (2015). Sae: An R package for Small Area Estimation, The R Journal of Statistics, 7, pp. 81–98.

MOLINA, I., RAO, J. N. K., DATTA, G. S., (2015). Small area estimation under a Fay-Herriot model with preliminary testing for the presence of random effects, Survey Methodology, 41, pp. 1–19.

PFEFFERMANN, D., BEN-HUR, D., (2018). Estimation of randomization mean squared error in small area estimation, International Statistical Review, 87, pp. S31–S49.

RAO, J. N. K., (2003). Small Area Estimation. Hoboken, NJ: Wiley.

RAO, J. N. K., (2020). On making valid inferences by integrating data from surveys and other sources, Sankhya, Series B (in press).

RAO, J. N. K., RUBIN-BLEUER, S., ESTEVAO, V. M., (2018). Measuring uncertainty associated with model-based small area estimators, Survey Methodology, 44, pp. 151–166.

SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation, Canadian Journal of Statistics, 37, pp. 381–399.

TZAVIDIS, N., ZHANG, L. C., LUNA, A., SCHMID, T., ROJAS-PERILLA, N., (2018). From start to finish: a framework for the production of small area official statistics, Journal of the Royal Statistical Society, Series A, 181, pp. 927–979.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 59–67, DOI 10.21307/stattrans-2020-030
Received – 31.01.2020; accepted – 30.06.2020

Rejoinder

Malay Ghosh[1]

I thank all the seven discussants for taking time to read the paper, and for their kind and valuable comments. In particular, they introduced some important current and potentially useful future topics of research, thus supplementing nicely the material covered in this article.

With the current exponential growth in the small area estimation (SAE) literature, I realized the near impossibility of writing a comprehensive review of the subject. Instead, I took the easier approach of tracing some of its early history, and bringing in only a few of the current day research topics, and that too reflecting my own familiarity and interest. I listed a number of uncovered topics in this paper, far outnumbering those that are covered. I am very glad to find that some of these topics are included in the discussion, in varied details.

I will reply to each discussant individually. Professor Molina and Dr. Newhouse have both discussed small area poverty indication, with some overlapping material. I will first discuss them jointly, and then individually on the distinct aspects of their discussion.

Gershunskaya

I thank Dr. Gershunskyaya for highlighting some of the potential problems that one may encounter in small area estimation. Yes, the assumption of known variances D_i, when indeed they are sample estimates, is a cause of concern. Joint modeling of $(y_i, \hat{D}_i)$, when possible, must be undertaken. Unfortunately, without the availability of microdata, especially for secondary users of surveys, modeling the $\hat{D}_i$ can be quite ad hoc, often resulting in very poor estimates. People in Federal Agencies, for example those in the BLS, US Census Bureau and others do have access to the microdata, which can facilitate their modeling. However, even then the issue may not always be completely resolved. I like the hierarchical Bayesian model of Dr. Gershunskaya, something similar to what I have used before. But I have always been concerned about the choice of hyperparameters. For example, in the inverse gamma hyperprior IG$(a_i, c_i\gamma)$, the choice of a_i and c_i can influence the inference considerably, and this demands sensitivity analysis. I wonder whether there is any real global justification of the choice $a_i = 2$ and $c_i = n_i^{-1}$ as proposed in Sugasawa et al. (2017). Added to this is modeling of the parameter γ, which enhances complexity.

[1]Department of Statistics, University of Florida, Gainesville, FL. USA. E-mail: ghoshm@stat.ufl.edu. ORCID: https://orcid.org/0000-0002-8776-7713

Following the same notations of Dr. Gershunskaya, another option may be to use a default half-Cauchy prior (Gelman, 2006) for $D_i^{1/2}$. This results in the prior $\pi(D_i) \propto D_i^{-1/2}(1+D_i)^{-1}$, the so-called "Horseshoe", which enjoys global popularity in these days. It may be noted though that the above prior is just a special case of a Type II beta prior $\pi(D_i) \propto D_i^{a-1}(1+D_i)^{-a-b}$ with $a=b=1/2$. In my own experience, even in the context of SAE research, the choice $a=b=1/2$ is not always the best choice. Other (a,b) combinations produce much better results.

I very much echo the sentiment of Dr. Gershunskaya that reliable estimates for thousands of small domains within a very narrow time frame is a real challenge for most Federal Agencies. With the present COVID-19 outbreak, the BLS is producing steady unemployment numbers for all the States in the US. In situations demanding a very urgent answer, I am quite in favor of a very pragmatic approach, for example, an empirical Bayes approach where one just uses estimates of the hyperparameters. Alternative frequentist approaches such as the jackknife and the bootstrap for mean squared error (MSE) estimation are equally welcome.

Dr. Gershunskaya has highlighted the importance of "external evaluation" of Current Employment Survey (CES) estimates, which I value as extremely important. However, is a six to nine month time lag on the availability of Quarterly Census of Employment and Wages (QCEW) seems a little too much for an ongoing survey like CES. Presumably, different QCEW data are used for production and evaluation. Otherwise, one is faced with the same old criticism of double use of the same data.

I agree wholeheartedly with Dr. Gershunskaya on the issue of robustness of models, and replacing the normal prior by mixtures of normals. In this article, I have mentioned the use of continuous "global-local shrinkage" priors which essentially attain the same goal and are easier for implementation.

Finaly, I thank Dr. Gershunskaya for bringing into our attention that the term "statistical engineering" was used by the late P.C. Mahalanobis, the founding father of statistics in India, back even in 1946 !

Han

I thank Dr. Han for her discussion of the current day research on probabilistic record linkage. While the theoretical framework of record linkage goes back to Fellegi and Sunter (1969), it seems that there was a long fallow period of research up until recent times. Indeed, in my opinion, research on record linkage has taken a giant leap in the last few years, mostly for catering to the needs of Federal Agencies, but its importance has been recognized by the industrial sectors as well.

While record linkage requires merging of two or more sources of data, often it is impossi-

ble to find a unique error-free identifier, for example, when there is an error in recording a person's Social Security Number. This necessitates the need for probabilistic record linkage.

While small area estimation seems to be a natural candidate for application of record linkage in merging survey and administrative data, research in this topic has taken off only very recently. I think that the major reason behind this is the formidable challenge of trustworthy implementation.

Let me elaborate this point a bit. It is universally recognized that small area estimators are model-based estimators. But as pointed out by Dr. Han, now one needs an integrated model based on three components: (1) a unit level SAE model, (2) a linkage error model and (3) a two-class mixture model on comparison vectors. Now, instead of model diagnostics for one single SAE model, one needs model diagnostics for all three models in order to have reliable SAE estimates. In my mind, this seems to be a formidable task. Nevertheless, I encourage Dr. Han and her advisor Partha Lahiri to pursue research in this very important area, and I am very hopeful that their joint venture will become a valuable resource for both researchers and practitioners.

I have some query regarding the assumptions (1)-(3) of Dr. Han. Can one always avoid duplicates in the source files ? Also, is the assumption $S_y \subset S_x$ always tenable ?

In summary, I thank Dr. Han again for her succinct discussion which will be a valuable source of information for the apparent two distinct groups of researchers, one on SAE and the other on record linkage.

Li

I congratulate Dr. Li for bringing in the very important issue of variable selection, a topic near and dear to me in these days. Variable selection is an essential ingredient of any model-based inference, and SAE is no exception.

Dr. Li has provided some very important information regarding necessary modifications of some of the standard criteria, such as the AIC, BIC, Mallows' C_p needed for variable selection in the SAE context. In my opinion though, AIC, BIC, C_p and their variants are more geared towards model diagnostics, and only indirectly towards variable selection. I admit that the two cannot necessarily be separated, but what I like in these days is a direct application of the LASSO (Least Absolute Shrinkage Selection Operator) which achieves simultaneously variable selection and estimation. This is achieved by getting some of the regression coefficients exactly equal to zero, which is extremely useful in the presence of sparsity. In some real life SAE examples that I have encountered, there is a host of independent variables. Rather than the classical forward and backward selection, LASSO and its variant such as LARS (Least Absolute Regrssion Shrinkage) can provide a very direct variable selection and estimation in one stroke.

For simpliicity of exposition, I restrict myself to linear regression models, although the application of LASSO can be extended to generalized linear models, Cox's proportional hazards models and others. For the familiar linear regression model given by $Y = X\beta + e$ notation. The LASSO estimator of β is given by

$$\hat{\beta}_{\text{LASSO}} = \text{argmin}_{\beta} \left[||Y - X\beta||^2 + \lambda \sum_j |\beta_j| \right],$$

where λ is the regularization or the penalty parameter. The choice of the penalty parameter can often become a thorny problem, and there are many proposals including an adaptive approach (Zou, 2006). It will be interesting to see an analog of LASSO in mixed effects models where there is a need for simultaneous selection of regression coefficients and random effects. Obviously, this is of direct relevance to small area estimation. The transformed model of Professor Li from random to fixed effects seems to facilitate the LASSO application in selecting the appropriate regression coefficients. I may add also that there is some recent work on the selection of random effects in the SAE context as discussed in the present paper. But the simultaneous selection problem can potentially be a valuable topic for future research.

I cannot resist the temptation of the well-known Bayesian interpretation of LASSO estimators. Interpreting the loss as the negative of the log-likelihood, and the regularization part as the prior, the LASSO estimator can be interpreted as the posterior mode of a normal likelihood with a double exponential prior. One interesting observation here is that the double exponential prior has tails heavier than that of the normal, but it is still exponential-tailed. Tang, Li and Ghosh (2018), pointed out that polynomial-tailed priors rectify certain deficiencies of exponential-tailed priors. Some of these priors were used in Tang, Ghosh, Ha and Sedransk (2018), as discussed in the present paper.

Molina and Newhouse

Both Professor Molina and Dr. Newhouse have presented very elegantly the current state of the art for estimation of small area poverty indicators. While Professor Molina has has provided a very up-to-date coverage of methodological advances in this area, Dr. Newhouse has focused very broadly on practical applications with examples, and finally a few pointers regarding possible alterations of the World Bank SAE methods with the advent of the so-called "big" data. As I mentioned at the beginning of this rejoinder, I will first present a few common things that I learnt from their discussion, and then reply separately to these two discussants.

One very interesting feature is that SAE of poverty indicators is based on unit level models, another good application of the classical model of Battese, Harter and Fuller (1988). Both discussants began their discussion mentioning the paper of Elbers, Lanjouw and Lanjouw (ELL, 2003), which in my mind, set the stage for further development. An

important piece of information here is that while the SAE indicators both use survey and census data, they cannot be *linked* together at a household level due to data confidentiality. As described in details by Professor Molina, and also hinted at by Dr. Newhouse, ELL circumvented this problem by first fitting the survey data to estimate the model parameters, and then generating multiple censuses to estimate the SAE poverty indicators and their MSE by some sort of averaging of these censuses.

The second important aspect of this research is that unlike most SAE problems which involve estimation of totals, means or proportions, one needs to face nonlinear estimation in addressing the poverty indication problem. This poses further challenge. Variable transformation seems to be a way to justify approximate normality of transformed variables, and I will comment more on this while discussing Professor Molina.

Now I will respond individually to Professor Molina and Dr. Newhouse. Maintaining the alphabetical order throughout this rejoinder, I will first discuss Professor Molina and then Dr. Newhouse.

Molina

Professor Molina has pointed out the distinction of her 2010 joint paper with Dr. Rao with that of ELL. The Molina-Rao (MR) paper is an important contribution, which attracted attention of conventional small area researchers. I am not quite sure what Professor Molina means by "unconditional expectation" in ELL. What I understand though, and also essentially pointed out in Molina, that ELL is producing a synthetic estimator in contrast to an optimal composite estimator, namely the EBLUP as given in MR. This optimality is achieved by combining two sources of information, quite in conformity with the usual Bayesian paradigm, which combines a likelihood with a prior.

There are some important issues stemming out of the ELL and MR papers. One, which seems to have been addressed already in the 2019 paper of Dr. Molina, is how best one can utilize both survey and census data when they cannot be linked together. The second pertains to the question of variable transformation. The log transformation is often useful, especially since the moments of a log-normal distribution can easily be calculated via moment generating function of a normal distribution. While the log transfomation reduces skewness, resulting normality can sometimes be put to question. Professor Molina has mentioned the Box-Cox transformation, which is definitely useful. So are the skewed normal and generalized beta of the second kind. But what about a Bayesian nonparametric approach?

The Bayesian approach has a very distinct advantage of providing some direct measure of uncertainty associated with a point estimate via posterior variance. As recognized by Professor Molina, a hierarchical Bayesian approach avoids much of the implementation complexity, when compared to procedures such as the jackknife and bootstrap. But a Bayesian nonparametric approach seems equally applicable here. MR considered a

general class of poverty measures given in Foster, Greer and Thorbecke (1984). These measures when simplified lead to estimation of either the distribution function or functionals of the distribution function. A Dirichlet process or its mixture with a normal or a heavy-tailed mixing distribution such as the double exponential can be used without much extra effort. This may be a potential topic of useful research.

Professor Molina has also pointed out that the revised World Bank approach of bootstrapped EB predictors can be severely biased. What about the double bootstrap of Hall and Maiti (2006)?

In summary, I thank Dr. Molina again for bringing in the salient features related to estimation of small area poverty indicators. There are potentials for further development, which I believe will take place in the next few years by Dr. Molina and her collaborators.

Newhouse

I thank Dr. Newhouse not only for bringing in the current World Bank practice of producing small area estimates of poverty indicators, but also for pointing out their global applications as well as some important directions for future research.

The World Bank produces small area estimates at a "subnational" level for 60 countries. Dr. Newhouse did not define subnational as its meaning inevitably varies from country to country. For me, it can be counties, census tracts, school districts, or sometimes even the states, depending on the problem at hand. What I admire though is the importance and relevance of this project from a global standpoint.

I agree with Dr. Newhouse about the need for separate models for urban and rural areas. In addition, in the US, variation between the states, for example, West Virginia and New York, also demands separate modeling. I do not think that this approach leads to reduction in efficiency. Rather, it has the potential to provide more meaningful measures of poverty indicators.

I agree wholeheartedly with Dr. Newhouse regarding the use of alternative sources of auxiliary data. But even there, one may often face the difficulty of proper linkage. Partha Lahiri and Ying Han are currently working quite extensively on probabilistic record linkage in the context of small area estimation. Some of their proposed methods may be helpful in other contexts as well.

"Big"" data offers a huge potential. Combining survey data with administrative data, whenever possible, is expected to provide better results than one that uses only one of these two sources of data. I may add that "non probability sampling" has started receiving attention as well because of the richness of administrative data. Whatever the source, model-based SAE is inevitable, and thus always has the potential danger of failing to provide the right answer. External evaluation of model-based procedures

against some "gold standard" seems to be a necessity. This may not be feasible all the time. As an alternative, one may think of cross-validation.

Finally, I like to point out that a model may need to go through a thorough overhaul in the event of a natural or social catastrophe, as we are witnessing now in COVID-19, a "shock" in the general terminology of Dr. Newhouse. Many small area models, by necessity are spatial, temporal or spatio-temporal. Any prediction based on these models, assuming a smooth continuum, will be severely compromised with the occurrence of "shock" events even though some of the auxiliary variables may not be affected.

I thank Dr. Newhouse again for bringing in the current World Bank approach to the production of small area poverty indicators, and his insight into how to improve these estimates in the future.

Pfeffermann

I really appreciate all the valuable comments made by Dr. Pfeffermann in my original text, and they are all incorporated in the revision of this paper. Dr. Pfeffermann has years of both academic and administrative experience, and this is clearly reflected in his discussion. I will try point by point response to his comments, even though I really do not know proper answer to many of the issues that he has raised.

1. I agree with Dr. Pfeffermann that response rate, unless mandatory, is declining fast in most surveys. Further, the simplifying assumption of missing completely at random (MCAR) or missing at random (MAR) is often not very tenable. However, with not missing at random (NMAR) data, I do not see any alternative other than modeling the missingness. In the SAE context, this becomes an extra modeling in addition to the usual SAE modeling, and one requires validation of the integrated model. SAE models with a combination of survey and administrative data, can admit model diagnostics, or sometimes even external evaluation, for example with the nearest census data. Is there a simple way to validate the missingness model in this context? I simply do not know.
2. Again, I agree with Dr. Pfeffermann that present-day surveys offer the option of response via internet, telephone or direct face to face interview. In this cell phone era, I am not particularly fond of telephone interviews. A person living in Texas may have a California cell number. In an ideal situation, for example, a survey designed only for obtaining some basic non sensitive data, the response may not depend much on the mode used. But that is not the case for most surveys, and then the answer may indeed depend on the chosen mode as pointed out very appropriately by Dr. Pfeffermann. What I wonder though is that when there is modal variation in the basic response, is it even possible to quantify the modal difference in the data analysis?
3. Research on measurement errors in covariates for generalized linear models in the SAE context has not possibly started as yet, but it seems feasible. The approach that comes to mind is a hierarchical Bayes approach, both for functional and structural measurement error models.

4. Benchmarking for GLMM is possibly quite challenging from a theoretical point of view in a frequentist set up. It is not at all a problem in a Bayesian framework. Indeed, in Datta et al. (2011), as cited in the present paper, Bayesian benchmarking with squared error loss can be implemented knowing only the posterior mean vector and the posterior variance-covariance matrix.
5. The final point of Dr. Pfeffermann is extremely important as it opens up a new avenue of research. There is always a need for providing uncertainty measures associated with model-based estimates. As George Box once said: "All models are wrong, but some are useful". As a safeguard against potential model uncertainty, one option is to derive design-based MSE of model-based SAE estimators. This also has the potential for convincing conventional survey analysts that model-based SAE or even model-based survey sampling, in general, is not just an academic exercise. Research seems to have just started in this area. A paper that I have just become aware of, courtesy of Dr. Pfeffermann, and mentioned in the current version of the paper, is Pfeffermann and Ben-Hur (2018). Lahiri and Pramanik (2019) addressed the issue of average design-based estimator of design-based MSE, when the average is taken over similar small areas.

Rao

I very much appreciate the kind remarks of Professor Rao. It is needless to say that he is one of the pioneers who brought SAE in the forefront of not just survey statisticians, but for the statistics community at large. I have had the fortune of collaborating with him in a paper only once. But I have had the fortune of getting his advice on a number of occasions in my SAE research.

Regarding the points that he has raised, I agree virtually with all of them. Without a hierarchical Bayesian procedure, it is quite possible to get zero estimates of A, the random effect variance, by any of the standard methods, be it method of moments, ML or REML. Adjusted ML by Li and Lahiri (2010), and subsequent development by Yoshimori and Lahiri (2014), Molina et al. (2015) and Hirose and Lahiri (2018) are indeed very welcome as they rectify this deficiency.

The second point regarding external evaluation is also very useful. Census figures have often been used as "gold standard", used by many researchers including myself. Unfortunately, in many SAE examples, one does not have this opportunity of external validity. I do not have a real idea of an alternative approach with firm footing in this case, but think that cross validation may be an option.

Professor Rao has mentioned the need for design-based MSE computation of model-based SAE estimators. I have emphasized its relevance and importance, while discussing Dr. Pfeffermann. I reiterate that this topic will possibly be a fruitful research topic in the next few years.

I have not seen yet the review article of Jiming Jiang and Sunil Rao, but can appreciate their viewpoint. I have cherished the view for a long time that outliers should not necessarily be discarded for inferential purposes. Rather they can very well be a part of a model, typically a mixture model, which was advocated by Tukey many years ago.

I endorse also that it is high time to go beyond estimation of small area means. Estimation of small area poverty indicators where the World Bank people as well as Professors Rao and Molina have made significant contribution, has taken off the ground and research is pouring in this area. Another potential topic seems to be estimation of quantiles in general, since these parameters are less vulnerable to outliers.

Finally, I thank all the discussants once again for their thorough and informative discussion, supplementing very well the topics not covered in this paper. It is needless to say there is a plethora of other uncovered topics in my paper. We may need another review paper (not by myself) with discussion fairly soon to cover some of these other topics.

REFERENCES

FOSTER, J., GREER, J., THORBECKE, E., (1984). A class of decomposable poverty measures. *Econometrika*, 52, pp. 761–766.

GELMAN, A., (2006). Prior distributions for variance parameters in hierarchical models. (Comment on article by Browne and Draper). *Bayesian Analysis*, 1, pp. 515–553.

HALL, P., MAITI, T., (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, B*, 68, pp. 221–238.

HIROSE, M. Y., LAHIRI, P., (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics*, 46, pp. 1721–1741.

TONG, X., XU, X., GHOSH, M., GHOSH, P., (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80, pp. 215–246.

YOSHIMORI, M., LAHIRI, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriott small area model. *Journal of Multivariate Analysis*, 101, pp. 1418–1429.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 68–83, DOI 10.21307/stattrans-2020-031
Received – 31.01.2020; accepted – 30.06.2020

Effective transformation-based variable selection under two-fold subarea models in small area estimation

Song Cai[1], **J. N. K. Rao**[2], **Laura Dumitrescu**[3], **Golshid Chatrchi**[4]

ABSTRACT

We present a simple yet effective variable selection method for the two-fold nested subarea model, which generalizes the widely-used Fay-Herriot area model. The two-fold subarea model consists of a sampling model and a linking model, which has a nested-error model structure but with unobserved responses. To select variables under the two-fold subarea model, we first transform the linking model into a model with the structure of a regular regression model and unobserved responses. We then estimate an information criterion based on the transformed linking model and use the estimated information criterion for variable selection. The proposed method is motivated by the variable selection method of Lahiri and Suntornchost (2015) for the Fay-Herriot model and the variable selection method of Li and Lahiri (2019) for the unit-level nested-error regression model. Simulation results show that the proposed variable selection method performs significantly better than some naive competitors, especially when the variance of the area-level random effect in the linking model is large.

Key words: bias correction, conditional AIC, Fay-Herriot model, information criterion.

1. Introduction

Small area estimation (SAE) aims to provide reliable estimates of some parameters of interest, such as means or totals, of subpopulations (areas). Sample surveys are usually carried out in some or all areas to collect unit-level data and design-based direct estimators of the parameters are obtained. A common practical issue in SAE is that the design-based direct estimators are usually unreliable because the sampled areas typically have small sample sizes. It is advantageous to use model-based approaches, which can incorporate auxiliary information through linking models to provide reliable estimates of small area parameters (Rao and Molina, 2015). In general, there are two types of small area models, unit-level models and area-level models. We focus on area-level models.

The celebrated Fay-Herriot (FH) area model (Fay and Herriot, 1979) combines direct estimators and auxiliary variables using a linking model to obtain accurate estimates of small area parameters. Let θ_i be the parameter of interest of a sampled area $i = 1, \ldots, m$

[1]Carleton University, Ottawa, ON, Canada. E-mail: scai@math.carleton.ca.
ORCID: https://orcid.org/0000-0003-1368-394X .
[2]Carleton University, Ottawa, ON, Canada. E-mail: jrao@math.carleton.ca.
ORCID: https://orcid.org/0000-0003-1103-5500.
[3]Victoria University of Wellington, Wellington, New Zealand. E-mail: laura.dumitrescu@vuw.ac.nz.
ORCID: https://orcid.org/0000-0002-9205-9151.
[4]Statistics Canada, Ottawa, Ontario, Canada. E-mail: golshid.chatrchi@canada.ca.
ORCID: https://orcid.org/0000-0002-2055-8605.

and x_i be an associated covariate vector. Let y_i be a direct estimator of θ_i, obtained using unit-level data. The FH model assumes that

$$y_i = \theta_i + e_i, \tag{1}$$
$$\theta_i = x_i^{\mathsf{T}}\beta + u_i, \tag{2}$$

where β is a parameter vector, u_i's are independent and identically distributed (iid) random effects following $\mathrm{N}(0, \sigma_u^2)$ with unknown σ_u^2, e_i's are independent (ind) sampling errors following $\mathrm{N}(0, \Psi_i)$ with known sampling variance Ψ_i, and u_i's are independent of e_i's. In practice, Ψ_i is obtained by smoothing the direct estimates of the sampling variances, based on the unit level data, and then treating the smoothed estimates as the true sampling variances. Model (1) is known as the "sampling model" and model (2) is called the "linking model". The empirical best linear unbiased prediction (EBLUP) estimator of θ_i for a sampled area is given by $\hat{\theta}_i = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i) x_i^{\mathsf{T}} \hat{\beta}$, where $\hat{\gamma}_i = \hat{\sigma}_u^2/(\Psi_i + \hat{\sigma}_u^2)$, $\hat{\beta}$ is the best linear unbiased estimator of β and $\hat{\sigma}_u^2$ is the maximum likelihood (ML) estimator or the restricted ML (REML) estiamtor or a method of moments (MM) estimator of σ_u^2 (Rao and Molina, 2015, Chapter 6). The EBLUP estimator of θ_i is a weighted sum of the direct estimator y_i and the so-called "synthetic estimator" $x_i^{\mathsf{T}}\hat{\beta}$.

When multiple auxiliary variables are available, selecting a parsimonious model that fits the data well is especially important for attaining high estimation accuracy for small area parameters. Han (2013) used a conditional Akaike information criterion (cAIC) to select variables under the FH model. Lahiri and Suntornchost (2015) proposed a variable selection method for the FH model by estimating information criteria under the linking model (2). For variable selection under the unit-level nested-error regression (NER) model (Rao and Molina, 2015, Section 4.3), Meza and Lahiri (2005) proposed a method based on the Fuller-Battese transformation (Fuller and Battese, 1973), which requires estimated values of the variance parameters. Li and Lahiri (2019) used a parameter-free transformation method to avoid estimating the variance parameters.

In many applications, data for the subpopulations of interest are collected using a two-fold setup. First, some areas, e.g. states, are sampled. Then, a sample of subareas, e.g. counties, is further selected from each sampled area. Unit-level data then are collected from the sampled subareas. The goal is to estimate a subarea parameter θ_{ij} where i denotes an area and j denotes a subarea. An example of this nested two-fold setup is given by Mohadjer et al. (2012). In the two-fold case, subareas within an area are likely to share some common features and hence the variables of interest are correlated among those subareas. Naively applying the FH model to the subarea-level data will not capture the correlation.

The two-fold subarea model generalizes the FH model and is tailored for the two-fold setup. Suppose that m areas, labelled as $i = 1, \ldots, m$, are sampled from M areas, and for the ith sampled area, n_i subareas, labelled as $j = 1, \ldots, n_i$, are further sampled from N_i subareas. Let y_{ij}, $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, be design-unbiased direct estimators

of θ_{ij}, and x_{ij} be associated covariate vectors. The two-fold subarea model consists of

$$\text{Sampling model: } y_{ij} = \theta_{ij} + e_{ij}, \tag{3}$$

$$\text{Linking model: } \theta_{ij} = x_{ij}^{\mathsf{T}}\beta + v_i + u_{ij}, \tag{4}$$

where $e_{ij} \overset{ind}{\sim} \mathrm{N}(0, \Psi_{ij})$ with known sampling variances Ψ_{ij}, β is a regression parameter vector, $v_i \overset{iid}{\sim} \mathrm{N}(0, \sigma_v^2)$ with unknown σ_v^2, and $u_{ij} \overset{iid}{\sim} \mathrm{N}(0, \sigma_u^2)$ with unknown σ_u^2. The random errors e_{ij}, v_i and u_{ij} are assumed to be independent. Different from the FH model, the linking model (4) under the two-fold subarea model has an area-level random effect v_i, which pools information across subareas within an area. Torabi and Rao (2014) developed the theory of EBLUP estimators under the two-fold subarea model.

Despite the fact that the two-fold subarea model is gaining popularity, little research has been conducted on variable selection under the model. In this paper, we propose a simple yet effective variable selection method for the two-fold subarea model, which combines and extends the variable selection method of Lahiri and Suntornchost (2015) for the FH model and the variable selection method of Li and Lahiri (2019) for the unit-level NER model.

The paper is organized as follows. In Section 2, we give a detailed review of some variable section methods for the FH model. In Section 3, we describe the proposed variable selection method for the two-fold subarea model. Simulation results for assessing the performance of the proposed method are provided in Section 4. Concluding remarks are given in Section 5. Proofs and additional simulation results are included in the Appendix.

2. Variable selection under the FH model

2.1. The Lahiri-Suntornchost method

Lahiri and Suntornchost (2015) developed a simple bias-correction method that can activate multiple information criteria for regular linear regression, including Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows' C_p and adjusted R^2, to be used for variable selection under the FH model. Note that the linking model (2) takes the form of a regular regression model although the response values θ_i are unobserved. A simple idea is to estimate an information criterion, for example BIC, for the linking model (2) and then use the estimated information criterion to carry out variable selection under the FH model.

To achieve this, Lahiri and Suntornchost (2015) proposed to estimate $\mathrm{MSE}_\theta := \frac{1}{m-p}\theta^{\mathsf{T}}(I_m - P)\theta$, where I_m is the m by m identity matrix, $\theta = (\theta_1 \ \dots \ \theta_m)^{\mathsf{T}}$, $P = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$, $X = (x_1 \ \dots \ x_m)^{\mathsf{T}}$, and p is the length of β under the FH model. The estimator of MSE_θ is given by

$$\widehat{\mathrm{MSE}}_\theta = \mathrm{MSE}_y - \overline{\Psi}_w,$$

where $\mathrm{MSE}_y = \frac{1}{m-p}y^{\mathsf{T}}(I_m - P)y$, $y = (y_1 \ \dots \ y_m)^{\mathsf{T}}$, $\overline{\Psi}_w = \frac{1}{m-p}\sum_{i=1}^{m}(1-h_{ii})\Psi_i$, and $h_{ii} =$

$x_i^{\mathsf{T}}(X^{\mathsf{T}}X)^{-1}x_i$. Observing that the BIC for the linking model (2) is a continuous function of MSE_θ, i.e. $\mathrm{BIC} = m\log\{(m-p)\,\mathrm{MSE}_\theta/m\} + p\log m$, one can estimate the BIC by plugging in $\widehat{\mathrm{MSE}}_\theta$,

$$\widehat{\mathrm{BIC}} = m\log\left\{(m-p)\widehat{\mathrm{MSE}}_\theta/m\right\} + p\log m.$$

Other information criteria, including AIC, Mallows' C_p and adjusted R^2 for the linking model (2), can be estimated similarly. Lahiri and Suntornchost (2015) also proposed a modification to $\widehat{\mathrm{MSE}}_\theta$ that leads to a strictly positive estimator of MSE_θ.

Lahiri and Suntornchost (2015) commented that the goal of their method is to make simple adjustments to the regression packages available to data users, and their objective is not to decide on the best possible regression model selection criterion, but to suggest ways to adjust a data user's favourite model selection criterion. Indeed, given the conceptual and computational simplicity of the method and wide availability of software packages for the regular regression model, this is a method that is likely to be adopted by users.

2.2. The cAIC method

Han (2013) adapted the cAIC method for linear mixed-effects models (Vaida and Blanchard, 2005) to select variables under the FH model. Han (2013) showed that the cAIC for the FH model is given by

$$\mathrm{cAIC} = -2\log f_c(y|\hat{\theta}) + 2\Phi_0,$$

where $\hat{\theta} = (\hat{\theta}_1 \ \ldots \ \hat{\theta}_m)^{\mathsf{T}}$, $\hat{\theta}_i$ is the EBLUP of θ_i, $f_c(y|\hat{\theta})$ is the conditional density of y given $\hat{\theta}$, and $\Phi_0 = \sum_{i=1}^m(\partial\hat{\theta}_i/\partial y_i)$. When comparing submodels, the submodel with the smallest cAIC value is chosen.

In the expression of the EBLUP $\hat{\theta}_i$, estimated model parameters β and σ_u^2 are required. As a consequence, different estimators of model parameters lead to different expressions for the penalty term Φ_0. Han (2013) derived the analytical expressions of Φ_0 for three frequently used estimators of model parameters: the unbiased quadratic (UQ) estimator, the REML estimator, and the ML estimator. In all three cases, the penalty term Φ_0 has complicated expressions. Compared to the cAIC method, the Lahiri-Suntornchost (2015) method would be more attractive to data users because of its simplicity.

3. Variable selection under two-fold subarea model

We now turn to variable selection under the two-fold subarea model. The two-fold subarea model defined by (3) and (4) can be rewritten in vector form as

$$\text{Sampling model: } y_i = \theta_i + e_i, \tag{5}$$

$$\text{Linking model: } \theta_i = X_i\beta + \tau_i \tag{6}$$

for $i = 1, \ldots, m$, where $y_i = (y_{i1} \ \ldots \ y_{in_i})^\mathsf{T}$, $X_i = (x_{i1} \ \ldots \ x_{in_i})^\mathsf{T}$, $\theta_i = (\theta_{i1} \ \ldots \ \theta_{in_i})^\mathsf{T}$, $e_i = (e_{i1} \ \ldots \ e_{in_i})^\mathsf{T}$, and $\tau_i = v_i \mathbb{1}_{n_i} + u_i$ with $\mathbb{1}_k$ denoting a k-vector of 1s and $u_i = (u_{i1} \ \ldots \ u_{in_i})^\mathsf{T}$. We have $\tau_i \sim \mathrm{N}(0, \Sigma_i)$, where

$$\Sigma_i = \sigma_v^2 \mathbb{1}_{n_i} \mathbb{1}_{n_i}^\mathsf{T} + \sigma_u^2 I_{n_i}. \tag{7}$$

The key difference between the linking model (6) and the linking model (2) under the FH model is that the random effect τ_i in (6) does not have a diagonal structure. If the covariance matrix Σ_i of τ_i can be transformed to have a diagonal structure with equal diagonal entries, then the Lahiri-Suntornchost method for the FH model can be applied. Our proposed method is based on this simple idea and is outlined in two steps below.

First, we linearly transform the linking model (6) into a model with iid random errors. Specifically, for each $i = 1, \ldots, m$, we find a non-random matrix A_i such that $\tau_i^* := A_i \tau_i$ has a diagonal covariance matrix with all diagonal entries equalling some positive constant c across i, and then transform the linking model (6) into

$$\theta_i^* = X_i^* \beta + \tau_i^*, \tag{8}$$

where $\theta_i^* = A_i \theta_i$ and $X_i^* = A_i X_i$. Model (8) takes the form of a regular regression model but with unknown θ_i^*, which is similar to the linking model (2) under the FH model. Second, we estimate information criteria for the transformed linking model (8) using a method similar to the Lahiri-Suntornchost (2015) method for the FH model. The estimated information criteria then can be used for model selection.

In what follows, we give two transformation methods in subsection 3.1, and then describe the proposed method of estimating information criteria in subsection 3.2.

3.1. Transformation

3.1.1 The parameter-free Lahiri-Li transformation

The purpose of the linear transformation A_i is to make $\mathrm{Var}(\tau_i^*) = A_i \Sigma_i A_i^\mathsf{T}$ a diagonal matrix with constant diagonal entries. Ideally, the transformation matrix A_i should not depend on any unknown parameters. Lahiri and Li (2009) proposed a parameter-free transformation method, which can achieve this purpose, and Li and Lahiri (2019) used that transformation method for variable selection under the unit-level NER model. The idea of the transformation is as follows. By (7),

$$\mathrm{Var}(\tau_i^*) = A_i \Sigma_i A_i^\mathsf{T} = \sigma_v^2 (A_i \mathbb{1}_{n_i})(A_i \mathbb{1}_{n_i})^\mathsf{T} + \sigma_u^2 A_i A_i^\mathsf{T}.$$

Hence, to make a constant-diagonal structure for $\mathrm{Var}(\tau_i^*)$, it suffices to find an A_i such that (a) $A_i \mathbb{1}_{n_i} = 0$, and (b) $A_i A_i^\mathsf{T}$ is a diagonal matrix with diagonal entries being constant across $i = 1, \ldots, m$. The conditions (a) and (b) do not involve any parameter, so any matrix A_i satisfying them can be parameter free. Note that the rank of such an A_i is at most $n_i - 1$ because of the linear constraint (a).

Particular examples of parameter-free A_i that satisfy the conditions (a) and (b) were given by Lahiri and Li (2009) and Li and Lahiri (2019), but no general method for

finding parameter-free A_i was suggested. Here, we complement their examples by giving a general method to construct a desired A_i as follows.

Step 1: Fix a set of n_i-1 linearly independent vectors of length n_i, denoted $b_1,\ldots,b_{n_i-1}$, which satisfies $b_k^\intercal \mathbb{1}_{n_i}=0$ for $k=1,\ldots,n_i-1$. For example, one can take b_k to be the vector with kth entry being 1, the last entry being -1 and all the other entries being 0, or, the vector with kth entry being 1, the $(k+1)$th entry being -1 and all the other entries being 0.

Step 2: Apply the Gram-Schmidt process to $b_1,\ldots,b_{n_i-1}$ to obtain a set of orthogonal vectors $a_1,\ldots,a_{n_i-1}$ with $a_1=b_1$ and $a_k=b_k-\sum_{l=1}^{k-1}\mathrm{Proj}_{a_l}(b_k)$ for $k=2,\ldots,n_i-1$, where $\mathrm{Proj}_y(x):=\frac{x^\intercal y}{y^\intercal y}y$ is the projection of vector x onto the line spanned by y. Take $A_i=\left(\frac{a_1}{\|a_1\|}\ \ldots\ \frac{a_{n_i-1}}{\|a_{n_i-1}\|}\right)^\intercal$, where $\|\cdot\|$ is the Euclidean norm.

The A_i constructed this way is parameter free and satisfies the requirements (a) and (b), and correspondingly $A_iA_i^\intercal=I_{n_i-1}$.

In spite of being parameter free, this transformation has two drawbacks: (i) Since the rank of A_i is n_i-1 instead of n_i, each area i loses one degree of freedom after transformation, which is undesirable when the number of sampled areas, m, is large. (ii) After transformation, the intercept term, if included in the original model, will be removed because of the requirement (a). Hence, this transformation method cannot be used if the intercept is to be selected. In practice, this is not an issue because the intercept is usually included in the model and only the other variables are to be selected. Moreover, transformation matrix that satisfies (a) and (b) is not unique, although we do not find that using different parameter-free transformation matrices affects variable selection results significantly. Overall, being simple and parameter-free is of practical importance and hence the Lahiri-Li transformation method is likely to be favoured by most data users.

3.1.2 The Fuller-Battese transformation

If not restricted to a parameter-free transformation, a straightforward idea to make $\mathrm{Var}(\tau_i^*)=A_i\Sigma_iA_i^\intercal$ a diagonal matrix with constant diagonal entries is to take $A_i=d\Sigma_i^{-1/2}$, where $\Sigma_i^{-1/2}$ is the positive definite square-root matrix of Σ_i^{-1} and d is a non-zero constant. Choosing $d=\sigma_u$ and working out $\Sigma_i^{-1/2}$, we get

$$A_i=I_{n_i}-\frac{1}{n_i}\left(1-\sqrt{\frac{1-\rho}{1+(n_i-1)\rho}}\right)\mathbb{1}_{n_i}\mathbb{1}_{n_i}^\intercal,$$

where $\rho=\sigma_v^2/(\sigma_v^2+\sigma_u^2)$, which depends on the model parameters σ_v^2 and σ_u^2. This is the same as the transformation used by Fuller and Battese (1973). Under the transformation, $\mathrm{Var}(\tau_i^*)=\sigma_u^2I_{n_i}$.

In practice, ρ has to be estimated, which is undesirable. One can use the estimating equation (EE) method by Torabi and Rao (2014) or the ML method to estimate ρ

under the two-fold subarea model. Meza and Lahiri (2005) used the Fuller-Battese transformation for variable selection under the unit-level NER model.

3.2. Estimating information criteria

The transformed linking model (8) is a regular regression model with unobserved responses θ_i^*. We now adapt the Lahiri-Suntornchost (2015) method to estimate AIC, BIC and Mallows' C_p under model (8).

Define the mean sum of squares of errors (MSE) of (8) as

$$\mathrm{MSE}_{\theta^*} = \frac{1}{n^* - p}\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*,$$

where $\theta^* = \left(\theta_1^{*\mathsf{T}} \dots \theta_m^{*\mathsf{T}}\right)^{\mathsf{T}}$, $P^* = X^*\left(X^{*\mathsf{T}}X^*\right)^{-1}X^{*\mathsf{T}}$ with $X^* = \left(X_1^{*\mathsf{T}} \dots X_m^{*\mathsf{T}}\right)^{\mathsf{T}}$, n^* is the length of θ^*, and p is the length of β. For a submodel of (8) with p_s covariates, the AIC, BIC and Mallows' C_p are given, respectively, by

$$\begin{aligned}
\mathrm{AIC}^{(s)} &= n^* \log\left\{\left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/n^*\right\} + 2p_s,\\
\mathrm{BIC}^{(s)} &= n^* \log\left\{\left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/n^*\right\} + p_s\log(n^*),\\
C_p^{(s)} &= \left(n^* - p_s\right)\mathrm{MSE}_{\theta^*}^{(s)}/\mathrm{MSE}_{\theta^*} + 2p_s - n^*,
\end{aligned}$$

where $\mathrm{MSE}_{\theta^*}^{(s)}$ is the MSE from the submodel. Since θ^* is unknown, the above information criteria cannot be calculated. To estimate them, we first propose an estimator of MSE_{θ^*}.

Transform the direct estimator vector y_i using the same transformation matrix A_i by letting $y_i^* = A_i y_i$ and $y^* = \left(y_1^{*\mathsf{T}} \dots y_m^{*\mathsf{T}}\right)^{\mathsf{T}}$. Define $\mathrm{MSE}_{y^*} = \frac{1}{n^*-p}y^{*\mathsf{T}}(I_{n^*} - P^*)y^*$. We propose to estimate MSE_{θ^*} by

$$\widehat{\mathrm{MSE}}_{\theta^*} = \mathrm{MSE}_{y^*} - \frac{1}{n^* - p}\,\mathrm{tr}\left\{(I_{n^*} - P^*)AV_eA^{\mathsf{T}}\right\}, \tag{9}$$

where $A = \mathrm{diag}(A_1, \dots, A_m)$ and $V_e = \mathrm{diag}(\Psi_{11}, \dots, \Psi_{mn_m})$. The second term on the right hand side of the above equation can be viewed as a bias-correction term. A simple modification to the MSE estimator as used by Lahiri and Suntornchost (2015) can be applied to $\widehat{\mathrm{MSE}}_{\theta^*}$ to ensure a strictly positive estimator of MSE_{θ^*}.

Theorem 1. *Suppose that the sampling variances Ψ_{ij} are bounded for all i and j, and $n_i \geq 2$ for all i. Then, as the number of areas $m \to \infty$,*

$$\widehat{\mathrm{MSE}}_{\theta^*} = \mathrm{MSE}_{\theta^*} + o_p(1).$$

The proof of Theorem 1 is given in Appendix A. Estimators of AIC, BIC and Mallows' C_p are obtained by plugging $\widehat{\mathrm{MSE}}_{\theta^*}$ into their corresponding expressions. Since all these information criteria are continuous functions of MSE_{θ^*}, by the continuous mapping theorem (van der Vaart, 1998, Theorem 2.3), the errors of the estimated information criteria are also of $o_p(1)$.

To carry out variable selection, one can choose one of the above information criteria and estimate its values for a set of submodels under consideration. The submodel with the smallest estimated information criterion value is selected as the final model.

4. Simulation study

We conducted a simulation study to assess the performance of the proposed variable selection method under the two-fold subarea model. In the simulation, the number of sampled areas m is set to 30. The number of sampled subareas is set to 8 for the first 10 sampled areas, 5 for the next 15 sampled areas, and 10 for the last 5 sampled areas. The sampling standard deviation $\sqrt{\Psi_{ij}}$ is generated from $\mathrm{Unif}(0.5, 1.5)$. We set $\sigma_u = 2$ and consider a few settings for the standard deviation of the area-level random effect with $\sigma_v = 2$, 3.5, 5, 6.5 and 8. In the linking model, we consider an intercept and eight covariates with

$$
\begin{aligned}
&x_{ij,1} \sim \text{Log-normal}(0.3, 0.5), \quad x_{ij,2} \sim \mathrm{Gamma}(1.5, 2), \quad x_{ij,3} \sim \mathrm{N}(0, 0.8),\\
&x_{ij,4} \sim \mathrm{N}(1, 1.5), \quad x_{ij,5} \sim \mathrm{Gamma}(0.6, 10), \quad x_{ij,6} \sim \mathrm{Beta}(0.5, 0.5),\\
&x_{ij,7} \sim \mathrm{Unif}(1, 3), \quad x_{ij,8} \sim \mathrm{Poisson}(1.5),
\end{aligned}
$$

where $x_{ij,k}$ represents the value of the kth covariate for the ith area and jth subarea, Log-normal(μ, σ) denotes the log-normal distribution with mean μ and standard deviation σ on the log-scale, $\mathrm{Gamma}(\alpha, \beta)$ denotes the gamma distribution with shape parameter α and rate parameter β, $\mathrm{Beta}(\kappa, \gamma)$ denotes the beta distribution with shape parameters κ and γ, $\mathrm{Unif}(a, b)$ denotes the uniform distribution on the interval (a, b), and $\mathrm{Poisson}(\lambda)$ denotes the Poisson distribution with mean parameter λ.

We consider two settings for the true underlying model. In the first setting (Setting I), the true regression parameter value is fixed to $\beta = (2, 0, 0, 4, 0, 8, 0, 0, 0)^\mathsf{T}$. The corresponding true model is the submodel with an intercept and covariates $x_{ij,3}$ and $x_{ij,5}$. In the second setting (Setting II), the true regression parameter value is set to $\beta = (2, 3, 0, 4, 0, 8, 0, 1, 0)^\mathsf{T}$, which corresponds to the true model with an intercept and covariates $x_{ij,1}$, $x_{ij,3}$, $x_{ij,5}$, and $x_{ij,7}$. When selecting variables, the intercept term is always included in the model, and we compare all submodels defined by inclusion/exclusion of $x_{ij,k}$, $k = 1, \ldots, 8$.

When generating data, the covariates are generated first and fixed throughout all simulation replications. Then in each simulation replication, y_i, $i = 1, \ldots, m$, are generated from the two-fold subarea model using the above settings. The total number of simulation replications is set to 10000.

We use the proposed method to select covariates by comparing all submodels defined

by the subsets of the eight covariates. We consider the proposed method using the parameter-free Lahiri-Li transformation ($\mathrm{TWOF}_{\mathrm{LL}}$), the Fuller-Battese transformation with the true ρ value ($\mathrm{TWOF}_{\mathrm{FB}}(\rho_0)$), that with the MLE of ρ ($\mathrm{TWOF}_{\mathrm{FB}}(\hat{\rho}_{mle})$), and that with the estimated ρ based on the estimating equation method of Torabi and Rao (2014) ($\mathrm{TWOF}_{\mathrm{FB}}(\hat{\rho}_{ee})$). For comparison, we consider three naive competitors, the method of Lahiri and Suntornchost (2015) for the FH model fitted naively to the data (Naive 1), information criterion approach for the regular linear regression model fitted naively to the data (Naive 2), and the cAIC method of Han (2013) for the FH model fitted naively to the data (Naive cAIC). Note that different information criteria can be used with Naive 1 and Naive 2 methods, but Naive 3 uses cAIC only.

The simulation results using BIC for variable selection under Setting I of the underlying model are reported in Table 1. All versions of the proposed method have significantly

Table 1: Percentage (%) of selecting the true model using BIC; True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
$\mathrm{TWOF}_{\mathrm{LL}}$	76.98	77.00	76.06	76.52	77.54
$\mathrm{TWOF}_{\mathrm{FB}}(\rho_0)$	78.92	78.58	77.74	77.94	78.46
$\mathrm{TWOF}_{\mathrm{FB}}(\hat{\rho}_{mle})$	78.12	78.16	77.02	77.62	78.46
$\mathrm{TWOF}_{\mathrm{FB}}(\hat{\rho}_{ee})$	78.56	78.34	77.22	76.52	78.70
Naive 1	71.42	49.92	29.48	18.80	11.92
Naive 2	73.90	49.52	29.08	18.34	11.66

higher percentages of selecting the true model in all cases. When the standard deviation σ_v of the area-level random effect increases, all versions of the proposed method exhibit stable rate of selecting the true model at approximately 77% level, while both naive methods show dramatic decay in performance from approximately 72% rate of selecting the true model when $\sigma_v = 2$ to nearly 12% when $\sigma_v = 8$. This suggests that when there is a strong area-level effect, as it commonly happens in practice, the proposed method is a clear choice over the naive ones. The proposed method based on the parameter-free Lahiri-Li transformation and that based on the Fuller-Battese transformation perform equally well. Moreover, using an estimated ρ instead of the true value of ρ in the Fuller-Battese transformation does not adversely affect the performance of variable selection in this case.

The simulation results using AIC and Naive cAIC for variable selection under Setting I are given in Table 2. Compared to BIC, AIC yields lower percentage of selecting the true model for all the methods. However, the comparison between the proposed method and the naive methods is similar to the case using BIC. All versions of the proposed method perform similarly and give stable results for different values of σ_v. The naive methods, on the other hand, have poorer performance, and their performance drops considerably as σ_v increases. The Naive cAIC method performs worse than the Naive 1 and Naive 2 methods, likely because the cAIC has a complicated expression.

The simulation results using Mallows' C_p for variable selection under Setting I are

Table 2: Percentage (%) of selecting the true model using AIC or Naive cAIC; True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
TWOF_{LL}	29.92	28.86	28.12	29.26	30.18
$\text{TWOF}_{\text{FB}}(\rho_0)$	30.30	28.52	28.52	29.32	29.94
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$	29.94	28.16	28.30	29.14	29.90
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{ee})$	30.02	28.52	28.74	29.56	30.46
Naive 1	27.40	24.94	19.88	15.36	12.82
Naive 2	29.92	26.20	20.04	15.52	12.92
Naive cAIC	22.90	19.18	16.50	12.51	11.44

reported in Table 3. These results are similar to those using AIC, and the same conclusion can be drawn: the proposed method has stable performance for different values of σ_v and it outperforms the Naive methods in all cases.

Table 3: Percentage (%) of selecting the true model using Mallows' C_p; True model, Setting I: $\beta = (2,0,0,4,0,8,0,0,0)^{\mathsf{T}}$

Method	σ_v				
	2	3.5	5	6.5	8
TWOF_{LL}	31.18	29.98	29.52	30.68	31.48
$\text{TWOF}_{\text{FB}}(\rho_0)$	31.60	29.76	29.62	30.88	31.06
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$	31.40	29.58	29.50	30.46	31.34
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{ee})$	31.40	29.82	29.82	30.84	31.58
Naive 1	28.40	25.92	20.70	15.84	13.24
Naive 2	31.02	27.22	21.20	16.30	13.22

The simulation results for variable selection using BIC, AIC/cAIC and Mallows' C_p under Setting II of the underlying model are reported in Table 4, Table 5 and Table 6, respectively, in Appendix B. The comparison among different methods is similar to that under Setting I. It is worth noting that, compared to Setting I, although more covariates are included in the true model under Setting II, the performance gap between the proposed method and the naive methods is larger, and the performance of the naive methods drops quicker as σ_v increases when using AIC, cAIC or Mallows' C_p.

5. Concluding remarks

We proposed a simple transformation-based variable selection method for the two-fold subarea model. This method is a blend of the variable selection method of Lahiri and Suntornchost (2015) for the FH model and the variable selection method of Li and Lahiri (2019) for the unit-level NER model. The proposed method can be used

with the parameter-free Lahiri-Li (Lahiri and Li, 2009) transformation or the Fuller-Battese transformation which requires estimating model parameters σ_v^2 and σ_u^2. The performance of the proposed method using two different transformations is found to be comparable and substantially better than some naive competitors, especially when the variance of the area-level random effect is large. In practice, using the proposed method with the parameter-free Lahiri-Li transformation is preferred because of the simplicity of the transformation.

Acknowledgements

This work was supported by research grants to Song Cai and J. N. K. Rao from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

FAY, R. E., 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), pp. 269–277.

FULLER, W. A., BATTESE, G. E., (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, 68(343), pp. 626–632.

HAN, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, 11, pp. 53–67.

LAHIRI, P., LI, Y., (2009). A new alternative to the standard F test for clustered data. *Journal of Statistical Planning and Inference*, 139(10), pp. 3430–3441.

LAHIRI, P., SUNTORNCHOST, J., (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhyā B*, 77(2), pp. 312–320.

LI, Y., LAHIRI, P., (2019). A simple adaptation of variable selection software for regression models to select variables in nested error regression models. *Sankhyā B*, 81(2), pp. 302–371.

MAGNUS, J. R., NEUDECKER, H., (2019). *Matrix Differential Calculus with Applications in Statistics and Econometrics, 3rd Edition*. Hoboken: Wiley.

MEZA, J. L., LAHIRI, P., (2005). A note on the C_p statistic under the nested error regression model. *Survey Methodology*, 31(1), pp. 105–109.

MOHADJER, L., RAO, J. N. K., LIU, B., KRENZKE, T., VAN DE KERCKHOVE, W., (2012). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *Journal of the Indian Society of Agricultural Statistics*, 66(1), pp. 55–63.

RAO, J. N. K., MOLINA, I., (2015). *Small Area Estimation, 2nd Edition*. Hoboken: Wiley.

TORABI, M., RAO, J. N. K., (2014). On small area estimation under a sub-area level model. *Journal of Multivariate Analysis*, 127, pp. 36–55.

VAIDA, F., BLANCHARD, S., (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92(2), pp. 251–370.

VAN DER VAART, A. W., (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

APPENDICES

A. Proof of Theorem 1

The idea of the proof is to show that

$$\mathrm{E}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big) = \mathrm{MSE}_{\theta^*} \tag{10}$$

and

$$\mathrm{E}\big\{\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big\} \to 0 \quad \text{as } m \to \infty. \tag{11}$$

Then, by (10) and Markov's inequality, for any given $\varepsilon > 0$, we have

$$\begin{aligned}\Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*} - \mathrm{MSE}_{\theta^*}\big| \geq \varepsilon \,\Big|\, \theta^*\Big) &= \Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*} - \mathrm{E}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big| \geq \varepsilon \,\Big|\, \theta^*\Big) \\ &\leq \frac{\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)}{\varepsilon^2}.\end{aligned}$$

Taking expectation on both sides of the above inequality and applying (11) gives

$$\Pr\Big(\big|\widehat{\mathrm{MSE}}_{\theta^*} - \mathrm{MSE}_{\theta^*}\big| \geq \varepsilon\Big) \leq \frac{\mathrm{E}\big\{\mathrm{Var}\big(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*\big)\big\}}{\varepsilon^2} \to 0$$

as $m \to \infty$, which proves the claimed. To complete the proof, we show (10) and (11) in the sequel.

Lemma 1. *Let C and D be real-valued matrices of the same order, then*

$$\mathrm{tr}\big\{\big(C^{\mathsf{T}}D\big)^2\big\} \leq \mathrm{tr}\big(C^{\mathsf{T}}CD^{\mathsf{T}}D\big)$$

and

$$\big\{\mathrm{tr}\big(C^{\mathsf{T}}D\big)\big\}^2 \leq \mathrm{tr}\big(C^{\mathsf{T}}C\big)\,\mathrm{tr}\big(D^{\mathsf{T}}D\big).$$

Lemma 1 is Theorem 11.2 of Magnus and Neudecker (2019). See a proof therein.

We now prove (10). By the sampling model (5) and the definite of y^*, we have $y^* = \theta^* + e^*$, where $e^* = \left(e_1^{*\mathsf{T}} \dots e_m^{*\mathsf{T}}\right)^{\mathsf{T}}$ with $e_i^* = A_i e_i$ for $i = 1, \dots, m$. This gives

$$\mathrm{MSE}_{y^*} = \frac{y^{*\mathsf{T}}(I_{n^*} - P^*)y^*}{n^* - p} \frac{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^* + 2\theta^{*\mathsf{T}}(I_{n^*} - P^*)e^* + e^{*\mathsf{T}}(I_{n^*} - P^*)e^*}{n^* - p}.$$

Because e^* is independent of θ^*, we have

$$\begin{aligned}\mathrm{E}\left(\mathrm{MSE}_{y^*} \,|\, \theta^*\right) &= \frac{1}{n^* - p}\left[\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^* + 2\theta^{*\mathsf{T}}(I_{n^*} - P^*)\,\mathrm{E}(e^*) + \mathrm{E}\left\{e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\}\right] \\ &= \mathrm{MSE}_{\theta^*} + \frac{1}{n^* - p}\left[2\theta^{*\mathsf{T}}(I_{n^*} - P^*)\,\mathrm{E}(e^*) + \mathrm{E}\left\{e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\}\right].\end{aligned}$$

Put $e = \left(e_1^{\mathsf{T}} \dots e_m^{\mathsf{T}}\right)^{\mathsf{T}}$. Then $\mathrm{E}(e) = 0$, $\mathrm{Var}(e) = V_e$ and $e^* = Ae$, where A and V_e are defined just before Theorem 1. Hence, $\mathrm{E}(e^*) = 0$ and $\mathrm{Var}(e^*) = AV_eA^T$, which implies that $\mathrm{E}\left\{e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\} = \mathrm{tr}\left\{(I_{n^*} - P^*)AV_eA^{\mathsf{T}}\right\}$ by a standard result in multivariate statistics. This leads to

$$\mathrm{E}\left(\mathrm{MSE}_{y^*} \,|\, \theta^*\right) = \mathrm{MSE}_{\theta^*} + \frac{1}{n^* - p}\,\mathrm{tr}\left\{(I_{n^*} - P^*)AV_eA^{\mathsf{T}}\right\}.$$

Then by the definition, (9), of $\widehat{\mathrm{MSE}}_{\theta^*}$, equation (10) is true.

Finally, we prove (11). With simple algebra, we obtain the following decomposition:

$$\mathrm{Var}\left(\widehat{\mathrm{MSE}}_{\theta^*} | \theta^*\right) = \frac{1}{(n^* - p)^2}(T_1 + T_2 + T_3), \tag{12}$$

where

$$\begin{aligned}T_1 &= \mathrm{Var}\left\{e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\}, \\ T_2 &= 4\theta^*(I_{n^*} - P^*)\,\mathrm{E}\left\{e^* e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\}, \\ T_3 &= 4\,\mathrm{E}\left[\left\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\}^2 \mid \theta^*\right].\end{aligned}$$

Since $e \sim \mathrm{N}(0, V_e)$, we have $e^* \sim \mathrm{N}(0, AV_eA^{\mathsf{T}})$. Then by a standard result for multivariate normal distribution, we have $\mathrm{E}\left\{e^* e^{*\mathsf{T}}(I_{n^*} - P^*)e^*\right\} = 0$, which gives $T_2 = 0$. In what follows, we derive upper bounds for T_1 and T_3.

By normality of e^*, we have

$$T_1 = 2\,\mathrm{tr}\left[\left\{(I_{n^*} - P^*)AV_eA^{\mathsf{T}}\right\}^2\right].$$

Noting that $I_{n^*} - P^*$ is symmetric and idempotent, and AV_eA^{T} is symmetric, by Lemma 1, we have

$$T_1 \le 2\,\mathrm{tr}\left\{(I_{n^*} - P^*)(AV_eA^{\mathsf{T}})^2\right\} = 2\,\mathrm{tr}\left\{(AV_eA^{\mathsf{T}})^2\right\} - 2\,\mathrm{tr}\left\{P^*(AV_eA^{\mathsf{T}})^2\right\}$$

Since P^* is symmetric and idempotent, by the cyclic property of trace, we have

$$\mathrm{tr}\Big\{P^*(AV_eA^\mathsf{T})^2\Big\} = \mathrm{tr}\Big\{P^{*2}(AV_eA^\mathsf{T})^2\Big\} = \mathrm{tr}\Big\{P^*(AV_eA^\mathsf{T})^2P^*\Big\} = \mathrm{tr}\Big\{Q^\mathsf{T}Q\Big\} \geq 0,$$

where $Q = (AV_eA^T)P^*$. Therefore,

$$T_1 \leq 2\,\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\}.$$

Noting that $AV_eA^\mathsf{T} = \mathrm{diag}\big\{\big(A_1V_{e_1}A_1^\mathsf{T}\big), \ldots, \big(A_mV_{e_m}A_m^\mathsf{T}\big)\big\}$ where $V_{e_i} = \mathrm{diag}(\Psi_{i1}, \ldots, \Psi_{in_i})$ for $i = 1, \ldots, m$, we have

$$\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\} = \sum_{i=1}^{m}\mathrm{tr}\Big\{\big(A_iV_{e_i}A_i^\mathsf{T}\big)^2\Big\} = \sum_{i=1}^{m}\mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\}.$$

By Lemma 1, we further have

$$\mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\} \leq \mathrm{tr}\Big\{V_{e_i}^2(A_i^\mathsf{T}A_i)^2\Big\} = \mathrm{tr}\Big\{V_{e_i}(A_i^\mathsf{T}A_i)^2V_{e_i}\Big\}.$$

Let λ_i be the largest eigenvalue of $(A_i^\mathsf{T}A_i)^2$. By an inequality about quadratic forms, we have that the jth diagonal entry of $V_{e_i}(A_i^\mathsf{T}A_i)^2V_{e_i}$ is bounded by $\lambda_i\Psi_{ij}^2$. For both the parameter-free Lahiri-Li transformation based on the proposed procedure using the Gram-Schmidt process and the Fuller-Battese transformation, it is easy to show that $\lambda_i = 1$. Then, since Ψ_{ij} is bounded by some constant Ψ_0 for all i and j, we have

$$\mathrm{tr}\Big\{\big(V_{e_i}A_i^\mathsf{T}A_i\big)^2\Big\} \leq \sum_{j=1}^{n_i}\lambda_i\Psi_{ij}^2 \leq n_i\Psi_0^2.$$

Therefore,

$$T_1 \leq 2\,\mathrm{tr}\Big\{(AV_eA^\mathsf{T})^2\Big\} \leq 2\sum_{i=1}^{m}n_i\Psi_0^2 = 2n\Psi_0^2. \tag{13}$$

We now turn to T_3. Because e^* is independent of θ^* and $\mathrm{E}(e^*) = 0$, we have

$$\begin{aligned} T_3 &= 4\,\mathrm{E}\Big[\big\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)e^*\big\}^2 \mid \theta^*\Big]. \\ &= 4\theta^{*\mathsf{T}}(I_{n^*} - P^*)\,\mathrm{E}\big(e^*e^{*\mathsf{T}}\big)(I_{n^*} - P^*)\theta^* \\ &= 4\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big(AV_eA\big)(I_{n^*} - P^*)\theta^*. \end{aligned}$$

Observing that $T_3 = \mathrm{tr}(T_3)$, we further have

$$\begin{aligned} T_3 &= 4\,\mathrm{tr}\big\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big(AV_eA\big)(I_{n^*} - P^*)\theta^*\big\} \\ &= 4\,\mathrm{tr}\big\{\big(AV_eA\big)(I_{n^*} - P^*)\theta^*\theta^{*\mathsf{T}}(I_{n^*} - P^*)\big\}. \end{aligned}$$

Then, by Lemma 1 and (13), we get

$$T_3 \le 4\sqrt{\mathrm{tr}\{(AV_eA^\mathsf{T})^2\}\,\mathrm{tr}\{(UU^\mathsf{T})^2\}} \le 4\sqrt{n}\Psi_0\sqrt{\mathrm{tr}\{(UU^\mathsf{T})^2\}},$$

where $U = (I_{n^*} - P^*)\theta^*$. In addition, by the cyclic property of trace, $\mathrm{tr}\{(UU^\mathsf{T})^2\} = \mathrm{tr}\{(U^\mathsf{T}U)^2\} = (U^\mathsf{T}U)^2$. Hence

$$T_3 \le 4\sqrt{n}\Psi_0(U^\mathsf{T}U) = 4\sqrt{n}\Psi_0\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*\}. \tag{14}$$

Combining (12), (13), (14) and the fact that $T_2 = 0$, we get

$$\begin{aligned}\mathrm{Var}(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*) &= \frac{1}{(n^*-p)^2}(T_1 + T_2 + T_3)\\ &\le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{(n^*-p)^2}4\Psi_0\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*\}.\end{aligned}$$

Therefore,

$$\mathrm{E}\{\mathrm{Var}(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*)\} \le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{(n^*-p)^2}4\Psi_0\,\mathrm{E}\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*\}.$$

Since θ^* is normally distributed with covariance matrix $\sigma_u^2 I_{n^*}$, $I_{n^*} - P^*$ is a symmetric idempotent matrix and $\mathrm{E}\{(I_{n^*} - P^*)\theta^*\} = 0$, $\frac{1}{\sigma_u^2}\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*$ has a chi-square distribution with $n^* - p$ degrees of freedom, and so $\mathrm{E}\{\theta^{*\mathsf{T}}(I_{n^*} - P^*)\theta^*\} = (n^* - p)\sigma_u^2$. Further recall that $n^* = n - m$ for the Lahiri-Li transformation, $n^* = n$ for the Fuller-Battese transformation, and $n_i \ge 2$. Then, under both transformations, we have $\frac{n}{(n^*-p)^2} \to 0$ and $\frac{\sqrt{n}}{(n^*-p)} \to 0$ as $m \to \infty$. With the above results, we conclude that

$$\mathrm{E}\{\mathrm{Var}(\widehat{\mathrm{MSE}}_{\theta^*}|\theta^*)\} \le \frac{n}{(n^*-p)^2}2\Psi_0^2 + \frac{\sqrt{n}}{n^*-p}4\Psi_0\sigma_u^2 \to 0$$

as $m \to \infty$, and hence Theorem 1 is proved. □

B. Simulation results under Setting II of the underlying model

Table 4: Percentage (%) of selecting the true model using BIC; True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^\mathsf{T}$

Method	σ_v				
	2	3.5	5	6.5	8
TWOF_{LL}	71.95	72.46	72.60	72.54	72.36
$\text{TWOF}_{\text{FB}}(\rho_0)$	73.67	73.22	73.29	73.40	72.76
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$	73.02	72.88	73.24	73.18	72.75
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{ee})$	73.15	73.02	73.28	73.12	72.66
Naive 1	53.53	23.20	9.75	4.04	2.06
Naive 2	50.64	21.52	8.91	3.86	1.95

Table 5: Percentage (%) of selecting the true model using AIC or Naive cAIC; True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^\mathsf{T}$

Method	σ_v				
	2	3.5	5	6.5	8
TWOF_{LL}	42.56	41.59	41.99	42.17	42.48
$\text{TWOF}_{\text{FB}}(\rho_0)$	42.34	42.27	42.26	42.49	42.05
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$	42.04	41.96	42.05	42.12	42.04
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{ee})$	42.25	42.16	41.37	42.37	42.43
Naive 1	39.20	27.72	18.88	12.18	8.59
Naive 2	41.37	28.32	19.11	12.17	8.64
Naive cAIC	37.11	22.77	14.46	7.71	8.57

Table 6: Percentage (%) of selecting the true model using Mallows' C_p; True model, Setting II: $\beta = (2,3,0,4,0,8,0,1,0)^\mathsf{T}$

Method	σ_v				
	2	3.5	5	6.5	8
TWOF_{LL}	43.83	43.01	43.35	43.49	43.78
$\text{TWOF}_{\text{FB}}(\rho_0)$	43.78	43.55	43.54	43.74	43.47
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{mle})$	43.44	43.27	43.36	43.51	43.39
$\text{TWOF}_{\text{FB}}(\hat{\rho}_{ee})$	43.77	43.69	43.68	43.85	43.91
Naive 1	40.51	28.20	19.17	12.15	8.67
Naive 2	42.46	28.92	19.35	12.16	8.67

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 84–102, DOI 10.21307/stattrans-2020-032
Received – 31.01.2020; accepted – 30.06.2020

Skew normal small area time models for the Brazilian annual service sector survey

André Felipe Azevedo Neves[1], Denise Britz do Nascimento Silva[2], Fernando Antônio da Silva Moura[3]

ABSTRACT

Small domain estimation covers a set of statistical methods for estimating quantities in domains not previously considered by the sample design. In such cases, the use of a model-based approach that relates sample estimates to auxiliary variables is indicated. In this paper, we propose and evaluate skew normal small area time models for the Brazilian Annual Service Sector Survey (BASSS), carried out by the Brazilian Institute of Geography and Statistics (IBGE). The BASSS sampling plan cannot produce estimates with acceptable precision for service activities in the North, Northeast and Midwest regions of the country. Therefore, the use of small area estimation models may provide acceptable precise estimates, especially if they take into account temporal dynamics and sector similarity. Besides, skew normal models can handle business data with asymmetric distribution and the presence of outliers. We propose models with domain and time random effects on the intercept and slope. The results, based on 10-year survey data (2007-2016), show substantial improvement in the precision of the estimates, albeit with presence of some bias.

Key words: Annual Service Sector Survey, hierarchical Bayesian model.

1. Introduction

Small area estimation approaches aim at obtaining precise estimates for geographic areas or domains for which sample sizes are not sufficient to yield satisfactory precision if direct estimators are used. The issue of small area estimation can arise from the demand for information on a specific group such as when estimates for an industrial district or other restricted segment are required.

The *small area (domain) estimation* problem has received much attention in recent decades, in which Fay and Herriot (1979) and Battese, Harter and Fuller (1988) are two key papers. The first considered an area level model in which the input response variable is the direct estimate and auxiliary information comes from area level variables. Battese, Harter and Fuller (1988) proposed a unit level model with both input and auxiliary variables considered to be available at the unit sample level. The Fay-Herriot model uses data at the domain level, with greater scope for application compared to models at the

[1]National School of Statistical Sciences. Brazil. E-mail: andre.neves@ibge.gov.br.
ORCID: https://orcid.org/0000-0001-9819-2300.
[2]National School of Statistical Sciences. Brazil. E-mail: denise.silva@ibge.gov.br.
ORCID: https://orcid.org 0000-0002-5514-7558.
[3]Statistics Department of Federal University of Rio de Janeiro. Brazil. E-mail: fmoura@im.ufrj.br.
ORCID: https://orcid.org/0000-0002-3880-4675.

sampling unit level since aggregated data are more accessible and are less subjected to statistical confidentiality restrictions. However, as pointed out by Moura *et al.* (2017), the Fay–Herriot model assumes conditional normality of the direct estimator which is not suitable for fitting skewed data, particularly for domains with very small sample sizes.

Neves *et al.* (2013) developed the first small domain estimation approach for Brazilian economic surveys. The authors proposed a Fay-Herriot model for the logarithmic transformation of the variable of interest to stabilize the variance resulting from the presence of outliers. However, due to difficulties when converting the results to the original scale, a better alternative is to use an asymmetric distribution to model the direct estimator. Ferraz and Moura (2012) modeled the direct survey estimator as skew normally distributed. They successfully fitted the skew normal model to head-of-household mean income for 140 enumeration areas in the scope of an experimental Brazilian demographic census. Moura *et al.* (2017) compared different small area approaches for fitting skewed data using real business survey data. It was the first experiment in which skew normal models in a Bayesian framework were tested to produce small area estimates for the Brazilian Annual Service Sector Survey (BASSS). The main objective was to develop models for estimating service revenue totals by economic activity at levels of aggregation not planned in the BASSS sampling design.

Considering earlier research and corresponding developments, the principal aim of this work is to extend the previous skew normal models to allow sharing information from repeated surveys, such as the BASSS. We consider models to estimate *gross service revenue* totals in specific groups of economic activities (class level four-digit codes of the International Standard Industrial Classification - ISIC) for states in the Northeast region of the country since these direct survey estimates are not currently published due to small sample size and low precision (Neves, 2012).

This paper is organized as follows. Section 2 presents the Brazilian Annual Service Sector Survey and the small domain estimation problem. Section 3 introduces the skew normal models and their extensions to skew normal time models whereas Section 4 displays results and related analysis. Section 5 contains final remarks and suggestions for future research.

2. Small Area Estimation for the Brazilian Annual Service Sector Survey

Service activity comprises the production of intangible goods for immediate consumption by individuals and institutions. Activities with these characteristics include commerce, transport, advertising, information and technology activities, health and education services, tourism and hospitality, financial and insurance services, and services provided by the public sector.

Although important to the Brazilian economy, the service sector occupies a less prominent position in public since industry is considered the most dynamic and important sector. However, as all sectors are vital for the efficient integrated functioning of the economy, reliable, detailed and timely statistics about the service sector are re-

quired. The BASSS is a non-financial services survey conducted by IBGE since 1998. It investigates economic and financial variables of companies, such as *revenues, costs and expenses, inventories, wages, number of employees* and *number of establishments*. Since firms control the accounting records of all their local units (establishments), where the economic and financial results are registered, the BASSS survey unit is the enterprise – the legally constituted unit that produces services.

Table 1. Disaggregation level of economic classification for which direct estimates are published and services in the scope of this study

Service	*Economic classification*	
	4-digit code (small domains)	*2-3-digit code (published estimates)*
Food and beverages	5611-2	561
Engineering and architecture	7111-4, 7112-0, 7119-7	711
Advertising	7311-4, 7312-2, 7319-0	731
Renting and leasing of personal and household goods	7722-5, 7723-3, 7729-2	772
Travel agency and tour operator activities	7911-2	79
Cleaning and pest control	8121-4, 8122-2	812
Foreign language instruction	8593-7, 8599-6	859
Creative, arts and entertainment activities	9001-9	90
Fitness centers and other physical activity providers	9313-1	931
Other personal services	9601-7, 9602-5, 9603-3	960

The survey frame is a business register comprised of administrative records with basic information about companies, such as wages, number of employees and number of establishments. The survey sample is stratified by economic activities and geographic areas (states), and also according to the number of employees. In addition, enterprises with 20 or more employees and those that operate in more than one Brazilian state are allocated in a *take-all stratum*. The survey publishes total estimates, and corresponding precision, by state and economic activity.

Here, we consider a subset of economic activities, focusing on activities in which the enterprises operate mainly in one state. Table 1 above shows the subset of domains in the scope of this study. Note that, for most of the country, direct survey estimates are only produced by group (3-digit code economic classification) due to the survey sampling design. Therefore, small domains are defined by the four-digit codes, listed in Table 1, in each of the nine Northeast Brazilian states.

Depending on the geographic region, the survey provides information at different levels of economic classification. For the South and Southeast regions, IBGE publishes *class level* data (four-digit codes) of the National Classification of Economic Activities (similar to ISIC). For the states of the North, Northeast and Midwest, survey results are only available at the *group level* (three-digit codes), therefore, at a lower level of activity breakdown (IBGE, 2018). Table 2 presents the number of enterprises and the sample sizes restricted to the services enumerated in Table 1. It also contains the number of small domains (defined by state and economic classification). We use 10-year data to develop models that can also borrow strength over time.

Table 2. Number of enterprises, sample sizes, number of domains and domain samples sizes in the scope of this study by year

Year	Population size	Sample size	Number of domains	Domain sample size	
				Median	Maximum
2007	46,056	730	81	9.0	17
2008	35,050	587	70	8.0	17
2009	37,733	637	72	9.0	16
2010	42,244	668	73	9.0	16
2011	46,501	675	74	9.0	15
2012	48,880	738	80	8.5	15
2013	48,976	665	76	8.0	16
2014	53,458	658	76	8.5	15
2015	52,019	660	80	8.0	13
2016	55,545	656	76	8.0	16

3. Skew normal small area models

Fay and Herriot (1979) developed a two-level linear model to estimate the average income per capita in small towns with less than 1,000 people in the United States. This model uses a direct estimator of the domain total and assumes residuals following a normal distribution, with zero mean and known sample variance.

The Fay-Herriot model incorporates random domain effects to capture variability between the domains that cannot be explained by fixed effects. The model is often cited in the literature of small domain estimation. Because the Fay-Herriot model uses data at the domain level, it allows a greater possibility of application when compared to unit level models considering that aggregated data are more easily accessible and are less subject to statistical confidentiality.

The basic Fay-Herriot model is defined in two stages. We denote by y_d the direct estimates of the true totals and as μ_d the response input variable of the model, where $d = 1,...,D$ are the domains of study. These estimates have a sampling error ε_d that depends on their respective sample sizes and the domain variability. Thus, the first stage model equation can be written as:

$y_d = \mu_d + \varepsilon_d$ – sampling model

$\varepsilon_d \overset{ind}{\sim} N(0, \phi_d),\ d = 1,...,D$

where ϕ_d is the sampling variance of the corresponding direct estimator, assumed known for all domains. In the second stage (linking model), the true values are assumed to be linearly related to a vector of auxiliary variables:

$\mu_d = \mathbf{x}_d^t \beta + \nu_d$ – linking model

$$v_d \overset{ind}{\sim} N(0, \sigma_0^2)$$

Errors ε_d and v_d are mutually independent. Substituting linking model equation in sampling model, we obtain:

$$y_d = \mathbf{x}_d^t \boldsymbol{\beta} + v_d + \varepsilon_d$$

Fay and Herriot (1979) assumed that the sampling variances are known and given by their respective sampling variance estimates. However, these estimates are unstable for areas with small sample sizes. There is a series of papers on joint modeling of survey-weighted estimates and sampling variances, see for example Arora and Lahiri (1997), and Gershunskaya and Savitsky (2019) for a recent discussion of this approach.

The Fay-Herriot model assumes that the sample size in each domain is large enough to apply the central limit theorem (CLT). However, in real situations, the response variable can be asymmetric, implying that assumptions of asymptotic normality are unreasonable in several domains. To overcome this problem, a response variable transformation, such as a logarithmic transformation, is commonly used. However, while the lognormal model makes the asymmetry hypothesis more plausible, an exponential function is required when estimates are converted to the original scale, increasing the variability of the estimates. Moreover, Moura *et al.* (2017) found that the lognormal model performs less well than the skew normal model in their application to BASSS data.

3.1. Skew normal model

Azzalini (1985) described the family of skew normal distributions that preserve some properties of the normal distribution except for the parameter that regulates the distribution's asymmetry. This class of distributions includes the normal distribution as a particular case and facilitates the transition from non-normality to normality. The properties of the skew normal distribution are suitable for asymmetric economic data. We adopt Azzalini's (1985) notation to describe the skew normal density function:

$$Y \sim SN(\mu, \sigma, \lambda) \Leftrightarrow f_Y(y) = \frac{2}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right)$$

where $\Phi(\cdot)$ denotes the cumulative distribution function, $\phi(\cdot)$ is the density function of the standard normal distribution, and the parameters μ,σ and λ are the *location*, *scale* and *asymmetry*, respectively. A particular case is the normal distribution when $\lambda = 0$. The skew normal distribution has interesting properties, some of which are shared with the normal distribution. The mean and variance of the skew normal distribution are given by:

$$E(Y) = \mu + \sigma\delta\sqrt{\frac{2}{\pi}} \quad \text{and} \quad V(Y) = \sigma^2\left\{1 - 2\pi^{-1}\delta^2\right\}$$

where δ is given by: $\delta = \lambda/\sqrt{1+\lambda^2}$.

Ferraz and Moura (2012) proposed the following model, here named Model 1, whose joint distribution of the direct estimator y_d and its sample variance estimator $\hat{\phi}_d$ are described in the following expressions:

$$
\begin{aligned}
y_d|\mu_d,\lambda,n_d,\phi_d &\sim SN(\mu_d,\sqrt{\phi}_d,\lambda/\sqrt{n_d}) \\
\hat{\phi}_d|n_d,\phi_d &\sim Ga\left\{\frac{1}{2}(n_d-1),\frac{1}{2}(n_d-1)\phi_d^{-1}\right\},\ d=1,...,D, \\
\phi_d^{-1}|a_\phi,b_\phi &\sim Ga(a_\phi,b_\phi) \\
\mu_d|\beta,\sigma_0^2 &\sim N(\mathbf{x}_d^t\beta,\sigma_0^2)
\end{aligned} \tag{1}
$$

where D is the number of small domains and n_d is the sample size in the d^{th} domain from a population of N_d units. They assume that the parameters ϕ_d, $d=1,\ldots,D$ are conditionally independent, following each an inverse-gamma distribution $\phi_d^{-1} \sim Ga(a_\phi,b_\phi)$, with unknown common hyperparameters a_ϕ and b_ϕ.

For BASSS survey data, μ_d can be written as a linear function of area-level auxiliary variables with unknown fixed coefficient and a random small domain effect β_{0d}, i.e., $\mu_d=\beta_0+\beta_{0d}+\beta_1 x_d$ where: i) the parameter β_0 is the global intercept; ii) β_{0d} is an intercept that varies by domain; iii) and β_1 is the slope. The auxiliary variable x_d is the total wage by domain, which comes from the business register used as the BASSS sample frame.

As the sample size grows, the skew normal distribution converges to the normal with mean μ_d and variance ϕ_d. Our main parameter of interest is $\theta_d^{sn}=E_d^{sn}(y_d)$, the expected value of y_d in the skew normal model, given by:
$\theta_d^{sn}=\mu_d+\delta_d\sqrt{2\phi_d/\pi}$ where $\delta_d=\lambda_d/\sqrt{1+\lambda_d^2}=\lambda/\sqrt{n_d+\lambda^2}$, with $\lambda_d=\lambda/\sqrt{n_d}$.

The sampling variance estimator $\hat{\phi}_d$ is assumed to be unbiased, providing information about the scale parameter ϕ_d. To borrow strength over domains, the model is completed through a hierarchical structure with respect to the parameters β_{0d} and ϕ_d. The parameters β_{0d} are hypothetically independent and distributed as $\beta_{0d} \sim N(0,\sigma_0^2)$.

The Ferraz and Moura model described by the equations in (1) is complemented by assigning a proper and independent prior distribution to the hyperparameters. When modeling the BASSS survey data, we assigned the following priors to these hyperarameters: $\beta=(\beta_0,\beta_1)^t \sim N_2(\mathbf{0},\Omega_\beta)$, $a_\phi \sim Ga(a,b)$, $b_\phi \sim Ga(c,d)$. To obtain relatively vague prior distributions, we set $\Omega_\beta=1000\mathbf{I}_2$, where $\mathbf{I}_2$ is an identity matrix of order 2 and $a=b=c=d=0.01$.

It is worth noting that Ferraz and Moura (2012) considered $\sigma_0^{-2} \sim Ga(a_0,a_0)$, with $a_0=0.01$. Since we experienced difficulties in fitting some models with this prior, we follow Gelman (2006) and placed a relative vague uniform prior on σ_0, i.e., $\sigma_0 \sim U(0,100)$.

The selection of a prior distribution to the λ parameter must be done carefully. Ferraz and Moura (2012), using results obtained in Sugden *et al.* (2000), proposed a normal distribution for the parameter λ, centered close to zero and with standard deviation given by $\sigma_\lambda=5.5a_\gamma/2.576$, where a_γ is an initial suggested value or estimate of the γ

asymmetry coefficient. For BASSS survey data, we estimated a_γ= 4.7. Therefore, the prior for γ was fixed at $\lambda \sim N(0, 100)$.

3.2. Skew Normal Time Models

In this section, we propose to generalize the skew normal model by introducing an extra random time effect (Models 2, 3 and 4). Models 2 to 4, showed in this section, take into account information from domains over time. As mentioned in Section 2, the BASSS data used here cover a 10-year period from 2007 to 2016. The models are developed to estimate the total *gross revenue* from services for 2016, the final year of this series. Therefore, Model 2 is written as:

$$y_{dt}|\mu_{dt}, \lambda, n_{dt}, \phi_d \sim SN(\mu_{dt}, \sqrt{\phi}_d, \lambda/\sqrt{n}_{dt})$$

$$\hat{\phi}_{dt}|n_{dt}, \phi_d \sim Ga\left\{\frac{1}{2}(n_{dt}-1), \frac{1}{2}(n_{dt}-1)\phi_d^{-1}\right\}$$

$$\phi_d^{-1}|a_\phi, b_\phi \sim Ga(a_\phi, b_\phi)$$

$$\mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt}$$

where $d = 1, ..., D$ denotes the domains of study in all years $t = 1, ..., T$ and n_{dt} is the sample size in the d^{th} domain in year t from the population of N_{dt} units. Note that μ_{dt} can be written as a linear function of area-level auxiliary variables with unknown fixed coefficients, a random small domain effect β_{0d} and a random time effect ζ_{0t}. Because the sample size for each domain does not vary much over the years, we assume that the true sampling variance of the direct estimator is constant over time.

The distributions of the inverse of scale parameter ϕ_d^{-1}, as well as the parameters a_ϕ and b_ϕ are the same as in Model 1. The distributions of the random coefficients under the influence of their respective random effects are defined by:

$$\beta_{0d} \sim N\left(0, \sigma_0^2\right) \text{ and } \zeta_{0t} \sim N\left(0, \sigma_{\zeta_0}^2\right)$$

We assigned a uniform prior distribution to the standard deviations σ_0 and σ_{ζ_0}. As discussed in Gelman (2006), the use of this prior guarantees a proper posterior density as well as other desirable properties. Thus, the relatively vague uniform priors for the standard deviations of both domain and time random effects on the intercept are:

$$\sigma_0 \sim U\left(0, 100\right) \text{ and } \sigma_{\zeta_0} \sim U\left(0, 100\right)$$

In addition, the following constraints are imposed to ensure identifiability of the parameters:

$$\beta_{01} = -\sum_{d=2}^{D} \beta_{0d} \text{ and } \zeta_{01} = -\sum_{t=2}^{T} \zeta_{0t}$$

3.3. Skew Normal Model with Random Effects on the Intercept and Slope

Following Moura and Holt (1999), Model 3 includes domain and time random effects on the intercept and a domain random effect on the slope, whereas Model 4 considers domain and time random effects on both intercept and slope:

$$\text{Model 3: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + \beta_{1d} x_{dt}$$
$$\text{Model 4: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + (\beta_{1d} + \zeta_{1t}) x_{dt}$$

As in Model 2, independent uniform priors with mean 50 are assigned to the standard deviations of both domain and time random effects, as follows:

σ_0^2 – variance of the domain random effect on the intercept,
σ_1^2 – variance of the domain random effect on the slope,
$\sigma_{\zeta_0}^2$ – variance of the time random effect on the intercept,
$\sigma_{\zeta_1}^2$ – variance of the time random effect on the slope.

The identifiability constraints are given by:

$$\beta_{01} = -\sum_{d=2}^{D} \beta_{0d},\ \zeta_{01} = -\sum_{t=2}^{T} \zeta_{0t},\ \beta_{11} = -\sum_{d=2}^{D} \beta_{1d} \text{ and } \zeta_{11} = -\sum_{t=2}^{T} \zeta_{1t}$$

3.4. Skew normal model with random walk effect

Rao and Yu (1994) proposed an extension of the Fay-Herriot model to handle cross-sectional and time-series data, see also Molina and Rao (2015) for further explanation and extensions. Unlike Rao and Yu (1994), Datta *et al.* (1999) employed a Bayesian method to implement a time series cross-sectional model with random walk component to estimate unemployment rates of U.S. states. Since it is reasonable to suppose influence of lag random effects when working with economic data, we also considered another model that includes an addictive random lag term effect of first order:

$$\mu_{dt} = \beta_0 + \beta_d + \beta_{0d,t} + \beta_1 x_{dt}$$

where $\beta_d \sim N(0, \sigma_0^2)$ and $\beta_{0d,t} \sim N(\beta_{0d,t-1}, \sigma_{\xi_0}^2)$ and they are all assumed independent. In Bayesian framework, it is also needed to assign prior distributions to $\beta_{0d,0}$ for $d = 1, \ldots, D$. We considered $\beta_{0d,0} \sim N(0, 100),\ \forall d$ and independently distributed. The other model components are analogously defined as the previous models. We named this Model 5 as "Skew normal model with random walk effect".

Therefore, the linear functions of area-level auxiliary variables for all five models are:

$$\text{Model 1: } \mu_d = \beta_0 + \beta_{0d} + \beta_1 x_d$$
$$\text{Model 2: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt}$$
$$\text{Model 3: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + \beta_{1d} x_{dt}$$

$$\text{Model 4: } \mu_{dt} = \beta_0 + \beta_{0d} + \zeta_{0t} + \beta_1 x_{dt} + (\beta_{1d} + \zeta_{1t}) x_{dt}$$
$$\text{Model 5: } \mu_{dt} = \beta_0 + \beta_d + \beta_{0d,t} + \beta_1 x_{dt}$$

The models are evaluated in Section 4. Model 1 is fitted based on 2016 survey data (direct estimates of total gross service revenue by domain) whereas Models 2 to 5 take into account 10-year (2007–2016) data. Model comparisons are carried out considering direct and model-based estimates for 2016.

4. Results

Parameter and small domain estimates for the models defined in Section 3.1 to 3.4 (Tables 3 and 4) were obtained via MCMC (*Markov chain Monte Carlo*). All results correspond to $100,000$ MCMC sweeps, after a burn-in of $50,000$ iterations. The chain was subsequently thinned by taking every 5^{th} sample value. The Gelman and Rubin (1992) statistics are less than 1.05 for all estimated coefficients and fitted models, showing convergence of chains. Computational details of how to implement MCMC estimation procedure and corresponding Winbugs code are displayed in the Appendix. It also contains the full conditionals of the model described by the equations in (1) as in Ferraz and Moura (2012).

The auxiliary information, such as number of employees, total wages and number of establishments, were obtained from the business register used as the BASSS sampling frame. Model selection procedures showed that simultaneous inclusion of those variables was not adequate since they are highly correlated. Taking into account economic analysis, total wages was chosen as the only explanatory variable for the small area estimation models.

Both response (total gross service revenue) and auxiliary variables (total wages) are expressed in millions of Brazilian currency (Reais-R$). The estimated *wages* coefficients are positive, as anticipated, since the total revenue per domain might increase with the investment in the labor factor. The estimates of the asymmetry coefficient are positive for all models in accordance with the usual pattern of economic data (positively asymmetrical distribution). Nevertheless, the estimated values of this coefficient in Models 2 to 5 are about half of the value in Model 1.

When estimates are compared, the highlight is the variance reduction of the domain random effect on the intercept in the presence of random effects on the slope in Models 3 or 4. Also, the domain random effect on the intercept in Model 2 is considerably greater (4.884) than the time random effect (0.267). Similarly, in Models 3 and 4, the domain random effects have higher coefficients than the estimates of time random effects. In addition, the posterior mean for the intercept parameter in Model 5 exceeds more than twice the estimated values for other models.

The noticeable reduction of the intercept domain random effect variance from Model 1 to Model 3 suggests the need for a domain random effect on the slope indicating that the relation between direct estimates and auxiliary variables is not the same for all domains.

Table 3. Summary of hyperparameters' posterior distributions – Models 1, 2 and 5 - domain and time random effects on the intercept

Parameter	Model 1				Model 2				Model 5			
	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%
β_0	3.404	1.592	0.334	6.619	3.085	0.460	2.207	4.007	8.185	2.783	2.434	13.510
β_1	1.953	0.159	1.660	2.300	2.379	0.075	2.234	2.526	2.561	0.099	2.372	2.760
λ	9.174	5.195	2.922	22.390	3.505	0.319	2.912	4.161	4.424	0.486	3.555	5.445
σ_0	5.452	1.589	2.370	8.799	4.884	0.721	3.666	6.473	2.442	1.920	0.109	6.936
σ_{ζ_0}	-	-	-	-	0.267	0.225	0.008	0.836	-	-	-	-
σ_{ξ_0}	-	-	-	-	-	-	-	-	2.411	0.344	1.758	3.111
a_ϕ	0.321	0.044	0.240	0.414	0.302	0.013	0.277	0.329	0.304	0.013	0.278	0.331
b_ϕ	14.267	4.334	7.193	23.990	4.210	0.430	3.411	5.096	4.054	0.416	3.288	4.914

Table 4. Summary of hyperparameters' posterior distributions – Models 3 and 4 - Domain and time random effects on the intercept and on the slope

Parameter	Model 3				Model 4			
	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%	Mean	Standard Deviation	Percentile 2.5%	Percentile 97.5%
β_0	1.757	0.469	0.902	2.737	2.365	0.519	1.422	3.432
β_1	2.959	0.174	2.619	3.302	2.707	0.182	2.351	3.073
λ	4.066	0.369	3.379	4.838	4.489	0.427	3.711	5.383
σ_0	1.583	0.446	0.765	2.498	1.846	0.456	1.003	2.800
σ_1	1.548	0.190	1.196	1.939	1.474	0.188	1.124	1.860
σ_{ζ_0}	0.598	0.367	0.040	1.435	0.222	0.205	0.007	0.760
σ_{ζ_1}	-	-	-	-	0.389	0.133	0.204	0.718
a_ϕ	0.304	0.013	0.278	0.331	0.304	0.013	0.278	0.331
b_ϕ	4.12	0.421	3.355	5.003	4.122	0.419	3.354	4.991

4.1. Model Comparison

Table 5 presents the deviance information criterion (DIC), the posterior mean of the deviance ($\bar{D}$) and the effect number of parameters (pD) for Models 1 to 5. Note that $DIC = \bar{D} + pD$, see Spiegelhalter *et al.* (2002) for further details about the meaning of these measures. Because the data are formed by the joint pairs $(y_d, \hat{\phi}_d)$, $d = 1, .., D$, all these measurements can be calculated separately and overall values, as presented in Table 5, were obtained by summation. The model with the smallest DIC should be the one that would best jointly predict a replicate data set of y_d and $\hat{\phi}_d$. It can be seen that Model 1 (with domain and time effects in the intercept) seems to fit the service revenue data better than its counterparts. However, the performance of Models 3, 4 and 5 is similar.

Table 5. Model selection – Deviance Information Criterion (DIC)

Model	*DIC*	*pD*	$\bar{D}$
Model 1	1,636.3	145.0	1,491.3
Model 2	1,705.5	103.7	1,601.8
Model 3	1,661.5	115.1	1,546.4
Model 4	1,655.9	119.9	1,536.0
Model 5	1,664.7	119.4	1,545.3

The posterior predictive p-values (Meng, 1994), given by $P(y_d^{rep} > y_d|Data)$, where y_d^{rep} is a predictive value of the observed y_d under the considered model, were also calculated for all models with 2016 data. Values around 0.5 indicate that the distributions of the replicate and the actual values are close. Figure 1 displays the boxplots of the posterior predictive p-values for all models. According to Figure 1, model 5 seems to fit best the 2016 BASSS data. Additional information on precision and bias of small domain estimates follows next to enhance the analysis.

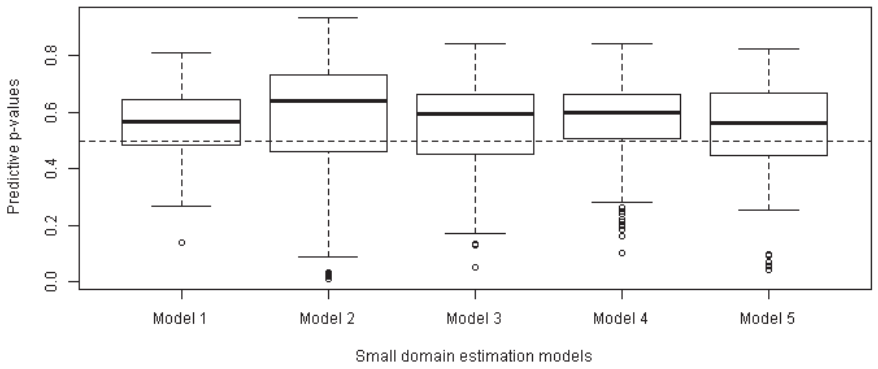


Figure 1 - Posterior predictive p-values of model-based estimates

Model-based estimates are biased, although more precise in general. The estimation procedure aims to balance the trade-off between variance and bias, producing estimates with good precision and little bias as possible. To compare the model performances, precision of estimates and relative differences of the model-based and direct estimates are presented. Figure 2 displays the improvement in coefficients of variation (CVs) for model-based estimates in relation to the direct estimates. Model 1 reduces the coefficients of variation of the small domain estimates with respect to the direct ones in 93.7% of the cases and Models 2 to 5 produce estimates with better precision for all domains. There is evidence that Model 2 provides estimates with better precision than the others. Nevertheless, considering that National Statistical Institutes may suppress the publication of estimates with CV greater than 20% as a quality threshold, Models 2,

3 and 4 do not differ in this aspect. Model 2 has 92.1% of domain estimates with CV below the threshold. This is achieved for 90.8% of the domains in the case of Models 3 and 4, but in only 81.6% of Model 5 estimates.

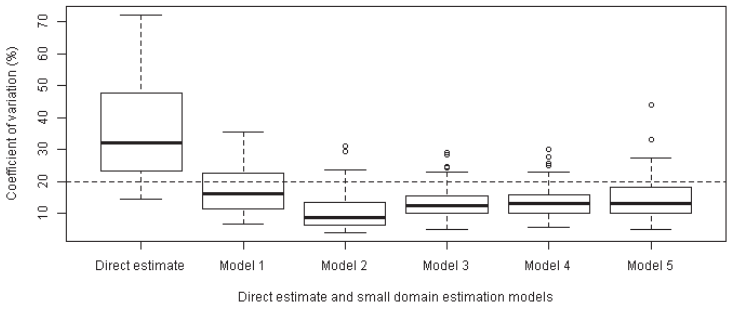


Figure 2 - Coefficients of variation of direct and model-based estimates

The analysis of the relative differences of model-based and the direct estimates $\left(\frac{Model-Direct}{Direct}\ \%\right)$ allows investigating the presence of bias. Relative differences for Models 3 and 4 that incorporate random slopes are closer to zero compared to those from models with random intercept only, as illustrated in Figure 3. In addition, the symmetric distribution for Model 5 relative differences, centered at zero, is good evidence against bias.

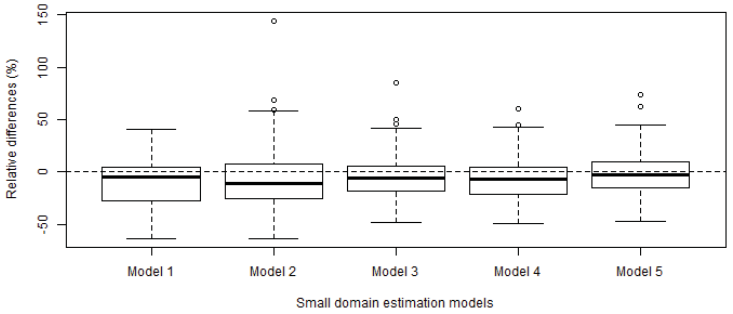


Figure 3 - Relative differences of model-based and direct estimates

The deviance information criterion and the posterior predictive p-values, together with precision and bias of small domain estimates, show that Models 3 and 4 exhibit similar performance. Results for Model 5, with comparable *DIC* value, indicate a slight improvement on the bias, but a disadvantage regarding the precision of estimates. However, Model 5 presents the best performance with respect to the predictive p-value statistics.

Considering all these measures when comparing Models 3, 4 and 5 and the quality threshold for the precision of estimates, the random walk time model (Model 5) can be recommended to produce small domain estimates for the service sector survey.

4.2. Model Diagnostics

We carried out an analysis of the standard residuals, $r_d = \frac{(y_d - \mu_d)}{\sqrt{\phi_d}}$. Since the parameters μ_d and ϕ_d are unknown, they were replaced by their respective posterior means to obtain the $\hat{r}_d$ statistics. According to Genton (2004), if y_d is skew normal distributed, the statistics $\hat{r}_d^2$ is approximately χ_1^2. Figure 4 exhibits residual plots for the application of Model 5 to the BASSS data. The histogram of the $\hat{r}_d$ statistics shows that they have positive skewness. QQ-plots and corresponding envelopes are also presented with lines for the 5^{th} percentile, the mean and the 95^{th} percentile of each observation based on the estimates of squared standard residuals, $\hat{r}_d^2$. The random variable $\hat{r}_d^2$ also enables marginal model checking and detection of outlying observations. The simulated envelope graph plotted to validate the skew-normal Model 5 indicates a few points outside the confidence bounds.

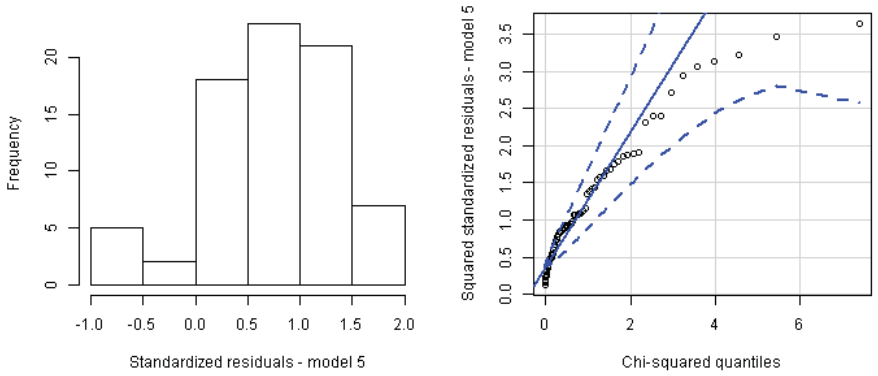


Figure 4 - Histogram and qqplot - Model 5 residuals

We also investigated the relationship between the relative differences and the domain sample sizes (Figure 5) for Model 5 estimates. Although the domain sample sizes are all very modest, with maximum value 16, large relative differences are associated with the smallest sample sizes. The negative relative difference of almost 40% for a sample size of 16 enterprises deserves mention. It refers to a domain whose economic activity is coded as 9001 - *Performing arts, shows and complementary activities*, with unstable demand since these services are not essential and, therefore, subject to income fluctuations and seasonality. Other domains with a sample size greater than 10 for which the relative differences are beyond the limits of 20% are related to economic activity 9313 - *Fitness activities,* which are constantly changing and very diverse (currently the traditional gym

centers coexist with other smaller businesses such as Pilates studios and the services of personal trainers).

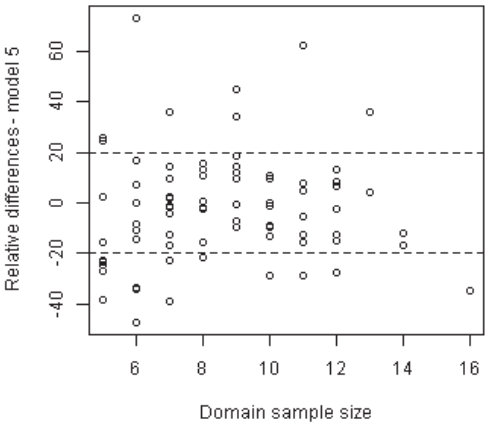


Figure 5 - Relative differences (%) by domain sample sizes - Model 5

5. Conclusions

The small domain estimation models proposed in this article showed good performance in improving the precision of estimates of *gross service revenue* by state and economic activity in the Brazilian Annual Service Sector Survey. The use of skew normal models leads to estimates with much better precision than the direct estimates. Moreover, for most domains, the coefficients of variation are below 20%, which could allow their publication. The skew normal time models with domain and time random effects on the intercept and slope exhibit promising performance. However, the presence of bias is still noted. This is better in Model 5 (Skew normal model with random walk effect), which shows some balance between estimates that exceed or not the direct estimates. Nevertheless, even considering the modest domain sample sizes, there are some domains for which values of relative differences are too high. Thus, despite the relevant gains in precision, the issue of controlling bias requires additional studies. It is important to highlight that this work was carried out using real survey data, focusing on the production of official statistics. Future work is planned to investigate new models to overcome the difficult problem of borrowing strength from domains associated with similar economic activities.

Acknowledgements

This research was supported by IBGE, the National School of Statistical Sciences (ENCE) and Federal University of Rio de Janeiro (UFRJ).

REFERENCES

ARORA, V., LAHIRI, P., (1997). On the superiority of the Bayesian methods over the BLUP in small area estimation problems. *Statistica Sinica* 7, pp. 1053–1063.

AZZALINI, A., (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

BATTESE, G. E., HARTER, R. M., 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*, Mar/1988, Vol.83, 401, pp. 28–36.

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

FAY, R. E., 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*, Vol. 74, 366, pp. 269–277.

FERRAZ, V. R. S., MOURA, F. A. S., (2012). Small area estimation using skew normal models. *Computational Statistics & Data Analysis* 56(10), pp. 2864–2874.

GELMAN, A., (2006). Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper). *Bayesian Analysis* 3, pp. 515–534.

GELMAN, A., RUBIN, D. B., (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 7(4), pp. 457–472.

GENTON, M. G., (2004). *Skew-elliptical distributions and their applications: a journey beyond normality*. Chapman & Hall/CRC.

GERSHUNSKAYA, J., SAVITSKY, T. D., (2019). Model-based screening for robust estimation in the presence of deviations from linearity in small domain models. *Journal of Survey Statistics and Methodology*, 8, 2, pp. 181–205.

IBGE – Instituto Brasileiro de Geografia e Estatística. Pesquisa Anual de Serviços 2016. Diretoria de Pesquisas, Coordenação de Serviços e Comércio. Rio de Janeiro, 2018.

MENG, X.-L., (1994). Posterior predictive p-values. *Annals of Statistics*, 22, pp. 1142–1160.

MOURA, F. A. S., HOLT, D., (1999). Small area estimation using multilevel models. *Survey Methodology*, June 1999, 73, Vol. 25, 1, pp. 73–80, Statistics Canada, Catalogue No. 12-001.

MOURA, F. A. S., NEVES, A. F. A., SILVA, D. B. N., (2017). Small area models for skewed Brazilian business survey data. *Journal of Royal Statistical Society*, 180,Part 4, pp. 1039–1055, serie A.

NEVES, A. F. A., (2012). Small domain estimation applied to Annual Service Sector Survey 2008. *Master dissertation of National School of Statistical Sciences (originally in Portuguese)*. Rio de Janeiro, jul/2012.

NEVES, A. F. A.,SILVA, D. B. N.,CORRÊA, S. T., (2013). Small domain estimation for the Brazilian Service Sector Survey. *Estadística, 65, 185, pp. 13–37, Instituto Interamericano de Estadística)*.

RAO, J. N. K., MOLINA, I., (2015). *Small area estimation*, 2nd ed., New York, Wiley.

RAO, J. N. K., YU, M., (1994). Small-Area Estimation by Combining Time-Series and Cross-Sectional Data. *The Canadian Journal of Statistics*, 22, 4, pp. 511–528.

SPIEGELHALTER, D. J., BEST, N. G., CARLIN, B. P., LINDE, A. V., (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, B 64, Part 4, pp. 583–639.

SUGDEN, R., SMITH, T., JONES, R., (2000). Cochran's rule for simple random sampling. *Journal of the Royal Statistical Society*: Series B 62, pp. 787–793.

COMPUTATIONAL APPENDIX

Stochastic representation

Samples from skew normal density can be generated using the following stochastic representation:

$$y_d|\eta_d,\mu_d,\lambda,\phi_d^2 \sim N(\mu_d+\phi_d\delta_d\eta_d,\phi_d^2(1-\delta_d^2)) \text{ and } \eta_d \sim HN\,(0,1),\ d=1,...,D$$

where $HN(a,b)$ denotes a half-normal distribution with location and scale parameters a and b, respectively. This stochastic representation is useful for implementing the skew normal distribution in statistical packages, such as *WinBUGS* (Spiegelhalter *et al.*, 2002).

Full conditional distributions for Model 1 as in Ferraz and Moura (2012)

$$\begin{aligned}
\pi(\sigma_0^2) &\sim IG\left[a_0+\frac{D}{2},a_0+\frac{1}{2}\sum_{d=1}^{D}(\mu_d-\mathbf{x}_d^t\beta)^2\right],\\
\pi(\beta) &\sim N\left(\left[\sigma_0^2\Omega_\beta^{-1}+\sum_{d=1}^{D}\mathbf{x}_d\mathbf{x}_d^t\right]^{-1}\sum_{d=1}^{D}\mathbf{x}_d\mu_d,\left[\sigma_0^2\Omega_\beta^{-1}+\sum_{d=1}^{D}\mathbf{x}_d\mathbf{x}_d^t\right]^{-1}\right),\\
\pi(\mu_d) &\sim N\left[\left(\frac{(y_d-\sqrt{\phi_d}w_d\delta_d)}{\phi_d(1-\delta_d^2)}+\frac{\mathbf{x}_d^t\beta}{\sigma_0^2}\right)\left(\frac{1}{\phi_d(1-\delta_d^2)}+\frac{1}{\sigma_0^2}\right)^{-1},\left(\frac{1}{\phi_d(1-\delta_d^2)}+\frac{1}{\sigma_0^2}\right)^{-1}\right],\\
\pi(W_d) &\sim N\left[\left(\frac{\delta_d(y_d-\mu_d)}{\sqrt{\phi_d}(1-\delta_d^2)}\right)\left(1+\frac{\delta_d^2}{(1-\delta_d^2)}\right)^{-1},\left(1+\frac{\delta_d^2}{(1-\delta_d^2)}\right)^{-1}\right]I_{(w_d>0)},
\end{aligned}$$

where the symbol $Y\sim IG(a,b)$ generically denotes that Y is inverse gamma distributed, that is, $Y^{-1}\sim Ga(a,b)$, and $N(a,b)I_{(w_d>0)}$ denotes a truncated normal distribution with parameters a and b.

There are no closed forms for the full conditional distributions of ϕ_d, a_ϕ, b_ϕ and λ. Nevertheless, Gibbs sampling with Metropolis-Hasting steps can be used to sample from them. The transition distribution for λ may be normal with the variance tuned for appropriate chain movements. The proposed distributions for ϕ_d, a_ϕ, b_ϕ can be gamma with the mean and variance updated with chain movement.

WinBUGS code

model
{
# Model 5
# Prior distributions
$aphi \sim dgamma(0.01, 0.01)$
$bphi \sim dgamma(0.01, 0.01)$
$beta0 \sim dnorm(0, 0.001)$
$beta1 \sim dnorm(0, 0.001)$
$sigmad0 \sim dunif(0, 100)$
$sigmadt0 \sim dunif(0, 100)$
$Lambda \sim dnorm(0, 0.01)$

Function of the hyperparameters
$sigma2d0 \leftarrow pow(sigmad0, 2)$
$taud0 \leftarrow 1/sigma2d0$
$sigma2dt0 \leftarrow pow(sigmadt0, 2)$
$taudt0 \leftarrow 1/sigma2dt0$

Model 5 description
$for(d\ in\ 1 : Ntot)\{$
$ytot[d] \sim dnorm(mus[d], taus[d])$
$mus[d] \leftarrow mu[d] + delta[d] * sqrt(1/invphi[d]) * t[d]$
$mu[d] \leftarrow beta0 + bd0[domid[d]] + bdt0[domid[d], timeid[d]]$
$+ beta1 * saltot[d]$
$delta[d] \leftarrow lambda[d]/(sqrt(1 + pow(lambda[d], 2)))$
$lambda[d] \leftarrow Lambda/sqrt(n[d])$
$t[d] \leftarrow dnorm(0, 1)I(0,)$
$thetasn[d] \leftarrow mu[d] + sqrt(2/3.14159265359) * delta[d] * sqrt(1/invphi[d])$
$as[d] \leftarrow (n[d] - 1)/2$
$bs[d] \leftarrow (n[d] - 1) * invphi[d]/2$
$phiest[d] \sim dgamma(as[d], bs[d])$
$phi[d] \leftarrow 1/invphi[d]$
$invphi[d] \sim dgamma(aphi, bphi)$
$taus[d] \leftarrow invphi[d] * (1/(1 - pow(delta[d], 2)))$
Standardized residuals
$res[d] \leftarrow (ytot[d] - mu[d]) * sqrt(invphi[d])$
Squared standardized residuals
$dest[d] \leftarrow pow((ytot[d] - mu[d]), 2) * invphi[d]$

DIC calculation
$D1[d] \leftarrow 1.837877 - log(taus[d]) + taus[d] * (pow(ytot[d] - mus[d], 2))$
$D2[d] \leftarrow -2 * as[d] * log(bs[d]) - 2 * (as[d] - 1) * log(phiest[d]) +$

$2 * bs[d] * phiest[d] + 2 * loggam(as[d])$
$D[d] \leftarrow D1[d] + D2[d]$

# Random walk
# Distributions of coefficients
}
$for(j\ in\ 1 : Ndom)\{$
$bd0[j] \sim dnorm(0, taud0)$
}
$for(l\ in\ 1 : Ndom)\{$
$for(k\ in\ 2 : Ntime)\{$
$bdt0[l,k] \sim dnorm(bdt0[l,k-1], taudt0)$
}
}
$for(l\ in\ 1 : Ndom)\{$
$bdt0[l,1] \sim dnorm(bdt0f[l], taudt0)$
}
$for(m\ in\ 1 : Ndom)\{$
$bdt0f[m] \sim dnorm(0, 0.001)$
}
# predictive p-value
$for(i\ in\ ii : ie)\{$
$ypred[i] \sim dnorm(mus[i], taus[i])$
$ppred[i] \leftarrow step(ytot[i] - ypred[i])$
}
}

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 103–122, DOI 10.21307/stattrans-2020-033
Received – 31.01.2020; accepted – 30.06.2020

A comparison of area level and unit level small area models in the presence of linkage errors

Loredana Di Consiglio[1], **Tiziana Tuoto**[2]

ABSTRACT

In Official Statistics, interest in data integration has grown enormously, but the effect of integration procedures on statistical analysis has not yet been sufficiently developed. Data integration is not an error-free procedure and linkage errors, as false links and missed links can invalidate standard estimates. Recently, increasing attention has been paid to the effect of linkage errors on the statistical analyses and on statistical predictions.

Recently, methods to adjust the unit level small area estimators for linkage errors have been proposed when the domains are correctly specified. In this paper we compare the naïve and the adjusted unit level estimators with the area level estimators that are not affected by the linkage errors. The comparison encourages the use of the adjusted unit level estimator.

Key words: linear mixed models, data integration, linkage errors.

1. Data integration and the impact of linkage errors

In Official Statistics, data integration has been acquiring more and more importance; the effect of this procedure on statistical analyses has long been disregarded for a long time but in recent years the impact of linkage errors, false links and missed links, on standard estimates has begun to be analysed. The effect of linkage errors on subsequent analyses has first been investigated by Neter et al. (1965) where first solutions can be found.

Scheuren and Winkler (1993, 1997) and Lahiri and Larsen (2005) analyse the problem from a primary user perspective; in this case the evaluation of the linkage errors is a by-product of the linkage procedure and they propose different methods to use this information to adjust for the linkage biases in subsequent analyses. Clearly, the resulting unbiased estimators depend on the parameters of the linkage model. Recently, Han and Lahiri (2018) propose a general framework for statistical analysis with linked data under general assumptions. A different perspective is in Chambers (2009); secondary data users generally do not have detailed information on linkage model and parameters, in this setting, Chambers (2009) suggests an approximated Best Linear Unbiased Estimator and its empirical version and proposes a maximum likelihood estimator with application to linear and logistic regression functions. An extension to sample-to-register linkage is also proposed.

[1]Istituto Nazionale di Statistica - Istat, Italy. E-mail: diconsig@istat.it
[2]Istituto Nazionale di Statistica - Istat, Italy. E-mail: tuoto@istat.it.
ORCID: https://orcid.org/0000-0003-3436-9474.

In the context of fitting mixed models with linked data, Samart and Chambers (2014) extend the settings in Chambers (2009) and suggest linkage error adjusted estimators of variance effects under alternative methods. In Official Statistics, mixed models are largely used for small area estimation to increase the detail of dissemination of statistical information at local level.

Administrative data can be used to augment the information collected by sample surveys. They can, therefore, increase the set of auxiliary variables and help to improve the model fitting for small area estimation. Linkage of external sources with basic statistical registers as well as with sample surveys can be carried out on different linkage scenarios, see section 2 for the linkage model and errors we adopt in this paper.

Di Consiglio and Tuoto (2016) extend the analysis on the effects of linkage errors on the predictors based on unit level mixed models for small area estimation when auxiliary variables are obtained through a linkage procedure with an external register.

Under the assumption that false matches occur only within the same small area - i.e. in Chambers's terminology the block coincides with the small area-, the linkage errors affect small area predictors both through the impact on the estimation of the fixed and random components, and through the impact on the variance matrix of the linked values. Finally, linkage errors also result in an erroneous evaluation of the covariates means over the sampled units and consequently of the unobserved population units.

Following Chambers (2009) in the sample-to-register linkage setting, and in particular, assuming that the sampling mechanism does not affect the outcome of the linkage process (see Chambers 2009 for details), a pseudo-EBLUP estimator based on the derived distribution of the linked variable can be obtained. Section 3.4 illustrates the method in more detail.

Briscolini et al. (2018) introduce a Bayesian approach that jointly solves the record linkage problem and the small area predictions. They also compare the Bayesian approach with the frequentist estimator proposed in Di Consiglio and Tuoto (2016). In the context of secondary data analysis, Han (2018) put forward an approach to solve small area estimation in presence of linkage errors.

The cited studies focus on the evaluating and the adjustment of linkage errors when small area prediction is performed by a unit level model. However, one might question whether the complexity of adjusting for linkage errors at unit level is in fact overwhelmed by the simplicity of area level models, which do not require unit level linkage for the estimation.

This paper aims at comparing the unit level estimator with the area level estimator in the presence of linkage errors, illustrating advantages and drawbacks by means of the application to real case and the simulation of various scenarios.

2. Linkage model and linkage errors

The reference theory for record linkage dates back to Fellegi and Sunter (1969). They consider the linkage between two lists, L_1 and L_2, of size N_1 and N_2 respectively. Within this context, we can consider, for instance, the linkage between a register and a sample. From a statistical viewpoint, the linking process is a classification problem; it aims to

classify all the pairs generated by the lists' comparison $\Omega = \{L_1 \times L_2\} = \{\omega = (i,j)\}$ where $i \in L_1$ and $j \in L_2$ into two independent and mutually exclusive subsets, M and U respectively;

- M is the set of links, grouping all the pairs composed by records belonging to the same unit $M = \{\omega = (i,j) \quad | \quad i = j\}$;
- U is the set of non-links $U = \{\omega = (i,j) \quad | \quad i \neq j\}$, where $i \in L_1$ and $j \in L_2$.

The classification decision is taken for each pair ω on the basis of the comparison on K linking variables, common to the two lists, e.g. name, surname, date of birth, address. The comparison on the linking variables results in a comparison vector γ_{ij}, e.g. $\gamma_{ij} = (1,1,0,1)$ if unit $i \in L_1$ and unit $j \in L_2$ present the same (or similar) values for the first, the second, and the forth linking variables and different (or quite dissimilar) value for the third linking variable. From the observed probability distribution of γ over the pair space Ω, two probability distributions are estimated:

- $m(\gamma_{ij})$, i.e. the probability of γ given that the pair (i,j) belongs to set M;
- $u(\gamma_{ij})$, i.e. the probability of γ given that the pair (i,j) belongs to set U.

To estimate the two distributions $m(\gamma_{ij})$ and $u(\gamma_{ij})$, and the prevalence of the links in the pairs $\pi = |M|/|\Omega|$ usually the EM algorithm is applied; details can be found in Jaro (1989), Herzog et al. (2007).

The classification procedure might produce two kinds of errors: the mismatch or false positive, when a pair (i,j) is classified as a link but in reality the two records i and j refer to different units, and the missing match or false negative, when the pair (i,j) is classified as a non-link but in reality the two records i and j belong to the same unit.

Linkage procedure aims at minimising both the probability of false match and the probability of missing match or, at least, to keep both below acceptable values. The classification procedure provides as a by-product the false positive rate and the false negative rate. For each pair, it also provides estimate of the probability of being a correct link given that the link is assigned:

$$\lambda_{ij} = \frac{m(\gamma_{ij})\pi}{m(\gamma_{ij})\pi + u(\gamma_{ij})(1-\pi)}. \tag{1}$$

The quantities λ_{ij} will be exploited for adjusting the linkage errors in the small area estimation framework described in the next section. It is worthwhile noting that accurate estimation of these probabilities is not a trivial task, even when the probabilistic linkage strategies are very effective in identifying the correct links. We will go back to this point in section 4, however the estimation of λ_{ij} is not the focus of this paper.

3. Small area estimation

When the survey is not planned to provide estimates at a very fine disaggregation (e.g. by geography or by a cross-classification such as gender and age), the standard estimates

are often too variable, because the sample size is too small or zero at the desired level. Small area estimation methods allow an improvement of the quality of the estimates exploiting relationships of the target variable with highly correlated auxiliary variables at unit level or area (domain) level. For an extensive review of small area methods, see Rao and Molina (2015).

In the following sub-sections we briefly overview the basic unit level (Battese-Harter-Fuller, 1988) and area level (Fay-Herriott, 1979) estimators. We describe how the former has to be modified to account for the linkage errors in the presence of auxiliary variables that are not recorded in the survey but obtained from an external source, such as administrative data.

3.1. The unit linear mixed model

Let the population units be partitioned into D different domains. Let Y be the target variable and X the auxiliary variables observed on the same units. Let us assume a linear mixed relationship between the target variable and the covariates

$$y_{id} = X_{id}^T \beta + u_d + e_{id}, \;\; i = 1, \dots, N_d, \;\; d = 1, \dots, D, \tag{2}$$

where β is a p-dimensional vector of fixed regression coefficients and u_d, $d = 1, \dots, D$, are the i.i.d. random variables related to the specific or domain contributions, with $E(u_d) = 0$ and $V(u_d) = \sigma_u^2$, independently distributed to the random errors e_{id} i.i.d. with $E(e_{id}) = 0$ and $V(e_{id}) = \sigma_e^2$. In matrix notation

$$Y = X\beta + Zu + e$$

where Z is the design matrix denoting the belonging of units to the areas: $Z = Blockdiag(Z_d = 1_{N_d}; d = 1 \cdots D)$.

The total variance is given by $V(Y) = V = \sigma_u^2 ZZ^T + \sigma_e^2 I$; equivalently, in matrix notation, $V = diag(V_d; d = 1 \cdots D)$ with $V_d = \sigma_e^2 I_{N_d} + \sigma_u^2 Z_d Z_d^T$. When σ_u^2 and σ_e^2 are known, the BLUP estimator of a small area mean or totals $\bar{Y}_d$, is given by

$$\hat{\bar{Y}}_d^{BLUP} = \frac{1}{N_d} \left(\sum_{i \in s_d} y_{id} + \sum_{i \in s_d^c} \hat{y}_{id}^{BLUP} \right) \tag{3}$$

where s_d is the sample in area d, $\hat{y}_{id}^{BLUP} = X_{id}^T \hat{\beta} + \tilde{u}_d$ with

$$\hat{\beta} = (X_s^T V_{ss}^{-1} X_s)^{-1} X_s^T V_{ss}^{-1} y$$

and $\tilde{u} = \sigma_u Z_s^T V_{ss}^{-1} (y - X_s \hat{\beta})$, where y is the sample vector of Y and denoting with the subscript s the portion of vector and matrices related to the sample observations.

In real cases, the estimates are given by the EBLUP that is obtained by plugging the estimates $\hat{\sigma}_u$ and $\hat{\sigma}_e$ into V and then into the previous expressions of $\hat{\beta}$ and $\tilde{u}$. See the section (sec.3.5) for a brief overview of the variance components estimation.

3.2. Area level small area predictor

The basic area level model (Fay and Herriot, 1979) relies on a linear relationship between the direct estimates $\hat{\bar{Y}}_d$ and the true finite population values $\bar{Y}_d$ in each area d, and a linear relationship among the true values and known area totals X_d:

$$\hat{\bar{Y}}_d = \bar{Y}_d + \varepsilon_d \qquad d = 1, \ldots, D, \tag{4}$$

where ε_d is the sampling error in the estimation of $\bar{Y}_d$, with mean zero and assumed known variance σ^2_{ed}, and

$$\bar{Y}_d = X_d\beta + u_d \qquad d = 1, \ldots, D, \tag{5}$$

where β is the vector of regression coefficients and u_d is assumed to be normal with zero mean and variance σ^2_u. Combining (4) and (5) one gets:

$$\hat{\bar{Y}}_d = X_d\beta + \varepsilon_d + u_d \qquad d = 1, \ldots, D, \tag{6}$$

where ε and u are assumed to be independent.

The BLUP estimator based on the model in (6) is given by:

$$\tilde{\bar{Y}}^{FH}_d = \gamma_d \hat{\bar{Y}}_d + (1-\gamma_d) X_d \hat{\beta} \qquad d = 1, \ldots, D, \tag{7}$$

where $\gamma_d = \sigma^2_u/(\sigma^2_u + \sigma^2_{ed})$. The EBLUP is obtained by replacing an estimate (e.g ML or REML estimate) of σ^2_u in formula (7). See Molina and Rao (2015) for more details. The FH model assumes known σ^2_{ed}. In practice it has to be estimated. See section (4) for more details on how it is estimated in the present work.

3.3. Linear mixed model under Record Linkage

When the auxiliary variables X and target variable Y are not jointly observed on the same data set but are obtained, for instance, by linking a sample with a register, the use of the relationship (2) and the corresponding estimator can produce biased estimates, if naively applied on linked data. Di Consiglio and Tuoto (2016) analyse the effect of linkage errors on unit level small area estimators and propose an adjustment to account for linkage errors, following the setting in Chambers (2009) and Samart and Chambers (2014).

The proposed adjustment, however, requires that no linkage errors occur between blocks/small areas. Under this assumption, the area level estimator is not affected by linkage errors and therefore linkage bias, since it only needs the mean value of X for each of the target domains. Hence, under the assumption of no linkage errors between areas, the standard Fay-Herriot estimator can be applied even in the presence of linkage errors within the small areas.

Let us first consider a register-register linkage and describe the linear mixed model and the proposed adjustment in this linkage setting.

Let us denote with y^*_{id} the value of the variable Y from one register that is matched with the value X_{id} in the other register, for unit i in domain d.

Let us assume that the blocking variable Z is measured without error on both the Y-register and the X-register, and that the partition of the registers introduced by Z is such that linkage errors only occur within this blocking variable.

Finally, let us assume an exchangeable linkage error model (see Chambers, 2009), i.e. the probability of correct linkage is the same for all records in block q, $q = 1, \cdots, Q$.

Under the following standard assumptions, as in Chambers (2009) and in Samart and Chamber (2010):

1. the linkage is complete, i.e. the X-register and Y-register refer to the same population and have no duplicates;

2. the linkage is one to one between the Y- and X-registers;

3. exchangeable linkage error model;

the observed linked variable Y^* is a permutation of the true one Y: $Y^* = AY$, where A is a random permutation matrix such that $E(A|X) = E$. The blocking index q is omitted in previous equations for simplicity of notation.
Being $Pr(a_{ii} = 1|X) = Pr(correct\ \ linkage) = \lambda$ and $Pr(a_{ij} = 1|X) = Pr(incorrect\ \ linkage) = \psi$, the expected value $E(A|X) = E$ can be written as:

$$E = (\lambda - \psi)I + \psi 1 1^T. \tag{8}$$

In this setting, Samart and Chambers (2014) proposed a ratio type corrected estimator for the regression coefficients β:

$$\tilde{\beta}_R = (X^T V^{-1} E X)^{-1} X^T V^{-1} y^* \tag{9}$$

following the same rationale of the bias correction estimator in the linear model (Chambers, 2009). They also proposed an approximation of the BLUE estimator by exploiting the new relationship between Y^* and X:

$$\tilde{\beta}_C = (X^T E^T \Sigma^{-1} E X)^{-1} X^T E^T \Sigma^{-1} y^* \tag{10}$$

where the derived variance $V(Y^*)$ of the observed y^* is considered:

$$V(Y^*) = \Sigma = \sigma_u^2 K + \sigma_e^2 I + W \tag{11}$$

with

$$W \approx diag((1-\lambda)(\lambda(f_i - \bar{f}) + \bar{f}^{(2)} - \bar{f}^2)) \tag{12}$$

being $f_i = X_i \beta$ and K a function of the number of areas within a block, block-group sizes and λs; see Samart and Chambers (2014) for more details. Clearly, the estimation of β requires an iterative process as Σ depends on β via the f. Moreover, the variance components are unknown and have to be estimates. The linkage errors can affect also

their estimation, see section 3.5 for a short description of how Samart and Chambers (2014) propose to deal with this issue.

3.4. Unit level small area predictor under linkage errors

Let us now consider the more realistic situation when the linkage is between a sample, where the variable Y is observed, and a register where X is recorded; this is the case where mixed models are useful for small area estimation.

In the sample-to-register setting, Chambers (2009) adds the assumption that the sampling does not change the outcome of the linkage process, i.e. selecting a record to be in sample does not change the register record to which it would be linked if all records were linked. Hence, the same permutation of the y described above would apply. This scheme works as if a hypothetical linkage can be performed before the sampling process and then we observe the sampled sub-set.

This assumption, as already pointed out by Chambers (2009), can be easily challenged as the sampling process may indeed affect the linkage process, but it is very useful in extending the register-register estimation setting to the survey-register situation.

Under the given conditions, the matrices E, V and Σ depend only on blocking variables and linkage errors, so there is no need to use sampling weights.

If the exchangeable linkage error model is assumed, as in section 3.3, the linkage errors occur only within the same block where records have the same probability of being correctly linked, then the mixed model can be fitted with the observed sample quantities applying the same argument as in the register-to-register case. See Chambers (2009) for more details.

Finally, for the small area estimation, we assume that small areas coincide with blocks. Note that with the latter assumption, the target mean of y is the same as the mean of the linked Y^*:

$$\hat{\bar{Y}}^* = \hat{\bar{Y}}.$$

Di Consiglio and Tuoto (2016) propose to exploit the distribution of Y^* to obtain the pseudo-BLUP estimator of $\bar{y}^*$ and then an estimation of $\bar{y}$:

$$\hat{\bar{Y}}_d^{*BLUP} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{id}^* + \sum_{i\in s_d^c} \hat{y}_{id}^{*BLUP}\right) \tag{13}$$

where $\hat{y}_{id}^{*BLUP} = EX\tilde{\beta}_C + \tilde{u}_d$, $\tilde{u} = \sigma_u Z^T \Sigma^{-1}(y^* - EX\tilde{\beta}_C)$ and $\tilde{\beta}_C$ is given in formula (10).

The pseudo-EBLUP estimator is given by replacing the estimates of the variance components (as in section 3.5) into the estimates of $\tilde{\beta}_C$ and $\tilde{u}$ and then in (13).

3.5. Estimation of variance components

The BLUE and the approximate BLUE estimators considered in the previous sections are based on known variance components. However, the variance components σ_u and σ_e are usually unknown, they are commonly estimated by methods of moments, ML or REML (Harville, 1977, Searle et al 2006). In Samart and Chambers (2014), a Pseudo-ML and

Pseudo-REML are proposed for adjusting variance component estimation for linkage errors. In the application and simulation study reported in section 4, we consider only ML approach, and pseudo-ML for the linkage error framework, assuming multivariate normal distribution.

In general, there is no analytical expression for the ML variance component estimator and the method of scoring is applied. When variables X and Y are both recorded on the sample, hence no linkage errors, the target variable is $y \sim N(X\beta; V)$. On the other hand, in the presence of linkage errors, we should use the modified distribution $y^* \sim N(Ef; \Sigma)$. The scoring algorithm can be applied on the derivatives of this likelihood rather than of the likelihood of the un-observed target variable y.

In the presence of linkage errors, estimates of β can be obtained from formulas (9) or (10) by replacing the variance components with their estimates. An iterative process is needed between the pseudo-ML estimates of the variance components and the estimate of β. See Samart and Chambers (2014) for more details.

4. Results on real and simulated data

Previous estimators are applied to a realistic case for estimating small areas in the presence of linkage errors. In addition, several synthetic populations have been generated based on two different mixed linear models to test the performance of estimators in a controlled environment. This section illustrates the real case and the data generation for the controlled experiment and describes the result.

4.1. The real case data

Microdata from the Survey on Household Income and Wealth, Bank of Italy, (SHIW), can be used to study the relationship between the consumption (the variable Y observed throughout the survey) and the net disposable income (the variable X available for the whole population). The survey sample is designed to produce reliable estimates at NUTS1 level, but the relevance of the topic prompts analysis of the results at the finer level, i.e. the NUTS2 administrative regions, which therefore represent a small area of estimation. In fact, variables Y and X are both observed by the survey: this allows us to compare different settings for linkage and mixed model estimation, knowing the true value of the regression model parameters. However, in principle one can imagine to study the relationship between the consumption recorded via the survey and the income from the tax register, available to the entire Italian population, thus overcoming the households' reluctance to provide information on income via a survey.

To overcome privacy issue and guarantee the reproducibility of the experiment, the record linkage procedure is applied to the fictitious population census data (McLeod et al. 2011) created for the ESSnet DI, an European project on data integration that run from 2009 to 2011. The population size is over 20000 records; data contain linking variables (names, dates of birth, addresses) for individual identification with missing values and typos, mimicking a real situation. The small domains are defined as aggregation of postal codes, assigning 18 areas. From this population, 100 replicated samples of size 1000 were

Table 1: True values of the correct linkage rates

Scenario	Min(λ)	Mean(λ)	Max(λ)	MMR
A	0.9525	0.9730	0.9834	0.0629
B	0.8430	0.8757	0.9043	0.0424

average values in 100 replications, over the 18 areas

independently randomly selected without replacement. Finally on each replication, the sample containing the Y variable was linked with the register reporting the X variables. The linkage was performed by means of the batch version of the software RELAIS (2015) that implements the probabilistic record linkage model (Fellegi and Sunter, 1969; Jaro, 1989).

We considered two linkage scenarios, characterized by two different sets of linking variables: in Scenario A we used "Day, Month, and Year of Birth"; in Scenario B we adopted "Day and Year of Birth", and "Gender". The first scenario uses linking variables with higher identifying power than the second scenario, producing fewer linkage errors in the results (both in terms of missing and false links). In both scenarios we assume that false linkage errors between different areas do not occur, in other words the administrative areas, i.e. the small domains are the blocking variable for the linkage procedures. Both scenarios also contain missing matches, mimicking the real outcomes of linkage procedures. Missing matches are mainly due to typos in the linking variables and hence they are independent from the target variable Y and the auxiliary variables X. In few words, they can be considered missing at random. However, they have the effect of reducing the sample size. Therefore, in the presence of linkage procedure, the estimators rely on the linked subset s_{Ld} of the sample s_d for the domain d.

True matches are known for the ESSnet DI data, so one can calculate the true value of the linkage errors for the proposed scenarios by comparing the obtained links with the true matches. Therefore, the value of the probability of correct link, λ, is calculated for each block (small area), as the ratio between the true matches in the linked set and the links within each area. Table 1 summarizes the results of the linkage procedures for the 100 replicas, showing the statistics for the probability of correct link λ, on average in the 18 areas. Moreover, Table 1 reports the average of the missing match rate, MMR, in the 18 areas for the 100 replicas, calculated as one minus the ratio between the numbers of identified links and the true matches. As expected, in the two scenarios, there is a trade-off between false matches and missing matches: scenario A has a lower false match rate but a higher missing match rate and vice-versa for scenario B.

For the adjusted estimator introduced in section 3.4, we use the true false linkage rate, $1-\lambda$, in each area. We do not simulate additional evaluation of λs, as the accurate estimation of λ is still an open research question in record linkage and it is not in the focus of this paper. However, at the end of the simulation study, we propose an insight into the behavior of the estimators when the linkage errors are overestimated.

The experiment considers five estimators for comparison:

1. BHF : is the EBLUP based on the Battese-Harter-Fuller model with X and Y observed on the same dataset, i.e. no linkage is assumed in this setting:

$$\hat{\bar{Y}}_d^{BHF} = \frac{1}{N_d}\left(\sum_{i\in s_d} y_{id} + \sum_{i\in s_d^c} \hat{y}_{id}^{EBLUP}\right),$$

where s_d is the sample in area d, $\hat{y}_{id}^{EBLUP} = X_{id}^T\hat{\beta} + \hat{u}_d$ with

$$\hat{\beta} = (X_s^T \hat{V}_{ss}^{-1} X_s)^{-1} X_s^T \hat{V}_{ss}^{-1} y$$

and $\hat{u} = \hat{\sigma}_u Z_s^T \hat{V}_{ss}^{-1}(y - X\hat{\beta})$.

2. BHF_L : is the EBLUP based on the Battese-Harter-Fuller model on the subset of linked records. In this estimator we reduce the sample size to the linked records but we do not introduce linkage errors; this is our benchmark:

$$\hat{\bar{Y}}_d^{BHF\_L} = \frac{1}{N_d}\left(\sum_{i\in s_{Ld}} y_{id} + \sum_{i\in s_{Ld}^c} \hat{y}_{id}^{EBLUP}\right),$$

where s_{Ld} is the sub-set of linked sample units in area d.

3. BHF_naive : is the naïve EBLUP based on the Battese-Harter-Fuller model on the subset of linked records, considering X and Y observed on two different datasets, without adjustment for linkage error:

$$\hat{\bar{Y}}_d^{BHF\_naive} = \frac{1}{N_d}\left(\sum_{i\in s_{Ld}} y_{id}^* + \sum_{i\in s_{Ld}^c} \hat{y}_{id}^{*EBLUP\_naive}\right),$$

where s_{Ld} is the sub-set of linked sample units in area d, $\hat{y}_{id}^{*EBLUP\_naive} = X_{id}^T\hat{\beta}^* + \hat{u}_d$ with

$$\hat{\beta}^* = (X_{s_L}^T \hat{V}_{s_L s_L}^{-1} X_{s_L})^{-1} X_{s_L}^T \hat{V_{s_L s_L}}^{-1} y^*$$

and $\hat{u} = \hat{\sigma}_u Z^T \hat{V}_{s_L s_L}^{-1}(y^* - X_{s_L}\hat{\beta}^*)$.

4. BHF_adj: is the adjusted EBLUP based on the Battese-Harter-Fuller model:

$$\hat{\bar{Y}}_d^{BHF\_adj} = \frac{1}{N_d}\left(\sum_{i\in s_{Ld}} y_{id}^* + \sum_{i\in s_{Ld}^c} \hat{y}_{id}^{*EBLUP}\right),$$

where $\hat{y}_{id}^{*EBLUP} = EX\hat{\beta}_C + \hat{u}_d$ and $\hat{u} = \hat{\sigma}_u Z^T \hat{\Sigma}^{-1}(y^* - EX\hat{\beta}_C)$, and $\hat{\beta}_C$ is given by

$$\hat{\beta}_C = (X_{s_L}^T E_{s_L}^T \hat{\Sigma}_{s_L s_L}^{-1} E_{s_L} X_{s_L})^{-1} X_{s_L}^T E_{s_L}^T \hat{\Sigma}_{s_L s_L}^{-1} y^*.$$

5. FH : is the EBLUP based on the Fay-Herriot model:

$$\tilde{\bar{Y}}^{FH} = \hat{\gamma}_d \hat{\bar{Y}} + (1 - \hat{\gamma}_d) X_d \hat{\beta},$$

where $\hat{\gamma}_d = \hat{\sigma}_u / (\hat{\sigma}_u + \hat{\sigma}_{ed})$. The FH model assumes known sampling variance σ^2_{ed}, however it needs to be estimated in practice. In this simulation, we used a simple minded smoothing method, which assumes that the population variances of all the domains are identical, σ^2_e. The variances of the direct estimators are then evaluated as $\hat{\sigma}^2_e / n_d$ where $\hat{\sigma}^2_e$ is estimated from the unit linear model.

It is worth noting that the five estimators are evaluated on different sub-sets; the BHF estimator and the FH estimator are evaluated on the sample s_d, the BHF_naive and the BHF_adj estimators are evaluated on the linked sub-set s_{Ld} that might include linkage errors; the estimator BHF_L is evaluated on the sub-sample s_{Ld} but the correct values of X in the register have been used.

Table 2 reports the average of the Absolute Relative Error (ARE) over the 18 areas, the average of the Standard Deviation (SD), and the average of the Mean Square Error (MSE). Results in table 2 show that in terms of bias the area level estimator outperforms the unit level estimators, even when linkage error correction is applied. However, in terms of variability, the area level estimator shows values considerably higher compared to the other estimators. We assumed equal population variances in all domains in the implementation of the Fay-Herriot model. This assumption may be not appropriate in our context, highlighting that sampling variance smoothing deserves great attention in the application of the FH estimator. We will return to this point in the concluding remarks, though the variance estimation is not the focus of this paper, see Hawala and Lahiri (2018) for some ideas on variance modeling.

Table 2 shows that the adjusted unit level EBLUP (BHF_adj) reduces the bias with respect to the naïve estimator (BHF_naive), at the price of an increase in variance that is, however, compensated at MSE level. In fact, the MSE of the adjusted unit level EBLUP (BHF_adj) is similar to that of the benchmark estimator (BHF_L), based on the linked sample without errors. Similar results are also in Di Consiglio and Tuoto (2016), and in Briscolini et al. (2018). It is worth noting that the adjustment for linkage errors does not completely eliminate the bias. We will return to this point in our concluding remarks.

4.2. Simulated data

In the previous subsection, the comparison of the unit level and area level estimators in the presence of linkage errors can be affected by the actual relationship between the variables, which are observed in the field and interpreted with linear mixed models to pursue our purposes.

To compare the unit level and the area level estimators in the presence of linkage errors in a fully controlled setting, we create two different models, Model1 and Model2,

Table 2: Average of the absolute relative error (ARE), standard deviation (SD), and Mean Square Error (MSE) for estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH

ARE

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.0330	0.0333	0.0350	0.0335	0.0231
B	0.0330	0.0334	0.0430	0.0347	0.0231

SD

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.4659	0.4820	0.4729	0.4762	2.3188
B	0.4659	0.5426	0.5107	0.5262	2.3188

MSE

Scenario	BHF	BHF_L	BHF_naive	BHF_adj	FH
A	0.6753	0.6906	0.6981	0.6913	2.3336
B	0.6753	0.7358	0.7938	0.7383	2.3336

based on the following linear mixed models:

$$Model1: X \sim [1, Uniform(0,1)], \quad \beta = [2,4], \quad u \sim N(0,1), \quad e \sim N(0,3),$$
$$RealizedVar(u) = 1.5728$$
$$Model2: X \sim [1, Uniform(0,1)], \quad \beta = [2,4], \quad u \sim N(0,3), \quad e \sim N(0,1),$$
$$RealizedVar(u) = 4.7186.$$

The variables from the two models have been attached to the ESSnet DI data, containing the linking variables. The previous linking scenarios, A and B, have been considered for each model. Then, for each model, 100 replicated samples of size 1000 were independently and randomly selected without replacement; finally, for each replication, the sample containing the variable Y was linked to the register that reported the variables X.

As in the previous section, five estimators are compared: BHF, BHF_L, BHF_naive, BHF_adj and FH. Table 3 reports the Absolute Relative Error (ARE), the Standard Deviation (SD), and the Mean Square Error (MSE), averaged over the 18 areas, for linkage scenario B. The results for linkage scenario A are substantially similar and are not presented here for the sake of brevity.

For BF estimators, bias and variance are smaller in Model 2 than in Model 1. This is not the case for the FH estimator. As already observed with real data, the bias reduction of the adjusted estimator BHF_adj more than offsets the increase in variance, so the

Table 3: Average of the absolute relative error (ARE), standard deviation (SD), and Mean Square Error (MSE) for estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH

	ARE				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.0412	0.0423	0.0476	0.0435	0.0401
Model2	0.0135	0.0137	0.0199	0.0161	0.0266

	SD				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.3265	0.3447	0.3349	0.3424	0.9108
Model2	0.2263	0.2412	0.2522	0.2519	1.0040

	MSE				
	BHF	BHF_L	BHF_naive	BHF_adj	FH
Model1	0.3837	0.4018	0.4060	0.4013	0.9240
Model2	0.2333	0.2476	0.2652	0.2595	1.0064

MSE of estimator BHF_adj is always smaller than the MSE of estimator BHF_naive. The improvement is quite small when the linkage errors are small. As far as the area level estimator is concerned, it performs better than the BHF estimators in terms of bias in Model 1, whilst the FH performs worse than the unit level estimators, including the not-adjusted estimator BHF_naive in Model 2. In terms of variability, as anticipated in the previous section on real data, the area level estimator FH performs worse than the others, in both scenarios and in both models. The boxplot in figure 1 shows the relative errors for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH, in the 18 areas.

Figure 2 shows the standard deviations for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas. The distribution over the areas basically confirms the behavior of the estimators highlighted in table 3.

These evidences do not allow us to answer in a definitive way to the initial question of the possible advantage of the FH which, unlike the unit level estimator in the presence of linkage errors, does not require unit linkage. This simulation study seems to suggest that there are situations (Model 1, real data of previous section) where the area level estimator can perform well enough and one can avoid to complicate the analysis introducing record linkage to apply an adjusted unit level estimator, if the FH guarantees enough accuracy. However, there are also contexts (e.g. Model 2) that show the advantages of considering auxiliary information at record level, even in the presence of uncertainty introduced by record linkage. As an aside, one should be careful on using an appropriate smoothing method for the variance of the direct for the FH estimator.

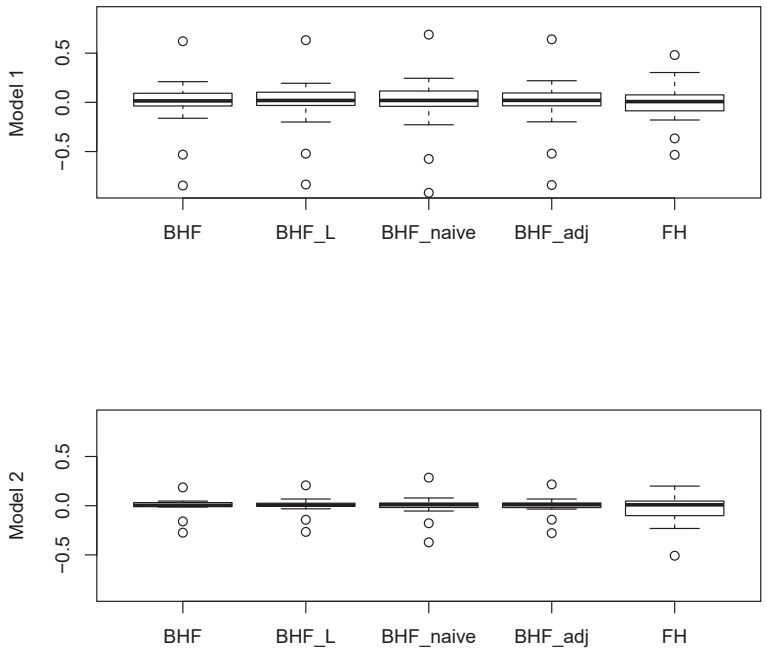


Figure 1: Boxplot of the relative errors for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas

A comparison can be made between unit level and area level estimators when linkage errors are not accurately evaluated. As already discussed in the previous section, in this analysis we know the true value of the linkage errors and use them for the adjustments. However, generally in real cases, assessing linkage errors is not an easy task, the research on the topic is still active, some proposals include Belin and Rubin (1995), Tuoto (2016), and Chipperfield and Chambers (2015). To account for difficulties in assessing linkage errors, we propose a sketch on the behavior of the small area estimators when linkage errors are not accurately evaluated. When linkage errors are underestimated, we tend to make estimates such as the naïve. So, let's focus on the behavior of unit level and area level estimators when linkage errors are overestimated. To overestimate the linkage errors, within each small domain we treat the observed range of false linkage rate as if it were normally distributed, then we evaluate a 95% normality-based confidence interval for $1-\lambda$, and we consider the superior extreme of the confidence intervals as values of $1-\lambda$ in the estimator BHF_adj for the 100 replications.

In this analysis, we only consider the Scenario B, which shows the highest linkage error levels. The boxplot of the values of λ within the 18 areas in the 100 replications is

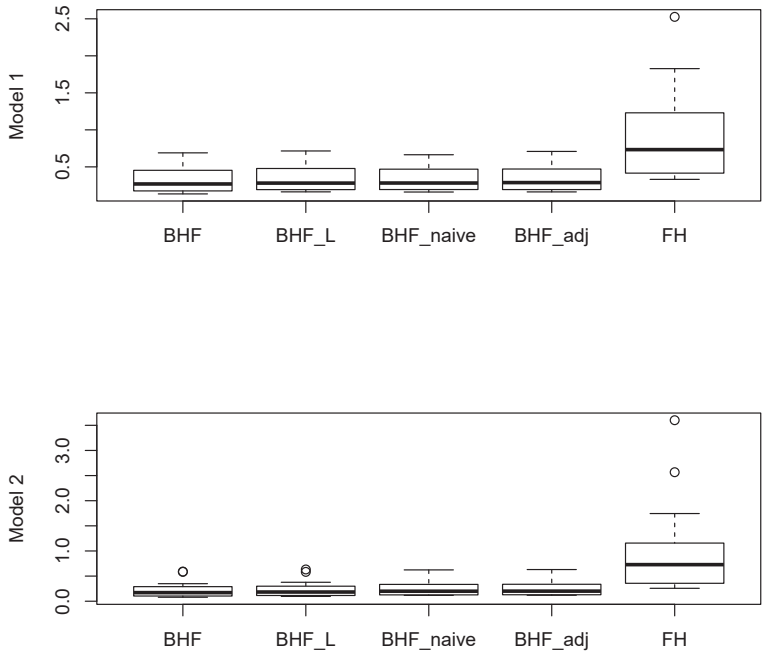


Figure 2: Boxplot of the standard deviations for the estimators BHF, BHF_L, BHF_naive, BHF_adj, and FH in the 18 areas

shown in Figure 3. It is worth noting that the areas with the lowest linkage errors (i.e. area M3 and area M7) are the smallest ones, both in terms of population and sample. No linkage errors in these areas is a realistic assumption, since the small size of the areas avoids false matches.

Table 4 shows the average of the Absolute Relative Error (ARE), the Standard Deviation (SD), and the Mean Square Error (MSE), in the 18 areas, for estimators FH and BHF_adj.

Table 4 confirms the observed behavior and the relationship between area level and unit level estimator, even when the linkage errors are not accurately measured. Still in terms of bias, the FH estimator is preferable to the adjusted BHF estimator in Model 1, whilst the vice-versa in observed for Model 2. In terms of variability, the BHF estimator outperforms the FH estimator in both models.

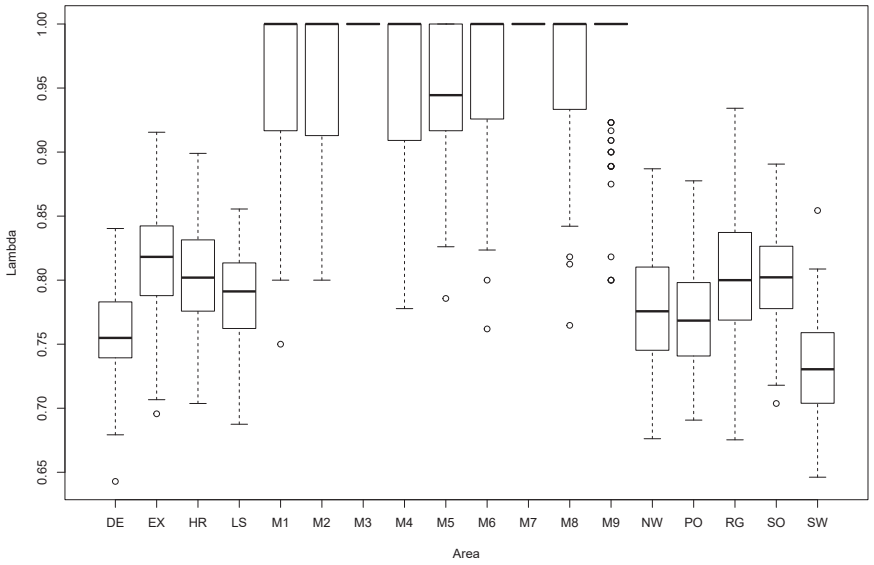


Figure 3: Boxplot of the values of λ in the 100 replications within the 18 small areas

5. Concluding remarks and future works

We explored the behavior of unit level and area level estimators in the presence of linkage errors. The area level is, in principle, quite attractive as it does not require record linkage at all. However, with both realistic and simulated data, we find that the use of auxiliary information at unit level is still useful, even if it exposes to the risk of unit identification errors.

As already noted, the implementation of the area level estimator under the Fay-Herriot model needs reliable smoothed estimates of the sampling variability. We used a simple minded smoothing method, which assumes that the population variances of all the domains are identical. This might be a strong assumption and it might have an

Table 4: Average of the Absolute Relative Error (ARE), Standard Deviation (SD), and Mean Square Error (MSE) for estimators BHF_adj and FH when linkage errors are over-estimated

ARE		SD		MSE		
	BHF_adj	FH	BHF_adj	FH	BHF_adj	FH
Model1	0.0422	0.0401	0.3470	0.9108	0.4009	0.9240
Model2	0.0142	0.0266	0.2554	1.0040	0.2608	1.0064

impact on our results. Further work is needed to improve the variance smoothing for the FH estimator.

In this work, the linkage error adjusted unit level estimator is the one suggested in Di Consiglio and Tuoto (2016) and Briscolini et al. (2018). In the adjustment, we assumed block specific probabilities of correct link are known and this is indeed a strong assumption (see remark 2 (3) of Han and Lahiri, 2018). Moreover, the proposed adjustment assumes the exchangeability of linkage errors, and the small areas coinciding with the blocks of the linkage process. As already noted in Di Consiglio and Tuoto (2016) and in Section 4, the adjustment at unit level does not completely remove the bias introduced by linkage errors. This can be the result of the fact that the exchangeability assumption is not perfectly met.

While our evaluation does not provide a definite answer, we hope our paper encourages others to design an extensive evaluation experiment in order to compare BHF estimator corrected for linkage error with the EBLUP under the Fay-Herriot model that does not require any correction for linkage errors.

In the future, we propose to expand our simulation experiment to include the framework proposed by Han and Lahiri (2018) to correct the unit level small area estimation and to benefit from the use of unit level information to improve estimators, even in the presence of linkage errors. One of the promising advantages of the Han and Lahiri's setting is that it does not require any exchangeability assumption. In Han's dissertation thesis (Han, 2018), she suggests an integrated model where the information about the linkage is carried by all record pairs (links and non-links). In this way all record pairs contribute to the estimation process and to correct for linkage bias. This model is different from the secondary data analysis, adopted in this paper, where only the designated links are considered. More in details, the linkage process is viewed as a permutation of the true covariates associated with the observed target variables within a block/small area. Under the assumption that the random errors and random effects are independent from the observed linked covariates and the comparison matrix of the linkage, given the true covariates values, an Empirical Best Predictor is derived.

Acknowledgments

We wish to thank the Guest Editor-in-Chief Prof. Partha Lahiri for suggesting this research problem and encouraging us to follow up this work.

REFERENCES

BELIN, T., RUBIN, D. B., (1995). A method for calibrating false - match rates in record linkage, *Journal of the American Statistical Association*, 90, pp. 694–707.

BATTESE, G. E., HARTER, R.M., FULLER, W. A., (1988). An Error-Components Model for Prediction of Crop Areas Using Survey and Satellite Data, Journal of the American Statistical Association, 83, pp. 28–36.

BRISCOLINI, D., DI CONSIGLIO, L., LISEO, B., TANCREDI, A., TUOTO, T., (2018). New methods for small area estimation with linkage uncertainty. International Journal of Approximate Reasoning, 94, pp. 30–42.

CHAMBERS, R., (2009). Regression analysis of probability-linked data, Official Statistics Research Series, Vol. 4.

CHIPPERFIELD, J. O., CHAMBERS, R. L., (2015). Using the Bootstrap to Account for Linkage Errors when Analysing Probabilistically Linked Categorical Data, Journal of Official Statistics, Vol. 31, No. 3,

DI CONSIGLIO, L., TUOTO, T., (2016). Small Area Estimation in the Presence of Linkage Errors. In International Conference on Soft Methods in Probability and Statistics, pp. 165–172. Springer, Cham.

DI CONSIGLIO, L., TUOTO, T., (2018). When adjusting for the bias due to linkage errors: A sensitivity analysis. Statistical Journal of the IAOS, 34(4), pp. 589–597.

FELLEGI, I. P., SUNTER, A. B., (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64, pp. 1183–1210.

FAY, HERRIOTT, (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association 74, pp. 269–277.

HAN, Y., (2018). Statistical Inference Using Data From Multiple Files Combined Through Record Linkage, PhD Dissertation thesis, downloadable at https: //drum.lib.umd.edu/bitstream/handle/1903/21155/HAN_umd_0117E_19360.pdf

HAN, Y., LAHIRI, P., (2018). Statistical analysis with linked data. International Statistical Review, 87, S139–S157.

HARVILLE, D. A., (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. Journal of American Statistical Association, 72, pp. 320– 338.

HAWALA, S., LAHIRI, P., (2018). Variance Modeling for Domains. Statistics and Applications, 16, pp. 399– 409.

HERZOG, T. N., SCHEUREN F.J., WINKLER, W. E., (2007). Data Quality and Record Linkage Techniques, Springer Science & Business Media.

JARO, M., (1989). Advances in record linkage methodology as applied to matching the 1985 test census of Tampa, Florida. Journal of American Statistical Association, 84, pp. 414–420.

LAHIRI, P., LARSEN, M. D., (2005). Regression Analysis With Linked Data. Journal of the American Statistical Association, 100, pp. 222–230.

MCLEOD, P., HEASMAN, D. and FORBES, I., (2011). Simulated data for the on the job training, http://www.cros-portal.eu/content/job-training.

NETER, J., MAYNES, E. S, RAMANATHAN, R., (1965). The effect of mismatching on the measurement of response errors. Journal of the American Statistical Association, 60, pp. 1005–1027.

RAO, J. N. K., MOLINA, (2015). Small Area Estimation, Second Edition, Wiley, New York.

RELAIS 3.0 User's Guide, (2015). available at http://www.istat.it/it/strumenti/metodi-e-strumenti-it/strumenti-di-elaborazione/relais.

SEARLE, S. R., CASELLA, G., MCCULLOCH, C. E., (2006). Variance Components, Wiley, New York.

SAMART, K., (2011). Analysis of probabilistically linked data, PhD thesis, School of Mathematics and Applied Statistics, University of Wollongong.

SAMART, K., CHAMBERS, R., (2010). Fitting Linear Mixed Models Using Linked Data, Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper pp. 18–10.

SAMART, K., CHAMBERS, R., (2014). Linear regression with nested errors using probability-linked data, Australian and New Zealand Journal of Statistics 56.

SCHEUREN, F., WINKLER, W. E., (1993). Regression analysis of data files that are computer matched – Part I. Survey Methodology, Volume 19, pp. 39–58.

SCHEUREN F., WINKLER W. E., (1997). Regression analysis of data files that are computer matched- part II, Survey Methodology, 23, pp. 157–165.

TANCREDI, A., LISEO, B., (2011) A hierachical Bayesian approach to record linkage and population size problems. Annals of Applied Statistics, 5, pp. 1553–1585.

TUOTO, T., (2016). New proposal for linkage error estimation" Statistical Journal of the IAOS, Vol 32, no. 2, pp. 1–8.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 123–143, DOI 10.21307/stattrans-2020-034
Received – 31.01.2020; accepted – 30.06.2020

High dimensional, robust, unsupervised record linkage

Sabyasachi Bera[1], **Snigdhansu Chatterjee**[2]

ABSTRACT

We develop a technique for record linkage on high dimensional data, where the two datasets may not have any common variable, and there may be no training set available. Our methodology is based on sparse, high dimensional principal components. Since large and high dimensional datasets are often prone to outliers and aberrant observations, we propose a technique for estimating robust, high dimensional principal components. We present theoretical results validating the robust, high dimensional principal component estimation steps, and justifying their use for record linkage. Some numeric results and remarks are also presented.

Key words: record linkage, principal components, high dimensional, robust.

1. Introduction

In recent times, owing to rapid advancement of a variety of technological resources and services, and increasingly digitally connected environment, numerous kinds of datasets are available. For example, for a given community of individual's, there may be very high dimensional data available on each individual's (i) online shopping patterns, as well as on their (ii) social media presence and usage. It may of interest to businesses to understand their customers better based on their social and cultural backgrounds, consequently it is of interest to link an online shopper's profile with their social media data. Owing to privacy rights of individuals and confidentiality concerns, identifying information may not be available to the statistician linking the records.

This paper is primarily on a methodology for linking high dimensional datasets of above type. Many existing approaches for entity resolution and record linkage are applicable only on low dimensional datasets, and where the datasets have shared features or variables. We do not require the two datasets to have a common set of features and in fact, present our discussion for the case where the datasets have no common variable.

In this context, we also develop a mathematical framework for the topic of record linkage, for better understanding and tractability of the theoretical properties of such linking algorithms. Parts of the existing literature on record linkage and entity resolution are based on *ad hoc* principles, and we hope to address some foundational challenges in this topic.

[1]University of Minnesota. USA. E-mail: berax008@umn.edu. ORCID: https://orcid.org/0000-0002-9053-4094

[2]University of Minnesota. USA. E-mail: chatt019@umn.edu. ORCID: https://orcid.org/0000-0002-7986-0470.

Additionally, it is routinely observed that high dimensional datasets can contain outliers or aberrant observations. Consequently a major aspect of our proposed methodology is to develop *robust* techniques for data linkage. Our proposal borrows recent studies on high-dimensional principal components and extends them to the case of robust principal components.

Another aspect of this paper is that we propose a computationally much simpler and easily implementable method for linking records than the available Bayesian approaches such as the ones given in LISEO and TANCREDI (2013). Other machine learning approaches involving graphs and networks that are sometimes adopted for entity resolution, also require heavy machinery computing. It is not clear if such extremely computation intensive methodology is either necessary, or whether there is a principled statistical foundation to such methodology. Apropos of this, the computational burden from our proposal is significantly lower.

This paper advocates a principled approach. Our approach is broadly as follows: we implement a robust, sparse, high-dimensional principal component analysis (PCA often hereafter) on both datasets, and consolidate the information about each observation (that is, each row of both matrices) into a low-dimensional vector (p_0 in the notations of this paper). Then, we compute correlations between these p_0 dimensional vectors from the two matrices, with the understanding that an existing linkage will show up as a highly correlated entry. The threshold for the correlation is based on the training set in the current paper, but our principle is workable even when there is no training set available. Owing to the facts that (a) our proposal requires no common features or variables common to both datasets, and (b) we do not require a training dataset, we call our proposal *unsupervised* record linkage.

In order to ensure clarity of our presentation and to keep the technicalities at a reasonable level, in this paper we only present results on unsupervised record linkage where all the variables are continuous in nature. In particular, commonly used variables for record linkage, like name, date of birth, address, do not satisfy our technical assumptions. In practice, we may use a traditional method using the nominal and ordinal variables to do a preliminary subsetting of potential linkages, after which the unsupervised record linkage method may be used on the continuous variables. Also, it is possible in some cases to use continuous variables as underlying latent variables governing the behavior of a categorical random variable. These directions of research will be part of our future work.

The rest of this paper is organized as follows: In Section 2 we present a brief and necessarily incomplete state of the literature on record linkage, in order to clarify how our contribution differs from the advancements on this topic thus far. Then, in Section 3 we discuss notations, the conceptual framework of the datasets that we propose to link, and the linking model. Following that, in Section 4 we present our statistical model and technical arguments. In particular, Section 4.1 contains the record linking algorithm, and Section 4.2 contains the theoretical framework and justifications for our statistical model and algorithmic steps. Then, in Section 5 we present some numeric results based on simulation studies, along with additional comments on practical implementation of our proposed methodology. A final Section 6 collects our concluding remarks.

2. A Broad Overview of Record Linkage

Record linkage, also often referred to entity resolution, de-duplication or co-reference is a widely used technique for identifying records referring to the same entity across different databases. Although this is a trivial task when unique error-free identifiers of the entities are recorded in the data files, in general it need to be solved in the absence of unique identifiers using other information that the sources have in common on the entities.

The seminal article by FELLEGI and SUNTER (1969) presented the first mathematical model for this topic, based on earlier work by NEWCOMBE and KENNEDY (1962) for one-to-one entity resolution across two databases in terms of Neyman-Pearson hypothesis testing.

In this brief review, we focus primarily on *bipartite record linkage*, where the key assumption is that each entity is recorded *at-most* once in each files. Most of the literature (including our set-up) on record linkage falls in this scenario. This assumption implies a maximum one-to-one restriction in the linkage, that is, a record from one file can be linked with maximum one record from the other file.

The main principle of bipartite record linkage may be described as follows: Consider two data files $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ that record information from two overlapping sets of individuals or entities. These data files contain n_ℓ records respectively (without loss of generality we assume $n_1 \leq n_2$) for $\ell = 1,2$ with n_0 being the number of entities simultaneously recorded in both files, hence $0 \leq n_0 \leq n_1$.

In the bipartite record linkage context, we can think of the records from files Y_1 and Y_2 as two disjoint sets of nodes, where an edge between two records represents them referring to the same entity, which we also call being co-referent or being a match. Formally, this match can be encoded into a matrix $\triangle_{n_1 \times n_2}$ as follows:

$$\triangle_{ij} = \begin{cases} 1 & \text{if records } i \in Y_1 \text{ and } j \in Y_2 \text{ represent the same entity} \\ 0 & \text{otherwise} \end{cases}$$

The characteristics of a bipartite matching imply that at-most one entry in each column and each row of $\triangle$ can be equal to 1. The goal of bipartite record linkage is to estimate $\triangle$ using the information contained in Y_1 and Y_2.

The set of ordered record pairs $Y_1 \times Y_2$ can be thought as the union of the set of matches $M = \{(i,j) : i \in Y_1,\ j \in Y_2,\ \triangle_{ij} = 1\}$ and the set of non-matches $U = \{(i,j) : i \in Y_1,\ j \in Y_2,\ \triangle_{ij} = 0\}$. Thus, the problem of estimating $\triangle$ from Y_1 and Y_2 can be seen as identifying the sets M and U. When record pairs are estimated to be matches they are called links and when estimated to be non-matches they are called non-links.

2.1. The Fellegi–Sunter Approach of Record Linkage

The key idea of the Fellegi-Sunter approach is as follows: Comparison vectors γ_{ij} are obtained for each record pair (i,j) in $Y_1 \times Y_2$ with the goal of finding evidence of whether they represent matches or not. These vectors can be written as $\gamma_{ij} = (\gamma_{ij}^1, \ldots, \gamma_{ij}^f, \ldots, \gamma_{ij}^F)$, where F denotes the number of criteria used to compare the records. Traditionally, these

F criteria correspond to one comparison for each variable that the data files have in common.

Let $S_f(i,j)$ denote a similarity measure computed from field f of records i and j. The range of S_f can be divided into L_f+1 intervals $I_{f0}, I_{f1}, \ldots, I_{fL_f}$, which represent different disagreement levels. In this construction, the interval I_{f0} represents the highest level of agreement, which includes total agreement, and the last interval I_{fL_f} represents the highest level of disagreement.

In their paper, FELLEGI and SUNTER (1969), the authors propose to the log-likelihood ratios

$$w_{ij} = \log \frac{\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]}{\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]}$$

as weights to estimate which record pairs are matches. The expression for w_{ij} assumes that γ_{ij} is a realization of a random vector, say, G_{ij} whose distribution depends on the matching status $\triangle_{ij}$ of the record pair. Similar to the Neyman-Pearson hypothesis testing, if this ratio is large we favor the hypothesis of the pair being a match.

When $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$ and $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ are known, the procedure orders the possible values of γ_{ij} by their weights w_{ij} in non-increasing order, indexing by the subscript h, and determines two values, $h^{'}$ and $h^{''}$, such that $\sum_{h\leq h'-1}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0] < \mu \leq \sum_{h\leq h'}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ and $\sum_{h\geq h''}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1] \geq \lambda > \sum_{h\geq h''+1}\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$, where $\mu = \mathbb{P}[\text{assign } (i,j) \text{ as link}|\triangle_{ij}=0]$ and $\lambda = \mathbb{P}[\text{assign } (i,j) \text{ as non-link}|\triangle_{ij}=1)$ are two admissible "type 1" and "type 2" error levels.

Finally, the record pairs are classified into 3 groups:

1. Those with $h \leq h^{'}-1$ are declared links

2. Those with $h \geq h^{''}+1$ are non-links and

3. Those with configurations between $h^{'}$ and $h^{''}$ require clerical review.

Fellegi and Sunter showed that this decision rule is optimal in the sense that for fixed values of μ and λ it minimizes the probability of sending a pair to clerical review.

However, in practice, $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=1]$ and $\mathbb{P}[\gamma_{ij}|\triangle_{ij}=0]$ are not known, and have to be estimated from Y_1 and Y_2. So, JARO (1989); LARSEN and RUBIN (2001) proposed to model the comparison data using mixture models of the type

$$\begin{aligned} G_{ij}|\triangle_{ij}=1 &\overset{iid}{\sim} M(m) \\ G_{ij}|\triangle_{ij}=0 &\overset{iid}{\sim} U(u), \\ \triangle_{ij} &\overset{iid}{\sim} \text{Bernoulli}(\theta) \end{aligned}$$

for comparison variables G_{ij}, some distributions $M(m)$ and $U(u)$, and $\theta \in (0,1)$. Estimation of M and U is usually done by EM-type algorithms.

2.2. Machine Learning/Classification Approach

In-general, record linkage task becomes quickly infeasible with size (n_ℓ) as well as the dimension (p_ℓ) of data files. A common solution to this problem is to partition the data files into blocks (e.g. geography, or gender and year of birth) of records determined by information that is thought to be accurately recorded in both data files, and then solve the task only within blocks. See CHRISTEN (2011); STEORTS et al. (2014) for extensive surveys. Earlier development into blocking is presented in HERZOG et al. (2007), who also discuss the use of blocking to identify duplicate list entires and for matching records between two sample surveys.

Recently a common approach of tackling the record linkage problems has been to treat it as a traditional supervised or semi-supervised classification problem: we need to classify record pairs into matches and non-matches. If we have a sample of record pairs for which the true matching statuses are known, we can train a classifier on this sample using comparisons between the pairs of records as our predictors, and then predict the matching statuses of the remaining record pairs. See MARTINS (2011); TORVIK and SMALHEISER (2009); TREERATPITUK and GILES (2009); VENTURA et al. (2015) for some examples.

2.3. Bayesian Methods

Bayesian methods have a long history of use in record linkage models. A major advantage of Bayesian methods is their natural handling of uncertainty quantification for the resulting estimates. For a review of recent development in Bayesian methods, see LISEO and TANCREDI (2013). While some of the Bayesian work incorporates the record data only through pairwise similarity scores (SADINLE, 2017; SADINLE and FIENBERG, 2013), other works (STEORTS et al., 2016) directly model the actual record data which usually requires crafting specific models for each type of field, and therefore mostly deal with categorical information. However, recently STEORTS et al. (2015) has generalized Bayesian methods to incorporate string variables such as addresses, phone numbers, or dates.

In addition, SADINLE and FIENBERG (2013) has extended the Fellegi-Sunter approach to linking records across more than two databases. Also, SINGLA and DOMINGOS (2006); ENAMORADO et al. (2018) generalized the underlying mixture models (specially the i.i.d. assumptions) in the Fellegi-Sunter approach.

3. Notations, the data and linking model

The main focus of this paper is to *link* observations from two datasets. Both datasets are matrices, with iid rows. However, the same observational units may have been used for both datasets. For example, the 17-th row of the first dataset and 47-th row of the second dataset may belong to the same individual. For a number of cases (n in the notation used below) we know the linkage, and thus can match and pair the information from both datasets. However, such linkage is not known for many other rows of both datasets. The main goal of a ***record linkage*** exercise is to establish such linkages.

It is also commonly understood that most observations from both datasets are linked, and only a handful of observations from either dataset do not have an entry on the other. For this paper, we assume that the datasets do not have any common variables. We also assume that both datasets considered here are high-dimensional.

3.1. Notations

Since the data, various parameters and latent variables will have multiple indices, we establish some notations first. The notation $a_\square$ denotes a vector, of dimension that will be determined by the context. All vectors are column vectors, and the notation $a_\square^T$ or a^T denotes the transpose, and $|a|$ denotes its Euclidean norm. The $n \times m$ matrix A has column vectors denoted by $A_{\square,1}, \ldots, A_{\square,m} \in \mathbb{R}^n$, and row vectors denoted by $A_{1,\square}, \ldots, A_{n,\square} \in \mathbb{R}^m$, thus

$$A = (A_{\square,1} : A_{\square,2} : \ldots : A_{\square,m})_{n\times m} = \begin{pmatrix} A_{1,\square} \\ A_{2,\square} \\ . \\ . \\ . \\ A_{n,\square} \end{pmatrix}_{n\times m}.$$

We index the datasets used in this paper by $\ell = 1,2$, and the notation Y_ℓ stands for the ℓ-th dataset with dimensions $n_\ell \times p_\ell$, consisting of n_ℓ independent observations of the p_ℓ features that are stacked as row vectors of the Y_ℓ matrix. The k-dimensional multivariate Normal distribution with mean $\mu \in \mathbb{R}^p$ and variance $\Sigma \in \mathbb{R}^{p\times p}$ will be denoted by $N_p(\mu,\Sigma)$. The notation $X_i \overset{i.i.d.}{=} \mathbb{F}$ denotes that the X_i's are independent, identically distributed according to $\mathbb{F}$.

3.2. The data and the linking framework

We consider two datasets for linkage, $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ for $\ell = 1,2$. Without loss of generality, $n_1 \le n_2$. In both datasets, each row represents an observation, and each column a feature. We assume, for mathematical simplicity, that there is no duplication of features in the two datasets. Within each matrix, the rows are independent. However, a pair of rows, one from each matrix, may have dependency.

For any positive integer k, let $\mathbb{N}_k = \{1,2,\ldots,k\}$, the set of positive integers or natural numbers up to and including k. For any finite set $\mathscr{S}$ (for example, $\mathbb{N}_k$), let $\sigma(\mathscr{S})$ be any permutation of the elements of $\mathscr{S}$.

Suppose that $n_0 \le n_1 (\le n_2)$ is the unknown number of *linked* observations between the datasets. Define $\tilde{n}_\ell = n_\ell - n_0$ for $\ell = 1,2$, denoting the number of unmatched observations from either dataset. Define $N = n_0 + \tilde{n}_1 + \tilde{n}_2$, this is the total number of observational units we consider, and we label the observational units with the index set $\mathbb{N}_N$. The i-th observation of the first dataset, $Y_{1,i,\square} \in \mathbb{R}^{p_1}$, corresponds to the unit whose index matches with the i-th element of $\sigma\big(\{1,\ldots,n_0+\tilde{n}_1\}\big)$, a random permutation of the index subset $\mathscr{S}_1 = \{1,\ldots,n_0+\tilde{n}_1\}$. The index subset of the second data is $\mathscr{S}_2 =$

$\{1,\ldots,n_0,n_0+\tilde{n}_1+1,\ldots,N\}$. The i-th observation of the second dataset, $Y_{2,i,\square}\in\mathbb{R}^{p_2}$, corresponds to the unit whose index matches with the i-th element of $\sigma\left(\mathscr{S}_2\right\}\big)$, a random permutation of the index subset $\mathscr{S}_2$. Note that $\mathscr{S}_1$ and $\mathscr{S}_2$ have exactly n_0 elements in common, reflecting the n_0 matched observational units for the two datasets.

In a few cases, the linkage between some observations units is known, and forms the training set. A training set is not needed for the present paper, but if indeed we have known and established linkages between, say $n\leq n_0$, observational units, without loss of generality we stack these known linkage cases as the first n rows of Y_ℓ, $\ell=1,2$.

We assume that the observations satisfy

$$Y_{\ell,i,\square}\overset{i.i.d.}{=}N_{p_\ell}\Big(0,\Sigma_\ell\Big),i=1,\ldots,n_\ell,\ \ \Sigma_\ell \text{ unknown},\ \ell=1,2.$$

We use the spectral representation

$$\begin{aligned}
&\Sigma_\ell=\Gamma_\ell\Lambda_\ell\Gamma_\ell^T,\text{ where}\\
&\Lambda_\ell=diag\Big(\lambda_{\ell,1},\ldots,\lambda_{\ell,p_\ell}\Big),\text{ with}\\
&\quad\lambda_{\ell,1}\geq\lambda_{\ell,2}\geq\ldots\geq\lambda_{\ell,p_0}\gg\lambda_{\ell,p_0+1}\geq\ldots\lambda_{\ell,p_\ell},\\
&\Gamma_\ell=\Big[\gamma_{\ell,\square,1}:\ldots:\gamma_{\ell,\square,p_\ell}\Big]\in\mathbb{R}^{p_\ell\times p_\ell}.
\end{aligned}$$

Thus, Λ_ℓ is the diagonal matrix of the eigenvalues of Σ_ℓ, and the columns of Γ_ℓ contain the corresponding eigenvectors. It is assumed that the first p_0 eigenvalues are considerably higher than the rest, and contain information relevant for linking records. Also for mathematical simplicity, we assume henceforth that $\lambda_{1,j}=\lambda_{2,j}$ for $j=1,\ldots,p_0$. That is, the top p_0 eigenvalues are the same. This is not a necessary assumption, but makes the presentation and technicalities of the developments presented below considerably simpler. We assume that the top p_0 eigenvectors of both $\Sigma_\ell,\ell=1,2$ are sparse, in that all but κ_ℓ of the entries in these eigenvectors are zero. This is a necessary assumption to obtain statistically consistent and computationally obtainable estimators of the principal components that we use in this paper, see WANG et al. (2016) for further details.

There are multiple record linkage contexts in which the above framework may be useful. First, traditional linkage techniques that rely on nominal and ordinal variables like names, addresses and so on often result in plausible subsets of observations from one dataset linked to each unit of the other dataset. At that stage, a further analysis based on the continuous variables as described here may be useful. Second, due to confidentiality and privacy considerations, datasets are often anonymized. In such cases, the model presented above may be extremely useful, either directly for modeling the reported continuous variables, or in conjunction with other continuous but latent variables. Third, our framework allows the scope of record linkage to extend beyond the traditional applications of linking sample surveys involving individuals or households, into linking data from multitude of sources, like social media, online shopping platforms, and electronic records of various kinds (LI et al., 2020; FATEMI et al., 2018). In many such contexts, the observed and often suitably anonymized data may be modeled us-

ing high-dimensional continuous (observed or latent) variables. With such an extended scope, record linkage may provide an increase in precision and accuracy of recommender systems (DRACHSLER et al., 2010; SHABTAI et al., 2013; SLOKOM, 2018), for providing online security and privacy (ZHU et al., 2016; SALAS, 2019), for transfer learning (RONG et al., 2012) and for distributed computing and related technical developments.

4. The statistical model

Without loss of generality and to considerably simplify the presentation below, we assume that the first n_0 rows of $Y_\ell, \ell = 1,2$ are linked. To relate the two datasets $Y_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$, we define the following quantities:

$$\begin{pmatrix} Z_{1,i,j} \\ Z_{2,i,j} \end{pmatrix} \overset{i.i.d.}{=} N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \text{ when } i = 1,\ldots,n_0 \text{ and } j = 1,\ldots,p_0,$$
$$Z_{\ell,i,j} \overset{i.i.d.}{=} N(0,1), \text{ for } \ell = 1,2, \text{ when } i > n_0 \text{ or } j > p_0.$$

Thus, $Z_{\ell,i,j}$ are all standard normal random variables, and for the case $i = 1,\ldots,n_0$ and $j = 1,\ldots,p_0$, the two random variables $Z_{1,i,j}$ and $Z_{2,i,j}$ share a correlation ρ between them. We arrange the $Z_{\ell,i,j}$ into two matrices of dimensions identical to those of our datasets Y_ℓ. Thus, $Z_{\ell,i,j}$ is the (i,j)-th element of the matrix $Z_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$. It can be seen that each matrix Z_ℓ has iid $N(0,1)$ entries, but the top left corners of Z_1 and Z_2 are related.

We model the data as

$$Y_{\ell,i,\square} = \Gamma_\ell \Lambda_\ell^{1/2} Z_{\ell,i,\square}, \text{ where}$$
$$Z_{\ell,i,\square} \overset{i.i.d.}{=} N_{p_\ell}\left(0, \mathbb{I}_{p_\ell}\right), i = 1,\ldots,n_\ell$$

described above. In matrix terms, this then translates to

$$Y_\ell = Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T, \text{ where}$$
$$Z_{\ell,i,j} \overset{i.i.d.}{=} N(0,1).$$

Note, however, that we do not imply with the above that the matrices $Z_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$ are independent of each other, and indeed they are not.

4.1. The record linking algorithm

Our proposed algorithm is as follows:

We now discuss in details the steps outlined above. First, the scaling $X_{\ell,i,\square} = Y_{\ell,i,\square}/|Y_{\ell,i,\square}|$ ensures that each $X_{\ell,i,\square}$ has unit norm, thus ensuring that outliers do not affect the PCA and subsequent computations. It is well known that PCA is very sensitive to outliers. The second step of the above algorithm is about computation of the high dimensional principal components using an established procedure. The third

Algorithm 1 Record Linking Algorithm

1. Scale each row of the two datasets by their respective norms, to get $X_{\ell,i,\square} = Y_{\ell,i,\square}/|Y_{\ell,i,\square}|$, for $\ell = 1,2$ and $i = 1,\ldots,n_l$. Collect these in the matrices $X_\ell \in \mathbb{R}^{n_\ell \times p_\ell}$, $\ell = 1,2$.

2. Run the high dimensional sparse PCA algorithm due to WANG et al. (2016) on $X_\ell, \ell = 1,2$. This obtains the leading eigenvalue and eigenvector for these matrices. Project the data on the orthogonal space to this estimated eigenvector, and repeat the process to obtain the leading p_0 eigenvectors.

3. Obtain the coefficients $W_{\ell,i,\square} \in \mathbb{R}^{p_0}$ (given in (4.1) below) for each $i = 1,\ldots,n_\ell$, $\ell = 1,2$ from the projections of the observations on the top p_0 eigenvectors.

4. Obtain the correlations $C(i,\tilde{i})$ of $W_{1,i,\square}$ and $W_{2,\tilde{i},\square}$.

5. Arrange the correlations in descending order. Based on the values corresponding to the training set and a pre-set value for the maximum proportion of false positive matches, select a correlation threshold. If there are multiple matches above this threshold for any $i \in \{1,\ldots,n_1\}$ or $\tilde{i} \in \{1,\ldots,n_2\}$, the match with the higher correlation value is chosen. The proportion of false negatives is estimated from the number of training sample matches below the threshold.

through last steps of the algorithm are about using the principal component scores to obtain the record linkage. There can be considerable variation in the details in these steps depending on the context, we have presented one simple procedure.

The above algorithm used the training data only for the last step of setting a threshold for the correlations. We can easily formulate a variation, where a training set is not needed. Since we compute $n_1 n_2$ correlations of which only $\min(n_1,n_2)$ can possibly correspond to linked data, a threshold can be determined based a change-point in the correlation values, or on multiple matches. We illustrate this aspect in simulations reported later in this paper.

4.2. Theoretical properties

The justification for using the above algorithm rests on the fact that there is a clear separation of the correlation values between the linked data-pairs, as opposed to the correlation between the not-linked cases. We establish this fact in the following result:

Theorem 4.1. *Under the conditions of the model, the population correlation value for each linked pair of observations is ρ, and is zero for two observations that are not linked.*

Thus, there is a clear separation of the correlation values from the linked data-pairs from the rest. There would be sample variations, and the value of ρ is not known. Consequently, either a training set-based threshold or a change detection technique can be used to sort the true linkages from the rest.

Proof of Theorem 4.1: Let

$$\tilde{\Gamma}_\ell = \left[\gamma_{\ell,\square,1} : \ldots : \gamma_{\ell,\square,p_0}\right] \in \mathbb{R}^{p_\ell \times p_0},$$

the first p_0 columns of Γ_ℓ, for which the eigenvalues are considerably higher than the rest. We project the datapoints $Y_{\ell,i,\square}$ on the column space of $\tilde{\Gamma}_\ell$, for $i = 1, \ldots, n_\ell$. Note that all columns of $\tilde{\Gamma}_\ell$ are orthonormal (by construction, since these are estimators of successive eigenvectors, so they are orthogonal to each other and have unit norm). Given this, it is easy to see that the projection of $Y_{\ell,i,\square}$ is

$$\begin{aligned}\tilde{Y}_{\ell,i,\square} &= \sum_{j=1}^{p_0} \left\langle Y_{\ell,i,\square}, \gamma_{\ell,\square,j} \right\rangle \gamma_{\ell,\square,j} \\ &= \tilde{\Gamma}_\ell \tilde{\Gamma}_\ell^T Y_{\ell,i,\square} \\ &= \tilde{\Gamma}_\ell W_{\ell,i,\square}.\end{aligned}$$

The relevant information about the projections is carried in the low-dimensional ***weights*** $W_{\ell,i,\square} \in \mathbb{R}^{p_0}$, consequently we develop our analysis based on these below. Putting these weights as rows in a matrix, we have

$$\begin{aligned}W_\ell &= Y_\ell \tilde{\Gamma}_\ell \\ &= Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T \tilde{\Gamma}_\ell \in \mathbb{R}^{n_\ell \times p_0}\end{aligned} \tag{4.1}$$

This last expression can be simplified further, since

$$\begin{aligned}\Gamma_\ell^T \tilde{\Gamma}_\ell &= \begin{pmatrix} \gamma_{\ell,\square,1}^T \\ \gamma_{\ell,\square,2}^T \\ . \\ . \\ . \\ \gamma_{\ell,\square,p_\ell}^T \end{pmatrix} \left[\gamma_{\ell,\square,1} : \ldots : \gamma_{\ell,\square,p_0}\right] \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ . & & & \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ . & & & \\ . & & & \\ 0 & 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{I}_{p_0} \\ \mathbf{0}_{(p_\ell - p_0) \times p_0} \end{pmatrix}.\end{aligned}$$

Thus we have

$$\begin{aligned}
&\Lambda_\ell^{1/2}\Gamma_\ell^T\tilde{\Gamma}_\ell \\
&\qquad = \begin{pmatrix}
\lambda_{\ell,1}^{1/2} & 0 & 0 & 0 \\
0 & \lambda_{\ell,2}^{1/2} & 0 & 0 \\
. & & & \\
0 & 0 & 0 & \lambda_{\ell,p_0}^{1/2} \\
0 & 0 & 0 & 0 \\
. & & & \\
. & & & \\
0 & 0 & 0 & 0
\end{pmatrix} \\
&\qquad = \begin{pmatrix} \tilde{\Lambda}_\ell^{1/2} \\ \mathbf{0}_{(p_\ell - p_0)\times p_0} \end{pmatrix} \in \mathbb{R}^{p_\ell \times p_0}.
\end{aligned}$$

Define

$$\tilde{Z}_\ell = \left[Z_{\ell,\square,1} : Z_{\ell,\square,2} : \ldots : Z_{\ell,\square,p_0} \right] \in \mathbb{R}^{n_\ell \times p_0}.$$

Consequently, we have

$$\begin{aligned}
W_\ell &= Y_\ell \tilde{\Gamma}_\ell \\
&= Z_\ell \Lambda_\ell^{1/2} \Gamma_\ell^T \tilde{\Gamma}_\ell \\
&= \left[\lambda_{\ell,1}^{1/2} Z_{\ell,\square,1} : \lambda_{\ell,2}^{1/2} Z_{\ell,\square,2} : \ldots : \lambda_{\ell,p_0}^{1/2} Z_{\ell,\square,p_0} \right] \\
&= \tilde{Z}_\ell \tilde{\Lambda}_\ell^{1/2} \in \mathbb{R}^{n_\ell \times p_0}.
\end{aligned}$$

We thus have, for any $i \in \{1, \ldots, n_0\}$

$$\begin{aligned}
\mathbb{E}W_{\ell,i,\square} &= 0, \\
\mathbb{V}W_{\ell,i,\square} &= \tilde{\Lambda}_\ell, \\
\mathbb{E}W_{1,i,\square}W_{2,i,\square}^T &= \tilde{\Lambda}_1 \mathbb{E}\tilde{Z}_{1,i,\square}\tilde{Z}_{2,i,\square}^T \tilde{\Lambda}_2, \\
\mathbb{E}W_{1,i,\square}^T W_{2,i,\square} &= \mathbb{E}\tilde{Z}_{1,i,\square}^T \tilde{\Lambda}_1 \tilde{\Lambda}_2 \tilde{Z}_{2,i,\square}^T = \sum_{j=1}^{p_0} \lambda_{1,j}^{1/2} \lambda_{1,j}^{1/2} \mathbb{E}\tilde{Z}_{1,i,j}\tilde{Z}_{2,i,j}.
\end{aligned}$$

Then it follows that

$$Cor\left(W_{1,i,\square}, W_{2,i,\square}\right) = \rho.$$

The algebra for the observations that are not linked is similar and omitted here. □

One important consideration for our framework is to ensure that the robustness procedure we implemented in the first step does not alter the eigenvector structure of the original data. That is, we need to ensure that the eigenvectors of Σ_ℓ match those of

the variance of X_ℓ, $\ell = 1,2$. This is ensured in the following result:

Theorem 4.2. *Suppose $X \in \mathbb{R}^p$ is a random vector with variance Σ_X, and let the variance of $U = X/|X|$ be denoted by Σ_U.*

(i) When X has an elliptically symmetric distribution and zero mean, the eigenvectors of Σ_X and Σ_U are identical.

(ii) If $\mathbb{E}U = 0 \in \mathbb{R}^p$, $\mathbb{E}U|X| = 0 \in \mathbb{R}^p$ and $\mathbb{E}|X|^2UU^T = \mathbb{E}|X|^2\mathbb{E}UU^T \in \mathbb{R}^{p\times p}$, then again the eigenvectors of Σ_X and Σ_U are identical. Moreover, if the eigenvalues of Σ_X are $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$, then the eigenvectors of Σ_U are $\frac{\lambda_1}{\sum_{i=1}^p \lambda_i} \geq \ldots \frac{\lambda_p}{\sum_{i=1}^p \lambda_i} > 0$.

Note that a p-dimensional random vector X is said to elliptically distributed if there exist a vector $\mu \in \mathbb{R}^p$, a positive semi-definite matrix $\Sigma \in \mathbb{R}^{p\times p}$ and a function $\phi : (0,\infty) \to \mathbb{R}$ such that the characteristic function of X is $\exp\{it^T\mu\}\phi(t^T\Sigma t)$ for $t \in \mathbb{R}^p$. See FANG et al. (1990) for several alternative and equivalent definitions of the elliptically contoured family, as well as for additional details. An example of elliptically contoured distribution is the multivariate Gaussian distribution, thus the framework adopted in this paper satisfies the first condition of Theorem 4.2. The second part of Theorem 4.2 is for general interest, in case an elliptic distributional assumption is not satisfied.

Proof of Theorem 4.2: First, consider the case where X has an elliptically symmetric distribution with mean zero. In such cases, we may write $X = R\Gamma\Lambda^{1/2}E$, where Γ is a rotation matrix, Λ is a diagonal matrix with positive elements, E is uniformly distributed on the unit sphere and R is a positive random variable that is independent of E. Then, we have $|X|^2 = R^2E^T\Lambda E$. Let $\tilde{E} = \frac{\Lambda^{1/2}E}{|\Lambda^{1/2}E|}$. Consequently, $U = X/|X| = \Gamma\tilde{E}$ is a function of E alone. Note that in the circularly symmetric case where $\Lambda = \lambda\mathbb{I}$, we now have $|X|$ independent of U, and the above conditions are trivially satisfied. For general Λ, note that

$$\mathbb{E}UU^T = \Gamma\mathbb{E}\tilde{E}\tilde{E}^T\Gamma^T,$$

and it can be easily shown that $\mathbb{E}\tilde{E}\tilde{E}^T$ is a diagonal matrix. Thus, $\Sigma_X = \Gamma\Lambda\Gamma^T$ and Σ_U have the same eigenvectors in this case. This proof is reminiscent of the arguments used in TASKINEN et al. (2012).

Under the assumptions of the second part, that is, $\mathbb{E}U = 0 \in \mathbb{R}^p$, $\mathbb{E}U|X| = 0 \in \mathbb{R}^p$ and $\mathbb{E}|X|^2UU^T = \mathbb{E}|X|^2\mathbb{E}UU^T \in \mathbb{R}^{p\times p}$, U and $|X|$ are uncorrelated, as is $|X|^2$ and UU^T. This immediately implies that $\mathbb{E}X = \mathbb{E}U|X| = 0$, thus we have $\Sigma_X = \mathbb{E}XX^T$ and $\Sigma_U = UU^T$. We also easily have

$$\begin{aligned}\Sigma_X &= \mathbb{E}XX^T \\ &= \mathbb{E}|X|^2UU^T \\ &= \mathbb{E}|X|^2\mathbb{E}UU^T.\end{aligned}$$

Thus, the eigenvectors of Σ_X and Σ_U are identical, since $\mathbb{E}|X|^2$ is a scalar.

Now, note that

$$\Sigma_U = \frac{\Sigma_X}{\mathbb{E}|X|^2} = \frac{\Sigma_X}{\mathbb{E}[\text{Trace } XX^T]} = \frac{\Sigma_X}{\text{Trace } \Sigma_X} = \frac{\Sigma_X}{\sum_{i=1}^p \lambda_i}.$$

□

The rest of this section is on the estimation of the high dimensional principal components. We present the results only for the first principal component. Our development here closely follows that of WANG et al. (2016), and we essentially make use of their theoretical machinery and algorithm for the rest of this paper. The results below are primarily designed to show that the technical conditions of WANG et al. (2016) hold for our case, and the eigenvector estimation algorithm they established also works for us. We omit many algebraic details, since they are similar to those of WANG et al. (2016).

Our first result is to show that U has a sub-Gaussian distribution. This is immediate, since U is bounded. We have multiple proofs of this result with sharp bounds on the constant σ^2, but present the simplest one here for clarity. Ensuring that U has a sub-Gaussian distribution facilitates the use of various known concentration inequality and other probabilistic results.

Lemma 4.1. $U \in$ *Sub-Gaussian*(2).

Proof of Lemma 4.1: We recall the definition of Sub-Gaussian distributions

$$X \in \text{Sub-Gaussian}(\sigma^2) \text{ if } \forall u \in \mathbb{R}^p,\ \mathbb{E}[e^{u^T X}] \leq e^{\frac{\sigma^2|u|^2}{2}}$$

Since $|U| = 1$, we have for $|u| \geq 1$

$$\begin{aligned} u^T U \leq |u|^2 (\text{C-S inequality}) \text{ for any } u \text{ on } \mathbb{R}^p \\ \Rightarrow e^{u^T U} \leq e^{|u|^2} \text{ for any } u \text{ on } \mathbb{R}^p \\ \Rightarrow U \in \text{Sub-Gaussian}(2). \end{aligned}$$

The case for $|u| < 1$ is more delicate, but can be handled with some routine algebra. We omit the details here. □

We now recall some definitions from WANG et al. (2016) for developing our next set of results. Since our framework is high dimensional, we need structural assumptions on the nature of the eigenvectors of Σ_U (or Σ_X), and the most common and convenient assumption here is one of sparsity. We define the sparse unit ball in p-dimensions having at most k non-zero entries as follows:

$$B_0(k) = \left\{ x \in \mathbb{R}^p : |x| = 1, \sum_{j=1}^p \mathscr{I}_{\{x_i \neq 0\} \leq k} \right\}.$$

Based on this and sample size n, for any $j \in \{1, \ldots, p\}$ and $C > 0$, a probability measure P is said the satisfy the *Restricted Covariance Condition* (RCC) with parameters p, n, j

and C, and written as $P \in RCC_p(n, j, C)$ if

$$P\{\sup_{u \in B_0(l)} |\hat{V}(u) - V(u)| \geq C \max(\sqrt{\frac{j \log(p/\delta)}{n}}, \frac{j \log(p/\delta)}{n})\} \leq \delta$$

for all $\delta > 0$, where $V(u) = \mathbb{E}u^T \Sigma u$, $\hat{V}(u) = \mathbb{E}u^T \hat{\Sigma} u$ and $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n U_i U_i^T$ for $U_1, U_2, \ldots, U_n \overset{iid}{\sim} P$. We also define

$$RCC_p(C) = \bigcap_{l=1}^{p} \bigcap_{n=1}^{\infty} RCC_p(n, l, C).$$

Suppose, associated with a generic distribution $\mathbb{P}$ on $\mathbb{R}^p$, is the variance matrix Σ with the j-th eigenvalue and eigenvector respectively being λ_j and $\gamma_{\square, j}$, $j = 1, \ldots, p$. The results for the rest of this section are valid for the following class of probability measures. For $\theta > 0$, define

$$\mathscr{P}_p(n, k, \theta) = \Big\{\mathbb{P} \in RCC_p(n, 2, 1) \cap RCC_p(n, 2k, 1) : \gamma_{\square,1} \in B_0(k), \lambda_1 - \lambda_2 \geq \theta\Big\}.$$

Our next result states that U, after suitable scaling, has a distribution that satisfies the restricted covariance condition with appropriate selection of constants.

Lemma 4.2. *For the random variable $U \in \mathbb{R}^p$ with $p \geq 2$, assume that $\gamma_{\square,1} \in B_0(k)$ and $\theta = \big(\lambda_1(\Sigma_X) - \lambda_2(\Sigma_X)\big) > 0$. Then $Z = \frac{U}{22} \in \mathscr{P}_p(n, k, \frac{\theta}{22\, Trace(\Sigma_X)})$.*

Proof of Lemma 4.2: Recall from Proposition 1 of WANG et al. (2016) that for every $\sigma > 0$,

$$\text{Sub-Gaussian}(\sigma^2) \subseteq RCC_p\big(16\sigma^2(1 + \frac{9}{\log(p)})\big).$$

Therefore using Lemma 4.1, we have that

$$\begin{aligned}
& U = \frac{X}{|X|} \in RCC_p\big(32\,(1 + \frac{9}{\log(p)})\big), \\
\Rightarrow & \frac{U}{\sqrt{32\,(1 + \frac{9}{\log(p)})}} \in RCC_p(1), \\
\Rightarrow & \frac{U}{22} \in RCC_p(1), \\
\Rightarrow & \frac{U}{22} \in \mathscr{P}_p(n, k, \frac{\theta}{22\text{Trace}(\Sigma_X)}).
\end{aligned}$$

□

For a symmetric matrix $A \in \mathbb{R}^{p \times p}$, let us define

$$\hat{\gamma}^k_{\square, max}(A) = \text{sargmax}_{u \in B_0(k)} u^T A u$$

to be the k-sparse maximum eigenvector of A, where *sargmax* denotes the smallest element of the argmax in the lexicographic ordering. We use A as an argument of

$\hat{\gamma}^k_{\square,max}(\cdot)$ here to distinguish between the estimated eigenvectors from various matrices. Also, between 2 unit vectors u and v , we define the loss function

$$L(u,v) = \left(1-(u^T v)^2\right)^{1/2}.$$

Note that $\hat{\gamma}^k_{\square,max}(A)$ are all identical for $A = \hat{\Sigma}_U, \hat{\Sigma}_X, \hat{\Sigma}_Z$. Our next result is the main consistency and probabilistic guarantee result on the sample version of the k-sparse maximum eigen-vector. This result ensures in particular that under suitable conditions with $\log(p) = o(n)$, the sample k-sparse maximum eigenvector is consistent for the population maximum eigenvector.

Theorem 4.3. *For $2k\log(p) \le n$, the k-sparse empirical maximum eigen-vector, $\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U)$ satisfies*

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U), \gamma_{\square,1}(\Sigma_U)\right) \le 44\sqrt{2}\left(1+\frac{1}{\log(p)}\right)\sqrt{\frac{k\log(p)}{n\theta^2}}\ \mathit{Trace}(\Sigma_X).$$

Proof of Theorem 4.3: We apply Theorem 2 of WANG et al. (2016) on Z and note that for $2k\log(p) \le n$, the k-sparse empirical maximum eigen-vector, $\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z)$ satisfies

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z), \gamma_{\square,1}(\Sigma_Z)\right) \le 2\sqrt{44}\left(1+\frac{1}{\log(p)}\right)\sqrt{\frac{k\log(p)}{n\theta^2}}\ \mathrm{Trace}(\Sigma_X).$$

Proof follows by noting that

$$\mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_Z), \gamma_{\square,1}(\Sigma_Z)\right) = \mathbb{E}L\left(\hat{\gamma}^k_{\square,max}(\hat{\Sigma}_U), \gamma_{\square,1}(\Sigma_U)\right).$$

□

Let $U_i = \frac{X_i}{|X_i|}$ for $i = 1,2,\ldots,n$ where $X_i \overset{iid}{\sim} \mathbb{P}$, and we denote $\hat{\gamma}^{SDP}_{\square,1}(U)$ for the output of the SDP algorithm of WANG et al. (2016). with input $U = (U_1,\ldots,U_n)^T$, $\lambda = 4\sqrt{\frac{\log(p)}{n}}$ and $\varepsilon = \frac{\log(p)}{4n}$.

While Theorem 4.3 provided a general probabilistic guarantee on the error of the sample k-sparse maximum eigenvector, we need a similar result for the sparse maximum eigenvector that is obtained using the SDP algorithm. Note that the SDP algorithm allows for computation of a sparse maximum eigenvector in real time, and is thus both practical and is of theoretical relevance. The following result establishes a probabilistic guarantee and consistency for this version of the sparse empirical maximum eigenvector.

Theorem 4.4. *If $4\log(p) \le n \le k^2p^2\theta^{-2}\log(p)$, then*

$$\mathbb{E}L\left(\hat{\gamma}^{SDP}_{\square,1}(U), \gamma_{\square,1}(\Sigma_U)\right) \le (352\sqrt{2}+44)\ \mathit{Trace}\Sigma_X\sqrt{\frac{k^2\log(p)}{n\theta^2}}.$$

Proof of Theorem 4.4: We apply Theorem 5 of WANG et al. (2016) on Z and note

that

$$\mathbb{E}L\big(\hat{\gamma}_{\square,1}^{SDP}(Z), \gamma_{\square,1}(\Sigma_Z)\big) \leq 22(16\sqrt{2}+2)\sqrt{\frac{k^2\log(p)}{n\theta^2}}\,\mathrm{Trace}(\Sigma_X).$$

for $Z = (Z_1, \ldots, Z_n)^T$ where $Z_i = \frac{U_i}{22}$. The result is immediate.

□

In general, $\hat{\gamma}_{\square,1}^{SDP}(U)$ is not a sparse estimator. However, it turns out that a k-sparse version of $\hat{\gamma}_{\square,1}^{SDP}(U)$, that is, some $\hat{\gamma}_{\square,1}^{SDP,k}(U) \in B_0(k)$, may be obtained by setting all but the top k coordinates of $\hat{\gamma}_{\square,1}^{SDP}(U)$ in absolute value to zero and renormalizing the vector. In particular, $\hat{v}_0^{SDP}$ is computable in polynomial time and under the same condition as in Theorem 4.4.

5. Some Simulation Results

In this section, we present a simulation exercise to illustrate the performance of the proposed record linkage methodology, and also to illustrate some practical implementation steps.

To generate data, we followed the framework laid out at the start of Section 4. That is, we generated a set of independent bivariate Gaussian random variables with common correlation ρ, and several independent univariate standard Normal random variables, and used these to populate the two data matrices. We tested various choices of sample sizes, dimensions, correlation ρ. For brevity, we report the case where the two matrix datasets that we use are of dimensions $n_1 = 60, p_1 = 100$, and $n_2 = 70, p_2 = 120$ of independent rows each, and $\rho = 0.8$. The first $n_0 = 50$ entries of these two matrices are linked to each other. The rest (10 for the first matrix, 20 for the second matrix) are not linked. We use the first $n = 20$ observations for training in the version of the algorithm where a training set is used, thus leaving the last 30 linked data points for testing. We also demonstrate the performance of our method when no training dataset is available. We fix $p_0 = 10$ for this exercise, and repeated the entire simulation 100 times.

As practical steps, we found that using a sparse version of the estimated eigenvalues, as proposed in Theorem 4.4, considerably improves performance, owing to reduction of the effect of noise terms in the eventual linkage. Also, the estimated principal components for the two datasets may not have the same orientation and may not appear in the same order. Hence, when needed, a principled permutation and sign reversal of the estimated eigenvectors of the second dataset is done to improve linkage accuracy. While in theory the estimators of the eigenvalues are not required for the linkage steps, using those as weights improves linkage.

Over 100 replications of the simulation experiment, the correct linkage established on the test set by our proposed method was about 43.5% times, with a standard error of about 5.37%. When no training sample is used and instead a threshold for the correlation estimated from the data, the correct linkage percentage increases to about 56%, however, the standard error also increases to about 14.8%. The estimated threshold for correlation was at a lower value than the case with training data: a pattern that we noticed in

multiple simulations, which we will study further later. The linkage accuracy may seem low, however, we need to remember that this is a ***unsupervised*** framework, involving high dimensional datasets with no common variables, and minimal or no training data.

6. Conclusions and Future Work

Record linkage is in-general a difficult exercise. Bayesian models are often too complex for practical purposes and some Bayesian formulations fail to accommodate non-categorical variables. Supervised classification methods assume the existence of large, accurate sets of training data, which are often difficult and/or expensive to obtain. Also, in those methodologies, it is difficult to guarantee maximum one-to-one assignment constrain of bipartite record linkage. While some theoretical modifications have been developed to ensure an one-to-one assignment(SADINLE, 2017), typically some subsequent post-processing step is required to solve these inconsistencies.

In view of these difficulties, in this paper we propose a completely new approach towards record linkage. We do not require a common set of variables between the two datasets, we do not require a training set, and the dimensionality and sample sizes can both be large. Naturally, our methodology extends to cases where a common set of variables exist, we will elaborate on this in a future work. If a training set exists, we can make use of it, as illustrated in this paper. We have presented the case for the bipartite record linkage, but our model conceptually extends to other cases as well.

Some of the technical assumptions of this paper, like the two covariances matrices having the same set of leading eigenvalues, or the number leading eigenvalues p_0 being known, or the latent random variables that link the two datasets having the same correlation ρ, can be addressed with some additional work and methodological developments. The assumption of multivariate normality of the data is not critical: our proposal only depends on robust, high dimensional principal components, and these are available for data from many distributions with both discrete and continuous components. The assumption of sparsity in the leading eigenvectors is owing to the fact that for high dimensional modeling, some structural assumptions are needed since the sample size is not adequate to estimate all relevant unknown parameters. In any case, there are considerable challenges to estimating high dimensional principal components, see PAUL (2007).

The robust, high dimensional principal component we use is built on the work by WANG et al. (2016). The credit for both the theoretical framework and the algorithm goes to that work primarily. Our setup differs from WANG et al. (2016) in the detail that for robustness purposes we transform each observation to be on the unit sphere. One future work for us is to establish the theoretical results under weaker assumptions than WANG et al. (2016), or to show better theoretical properties.

We have used a simple method for linking observations in this paper, using correlations. A correlation-based linkage is not critical to our primary methodological steps. More complex and realistic measures of linkage will be studied in the future. The case where no training data is present needs further investigation, which will also be part of our future work. In absolute terms, our simulation results are not excellent; however,

we do not have a baseline for comparison since most other papers on record linkage do not use as general a framework as ours with (*i*) high dimensional data, (*ii*) no common set of features, and (*iii*) possibly no training set. Our framework may be termed *unsupervised learning of record linkage*, and in the unsupervised learning framework, our numeric results are perhaps acceptable. However, considerable fine tuning and experimentation with the algorithm for record linkage is needed. We have ensured that our high-dimensional, robust and potentially sparse principal component estimator is highly accurate, and some of our studies (not reported here) suggest that using a small number of common features dramatically increases linkage accuracy. A part of our future work is on including nominal and categorical variables for linkages in our framework, which will make our proposed approach more aligned with traditional record linkage techniques. In this context, we will also investigate how much additional gain results from using PCA in addition to available matching fields, compared to the traditional Felligi-Sunter method.

An important topic to consider in future from this paper is on statistical inference based on linked datasets. This is a non-trivial task, since the datasets are used multiple times in the process of linking, estimation of various quantities of interest, and then inference. The article HAN and LAHIRI (2019) provides review of the current state of the art in this direction of work. Some alternatives to fully Bayesian methods, for example regression analysis using linked data LAHIRI and LARSEN (2005); SCHEUREN and WINKLER (1997, 1993), have both computational efficiency and analytical tractability, which may make them attractive practical choices for applications. Comparisons with such alternatives is an additional future work.

An additional future work for us is to extend the methodology proposal here to multiple datasets. We will also work on real data examples, which has not been possible for this paper owing to data access limitations. It will be of interest to compare our unsupervised record linkage approach with more traditional record linking algorithms.

Acknowledgements

This research is partially supported by the US National Science Foundation (NSF) under grants # DMS-1622483, # DMS-1737918, # OAC-1939916 and #DMR-1939956.

References

CHRISTEN, P., (2011). A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9), pp. 1537–1555.

DRACHSLER, H., BOGERS, T., VUORIKARI, R., VERBERT, K., DUVAL, E., MANOUSELIS, N., BEHAM, G., LINDSTAEDT, S., STERN, H., FRIEDRICH, M., et al., (2010). Issues and considerations regarding sharable data sets for recommender systems in technology enhanced learning. *Procedia Computer Science*, 1(2), pp. 2849–2858.

ENAMORADO, T., FIFIELD, B., and IMAI, K., (2018). Using a probabilistic model to assist merging of large-scale administrative records. *Available at SSRN 3214172*.

FANG, K.-T., KOTZ, S., and NG, K.-W., (1990). *Symmetric Multivariate and Related Distributions*. CRC Press.

FATEMI, B., KAZEMI, S. M., and POOLE, D., (2018). Record linkage to match customer names: A probabilistic approach. *arXiv preprint arXiv:1806.10928*.

FELLEGI, I. P. and SUNTER, A. B., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

HAN, Y. and LAHIRI, P., (2019). Statistical analysis with linked data. *International Statistical Review*, 87, pp. S139–S157.

HERZOG, T. N., SCHEUREN, F. J., and WINKLER, W. E., (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

JARO, M. A., (1989). Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406), pp. 414–420.

LAHIRI, P. and LARSEN, M. D., (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469), pp. 222–230.

LARSEN, M. D. and RUBIN, D. B., (2001). Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453), pp. 32–41.

LI, J., DOU, Z., ZHU, Y., ZUO, X., and WEN, J.-R., (2020). Deep cross-platform product matching in e-commerce. *Information Retrieval Journal*, 23(2), pp. 136–158.

LISEO, B. and TANCREDI, A., (2013). Some advances on Bayesian record linkage and inference for linked data. *URL http://www. ine. es/e/essnetdi_ws2011/ppts/Liseo_Tancredi. pdf*.

MARTINS, B., (2011). A supervised machine learning approach for duplicate detection over gazetteer records. In *International Conference on GeoSpatial Sematics*, pp. 34–51, Springer.

NEWCOMBE, H. B. and KENNEDY, J. M., (1962). Record linkage: making maximum use of the discriminating power of identifying information. *Communications of the ACM*, 5(11), pp. 563–566.

PAUL, D., (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4), pp. 1617–1642.

RONG, S., NIU, X., XIANG, E. W., WANG, H., YANG, Q., and YU, Y., (2012). A machine learning approach for instance matching based on similarity metrics. In *International Semantic Web Conference*, pp. 460–475, Springer.

SADINLE, M., (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112(518), pp. 600–612.

SADINLE, M. and FIENBERG, S. E., (2013). A generalized fellegi–sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108(502), pp. 385–397.

SALAS, J., (2019). Sanitizing and measuring privacy of large sparse datasets for recommender systems. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12.

SCHEUREN, F. and WINKLER, W. E., (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, pp. 39–58.

SCHEUREN, F. and WINKLER, W. E., (1997). Regression analysis of data files that are computer matched-ii. *Survey Methodology*, 23, pp. 157–165.

SHABTAI, A., ROKACH, L., and ELOVICI, Y., (2013). Occt: A one-class clustering tree for implementing one-to-many data linkage. *IEEE Transactions on Knowledge and Data Engineering*, 26(3), pp. 682–697.

SINGLA, P. and DOMINGOS, P., (2006). Entity resolution with markov logic. In *Sixth International Conference on Data Mining (ICDM'06)*, pp. 572–582, IEEE.

SLOKOM, M., (2018). Comparing recommender systems using synthetic data. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pp. 548–552.

STEORTS, R. C. et al., (2015). Entity resolution with empirically motivated priors. *Bayesian Analysis*, 10(4), pp. 849–875.

STEORTS, R. C., HALL, R., and FIENBERG, S. E., (2016). A bayesian approach to graphical record linkage and deduplication. *Journal of the American Statistical Association*, 111(516), pp. 1660–1672.

STEORTS, R. C., VENTURA, S. L., SADINLE, M., and FIENBERG, S. E., (2014). A comparison of blocking methods for record linkage. In *International Conference on Privacy in Statistical Databases*, pp. 253–268, Springer.

TASKINEN, S., KOCH, I., and OJA, H., (2012). Robustifying principal component analysis with spatial sign vectors. *Statistics & Probability Letters*, 82(4), pp. 765–774.

TORVIK, V. I. and SMALHEISER, N. R., (2009). Author name disambiguation in medline. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(3), pp. 1–29.

TREERATPITUK, P. and GILES, C. L., (2009). Disambiguating authors in academic publications using random forests. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*, pp. 39–48.

VENTURA, S. L., NUGENT, R., and FUCHS, E. R., (2015). Seeing the non-stars:(some) sources of bias in past disambiguation approaches and a new public tool leveraging labeled records. *Research Policy*, 44(9), pp. 1672–1701.

WANG, T., BERTHET, Q., and SAMWORTH, R. J., (2016). Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics*, 44(5), pp. 1896–1930.

ZHU, J., ZHANG, S., SINGH, L., YANG, G. H., and SHERR, M., (2016). Generating risk reduction recommendations to decrease vulnerability of public online profiles. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 411–416, IEEE.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 144–158, DOI 10.21307/stattrans-2020-035
Received – 31.01.2020; accepted – 30.06.2020

Confidence bands for a distribution function with merged data from multiple sources

Takumi Saegusa[1]

ABSTRACT

We consider nonparametric estimation of a distribution function when data are collected from multiple overlapping data sources. Main statistical challenges include (1) heterogeneity of data sets, (2) unidentified duplicated records across data sets, and (3) dependence due to sampling without replacement from a data source. The proposed estimator is computable without identifying duplication but corrects bias from duplicated records. We show the uniform consistency of the proposed estimator over the real line and its weak convergence to a Gaussian process. Based on these asymptotic properties, we propose a simulation-based confidence band that enjoys asymptotically correct coverage probability. The finite sample performance is evaluated through a simulation study. A Wilms tumor example is provided.

Key words: confidence band, data integration, Gaussian process.

1. Introduction

We consider nonparametric estimation of a distribution function F of a random variable X when data are collected from multiple overlapping data sources. Inference on F is a rather simple problem if data are independent and identically distributed (i.i.d.). When data sets are merged from various sources, this basic question faces a significant challenge from both theoretical and methodological perspectives. Statistical issues we address in this paper is (1) heterogeneity of data sources, (2) unidentified duplicated records in multiple data sets, and (3) finite population sampling from each data source. Without proper care, these issues yield bias in estimation and wrong quantification of uncertainty.

The following setting (schematically shown in Figure 1) is considered:

• The variables of interest for data integration is a random vector $W = (X, Y)$ taking values in a measurable space $(\mathscr{W}, \mathscr{A})$. In this paper, we focus on inference regarding X but inference on X and Y is of general interest in data integration.

• Let $V = (\tilde{W}, Z) \in \mathscr{V}$ where $\tilde{W}$ is a coarsening of W and Z is a vector of auxiliary variables. The variables Z do not involve inference on W but help to create data sources. The space $\mathscr{V}$ is composed of J overlapping population data sources $\mathscr{V}^{(1)}, \ldots, \mathscr{V}^{(J)}$ with $\mathscr{V} = \cup_j \mathscr{V}^{(j)}$ and $\mathscr{V}^{(j)} \cap \mathscr{V}^{(k)} \neq \emptyset$ for some (j,k). Values of V determine membership of data sources.

• Data collection is carried out in a two-stage framework. First, a large i.i.d. sample of $V_1, \ldots, V_N$ is collected from a population. The unit i is distributed to data source j if

[1]University of Maryland. USA. E-mail: tsaegusa@umd.edu. ORCID: https://orcid.org/0000-0001-6869-2451.

$V_i \in \mathscr{V}^{(j)}$. Because data sources overlap, the unit i may belong to multiple sources. The sample size of data source $\mathscr{V}^{(j)}$ is denoted as $N^{(j)}$.

• Next, a random sample of size $n^{(j)}$ is selected without replacement from data source $\mathscr{V}^{(j)}$. The selection probability for this data source is $\pi^{(j)}(V_i) = (n^{(j)}/N^{(j)})I\{V_i \in \mathscr{V}^{(j)}\}$ where I is the indicator function. For selected items, variables W_i are observed.

• The above procedure is repeated for all data sources. Data sets from each data source are then combined and statistical analysis is conducted. If the unit i is selected multiple times, its duplication is not identified.

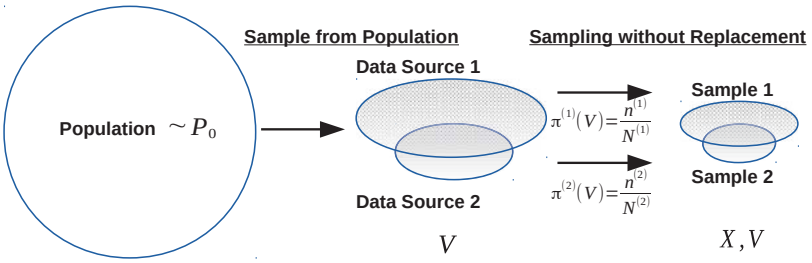


Figure 1: Sampling scheme for merged data from multiple sources with $J = 2$.

This two-stage formulation is essential in describing duplicated records in multiple data sets. Duplication naturally occurs in public health data integration. Clinical studies have their own target populations defined by the inclusion and exclusion criteria. When these studies are combined with national disease registries, a patient in a study is also in a national database. Duplicated records are difficult to identify in practice because key identifiers such as names and addresses are often not disclosed for privacy protection in public health data. Instead, the membership of selected items in the final sample is assumed known (e.g., the selected item i from source $\mathscr{V}^{(j)}$ is also known to belong to $\mathscr{V}^{(k)}$). This is plausible because one can compare inclusion and exclusion criteria. For more detailed discussion on practical issues of our setting, see SAEGUSA (2019).

The final sample is a biased and dependent sample with duplication. There are two sources of bias in our setting. Certain data sources are over/under-represented in the final sample due to biased sampling with different selection probabilities $\pi^{(j)}$. Duplicated records from overlapping data sources enter statistical analysis without identification. Dependence also comes from two sources. Multiple data sets are dependent through duplicated records while items in the same data source are dependent due to sampling without replacement. These characteristics well capture the challenging issue of heterogeneity in data integration problems. Our framework covers the number of examples including opinion polls (BRICK et al., 2006), public health surveillance (HU et al., 2011), and health interview surveys (CERVANTES et al., 2006), and the synthesis of existing clinical and epidemiological studies with surveys, disease registries, and healthcare databases (CHATTERJEE et al., 2016; KEIDING and LOUIS, 2016; METCALF and SCOTT, 2009).

In this paper, we propose and study a nonparametric estimator of the distribution

function F. Our estimator is motivated by Hartley's estimator for multiple-frame surveys in sampling theory (HARTLEY, 1962, 1974). We provide a rigorous asymptotic theory to its uniform consistency over the real line and weak convergence to a Gaussian process. Based on the limiting distribution, we propose a Monte Carlo based method to construct confidence bands for F. We verify the validity of our methodology theoretically and through a simulation study for both continuous and discrete random variables.

Recently SAEGUSA (2019) studied the same data integration setting and derived the law of large numbers and the central limit theorem. Asymptotic results are then applied to infinite-dimensional M-estimation to study the Cox proportional hazards model (COX, 1972). These results are useful to compute the limiting distribution of our estimator but not sufficient for constructing confidence bands.

Typically, confidence bands for F are obtained from a rather simple limiting distribution or bootstrap. In the i.i.d. setting, the Kolmogorov-Smirnov statistic is used to compute confidence bands for continuous random variables. Its limiting distribution is the supremum of Brownian bridge, whose quantile is analytically obtained (KOLMOGOROV, 1933; SMIRNOV, 1944). For non-continuous random variables, confidence bands can be obtained by inverting the Dvoretzky–Kiefer–Wolfowitz inequality (DVORETZKY et al., 1956) with a tight constant obtained by MASSART (1990). An alternative way explored by BICKEL and FREEDMAN (1981) is to bootstrap the Kolmogorov-Smirnov statistic to estimate its quantiles. For stratified sampling from a finite population where X_i is treated as fixed, BICKEL and KRIEGER (1989) apply bootstrap methods for finite population sampling to the weighted Kolmogorov-Smirnov statistic to obtain valid confidence bands. These bootstrap methods cover the distribution function for non-continuous random variables.

In our data integration setting, randomness comes from (1) sampling from population and (2) subsequent sampling from data sources. A valid confidence band should reflect both types of uncertainty. The previous methods described above focus on randomness due to either sampling from population or finite population sampling, and cannot be applied to our data integration problem. The corresponding limiting distribution in our setting is the supremum of the linear combination of independent Gaussian processes. This process cannot be reduced to other well-known processes in general. Also, our formulation of the data integration problem is rather new and a valid bootstrap method is not available.

Methods for confidence bands for the distribution function have been studied in various ways other than analytical computation of quantiles of the limiting distribution and bootstrap. Confidence bands for parametric models are considered for normal distributions (KANOFSKY and SRINIVASAN, 1972), Weibull distributions (SCHAFER and ANGUS, 1979), and the location scale parameter model (CHENG and ILES, 1983). Bayesian approach with the Dirichlet prior was studied by BRETH (1978). OWEN (1995) considered inverting a nonparametric likelihood test of uniformity by BERK and JONES (1978). FREY (2008) proposed the narrowness criterion to derive optimal confidence bands. WANG et al. (2013) developed a smooth confidence band based on the kernel smoothed estimator of a distribution function.

The rest of the paper is organized as follows. In Section 2, we introduce our esti-

mator of F and derive its limiting distribution. We present the algorithm to compute the confidence band and study its asymptotic property in Section 3. We extend our methodology to conditional distribution functions in Section 4. The performance of the proposed method is evaluated through a simulation study in Section 5. We discuss a data example from the national Wilms tumor study in Section 6. All proofs are deferred to the appendix.

2. Estimator and its asymptotic properties

We introduce additional notation for our estimator. Let $R_i^{(j)} \in \{0,1\}$ be the selection indicator from data source $\mathscr{V}^{(j)}$. The item i has a vector of selection indicators $R_i = (R_i^{(1)}, \ldots, R_i^{(J)})$ but $R_i^{(j)} = 0$ if the item i does not belongs to source $\mathscr{V}^{(j)}$. For the items i in data source j (i.e., $V_i \in \mathscr{V}^{(j)}$), $R_i^{(j)}$s follow the distribution of sampling without replacement where $n^{(j)}$ is selected out of $N^{(j)}$. Since data collection procedures are carried out independently, selection indicators $(R_1^{(j)}, \ldots, R_N^{(j)})$ and $(R_1^{(k)}, \ldots, R_N^{(k)})$ are conditionally independent given $V_1, \ldots, V_N$ if $j \neq k$. For $V \in \mathscr{V}^{(j)}$, we assume the selection probability $\pi^{(j)}(V) = n^{(j)}/N^{(j)}$ converges to $p^{(j)} > 0$ as $N \to \infty$. We write the membership probability in source $\mathscr{V}^{(j)}$ as $\nu^{(j)} = P(V \in \mathscr{V}^{(j)})$ and the conditional expectation given membership in source $\mathscr{V}^{(j)}$ as $E^{(j)}$.

The desirable properties that an estimator of F in our data integration setting should satisfy are that (1) the estimator corrects bias due to biased sampling and duplication, and that (2) the estimator is computable without identification of duplicated records. To describe our estimator, we begin with $J = 2$ data sources. The key component of our estimator is

$$\rho(v) = (\rho^{(1)}(v), \rho^{(2)}(v)) \equiv \begin{cases} (1,0) & \text{if} \quad v \in \mathscr{V}^{(1)} \text{ and } v \notin \mathscr{V}^{(2)}, \\ (0,1) & \text{if} \quad v \notin \mathscr{V}^{(1)} \text{ and } v \in \mathscr{V}^{(2)}, \\ (c^{(1)}, c^{(2)}) & \text{if} \quad v \in \mathscr{V}^{(1)} \cap \mathscr{V}^{(2)}, \end{cases}$$

for positive constants $c^{(1)}, c^{(2)}$ with $c^{(1)} + c^{(2)} = 1$. The evaluation of this function only requires the membership in the mutually exclusive subsets of $\mathscr{V}$ based on data sources $\mathscr{V}^{(1)}$ and $\mathscr{V}^{(2)}$. We can compute the value of ρ for selected items because we assume information on data source membership is available for selected items. The choice of ρ is at the disposal of a data analyst. The optimal choice of ρ is considered by SAEGUSA (2019) and we use them in a simulation study and data example below.

Using the function ρ, we propose the following estimator of F given by

$$\mathbb{F}_N(x) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) I\{X_i \leq x\}.$$

Here we use the convention $0/0 = 0$ for the inverse probability weighting $R^{(j)}/\pi^{(j)}(V)$. This estimator is unbiased for F because inverse probability weighting $R^{(j)}/\pi^{(j)}(V)$ has conditional expectation 1 given $V_1, \ldots, V_N$ and $X_1, \ldots, X_N$ and because $\rho^{(1)}(v) + \rho^{(2)}(v) =$

1 for every v. Moreover, the estimator can be computed separately based on two subsamples through the expression

$$\mathbb{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{R_i^{(1)}\rho^{(1)}(V_i)}{\pi^{(1)}(V_i)} I\{X_i \le x\} + \frac{1}{N}\sum_{i=1}^{N} \frac{R_i^{(2)}\rho^{(2)}(V_i)}{\pi^{(2)}(V_i)} I\{X_i \le x\}.$$

The proposed estimator can be considered as the weighted empirical distribution with weights computed from the selection probability and the function ρ. A difference from the empirical distribution is that our estimator may not have $\mathbb{F}_N(x) = 1$ for x greater than the largest selected X_i unless all the items i in $\mathcal{V}^{(1)} \cap \mathcal{V}^{(2)}$ selected from source $\mathcal{V}^{(1)}$ are also selected from $\mathcal{V}^{(2)}$. If $\mathbb{F}_N(x) > 1$ we can modify our estimator to $\tilde{\mathbb{F}}_N(x) = \min\{\mathbb{F}_N(x), 1\}$. For brevity of the presentation, we study $\mathbb{F}_N(x)$ but all properties below are satisfied for $\tilde{\mathbb{F}}_N(x)$.

The extension to more than two data sources is straightforward. Let $\rho = (\rho^{(1)}, \ldots, \rho^{(J)}) : \mathcal{V} \mapsto [0,1]^J$ where

$$\rho^{(j)}(v) = \begin{cases} 1, & v \in \mathcal{V}^{(j)} \cap \left(\cup_{m \ne j} \mathcal{V}^{(m)}\right)^c, \\ c^{(j)}_{k_1,\ldots,k_l}, & v \in \mathcal{V}^{(j)} \cap \left(\cap_{m=1}^{l} \mathcal{V}^{(k_m)}\right) \cap \left(\cup_{m \notin \{j,k_1,\ldots,k_l\}} \mathcal{V}^{(m)}\right)^c, \\ 0, & v \notin \mathcal{V}^{(j)}, \end{cases}$$

with $j, k_1, \ldots, k_l$ all different and $\sum_{j=1}^{J} \rho^{(j)}(v) = 1$. The proposed estimator is

$$\mathbb{F}_N(x) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J} \frac{R_i^{(j)}\rho^{(j)}(V_i)}{\pi^{(j)}(V_i)} I\{X_i \le x\}.$$

Now, we develop asymptotic properties of our estimator. As the uniform consistency of the empirical distribution follows from the Glivenko-Cantelli theorem, the uniform consistency for our estimator follows from the uniform law of large numbers for data integration (SAEGUSA, 2019).

Theorem 2.1. *The estimator $\mathbb{F}_N$ is uniformly consistent for F over $\mathbb{R}$. That is,*

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_N(x) - F(x)| \to_P 0.$$

As the Donsker theorem yields the weak convergence of the empirical distribution to the Brownian bridge process, the weak convergence for our estimator follows from the uniform central limit theorem for data integration (SAEGUSA, 2019). Its limiting distribution is still a Gaussian process, but not the Brownian bridge process.

Theorem 2.2. *Let $D(\mathbb{R})$ be the class of cadlag functions on $\mathbb{R}$. Our estimator $\sqrt{N}(\mathbb{F}_N - F)$ weakly converges to the Gaussian process $\mathbb{G}$ in $D(\mathbb{R})$ given by*

$$\mathbb{G} = \mathbb{G}_0 + \sum_{j=1}^{J} \sqrt{\nu^{(j)}} \sqrt{\frac{1-p^{(j)}}{p^{(j)}}} \mathbb{G}_j,$$

where $\mathbb{G}_j, j = 0,1,\ldots,J$, *are independent Gaussian processes with covariance functions*

$$\begin{aligned} k^{(0)}(s,t) &= F(s\wedge t) - F(s)F(t), \\ k^{(j)}(s,t) &= E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le s\wedge t\}\right] \\ &\quad - E^{(j)}\left[\rho^{(j)}(V) I\{X \le s\}\right] E^{(j)}\left[\rho^{(j)}(V) I\{X \le t\}\right], \end{aligned}$$

for $s,t \in \mathbb{R}$ *and* $j = 1,\ldots,J$.

An immediate consequence of this theorem is that $\sqrt{N}(\mathbb{F}_N(x) - F(x))$ converges in distribution to the zero-mean normal random variable with variance as the sum of $P(X \le x)\{1 - P(X \le x)\}$ and

$$\sum_{j=1}^{J}\left(\nu^{(j)}\frac{1-p^{(j)}}{p^{(j)}}E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le x\}\right] - \left\{E^{(j)}\left[\rho^{(j)}(V) I\{X \le x\}\right]\right\}^2\right).$$

Note that $P(X \le x)\{1 - P(X \le x)\}$ is asymptotic variance which we would obtain from the analysis of i.i.d. data. Merging samples from overlapping sources increases additional uncertainty in our estimator. If we select all items from each source without identifying duplication, then $p^{(j)} = 1, j = 1,\ldots,J$, yield the same variance as in the i.i.d. case. Hence, we see that the additional variance comes from additional selection, not duplication. The effect of duplication appear only through the variable $\rho^{(j)}(V)$. Uncertainty in large data source (i.e., $\nu^{(j)} = P(V \in \mathscr{V}^{(j)})$ contributes more to the asymptotic variance.

3. Confidence band

The basic idea to obtain a confidence band is to obtain $q_{1-\alpha}$ such that

$$P\left(\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)| \le q_{1-\alpha}\right) \to 1-\alpha, \quad n \to \infty,$$

from which the large sample $100(1-\alpha)\%$ confidence band is obtained as

$$\mathbb{F}_N(x) - q_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + q_{1-\alpha}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}.$$

One potential approach is to use an analytical expression of quantiles of the limiting distribution of $\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)|$ but this limiting distribution $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$ obtained from Theorem 2.1 is the supremum of the complicated Gaussian process whose quantiles cannot be analytically derived in general. Another approach is to estimate $q_{1-\alpha}$ by nonparametrically bootstrapping $\sup_{x\in\mathbb{R}} \sqrt{N}\,|\mathbb{F}_N(x) - F(x)|$ but there is no known valid bootstrap method for our setting. Generating data from $\mathbb{F}_N$ would be another alternative but it is not clear how to simultaneously generate V to mimic the data integration process.

The proposed methodology does not analytically compute $q_{1-\alpha}$ from the limiting distribution nor simulating data generating mechanism. Instead, we directly simulate

the limiting distribution to estimate its quantiles. The distribution of the zero-mean Gaussian process $\mathbb{G}$ is completely determined by the unknown covariance function

$$k(s,t) = k^{(0)}(s,t) + \sum_{j=1}^{J} \nu^{(j)} \frac{1-p^{(j)}}{p^{(j)}} k^{(j)}(s,t).$$

We estimate this covariance function $k(s,t)$ as follows. For data source membership probability $\nu^{(j)}$ and selection probability $p^{(j)}$, we estimate them by $N^{(j)}/N$ and $n^{(j)}/N^{(j)}$ respectively. For $k^{(0)}(s,t)$, an obvious estimator is $\mathbb{F}_N(s \wedge t) - \mathbb{F}_N(s)\mathbb{F}_N(t)$. For $k^{(j)}(s,t)$, conditional expectations given membership in $\mathscr{V}^{(j)}$ are estimated by inverse probability weighting based on a sample selected from source $\mathscr{V}^{(j)}$ (i.e., items i with $R_i^{(j)} = 1$). Specifically, the first term in $k^{(j)}(s,t)$ is estimated by

$$\frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \{\rho^{(j)}(V_i)\}^2 I\{X_i \le s \wedge t\},$$

and the second term in $k^{(j)}(s,t)$ is estimated by

$$\left\{ \frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) I\{X_i \le s\} \right\} \left\{ \frac{1}{N^{(j)}} \sum_{i=1}^{N} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) I\{X_i \le t\} \right\}.$$

We denote our estimator of $k(s,t)$ by $\hat{k}_N(s,t)$.

The zero-mean Gaussian process $\hat{\mathbb{G}}_N$ with covariance function $\hat{k}_N(s,t)$ weakly converges to the limiting process $\mathbb{G}$. However, the supremum of $|\mathbb{G}(x)|$ may have a jump at the lower end of the support of X (TSIRELSON, 1975). To avoid the possibility that the jump occurs at its $100(1-\alpha)$%tile, we assume the following condition. The same condition is imposed by BICKEL and KRIEGER (1989) for finite population sampling.

Condition 3.1. *The distribution of* $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$ *is continuous.*

Under this condition, we have the following result.

Theorem 3.1. *Let* $q \in \mathbb{R}$. *Let* $\hat{\mathbb{G}}_N$ *be the zero-mean Gaussian process with covariance function* $\hat{k}_N(s,t)$, *Under Condition 3.1, as* $N \to \infty$,

$$P\left(\sup_{x\in\mathbb{R}} |\hat{\mathbb{G}}_N(x)| \le q \right) \to P\left(\sup_{x\in\mathbb{R}} |\mathbb{G}(x)| \le q \right).$$

We propose the following procedure to construct a confidence band of F:

- Generate the first zero-mean Gaussian process $\hat{\mathbb{G}}_N$ with covariance function $\hat{k}_N(s,t)$, and compute the supremum s_1 of $|\hat{\mathbb{G}}_N|$
- Repeat this procedure B times to obtain $s_1, \ldots, s_B$, and compute their $100(1-\alpha)$%tile $\hat{q}_{1-\alpha}$.

- Compute the $100(1-\alpha)\%$ confidence band of F by

$$\mathbb{F}_N(x) - \hat{q}_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + \hat{q}_{1-\alpha}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}. \tag{1}$$

The proposed confidence band has the correct coverage probability asymptotically.

Theorem 3.2. *Under Condition 3.1, as* $N, B \to \infty$,

$$P\left(\mathbb{F}_N(x) - \hat{q}_{1-\alpha}/\sqrt{N} \le F(x) \le \mathbb{F}_N(x) + \hat{q}_{1-\alpha}/\sqrt{N}, \quad \textit{all } x \in \mathbb{R}\right) \to 1-\alpha.$$

4. Extension to conditional distribution given discrete variables

In practice, it is of interest to compare different groups through graphical comparison of distribution functions. An extension of our method to conditional distributions given a discrete random variable is straightforward. Let U be a discrete random variable. First, we estimate the sub-distribution function $F(x,u) = P(X \le x, U = u)$ by

$$\mathbb{F}_N(x,u) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right) I\{X_i \le x, U_i = u\}.$$

The limiting distribution is similar to the one in Theorem 2.2 but covariance functions are now

$$\begin{aligned} k_u^{(0)}(s,t) &= P(X \le s \wedge t, U = u) - P(X \le s, U = u)P(X \le t, U = u), \\ k_u^{(j)}(s,t) &= E^{(j)}\left[\left\{\rho^{(j)}(V)\right\}^2 I\{X \le s \wedge t, U = u\}\right] \\ &\quad - E^{(j)}\left[\rho^{(j)}(V) I\{X \le s, U = u\}\right] E^{(j)}\left[\rho^{(j)}(V) I\{X \le t, U = u\}\right]. \end{aligned}$$

This covariance function can be similarly estimated and the same procedure described above yields the confidence band given by

$$\mathbb{F}_N(x,u) - \hat{q}_{1-\alpha,u}/\sqrt{N} \le F(x,u) \le \mathbb{F}_N(x,u) + \hat{q}_{1-\alpha,u}/\sqrt{N}, \quad \text{all } x \in \mathbb{R}.$$

Since $F(x|u) = P(X \le x|U = u) = P(X \le x, U = u)/P(U = u)$, we estimate $p_u = P(U = u)$ by a consistent estimator

$$\hat{p}_u = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right) I\{U = u\}.$$

Now we propose the confidence band for $F(x|u)$ given by

$$\frac{\mathbb{F}_N(x,u)}{\hat{p}_u} - \frac{\hat{q}_{1-\alpha,u}}{N^{1/2}\hat{p}_u} \le F(x|u) \le \frac{\mathbb{F}_N(x,u)}{\hat{p}_u} + \frac{\hat{q}_{1-\alpha,u}}{N^{1/2}\hat{p}_u}, \quad \text{all } x \in \mathbb{R}. \tag{2}$$

Table 1: Sample sizes for three scenarios.

	Scenario 1 & 2			Scenario 3		
N	100	250	500	100	250	500
$N^{(1)}$	79	197	395	78	197	395
$N^{(2)}$	51	127	255	51	127	255
$N^{(3)}$				28	70	141
$n^{(1)}$	16	40	80	16	40	80
$n^{(2)}$	16	38	77	16	38	77
$n^{(3)}$				14	35	71
Duplication (2 sources)	2	5	9	6	2	27
Duplication (3 sources)				0	1	1

This confidence band has the correct coverage probability asymptotically. The proof is similar to that of Theorem 3.2, and omitted.

5. Simulation study

We carry out a simulation study to evaluate the finite-sample performance of the proposed confidence band. We consider three different scenarios. The first two scenarios concern two partially overlapping data sources. The third scenario deals with three data sources with one data source contained in other two. The distributions considered are mixtures of beta distributions, Poisson distributions, and normal distributions, respectively.

In the first scenario, the variable Y is a Bernoulli random variable with $p = 0.3$. The variable X of interest follows the beta distribution with $\alpha = 5$ and $\beta = 2$ if $Y = 0$ and the beta distribution with $\alpha = 2$ and $\beta = 5$ if $Y = 1$. The variables $W = (X, Y)$ are not available at the first stage of sampling. The auxiliary binary variable V is correlated with Y with sensitivity 0.9 and specificity 0.9. Data sources are created by values of V. If $V = 0$, the item belongs to data source 1 and if $V = 1$ it belongs to data source 2. In both situations, the item belongs to the intersection of two data sources with probability 0.3. Selection probabilities are 0.2 from data source 1 and 0.3 from data source 2. The second scenario is the same as the first except that the variable X follows the Poisson distribution with $\lambda = 2$ if $Y = 0$ and the Poisson distribution with $\lambda = 4$ if $Y = 1$. In the third scenario, variables Y and V and data sources 1 and 2 are similarly generated as in the other two cases. The variable X follows the normal distribution with $\mu = 1$ and $\sigma^2 = 1$ if $Y = 0$ and the normal distribution with $\mu = 3$ and $\sigma^2 = 1$ if $Y = 1$. The data source 3 consists of items with $X \in [1, 2]$. Selection probabilities are 0.2 from data source 1, 0.3 from data source 2, and 0.5 from data source 3.

Data were generated 500 times in each scenario with sample size $N = 100$, $N = 250$, and $N = 500$. In each data set, the 95% confidence band was constructed based on 2000 simulated Gaussian processes with the formula (1). Table 1 summarizes average sample sizes for each data source before and after the selection into the final sample. Note that the proposed estimator is based on 30 items for scenarios 2 and 3, and 40

Table 2: Simulated coverage probabilities for the confidence bands.

	Scenario 1		Scenario 2		Scenario 3	
	Coverage	Width	Coverage	Width	Coverage	Width
$N = 100$	0.940	0.454	0.936	0.442	0.920	0.464
$N = 250$	0.944	0.295	0.954	0.286	0.956	0.304
$N = 500$	0.952	0.211	0.944	0.203	0.956	0.217

items for scenario 3 on average without duplication when $N = 100$. Table 2 shows simulated coverage probabilities and average width based on 500 data simulated data sets. Coverage probabilities are close to the nominal level in all scenarios when N is greater than 250 while we see under-coverage when $N = 100$. Confidence bands are wide for $N = 100$ but the width becomes reasonable as N increases. Overall, our methodology shows reasonable performance for a practical use.

6. Application

We illustrate the proposed method using data from the national Wilms tumor study (D'ANGIO et al., 1989). Wilms tumor is a rare kidney cancer for children. The predictor of relapse includes histology of cancer, age at diagnosis, and tumor diameter. Data for all 3915 patients are available and were used to compare different designs (BRESLOW and CHATTERJEE, 1999; BRESLOW et al., 2009; SAEGUSA, 2019). In our analysis, we check if the empirical distributions based on the entire cohort are contained in the proposed confidence bands based on a smaller biased sample with duplication. Three data source are deceased patients, patients with unfavorable histology measured at the hospital, and the entire cohort. Selection probabilities 100%, 50%, and 10%, respectively, yielding the sample size 1027 in the final sample (885 patients without duplication). For selected patients, tumor diameter is measured and histology is re-examined at the central reference laboratory. Our goal is to create two distribution functions of tumor diameter based on the histology information measured at the second time. Among selected patients, 646 (603 without duplication) patients have favorable histology and 382 (282 without duplication) patients have unfavorable histology.

Figure 2 shows the confidence bands for the conditional distributions of tumor diagmeter given histology based on the formula (2). The solid line is smoothed empirical distribution based on the entire cohort of size 3915. Our estimators are close to empirical distributions. Moreover, the proposed confidence bands successfully contain empirical distributions. The difference in sample sizes based on histology is reflected in the difference of widths. The confidence band for favorable histology has width 0.133 while the band for unfavorable histology has width 0.307. Graphical comparison of both estimators with the help of confidence bands shows that there is no striking difference between distributions of tumor diameter in different histology groups. In fact, empirical quartiles of tumor diameter for both groups agree well. A similar analysis (not shown here) conditional on survival status led to the same conclusion. In the proportional hazards regression analysis, SAEGUSA (2019) shows that tumor diameter has a small effect on

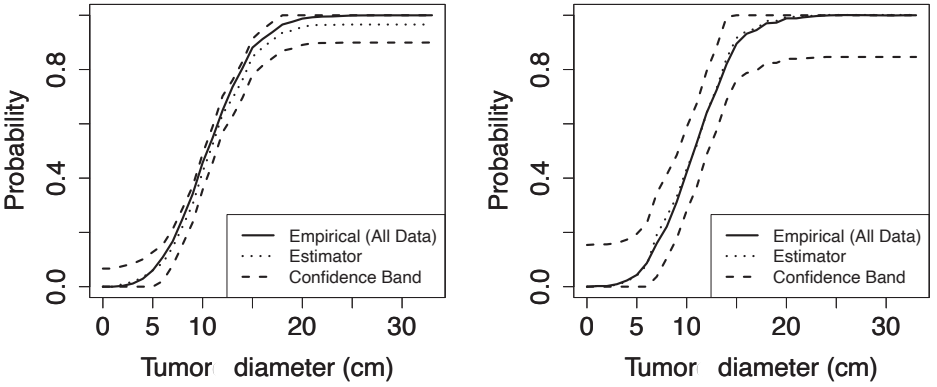


Figure 2: Confidence bands for conditional distribution functions of tumor diameter given favorable histology (left panel) and unfavorable histology (right panel).

tumor relapse while histology is statistically significant.

Acknowledgements

We thank Partha Lahiri for helpful discussions and encouragement for this project.

References

BERK, R. H. JONES, D. H., (1978). Relatively optimal combinations of test statistics. *Scand. J. Statist.*, 5(3), pp. 158–162.

BICKEL, P. J. FREEDMAN, D. A., (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6), pp,1196–1217.

BICKEL, P. J. KRIEGER, A. M., (1989). Confidence bands for a distribution function using the bootstrap. *J. Amer. Statist. Assoc.*, 84(405), pp. 95–100.

BRESLOW, N. E. CHATTERJEE, N., (1999). Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4), pp. 457–468.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C., CHAMBLESS, L., KULICH, M., (2009). Using the whole cohort in the analysis of case-cohort data. *American J. Epidemiol.*, 169, pp. 1398–1405.

BRETH, M., (1978). Bayesian confidence bands for a distribution function. *Ann. Statist.*, 6(3), pp. 649–657.

BRICK, J. M., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *The Public Opinion Quarterly*, 70(5), pp. 780–793.

CERVANTES, I., JONES, M., ROJAS, L., BRICK, J., KURATA, J., GRANT, D., (2006). A review of the sample design for the california health interview survey. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 3023–3030.

CHATTERJEE, N., CHEN, Y.-H., MAAS, P., CARROLL, R. J., (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.*, 111(513), pp. 107–117.

CHENG, R. C. H. ILES, T. C., (1983). Confidence bands for cumulative distribution functions of continuous random variables. *Technometrics*, 25(1), pp.77–86.

COX, D. R., (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, 34, pp. 187–220.

D'ANGIO, G. J., BRESLOW, N., BECKWITH, J. B., EVANS, A., BAUM, H., DE-LORIMIER, A., FERNBACH, D., HRABOVSKY, E., JONES, B., KELALIS, P., (1989). Treatment of Wilms' tumor. Results of the Third National Wilms' Tumor Study. *Cancer*, 64(2), pp. 349–360.

DVORETZKY, A., KIEFER, J., WOLFOWITZ, J., (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27, pp. 642–669.

FREY, J., (2008). Optimal distribution-free confidence bands for a distribution function. *J. Statist. Plann. Inference*, 138(10), pp. 3086–3098.

GINÉ, E. NICKL, R., (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York.

HARTLEY, H. O., (1962). Multiple frame surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 203–206.

HARTLEY, H. O., (1974). Multiple frame methodology and selected applications. *Sankhyā Ser. C*, 36, pp. 99–118.

HU, S. S., BALLUZ, L., BATTAGLIA, M. P., FRANKEL, M. R., (2011). Improving public health surveillance using a dual-frame survey of landline and cell phone numbers. *American Journal of Epidemiology*, 173(6), pp. 703–711.

KANOFSKY, P. SRINIVASAN, R., (1972). An approach to the construction of parametric confidence bands on cumulative distribution functions. *Biometrika*, 59, pp. 623–631.

KEIDING, N. LOUIS, T. A., (2016). Perils and potentials of self-selected entry to epidemiological studies and surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 179(2), pp. 319–376.

KOLMOGOROV, A. N., (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4, pp. 83–91.

MASSART, P., (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3), pp. 1269–1283.

METCALF, P. SCOTT, A., (2009). Using multiple frames in health surveys. *Statistics in Medicine*, 28(10), pp. 1512–1523.

OWEN, A. B., (1995). Nonparametric likelihood confidence bands for a distribution function. *J. Amer. Statist. Assoc.*, 90(430), pp. 516–521.

SAEGUSA, T., (2019). Large sample theory for merged data from multiple sources. *Ann. Statist.*, 47(3), pp. 1585–1615.

SAEGUSA, T. WELLNER, J. A., (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.*, 41(1), pp. 269–295.

SCHAFER, R. E. ANGUS, J. E., (1979). Estimation of weibull quantiles with minimum error in the distribution function. *Technometrics*, 21(3), pp. 367–370.

SMIRNOV, N. V., (1944). Approximate laws of distribution of random variables from empirical data. *Uspehi Matem. Nauk*, 10, pp. 179–206.

TSIRELSON, V. S., (1975). The density of the distribution of the maximum of a Gaussian process. *Theory of Probability and its Applications*, 20, pp. 847–865.

WANG, J., CHENG, F., YANG, L., (2013). Smooth simultaneous confidence bands for cumulative distribution functions. *J. Nonparametr. Stat.*, 25(2), pp. 395–407.

APPENDIX

Proof of Theorem 2.1. Because the class of functions $\mathscr{F} = \{f_t(x) = I(x \le t) : t \in \mathbb{R}\}$ is a Glivenko-Cantelli class, apply the uniform law of large numbers for data integration (Theorem 3.1 of SAEGUSA (2019)) to $\mathscr{F}$ to obtain the desired result. □

Proof of Theorem 2.2. Because the class of functions $\mathscr{F} = \{f_t(x) = I(x \le t) : t \in \mathbb{R}\}$ is also a Donsker class, apply the uniform central limit theorem for data integration (Theorem 3.2 of SAEGUSA (2019)) to $\mathscr{F}$. The computation of the covariance function is straightforward. □

Proof of Theorem 3.1. We show the weak convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$. First we consider the finite dimensional convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$. As in the proof of Theorem 2.1, the law of large numbers for data integration yields

$$\sup_{s,t\in\mathbb{R}} |\mathbb{F}_N(s \wedge t) - \mathbb{F}_N(s)\mathbb{F}_N(t) - k^{(0)}(s,t)| \to_P 0.$$

For $k^{(j)}(s,t), j = 1,\ldots,J$, the law of large numbers for sampling without replacement (Theorem 5.1 of SAEGUSA and WELLNER (2013)) yields the uniform consistency over $s,t \in \mathbb{R}$. Since $n^{(j)}/N^{(j)} \to p^{(j)}$ by assumption and $N^{(j)}/N \to_P \nu^{(j)}$ by the weak law of large numbers, we conclude

$$\sup_{s,t\in\mathbb{R}} |\hat{k}_N(s,t) - k(s,t)| \to_P 0.$$

This implies the desired finite dimensional convergence.

Second, we consider asymptotic equicontinuity and total boundedness of $\mathbb{R}$ with respect to a constant multiple of

$$d^{(0)}(s,t) = k^{(0)}(s,s) + k^{(0)}(t,t) - 2k^{(0)}(s,t).$$

Note that the intrinsic metric $d(s,t) = k(s,s) + k(t,t) - 2k(s,t)$ to the limiting process $\mathbb{G}$ is equivalent to $d^{(0)}(s,t)$ (i.e.,$C_1 d(s,t) \le d^{(0)}(s,t) \le C_2 d(s,t)$ for some constants $C_1, C_2 > 0$) because $\rho^{(j)}(v)$ is bounded. Also, on the event A that $\sup_{s,t\in\mathbb{R}} |\hat{k}_N(s,t) - k(s,t)| < C_3$ for some small fixed constant $C_3 > 0$, $\hat{d}_N(s,t) = \hat{k}_N(s,s) + \hat{k}_N(t,t) - 2\hat{k}_N(s,t)$ is equivalent to $d(s,t)$ since $d(s,t)$ is bounded over $\mathbb{R}^2$. These observations imply that the process $\mathbb{G}$ and $\hat{\mathbb{G}}_N$ are sub-Gaussian processes with respect to $Cd^{(0)}(s,t)$ for some constant $C > 0$ on the event A. As a consequence, the property of the sub-Gaussian process (see e.g. Theorem 2.3.7 of GINÉ and NICKL (2016)) implies that

$$E\left[\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \varepsilon \middle| A\right] \le K \int_0^\delta \sqrt{\log 2N(\mathbb{R}, Cd^{(0)}, \varepsilon)}\, d\varepsilon \tag{3}$$

for some constant $K > 0$ as long as the integral on the right hand side is finite. Here $N(\mathbb{R}, Cd^{(0)}, \varepsilon)$ is the covering number of $\mathbb{R}$ with respect to the metric $Cd^{(0)}$ with radius ε.

For asymptotic equicontinuity, let $\eta > 0$ be arbitrary. We have

$$\limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta \right)$$
$$\le \limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta, A \right) + P(A^c).$$

where A^c is the complement of A. Since $P(A^c) \to 0$, we bound the first term by the Markov inequality and the inequality (3) to obtain

$$\limsup_{n\to\infty} P\left(\sup_{d^{(0)}(s,t)\le\delta} \left|\hat{\mathbb{G}}_N(s) - \hat{\mathbb{G}}_N(t)\right| > \eta \,\middle|\, A \right) P(A)$$
$$\le \limsup_{n\to\infty} \eta^{-1} K \int_0^\delta \sqrt{\log 2N(\mathbb{R}, Cd^{(0)}, \varepsilon)}\, d\varepsilon \to 0, \quad \text{as } \delta \downarrow 0,$$

assuming the integral on the right hand side is finite for any δ, which we will show next.

To compute the covering number with radius ε, create l subintervals $[I_i, I_{i+1}]$ of [0,1] with length less than ε with $I_0 = 0 < I_1 < \cdots < I_{l+1} = 1$. Note that we do not consider $\varepsilon \ge 1$ since we take $\delta \downarrow 0$. Let $q_i = F^{-1}(I_i)$. Then $F(q_{i+1}) - F(q_i) \le \varepsilon$. If $t \in [I_i, I_{i+1})$, we have

$$d^{(0)}(q_i, t) = F(q_i)\{1 - F(q_i)\} + F(t)\{1 - F(t)\} - 2\{F(q_i) - F(q_i)F(t)\} \le 4\varepsilon.$$

This means t is in the $d^{(0)}$-ball with center q_i and radius 4ε. This implies that the covering number with radius ε is proportional to $1/\varepsilon$, and hence the entropy integral converges. This computation also shows that $\mathbb{R}$ is totally bounded with respect to $d^{(0)}$. Because asymptotic equicontinuity and total boundedness imply asymptotic tightness, we now conclude the weak convergence of $\hat{\mathbb{G}}_N$ to $\mathbb{G}$.

The continuous mapping theorem yields that $\sup_{x\in\mathbb{R}} |\hat{\mathbb{G}}_N(x)|$ converges in distribution to $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Thus, the desired result follows from Condition 3.1. □

Proof of Theorem 3.2. Theorem 2.2 and continuous mapping theorem imply $\sup_{x\in\mathbb{R}} \sqrt{N}|\mathbb{F}_N(x) - F(x)|$ converges in distribution to $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Theorem 3.1 implies that $\hat{q}_{1-\alpha}$ converges in probability to the $100(1-\alpha)$%tile $q_{1-\alpha}$ of $\sup_{x\in\mathbb{R}} |\mathbb{G}(x)|$. Combining these results completes the proof. □

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 159–165, DOI 10.21307/stattrans-2020-036
Received – 31.01.2020; accepted – 30.06.2020

Model selection in radon data fusion

Xuze Zhang[1], **Saumyadipta Pyne**[2], **Benjamin Kedem**[3]

ABSTRACT

Fitting parametric models or the use of the empirical cumulative distribution function are problematic when it comes to the estimation of tail probabilities from small samples. A possible remedy is to fuse or combine the small samples with additional data from external sources and base the inference on the so called density ratio model with variable tilt functions, which widens the support of the estimated distribution of interest. This approach is illustrated using residential radon concentration data collected from western Pennsylvania.

Key words: Tail probabilities, density ratio model, variable tilt functions, Appalachian Plateau, Forest County, Pennsylvania.

1. Introduction

In general, the estimation of tail probabilities requires large samples. However, in many cases the available samples are relatively small, a problem which can be overcome to a reasonable extent by fusing the available data from several independent sources. This is illustrated here using residential radon concentration data collected from counties in western Pennsylvania (PA). We used county-level indoor radon concentrations based on records collected by the Pennsylvania Department of Environmental Protection (PA DEP), Bureau of Radiation Protection, Radon Division. For more details about the data see Zhang, Pyne, and Kedem (ZPK) (2019), and the appropriate references including PA Department of Environmental Protection, Rack-Amber (2013), Wikipedia contributors (2019).

The range of values of a small sample may not be large enough to shed light on the tail behavior of the distribution which gave rise to the sample. In that case more data are needed. However, in many cases, more data are not available. Our goal is to demonstrate that the problem can be ameliorated to a reasonable extent when the sample is fused or combined with data from other sources, as the range of values of the combined data is larger. Technically, this can be achieved by appealing to the so called *density ratio model* (DRM), where the distributions of the various sources are connected by fixed *tilt functions*. The novelty of the paper is the use of *variable tilts* obtained by model selection.

[1]Department of Mathematics and Institute for Systems Research, University of Maryland, College Park. USA. E-mail: xzhang51@umd.edu. ORCID: https://orcid.org/0000-0002-8672-8515.

[2]Public Health Dynamics Laboratory, and Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh. USA. E-mail: spyne@pitt.edu. ORCID: https://orcid.org/0000-0003-3470-2345.

[3]Department of Mathematics and Institute for Systems Research, University of Maryland, College Park. USA. E-mail: bnk@umd.edu. ORCID: https://orcid.org/0000-0001-8720-3465.

In this paper, we apply a data fusion method in the estimation of residential radon levels in Forest County, located in the Appalachian Plateau in western PA. Its population is small, with 2000 households as per the 2010 census, yielding a small sample of 47 homes only, insufficient for the estimation of tail probabilities, and hence qualifying it as a "small area" problem. To overcome the small sample size, we fuse the Forest data with samples from the two adjacent counties Elk and Warren whose populations are much larger. Tail probabilities can then be estimated by using the density ratio model (DRM) with *variable tilt functions* ZPK (2019). This formulation requires the selection of optimal models out of a large number of models. In ZPK (2019), the selection of tilt function was done via a long process of hypothesis testing while here we use a more efficient model selection advocated in Fokianos (2007). The DRM is discussed in detail in Kedem, De Oliveira and Sverchkov (KDS) (2017) and Qin (2017).

Fusing data from Forest, Elk, and Warren counties is sensible as they share the geographical features of the "High Plateau Section" in northwestern PA in the region of Appalachian Plateau (Rack-Amber 2013, Wikipedia contributors 2019).

Radon is an odorless cancer-causing radioactive gas released from decaying uranium, thorium and radium in rocks and soil, and is the cause of thousands of deaths each year (Rack-Amber 2013). Approximately 40% of PA homes have radon levels exceeding EPA's action guideline of 4 picocuries (pCi) per liter (PA Department of Environmental Protection).

Therefore, it is of great importance to public health and policy that the residential radon exposure data be analyzed to produce robust tail or exceedance probabilities.

The organization of the paper is as follows. Section 2 deals with the semi-parametric estimation of the parameters and the probability densities of the density ratio model. It also addresses the selection of the tilt functions. A case in point in terms of residential radon is discussed in Section 3. A summary is provided in Section 4.

2. Methodology

2.1. Density Ratio Model

To make use of the data from neighboring counties, a multi-sample DRM is proposed to fuse the data from the county of interest and its m neighbors such that

$$\frac{g_k(x)}{g(x)} = \exp(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(x)) \qquad k = 1, \ldots, m \tag{1}$$

where g represents the density of residential radon levels of the county of interest and $g_1, \ldots, g_m$ represent the densities of its m neighbors.

The semi-parametric estimation of the parameters and densities in (1) is discussed in the next section using the empirical likelihood (Owen 2001). Model (1) was found adequate by a graphical goodness of fit test discussed briefly in Section 3. The model is discussed extensively in the recent books by KDS (2017) and in Qin (2017), which also describe quite a few applications from case-control tests of equidistribution to time series prediction.

Instead of making parametric assumptions on these densities, we propose a parametric structure of their ratios by DRM (KDS 2017, Qin 2017). A proper choice of the tilt functions $\boldsymbol{h}_k$'s is imperative since misspecification of the tilt functions leads to bias, large standard errors, and power loss (Fokianos and Kaimi 2006). We shall commence with a possibly redundant or "global" tilt and then select a reduced form of this tilt. Such a tilt function is specified in section 3.

2.2. Estimation and Asymptotic Result

Let $\boldsymbol{X}_0, \ldots, \boldsymbol{X}_m$ be the samples from the county of interest and its m neighbors with sample sizes $n_0, \ldots, n_m$, respectively. The sample $\boldsymbol{X}_0$ is referred to as the reference sample and we shall denote by G the corresponding reference cumulative distribution function (CDF). The fused sample is defined as $\boldsymbol{t} = (\boldsymbol{X}_0^T, \ldots, \boldsymbol{X}_m^T)^T$, with size $n = \sum_{k=0}^{m} n_k$.

Inference can be based on the following empirical likelihood obtained from the fused sample $\boldsymbol{t}$:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{k=1}^{m} \prod_{j=1}^{n_k} \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(X_{kj})\right) \tag{2}$$

where $p_i = dG(t_i)$ and the estimates $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and hence the $\tilde{p}_i$'s, are obtained by maximizing (2) with constraints

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(t_i)\right) = 1 \qquad k = 1, \ldots, m. \tag{3}$$

Subsequently, we obtain the estimated reference CDF $\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}_i I[t_i \le t]$ and the asymptotic result

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{d} N(0, \sigma(t)), \qquad \text{as } n \to \infty. \tag{4}$$

The expression of $\sigma(t)$ and other details regarding estimation and asymptotic result can be found in KDS (2017), Qin (2017) and ZPK (2019). Therefore, we can construct a 95% confidence interval of the tail probability $1 - G(T)$ for a given threshold T based on (4)

$$(1 - \tilde{G}(T) - z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}, 1 - \tilde{G}(T) + z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}). \tag{5}$$

2.3. Model Selection

As mentioned in 2.1, we aim to select tilt functions that can better specify the density ratio structure. Such selection can be made based on the AIC criterion given by

$$-2\log L(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}}, \tilde{G}) + 2q \tag{6}$$

where q is the number of free parameters in the model (Fokianos 2007). Note that the number of free parameters is equal to the number of β's due to the constraints (3).

3. Illustrative Example: Forest County Radon Data Fusion

Here Forest county is the county of interest. Denote the Forest sample by $\boldsymbol{X}_0$ and its size by n_0. The sample size $n_0 = 47$ is relatively small so that the empirical estimate of the CDF $\hat{G}(t) = \frac{1}{n_0}\sum_{i=1}^{n_0} I[X_{0i} \le t]$ may not be satisfactory for the estimation of tail probabilities. That is, we cannot make inference about G based on $\hat{G}$ outside of the range of $\boldsymbol{X}_0$. Also, a smaller sample size leads to higher standard errors and hence wider confidence intervals, and may not be adequate for the estimation of small tail probabilities.

We wish to mitigate these issues by fusing $\boldsymbol{X}_0$ with samples from its two neighboring counties Warren and Elk to obtain an estimate of the reference CDF $\tilde{G}$ based on the DRM (1). The samples from Warren and Elk are denoted as $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, respectively. The corresponding sample sizes are $n_1 = 837$ and $n_2 = 1191$.

Observing that the data in the three counties are positive and right skewed, the global or redundant tilt function $(x, \log(x), \log^2(x))^T$ is a sensible choice based on ZPK (2019). Hence, we initially assume that $\boldsymbol{h}_k = (x, \log(x), \log^2(x))^T$ for $k = 1, 2$, and then curtail it using the AIC model selection criterion. The AIC values corresponding to different tilts are shown in Table 1.

Table 1: AIC values of models based on different tilt choices. A hyphen "-" indicates that $\boldsymbol{h}_k(x) \equiv \boldsymbol{0}$ and therefore g_0 and g_k are identical.

AIC $\boldsymbol{h}_2$ \ $\boldsymbol{h}_1$	-	x	$\log(x)$	$\log^2(x)$	$(x, \log(x))$	$(x, \log^2(x))$	$(\log(x), \log^2(x))$	$(x, \log(x), \log^2(x))$
-	31696.52	31697.86	31694.68	31697.54	31686.85	31682.73	31694.35	31684.24
x	31698.24	31691.11	31695.63	31699.20	31680.96	31677.07	31696.32	31678.58
$\log(x)$	31693.46	31685.55	31695.07	31692.86	31687.35	31683.05	31694.81	31684.70
$\log^2(x)$	31695.67	31680.36	31696.67	31694.28	31680.14	31680.10	31691.31	31681.62
$(x, \log(x))$	31693.43	31684.21	31695.04	31694.63	31682.37	31679.01	31696.63	31680.02
$(x, \log^2(x))$	31693.13	31682.36	31695.03	31691.36	31681.38	31678.75	31690.98	31680.26
$(\log(x), \log^2(x))$	31695.11	31681.91	31696.71	31693.58	31680.03	31682.06	31691.40	31681.93
$(x, \log(x), \log^2(x))$	31694.44	31683.83	31696.05	31692.66	31680.67	31680.48	31690.13	31682.01

It is observed that the smallest AIC value of 31677.07 is achieved by the model with tilts $\boldsymbol{h}_1(x) = (x, \log^2(x))$ and $\boldsymbol{h}_2(x) = x$.

We proceed to estimate the parameters and reference CDF according to 2.2 with the chosen tilts. The confidence intervals of the tail probabilities for different thresholds obtained from both $\tilde{G}$ and $\hat{G}$ are shown in Table 2.

Table 2: Tail probability $1-G(T)$ estimates and 95% confidence intervals for threshold $T = 5, 10, 25, 50, 100, 150, 200, 250$.

T	$1-\tilde{G}(T)$	95% CI	Length of CI
5	0.4447	$(0.3773, 0.5121)$	0.1349
10	0.2790	$(0.2004, 0.3577)$	0.1573
25	0.1482	$(0.0693, 0.2271)$	0.1578
50	0.0915	$(0.0201, 0.1629)$	0.1429
100	0.0548	$(-0.0041, 0.1138)$	0.1178
150	0.0303	$(-0.0125, 0.0732)$	0.0857
200	0.0264	$(-0.0135, 0.0662)$	0.0798
250	0.0121	$(-0.0142, 0.0384)$	0.0526
T	$1-\hat{G}(T)$	95% CI	Length of CI
5	0.3191	(0.1859,0.4524)	0.2665
10	0.2553	$(0.1307, 0.3800)$	0.2493
25	0.1277	$(0.0323, 0.2231)$	0.1908
50	0.0851	$(0.0053, 0.1649)$	0.1595
100	0.0851	$(0.0053, 0.1649)$	0.1595
150	0.0426	$(-0.0152, 0.1003)$	0.1154
200	0.0213	$(-0.0200, 0.0625)$	0.0825
250	0.0000	-	-

From Table 2, it is readily seen that the lengths of the confidence intervals obtained by the DRM are significantly shorter than those obtained by the empirical CDF for a given threshold T. The slightly negative lower bounds are due to computational problems with small probabilities and should be replaced by 0's.

It is worth noting that $1-\hat{G}(250) = 0$ while $1-\hat{G}(50) = 1-\hat{G}(100)$. This is due to the fact that $\boldsymbol{X}_0$ does not contain observations between $(50, 100)$ or larger than 207. However, we can make inferences on these regions based on $\tilde{G}$ since $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ do contain observations between $(50, 100)$ or larger than 207.

Remark: The use of the DRM requires a justification in terms of goodness-of-fit tests discussed in KDS (2017) and in Qin (2017). As argued in Voulgaraki, Kedem, and Graubard (VKG) (2012), the DRM may not be valid for heavy tailed distributions. Examples include attempts to fit the model to data from two Cauchy distributions and from Cauchy and uniform distributions.

The graphical checking technique proposed in VKG (2012) is applied to check the goodness-of-fit of the selected model. From Figure 1, it is readily seen that the points roughly form a 45°-line, indicating the closeness of $\hat{G}$ and $\tilde{G}$ and hence an adequate DRM. A simulation of fusing absolute data from three Cauchy distributions, Cauchy(0,1), Cauchy(1,2) and Cauchy(2,3) with respective sample sizes 47, 837, 1191, and tilts $\boldsymbol{h}_1(x) = (x, \log^2(x))$ and $\boldsymbol{h}_2(x) = x$, has been conducted where the reference sample contains the absolute data from Cauchy(0,1). These are the sample sizes and tilts used in the analysis of the Forest radon data. It is observed in Figure 2 that the points are far away from a 45°-line, which indicates that the DRM is inappropriate. Such a result agrees with the examples in VKG (2012) mentioned above.

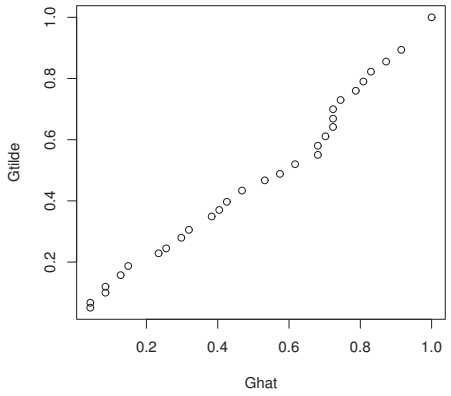


Figure 1: Pairs $(\tilde{G}(T),\hat{G}(T))$ from the selected radon data model

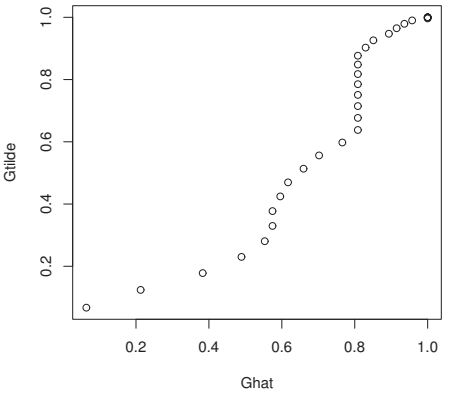


Figure 2: Pairs $(\tilde{G}(T),\hat{G}(T))$ from the DRM fit using absolute data from the three Cauchy distributions

4. Summary

When the size of a sample is relatively small, the empirical CDF might be inadequate for inference on distributions, while making parametric assumptions on the distributions can lead to misspecification. The DRM enables us to make semi-parametric inference about the reference distribution based on more observations, that is, based on fused samples with parametric assumptions on the ratios of the densities. These assumptions are generally weaker than the parametric assumptions on the distribution (ZPK 2019). Furthermore, an AIC based model selection renders the assumptions more sensible and hence it mitigates the problem of misspecification.

In the present residential radon application, we have seen that the lengths of the confidence intervals for tail probabilities obtained by the DRM are shorter than those obtained by the empirical CDF for a given threshold T.

Acknowledgements

Research supported by a Faculty-Student Research Award, University of Maryland, College Park.

REFERENCES

FOKIANOS, K., (2007). Density ratio model selection. *Journal of Statistical Computation and Simulation*, 77(9), pp. 805–819.

FOKIANOS, K., KAIMI, I., (2006). On the effect of misspecifying the density ratio model. *AISM*, 58, pp. 475–497.

KEDEM, B., DE OLIVEIRA, V., SVERCHKOV, M., (2017). *Statistical Data Fusion*, World Scientific, Singapore.

OWEN, A., (2001). *Empirical Likelihood*, Chapman & Hall/CRC, Boca Raton, Fl.

PA DEPARTMENT OF ENVIRONMENTAL PROTECTION. <https://www.dep.pa.gov /Business/RadiationProtection/RadonDivision/Pages/ Radon-in-the-home.aspx>

QIN, J., (2017)., *Biased Sampling, Over-identified Parameter Problems and Beyond*, Springer, Singapore.

RACK-AMBER, T., (2013)., American Lung Association in Pennsylvania to provide free radon testing kits. <https://www.heraldstandard.com/healthy_ living/american-association-in-pennsylvania-to-provide-free-radon-testing/article_ dc55f66a-4c36-588a-87b4-6c4d7448>

VOULGARAKI, A., KEDEM, B., and GRAUBARD, B. I., (2012). Semiparametric regression in testicular germ cell data. *Annals of Applied Statistics*, 6, pp. 1185–1208.

WIKIPEDIA CONTRIBUTORS, (2019). Geology of Pennsylvania - Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/wiki/Geology_of_Pennsylvania> [Online; accessed 20-December-2019].

ZHANG, X., PYNE, S., KEDEM, B., (2019). Estimation of Radon Concentration in Pennsylvania Counties by Data Fusion. *arXiv e-prints*, arXiv:1912.08149.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 166–190, DOI 10.21307/stattrans-2020-037
Received – 31.01.2020; accepted – 30.06.2020

An evaluation of design-based properties of different composite estimators

Daniel Bonnéry[1], Yang Cheng[2], Partha Lahiri[3]

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

ABSTRACT

For the last several decades, the US Census Bureau has been applying AK composite estimation method for estimating monthly levels and month-to-month changes in unemployment using data from the Current Population Survey (CPS), which uses a rotating panel design. For each rotation group, survey-weighted totals, known as month-in-sample estimates, are derived each month to estimate population totals. Denoting the vector of month-in-sample estimates by Y and the design-based variance-covariance matrix of Y by Σ, one can obtain a class of AK estimators as linear combinations of Y, where the coefficients of a linear combination in the class are functions of the two coefficients A and K. The coefficients A and K were optimized by the Census Bureau under rather strong assumptions on Σ such as the stationarity of Σ over a decade. We devise an evaluation study in order to compare the AK estimator with a number of rival estimators. To this end, we construct three different synthetic populations that resemble the Current Population Survey (CPS) data. To draw samples from these synthetic populations, we consider a simplified sample design that mimics the CPS sample design with the same rotation pattern. Since the number of possible samples that can be drawn from each synthetic population is not large, we compute the exact Σ and the exact mean squared error of all estimators considered to facilitate comparison. To generate the first set of rival estimators, we consider certain subclasses of the broader class of linear combinations of month-in-sample estimates. For each subclass, when Σ is known, the optimum estimator is obtained as a function of Σ. An estimated optimal estimator in each subclass is then obtained from the corresponding optimal estimator when Σ is replaced by an estimator. Neither the AK estimator nor the estimated optimal estimators for these subclasses performed well in our evaluation study. In our real life data analysis, the AK estimates are constantly below the survey-weighted estimates, indicating potential bias. Our study indicates limitations of the approach that generate an estimated optimal estimator by first obtaining the optimal estimator in a class of linear combination of Y and then substituting in the optimal estimator an estimate of Σ.

Any attempt to improve on the estimated optimal estimator in any given class would require a thorough investigation of the highly non-trivial problem of estimation

[1]University of Cambridge. UK. E-mail: dbb31@cam.ac.uk.
ORCID: https://orcid.org/0000-0001-8582-7856.
[2]US Census Bureau. USA. E-mail: yang.cheng@verizon.net.
[3]JPSM, University of Maryland. USA. E-mail: plahiri@umd.edu.
ORCID: https://orcid.org/0000-0002-7103-545X.

of Σ for a complex setting like the CPS. We have not discussed this problem in this paper. Instead, we adapted the regression composite estimator used by Statistics Canada in the CPS setting. Unlike the estimated optimal estimators, the regression composite estimator does not require estimation of Σ and is less sensitive to the rotation group bias in our simulation study. Our study indicates that there is a great potential for regression composite estimation technique in improving estimation of both levels and month-to-month changes in the unemployment rates.

Key words: calibration, estimated controls, longitudinal survey, labor force statistics.

1. Introduction

In repeated surveys, including rotating panel surveys, statistical data integration plays an important role in producing efficient estimators by extracting relevant information over time. To this end, various composite estimators have been proposed; see Jones (1980), Yansaneh and Fuller (1998), Bell (2001), Singh et al. (2001), Fuller and Rao (2001) and others. Such composite estimators typically improve on the standard direct survey-weighted estimators in terms of mean squared error (MSE) and are commonly used by different government agencies for producing official labor force statistics. For example, to produce national employment and unemployment levels and rates, the U.S. Census Bureau uses the AK composite estimation technique developed using the ideas given in Gurney and Daly (1965).

Motivated by a Statistics Canada application, Singh and Merkouris (1995) introduced an ingenious idea for generating a composite estimator that can be computed using Statistics Canada's existing software for computing generalized regression estimates. The key idea in Singh and Merkouris (1995) is to create a proxy (auxiliary) variable that uses information at the individual level as well as estimates at the population level from both previous and current periods. Using this proxy variable, Singh and Merkouris (1995) obtained a composite estimator, referred to as Modified Regression 1 estimator (MR1) in the literature. However, Singh et al. (1997) noted that MR1 does not perform well in estimating changes in labor force statistics, which motivated them to propose a different composite estimator, called MR2, using a new proxy variable. Singh et al. (2001) generalized the idea of MR1 and MR2 estimators by suggesting a general set of proxy variables.

Fuller and Rao (2001) noted that the regression composite estimator proposed by Singh et al. (1997) is subject to an undesirable drift problem, i.e., it may produce estimates that drift away from the real value suggested by the underlying model as time progresses. They proposed an alternative regression composite method to rectify the drift problem. Their method differs from the method of Singh et al. (2001) in two directions. First, the idea of rectifying the drift problem by a weighted combination of the two proxy variables used for MR1 and MR2 is new. Secondly, their final regression composite estimator involves estimation of the weight assigned to MR1 or MR2 control variable in the weighted combination — this idea was not discussed in Singh et al. (2001). In short, the Fuller-Rao regression composite estimator with estimated weight cannot be viewed as a special case of Singh et al. (2001) and vice versa.

Gambino et al. (2001) conducted an empirical study to evaluate the Fuller-Rao regression composite estimator, offered missing value treatment and listed several advantages (e.g. weighting procedure, consistency, efficiency gain, etc.) of the Fuller-Rao regression composite estimator over the AK estimator. Statistics Canada now uses the Fuller-Rao method for their official labor force statistics production. Salonen (2007) conducted an empirical study to compare the currently used Finnish labor force estimator with the Fuller-Rao's regression composite and other estimators. Bell (2001) applied the generalized regression technique to improve on the Best Linear Unbiased Estimator (BLUE) based on a fixed window of time points and compared his estimator with the AK composite estimator of Gurney and Daly (1965) and with the modified regression estimator of Singh et al. (1997), using data from the Australian Labour Force Survey. Beaumont and Bocci (2005) proposed a regression composite estimator with missing covariates defined using variables of interest from the previous month.

The main goal of this paper is to compare the design-based properties of the AK estimator with different rival estimators using the CPS data. To this end, we first expand the list of potential estimators by considering two new classes of composite estimators. The first class includes the AK estimator as a member. The second class generalizes the class of estimators considered earlier by Yansaneh and Fuller (1998) to incorporate multiple categories of employment status (e.g., employed, unemployed, and not in the labor force). We obtain the best linear unbiased estimator (BLUE) for each class of estimators. We call them the best AK estimator and multivariate BLUE, respectively. As special cases of the multivariate BLUE, one can generate the univariate BLUE and the best AK estimator. If the covariance matrix between two vectors of observations corresponding to any two different variables is a null matrix, then multivariate BLUE is identical to the univariate BLUE when the design matrix is the same for the variables. However, in general they are not identical when we do not have a block-diagonal covariance structure as is the case in our problem.

The optimal estimator for a given class of estimators, derived under given model and optimality condition, cannot be used as it involves unknown model parameters (e.g., variances and covariances). The AK estimator used by the Census Bureau is obtained from the optimal estimator when variances and covariances are substituted by estimators justified under a rather strong stationary assumption. We devise an evaluation study in order to assess the exact design-based properties of different composite estimators using the CPS data and CPS sample design. We demonstrate that the optimal estimator for a given model with estimated variances and covariances can perform poorly even when the modeling assumptions are valid. We included the multivariate BLUE with estimated variances and covariances for completeness of this research. While the multivariate BLUE performs the best under the model that generates it, it performed worse than the univariate BLUE with estimated variances and covariances. Overall, we found that the Fuller-Rao estimator performed the best among all composite estimators considered in our study.

In Section 2, we discuss the population and sample design. In Section 3, we review different classes of estimators and the optimal estimator within each class. In Section 4, we describe our evaluation study to assess the design-based properties of different

estimators. In Section 5, we report the CPS data analysis. Some discussion and future research topics are given in Section 6. We defer the proofs of relevant results and description of CPS design to the Appendix. To facilitate reading of the paper, we list all the notation used in the paper in the appendix.

2. Notations

2.1. Population

Our theoretical framework uses three indices to identify three dimensions: m for month, k for individual and e for an employment status category. In this paper, we will consider three categories of employment status: employed, unemployed and not in the labor force. The theory and methods developed in this paper, however, extend to more than 3 categories of employment status. Consider a sequence of finite populations of individuals $(U_m)_{m\in\{1...M\}}$, where U_m refers to the finite population for month m. Let N denote the cardinality of $U=\bigcup_{m=1}^{M}U_m$. Let $\mathbf{y}_{m,k,e}=1$ if the kth individual belongs to U_m and has eth employment status and $\mathbf{y}_{m,k,e}=0$ otherwise, $m\in\{1,\cdots,M\}$, $k\in\{1,\cdots,N\}$, $e\in\{1,2,3\}$. Because of our three dimensional data structure, we find it convenient to introduce arrays in developing our methodology and theory. Let $\mathbf{y}=[\mathbf{y}_{m,k,e}]_{m\in\{1,...,M\},k\in\{1,...,N\},e\in\{1,2,3\}}$ denote a three dimensional $(M,N,3)$-sized array. We also define $\mathbf{x}$ as a 3-dimensional array of auxiliary variables indexed by month, individual and auxiliary variable, and an array $\mathbf{z}$, indexed the same way, which contains endogenous variables in the sense that $\mathbf{z}$ is a function of $\mathbf{x}$ and $\mathbf{y}$. Any element of an array with (m,k)-index satisfying $k\notin U_m$ is equal to 0 by convention.

2.2. Notational conventions on arrays

Given subsets A, B, C of $\{1,\ldots,M\}$, $\{1,\ldots,N\}$, $\{1,2,3\}$, respectively (including the full set), we use the following notation for sub-arrays: $\mathbf{y}_{A,B,C}=[\mathbf{y}_{a,b,c}]_{a\in A,b\in B,c\in C}$, and may replace A, B, or C by "." when $A=\{1,\ldots,M\}$, $B=\{1,\ldots,N\}$ or $C=\{1,2,3\}$, respectively: for example, $\mathbf{y}=\mathbf{y}_{.,.,.}$. Let $\mathbf{t}_{\mathbf{y}}=\left[\sum_{k\in U}\mathbf{y}_{m,k,e}\right]_{m\in\{1,...,M\},e\in\{1,2,3\}}$ be the two dimensional $(M,3)$-sized array of population totals indexed by month m and employment status e. We now show we can form a vector or matrix from an array. For a p-dimensional $(a_1,\ldots,a_p)$-sized array A, define $\vec{A}$ as the vector $\left(\vec{A}_1,\ldots,\vec{A}_{\prod_{l=1}^{p}a_l}\right)$, where $\forall(i_1,\ldots,i_p)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $\vec{A}_{1+\sum_{l=1}^{p}\left[\prod_{l'<l}(a_{l'}-1)i_1\right]}=A_{i_1,\ldots,i_p}$, with the convention that a product over the empty set equals 1. By convention, when an array B is defined as an $((a_1,\ldots,a_p),(b_1,\ldots,b_q))$-sized array (with two vector of indexes), $\vec{A}$ is the matrix $\left[\vec{A}_{i,j}\right]_{i\in\{1,\ldots,\prod_{l=1}^{p}a_l\},j\in\{1,\ldots,\prod_{l=1}^{q}b_l\}}$ such that $\forall(i_1,\ldots,i_p)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $(j_1,\ldots,j_q)\in\prod_{l=1}^{p}\{1,\ldots,a_l\}$, $\vec{A}_{1+\sum_{l=1}^{p}\left[(i_l-1)\prod_{l'<l}(a_{l'})\right],1+\sum_{l=1}^{q}\left[(j_l-1)\prod_{l'<l}(b_{l'})\right]}=A_{(i_1,\ldots,i_p),(j_1,\ldots,j_p)}$. Given A an $((a_1,\ldots,a_n),(b_1,\ldots b_l))$ array and B a $((b_1,\ldots,b_l),(c_1,\ldots c_p))$ array, $C=A\times B$ is the $((a_1,\ldots,a_n),(c_1,\ldots c_p))$ array defined by $C_{(i_1,\ldots,i_n),(k_1,\ldots,k_n)}=\sum_{j_1,\ldots,j_l}A_{(i_1,\ldots,i_n),(j_1,\ldots,j_l)}B_{(j_1,\ldots,j_l),(k_1,\ldots,k_n)}$.

2.3. The sample design

The CPS monthly sample comprises about 72,000 housing units and is collected for 729 areas (Primary Sampling Units) consisting of more than 1,000 counties covering every state and the District of Columbia. The CPS, conducted by the Census Bureau, uses a 4-8-4 rotating panel design. For any given month, the CPS sample can be grouped into eight subsamples corresponding to the eight rotation groups. All the units belonging to a particular rotating panel enter and leave the sample at the same time. A given rotating panel (or group) stays in the sample for four consecutive months, leaves the sample for the eight succeeding months, and then returns for another four consecutive months. It is then dropped from the sample completely and is replaced by a group of nearby households. Of the two new rotation groups that are sampled each month, one is completely new (their first appearance in the panel) and the other is a returning group, which has been out of the sample for eight months. Thus, in the CPS design, six out of the eight rotation groups are common between two consecutive months (i.e., 75% overlap), and four out of eight are common between the same month of two consecutive years (i.e., 50% overlap) respectively; see Hansen et al. (1955). For month m, let S_m denote the sample of respondents. Let $S_{m,g}$ denote the set of sampled respondents in the gth sample rotation group for month m and $S_m = \bigcup_{g=1}^{8} S_{m,g}$. For a given month m, the rotation groups $S_{m,g}$, $g = 1,\ldots,8$ are indexed so that g indicates the number of times that rotation group $S_{m,g}$ has been a part of the sample in month m and before. In the US Census Bureau terminology, g is referred to as the month-in-sample (mis) index and $S_{m,g}$ as the month-in-sample g rotation group (more details on this design are given in Section 4.3). We adopt a design-based approach in this study in which variables $\mathbf{x}$ and $\mathbf{y}$ are considered fixed parameters of the underlying fixed population model for design-based inference (Cassel et al., 1977, p. 2).

3. Estimation

3.1. Direct and month-in-sample estimators

Let $\mathbf{w}_{m,k}$ denote the second stage weight of individual k in month m, obtained from the basic weight (that is, the reciprocal of the inclusion probability) after standard non-response and post-stratification adjustments. By convention, $\mathbf{w}_{m,k} = 0$ if $k \notin S_m$. Let $\mathbf{w}$ be the (M,N)-sized array indexed by m and k of $\mathbf{w}_{m,k}$. We refer to CPS Technical Paper (2006) for a detailed account of weight construction. The array of direct survey-weighted estimator of $\mathrm{t}_{\mathbf{y}}$ is given by $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}} = \left[\sum_{k\in S_m} \mathbf{w}_{m,k}\mathbf{y}_{m,k,e}\right]_{m\in\{1,\ldots,M\},e\in\{1,2,3\}}$. Define the $(M,8,3)$-sized array of month-in-sample estimates: $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}} = \left[8\times\sum_{k\in S_{m,g}} \mathbf{w}_{m,k}\mathbf{y}_{m,k,e}\right]_{m\in\{1,\ldots,M\},g\in\{1,\ldots,8\},e\in\{1,2,3\}}$. For a month-in-sample number g, $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{.,g,.}$ is called the month-in-sample g estimator of $\mathrm{t}_{\mathbf{y}}$.

3.2. An extended Bailar model for the rotation group bias

Because of differential non-response and measurement errors across different rotation groups, the direct and month-in-sample estimators are subject to a bias, commonly referred to as the rotation group bias. Bailar (1975) proposed a class of semi-parametric models on the expected values of the month-in-sample estimators. Under a model in this class, (i) the bias of each month-in-sample estimator of total of unemployed depends on the month-in-sample index g only, (ii) the bias is invariant with time, and (iii) the vector of month-in-sample biases are bounded by a known linear constraint (without this binding linear constraint, month-in-sample rotation group biases could only be estimated up to an additive constant). Note that these very strong assumptions were made in order to reveal the existence of what is known as the rotation group bias in US Census Bureau terminology. It would be highly questionable to use this model for rotation group bias correction because (i) the choice of the linear constraint would be totally arbitrary in the absence of a re-interview experiment and (ii) the stationarity assumptions are unreasonable. We propose the following model in order to extend the Bailar model to account for the rotation group biases of the multiple categories:

$$\mathrm{E}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m,g,e}\right] = (\mathrm{t}_{\mathbf{y}})_{m,e} + b_{g,e}, \tag{1}$$

where b is a two-dimensional $(8,p)$-sized array of biases such that $\forall e, C_e b_{.,e} = 0$, C_1, C_2, C_3 being known linear forms satisfying $C_e\,(1,\ldots,1)^{\mathrm{T}} \neq 0$.

3.3. Estimation of unemployment rate and variance approximation

We define the function $\mathrm{R} : (0,+\infty)^3 \to [0,1], x \mapsto x_2/(x_1+x_2)$. By convention, when applied to an array with employment status as an index, x_1, x_2 denote the subarrays for employment status 1 and 2, respectively, and / denotes the term by term division. The unemployment rate vector is defined as $\mathrm{r} = \mathrm{R}\,(\mathrm{t}_{\mathbf{y}}) = (\mathrm{t}_{\mathbf{y}})_{.,1} / \left((\mathrm{t}_{\mathbf{y}})_{.,1} + (\mathrm{t}_{\mathbf{y}})_{.,2}\right)$.
Given an estimator $\hat{\mathrm{t}}_{\mathbf{y}}^{\star}$ of $\mathrm{t}_{\mathbf{y}}$, we derive the following estimator of r from $\hat{\mathrm{t}}_{\mathbf{y}}^{\star}$: $\hat{\mathrm{r}}^{\star} = \mathrm{R}(\hat{\mathrm{t}}_{\mathbf{y}}^{\star})$. Using the linearization technique, we can approximate the variance $\mathrm{Var}[\hat{\mathrm{r}}_m^{\star}]$ of the unemployment rate estimator for month m by $J_1 \mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m,.}\right] J_1{}^{\mathrm{T}}$, where J_1 is the Jacobian matrix: $J_1 = \left(\frac{\mathrm{d}\ \mathrm{R}(t)}{\mathrm{d}\ t}\right)\left((\mathrm{t}_{\mathbf{y}})_{m,.}^{\star}\right) = \left[(\mathrm{t}_{\mathbf{y}})_{m,1}^{-1}, -(\mathrm{t}_{\mathbf{y}})_{m,1}\,(\mathrm{t}_{\mathbf{y}})_{m,2}^{-2}, 0\right]$, and the variance of the estimator of change of the employment rate between two consecutive months by $J_2 \mathrm{Var}\left[\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m,.}, \left(\hat{\mathrm{t}}_{\mathbf{y}}^{\star}\right)_{m-1,.}\right)\right] J_2{}^{\mathrm{T}}$, where

$$\begin{aligned} J_2 &= \left(\frac{\mathrm{d}\ \mathrm{R}(t) - \mathrm{R}(t')}{\mathrm{d}\ (t,t')}\left((\mathrm{t}_{\mathbf{y}})_{m,.}, (\mathrm{t}_{\mathbf{y}})_{m-1,.}\right)\right) \\ &= \left[(\mathrm{t}_{\mathbf{y}})_{m,1}^{-1}, -(\mathrm{t}_{\mathbf{y}})_{m,1}\,(\mathrm{t}_{\mathbf{y}})_{m,2}^{-2}, 0, -(\mathrm{t}_{\mathbf{y}})_{m-1,1}^{-1}, (\mathrm{t}_{\mathbf{y}})_{m-1,1}\left((\mathrm{t}_{\mathbf{y}})_{m-1,2}\right)^{-2}, 0\right]. \end{aligned}$$

3.4. The class of linear combinations of month-in-sample estimators

Here, as in Yansaneh and Fuller (1998), we consider the best estimator of counts by employment status in the class of linear combinations of month-in-sample estimators. Generalizing Yansaneh and Fuller (1998), the unbiasedness assumption of all month-in-sample estimators is:

$$\mathrm{E}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = \vec{X}\vec{\mathrm{t}}_{\mathbf{y}}, \tag{2}$$

where X is the $((M,8,3),(M,3))$-sized array with rows indexed by the triplet (m,g,e) and columns indexed by the couple (m,e) such that $X_{(m,g,e),(m',e')} = 1$ if $m' = m$ and $e' = e$, 0 otherwise. Let L be a $(p,(M,3))$-sized array with $p \in \mathbb{N}\setminus\{0\}$ and rows indexed by (m,e). By class of linear estimators of $L\mathrm{t}_{\mathbf{y}}$, we will designate the class of estimators that are linear combinations of the month-in-sample estimators, i.e., of the form $W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$, where W is a fixed (does not depend on the observations) $(p,(M\times 8\times 3))$-sized matrix.

Best linear estimator

Let $\Sigma_{\mathbf{y}} = \mathrm{Var}_{\mathbf{y}}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right]$. In the design-based approach, $\Sigma_{\mathbf{y}}$ is a function of the finite population $\mathbf{y}$. The variance of a linear transformation $W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$ of $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}$ is: $\mathrm{Var}\left[W\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = W^T\Sigma_{\mathbf{y}}W$. When month-in-sample estimators are unbiased, $\Sigma_{\mathbf{y}}$ is known, and only $\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$ is observed, and $\vec{X}^{+}\vec{X} = I$, the Gauss-Markov theorem states that the BLUE of $\mathrm{t}_{\mathbf{y}}$ uniformly in t_y is the $(M,3)$-sized matrix $\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{BLUE}}$ defined by

$$\vec{X}^{+}(\vec{X}\vec{X}^{+})\left(I-\Sigma_{\mathbf{y}}((I-\vec{X}\vec{X}^{+})^{+}\Sigma_{\mathbf{y}}(I-\vec{X}\vec{X}^{+}))^{+}\right)\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}, \tag{3}$$

where the $^{+}$ operator denotes the Moore-Penrose pseudo inversion, I is the identity matrix. Here the minimization is with respect to the order on the space of symmetric positive definite matrices: $M_1 \leq M_2 \Leftrightarrow M_2 - M_1$ is positive. It can be shown that $\vec{X}^{+} = \vec{X}^{\mathrm{T}}/8$ in our case and that $\vec{X}^{+}\vec{X} = I$. For more details about the Gauss-Markov result under singular linear model, one may refer to (Searle, 1994, p. 140, Eq. 3b). This is a generalization of the result of Yansaneh and Fuller (1998), as it takes into account the multi-dimensions of $\mathbf{y}$ and non-invertibility of $\Sigma_{\mathbf{y}}$. Note that $\Sigma_{\mathbf{y}}$ can be non-invertible, especially when the sample is calibrated to a given fixed population size, considered non-random, because of an affine relationship between month-in-sample estimates (e.g., $\sum_{g=1}^{8}\sum_{e=1}^{3}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m,g,e}$ is not random).

We recall the following:

(i) For any linear transformation L applicable to $\vec{\mathrm{t}}_{\mathbf{y}}$, the best linear unbiased estimator of $L\vec{\mathrm{t}}_{\mathbf{y}}$ uniformly in $\mathrm{t}_{\mathbf{y}}$ is $L\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{BLUE}}$, which ensures that the BLUE of month-to-month change can be simply obtained from the BLUE of level. Thus, there is no need for searching a compromise between estimation of level and change.

(ii) For any linear transformation L applicable to $\vec{\mathrm{t}}_{\mathbf{y}}$, any linear transformation J applicable to $L\vec{\mathrm{t}}_{\mathbf{y}}$, $L\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{BLUE}} \in \mathrm{argmin}\left\{JW\Sigma_{\mathbf{y}}(JW)^{\mathrm{T}}\middle| W, W\vec{X} = L\right\}$.Thus, the plug-in esti-

mators for unemployment rate and month-to-month unemployment rate change derived from the BLUE are also optimal in the sense that they minimize the linearized approximation of the variance of such plug-in estimators, which can be written in the form $JW\Sigma_{\mathbf{y}}(JW)^{\mathrm{T}}$.

Remark: BLUE under Bailar rotation bias model

Here, we give the expression of the BLUE under the general Bailar rotation bias model. Bailar's rotation bias model can be written in the following matrix notation:

$$\mathrm{E}\left[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}\right] = \vec{X}\vec{\mathrm{t}}_{\mathbf{y}} + \vec{X}'\vec{b}, \tag{4}$$

where X' is a fixed known array; see also Yansaneh and Fuller (1998, equation 8). For example under Model (1), with $C_1 = C_2 = C_3 = (1,\ldots,1)$, X' is the $((M,8,3),(7,2))$-sized array such that for $m \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, $g' \in \{1,\ldots,7\}, e \in \{1,2,3\}$, $e' \in \{1,2,3\}$, $X'_{(m,g,e),(g',e')} = 1$ if $g = g' < 8$ and $e = e'$, -1 if $g = 8$ and $e = e'$, 0 otherwise. We can reparametrize Model (4) as $\mathrm{E}[\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}] = X^{\star}\mu$, where $X^{\star} = [\vec{X} \mid \vec{X}']$, and the parameter $\mu = [\vec{\mathrm{t}}_{\mathbf{y}} \mid \vec{b}]^{\mathrm{T}}$. The best linear unbiased estimator of $\vec{\mathrm{t}}_{\mathbf{y}}$ under this rotation bias model is given by

$$LX^{\star+}(X^{\star}X^{\star+})\left(I - \Sigma_{\mathbf{y}}(I - X^{\star}X^{\star+})^{+}\Sigma_{\mathbf{y}}(I - X^{\star}X^{\star+}))^{+}\right)\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}},$$

with L satisfying $LX^{\star} = \vec{X}$. This is a generalization of Yansaneh and Fuller (1998) because it (i) considers non-invertible $\Sigma_{\mathbf{y}}$, (ii) does not limit to a unidimensional variable and (iii) is generalized to general Bailar's model.

3.5. AK composite estimation

Definition

We define a general class of AK composite estimators. Let $A = \mathrm{diag}(a_1, a_2, a_3)$ and $K = \mathrm{diag}(k_1, k_2, k_3)$ denote two diagonal matrices of dimension 3. The AK estimator with coefficients A and K is defined as follows: first define $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{1,.} = \left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}}\right)_{1,.}$, then recursively define for $m \in 2,\ldots,M$,

$$\begin{aligned}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.} &= (I-K)\times\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{direct}}\right)_{m,.} \\ &+ K\times\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m-1,.} + \frac{4}{3}\sum_{k\in S_m\cap S_{m-1}}\left(\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.} - \mathbf{w}_{m-1,k,.}\mathbf{y}_{m-1,k,.}\right)\right) \\ &+ A\times\left(\sum_{k\in S_m\setminus S_{m-1}}\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.} - \frac{1}{3}\sum_{k\in S_m\cap S_{m-1}}\mathbf{w}_{m,k,.}\mathbf{y}_{m,k,.}\right),\end{aligned} \tag{5}$$

where $\setminus$ denotes the set difference operator and I is the identity matrix of dimension 3. The sum of the first two terms of the AK estimator is indeed a weighted average of the

current month direct estimator and the previous month AK estimator suitably updated for the change. The last term of the AK estimator is correlated to the previous terms and has an expectation 0 with respect to the sample design. Gurney and Daly (1965) explained the benefits of adding the third term in reducing the mean squared error. The Census Bureau uses specific values of A and K, which were empirically determined in order to arrive at a compromise solution that worked reasonably well for both employment level and rate estimation; see, e.g., Lent et al. (1999). The corresponding unemployment rate estimator is obtained as: $\hat{\mathrm{r}}_m^{\mathrm{AK}} = \mathrm{R}\left(\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.}\right)$. Note that $\hat{\mathrm{r}}_m^{\mathrm{AK}}$ depends on a_1, a_2, k_1, k_2, but not on a_3 and k_3. Note that the class of AK estimators is a sub class of the class of linear estimators, as the AK estimator can be written as a linear combination of the month-in-sample estimators: $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,.} = \sum_{m'=1}^{m}\sum_{g=1}^{8} c_{m,m',g}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis}}\right)_{m',g,.}$, where the $(3,3)$ matrices $c_{m,m,g}$ are defined recursively: $\forall g \in \{1,\ldots,8\}, c_{1,1,g} = (1/8)\times I$ and

$$\forall m \in \{2,\ldots,M\}, \begin{cases} \forall g \in \{1,5\} & c_{m,m,g} & = ((I-K)+A)/8 \\ \forall g \in \{2,3,4,6,7,8\} & c_{m,m,g} & = ((I-K)+4K/3-A/3)/8 \\ \forall g \in \{1,2,3,5,6,7\} & c_{m,m-1,g} & = c_{m-1,m-1,g}\times K-(4K/3)/8 \\ \forall g \in \{4,8\} & c_{m,m-1,g} & = c_{m-1,m-1,g}\times K \\ \forall 1 \le m' < m-1 & c_{m,m',g} & = c_{m-1,m',g}\times K \end{cases} \tag{6}$$

$\forall m' > m, g \in \{1,\ldots,8\}, c_{m,m',g} = 0.$

Let W^{AK} be the $((M,3),(M,8,3))$ array such that for $m,m' \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, $e,e' \in \{1,2,3\}$, $W^{\mathrm{AK}}_{(m,e),(m',g,e')} = c_{m,m',g}$ if $e = e'$, 0 otherwise. Then $\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{AK}} = \vec{W}^{\mathrm{AK}}\vec{\hat{\mathrm{t}}}_{\mathbf{y}}^{\mathrm{mis}}$.

Notes on AK estimator

In presence of rotation bias, the bias of the AK estimator is $\vec{W}^{\mathrm{AK}}\vec{X}'\vec{b}$, which may not be equal to 0. Depending on the rotation bias model, an unbiased version of the AK estimator may not exist. Furthermore, contrary to the BLUE, the best A and K coefficients for estimation of one particular month and status may not be optimal for another month and status. Moreover, the best A and K coefficients for estimation of level may not be optimal for estimation of change. For example, it is possible to find A,K,m,e,A',K',m',e' such that $\mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}}\right)_{m,e}\right] < \mathrm{Var}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{A}',\mathrm{K}'}\right)_{m,e}\right]$ and $\mathrm{Var}\left[\hat{\mathrm{t}}_{\mathbf{y}_{m',e'}}^{\mathrm{AK}}\right] > \mathrm{Var}\left[\hat{\mathrm{t}}_{\mathbf{y}_{m',e'}}^{\mathrm{A}',\mathrm{K}'}\right]$.

When $\Sigma_{\mathbf{y}}$ is known, let $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,level}$ and $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,change}$ denote the AK estimators obtained by minimizing (with respect to A and K) the average approximated variance of level estimators $\sum_{m=1}^{M} J_1 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{m,.}\right] J_1{}^{\mathrm{T}}$ and of change estimators $\sum_{m=1}^{M} J_2 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{\{m-1,m\},.}\right] J_2{}^{\mathrm{T}}$, respectively; let $\hat{\mathrm{t}}_{\mathbf{y}}^{BAK,compromise}$ denote the AK estimator obtained by minimizing the averaged variance

$\sum_{m=1}^{M}\left(J_1 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{m,.}\right] J_1{}^{\mathrm{T}} + J_2 \mathrm{Var}_{\mathbf{y}}\left[\left(\hat{\mathrm{t}}_{\mathbf{y}}^{A,K}\right)_{\{m-1,m\},.}\right] J_2{}^{\mathrm{T}}\right)$. For AK estimation, note that the three objective functions are polynomial functions of A and K whose coefficients are functions of $\Sigma_{\mathbf{y}}$. By using a standard numerical method (Nelder-Mead) we can obtain the optimal coefficients.

3.6. Empirical best linear estimator and empirical best AK estimator.

Let $\hat{\Sigma}$ be an estimator of $\Sigma_{\mathbf{y}}$, and let $\hat{\mathrm{t}}_{\mathbf{y}}^{EBLUE}$ be the estimator of $\mathrm{t}_{\mathbf{y}}$ obtained from (3) when $\Sigma_{\mathbf{y}}$ is replaced by $\hat{\Sigma}$. In the same manner, we can define the empirical best AK estimators for change, level and compromise. For the CPS, optimal A and K coefficients were determined so that a compromise objective function, accounting for the variances of the month-to-month change and level estimates, would be minimum. The variances were estimated under the assumption of a stationary covariance of month-in-sample estimators; see Lent and Cantwell (1996). The method used in the Census Bureau consists in choosing the best coefficients a_1, a_2, k_1, k_2 on a grid with 9 possible values for each coefficient $(0.1,\ldots,0.9)$.

3.7. Regression Composite Estimation

In this section we elaborate on the general definition of the class of regression composite estimators parametrized by a real number $\alpha \in [0,1]$ as proposed by Fuller and Rao (2001). This class includes regression composite estimators MR1 (for $\alpha = 0$) and MR2 (for $\alpha = 1$) as defined by Singh and Merkouris (1995) and Singh et al. (2001). For $\alpha \in [0,1]$, the regression composite estimator of $\mathrm{t}_{\mathbf{y}}$ is a calibration estimator $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,.}$ defined as follows: provide calibration totals $\left(\mathrm{t}_{\mathbf{x}}^{adj}\right)_{m,.}$ for the auxiliary variables (they can be equal to the true totals when known or estimated), then define $\left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{1,.} = \left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{direct}}\right)_{1,.}$, and $\mathbf{w}_{1,k}^{\mathrm{r.c.},\alpha} = \mathbf{w}_{1,k}$ if $k \in S_1$, 0 otherwise. For $m \in \{2,\ldots,M\}$, recursively define

$$\mathbf{z}_{m,k,.}^{\mathrm{r.c.}(\alpha)} = \begin{cases} \alpha\left(\tau_m^{-1}\left(\mathbf{z}_{m-1,k,.} - \mathbf{z}_{m,k,.}\right) + \mathbf{z}_{m,k,.}\right) + (1-\alpha)\ \mathbf{z}_{m-1,k,.} & \text{if } k \in S_m \cap S_{m-1}, \\ \alpha\ \mathbf{z}_{m,k,.} + (1-\alpha)\ \left(\sum_{k\in S_{m-1}} \mathbf{w}_{m-1,k}^{\mathrm{r.c.},\alpha}\right)^{-1}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{c}}\right)_{m-1,.} & \text{if } k \in S_m \setminus S_{m-1}, \end{cases} \tag{7}$$

where $\tau_m = \left(\sum_{k\in S_m\cap S_{m-1}} \mathbf{w}_{m,k}\right)^{-1} \sum_{k\in S_m} \mathbf{w}_{m,k}$. Then the regression composite estimator of $(\mathrm{t}_{\mathbf{y}})_{m,.}$ is given by $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,.} = \sum_{k\in S_m} \mathbf{w}_{m,k}^{\mathrm{r.c.},\alpha} \mathbf{y}_{m,k}$, where

$$\left(\mathbf{w}_{m,.}^{\mathrm{r.c.},\alpha}\right) = \mathrm{argmin}\left\{ \sum_{k\in U} \frac{\left(\mathbf{w}_k^{\star} - \mathbf{w}_{m,k}\right)^2}{\mathbb{1}(k \notin S_m) + \mathbf{w}_{m,k}} \,\middle|\, \mathbf{w}^{\star} \in \mathbb{R}^U,\ \begin{array}{l} \sum_{k\in S_m} \mathbf{w}_k^{\star} \mathbf{z}_{m,k,.}^{\mathrm{r.c.}(\alpha)} = \left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{m-1,.} \\ \sum_{k\in S_m} \mathbf{w}_k^{\star} \mathbf{x}_{m,k,.} = \left(\mathrm{t}_{\mathbf{x}}^{adj}\right)_{m,.} \end{array} \right\}, \tag{8}$$

and $\left(\hat{\mathrm{t}}_{\mathbf{z}}^{\mathrm{r.c.},\alpha}\right)_{m,.} = \sum_{k\in S_m} \mathbf{w}_{m,k}^{\mathrm{r.c.},\alpha} \mathbf{z}_{m,k}^{\mathrm{r.c.}(\alpha)}$, where $\mathbb{1}(k \notin S_m) = 1$ if $k \notin S_m$ and 0 otherwise. Our definition of regression composite estimator is more general than the one in Fuller and Rao (2001) as it takes into account a multivariate version of $\mathbf{y}$. Modified Regression 3 (MR3) of Gambino et al. (2001) does not belong to the class of regression composite

estimators. The MR3 estimator imposes too many constraints in the calibration procedure, which leads to a high variability of the calibration weights; consequently, MR3 estimator has a larger MSE than composite regression estimators.

Choice of z and choice of α

Fuller and Rao (2001) studied the properties of the estimator $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{r.c.},\alpha}\right)_{m,1}$ for the choice of $\mathbf{z} = \mathbf{y}_{.,1}$. As the employment rate is a function of $\mathbf{y}_{m,1}$ and $\mathbf{y}_{m,2}$, we investigate the properties of Regression Composite Estimator with the choice $\mathbf{z} = \mathbf{y}$. Fuller and Rao (2001) proposed a method that allows for an approximation to the optimal α coefficient for month-to-month change and level estimation, under a specific individual level superpopulation model for continuous variables. They proposed this superpopulation model to explain the drift problem of MR2 (regression composite estimator for $\alpha = 1$) and obtained the best coefficient α. Since we deal with a discrete multidimensional variable, the continuous superpopulation model assumed by Fuller and Rao (2001) is not appropriate in our situation. It will be interesting to propose an approach to estimate the best α in our situation. For our preliminary study, we examine a range of known α values in our simulations and in the CPS data analysis.

4. Simulation Experiment

4.1. Description of Simulation Study

We conducted a simulation study to enhance our understanding of the finite sample properties of different composite estimators. We generated three synthetic finite populations, each with size 100,000. In order to make the simulation experiment meaningful, we generated employment statuses for each finite population in a manner that attempts to capture the actual U.S. national employment rate dynamics during the study period 2005-2012. Moreover, in order to understand the maximum gain from the composite estimation, we induced high correlation in the employment statuses between two consecutive months subject to a constraint on the global employment rate evolution. We set the probability of month-to-month changes in employment statuses for an individual to zero in case of no change in the corresponding direct national employment rates. Samples were selected according to a rotating design with systematic selection that mimics the CPS sample design. Since the number of possible samples is only 1000, we are able to compute the exact design-based bias, variance and mean squared error of different estimators, and subsequently, the optimal linear and optimal AK estimators. We compute employment rate, total employed, and total unemployed over the 85-month period using the direct, AK and the Fuller-Rao regression composite methods. We then compare the optimal estimator in the class of regression composite estimators to those in the class of the AK and best linear estimators. Note that the simulation study can be reproduced using the R package we created for this purpose; see Bonnéry (2016c).

4.2. Populations generation

We created three synthetic populations each with $N = 100,000$ individuals indexed by $1,\ldots,N$. For individual k of each population, we created a time series $(\mathbf{y}_{m,k,.})_{m\in 1,\ldots,M}$, where $\mathbf{y}_{m,k,.} \in \{(1,0,0),(0,1,0),(0,0,1)\}$ (for unemployed, not in labor force, employed), and $M = 85$. Each individual belongs to one household consisting of 5 individuals. The number of all households is $H = 20,000$; the set of all households is given by $\{h_i = \{(5\times(i-1)+1),\ldots,(5\times i)\} \mid i = 1,\ldots,H\}$. We created time series data under certain constraints at the population level. For each population, unemployment rates are the same as the direct estimates obtained from the CPS data. In population 1, the proportion of people who change status between two consecutive months is minimal. In populations 2 and 3, the proportions of persons who change from one status to another between two consecutive months are equal to those estimated from the CPS data. In population 2, a person with a small index has a higher probability to change status. In population 3, the probability to change status between two consecutive months is the same for all individuals with the same status.

4.3. Repeated design

We mimic the CPS design, which is described in appendix A. For month m, a sample S_m is the union of 8 rotation groups. The design and the creation of rotation groups are explained below. Rotation groups are made of $n = 20$ households with a total of 100 individuals. So for each month m, there are $\#(S_m) = 800$ individuals in the sample, and the inclusion probability of any unit is $1/125$. The selection of longitudinal sample $S_1,\ldots S_m$ is made in 3 steps:

1. Draw an integer number η between 1 and 1,000 from a uniform distribution.
2. For $\ell \in 1,\ldots,(M+15)$, create a cluster of households $\mathrm{Clu}_\ell = \bigcup_{j=1}^{n} h_{i_{\ell,j}}$, where $i_{\ell,j} = \mathrm{rem}\left((r-1+\ell-1)+\frac{H}{n}\times(j-1),H\right)+1$, and $\mathrm{rem}(a,b)$ denotes the remainder of the Euclidean division of a by b.
3. Let $\delta_1 = 0, \delta_2 = 1,\ \delta_3 = 2,\ \delta_4 = 3,\ \delta_5 = 12,\ \delta_6 = 13,\ \delta_7 = 14,\ \delta_8 = 15$. For $m \in \{1,\ldots,M\}$, $g \in \{1,\ldots,8\}$, create the samples $S_{m,g} = \mathrm{Clu}_{m+\delta_g}$, and $S_m = \bigcup_{g=1}^{8} S_{m,g}$.

We can compute exact design-based moments by drawing all the 1000 possible samples under our sample design. For example, for $\eta = 506,\ m = 12,\ g = 3$, we have $S_{m,g} = \mathrm{Clu}_{12+\delta_3} = \mathrm{Clu}_{14}$, and $\mathrm{Clu}_{14} = \{h_{\mathrm{rem}((506-1+14-1)+\frac{20000}{20}\times(k-1),20000)+1} \mid k = 1\ldots 20\} = \{h_{19},h_{1019},h_{2019},h_{3019},\ldots,h_{19019}\}$. Table 1 displays the rotation chart for our simulation, which is identical to the CPS rotation chart (CPS Technical Paper, 2006, Figure 3-1).

4.4. Rotation bias

In each sample, we introduced a measurement error by changing employment status of 20% of employed individuals in month-in-sample group 1 from employed to unemployed, which leads to an overestimation of the unemployment rate.

Table 1: The CPS Rotation chart

	Clu$_1$	Clu$_2$	Clu$_3$	Clu$_4$	Clu$_5$	Clu$_6$	Clu$_7$	Clu$_8$	Clu$_9$	Clu$_{10}$	Clu$_{11}$	Clu$_{12}$	Clu$_{13}$	Clu$_{14}$	Clu$_{15}$	Clu$_{16}$	Clu$_{17}$	Clu$_{18}$	Clu$_{19}$	Clu$_{20}$
Jan 05	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$									$S_{1,5}$	$S_{1,6}$	$S_{1,7}$	$S_{1,8}$				
Feb 05		$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$									$S_{2,5}$	$S_{2,6}$	$S_{2,7}$	$S_{2,8}$			
Mar 05			$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$									$S_{3,5}$	$S_{3,6}$	$S_{3,7}$	$S_{3,8}$		
Apr 05				$S_{4,1}$	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$									$S_{4,5}$	$S_{4,6}$	$S_{4,7}$	$S_{4,8}$	
May 05					$S_{5,1}$	$S_{5,2}$	$S_{5,3}$	$S_{5,4}$									$S_{5,5}$	$S_{5,6}$	$S_{5,7}$	$S_{5,8}$
Jun 05						$S_{6,1}$	$S_{6,2}$	$S_{6,3}$	$S_{6,4}$									$S_{6,5}$	$S_{6,6}$	$S_{6,7}$
Jul 05							$S_{7,1}$	$S_{7,2}$	$S_{7,3}$	$S_{7,4}$									$S_{7,5}$	$S_{7,6}$
Aug 05								$S_{8,1}$	$S_{8,2}$	$S_{8,3}$	$S_{8,4}$									$S_{8,5}$
Sep 05									$S_{9,1}$	$S_{9,2}$	$S_{9,3}$	$S_{9,4}$								
Oct 05										$S_{10,1}$	$S_{10,2}$	$S_{10,3}$	$S_{10,4}$							
Nov 05											$S_{11,1}$	$S_{11,2}$	$S_{11,3}$	$S_{11,4}$						
Dec 05												$S_{12,1}$	$S_{12,2}$	$S_{12,3}$	$S_{12,4}$					
Jan 06													$S_{13,1}$	$S_{13,2}$	$S_{13,3}$	$S_{13,4}$				
Feb 06														$S_{14,1}$	$S_{14,2}$	$S_{14,3}$	$S_{14,4}$			
Mar 06															$S_{15,1}$	$S_{15,2}$	$S_{15,3}$	$S_{15,4}$		
Apr 06																$S_{16,1}$	$S_{16,2}$	$S_{16,3}$	$S_{16,4}$	
May 06																	$S_{17,1}$	$S_{17,2}$	$S_{17,3}$	$S_{17,4}$
Jun 06																		$S_{18,1}$	$S_{18,2}$	$S_{18,3}$
Jul 06																			$S_{19,1}$	$S_{19,2}$
Aug 06																				$S_{20,1}$

Source: CPS Technical Paper (2006, Figure 3-1)

4.5. Variance on month-in-sample estimators computation

As we draw all the possible samples, we are able to compute the exact variance of any estimator. Moreover, we are able to compute the true $\Sigma_{\mathbf{y}}$, which yields both the optimal best linear and AK estimators.

4.6. Estimation of $\Sigma_{\mathbf{y}}$

Define

$$\sigma^2_{m,m'} = \frac{\sum_{i=1}^{H} \left(\sum_{k\in h_i} \mathbf{y}_{m,k,.} - \frac{\sum_{i'=1}^{H} \sum_{k'\in h_{i'}} \mathbf{y}_{m,k',.}}{H} \right) \left(\sum_{k\in h_i} \mathbf{y}_{m',k,.} \right)^{\mathrm{T}}}{H-1}.$$

We estimate $\sigma^2_{m,m'}$ by

$$\hat{\sigma}^2_{m,m'} = \frac{\sum\limits_{i\in\{1,\ldots,H\} | h_i \subset S_m \cap S_{m'}} \left(\sum\limits_{k\in h} \mathbf{y}_{m,k,.} - \frac{\sum_{i=1}^{H} \sum_{k\in h_i} \mathbf{y}_{m',k,.}}{\#\{i\in\{1,\ldots,H\} | h_i \subset S_m \cap S_{m'}\}} \right) \left(\sum\limits_{k\in h} \mathbf{y}_{m',k,.} \right)^{\mathrm{T}}}{\#\{i \in \{1,\ldots,H\} \mid h_i \subset S_m \cap S_{m'}\} - 1}$$

if $S_m \cap S_{m'} \neq \emptyset$, 0 otherwise. Let $m, m' \in \{1,\ldots,M\}$, $g, g' \in \{1,\ldots,8\}$. If $m' + \delta_{g'} = m + \delta_g$ then $S_{m,g} = S_{m',g'}$ and we approximate the distribution of $S_{m',g'}$ by a cluster sampling, where the first stage is simple random sampling. We estimate $\mathrm{Cov}\left[\hat{\mathrm{t}}^{\mathrm{mis},g}_m, \hat{\mathrm{t}}^{\mathrm{mis},g}_{m'}\right]$ by $\widehat{\mathrm{Cov}}\left[\hat{\mathrm{t}}^{\mathrm{mis},g}_{\mathbf{y}_{m,e}}, \hat{\mathrm{t}}^{\mathrm{mis},g'}_{\mathbf{y}_{m',e'}}\right] = (H)^2 \left(1 - \frac{n}{H}\right) \frac{\hat{\sigma}^2_{m,m'}}{n/8}$. If $m' + \delta_{g'} \neq m + \delta_g$, then $S_{m,g} \cap S_{m',g'} = \emptyset$ and we approximate the distribution of $(S_{m,g}, S_{m',g'})$ by the distribution of two independent simple random samples of clusters conditional to non-overlap of the two samples.

Table 2: Optimal (a_1,k_1) and (a_2,k_2) values for the three synthetic populations

	Population 1	Population 2	Population 3
(a_1,k_1) (unemployed)			
Level	$(0.0471,0.85)$	$(0.0395,0.398)$	$(-0.0704,-0.619)$
Compromise	$(0.029,0.895)$	$(0.00175,0.0551)$	$(0.0038,0.0253)$
Change	$(0.0243,0.89)$	$(0.0358,0.362)$	$(-0.0239,-0.445)$
(a_2,k_2) (employed)			
Level	$(0.0714,0.752)$	$(0.0453,0.73)$	$(-0.0354,0.825)$
Compromise	$(-0.0075,-0.232)$	$(0.002,0.0598)$	$(0.0464,0.0482)$
Change	$(-0.0187,-0.256)$	$(0.0658,0.723)$	$(-0.0529,0.836)$

We estimate $\mathrm{Cov}\left[\hat{t}^{\mathrm{mis}}_{m,g,.},\hat{t}^{\mathrm{mis}}_{m',g',.}\right]$ by $\widehat{\mathrm{Cov}}\left[\hat{\mathrm{t}}^{\mathrm{mis}}_{\mathbf{y}m,g,.},\hat{\mathrm{t}}^{\mathrm{mis}}_{\mathbf{y}m',g',.}\right] = -H\hat{\sigma}^2_{m,m'}$.

4.7. Choice of optimal estimator in each class

In our simulations, the best linear unbiased estimator turned out to be exact in the sense that for the three different choices of $\mathbf{y}$ (population 1, population 2, population 3), the $(1000,2040)$-matrix Y whose rows are the 1000 probable values of $\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}}$ is of rank 1000, so for all (m,e), we can find a 2040-sized vector $x_{m,e}$ such that $Yx_{m,e} = (\mathrm{t}_{\mathbf{y}})_{m,e}.\mathbb{1}$, where $\mathbb{1}$ is 1000-sized vector of ones. Then, we define W_o as the $((M\times 8\times 3),(M\times 3))$-sized array whose rows are the vectors $x_{m,g}$ such that $W_0\, Y^{\mathrm{T}} = \vec{\mathrm{t}}_{\mathbf{y}}$. This surely implies $W_o\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}} = \vec{\mathrm{t}}_{\mathbf{y}}$, and hence the BLUE is necessarily equal to $W_o\vec{\hat{\mathrm{t}}}^{\mathrm{mis}}_{\mathbf{y}}$, a result that we were able to reproduce in our simulations. This situation is particular to our simulation setup, which allows a small number of possible samples, but with a design for which the number of probable samples is larger than the number of month-in-sample estimates, the best linear unbiased estimator would likely have a strictly positive variance. We computed the objective functions for $\alpha \in \{0,0.05,\ldots,1\}$ only. Table 2 shows the optimal values for a_1, k_1, a_2, and k_2 for the three different populations and the best empirical estimator for level, change and compromise. The Census Bureau uses the coefficients $a_1 = 0.3$, $k_1 = 0.4$, $a_2 = 0.4$ and $k_2 = 0.7$ for the CPS. We notice that for each population, the best set of coefficients for change, level and compromise is very close, which means that the optimal choice for level is also almost optimal for change for those three populations. Table 3 shows the best coefficient α for the regression composite estimators.

4.8. Analysis without measurement error

Figure 1 displays the relative mean squared errors of different estimators of unemployment level and change over time: $\left(\frac{\mathrm{MSE}[\hat{\mathrm{r}}^{\star}_m]}{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\mathrm{direct}}_m\right]}\right)_{m\in\{1,\ldots,M\}}$, and $\left(\frac{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\star}_m-\hat{\mathrm{r}}^{\star}_{m-1}\right]}{\mathrm{MSE}\left[\hat{\mathrm{r}}^{\mathrm{direct}}_m-\hat{\mathrm{r}}^{\mathrm{direct}}_{m-1}\right]}\right)_{m\in\{2,\ldots,M\}}$, for $\star\in\{\mathrm{direct},\mathrm{AK},\mathrm{r.c.}\}$. In this figure, the best representative in each class is chosen in the sense that the coefficients of Tables 2 and 3 are used.

Table 3: Optimal regression composite estimator's α parameter value for three synthetic populations

	Population 1	Population 2	Population 3
Level	0.55 (0.6)	0.45 (0.6)	0
Change	1	0.75	0.8
Compromise	0.55 (0.6)	0.45 (0.6)	0

Numbers in the parentheses indicate parameter values in presence of rotation with bias when different

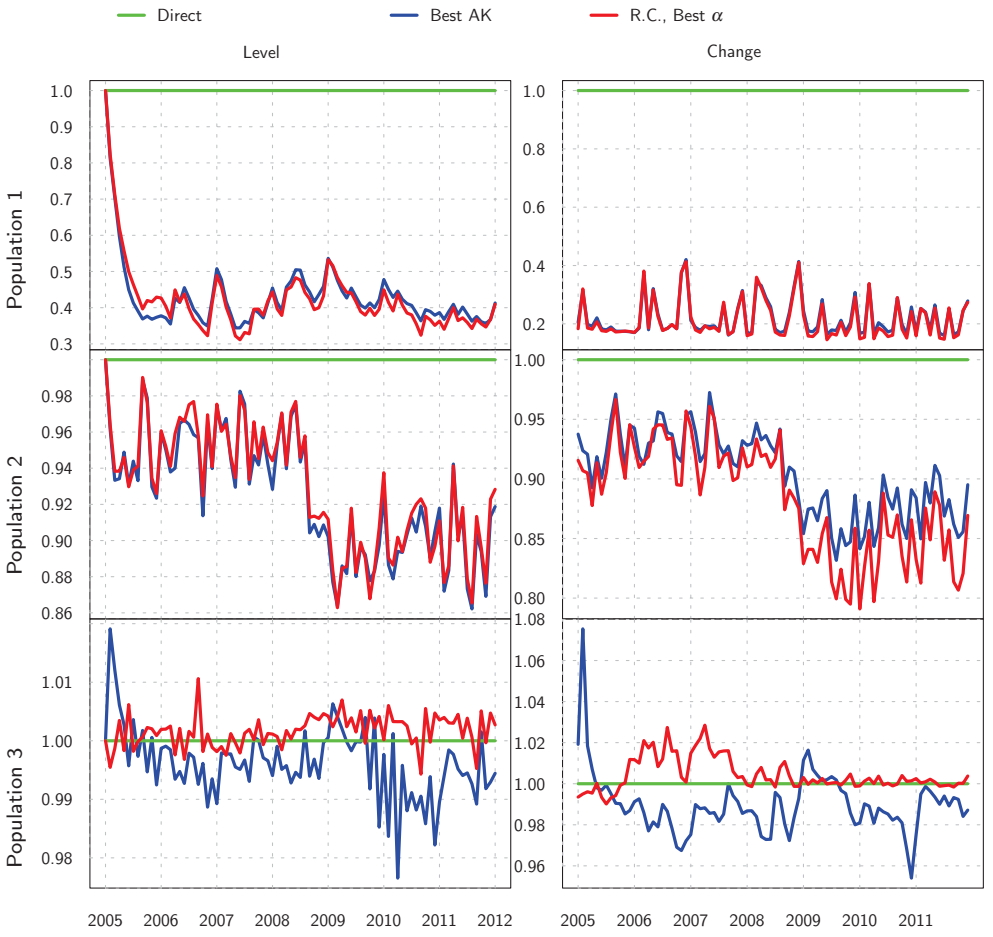
Table 4: For three synthetic populations, quantiles and means (over months) of the relative mean squared errors of unemployment level estimators

	Population 1					Population 2					Population 3				
	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.
0%	0.318	1	1	0.322	0.477	0.87	1	1	0.863	0.885	0.983	1	1	0.994	0.994
25%	0.377	1.52	2.59	0.38	0.546	0.906	1.35	2.56	0.913	0.94	0.996	1.08	1.03	1	1.01
50%	0.409	1.6	2.64	0.42	0.591	0.929	1.41	2.7	0.945	0.974	0.997	1.14	1.04	1	1.02
75%	0.454	1.95	2.74	0.472	0.663	0.951	1.49	2.79	0.969	0.989	1	1.26	1.07	1	1.02
100%	1	2.09	2.86	1	1	1	1.68	3.08	1	1.02	1.01	1.65	1.14	1.01	1.15
Mean	0.431	1.72	2.64	0.443	0.613	0.926	1.42	2.66	0.94	0.966	0.997	1.19	1.05	1	1.02

Note that in the absence of measurement error, the performances of all best "estimators" are comparable.

When trying to estimate the best A and K, the results differ. For different synthetic populations, Table 4 and 5 report the quantiles of relative mean squared errors of the best AK estimator, the empirical best AK estimator, the AK estimator with coefficient taken arbitrarily equal to the CPS AK coefficients (Arb. AK column), the best regression composite estimator (r.c.column) and the Regression Composite estimator with α taken arbitrarily equal to 0.75 (Arb. AK column) for the level and change estimation, respectively. For all three synthetic populations, both the estimated best AK estimator and arbitrary AK estimator perform worse than the direct estimator. Moreover, the arbitrary regression composite estimator seems to behave much better than the estimated best AK estimator and arbitrary AK estimators. We observe (not reported here) that the estimated best linear estimator performs worse than the estimated best AK estimator. This underlines the weakness of the AK and Yansaneh-Fuller type estimators: without a good estimator of the variance-covariance matrix, they perform very poorly. We note that the regression composite estimator with arbitrary α performs better without requiring any estimation of the variance.

Figure 1: Relative mean squared errors of different estimated series of unemployment level and of month-to-month changes



4.9. Analysis with measurement error

Under (2), a solution to the rotation group bias for adapting the AK estimator consists in estimating the rotation bias parameter vector b and then applying AK coefficients to corrected month-in-sample estimates, to obtain $\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{AK}*}\right)_{m,.} = \sum_{m'=1}^{m}\sum_{m'=1}^{m}\left(c_{m,m',g}\left(\hat{\mathrm{t}}_{\mathbf{y}}^{\mathrm{mis},g}\right)_{m,g,.} - \hat{b}_g\right)$. The question of how to adapt the regression composite estimator to take into account measurement error is more complicated. Besides, the model used for rotation bias is itself questionable. The linear constraint on b ($\sum b_{g,.} = 0$ or $b_{1,.} = 0$) is imposed to address an identifiability problem, but one cannot assess its validity. As a result we think it is not a good way to deal with the rotation bias. We have not investigated how to adapt the regression composite estimator to address the problem of rotation bias. Instead we studied its behaviour in presence of rotation bias. To this end, we systematically (for all months, all samples) changed the status of up to 2 unemployed persons of month-in-sample group 1 from unemployed to employed. For different populations, Tables 6 and 7 display quantiles and means of the relative mean squared errors of the best AK estimator and the best regression composite estimator for both level and change. We applied the best AK and best regression composite estimators to the cases without measurement error and with measurement error. We notice that AK estimator is very sensitive to rotation bias, whereas regression composite estimator is not. A reason may be that introducing a variable not correlated to the study variables in the calibration procedure does not much change the estimation of the study variable. Rotation bias weakens the correlation between $\mathbf{z}$ and $\mathbf{y}$, and yet the performance of the regression composite estimator is comparable to the performance of the direct.

Table 5: Quantiles and means (over months) of the relative mean squared errors for different populations and unemployment month-to-month change estimators

	Population 1					Population 2					Population 3				
	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.	Best AK	Arb. AK	Emp. AK	Best r.c.	Arb. r.c.
0%	0.0959	2.77	5.43	0.0279	0.0936	0.845	0.872	2.72	0.774	0.791	0.973	0.998	1.01	0.984	0.994
25%	0.123	3.31	6.35	0.0455	0.112	0.887	0.953	3.07	0.835	0.847	0.99	1.02	1.03	0.992	1
50%	0.142	3.68	6.64	0.0552	0.127	0.914	0.998	3.33	0.885	0.89	0.993	1.02	1.04	0.997	1
75%	0.215	5.21	6.93	0.146	0.201	0.932	1.03	3.62	0.916	0.919	0.996	1.03	1.06	1	1
100%	0.395	6.12	7.59	0.355	0.383	0.971	1.13	3.92	0.965	0.967	1.04	1.06	1.14	1.11	1.01
Mean	0.174	4.21	6.68	0.102	0.163	0.909	0.993	3.33	0.876	0.883	0.993	1.03	1.04	1	1

Table 6: Quantiles and means (over months) of the relative mean squared errors of unemployment level estimators for different populations.

	Population 1		Population 2		Population 3		Pop. 1 (bias)		Pop. 2 (bias)		Pop. 3 (bias)	
	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.
0%	0.318	0.322	0.87	0.863	0.983	0.994	1	0.0521	1	0.117	0.919	0.158
25%	0.377	0.38	0.906	0.913	0.996	1	46.1	0.0735	2.78	0.155	1.5	0.739
50%	0.409	0.42	0.929	0.945	0.997	1	47.9	0.0949	2.81	0.179	1.59	0.768
75%	0.454	0.472	0.951	0.969	1	1	52.5	0.115	2.86	0.254	1.86	0.786
100%	1	1	1	1	1.01	1.01	57.6	0.162	2.92	0.3	2.18	0.843
Mean	0.431	0.443	0.926	0.94	0.997	1	45.6	0.0957	2.77	0.203	1.64	0.754

5. The CPS Data Analysis

5.1. Implementation of regression composite estimator for the CPS

5.1.1 Choice of α

Under a simple unit level times series model with auto-regression coefficient ρ, Fuller and Rao (2001) proposed a formal expression for an approximately optimal α as a function of ρ and studied the so-called drift problem for the MR2 choice: $\alpha = 1$. They also proposed approximate expressions for variances of their estimators for the level and change. For various reasons, it seems difficult to obtain the optimal or even an approximately optimal α needed for the Fuller-Rao type regression composite estimation technique to produce the U.S. employment and unemployment rates using the CPS data. First of all, the simple time series model used by Fuller and Rao (2001) is not suitable to model a nominal variable (employment status) with several categories. Secondly, the complexity of the CPS design poses a challenging modeling problem. Before attempting to obtain the optimal or even an approximately optimal choice of α required for the Fuller-Rao type regression composite method, it will be instructive to evaluate regression composite estimators for different known choices of α. This is the focus of this section.

5.1.2 Choice of $\mathbf{x}$ and $\mathbf{z}$

In our study, we considered two options for $\mathbf{z}$: (i) $\mathbf{z} = \mathbf{y}$, (ii) a more detailed employment status variable with 8 categories. As the use of this more detailed variable reduces

Table 7: Quantile and means (over months) of the relative mean squared errors of unemployment month-to-month change estimators for different populations.

	Population 1		Population 2		Population 3		Pop. 1 (bias)		Pop. 2 (bias)		Pop. 3 (bias)	
	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.	AK	r.c.
0%	0.0959	0.0936	0.845	0.791	0.973	0.994	0.422	0.0298	0.898	0.457	1.19	0.935
25%	0.123	0.112	0.887	0.847	0.99	1	0.477	0.0385	0.938	0.552	1.48	0.994
50%	0.142	0.127	0.914	0.89	0.993	1	0.515	0.05	0.954	0.583	1.5	1.01
75%	0.215	0.201	0.932	0.919	0.996	1	0.563	0.093	0.971	0.613	1.52	1.02
100%	0.395	0.383	0.971	0.967	1.04	1.01	3.96	0.209	1.25	0.673	1.58	1.05
Mean	0.174	0.163	0.909	0.883	0.993	1	0.665	0.0671	0.958	0.581	1.48	1

the degrees of freedom in the calibration procedure and leads to estimates with a higher mean squared error, we report results for option (i) only. For an application of the Fuller-Rao method, one might think of including all the variables that have already been used for the weight adjustments in the $\mathbf{x}$ variables. However, this would introduce many constraints on the coefficients and thus is likely to cause a high variability in the ratio of $\mathbf{w}_{m,k}$ and $\mathbf{w}^{\text{r.c.}}_{m,k}$. The other extreme option is not to use any of the auxiliary variables, but then the final weights would not be adjusted for the known totals of auxiliary variables $\mathbf{x}$. As a compromise, we selected only two variables: gender and race.

5.2. Results

Figure 2(a) displays the difference $\widehat{\text{r}}^{\text{AK}}_m - \widehat{\text{r}}^{\text{direct}}_m$ between different composite estimates and the corresponding direct estimates against months m. For the regression composite estimator, we considered three choices: (i) $\alpha = 0.75$ (suggested by Fuller and Rao), (ii) $\alpha = 0$ (corresponding to MR1), and (iii) $\alpha = 1$ (corresponding to MR2). We display similar graphs for month-to-month change estimates in Figure 2(b). Notice that $\alpha = 0$ and $\alpha = 1$ correspond to MR1 and MR2, respectively. We display similar graphs for month-to-month change estimates in Figure 2.

It is interesting to note that the AK composite estimates of unemployment rates are always lower than the corresponding direct estimates in Figure 2(a). To our knowledge, this behavior of AK composite estimates has not been noticed earlier. In contrast, the regression composite estimates MR1 are always higher than the corresponding direct estimates. However, such deviations decrease as α gets closer to 1 as shown in Figure 2(a). Application of the Fuller-Rao method at the household level causes an increase in the distance between the original and calibrated weights and one may expect an increase in the variances of the estimates. Figure 2(b) does not indicate systematic deviations of the composite estimates of the month-to-month changes from the corresponding direct estimates. Deviations of the regression composite estimates from the corresponding direct estimates seem to decrease as α approaches 1.

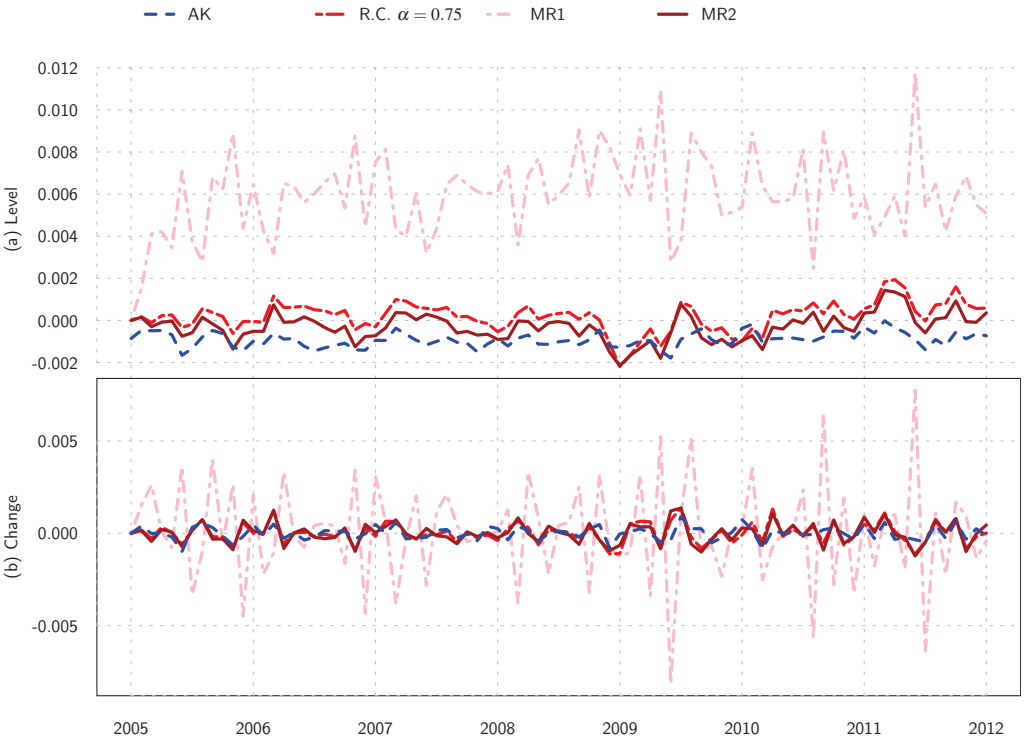
6. Discussion

Our study reveals that there is ample scope for improving the AK estimator used by the Census Bureau. We would like to emphasize the following undesirable features of the AK estimation method:

(i) The method used to compute optimal coefficient is crude — the best coefficients are selected from 9 different values. Our R package, based on the built in R Nelder-Mead algorithm, can provide the optimal coefficients within 8 digits of precision in a reasonable time.

(ii) The stationarity assumption on the variances and covariances of the month-in-sample estimators over a period of 10 years does not seem realistic, and to our knowledge, has not been tested before. Moreover, even if the stationary model was reasonable, the complexity of the CPS sample design makes it difficult to evaluate the quality of the estimators used for that model. The difficulty in proposing a stochastic model for the

Figure 2: Estimated series of differences between different composite estimates and the corresponding direct estimates



best linear estimators in the CPS was pointed out earlier by (Jones, 1980, Sec. 4). Our evaluation study shows that the AK estimators is very sensitive to the choices of A and K and that the errors in the estimation of the variances and covariances may lead to poor performance of the AK estimators. Moreover, estimators of variances and covariances of month-in-sample estimators affect the performances of empirical best linear unbiased estimators.

(iii) Using the Bailar model for the bias in our study, we showed that AK estimator is very sensible to rotation group bias. There is currently no satisfactory way to correct the AK estimator for the rotation bias. The Bailar model relies on an arbitrary constraint on the month-in-sample biases and a strong stationarity assumption of the month-in-sample bias and should not be used unless some re-interview study can justify the Bailar's model. One possible option would be to study the rotation bias at the individual level using resampling method. In this paper, we have not investigated how to adapt the regression composite estimator to address the problem of rotation bias. This could be a good problem for future research.

(iv) The computation of composite weights in CPS to calibrate the weights on the AK

estimators will affect all other weighted estimators. Although Lent and Cantwell (1996) showed that there was not a big effect on the estimates, considering the concerns about AK estimators listed before, we do not think that the use of those composite weights is a good option.

(v) The CPS data analysis shows that the AK estimates are consistently smaller than the corresponding direct survey-weighted estimates for the period 2005-2012. This is also a source of concern.

The composite regression estimator does not rely on an estimation of the variances and covariances matrix. In our simulation study, it appears to be less sensitive to rotation group bias, and bounces around the survey-weighted estimates when applied to the real CPS data. Our study encourages the use of the regression composite method in the US labor force estimation.

To facilitate and encourage further research on this important topic, we make the following three R packages, developed under this project, freely available: (i) the package dataCPS can be used to download CPS public data files and transform them into R data set (Bonnéry (2016b)); (ii) the package CompositeRegressionEstimation can be used to compute the AK, best AK, composite regression, linear and best linear estimators (Bonnéry (2016a)); (iii) the package pubBonneryChengLahiri2016 can be used to reproduce all computations and simulations of this paper (Bonnéry, 2016c).

Acknowledgements

We thank Editor-in-Chief Professor Wlodzimierz Okrasa and an anonymous referee for reading an earlier version of the article carefully and offering a number of constructive suggestions, which led to a significant improvement of our article. We thank Ms. Victoria Cheng for helping us proofreading the entire manuscript. The research of the first and third authors has been supported by the U.S. Census Bureau Prime Contract No: YA1323-09-CQ-0054 (Subcontract No: 41-1016588). The programs used for the simulations have been made available on the github repository Bonnéry (2016c). The first author completed the research as a postdoctoral research associate of P. Lahiri at the University of Maryland, College Park, USA.

References

BAILAR, B. A., (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70(349), pp. 23–30.

BEAUMONT, J.-F. and BOCCI, C., (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey for Change Estimates. *SSC Annual Meeting, Proceedings of the Survey Methods Section*, (June), pp. 1–6.

BELL, P., (2001). Comparison of alternative labour force survey estimators. *Survey Methodology*, 27(1), pp. 53–63.

BONNÉRY, D. B., (2016a). R package CompositeRegressionEstimation. https://github.com/DanielBonnery/CompositeRegressionEstimation.

BONNÉRY, D. B., (2016b). R package dataCPS. https://github.com/DanielBonnery/dataCPS.

BONNÉRY, D. B., (2016c). R package pubBonneryChengLahiri2016. https://github.com/DanielBonnery/pubBonneryChengLahiri2016.

CASSEL, C., SÄRNDAL, C., and WRETMAN, J., (1977). *Foundations of inference in survey sampling*.

CPS Technical Paper, (2006). Design and Methodology of the Current Population Survey. Technical Report 66, U.S. Census Bureau.

FULLER, W. A. and RAO, J. N. K., (2001). A regression composite estimator with application to the Canadian Labour Force Survey. *Survey Methodology*, 27(1), pp. 45–51.

GAMBINO, J., KENNEDY, B., and SINGH, M. M. P. M. M. P., (2001). Regression composite estimation for the Canadian labour force survey: Evaluation and implementation. *Survey Methodology*, 27(1), pp. 65–74.

GURNEY, M. and DALY, J. F., (1965). A multivariate approach to estimation in periodic sample surveys. In *Proceedings of the Social Statistics Section, American Statistical Association*, volume 242, p. 257.

HANSEN, M. H., HURWITZ, W. N., NISSELSON, H., and STEINBERG, J., (1955). The redesign of the census current population survey. *Journal of the American Statistical Association*, 50(271), pp. 701–719.

JONES, R. G., (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2), pp. 221–226.

LENT, J. and CANTWELL, S., (1996). Effect of Composite Weights on Some Estimates from the Current Population Survey. *Proceedings the American Statistical Association, Section on Survey Research Methods*, pp. 130–139.

LENT, J., MILLER, S. M., CANTWELL, P. J., and DUFF, M., (1999). Effects of composite weights on some estimates from the current population survey. *Journal of Official Statistics-Stockholm*, 15(1), pp. 431–448.

SALONEN, R., (2007). Regression Composite Estimation with Application to the Finnish Labour Force Survey. *Second Baltic-Nordic Conference on Survey Sampling, 2.–7. June 2007, Kuusamo, Finland*, 8(3): 503–517.

SEARLE, S., (1994). Extending some results and proofs for the singular linear model. *Linear Algebra and its Applications*, 210, pp. 139–151.

SINGH, A. C., KENNEDY, B., and WU, S., (2001). Regression composite estimation for the Canadian Labour Force Survey: evaluation and implementation. *Survey Methodology*, 27(1), pp. 33–44.

SINGH, A. C., KENNEDY, B., WU, S., and BRISEBOIS, F., (1997). Composite estimation for the Canadian Labour Force Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 300–305.

SINGH, A. C. and MERKOURIS, P., (1995). Composite estimation by modified regression for repeated surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 420–425.

YANSANEH, I. S. and FULLER, W. A., (1998). Optimal recursive estimation for repeated surveys. *Survey Methodology*, 24, pp. 31–40.

APPENDIX

A. Description of CPS design

This section uses CPS notations for rotation groups. Let U be the intersection of a given basic primary sampling unit component (BPC) and one of the frames used in CPS (see CPS Technical Paper (2006)). The BPC is a set of clusters of about four housing units, the clusters are the ultimate sampling units (USU). Let N be the number of clusters in U. The clusters in U are sorted according to geographical and demographic characteristics and then indexed by $k = 1 \ldots N$. In the sequence, we will designate a cluster by its index. Let SI_w be the adjusted within-PSU sampling interval, as defined in CPS Technical Paper (2006, p. 3-11). Let $n = \lfloor (21 \times 8 * SI_w)^{-1} N \rfloor$, where $\lfloor . \rfloor$ is the floor function. The number n is the sample size for a sample rotation group. The drawing of the USU within the PSU consists in the generation of a random number X according to the uniform law on $[0,1]$. For $i = 1 \ldots n$, $j = 1 \ldots 8$, $\ell = 85 \ldots (85+15)$, let $k_{i,j,\ell}$ denote the cluster $k_{i,j,\ell} = \lfloor (X + 8 \times (i-1) + j) \times SI_w + (\ell - 85) \rfloor$. Then, with the notations of CPS Technical Paper (2006) for $\ell = 85 \ldots 100$, $j = 1 \ldots 8$, the rotation group j of sample A_ℓ is given by

$$A_{\ell,j} = \left\{ k_{i,j,\ell} \mid i = 1 \ldots n \right\}.$$

For a given month the sample consists of 8 rotation groups. There are 120 months in a period of 10 years. For $m = 1 \ldots 120$, $j' \in \{1, \ldots, 8\}$, $\ell_{m,j'}$ and $j_{m,j'}$ are given by: $j_{m,j'} = t + j' - 1 - 8 \times \lfloor (t + j' - 2)/8 \rfloor$. If $j' \in \{1, \ldots, 4\}$, $\ell_{m,j'} = 85 + \lfloor (t + j' - 2)/8 \rfloor$. If $j' \in \{5, \ldots, 8\}$, $\ell_{m,j'} = 86 + \lfloor (t + j' - 2)/8 \rfloor$.

The sample of the mth month, counting from November 2009, is given by

$$s_m = \bigcup_{j'=1}^{8} A_{\ell_{m,j'}, j_{m,j'}}.$$

For example, June 2013 corresponds to $m = 44$, counting from November 2009. Then

$$\begin{aligned}
\ell_{m,1} &= 85 + \lfloor 43/8 \rfloor = 90 & j_{m,1} &= 44 - 8 \times \lfloor 43/8 \rfloor = 4 \\
\ell_{m,2} &= 85 + \lfloor 44/8 \rfloor = 90 & j_{m,2} &= 45 - 8 \times \lfloor 44/8 \rfloor = 5 \\
\ell_{m,3} &= 85 + \lfloor 45/8 \rfloor = 90 & j_{m,3} &= 46 - 8 \times \lfloor 45/8 \rfloor = 6 \\
\ell_{m,4} &= 85 + \lfloor 46/8 \rfloor = 90 & j_{m,4} &= 47 - 8 \times \lfloor 46/8 \rfloor = 7 \\
\ell_{m,5} &= 86 + \lfloor 47/8 \rfloor = 91 & j_{m,5} &= 48 - 8 \times \lfloor 47/8 \rfloor = 8 \\
\ell_{m,6} &= 86 + \lfloor 48/8 \rfloor = 92 & j_{m,6} &= 49 - 8 \times \lfloor 48/8 \rfloor = 1 \\
\ell_{m,7} &= 86 + \lfloor 49/8 \rfloor = 92 & j_{m,7} &= 50 - 8 \times \lfloor 49/8 \rfloor = 2 \\
\ell_{m,8} &= 86 + \lfloor 50/8 \rfloor = 92 & j_{m,8} &= 51 - 8 \times \lfloor 50/8 \rfloor = 3
\end{aligned}$$

We can check from the CPS rotation chart (CPS Technical Paper, 2006, Fig. 3-1) that the sample of June 2013 consists of the 4th, 5th, 6th, 7th rotation groups of A90, of the 8th rotation group of A91, and of the 1st, 2d and 3rd rotation groups of A92:

$$S_{\text{June 2013}} = A_{90,4} \cup A_{90,5} \cup A_{90,6} \cup A_{90,7} \cup A_{91,8} \cup A_{92,1} \cup A_{92,2} \cup A_{92,3}.$$

Index

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 191–211, DOI 10.21307/stattrans-2020-038
Received – 31.01.2020; accepted – 30.06.2020

A generic business process model for conducting microsimulation studies

Jan Pablo Burgard[1], Hanna Dieckmann[2], Joscha Krause[3], Hariolf Merkle[4], Ralf Münnich[5], Kristina M. Neufang[6], Simon Schmaus[7]

Abstract

Microsimulations make use of quantitative methods to analyze complex phenomena in populations. They allow modeling socioeconomic systems based on micro-level units such as individuals, households, or institutional entities. However, conducting a microsimulation study can be challenging. It often requires the choice of appropriate data sources, micro-level modeling of multivariate processes, and the sound analysis of their outcomes. These work stages have to be conducted carefully to obtain reliable results. We present a generic business process model for conducting microsimulation studies based on an international statistics process model. This simplifies the comprehensive understanding of dynamic microsimulation models. A nine-step procedure that covers all relevant work stages from data selection to output analysis is presented. Further, we address technical problems that typically occur in the process and provide sketches as well as references of solutions.

Keywords: multi-source analysis, multivariate modeling, social simulation, synthetic data generation

1. Introduction

Microsimulation studies represent a powerful tool for the multivariate analysis of populations (Merz, 1993; O'Donoghue, 2001; O'Donoghue and Dekkers, 2018; Burgard et al., 2019a). While macrosimulation methods are limited to selected population characteristics on an aggregated level, microsimulation methods are capable of considering

[1]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: burgardj@uni-trier.de. ORCID: https://orcid.org/0000-0002-5771-6179.
[2]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: dieckmann@uni-trier.de. ORCID: https://orcid.org/0000-0002-6455-1210.
[3]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: krause@uni-trier.de. ORCID: https://orcid.org/0000-0002-2473-1516.
[4]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: merkle@uni-trier.de. ORCID: https://orcid.org/0000-0001-7653-383X.
[5]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: muennich@uni-trier.de. ORCID: https://orcid.org/0000-0001-8285-5667.
[6]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: neufang@uni-trier.de. ORCID: https://orcid.org/0000-0002-8241-4818.
[7]Trier University, Department of Economic and Social Statistics, Germany.
E-mail: schmaus@uni-trier.de. ORCID: https://orcid.org/0000-0002-2037-4312.

individual characteristics and interactions. This allows for a more comprehensive understanding of the population and sophisticated projections on its development. As a result, microsimulations are increasingly applied for the analysis of complex systems. Exemplary applications are provided by Bourguignon and Spadaro (2006), Pichon-Riviere et al. (2011), Li and O'Donoghue (2013), Markham et al. (2017), O'Donoghue and Dekkers (2018), and Burgard et al. (2020).

Microsimulation studies are often performed according to a basic procedure. First, an adequate base dataset as a representation of the target population is needed. This requires either synthetic data or empirical observations from administrative records and surveys (Li and O'Donoghue, 2013). Next, selected features that characterize the population in its initial state are altered in scenarios. The scenarios are projected into future periods and construct individual branches in the evolution of the base population. After a sufficient number of simulation periods, the branches are compared. The comparison provides insights into essential dynamics and interdependencies within the population that typically cannot be assessed otherwise (Li and O'Donoghue, 2013; Burgard et al., 2019b).

However, there is a lack in generic descriptions on how to construct, implement, and evaluate microsimulations. This makes it difficult for researchers that are new to the field to properly conduct their own studies. Microsimulations require the statistically sound combination of multiple data sets, the construction of a sophisticated simulation infrastructure, as well as the careful analysis of simulation outcomes. If these challenges are not addressed properly, microsimulation results are not reliable and may lead to false conclusions in the analysis.

In this paper, we present a generic business process model for conducting microsimulation studies. We develop a coherent framework that can be used as instruction for all relevant work stages, including data generation, population projection, and output analysis. Drawing from the generic statistical business process model by UNECE (2013), our model consists of nine sequential steps. For each step, we elaborate on data requirements, methodological challenges, as well as possible solutions. Our descriptions can be broadly used as guidance to properly perform microsimulation research for various applications.

The remainder of the paper is organized as follows. In Section 2, we cover the specification of needs, data selection and preparation. Section 3 describes the population projection. In particular, we look at the design of the microsimulation model, population dynamics, as well as the actual simulation. In Section 4, we address output analysis. Here, relevant aspects are the analysis of simulation results, dissemination strategies, and evaluation. Section 5 closes with some concluding remarks and an outlook on future research.

2. Requirements and data selection

2.1. Step 1: Specification of needs

The underlying concept of microsimulations is to model the actions and interactions of micro-level units in a population to analyze their impact on the macro-level (Spielauer, 2011). For instance, micro-level units may represent individuals in the context of social change, firms in a competitive market situation, and cars as part of traffic or transport systems. Thus, in order to conduct a microsimulation study that allows for reliable results, a suitable simulation frame and clear research questions have to be defined. This can be done by answering the following questions:

- What kind of system shall be simulated?
- What are the characteristics of interest?
- Under which scenarios shall these features be studied?
- Which hypotheses shall be investigated?
- What are the smallest relevant entities for this purpose?
- What are potentially relevant processes and interdependencies?
- What temporal frequency for projection has to be considered?

An important distinction is between static and dynamic microsimulations (Rahman and Harding, 2017; Hannappel and Troitzsch, 2015). Static microsimulations typically have fewer data requirements and demand less computational resources than dynamic microsimulations. They are suitable for applications where the immediate effect of a clearly defined external change on micro-level units is of interest. The attributes associated with micro-level units are mainly persistent over the simulation process. In this setting, the temporal change of micro-level attributes can be modeled indirectly via reweighting and uprating (inflating/deflating) of variables (Dekkers, 2015). Prominent models such as EUROMOD (Sutherland and Figari, 2013) commonly focus on the impact of possible (e.g. tax-related) policy changes.

Dynamic microsimulation models such as DYNASIM (Favreault et al., 2015) allow for a more sophisticated evolution of the population on the micro-level. A given micro-level characteristic is an endogenous factor in the simulation. The probability for a specific realization depends on both the simulated time and the realizations of other micro-level characteristics. Likewise, dynamic microsimulations are characterized by stochastic transitions and direct temporal changes of micro-level unit attributes. They are suitable for applications where multidimensional dependencies between micro-level units are relevant for the simulation outcomes. For instance, Burgard et al. (2019a) used a dynamic model for investigating future long-term care demand in a city, which required the anticipation of family structures and neighborhood characteristics. Naturally, this simulation type can be very resource-intensive.

A further distinction of dynamic models is with respect to the representation of time: discrete and continuous. In discrete-time dynamic microsimulations, temporal changes occur at predefined time intervals, such as simulated months or years. In continuous-time dynamic microsimulations, temporal changes occur at any given time within the simulated time domain (simulation horizon). Conceptually, the choice between these modes depends on whether it is necessary to account for interperiodic events in light of the research questions. Methodologically, the choice should be based on the assumptions regarding transition dynamics the researcher is willing to make. Discrete dynamics require less assumptions for the modeling of a given transition, but are also less flexible in accounting for complex interdependent event sequences (Willekens, 2017). Continuous dynamics typically require far-reaching assumptions on conditional transition rates, but are generally capable of displaying highly complex temporal event dependencies. For deeper insights into dynamic microsimulation modeling, see for example Li and O'Donoghue (2013), O'Donoghue and Dekkers (2018), and Willekens (2017).

Another crucial distinction is between open and closed population microsimulations (Spielauer, 2009). It refers to the question of whether micro-level units can interact with other micro-level units that are not initially part of the system of interest. In a closed-population microsimulation model, interactions are restricted to units that are part of the base population prior to projection. In an open population microsimulation model, new units can be generated that are added to the base population during the simulation. For instance, if a demographic projection of a regional population shall be performed, then this may correspond to migration from other regions. Conceptually, the closed approach is sensible when the research focus is on the regional population in its current state. Any effect that unfolds under a particular projection scenario is exclusively intrinsic given the initial base population. The open approach can be used when the focus is on the evolution of the region in which the base population is located. Modeling the corresponding domain as an entity requires the consideration of migration in order to be realistic. Naturally, open-population microsimulations need detailed migration data for this purpose.

After a suitable variant has been determined, the researcher has to define several simulation scenarios. They should be constructed such that they meet population characteristics that are essential in light of the research questions. A key aspect of microsimulation is to examine how target variables change under various theoretical social, economic or policy-related developments. For instance, demographic scenarios or alternative policies (e.g. tax-benefit systems) might be relevant for the research context and, therefore, be integrated into the simulation process.

2.2. Step 2: Data selection & Step 3: Data preparation

After determining research objectives and the model variant, data requirements have to be specified. The methodological challenges associated with these work stages directly depend on each other. Therefore, we address these steps jointly.

We introduce some notation and a basic data setting that helps us to illustrate the relevant aspects. Let U denote a real-world population of $|U| = N$ individuals indexed

by $i = 1,...,N$. The objective is to analyze this population via microsimulation methods. Thus, in light of the comments from Section 1, it represents the system of interest. Let $\tilde{U}$ be the base population of $|\tilde{U}| = \tilde{N}$ micro-level units indexed by $u = 1,...,\tilde{N}$. It may be viewed as a digital replica of U that we can project into future periods. Further, let $D \subset U$ be a random sample of $|D| = n$ individuals indexed by $i = 1,...,n$. Denote p_i as the inclusion probability associated with $i \in U$ given the sampling design. The sample represents an exemplary data input for the microsimulation. It can be used to construct the base population $\tilde{U}$ and to obtain empirical parameters for the projection of $\tilde{U}$. In what follows, we elaborate on potential data sources that a researcher may consider as a base population directly or for the creation of such a population.

Data Type	Characteristics	Formalization	p_i known?	Example(s)
Administrative Data	All units of a population of interest available in its entirety	$i \in U$	$p_i = 1$	Register of residents, register of taxpayers
Census Data	Usually person- and household-level data	$i \in U$	$p_i = 1$	Data collected from a census
Survey Data	A random sample of the units of the population of interest is available	$i \in D \subset U$	$p_i \in (0,1]$	Survey of units of interest, e.g. households, persons, firms
Synthetic Data	A synthetic population of interest containing (partially) synthetic units	$u \in \tilde{U}$	Yes / No	Generated data based on other data sources
Big Data	Huge, complex or steadily fast generated data	$i \in D \subset U$	No	Remote sensing data or data collected using phones

Table 1: Datatypes and their properties

A crucial point for the assessment of data quality is to know about the data production process. Since data serves as input for microsimulation models, the data quality determines also the quality of the microsimulation model. Table 1 provides a generic overview of exemplary data sources and their associated properties. The most relevant data sources are administrative data, census data, household, and survey data, as well as synthetic data (Li and O'Donoghue, 2013). The use of big data sources is not yet established in the microsimulation literature, but marks a relevant option for future research (O'Donoghue and Dekkers, 2018).

We start with administrative and census data. In the best case, these data sets cover the entire population U and there is no sampling process that has to be anticipated.

Further, they are rarely subject to measurement errors, such as inaccurate reportings of sampled individuals. Therefore, they can often be used directly. If the data sets cover all relevant characteristics in light of the research questions, then the researcher can use them as base population $\tilde{U}$ for projection. However, if essential characteristics are missing, then the data sets may be extended artificially via synthetic data generation. Further, please note that there are also occasions where administrative data does not cover the entire population, but only a subset of it given the administrative purpose. A corresponding example would be administrative data on taxation, where only tax-payers are included. In these cases, issues like coverage problems have to be accounted for in order to create the base population. For further details, see for instance Smith et al. (2009).

In the case of survey data, the researcher must be aware of the sampling design in order to use the data correctly (Dekkers and Cumpston, 2012). Depending on the application, it is necessary to apply weighting and imputation procedures, provided that they are not already implemented by the data producer. These steps involve the adjustment for possible nonresponse. For unit-nonresponse, the design weights (typically inverse inclusion probabilities) are altered such that relevant sample totals reproduce known population totals (Haziza and Beaumont, 2017). This is achieved via calibration methods, such as the generalized regression estimator (Deville and Särndal, 1992; Särndal, 2007) and empirical likelihood techniques (Chen and Quin, 1993). For item-nonresponse, the missing observations are imputed, for instance via multiple imputation (Schafer and Graham, 2002). Once the data set is adjusted, it can either be directly used as base population or has to be expanded by means of adding synthetic individuals.

However, often the required data might not be available due to disclosure control, as the data provider is obligated to delete regional identifiers. In this case, the generation of synthetic data is an option (Drechsler, 2011). For this, often multiple data sources (e.g. survey data and known totals) can be combined to construct a synthetic population based on real-world observations. For instance, the researcher may calculate calibration weights (Deville and Särndal, 1992; Burgard et al., 2019c) for survey observations such that (synthetic) marginal distributions reproduce known population totals for a set of relevant characteristics. The synthetic population then consists of units allocated (with replacement) to spatial regions according to their newly calculated weight (Williamson, 2013; Tanton et al., 2014; Lovelace, 2016; Rahman and Harding, 2017; Tanton, 2018). Alternatively, a synthetic population can be modeled by estimating distribution or model parameters from survey data and actually reconstruct the population (Huang and Williamson, 2001; Münnich and Schürle, 2003; Alfons et al., 2011a, Alfons et al., 2011b). This can avoid cluster effects arising from units that are repeated frequently within a region. In conclusion, there is a reweighting and an imputation approach to generate synthetic data. For the imputation approach, one considers to apply editing procedures to avoid implausible variable outcomes (Drechsler, 2011). After the synthetic population has been generated, it can be used as base population for projection.

Although not yet established, using big data for microsimulation research is an important topic. These data sets are typically very rich in detail and allow to survey complex

phenomena, such as network structures. As a result, social media data is already used in humanity fields like sociology (Murthy, 2012). Microsimulations could greatly benefit from corresponding data in order to improve the modeling of network structures or social behavior. However, big data sources also impose several methodological challenges, such as coverage problems or unknown inclusion probabilities. These issues mark a central subject for future research in the field.

3. Population projection

3.1. Step 4: Design of the microsimulation model

In Section 2.2, we stated the importance of constructing suitable scenarios given the research questions. However, not only the scenario design is crucial, but also the design of an overall functional simulation infrastructure. There are different approaches to ensure that the infrastructure works as desired. Naturally, these approaches depend on the type and complexity of the microsimulation variant chosen in Step 1. We elaborate on this aspect hereafter.

Depending on the requirements concerning performance, flexibility, additional features and costs, researchers are offered different software solutions to conduct their microsimulation study. Following Li and O'Donoghue (2013), packages to program microsimulation models can be categorized according to their development environment, having pros and cons. General-purpose programming languages (such as C/C++/C#, Python, or Java) offer high flexibility, but also require high programming skills. General-purpose statistical or mathematical packages (such as Stata, SAS, or R) might be less efficient in computing the model, but provide pre-implemented statistical operations that can be applied for simulation. There are also simulation modeling packages that focus exclusively on setting up microsimulations, such as EUROMOD (Sutherland and Figari, 2013), Modgen (Spielauer, 2006; Bélanger and Sabourin, 2017), JAMSIM (Mannion et al., 2012) or LIAM2 (de Menten et al., 2014). These packages are typically less flexible, but easier to use for applied researchers without advanced knowledge in statistical programming.

When creating microsimulations, it is recommendable to use a modular structure as basis for the implementation of population dynamics. Population dynamics are driven by multiple subprocesses that are usually organized independently. Note that an independent organization does not necessarily imply that state transitions within corresponding subprocesses are stochastically independent. We will address that aspect later in this section. The conceptual distinction between these points can be made according to certain transitions or content groups. O'Donoghue et al. (2009, p. 20) describe modules as "the components where calculations take place, each with its own parameters, variable definitions and self-contained structure, with fixed inputs and outputs."

In a given programming language, the modules may correspond to functions that require the base population as input. Figure 1 shows a four-step process that takes place within an exemplary module for discrete-time dynamic microsimulation. In the first step, the individuals have to be selected regarding their eligibility for a change. This is

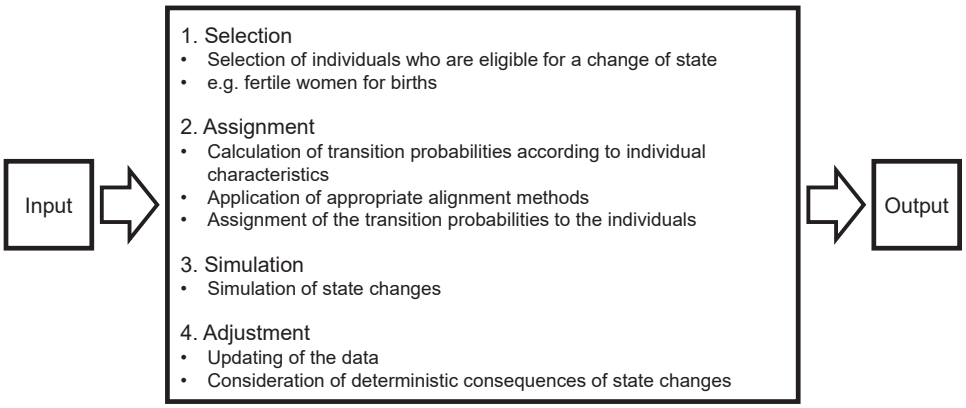


Figure 1: Module structure for a discrete-time dynamic microsimulation

to prevent implausible changes of state and to ensure the consistency of the population. The potential subpopulation for the event of birth includes, for example, women of fertile age. In the second step, the transition probabilities are calculated according to individual combinations of characteristics and linked to the individuals. If external benchmarks are not reached, calibration methods for an adjustment (so-called alignment methods) can be applied. Then, the state of the following period is simulated based on stochastic processes. This part corresponds to the simulation in the actual sense, since the concrete change of state is conducted. Finally, the population is updated according to all direct and indirect consequences of the simulated state changes. It should be noted that the exact structure of a module is individually designed in different models. Likewise, probabilities or transition matrices can serve as module output (O'Donoghue et al., 2009).

The modular structure plays a major role, especially in discrete-time dynamic microsimulations, since the changes of states have to be simulated successively. In continuous-time dynamic microsimulations, however, state changes cannot be determined independently of each other. Therefore, the estimated waiting times in the individual states could be specified as module output. The state changes are then carried out in an extra step after the simulation of all waiting times. While the structure of the simulation in continuous-time variants should not influence the simulation results, it heavily influences the dynamics in discrete-time models. This is briefly demonstrated hereafter. Let Y and X be two random events that may represent state changes within the microsimulation study. There are two different approaches to obtain the joint probability $P(Y,X)$ (Schaich and Münnich, 2001):

$$P(Y,X) = P(Y) \times P(X|Y) = P(X) \times P(Y|X). \tag{1}$$

We see that $P(Y,X)$ can be obtained by means of the conditional probability $P(X|Y)$, but also via the conditional probability $P(Y|X)$. In general, discrete dynamics do not provide

an exact point of time when a given transition is realized. Thus, in the case of two interdependent or competing events, it is necessary to determine which event occurrs first in light of the simulation context. For instance, if the event of a birth is simulated prior to the event of a marriage, the probability of marriage can be conditioned on the event of birth (Burgard et al., 2020). That is to say, the order of the simulation modules has a direct impact on the simulation outcomes. This motivated van Imhoff and Post (1998) to investigate three different strategies for organizing the modular structure. The first strategy is a randomized order of events, which, however, is hardly used in practice. Another possibility is a two-stage simulation of competing events, whereby the first stage simulates the occurrence of at least one of the competing events and the second stage the concrete event. As a third way, the sequence of modules or events of the microsimulation is considered in the modeling process (van Imhoff and Post, 1998).

For the basic functionality of microsimulations, it is generally not necessary to divide the population dynamics into different modules. However, the modularization provides clear practical advantages for the handling and the transparency of the simulation. Modularization allows individual modules to be easily adapted, exchanged and compared. It creates a flexible structure that allows the model to be further developed and adapted for further research questions. Moreover, it also facilitates working on different modules individually as well as in project teams (Lawson, 2014). In addition to that, modularization allows for the inclusion of module-specific debugging devices (O'Donoghue et al., 2009). The module can be written such that potential errors are detectable and precisely displayable. Ideally, the user can be informed about the reason for termination, otherwise at least about the exact position within the module. Additionally, plausibility checks within the modules are a useful extension to ensure data consistency. These checks verify whether the status changes have occurred even in the predefined subpopulation and whether implausible combinations of characteristics occur. Naturally, a modular structure is implemented as standard in many existing microsimulation tools such as LIAM2 (O'Donoghue et al., 2009; de Menten et al., 2014).

Nevertheless, using a modularized simulation structure also has some downsides. As mentioned before, the order of modules directly influences the simulation outcomes. Thus, the segmentation of population dynamics has to be conducted carefully with suitable theoretical justification. What is more, there is an ongoing debate to what extent probability estimation methods that are applied within each of the modules induce systematic errors across modules. For instance, a regression model may produce independent error terms in a given module. Still, these errors may not be independent from the error terms of another module, which may cause inferential problems. Hence, if a modularized structure is implemented, the simulation outcomes have to be carefully investigated with respect to these issues.

3.2. Step 5: Population dynamics

After the module sequence is defined and the modules are created, the dynamics for the projection of $\tilde{U}$ have to be established. They mark the underlying processes that drive the evolution of $\tilde{U}$ over the simulation horizon S. The nature and data requirements for

the projection depend on the type of microsimulation that is chosen in Step 1. In case of static microsimulations, only a few selected population characteristics evolve over time. The projection is often deterministic and can be performed without additional data sources. Typically, a set of scenarios for the selected variables is created. The base population evolves through the interaction of the remaining variables with them (Li and O'Donoghue, 2013).

In case of dynamic microsimulations, the projection is more sophisticated. Since all population characteristics evolve over time, the initialization of multivariate stochastic processes for $\tilde{U}$ is necessary. These processes need to resemble all relevant dynamics of the real population U as closely as possible to allow for genuine simulation outcomes. An essential concept for this purpose is called state transition, which we briefly explain hereafter. Let Y be a population characteristic with J different realizations within the finite state space $\mathscr{Y} = \{Y_1, ..., Y_J\}$. For instance, if a microsimulation on long-term demand is conducted Y may correspond to micro-level care dependency and its realizations could resemble different degrees of care dependency. Based on the theoretical developments of Burgard et al. (2019b), a state transition is defined as follows:

Definition 1 *Let $y_u^{(s)}$ be the realized value of Y for a unit $u \in \tilde{U}$ in period $s \in S$. A state transition is the outcome of a stochastic process where $y_u^{(s+1)} = Y_j$ and $y_u^{(s)} = Y_k$ with $Y_j, Y_k \in \mathscr{Y}$ and $Y_j \neq Y_k$. Its probability is given by $\pi_u^{(s+1)jk} := P(y_u^{(s+1)} = Y_j | y_u^{(s)} = Y_k)$.*

Accordingly, a state transition is a change in the realized value of a population characteristic for a given unit from one simulation period to the next. Recalling the long-term care example, a state transition would then correspond to a change of micro-level care dependency. In light of the previous comments, the probability $\pi_u^{(s+1)jk}$ must be determined such that the overall evolution of $\tilde{U}$ is realistic with respect to U. This is achieved by considering suitable data sources, such as a (panel) survey sample $D \subset U$. If a corresponding data set is available, transition probabilities can be quantified based on statistical models. In the first step, the statistical relation between transition probabilities and observed auxiliary variable realizations is estimated over all sampled individuals $i \in D$. In the next step, $\pi_u^{(s+1)jk}$ is determined via model prediction by using the realized values of the auxiliary variables for $u \in \tilde{U}$ in the simulation period $s+1$.

However, the exact methodology for estimation and projection depends on the concept of time that is chosen in Step 1. Recall that we distinguish between discrete-time and continuous-time dynamic simulations. For discrete-time, the simulation horizon $S := \{1, ..., T\}$ is a finite set of periods, such as months within a year. State transitions can only occur from one period to the next. In this setting, common approaches are generalized linear (mixed) models for the quantification of odds, such as logit models (McCullagh and Nelder, 1989; Greene, 2003). For continuous-time, the simulation horizon $S := [1, T]$ is a closed interval. State transitions may occur at any given point within this interval. In that case, estimation and prediction are performed using survival analysis, for instance via proportional hazard models (Cox, 1972; McCullagh and Nelder, 1989). Further, note that there are also models whose dynamics rely on Markovian processes with infinite state spaces, such as random walks for income simulation (Muennig et al., 2016).

Another important aspect of population projection is the consistency of simulated transition rates in $\tilde{U}$ to observed real-world realization frequencies in U. Let

$$\tau^{(t)k} := \sum_{i \in U} y_i^{(t)k}, \quad y_i^{(t)k} = \begin{cases} 1 & \text{if } y_i^{(t)} = Y_k \\ 0 & \text{else} \end{cases} \tag{2}$$

be the absolute frequency of Y_k in the real population U for a point of time t related to the simulation period $s+1$. Revisiting the long-term care example again, this figure may correspond to the number of individuals in a population that have a specific degree of care dependency. A corresponding figure could be known, for instance, from administrative records. In dynamic microsimulations, it is often the case that

$$\sum_{u \in \tilde{U}} \sum_{j \in \mathscr{Y}} \pi_u^{(s+1)jk} \neq \tau^{(t)k}. \tag{3}$$

The formula indicates that the simulated transition dynamics do not reproduce the empirically observed frequency for Y_k properly. This inconsistency may intensify over subsequent simulation periods and can lead to an implausible evolution of $\tilde{U}$. The latter ultimately causes the simulation outcomes to be not reliable for U, which is the main purpose of microsimulation studies. In order to ensure consistency in this case, so-called alignment methods are often applied (Li and O'Donoghue, 2014). These are (algorithmic) procedures that modify the transition probabilities such that they fit external benchmarks. Recently, several methodologies to achieve this have been proposed. Exemplary approaches are ex-post alignment via logit scaling (Stephensen, 2016) and parameter alignment via constrained maximum likelihood estimation (Burgard et al., 2019b).

3.3. Step 6: Performing the simulation

As dynamic microsimulations are based on stochastic processes, new populations are generated in each simulation run. Especially, if there are many individuals in the base population, it is not often possible to save them separately for each period and simulation run. Still, it is necessary to be able to reproduce the simulation results at any time. When conducting simulation studies, it is common practice to set seeds in order to repeat the random processes. In the sense of open and reproducible research, it is desirable to publish the seeds with the simulation code (Kleiber and Zeileis, 2012). In the case of error messages during the simulation, setting seeds enables the subsequent replication and analysis of the whole process.

Checking for plausibility and possible errors plays an important role not only within modules but also during the entire simulation. In order to identify potential causes in a targeted manner, predefined queries should be implemented at several points during the simulation process. The focus is on the functionality of the combination and interaction of different modules.

4. Output analysis

4.1. Step 7: Analysis of results

A big advantage of microsimulation models is providing information about possible impacts on a population, given the implemented scenarios. These advantages can only be converted into practical use if the analysis of the produced information is done properly. Several aspects have to be taken into account. First, the data has to be analyzed to prevent programming errors or logical errors in the simulation. Second, an uncertainty analysis should be performed to identify different sources of variation. And finally, the output of the simulation has to be analyzed concerning the research question. This includes both, the analysis of the final simulation states but also the processes that lead to the final results. Fourth, the analysis results have to be visualized. The visualization helps to understand the output and provides a good basis for the dissemination of the results.

4.1.1 Programming and logical errors

Even though Step 3 and Step 4 already include several plausibility checks, oftentimes problems in the coding or setup of the simulation only become apparent after a full simulation run. Surprising results may stem from non-linear population dynamics or errors in the code or setup of the simulation. It is therefore of utmost importance to first investigate the results of the simulation to the extent that outcomes seem sensible and the inner logic of the data set are met. If this is not the case, it is necessary to revisit the code and to explore how the results may be explained by the given process.

4.1.2 Uncertainty analysis

One major challenge in microsimulation modeling is the assessment of uncertainty. Typically, when analysing estimates, confidence intervals are calculated to quantify the uncertainty. Especially in dynamic microsimulations, the degree of complexity is high making a simple determination of confidence intervals hardly possible (Lappo, 2012). First of all, the potential sources of uncertainty should be identified. These depend on the type of modeling. Different types of uncertainty in microsimulation models can be distinguished (e.g. Lappo, 2015; Godemé et al. 2013, Sharif et al., 2012):

- Monte Carlo error
- Parameter uncertainty
- Structural uncertainty
- Uncertainty from the base population

The Monte Carlo error is a result of the stochastic processes and therefore occurs especially in case of dynamic microsimulations. However, behavioral changes in static simulations can also cause Monte Carlo errors. Parameter uncertainty is directly linked to

the models on which the microsimulation processes are based. If these are estimated on sample data, they are directly related to sampling uncertainty. Even assumption-based parameters are associated with, in this case subjective, uncertainty (Sharif et al., 2012). Structural uncertainty is primarily due to the type of modeling. This can be the type of estimation of transition probabilities or survival times on the one hand, but also the type of the entire microsimulation on the other hand. Since many microsimulations use survey data as base population, the uncertainty of the sampling must be taken into account. In the case of synthetic data sets, in turn, different sources of uncertainty arise, for example, from underlying data sources, used methods, parameters and stochastic processes in the preparation.

For the consideration of sampling uncertainty in static microsimulations through the application of standard variance estimation techniques, there are already useful examples (Lappo, 2012, Godemé et al. 2013). In the case of dynamic modeling the estimation of confidence intervals is much more difficult due to the complexity of the different sources. Sharif et al. (2012) and Sharif et al. (2017) propose techniques for the estimation of confidence intervals for the consideration of parameter uncertainty in dynamic disease microsimulation models. Petrik et al. (2018) estimate parameter uncertainty for an activity-based microsimulation model.

A possibility for quantifying the influence of various factors on univariate target values is variance-based sensitivity analysis as described in Burgard and Schmaus (2019). Here, the focus is not on estimating confidence intervals, but on measuring and comparing different influencing variables. The influencing variables can be selected variably, but must be pairwise independent. These factors may encompass all inputs that are to some extent wake. The goal of the sensitivity analysis is to attribute to the input factors a certain amount of variation observed in the target variable. For example different choices of scenarios or different parameter modeling strategies. Thus, sensitivity analyses are ideally suited for the selection of influential models for the later determination of confidence intervals. See Saltelli et al. (2008) for a comprehensive study of sensitivity analysis methods.

4.2. Hypothesis evaluation and result visualization

The hypotheses stated in Step 1 have to be checked. After conducting the microsimulation, it is necessary to evaluate whether the hypothesized outcome is a realistic development or not. It is possible to state the probability of the hypothesis to be true given the simulation evolves as the population will evolve. Of course, this condition is rarely possible to assume, and ex-ante, impossible to check in most cases. Especially, if the microsimulation is projecting the population for a long time horizon. The visualization of the results can considerably help the understanding of the simulation. Besides easing the simultaneous consideration of the measures used for the analysis it also helps to communicate the results to third persons and hence is also necessary for the dissemination of the results.

4.3. Step 8: Dissemination

The dissemination stage aims at disclosing knowledge acquired throughout the microsimulation study. To disseminate the study, it must be ensured that planned dissemination products, such as code, data and project reports are updated. In addition, the products must be available in such a way that they are comprehensible to outsiders and comply with legal requirements, such as publication standards.

The main focus of dissemination is to provide all interested parties with open access to resources related to the microsimulation study while respecting intellectual property rights. This includes, in particular, the provision of open access to peer-reviewed scientific publications, to research data and archival facilities for research results (European Commission, 2008). In particular in the case of microsimulation studies, however, open access to data cannot be granted for reasons of data protection and potential property rights to the data. The development of a security concept to guarantee privacy protection is to make data accessible through a research data center.

Additional dissemination strategies include the presentation of project research at conferences, organization of workshops and maintenance of a project website providing information about the project in general, conference contributions and publications related to the project. A project website also offers the possibility of setting up a mailing list to keep the interested public up to date. Furthermore, there are also associations such as the *International Microsimulation Association* that specifically aim at the dissemination of knowledge in the area of microsimulation (e.g. IMA, n.d.). For all dissemination strategies, especially when providing code and data, it is essential to have a contact person who accepts inquiries and supports users in the case of problems.

4.4. Step 9: Evaluation

The evaluation assesses all steps of the microsimulation study. It can be conducted either at the end or on an ongoing basis. The evaluation is based on the information gathered at the various steps and takes the experience from users, contributors and researchers into account. Continuously collected quality indicators are compiled to assess the quality of the individual preceding steps of the microsimulation study. Some steps, however, require specific measures such as the use of questionnaires to obtain information on the user-friendliness of the microsimulation study or to assess the effectiveness of the chosen dissemination strategies. As a result of the evaluation, an action plan is agreed upon. The implementation of the adopted actions will then again be part of the next round of evaluation (UNECE, 2013).

The complete business process model for conducting microsimulation studies is summarized in Figure 2.

5. Conclusion

Microsimulation methods play a more and more important role in policy support as well as in economic and social research. Major emphasis by now was laid on developing

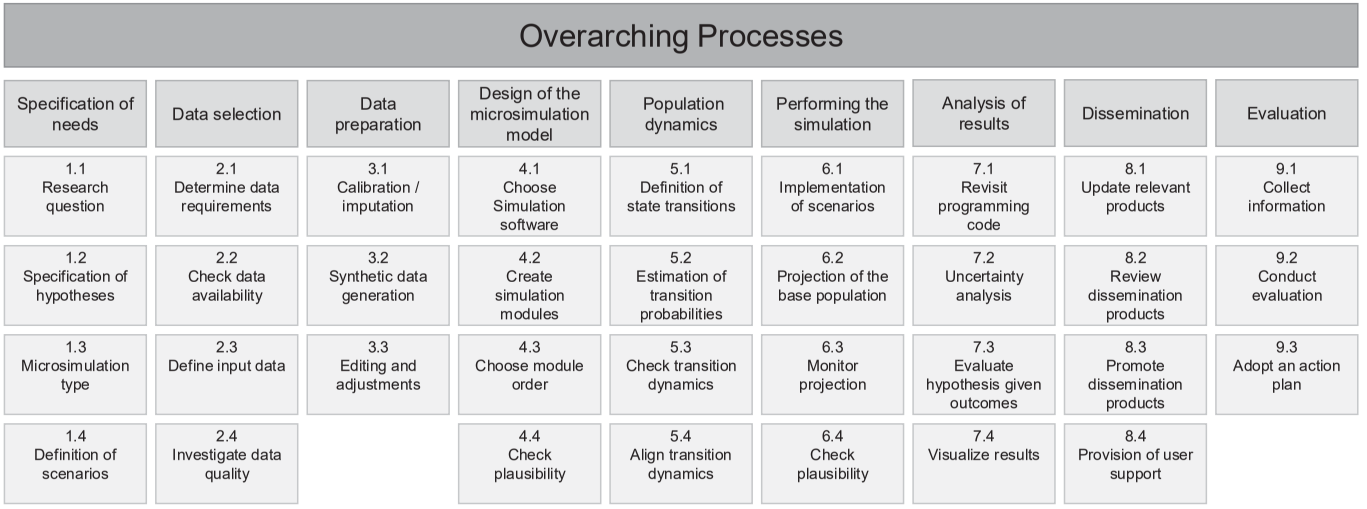


Figure 2: Generic business process model

important applications in many different areas. Less attention was put on the entire statistical production process. This becomes essentially important since the accuracy of the microsimulation heavily depends on data availability, data use, the core simulation, as well as the analysis considering all preceding steps.

The present article provided a general view of implementing a statistics business model that includes the different steps that have to be considered to establish an accurate microsimulation. The proposed model is based on UNECE (2013) on behalf of the international statistical community aiming at providing a general procedure that is widely accepted in the international statistical system. Further, it furnished the implementation of open and reproducible microsimulations as research and policy tool.

Acknowledgements

This research was conducted within the research group FOR 2559 *Sektorenübergreifendes kleinräumiges Mikrosimulationsmodell (MikroSim)*, which is funded by the German Research Foundation. We kindly thank for the financial support.

REFERENCES

ALFONS, A., FILZMOSER, P., HULLIGER, B., KOLB, J. P., KRAFT, S., MÜNNICH, R., TEMPL, M., (2011a). Synthetic data generation of SILC data. AMELI Research Project Report WP6-D6, 2.

ALFONS, A., KRAFT, S., TEMPL, M., FILZMOSER, P., (2011b). Simulation of close-to-reality population data for household surveys with application to EU-SILC. *Statistical Methods & Applications*, 20(3), pp. 383–407.

BÉLANGER, A., SABOURIN, P., (2017). Microsimulation and Population Dynamics: An Introduction to Modgen 12. Springer.

BOURGUIGNON, F., SPADARO, A., (2006). Microsimulation as a tool for evaluating redistribution policies. *The Journal of Economic Inequality*, Vol. 4, pp. 77–106.

BURGARD, J. P., KRAUSE, J., MERKLE, H., MÜNNICH, R., SCHMAUS, S., (2019a). Conducting a dynamic microsimulation for care research: Data generation, transition probabilities and sensitivity analysis. In *Stochastic Models, Statistics and Their Applications*. A. Steland, E. Rafajłowicz and O. Okhrin (eds.) Cham: Springer International Publishing, pp. 269–290.

BURGARD, J. P., KRAUSE, J., SCHMAUS, S., (2019b). Estimation of regional transition probabilities for spatial dynamic microsimulations from survey data lacking in regional detail. *Research Papers in Economics*, No. 12/19, Trier University.

BURGARD, J. P., KRAUSE, J., MERKLE, H., MÜNNICH, R., SCHMAUS, S., (2020). Dynamische Mikrosimulationen zur Analyse und Planung regionaler Versorgungsstrukturen in der Pflege. In *Mikrosimulationen - Methodische Grundlagen und ausgewählte Anwendungsfelder*. M. Hannappel and J. Kopp (eds.) Wiesbaden: Springer VS, pp. 283–313.

BURGARD, J. P., MÜNNICH, R. T., RUPP, M., (2019c). A generalized calibration approach ensuring coherent estimates with small area constraints (No. 10/19). *Research Papers in Economics*.

BURGARD, J. P., SCHMAUS, S., (2019). Sensitivity analysis for dynamic microsimulation models (No. 15/19). Research Papers in Economics, Trier University.

CHEN, J., QIN, J., (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, Vol. 80, pp. 107–116.

COX, D. R., (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 9, pp. 439–455.

DE MENTEN, G., DEKKERS, G., BRYON, G., LIÉGEOIS, P., O'DONOGHUE, C., (2014). Liam2: a new open source development tool for discrete-time dynamic microsimulation models. *Journal of Artificial Societies and Social Simulation*, Vol. 17, p. 9.

DEKKERS, G., (2015). The simulation properties of microsimulation models with static and dynamic ageing – a brief guide into choosing one type of model over the other.*International Journal of Microsimulation*, Vol. 8, pp. 97–109.

DEKKERS, G., CUMPSTON, R., (2012). On weights in dynamic-ageing microsimulation models. *The International Journal of Microsimulation*, Vol. 5(2), pp. 59–65.

DEVILLE, J., SÄRNDAL, C., (1992). Calibration estimators in survey sampling.*Journal ofthe American Statistical Association*, 87, pp. 376–382.

DRECHSLER, J., (2011). Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Vol. 201. Springer Science & Business Media.

EUROPEAN COMMISSION, (2008). Commission Recommendation of 10 April 2008 on the management of intellectual property in knowledge transfer activities and Code of Practice for universities and other public research organisations (notified under document number C(2008) 1329) (Text with EEA relevance). *Official Journal of the European Union (L 146)*, Vol. 51, pp. 19–24.

FAVREAULT, M. M., SMITH, K. E., JOHNSON, R. W., (2015). The dynamic simulation of income model (DYNASIM). Research Report at Urban Institute, Washington DC.

GOEDEME, T., VAN DEN BOSCH, K., SALANAUSKAITE, L., VERBIST, G., (2013). Testing the statistical significance of microsimulation results: A plea. International Journal of Microsimulation, 6(3), pp. 50–77.

GREENE, W. H., (2003). *Econometric analysis* (5 ed.) New Jersey: Prentice Hall.

HANNAPPEL, M., TROITZSCH, K. G., (2015). Mikrosimulationsmodelle. In N.Braun, N.J.Saam (eds): *Modellbildung und Simulation in den Sozialwissenschaften,* (pp. 455–489). Springer VS, Wiesbaden.

HAZIZA, D., BEAUMONT, J. F., (2017). Construction of weights in surveys: A review. *Statistical Science*, Vol. 32, pp. 206–226.

HUANG, Z.; WILLIAMSON, P., (2001). A Comparison of Synthetic Reconstruction and Combinatorial Optimisation Approaches to the Creation of Small-Area Microdata. University of Liverpool. Department of Geography. Working Paper 2001/2.

KLEIBER, C., ZEILEIS, A., (2013). Reproducible econometric simulations. *Journal of Econometric Methods*, Vol. 2, pp. 89–99.

LAPPO, S., (2015). Uncertainty in microsimulation, Master's Thesis, University of Helsinki.

LI, J., O'DONOGHUE, C., (2013). A survey of dynamic microsimulation models. Uses, model structure and methodology. *International Journal of Microsimulation*, Vol. 6, pp. 3–55.

LI, J., O'DONOGHUE, C., (2014). Evaluating binary alignment methods in microsimulation models. *Journal of Artificial Societies and Social Simulation*, Vol. 17, pp. 1–15.

LOVELACE, R., DUMONT, M., (2016). Spatial microsimulation with R. Chapman and Hall/CRC.

MANNION, O., LAY-YEE, R., WRAPSON, W., DAVIS, P., PEARSON, J., (2012). JAMSIM: A microsimulation modelling policy tool. *Journal of Artificial Societies and Social Simulation*, Vol. 15, p. 8.

MARKHAM, F., YOUNG, M., DORAN, B., (2017). Improving spatial microsimulation estimates of health outcomes by including geographic indicators of health behaviour: The example of problem gambling. *Health & Place*, Vol. 46, pp. 29–36.

MCCULLAGH, P., NELDER, J. A., (1989). *Generalized linear models* (2 ed.), Vol. 37 of *Monographs on Statistics and Applied Probability* London: Chapman and Hall.

MUENNIG, P.A., MOHIT, B., WU, J., JIA, H., ROSEN, Z., (2016). Coest effectiveness of the earned income tax credit as health policy investiment. *American Journal of Preventive Medicine*, Vol. 51(6), pp. 874–881.

MÜNNICH R, SCHÜRLE J., (2003). On the simulation of complex universes in the case of applying the GermanMicrocensus. DACSEIS research paper series No. 4, University of Tübingen.

MURTHY, D., (2012). Towards a sociological understanding of social media: theorizing Twitter. *Sociology*, Vol. 46(6), pp. 1–15.

O'DONOGHUE, C., (2001). Dynamic Microsimulation: A Methodological Survey. *Brazilian Electronic Journal of Economics*, Vol. 4, p. 77.

O'DONOGHUE, C., LENNON, J., HYNES, S., (2009). The Life-cycle Income Analysis Model (LIAM): a study of a flexible dynamic microsimulation modelling computing framework. *International Journal of Microsimulation*, Vol. 2, pp. 16–31.

O'DONOGHUE, C., DEKKERS, G., (2018). Increasing the impact of dynamic microsimulation modelling. *International Journal of Microsimulation*, Vol. 11, pp. 61–96.

ORCUTT, G. H., (1957). A new type of socio-economic system. *The review of economics and statistics*, 58, pp. 116–123.

PETRIK, O., ADNAN, M., BASAK, K., BEN-AKIVA, M., (2018). Uncertainty analysis of an activity-based microsimulation model for Singapore. Future Generation Computer Systems.

PICHON-RIVIERE, A., AUGUSTOVSKI, F., BARDACH, A., COLANTONIO, L., (2011). Development and validation of a microsimulation economic model to evaluate the disease burden associated with smoking and the cost-effectiveness of tobacco control interventions in Latin America. *Value in Health*, Vol. 14, S51–S59.

RAHMAN, A., HARDING, A., (2017). Small area estimation and microsimulation modeling. Boca Raton: CRC Press, Taylor & Francis Group.

SALTELLI, A., RATTO, M., TERRY, A., CAMPOLOGNO, F., CARIBONI, J., GATELLI, D., SAISANA, M., TARANTOLA, S., (2008). Global sensitivity analysis. The Primer. Chichester: John Wiley & Sons.

SÄRNDAL, C.-E., (2007). The calibration approach in survey theory and practice. *Survey Methodology*, Vol. 33, pp. 99–119.

SCHAFER, J.L., GRAHAM, J. W., (2002). Missing data: Our view of the state of the art. *Psychological Methods*, Vol. 7, pp. 147–177.

SCHAICH, E., MÜNNICH, R., (2001). Mathematische Statistik für Ökonomen. Vahlen.

SHARIF, B., KOPEC, J. A., WONG, H., FINES, P., SAYRE, E. C., LIU, R. R., WOLFSON, M. C., (2012). Uncertainty analysis in population-based disease microsimulation models. Epidemiology Research International, 2012.

SHARIF, B., WONG, H., ANIS, A. H., KOPEC, J. A., (2017). A practical ANOVA approach for uncertainty analysis in population-based disease microsimulation models. *Value in Health*, Vol. 20(4), pp. 710–717.

SMITH, D.M., CLARKE, G.P., HARLAND, K., (2009). Improving the synthetic data generation process in spatial microsimulation models. *Environment and Planning A: Economy and Space*, Vol. 41, pp. 1251–1268.

SPIELAUER, M., (2006). The "Life Course" model, a competing risk cohort microsimulation model: source code and basic concepts of the generic microsimulation programming language Modgen, MPIDR WORKING PAPER 2006–046.

SPIELAUER, M., (2009). Microsimulation approaches. Technical Report, Statistics Canada, Modeling Division.

SPIELAUER, M., (2011). What is Social Science Microsimulation? *Social Science Computer Review*, Vol. 29, pp. 9–20.

STEPHENSEN, P., (2016). Logit scaling: A general method for alignment in microsimulation models. *International Journal of Microsimulation*, Vol. 9, pp. 89–102.

SUTHERLAND, H., FIGARI, F., (2013). EUROMOD: the European Union tax-benefit microsimulation model. *International Journal of Microsimulation*, Vol. 6, pp. 4–26.

TANTON, R., (2018). Spatial microsimulation: Developments and potential future directions. *International Journal of Microsimulation*, Vol. 11(1), pp. 143–161.

TANTON, R., WILLIAMSON, P., HARDING, A., (2014). Comparing two methods of reweighting a survey file to small area data. International Journal of Microsimulation, 7(1), pp. 76–99.

UNECE, (2013). Generic statistical business process model. Version 5.0 – December 2013. The United Nations Economic Commission for Europe (UNECE). URL: http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0.

VAN IMHOFF, E., POST, W., (1998). Microsimulation methods for population projection. *Population: An English Selection*, Vol. 10, pp. 97–138.

WILLEKENS, F., (2017). Continuous-time microsimulation in longitudinal analysis. In *New Frontiers in Microsimulation Modelling*. A. Zaidi, A. Harding and P. Williamson (eds.), Routledge, pp. 413–436.

WILLIAMSON, P., (2013). An evaluation of two synthetic small-area microdata simulation methodologies: Synthetic reconstruction and combinatorial optimisation. In In Tanton and Edwards (eds): *Spatial microsimulation: A reference guide for users*. Springer, Dordrecht.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 212–236, DOI 10.21307/stattrans-2020-039
Received – 31.01.2020; accepted – 30.06.2020

Applying data synthesis for longitudinal business data across three countries

M. Jahangir Alam[1], Benoit Dostie[2], Jörg Drechsler[3], Lars Vilhuber[4]

ABSTRACT

Data on businesses collected by statistical agencies are challenging to protect. Many businesses have unique characteristics, and distributions of employment, sales, and profits are highly skewed. Attackers wishing to conduct identification attacks often have access to much more information than for any individual. As a consequence, most disclosure avoidance mechanisms fail to strike an acceptable balance between usefulness and confidentiality protection. Detailed aggregate statistics by geography or detailed industry classes are rare, public-use microdata on businesses are virtually inexistant, and access to confidential microdata can be burdensome. Synthetic microdata have been proposed as a secure mechanism to publish microdata, as part of a broader discussion of how to provide broader access to such data sets to researchers. In this article, we document an experiment to create analytically valid synthetic data, using the exact same model and methods previously employed for the United States, for data from two different countries: Canada (Longitudinal Employment Analysis Program (LEAP)) and Germany (Establishment History Panel (BHP)). We assess utility and protection, and provide an assessment of the feasibility of extending such an approach in a cost-effective way to other data.

Key words: business data, confidentiality, LBD, LEAP, BHP, synthetic.

1. Introduction

There is growing demand for firm-level data allowing detailed studies of firm dynamics. Recent examples include Bartelsman et al. (2009), who use cross-country firm-level data to study average post-entry behavior of young firms. Sedláček et al. (2017) use the Business Dynamics Statistics (BDS) to show the role of firm size in firm dynamics. However, such studies are made difficult due to the limited or restricted access to firm-level data.

Data on businesses collected by statistical agencies are challenging to protect. Many businesses have unique characteristics, and distributions of employment, sales and profits are highly skewed. Attackers wishing to conduct identification attacks often have access

[1]Department of Applied Economics, HEC Montréal, Canada & Department of Economics, Truman State University. USA. E-mail: jalam@truman.edu. ORCID: https://orcid.org/0000-0001-6478-114X.
[2]Department of Applied Economics, HEC Montréal. Canada. E-mail: benoit.dostie@hec.ca. ORCID: https://orcid.org/0000-0002-4133-2365.
[3]Institute for Employment Research. Germany. E-mail: joerg.drechsler@iab.de
[4]Cornell University. USA. E-mail: lars.vilhuber@cornell.edu. ORCID: https://orcid.org/0000-0001-5733-8932.

to much more information than for any individual. It is easy to find examples of firms and establishments that are so dominant in their industry or location that they would be immediately identified if data that included their survey responses or administratively collected data were publicly released. Finally, there are also greater financial incentives to identifying the particulars of some firms and their competitors.

As a consequence, most disclosure avoidance mechanisms fail to strike an acceptable balance between usefulness and confidentiality protection. Detailed aggregate statistics by geography or detailed industry classes are rare, public-use microdata on business are virtually inexistant,[5] and access to confidential microdata can be burdensome. It is not uncommon that access to establishment microdata, if granted at all, is provided through data enclaves (Research Data Centers), at headquarters of statistical agencies, or some other limited means, under strict security conditions. These restrictions on data access reduce the growth of knowledge by increasing the cost to researchers of accessing the data.

Synthetic microdata have been proposed as a secure mechanism to publish microdata (Drechsler et al., 2008; Drechsler, 2012; National Research Council, 2007; Jarmin et al., 2014), based on suggestions and methods first proposed by Rubin (1993) and Little (1993). Such data are part of a broader discussion of how to provide improved access to such data sets to researchers (Bender, 2009; Vilhuber, 2013; Abowd et al., 2004; Abowd et al., 2015).[6] For business data, synthetic business microdata were released in the United States (Kinney et al., 2011b) and in Germany (Drechsler, 2011b) in 2011. The former data set, called Synthetic Longitudinal Business Database (LBD) (SynLBD), was released to an easily web-accessible computing environment (Abowd et al., 2010), and combined with a validation mechanism. By making disclosable synthetic microdata available through a remotely accessible data server, combined with a validation server, the SynLBD approach alleviates some of the access restrictions associated with economic data. The approach is mutually beneficial to both agency and researchers. Researchers can access public use servers at little or no cost, and can later validate their model-based inferences on the full confidential microdata. Details about the modeling strategies used for the SynLBD can be found in Kinney et al. (2011b) and Kinney et al. (2011a).

In this article, we document an experiment to create analytically valid synthetic data, using the exact same model and methods previously used to create the SynLBD, but applied to data from two different countries: Canada (Longitudinal Employment Analysis Program (LEAP)) and Germany (Establishment History Panel (BHP)). We describe all three countries' data in Section 2.

In Canada, the Canadian Center for Data Development and Economic Research (CDER) was created in 2011 to allow Statistics Canada to make better use of its business data holdings, without compromising security. Secure access to business microdata for approved analytical research projects is done through a physical facility located in

[5] See Guzman et al. (2016) and Guzman et al. (2020) for an example of scraped, public-use microdata.

[6] For a recent overview of some, see Vilhuber et al. (2016b). See Drechsler (2011a) for a review of the theory and applications of the synthetic data methodology. Other access methods include secure data enclaves (e.g., research data centers of the U.S. Federal Statistical System, of the German Federal Employment Agency, others), and remote submission systems. We will comment on the latter in the conclusion.

Statistics Canada's headquarters.

CDER implements many risk mitigation measures to alleviate the security risks specific to micro-level business data including limits on tabular outputs, centralized vetting, monitoring of program logs. Access to the data is done through a Statistics Canada designed interface, in which actual observations cannot be viewed. But the cost of traveling to Ottawa remains the most significant barrier to access.

The Institute for Employment Research (IAB) in Germany also strictly regulates the access to its business data. All business data can be accessed exclusively onsite at the research data center (RDC) and only after the research proposal has been approved by the Federal Ministry of Labour and Social Affairs. All output is carefully checked by staff at the RDC and only cleared output can be removed from the RDC.

The experiment described in this paper aims not so much at finding the *best* synthetic data method for each file, but rather to assess the effectiveness of using a 'pre-packaged' method to cost-effectively generate synthetic data. In particular, while we could have used newer implementations of methods combined with a pre-defined or automated model (Nowok et al., 2016; Raab et al., 2018), we chose to use the exact SAS code used to create the original SynLBD. A brief synopsis of the method, and any adjustments we made to take into account structural data differences, are described in Section 3.

We verify the analytical validity of the synthetic data files so created along a variety of measures. First, we show how well average firm characteristics (gross employment, total payroll) in the synthetic data match those from the original data. We also consider how well the synthetic data replicates various measures of firm dynamics (entry and exit rates) and job flows (job creation and destruction rate). Second, we assess whether measures of economic growth vary between both data sets using dynamic panel data models. Finally, to assess the analytical validity from a more general perspective, we compute global validity measures based on the ideas of propensity score matching as proposed by Woo et al. (2009) and Snoke et al. (2018a).

To assess how protective the newly created synthetic database is, we estimate the probability that the synthetic first year equals the true first year given the synthetic fist year.

The rest of the paper is organized as follows. Section 2 describes the different data sources and summarizes which steps were taken to harmonize the data sets prior to the actual synthesis. Section 3 provides some background on the synthesis methods, limitations in the applications, and a discussion of some of the measures, which are used in Section 4 to evaluate the analytical validity of the generated data sets. Preliminary results regarding the achieved level of protection are included in Section 5. The paper concludes with a discussion of the implications of the study for future data synthesis projects.

2. Data

In this section, we briefly describe the structure of the three data sources.

2.1. United States: Longitudinal Business Database (LBD)

The LBD (U.S. Census Bureau, 2015) is created from the U.S. Census Bureau's Business Register (BR) by creating longitudinal links of establishments using name and address matching. The database has information on birth, death, location, industry, firm affiliation of employer establishments, and ownership by multi-establishment firms, as well as their employment over time, for nearly all sectors of the economy from 1976 through 2015 (as of this writing). It serves as a key linkage file as well as a research data set in its own right for numerous research articles, as well as a tabulation input to the U.S. Census Bureau's Business Dynamics Statistics (U.S. Census Bureau, 2017, BDS). Other statistics created from the underlying Business Register include the County Business Patterns (U.S. Census Bureau, 2016a, CBP) and the Statistics of U.S. Businesses (U.S. Census Bureau, 2016b, SBUSB). For a full description, readers should consult Jarmin et al. (2002). The key variables of interest for this experiment are birth and death dates, payroll, employment, and the industry coding of the establishment. Kinney et al. (2014b) explore a possible expansion of the synthesis methods described later to include location and firm affiliation. Note that information on payroll and employment does not come from individual-level wage records, as is the case for both the Canadian and German data sets described below, as well as for the Quarterly Workforce Indicators (Abowd et al., 2009) derived from the Longitudinal Employer-Household Dynamics (Vilhuber, 2018, LEHD) in the United States. Thus, methods that connect establishments based on labor flows (Benedetto et al., 2007; Hethey et al., 2010) are not employed. We also note that payroll is the cumulative sum of wages paid over the entire calendar year, whereas employment is measured as of March 12 of each year.

2.2. Canada: Longitudinal Employment Analysis Program (LEAP)

The LEAP (Statistics Canada, 2019b) contains information on annual employment for each employer business in all sectors of the Canadian economy. It covers incorporated and unincorporated businesses that issue at least one annual statement of remuneration paid (T4 slips) in any given calendar year. It excludes self-employed individuals or partnerships with non-salaried participants.

To construct the LEAP, Statistics Canada uses three sources of information: (1) T4 administrative data from the Canada Revenue Agency (CRA), (2) data from Statistics Canada's Business Register (Statistics Canada, 2019c), and (3) data from Statistics Canada's Survey of Employment, Payrolls and Hours (SEPH) (Statistics Canada, 2019a). In general, all employers in Canada provide employees with a T4 slip if they paid employment income, taxable allowances and benefits, or any other remuneration in any calendar year. The T4 information is reported to the tax agency, which in turn provides this information to Statistics Canada. The Business Register is Statistics Canada's central repository of baseline information on businesses and institutions operating in Canada. It

is used as the survey frame for all business related data sets. The objective of the SEPH is to provide monthly information on the level of earnings, the number of jobs, and hours worked by detailed industry at the national and provincial levels. To do so, it combines a census of approximately one million payroll deductions provided by the CRA, and the Business Payrolls Survey, a sample of 15,000 establishments.

The core LEAP contains four variables (1) a longitudinal Business Register Identifier (LBRID), (2) an industry classification, (3) payroll and (4) a measure of employment. The LBRID uniquely identifies each enterprise and is derived from the Business Register. To avoid "false" deaths and births due to mergers, restructuring or changes in reporting practices, Statistics Canada uses employment flows. Similar to Benedetto et al. (2007) and Hethey et al. (2010), the method compares the cluster of workers in each newly identified enterprise with all the clusters of workers in firms from the previous year. This comparison yields a new identifier (LBRID) derived from those of the BR. The industry classification comes from the BR for single-industry firms. If a firm operates in multiple industries, information on payroll from the SEPH is used to identify the industry in which the firm pays the highest payroll. Prior to 1991, information on industry was based on the SIC, but it is currently based on the North American Industrial Classification System (NAICS). We use the information at the NAICS four-digit (industry group) level. The firm's payroll is measured as the sum of all T4s reported to the CRA for the calendar year. Employment is measured either using Individual Labour Unit (ILU) or Average Labour Unit (ALU). ALUs are obtained by dividing the payroll by the average annual earnings in its industry/province/class category computed using the SEPH. ILUs are a head count of the number of T4 issued by the enterprise, with employees working for multiple employers split proportionately across firms according to their total annual payroll earned in each firm.

For the purpose of this experiment, we exclude the public sector (NAICS 61, 62, and 91), even though it is contained in the database, because it may not be accurately captured (Statistics Canada, 2019b). Statistics Canada does not publish any statistics for those sectors.

2.3. Germany: Establishment History Panel (BHP)

The core database for the Establishment History Panel is the German Social Security Data (GSSD), which is based on the integrated notification procedure for the health, pension and unemployment insurances, introduced in 1973. Employers report information on all their employees. Aggregating this information via an establishment identifier yields the Establishment History Panel (Bundesagentur für Arbeit, 2013, German abbreviation: BHP). We used data from 1975 until 2008, which at the time this project started was the most current data available for research. Information for the former Eastern German States is limited to the years 1992-2008.

Due to the purpose and structure of the GSSD, some variables present in the LBD are not available on the BHP. Firm-level information is not captured, and it is thus not known whether establishments are part of a multi-establishment employer. In 1999, reporting requirements were extended to all establishments; prior to that date, only es-

tablishments that had at least one employee covered by social security on the reference date June 30 of each year were subject to filing requirements. Payroll and employment are both based on a reference date of June 30, and are thus consistent point-in-time measures. Industries are identified according to the WZ 2003 classification system (Statistisches Bundesamt, 2003) at the five digit level.[7] We aggregated the industry information for this project using the first four digits of the coding system.

2.4. Harmonizing and Preprocessing

In all countries, the underlying data provide annual measures. However, SynLBD assumes a longitudinal (wide) structure of the data set, with invariant industry (and location). In all cases, the modal industry is chosen to represent the entity's industrial activity. Further adjustments made to the BHP for this project include estimating full-year payroll, creating time-consistent geographic information, and applying employment flow methods (Hethey et al., 2010) to adjust for spurious births and deaths in establishment identifiers. Drechsler et al. (2014b) provide a detailed description of the steps taken to harmonize the input data.

In both Canada and Germany, we encountered various technical and data-driven limitations. In all countries, data in the first year and last year are occasionally problematic, and such data were dropped. Both the German and the Canadian data experience some level of industry coding change, which may affect the classification of some entities. Furthermore, due to the nature of the underlying data, entities are establishments in Germany and the US, but employers in Canada.

After the various standardizations and choices made above, the data structure is intended to be comparable, as summarized in Table 1. The column "Nature" identifies the treatment of the variable in the synthesis process SynLBD.

Table 1: Variable descriptions and comparison

Name	Type	Description	US	Canada	Germany	Nature
Entity Identifier	identifier		Establishment	Employer	Establishment	Created
Industry code	Categorical	Various across countries	SIC3 (3-digit)	NAICS4 (4-digit)	WZ2003 (4-digit)	Unmodified
First year	Categorical	First year entity is observed		— firstyear —		Synthesized
Last year	Categorical	Last year entity is observed		— lastyear —		Synthesized
Year	Categorical	Year dating of annual variables		— year —		Derived
Employment	Continuous	Employment measure	Count (March 15)	ALU* (annual)	Count (June 30)	Synthesized
Payroll	Continuous	Payroll (annual)	Reported	Computed	Computed, Adjusted	Synthesized

* ALU = Average Labour Unit. See text for additional explanations.

[7] The WZ 2003 classification system is compliant with the requirements of the Statistical Classification of Economic Activities in the European Community (NACE Rev. 1.1), which is based on the International Standard Industrial Classification (ISIC Rev. 3.1).

3. Methodology

To create a partially synthetic database with analytic validity from longitudinal establishment data, Kinney et al. (2011a) synthesize the life-span of establishments, as well as the evolution of their employment, conditional on industry over that synthetic lifespan. Geography is not synthesized, but is suppressed from the released file (Kinney et al., 2011a). Applying this to the LBD, Kinney et al. (2011b) created the current version of the Synthetic LBD, based on the Standard Industrial Classification (SIC) and extending through 2000. Kinney et al. (2014a) describe efforts to create a new version of the Synthetic LBD, using a longer time series (through 2010) and newer industry coding (NAICS), while also adjusting and extending the models for improved analytic validity and the imputation of additional variables. In this paper, we refer to and re-use the older methodology, which we will call SynLBD. Our emphasis is on the comparability of results obtained for a given methodology across the various applications.

The general approach to data synthesis is to generate a joint posterior predictive distribution of $Y|X$ where Y are variables to be synthesized and X are unsynthesized variables. The synthetic data are generated by sampling new values from this distribution. In SynLBD, variables are synthesized in a sequential fashion, with categorical variables being generally processed first using a variant of Dirichlet-Multinomial models. Continuous variables are then synthesized using a normal linear regression model with kernel density-based transformation (Woodcock et al., 2009).[8] The synthesis models are run independently for each industry. SynLBD is implemented in SAS™, which is frequently used in national statistical offices.

To evaluate whether synthetic data algorithms developed in the U.S. can be adapted to generate similar synthetic data for other countries, Drechsler et al. (2014a) implement SynLBD to the German Longitudinal Business Database (GLBD). In this paper, we extend the analysis from the earlier paper, and extend the application to the Canadian context (SynLEAP).

3.1. Limitations

In all countries, the synthesis of certain industries failed to complete. In both Canada and the US, this number is less than 10. In Canada, they account for about 7 percent of the total number of observations (see Table 13 in the Online Appendix).

In the German case, our experiments were limited to only a handful of industries, due to a combination of time and software availability factors. The results should still be considered preliminary. In both countries, as outlined in Section 2, there are subtle but potentially important differences in the various variable definitions. Industry coding differs across all three countries, and the level of detail in each of the industry codings may affect the success and precision of the synthesis.[9]

[8] Kinney et al. (2014a) shift to a Classification and Regression Trees (CART) model with Bayesian bootstrap.

[9] STATISTICS CANADA et al. (1991), when comparing the 1987 US Standard Industrial Classification (SIC) to the 1980 Canadian SIC, already pointed out that the degree of specialization, the organization of production, and the size of the respective markets differed. Thus, the density of estab-

As noted in Section 2, entities are establishments in Germany and the US, but employers in Canada. SynLBD should work on any level of entity aggregation (see Kinney et al. (2014a) for an application to hierarchical firm data with both firm/employer and establishment level imputation). However, these differences may affect the observed density of the data within industry-year categories, and therefore the overall comparability.

Finally, due to a feature of SynLBD that we did not fully explore, synthesis of data in the last year of the data generally was of poor quality. For some industry-country pairs, this also happened in the first year. We dropped those observations.

3.2. Measuring outcomes

In order to assess the outcomes of the experiment, we inspect analytical validity by various measures and also evaluate the extent of confidentiality protection. To check analytical validity, we compare basic univariate time series between the synthetic and confidential data (employment, entity entry and exit rates, job creation and destruction rates), and the distribution of entities (firms and establishment, depending on country), employment, and payroll across time by industry. For a more complex assessment, we compute a dynamic panel data model of economic (employment) growth on each data set. We computed, but do not report here the confidence interval overlap measure (CIO) proposed by Karr et al. (2006) in all these evaluations.[10] The CIO is a popular measure when evaluating the validity for specific analyses. It evaluates how much the confidence intervals of the original data and the synthetic data overlap. We did not find this measure to be useful in our context. Most of our analyses are based on millions of records, and observed confidence intervals were so small that confidence intervals (almost) never overlap even when the estimates between the original data and the synthetic data are quite close.

To provide a more comprehensive measure of quality of the synthetic data relative to the confidential data, we compute the $pMSE$ (propensity score mean-squared error, Woo et al., 2009; Snoke et al., 2018b; Snoke et al., 2018a): the mean-squared error of the predicted probabilities (i.e., propensity scores) for those two databases. Specifically, $pMSE$ is a metric to assess how well we are able to discern the high distributional similarity between synthetic data and confidential data. We follow Woo et al. (2009) and Snoke et al. (2018b) to calculate the $pMSE$, using the following algorithm:

1. Append the n_1 rows of the confidential database X to the n_2 rows of the synthetic database X^s to create X^{comb} with $N = n_1 + n_2$ rows, where both X and X^s are in the long format.

2. Create a variable I_{et} denoting membership of an observation for entity e, $e = 1,\ldots,E$, at time point t, $t = 1,\ldots,T$, in the component databases, $I_{et} = \{1 : X^{comb}_{et} \in X^s\}$. I_{et} takes on values of 1 for the synthetic database and 0 for the confidential database.

lishments within each of the chosen categories is likely to affect the quality of the synthesis.

[10] The full parameter estimates and the computed CIO are available in our replication materials (Alam et al., 2020).

3. Fit the following generalised linear model to predict I

$$P(I_{et}=1) \quad = \quad g^{-1}(\beta_0+\beta_1 Emp_{et}+\beta_2 Pay_{et}+Age_{et}^T\beta_3+\lambda_t+\gamma_i), \qquad (1)$$

where Emp_{et} is log employment of entity e in year t, Pay_{et} is log payroll of entity e in year t, Age_{et} is a vector of age classes of entity e in year t, λ_t is a year fixed effect, γ_i is an time-invariant industry-specific effect for the industry classification i of entity e, and g is an appropriate link function (in this case, the logit link).

4. Calculate the predicted probabilities, $\hat{p}_{et}$.
5. Compute $pMSE = \frac{1}{N}\sum_{t=1}^{T}\sum_{e=1}^{E}(\hat{p}_{et}-c)^2$, where $c = n_2/N$.

If $n_1 = n_2$, $pMSE = 0$ means every $\hat{p}_{et} = 0.5$, and the two databases are distributionally indistinguishable, suggesting high analytical validity. While the number of records in the synthetic data typically matches the number of records in the original data, i.e., $n_1 = n_2$, this does not necessarily hold in our application. Although the synthesis process ensures that the total number of entities is the same in both data sets, the years in which the entities are observed will generally differ between the original data and the synthetic data and thus the number of records in the long format will not necessarily match between the two data sets. For this reason we follow Woo et al. (2009) and Snoke et al. (2018a) and use $c = n_2/N$ instead of fixing c at 0.5. Using this more general definition, c will always be the mean of the predicted propensity scores so that the $pMSE$ measures the average of the squared deviations from the mean, as intended.

Since the $pMSE$ depends on the number of predictors included in the propensity score model, Snoke et al. (2018a) derived the expected value and standard deviation for the $pMSE$ under the null hypothesis ($pMSE_0$) that the synthesis model is correct, i.e., it matches the true data generating process (Snoke et al., 2018a, Equation 1):

$$E[pMSE_0] = (k-1)(1-c)^2\frac{c}{N}$$

and

$$StDev[pMSE_0] = \sqrt{2(k-1)}(1-c)^2\frac{c}{N}$$

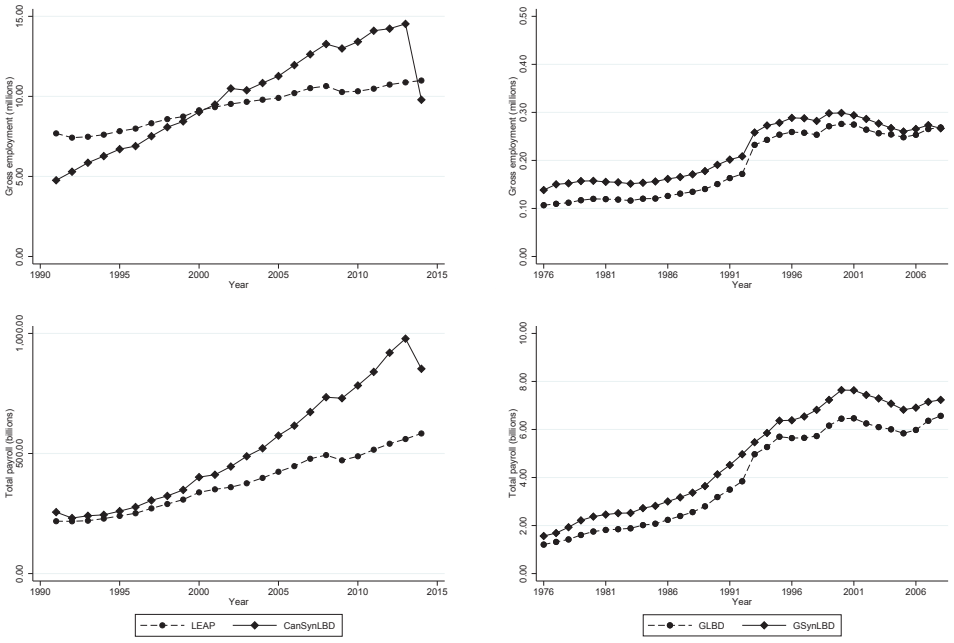
where k is the number of synthesized variables used in the propensity model. To measure the analytical validity of the synthetic data, they suggest looking at the *pMSE ratio*

$$pMSEratio = \frac{\widehat{pMSE}}{E[pMSE_0]}$$

and the *standardized pMSE*

$$pMSE_s = \frac{\widehat{pMSE} - E[pMSE_0]}{StDev[pMSE_0]},$$

where $\widehat{pMSE}$ is the estimated pMSE based on the data at hand. Under the null hypothesis, the $pMSE$ ratio has an expectation of 1 and the expectation of the standardized



(a) CanSynLBD (b) GSynLBD

Figure 1: Gross employment level (upper panels) and total payroll (lower panels) by year.

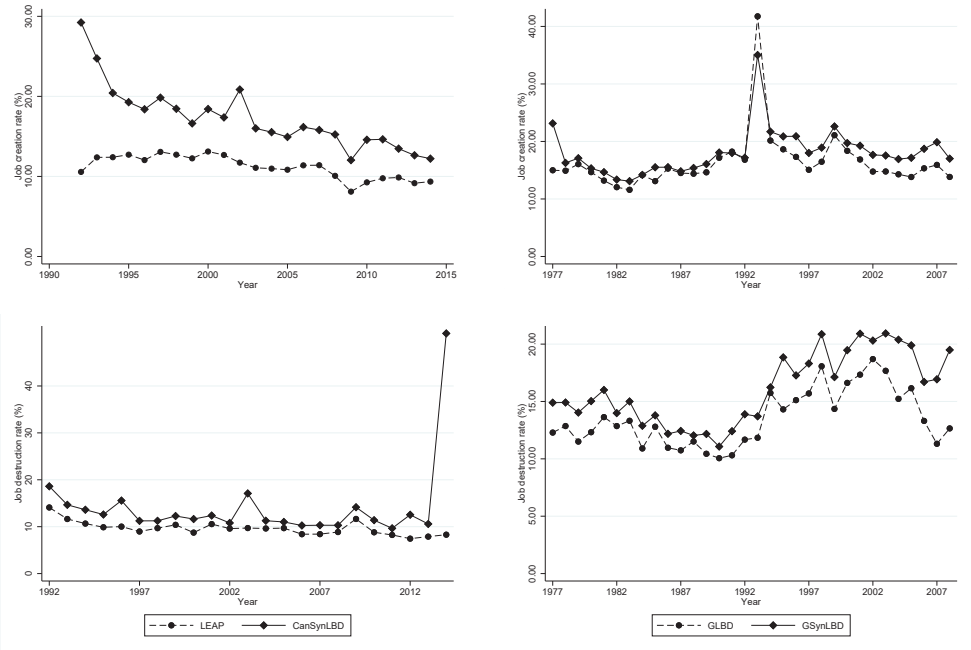
$pMSE_s$ is zero.

4. Analytical validity

In the following figures, the results for the Canadian data are shown in the left panels, and the German data in the right panels. In all cases, the Canadian data are reported for the entire private sector, including the manufacturing sector but excluding the public sector industries (NAICS 61, 62, and 91). German results are for two WZ2003 industries.

4.1. Entity Characteristics

Figure 1 shows a comparison between the synthetic data and the original data for gross employment level (upper panels) and total payroll (lower panels) by year. While the general trends are preserved for both data sources, the results for the German synthetic data resemble the trends from the original data more closely. For the Canadian data the positive trends over time are generally overestimated. However, in both cases, levels are mostly overestimated. These patterns are not robust. When considering the manufacturing sector in Canada (Figure 8 in the Online Appendix), trends are better matched, but a significant *negative* bias is present in levels.



(a) CanSynLBD (b) GSynLBD

Figure 2: Job creation rates (upper panels) and job destruction rates (lower panels) by year.

4.2. Dynamics of Job Flows

Key statistics commonly computed from business registers such as the LEAP or the BHP include job flows over time. Following Davis et al. (1996), job creation is defined as the sum of all employment gains from expanding firms from year $t-1$ to year t including entry firms. The job destruction rate is defined as the sum of all employment losses from contracting firms from year $t-1$ to year t including exiting firms. Figure 2 depicts job creation rates (upper panels) and destruction rates (lower panels). The general levels and trends are preserved for both data sources, but the time-series align more closely for the German data. Even the substantial increase in job creations in 1993, which can be attributed to the integration of the data from Eastern Germany after reunification, is remarkably well preserved in the synthetic data. Still, there seems to be a small but systematic overestimation of job creation and destruction rates in both synthetic data sources. The substantial deviation in the job destruction rate in the last year of CanSynLBD is an artefact requiring further investigation.[11]

[11] The results for the Canadian manufacturing sector are included in Figure 9 in the Online Appendix, and are comparable to the results for the entire private sector.

4.3. Entity Dynamics

To assess how well the synthetic data capture entity dynamics, we also compute entry and exit rates, i.e. how many new entities appear in the data and how many cease to exist relative to the population of entities in a specific year.[12] Figure 3 shows that those rates are very well preserved for both data sources.

Only the (delayed) re-unification spike in the entry rates in the German data is not preserved correctly. The confidential data show a large spike in entry rates in 1993. In that year, detailed information about Eastern German establishments was integrated for the first time. However, the synthetic data shows increased entry rates in the two previous years. We speculate that this occurs due to incomplete data in the confidential data: Establishments were successively integrated into the data starting in 1991, but many East German establishments did not report payroll and number of employees in the first two years. Thus, records existed in the original data, but the establishment size is reported as missing. Such a combination is not possible in the synthetic data. The synthesis models are constructed to ensure that whenever an establishment exists, it has to have a positive number of employees. Since entry rates are computed by looking at whether the employment information changed from missing to a positive value, most of the Eastern German establishments only exist from 1993 on-wards in the original data, but from 1991 in the synthetic data.

The second, smaller spike in the entry rate in the German data occurs in 1999. In that year, employers were required to report marginally employed workers for the first time. Some establishments exclusively employ marginally employed workers, and will thus appear for the first time in the data after 1999. The synthetic data preserves this pattern.

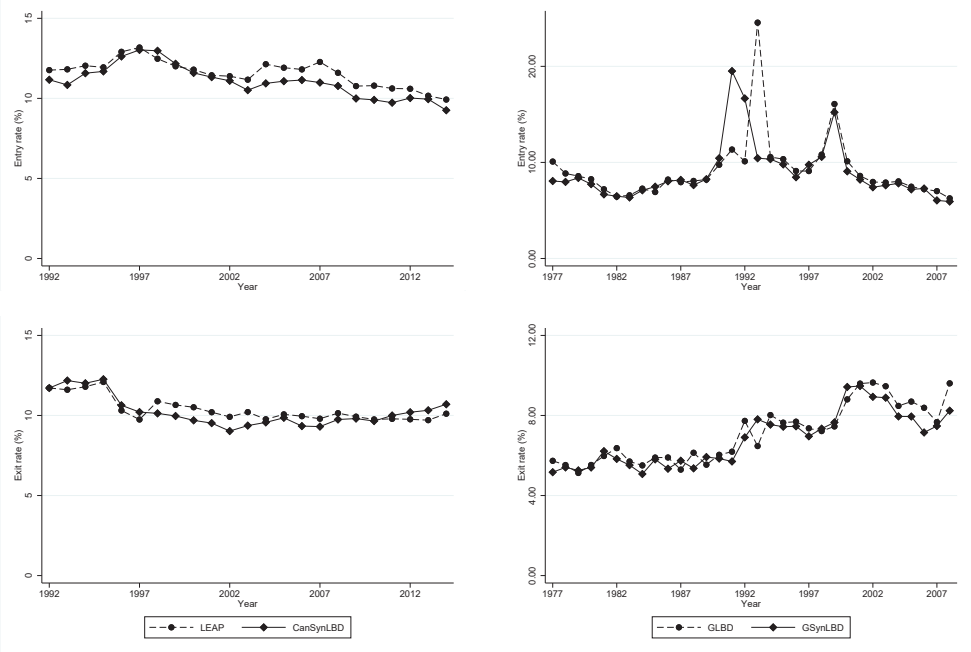
4.4. Distribution of variables across time and industry

The SynLBD code ensures that the total number of entities that ever exist within the considered time frame matches exactly between the original data and the synthetic data. But each entity's entry and exit date are synthesized, and the total number of entities at any particular point in time may differ, and with it employment and payroll. To investigate how well the information is preserved at any given point in time, we compute the following statistic:

$$x_{its} = X_{its} / \sum_i \sum_t X_{its}, \tag{2}$$

where i is the index for the industry (aggregated to the two digit level for the Canadian data), t is the index for the year and s denotes the data source (original or synthetic). $X_{its} = \sum_j X_{itsj},\ j = 1, \ldots, n_{its}$ is the variable of interest aggregated at the industry level and n_{its} is the number of entities in industry i at time point t in data source s. To compute the statistic provided in Equation (2), this number is then divided by the total of the variable of interest aggregated across all industries and years. Figure 4 plots the

[12] As described in Section 2, for both countries' data, corrections based on worker flows have been applied, correcting for any bias due to legal reconfiguration of economic entities.



(a) CanSynLBD (b) GSynLBD

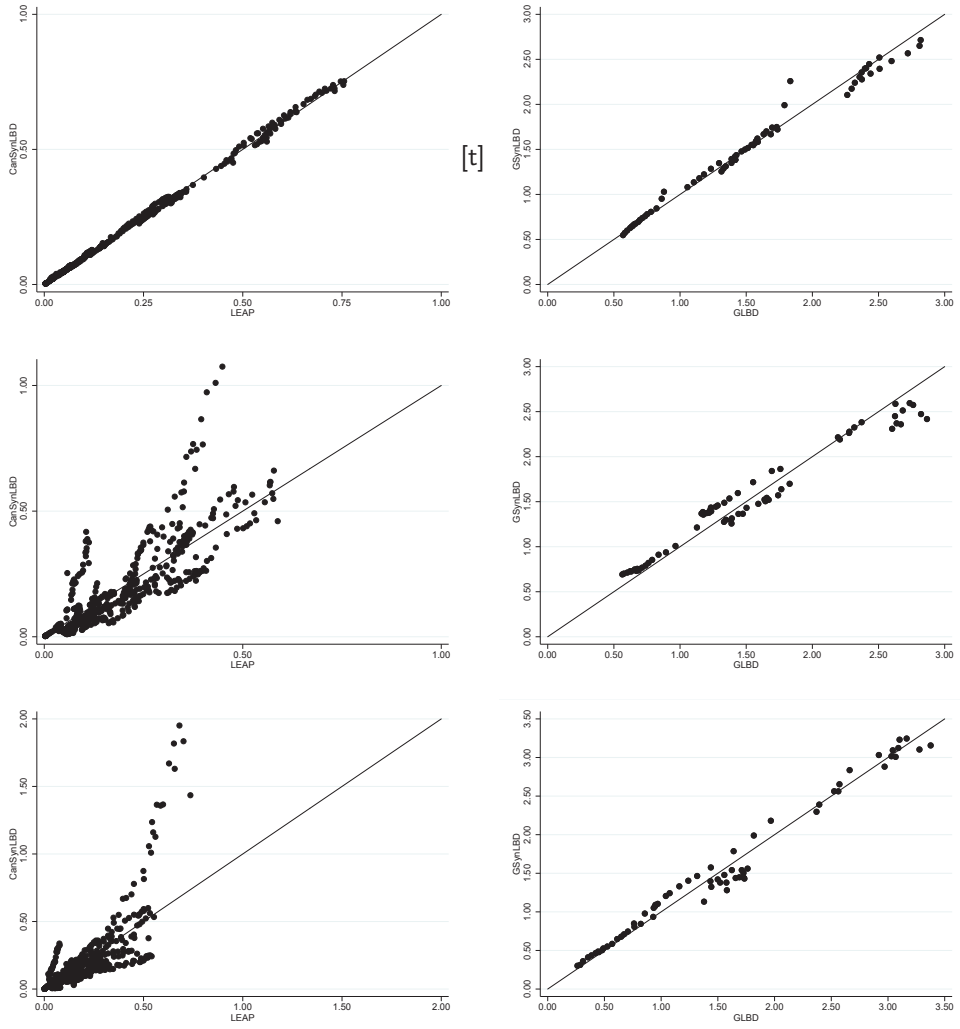
Figure 3: Entry rates (upper panels) and exit rates (lower panels) by year.

results from the original data against the results from the synthetic data for the number of entities, employment, and payroll. If the information is well preserved, all points should be close to the 45 degree line.

We find that the share of entities is well preserved for both data sources, but share of employment and share of payroll vary more in the Canadian data with an upward bias for the larger shares. It should be noted that the German data shown here and elsewhere in this paper only contain data from two industries, whereas the Canadian data contains nearly all available industry codes at the two digit level. Thus, results from Canada are expected to be more diverse. When only considering the Canadian manufacturing sector (see Figure 10 in the Online Appendix), less bias is present.

4.5. Modelling strategy

To assess how well the synthetic data perform in a more complex model and in the context of an analyst's modelling strategy, we simulate how a macroeconomist (the typical user of these data) might approach the problem of estimating a model for the evolution of employment if only the synthetic data are available. The analyst will consider both the literature and the data to propose a meaningful model. In doing so, a sequence of models will be proposed, and tests or theory brought to bear on their merits, potentially rejecting their appropriateness. In doing so, the outcome that the analyst obtains from



[t]

(a) CanSynLBD (b) GSynLBD

Figure 4: Share of entities (upper panels), share of employment (middle panels), and share of payroll (lower panels) by year and industry.

following that strategy using the synthetic data should not diverge substantially from the outcome they would obtain when using the (inaccessible) confidential data. The specific parameter estimates obtained, and the actual model retained, are not the goal of this exercise — the focus is on the process.

To do so, our analyst would start by using a base model (typically OLS), and then let economic and statistical theory suggest more appropriate models. In this case, we will estimate variants of a dynamic panel data model for the evolution of employment. For each model, tests can be specified to check whether the model is an appropriate

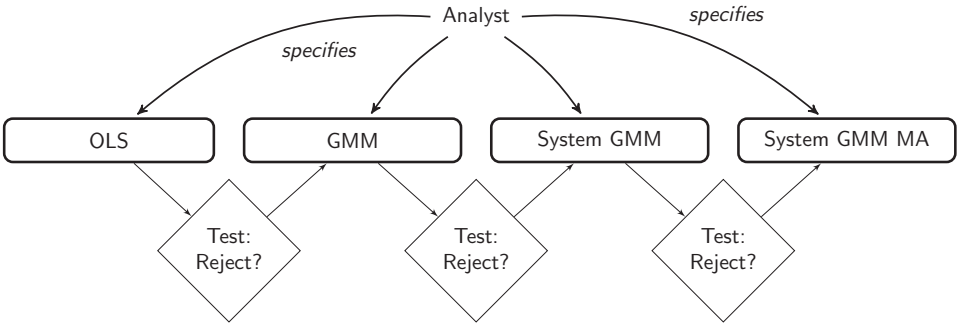


Figure 5: Modelling strategy of a hypothetical analyst

fit under a certain hypothesis.[13] The outcome of this exercise, illustrated by Figure 5, allows us to assess whether the synthetic data capture variability in economic growth due to industry, firm age and payroll — the key variables in the data — and whether the analyst might reasonable choose the same, or a closely related modelling strategy.

The base model is an OLS specification:

$$Emp_{et} = \beta_0 + \theta Emp_{e,t-1} + \eta Pay_{et} + Age_{et}^T \beta + \gamma_i + \lambda_t + \varepsilon_{et} \quad (3)$$

where Emp_{et} is log employment of entity e in year t, $Emp_{e,t-1}$ is its one year lag, Pay_{et} is the logarithm of payroll of entity e in year t, Age_{et} is a vector of dummy variables for age of entity e in year t, λ_t is a year effect, γ_i is a time-invariant industry-specific effect for each industry i, and ε_{et} is the disturbance term of entity e in year t. As $Emp_{e,t-1}$ is correlated with γ_i because $Emp_{e,t-1}$ is itself determined by time-invariant γ_i, OLS estimators are biased and inconsistent. To obtain consistent estimates of the parameters in the model, Arellano et al. (1991) suggest using generalized method of moments (GMM) estimation methods, as well as associated tests to assess the validity of the model. We also estimate the model using system GMM methods proposed by Arellano et al. (1995) and Blundell et al. (1998) (System GMM), as well as a variant of equation (3) that includes a first-order moving average in the error term ε_{et} (System GMM MA):

$$Emp_{et} = \beta_0 + \theta Emp_{e,t-1} + \eta Pay_{et} + Age_{et}^T \beta + \lambda_t + \alpha_e + \varepsilon_{et} + \varepsilon_{e,t-1} \quad (4)$$

where α_e is a time-invariant entity effect, which includes any time-invariant industry effects.

The Sargan test (Hansen, 1982; Arellano et al., 1991; Blundell et al., 2001) is used to assess the validity of the over-identifying restrictions. We also compute the z-score for the $m2$ test for zero autocorrelation in the first-differenced errors of order two (Arellano et al., 1991).

[13] We do not describe these models in more detail here, referring the reader to the literature instead, in particular Arellano et al. (1995) and Blundell et al. (1998).

An interesting derived effect is to consider the long-run effect of (log) payroll on (log) employment, or the elasticity of employment with respect to payroll. This can be estimated as

$$\eta^{\star} = \frac{\hat{\eta}}{1-\hat{\theta}}.$$

It is important that this model is close, but not identical to the model used to *synthesize* the data. In SynLBD, Emp_{et} is synthesized as $f(Emp_{e,t-1}, X_{et})$ (where X_{et} does not contain Pay_{et}), and $Pay_{et} = f(Pay_{e,t-1}, Emp_{et}, X_{et})$ (Kinney et al., 2011b, pg. 366). Thus, the model we chose is purposefully not (completely) congenial with the synthesis model, but the synthesis process of the SynLBD should preserve sufficient serial correlation in the data to be able to estimate these models.

We estimate each model and test statistics separately on confidential and synthetic data for the private sector (and for Canada, for the manufacturing sector). Detailed estimation results are reported in the Online Appendix. Here we focus on the two regression coefficients of major interest: θ and η, the coefficients for lagged employment and payroll, as well as the elasticity $\eta^{\star}$. Figure 6 plots the bias in the synthetic coefficients, i.e., $\theta_{synth} - \theta_{conf}$ and $\eta_{synth} - \eta_{conf}$, for all four models. While the detailed results in the Online Appendix confirm that all regression coefficients still have the same sign, all estimates plotted in Figure 6 show substantial bias in all models in all datasets (the OLS model for the German data being the only exception). Still, the computed elasticity $\eta^{\star}$ has very little bias in most models.

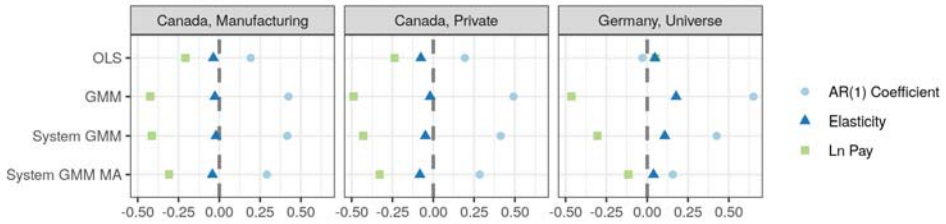


Figure 6: Bias in estimates of coefficients on pay and lagged employment
Note: For details on the estimated coefficients, see the Online Appendix.

However, we observe a striking pattern: The biases of the two regression coefficients are always symmetric, i.e. the sum of the biases of θ_{synth} and η_{synth} is close to zero in all models (and mostly cancel out in the computation of $\eta^{\star}$). This may simply be a feature of the modeling strategy pointed out earlier, which generates serial correlation with a slightly different structure. Another possible explanation could be that the model is poorly identified because of multicollinearity generating a ridge for the estimated coefficients. The estimated coefficients would be highly unstable in this case even in the original data and thus it would not be surprising to find substantial differences between the coefficients from the original data and the coefficients from the synthetic data. Better understanding this phenomenon will be an interesting area of future research.

While the bias in coefficients is quite consistent across countries and models, speci-

Table 2: m2 and Sargan tests by country

		Canada		Germany	
Model	Test	Confidential	Synthetic	Confidential	Synthetic
GMM	m2	-14.5	-27.54	-2.51	-4.13
	Sargan test	69000	15000	3600	2000
System GMM	m2	-11.43	-41.6	19.49	-8.83
	Sargan test	77000	18000	4500	2800
System GMM MA	m2	8.2	-40.03	19.03	-11.69
	Sargan test	28000	17000	3100	2500

Note: The Sargan test (Blundell et al., 2001; Arellano et al., 1991) is used to assess the validity of the over-identifying restrictions. The z-score for the $m2$ tests for zero autocorrelation in the first-differenced errors of order two (Arellano et al., 1991). See text for additional information.

fication tests such as the $m2$ test for autocorrelation and the Sargan test paint a slightly less consistent picture. Table 2 shows the two tests for each of the models estimated by country, synthetic status, and model. The Sargan test rejects the null in both countries and for all models, consistently for confidential and synthetic data. But the $m2$ test is of opposite signs for half of the comparisons.

4.6. pMSE

To compute the $pMSE$, we estimate Equation (1) using logit models. The estimated $pMSE$ is 0.0121 for the Canadian data (0.0041 for the manufacturing sector) and 0.0013 for the German data (see Table 3). While these numbers may seem small, the $pMSE$ ratio and the standardized $pMSE$ are large, indicating that the null hypothesis that the synthetic data and the original data stem from the same data generating process should be rejected. The expected $pMSE$ is quite sensitive to sample size N. Even small differences between the original and synthetic data will lead to large values for this test statistic. In both countries, the confidential data files are quite large (about 2 million cases for Germany and the manufacturing sector in Canada and about 34.5 million cases for the full Canadian data sets). In practice, therefore, it is quite likely to reject the null of equivalence given this test's very high power.

Table 3: pMSE by sector and country

Country	Sector	pMSE	pMSE ratio	standardized pMSE
Canada	Manufacturing	0.0041	656.88	4908.17
Canada	Private	0.0121	10957.61	135525.77
Germany	Universe	0.0013	725.21	2896.85

5. Confidentiality protection

To assess the risk of disclosure, we use a measure proposed by Kinney et al. (2011b): For each industry, we estimate the fraction of entities for which the synthetic birth year equals the true birth year, conditional on the synthetic birth year, and interpret it as a probability. Tables 14 and 15 in the Online Appendix show the minimum, maximum, and mean of these probabilities, by year. Figure 7 shows the maximum and average values across time, for each country.[14] The figure shows that these probabilities are quite low except for the first year. Entry rates in the first year are much larger than in any other year due to censoring. It is therefore quite likely that the (left-censored) entry year of the synthetic record matches that of the (left-censored) original record if the synthetic entry year is the first year observed in the data. A somewhat more muted version of this effect can be seen for Germany in the years 1991 and 1992, when the lower panel of Figure 7 shows another spike. These are the years in which data from Eastern Germany were added to the database successively, leading to new sets of (left-censored) entities.

With the exception of the first year in the data, the average rate of concordance between synthetic and observed birth year of an establishment in the Canadian data is below 5%, and the maximum is never above 50%. The German data reflect results from a smaller set of industries, and while the average concordance is higher (never above 10%), the maximum is never above 6% other than during the noted entry spikes. This suggests that the synthetic lifespan of any given entity is highly unlikely to be matched to its confidential real lifespan. This is generally considered to be a high degree of confidentiality.

[14] The Canadian manufacturing sector is not shown. In the German case, we only use two industries, but we show the average of the two, rather than the values for both industries, to maintain comparability with the Canadian plot.

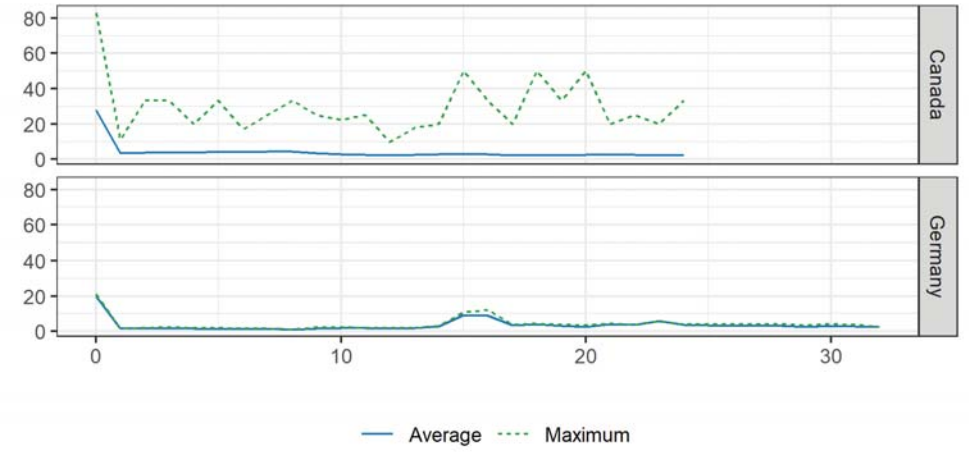


Figure 7: Average and maximum likelihood that synthetic birthyear matches actual birthyear

Note: Plot shows fraction of entities by industry for which the synthetic birth year equals the true birth year, conditional on the synthetic birth year. Plot has been rescaled to be relative to the first year observed in the data.

6. Conclusion

In this paper, we presented results from two projects that evaluated whether the code developed to synthesize the U.S. LBD can easily be adapted to create synthetic versions of similar data from Canada and Germany. We considered both univariate time-series comparisons as well as model-based comparisons of coefficients and model fit. In general, utility evaluations show significant differences between each country's synthetic and confidential data. Frequently-used measures such as confidence interval overlap and $pMSE$ suggest that the synthetic data are an unreliable image of the confidential data. Less formal comparisons of specification test scores suggest that the synthetic data do not reliably lead to the same modeling decisions.

Interestingly, the utility of the German synthetic data was higher than the utility of the Canadian data in almost all dimensions evaluated. At this point we can only speculate about potential reasons. The most important difference between the two data sources is that the German data comprises only a handful of industries while almost all industries have been included in the Canadian evaluation. Given that the industries included in the German data were rather large, and synthesis models are run independently for each industry, it might have been easier to preserve the industry level statistics for the German data. We cannot exclude the possibility that the structure of the German data aligns more closely with the LBD and thus the synthesis models tuned on the LBD data provide better results on the (adjusted) BHP than on the LEAP. We note that both the LBD

and the BHP are establishment-level data sets, whereas the LEAP is an employer-level data set.

We emphasize that adjustments to the original synthesis code were explicitly limited to ensuring that the code runs on the new input data. The validity of the synthetic data could possibly be improved by tuning the synthesis models to the particularities of the data at hand, such as the non-standard dynamics introduced into the German data by reunification. However, the aim of this project was to illustrate that the high investments necessary for developing the synthesis code for the LBD offered additional payoffs as the re-use of the code substantially reduced the amount of work required to generate decent synthetic data products for other business data. One of the major criticisms of the synthetic data approach has been that investments necessary to develop useful synthesizers are substantial. This project illustrated that substantial gains can be achieved when exploiting knowledge from previous projects. With the advent of tailor-made software such as the *synthpop* package in R (Nowok et al., 2016), the investments for generating useful synthetic data might be further reduced in the future.

However, even without fine-tuning or customization of models, the current synthetic data have, in fact, proven useful. De facto, many deployments of synthetic data, including the Synthetic LBD in the US, have been used for model preparation by researchers in a public or lower-security environment, with subsequent remote submission of prepared code for validation against the confidential data. When viewed through the lens of such a validation system, the synthetic data prepared here would seem to have reasonable utility. While time series dynamics are not the same, they are broadly similar. Models converged in similar fashions, and while coefficients were strictly different, they were broadly similar and plausible. Specification tests did not lead to the same conclusions, but they also did not collapse or yield meaningless conclusions. Thus, we believe that the synthetic data, despite being different, have the potential to be a useful tool for analysts to prepare models without direct access to the confidential data. Vilhuber et al. (2016a) and Vilhuber (2019) come to a similar conclusion when evaluating usage of the synthetic data sets available through the Synthetic Data Server (Abowd et al., 2010), including the Synthetic LBD. A more thorough evaluation would need to explicitly measure the investment in synthetic data generation, the cost of setting up a validation structure, and the number of studies enabled through such a setup. We note that such an evaluation is non-trivial: the counter-factual in many circumstances is that no access is allowed to sensitive business microdata, or that access occurs through a secure research data system that is also costly to maintain. This study has contributed to such a future evaluation by showing that plausible results can be achieved with relatively low up-front investments.

The use of synthetic data sets to broaden access to confidential microdata is likely to increase in the near future, with increasing concerns by statistical agencies regarding the disclosure risks of releasing microdata. The resulting reduction in access to scientific microdata is overwhelmingly seen as problematic. Broadly "plausible" if not analytically valid synthetic data sets such as those described in this paper, combined with scalable remote submission systems that integrate modern disclosure avoidance mechanisms, may be a feasible mitigation strategy.

Acknowledgements

The opinions expressed here are those of the authors, and do not reflect the opinions of any of the statistical agencies involved. All results were reviewed for disclosure risks by their respective custodians, and released to the authors. Alam thanks Claudiu Motoc and Danny Leung for help with the Canadian data. Vilhuber acknowledges funding through NSF Grants SES-1131848 and SES-1042181, and a grant from Alfred P. Sloan Grant (G-2015-13903). Alam and Dostie acknowledge funding through SSHRC Partnership Grant "Productivity, Firms and Incomes". The creation of the Synthetic LBD was funded by NSF Grant SES-0427889.

References

ABOWD, J. M. and J. I. LANE (2004). "New Approaches to Confidentiality Protection Synthetic Data, Remote Access and Research Data Centers". In: *Privacy in Statistical Databases*. Ed. by J. DOMINGO-FERRER and V. TORRA. Vol. 3050. Lecture Notes in Computer Science. Springer, pp. 282–289. DOI: 10.1007/978-3-540-22118-0. URL: http://www.springer.com/la/book/9783540221180.

ABOWD, J. M. and I. SCHMUTTE (2015). "Economic analysis and statistical disclosure limitation". In: *Brookings Papers on Economic Activity* Fall 2015. URL: http://www.brookings.edu/about/projects/bpea/papers/2015/economic-analysis-statistical-disclosure-limitation.

ABOWD, J. M., B. E. STEPHENS, L. VILHUBER, F. ANDERSSON, K. L. MCKINNEY, M. ROEMER, and S. D. WOODCOCK (2009). "The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators". In: *Producer Dynamics: New Evidence from Micro Data*. Ed. by T. DUNNE, J. B. JENSEN, and M. J. ROBERTS. University of Chicago Press. URL: http://www.nber.org/chapters/c0485.

ABOWD, J. M. and L. VILHUBER (2010). *VirtualRDC - Synthetic Data Server*. Cornell University, Labor Dynamics Institute. URL: http://www.vrdc.cornell.edu/sds/.

ALAM, M. J., B. DOSTIE, J. DRECHSLER, and L. VILHUBER (2020). *Replication archive for: Applying Data Synthesis for Longitudinal Business Data across Three Countries*. Code and data. Zenodo. DOI: 10.5281/zenodo.3785744.

ARELLANO, M. and S. BOND (1991). "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". In: *Review of Economic Studies* 58.2, pp. 277–297. URL: https://EconPapers.repec.org/RePEc:oup:restud:v:58:y:1991:i:2:p:277-297..

ARELLANO, M. and O. BOVER (1995). "Another look at the instrumental variable estimation of error-components models". In: *Journal of Econometrics* 68.1, pp. 29–51. URL: https://EconPapers.repec.org/RePEc:eee:econom:v:68:y:1995:i:1:p:29-51.

BARTELSMAN, E., J. HALTIWANGER, and S. SCARPETTA (2009). "Measuring and Analyzing Cross-country Differences in Firm Dynamics". In: DUNNE, T., J. B. JENSEN, and M. J. ROBERTS. *Producer Dynamics: New Evidence from Micro*

Data. University of Chicago Press, pp. 15–76. URL: http://www.nber.org/chapters/c0480.

BENDER, S. (2009). "The RDC of the Federal Employment Agency as a part of the German RDC Movement". In: *Comparative Analysis of Enterprise Data, 2009 Conference*. Comparative Analysis of Enterprise Data, 2009 Conference (Tokyo). URL: http://gcoe.ier.hit-u.ac.jp/CAED/index.html (visited on 05/05/2014).

BENEDETTO, G., J. HALTIWANGER, J. LANE, and K. MCKINNEY (2007). "Using Worker Flows in the Analysis of the Firm". In: *Journal of Business and Economic Statistics* 25.3, pp. 299–313.

BLUNDELL, R. and S. BOND (1998). "Initial conditions and moment restrictions in dynamic panel data models". In: *Journal of Econometrics* 87.1, pp. 115–143. URL: https://ideas.repec.org/a/eee/econom/v87y1998i1p115-143.html.

BLUNDELL, R., S. BOND, and F. WINDMEIJER (2001). "Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator". In: *Nonstationary Panels, Panel Cointegration, and Dynamic Panels*. Ed. by B. H. BALTAGI, T. B. FOMBY, and R. CARTER HILL. Vol. 15. Advances in Econometrics. Emerald Group Publishing Limited, pp. 53–91. DOI: 10.1016/S0731-9053(00)15003-0. URL: https://doi.org/10.1016/S0731-9053(00)15003-0 (visited on 04/30/2020).

BUNDESAGENTUR FÜR ARBEIT (2013). *Establishment History Panel (BHP)*. [Computer file]. Nürnberg, Germany: Research Data Centre (FDZ) of the German Federal Employment Agency (BA) at the Institute for Employment Research (IAB) [distributor].

DAVIS, S. J., J. C. HALTIWANGER, and S. SCHUH (1996). *Job creation and destruction*. Cambridge, MA: MIT Press.

DRECHSLER, J. (2011a). *Synthetic Datasets for Statistical Disclosure Control–Theory and Implementation*. New York: Springer. DOI: 10.1007/978-1-4614-0326-5.

DRECHSLER, J. (2011b). *Synthetische Scientific-Use-Files der Welle 2007 des IAB-Betriebspanels*. FDZ Methodenreport 201101_de. Institute for Employment Research, Nuremberg, Germany. URL: http://ideas.repec.org/p/iab/iabfme/201101_de.html.

— (2012). "New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey". In: *Journal of Applied Statistics* 39.2, pp. 243–265. URL: http://ideas.repec.org/a/taf/japsta/v39y2012i2p243-265.html.

DRECHSLER, J. and L. VILHUBER (2014a). *A First Step Towards A German Synlbd: Constructing A German Longitudinal Business Database*. Working Papers 14-13. Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org/p/cen/wpaper/14-13.html.

DRECHSLER, J., A. DUNDLER, S. BENDER, S. RÄSSLER, and T. ZWICK (2008). "A new approach for disclosure control in the IAB establishment panel—multiple imputation for a better data access". In: *AStA Advances in Statistical Analysis* 92.4, pp. 439–458.

DRECHSLER, J. and L. VILHUBER (2014b). "A First Step Towards A German SynLBD: Constructing A German Longitudinal Business Database". In: *Statistical Journal of*

the IAOS: Journal of the International Association for Official Statistics 30.2. DOI: 10.3233/SJI-140812. URL: http://iospress.metapress.com/content/X415V18331Q33150.

GUZMAN, J. and S. STERN (2016). *The State of American Entrepreneurship: New Estimates of the Quality and Quantity of Entrepreneurship for 32 US States, 1988-2014*. Working Paper 22095. National Bureau of Economic Research. DOI: 10.3386/w22095. URL: http://www.nber.org/papers/w22095.

— (2020). *Startup Cartography*. URL: https://www.startupcartography.com/ (visited on 01/26/2020).

HANSEN, L. P. (1982). "Large Sample Properties of Generalized Method of Moments Estimators". In: *Econometrica* 50.4, p. 1029. DOI: 10.2307/1912775. URL: https://www.jstor.org/stable/1912775?origin=crossref (visited on 04/30/2020).

HETHEY, T. and J. F. SCHMIEDER (2010). *Using worker flows in the analysis of establishment turnover: Evidence from German administrative data*. FDZ Methodenreport 201006_en. Institute for Employment Research, Nuremberg, Germany. URL: http://ideas.repec.org/p/iab/iabfme/201006_en.html.

JARMIN, R. S. and J. MIRANDA (2002). *The Longitudinal Business Database*. Working Papers 02-17. Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org/p/cen/wpaper/02-17.html.

JARMIN, R. S., T. A. LOUIS, and J. MIRANDA (2014). "Expanding The Role Of Synthetic Data At The U.S. Census Bureau". In: *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 30.2. DOI: 10.3233/SJI-140813. URL: http://iospress.metapress.com/content/fl8434n4v38m4347/?p=00c99b98bf2f4701ae806ee638594915&pi=0.

KARR, A. F., C. N. KOHNEN, A. OGANIAN, J. P. REITER, and A. P. SANIL (2006). "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality". In: *The American Statistician* 60.3, pp. 1–9. DOI: 10.1198/000313006X124640.

KINNEY, S. K., J. P. REITER, and J. MIRANDA (2014a). *Improving The Synthetic Longitudinal Business Database*. Working Papers 14-12. Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org/p/cen/wpaper/14-12.html.

— (2014b). "Improving The Synthetic Longitudinal Business Database". In: *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics* 30.2. DOI: 10.3233/SJI-140808.

KINNEY, S. K., J. P. REITER, A. P. REZNEK, J. MIRANDA, R. S. JARMIN, and J. M. ABOWD (2011a). *LBD Synthesis Procedures*. CES Technical Notes Series 11-01. Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org/p/cen/tnotes/11-01.html.

— (2011b). "Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database". In: *International Statistical Review* 79.3, pp. 362–384. DOI: j.1751-5823.2011.00152.x. URL: https://ideas.repec.org/a/bla/istatr/v79y2011i3p362-384.html.

LITTLE, R. J. (1993). "Statistical Analysis of Masked Data". In: *Journal of Official Statistics* 9.2, pp. 407–426.

NATIONAL RESEARCH COUNCIL (2007). *Understanding Business Dynamics: An Integrated Data System for America's Future*. Ed. by J. HALTIWANGER, L. M. LYNCH, and C. MACKIE. Washington, DC: The National Academies Press. DOI: 10.17226/11844. URL: https://www.nap.edu/catalog/11844/understanding-business-dynamics-an-integrated-data-system-for-americas-future.

NOWOK, B., G. RAAB, and C. DIBBEN (2016). "synthpop: Bespoke Creation of Synthetic Data in R". In: *Journal of Statistical Software, Articles* 74.11, pp. 1–26. DOI: 10.18637/jss.v074.i11. URL: https://www.jstatsoft.org/v074/i11.

RAAB, G. M., B. NOWOK, and C. DIBBEN (2018). "Practical Data Synthesis for Large Samples". In: *Journal of Privacy and Confidentiality* 7.3, pp. 67–97. DOI: 10.29012/jpc.v7i3.407. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/407.

RUBIN, D. B. (1993). "Discussion of Statistical Disclosure Limitation". In: *Journal of Official Statistics* 9.2, pp. 461–468.

SEDLÁČEK, P. and V. STERK (2017). "The Growth Potential of Startups over the Business Cycle". In: *American Economic Review* 107.10, pp. 3182–3210. DOI: 10.1257/aer.20141280. URL: http://www.aeaweb.org/articles?id=10.1257/aer.20141280.

SNOKE, J., G. M. RAAB, B. NOWOK, C. DIBBEN, and A. SLAVKOVIC (2018a). "General and specific utility measures for synthetic data". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.3, pp. 663–688. DOI: 10.1111/rssa.12358. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12358. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12358.

SNOKE, J. and A. SLAVKOVIC (2018b). "pMSE Mechanism: Differentially Private Synthetic Data with Maximal Distributional Similarity: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26-28, 2018, Proceedings". In: pp. 138–159. DOI: 10.1007/978-3-319-99771-1_10.

STATISTICS CANADA (2019a). *Business Register (BR)*. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=1105 (visited on 01/30/2020).

— (2019b). *Longitudinal Employment Analysis Program (LEAP)*. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=8013 (visited on 01/30/2020).

— (2019c). *Survey of Employment, Payrolls and Hours (SEPH)*. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey%5C&SDDS=2612 (visited on 01/30/2020).

STATISTICS CANADA and BUREAU OF THE CENSUS (1991). *Concordance between the Standard Industrial Classifications of Canada and the United States, 1980 Canadian SIC - 1987 United States SIC*. Catalogue No. 12-574E. Statistics Canada. URL: http://publications.gc.ca/site/eng/9.847987/publication.html (visited on 01/30/2020).

STATISTISCHES BUNDESAMT (2003). *Classification of Economic Activities, issue 2003 (WZ 2003)*. Statistisches Bundesamt (Federal Statistical Office) of Germany.

URL: https://www.klassifikationsserver.de/klassService/index.jsp?variant=wz2003 (visited on 02/02/2020).

U.S. CENSUS BUREAU (2015). *Longitudinal Business Database 1975-2015 [Data file]*. Tech. rep. URL: https://www.census.gov/programs-surveys/ces/data/restricted-use-data/longitudinal-business-database.html (visited on 01/26/2020).

— (2016a). *County Business Patterns (CBP)*. U.S. Census Bureau. URL: https://www.census.gov/programs-surveys/cbp.html (visited on 01/26/2020).

— (2016b). *Statistics of U.S. Businesses (SUSB)*. U.S. Census Bureau. URL: https://www.census.gov/programs-surveys/susb.html (visited on 01/26/2020).

— (2017). *Business Dynamics Statistics (BDS)*. U.S. Census Bureau. URL: https://www.census.gov/programs-surveys/bds.html (visited on 01/26/2020).

VILHUBER, L. (2013). *Methods for Protecting the Confidentiality of Firm-Level Data: Issues and Solutions*. Document 19. Labor Dynamics Institute. URL: http://digitalcommons.ilr.cornell.edu/ldi/19/.

— (2018). *LEHD Infrastructure S2014 files in the FSRDC*. Working Papers 18-27. Center for Economic Studies, U.S. Census Bureau. URL: https://ideas.repec.org/p/cen/wpaper/18-27.html.

— (2019). *Utility of two synthetic data sets mediated through a validation server: Experience with the Cornell Synthetic Data Server*. Presentation. Conference on Current Trends in Survey Statistics. URL: https://hdl.handle.net/1813/43883.

VILHUBER, L. and J. M. ABOWD (2016a). *Usage and outcomes of the Synthetic Data Server*. Presentation. Meetings of the Society of Labor Economists. URL: https://hdl.handle.net/.

VILHUBER, L., J. M. ABOWD, and J. P. REITER (2016b). "Synthetic establishment microdata around the world". In: *Statistical Journal of the International Association for Official Statistics* 32.1, pp. 65–68. DOI: 10.3233/SJI-160964.

WOO, M.-J., J. P. REITER, A. OGANIAN, and A. F. KARR (2009). "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation". In: *Journal of Privacy and Confidentiality* 1.1. DOI: 10.29012/jpc.v1i1.568. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/568.

WOODCOCK, S. D. and G. BENEDETTO (2009). "Distribution-preserving statistical disclosure limitation". In: *Computational Statistics & Data Analysis* 53.12, pp. 4228–4242. DOI: https://doi.org/10.1016/j.csda.2009.05.020. URL: http://www.sciencedirect.com/science/article/pii/S0167947309002011.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 237–253, DOI 10.21307/stattrans-2020-040
Received – 31.01.2020; accepted – 30.06.2020

A general Bayesian approach to meet different inferential goals in poverty research for small areas

Partha Lahiri[1], **Jiraphan Suntornchost**[2]

ABSTRACT

Poverty mapping that displays spatial distribution of various poverty indices is most useful to policymakers and researchers when they are disaggregated into small geographic units, such as cities, municipalities or other administrative partitions of a country. Typically, national household surveys that contain welfare variables such as income and expenditures provide limited or no data for small areas. It is well-known that while direct survey-weighted estimates are quite reliable for national or large geographical areas they are unreliable for small geographic areas. If the objective is to find areas with extreme poverty, these direct estimates will often select small areas due to the high variability in the estimates. Empirical best prediction and Bayesian methods have been proposed to improve on the direct point estimates. These estimates are, however, not appropriate for different inferential purposes. For example, for identifying areas with extreme poverty, these estimates would often select areas with large sample sizes. In this paper, using real life data, we illustrate how appropriate Bayesian methodology can be developed to address different inferential problems.

Key words: Bayesian model, cross-validation, hierarchical models, Monte Carlo simulations

1. Introduction

Eradication of poverty, one of the greatest challenges facing humanity, has been the central tool to guide various public policy efforts in many countries. According to the United Nations [3]: "Extreme poverty rates have fallen by more than half since 1990. While this is a remarkable achievement, one-in-five people in developing regions still live on less than $1.90 a day. Millions more make little more than this daily amount and are at risk of slipping back into extreme". On September 25, 2015, the United Nations adopted the 2030 Agenda for Sustainable Development with 17 new Sustainable Development Goals (SDGs), beginning with a historical pledge to end poverty in all forms and dimensions by 2030 everywhere permanently.[4] In order to achieve these goals, basic resources and services need to be more accessible to people living in vulnerable situations. Moreover, support for communities affected by conflict and climate related disasters needs to be raised.

[1]Department of Mathematics & Joint Program in Survey Methodology, University of Maryland, College Park, USA. E-mail: plahiri@umd.edu. ORCID: https://orcid.org/0000-0002-7103-545X.
[2]Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Thailand. E-mail: jiraphan.s@chula.ac.th. ORCID: https://orcid.org/0000-0001-5410-9659.
[3]see https://sdg-tracker.org/no-poverty
[4]see https://www.un.org/sustainabledevelopment/

National estimate of an indicator usually hides important differences among different regions or areas with respect to that indicator. In almost all countries, these differences exist and can often be substantial. The smaller the geographic regions for which indicators are available, the greater the effectiveness of interventions. Indeed this allows to reduce transfers to the non-poor and minimizes the risk that a poor person will be missed by the program. Ravallion (1994) found that Indian and Indonesian states or provinces are too heterogeneous for targeting to be effective. This underlines the need for production of estimates of indicators for small areas that are relatively homogenous.

It is now widely accepted that direct estimates of poverty based on household survey data are unreliable. There are now several papers available in the literature that attempt to improve on the direct estimates by borrowing strength from multiple relevant databases. Hierarchical models that combine information from different databases are commonly used to achieve the goal because such models not only provide improved point estimates but also incorporate different sources of variabilities. These models can be implemented using a synthetic approach (e.g., Elbers et al. 2003), empirical best prediction approach (Fay and Herriot 1979; Franco and Bell 2015; Bell et al. 2016; Molina and Rao 2010; Casas-Cordero et al. 2016), and Bayesian approach (Molina et al. 2014). See Jiang and Lahiri (2006), Pfeffermann (2013) and Rao and Molina (2015) for a review of different small area estimation techniques.

Empirical best prediction and hierarchical Bayesian methods (also shrinkage methods) have been employed in numerous settings, including studies of cancer incidence in Scotland (Clayton and Kaldor 1987), cancer mortality in France (Mollie and Richardson 1991), stomach and bladder cancer mortality in Missouri cities (Tsutukawa et al. 1985), toxoplasmosis incidence in El Salvador (Efron and Morris 1975), infant mortality in New Zealand (Marshall 1991), mortality rates for chronic obstructive pulmonary disease (Nandram et al. 2000), poverty research (Molina and Rao 2010; Molina et al. 2014; Bell et al. 2016; Casas-Cordero et al. 2016). The basic approach in all these applications is the same: a prior distribution of rates is posited and is combined with the observed rates to calculate the posterior, or stabilized, rates.

All the papers cited in the previous paragraph deal with the point estimation and the associated measure of uncertainty. However, in many cases, there could be different inferential goals where point estimates, whether empirical best prediction estimates or posterior means, though they can provide a solution, are not efficient. For example, one inferential goal could be to flag a geographical area (e.g., municipality) for which the true poverty measure of interest exceeds a pre-specified standard. The point estimates can certainly flag such areas but do not provide any reasonable uncertainty measure to assess the quality of such action. It is not clear how to propose such a measure for a method based on direct estimates. For the method based on posterior mean, one can perhaps propose a normal approximation using the posterior mean and posterior standard deviation to approximate the posterior probability of the true poverty measure exceeding the pre-specified standard. In some cases, quality of such approximation could be questionable. One can have a more complex inferential goal. For example, we may be interested in identifying the worst geographical area with respect to the poverty measure. In this case, the use of direct estimates for identification can be misleading when the

sample sizes vary across the geographic units. The regions with small sample sizes will tend to have both high and low poverty indices merely because they have the largest variability. The method based on posterior means is not good either since it tends to identify areas with more samples; see Gelman and Price (1999). Moreover, in either case there does not seem to be a clear way to produce a reasonable quality measure.

Gelman and Price (1999), Morris and Christiansen (1996), Langford and Lewis (1998), Jones and Spigelhalter (2011) discussed various inferential problems other than the point estimation. For example, Morris and Christiansen (1996) outlined inferential procedures for identifying areas with extreme indicator. Our approach is similar to theirs but applied for different complex parameters and used much more complex hierarchical model that is appropriate for survey data.

We would like to stress that our proposed approach for solving different inferential problems is fundamentally different from the constrained empirical hierarchical Bayesian (Louis 1984; Ghosh 1992; Lahiri 1990) and the triple-goal (Shen and Louis, 1998) approaches where the goal is to produce one set of estimates for different purposes. In contrast, we propose to use the same synthetic data matrix generated from the posterior predictive distribution for different inferential purposes. Other than this fundamental difference, our approach has a straightforward natural way to produce appropriate quality measures. In poverty research, Molina et al. (2014) suggested an interesting approach for estimating different poverty indices by generating a synthetic population from a posterior predictive density. However, they restricted themselves to point estimates and their associated uncertainty measures and did not discuss how one would solve a variety of statistical inferences.

We outline a general approach to deal with different inferential goals and illustrate our methodology using the data used by the Chilean government for their small area poverty mapping system. Section 2 describes the Chilean data, the hierarchical model, a general poverty index, inferential approach for achieving various goals and data analysis. In Section 3, we provide some concluding remarks and a direction for future research.

2. Illustration of the proposed methodology using Chilean poverty data

There has been a consistent downward trend in the official poverty rate estimates, which are the usual national survey-weighted direct estimates, in Chile since the early 90's. While this national trend is encouraging, there is an erratic time series trend in the direct estimates for small comunas (municipalities) - the smallest territorial entity in Chile. Moreover, for a handful of extremely small comunas, survey estimates of poverty rates are unavailable for some or all time points simply because the survey design, which traditionally focuses on obtaining precise estimates for the nation and large geographical areas, excludes these comunas for some or all of the time points. In any case, direct survey estimates of poverty rates typically do not meet the desired precision for small comunas and thus the assessment of implemented policies is not straightforward at the comuna level. In order to successfully monitor trends, identify influential factors,

develop effective public policies and eradicate poverty at the comuna level, there is a growing need to improve on the methodology for estimating poverty rates at this level of geography. In this section, we use the Chilean case to illustrate our Bayesian approach to answer a variety of research questions.

2.1. Data used in the Analysis

To illustrate our Bayesian approach, we use a household survey data as the primary source of information and comuna level summary statistics obtained from different administrative data sources as supplementary sources of information. We now provide a brief description of the primary and supplementary databases. Further details can be obtained from Casas-Cordero et al. (2016).

2.1.1 The Primary Data Source: The CASEN 2009 Data

The Ministry of Social Development estimates the official poverty rates using the National Socioeconomic Characterization Survey, commonly referred to as the CASEN. The Ministry has been conducting CASEN since 1987 every two or three years. The CASEN is a household survey collecting a variety of information of Chilean households and persons, including information about income, work, health, subsidies, housing and others. The Ministry calculates poverty rate estimates at national, regional and municipality (comuna) levels. The Ministry is the authority specified by the Chilean law to deliver poverty estimates for all the 345 comunas in Chile. These estimates are used, along with other variables, to allocate public funding to municipalities.

In a joint effort by the Ministry and United Nations Development Programme (UNDP), a Small Area Estimation (SAE) official system was developed for estimating poverty rates at comuna level using the CASEN 2009 survey. The Chilean method is based on an empirical Bayesian method using an area level Fay-Herriot Model (Fay and Herriot 1979) to combine the CASEN survey data with a number of administrative databases. The SAE system provides point estimates and parametric bootstrap confidence intervals (see, e.g., Chatterjee et al. 2008; Li and Lahiri 2010) for the Chilean comunas.

The CASEN 2009 used a stratified multistage complex sample of approximately $75,000$ housing units from $4,156$ sample areas. The entire Chile was divided into a large number of sections (Primary Stage Units or PSUs). The PSUs were then grouped into strata on the basis of two geographic characteristics: comuna and urban/rural classification. Overall, there were 602 strata in the CASEN 2009 survey and multiple PSUs were sampled per stratum. The probability of selection for each PSU in a stratum was proportional to the number of housing units in the (most recently updated) 2002 Census file.

Prior to the second stage of sampling, listers were sent to the sampled PSUs to update the count of housing units. This procedure was implemented in both urban and rural areas. In the second stage of sampling, a sample of housing units was selected within the sampled PSUs. The probability of selection for each housing unit (Secondary Stage Unit or SSU) is the same within each PSU. On the average, 16-22 housing units

were selected within each PSU by implementing a procedure that used a random start and a systematic interval to select the units to be included in the sample.

2.1.2 Administrative Data at the Comuna Level

Casas-Cordero et al. (2016) carried out an extensive task to identify a set of auxiliary variables derived from different administrative records of different agencies. In this paper, we use the same set of comuna level auxiliary variables for illustrating our approach. For completeness, we list them below:

(1) Average wage of workers who are not self-employed,

(2) Average of the poverty rates from CASEN 2000, 2003, and 2006,

(3) Percentage of population in rural areas,

(4) Percentage of illiterate population,

(5) Percentage of population attending school.

Like in Casas-Cordero et al. (2016), we also use arcsine square-root transformation for all the auxiliary variables except the first one, for which we use logarithmic transformation. We note that our approach is general and can use a different set of auxiliary variables that may be deemed appropriate in the future.

2.2. The Foster-Greer-Thorbecke (FGT) poverty indices

Currently, the Chilean government publishes headcount ratios or poverty rates for the nation and its comunas. To present our approach in a general setting, we consider a general class of poverty indices commonly referred to as the FGT indices, after the names of the three authors of the paper by Foster et al. (1984). To describe the FGT index, we first introduce the following notations:

N_c: total number of households in comuna c,

U_c: number of urbanicity statuses for comuna c; since for urbanicity status, we use urban and rural statuses only, U_c is either 1 or 2 for a given comuna,

k_u: fixed poverty line for urban-rural classification u ($u = 1$ and $u = 2$ for urban and rural, respectively),

M_{cu}: total number of PSUs in the universe for urban-rural classification u of comuna c,

N_{cup}: total number of households in the universe of the PSU p belonging to the urban-rural classification u of comuna c,

y_{cuph}: per-capita income of household h (that is, total income of the household divided by the number of household members) in PSU p, urban-rural classification u, and comuna c.

In our context, the class of FGT indices for comuna c is given by

$$Q_{c;\alpha} = \frac{1}{N_c} \sum_{u=1}^{U_c} \sum_{p=1}^{M_{cu}} \sum_{h=1}^{N_{cup}} g_\alpha(y_{cuph}),$$

where

$g_\alpha(y_{cuph}) = \left(\frac{k_u - y_{cuph}}{k_u}\right)^\alpha I(y_{cuph} < k_u)$, α is a "sensitivity" parameter ($\alpha = 0, 1, 2$ corresponding to poverty ratio, poverty gap, and poverty severity, respectively).

2.3. Hierarchical Model

A hierarchical model could be effective in capturing different salient features of the CASEN survey data and in linking comuna level auxiliary variables derived from different administrative records. We consider the following working hierarchical model to illustrate our general approach for inference. We call the model a working model because we recognize that it is possible to improve on it in the future. But this model will suffice to illustrate the central theme of the paper, i.e., how to carry out a particular inferential procedure given a hierarchical model.

Let $T_{cuph} = T(y_{cuph})$ be a given transformation on the study variable y_{cuph}. For the application of this paper, we consider $T(y_{cuph}) = \ln(y_{cuph} + 1)$. We consider the following hierarchical model for the sampled units:

$$\begin{aligned}
\text{Level 1:} \quad & T_{cuph}|\theta_{cup}, \sigma_T \overset{ind}{\sim} N\left(\theta_{cup}, \sigma_T^2\right), \\
\text{Level 2:} \quad & \theta_{cup}|\mu_{cu}, \sigma_\theta \overset{ind}{\sim} N\left(\mu_{cu}, \sigma_\theta^2\right), \\
\text{Level 3:} \quad & \mu_{cu}|\beta_u, \sigma_\mu \overset{ind}{\sim} N\left(\mathbf{x}_c^T \beta_u, \sigma_\mu^2\right),
\end{aligned}$$

where $\mathbf{x}_c$ is a vector of comuna level known fixed auxiliary variables; θ_{cup} and μ_{cu} are random effects; $\beta_u, \sigma_T^2, \sigma_\theta^2$ and σ_μ^2 are unknown hyperparameters.

We follow the recommendation of Gelman (2015) in assuming weakly informative priors for the hyperparameters. For example, we assume independent $N(0,1)$ prior for all regression coefficients and independent half normal prior for the standard deviations.

2.4. Inferential Approach

We first note that the inference on $Q_{c,\alpha}$ is equivalent to that of

$$Q_{c;\alpha} = \frac{1}{N_c} \sum_{u=1}^{U_c} \sum_{p=1}^{M_{cu}} \sum_{h=1}^{N_{cup}} g_\alpha\left(T^{-1}(T_{cuph})\right),$$

where T is a monotonic function (e.g., logarithm). Under full specification of the model for the finite population, one can make inferences about $Q_{c;\alpha}$ in a standard way. However, full specification of model for the unobserved units of the finite population seems to be a challenging task. To this end, appealing to the law of large numbers, we first approximate

$Q_{c;\alpha}$ by $\tilde{Q}^P_{c;\alpha}$, where

$$\tilde{Q}^P_{c;\alpha} = \frac{1}{N_c}\sum_{u=1}^{U_c}\sum_{p=1}^{M_{cu}}\sum_{h=1}^{N_{cup}} \mathbb{E}\left\{g_\alpha\left(T^{-1}(T_{cuph})\right)|\theta_{cup},\sigma_T\right\}.$$

This is reasonable under Level 1 of the hierarchical model (even without the normality assumption) since N_c is typically large. We then propose the following approximation to $\tilde{Q}^P_{c;\alpha}$.

$$\tilde{Q}_{c;\alpha} \equiv \tilde{Q}_{c;\alpha}(\theta_c,\sigma_T) = \sum_{u=1}^{U_c}\sum_{p=1}^{m_{cu}}\sum_{h=1}^{n_{cup}} w_{cuph}\mathbb{E}\left\{g_\alpha\left(T^{-1}(T_{cuph})\right)|\theta_{cup},\sigma_T\right\}, \tag{1}$$

where

w_{cuph} is the survey weight for the household h in the PSU p within urbanicity u of comuna c,

$\theta_c = \text{col}_{u,p}\theta_{cup}$; a $\sum_{u=1}^{U_c} m_{cu}\times 1$ column vector (we follow the notation of Prasad and Rao 1990),

$$g_\alpha\left(T^{-1}(T_{cuph})\right) = \left\{\frac{k_u - \left(T^{-1}(T_{cuph})\right)}{k_u}\right\}^\alpha I\left(T_{cuph}\le l_u\right),$$

$l_u = \ln(k_u+1)$, the poverty line of the urbanicity u in the transformed scale.

The weights are scaled within each comuna so that the sum of the weights for all households equals 1. In the last approximation, we assume that the scaled survey weight w_{cuph} represents proportion of units in the finite population (including the unit $cuph$) of comuna c that are similar to the unit $cuph$.

The calculations of (1) under the model described in Section 2.3 can be done through the following formula:

For $\alpha = 0$,

$$\begin{aligned}\mathbb{E}\left\{g_0(\left(T^{-1}(T_{cuph})\right)|\theta_{cup},\sigma_T^2\right\} &= \int_{-\frac{\theta_{cup}}{\sigma_T}}^{\frac{l_u-\theta_{cup}}{\sigma_T}} \phi(z|\theta_{cup},\sigma_T^2)dz \\ &= \Phi\Big(\frac{l_u-\theta_{cup}}{\sigma_T}\Big) - \Phi\Big(-\frac{\theta_{cup}}{\sigma_T}\Big),\end{aligned}$$

where ϕ and Φ are the density function and the distribution function of the standard normal distribution, respectively.

For $\alpha = 1$,

$$\begin{aligned}
&\mathbb{E}\{g_1(T^{-1}(T_{cuph}))|\theta_{cup},\sigma_T^2\} \\
&\qquad = \mathbb{E}\left\{\left(\frac{\exp(l_u)-\exp(T_{cuph})}{k_u}\right) I\big(T_{cuph}\le l_u\big)\right\} \\
&\qquad = \frac{1}{k_u}\int_{-\frac{\theta_{cup}}{\sigma_T}}^{\frac{l_u-\theta_{cup}}{\sigma_T}} (\exp(l_u)-\exp(\sigma_T z+\theta_{cup}))\,\phi(z|\theta_{cup},\sigma_T^2)dz \\
&\qquad = \frac{\exp(l_u)}{k_u}\Big[\Phi\Big(\frac{l_u-\theta_{cup}}{\sigma_T}\Big)-\Phi\Big(-\frac{\theta_{cup}}{\sigma_T}\Big)\Big] \\
&\qquad\quad - \frac{\exp\left(\theta_{cup}+\frac{\sigma_T^2}{2}\right)}{k_u}\left[\Phi\Big(\frac{l_u-\theta_{cup}-\sigma_T^2}{\sigma_T}\Big)-\Phi\Big(-\frac{\theta_{cup}+\sigma_T^2}{\sigma_T}\Big)\right],
\end{aligned}$$

where we use the fact that $\int_a^b \exp(\sigma z)\,\phi(z)dz = \exp\left(\frac{\sigma^2}{2}\right)\Big[\Phi(b-\sigma)-\Phi(a-\sigma)\Big]$ to obtain the last equation.

For $\alpha = 2$,

$$\begin{aligned}
&\mathbb{E}\{g_2(T^{-1}(T_{cuph}))|\theta_{cup},\sigma_T^2\} \\
&\qquad = \mathbb{E}\left\{\left(\frac{\exp(l_u)-\exp(T_{cuph})}{k_u}\right)^2 I\big(T_{cuph}\le l_u\big)\right\} \\
&\qquad = \frac{1}{k_u^2}\int_{-\frac{\theta_{cup}}{\sigma_T}}^{\frac{l_u-\theta_{cup}}{\sigma_T}} (\exp(l_u)-\exp(\sigma_T z+\theta_{cup}))^2\,\phi(z|\theta_{cup},\sigma_T^2)dz \\
&\qquad = \frac{\exp(2l_u)}{k_u^2}\Big[\Phi\Big(\frac{l_u-\theta_{cup}}{\sigma_T}\Big)-\Phi\Big(-\frac{\theta_{cup}}{\sigma_T}\Big)\Big] \\
&\qquad\quad + \frac{\exp(2\theta_{cup}+2\sigma_T^2)}{k_u^2}\left[\Phi\Big(\frac{l_u-\theta_{cup}-2\sigma_T^2}{\sigma_T}\Big)-\Phi\Big(-\frac{\theta_{cup}+2\sigma_T^2}{\sigma_T}\Big)\right] \\
&\qquad\quad - 2\frac{\exp(l_u+\theta+\frac{\sigma_T^2}{2})}{k_u^2}\left[\Phi\Big(\frac{l_u-\theta_{cup}-\sigma_T^2}{\sigma_T}\Big)-\Phi\Big(-\frac{\theta_{cup}+\sigma_T^2}{\sigma_T}\Big)\right].
\end{aligned}$$

In order to carry out a variety of inferential problems about $\tilde{Q}_{c;\alpha}$ for a given α, we use the Monte Carlo Markov Chain (MCMC). The procedures are described below.

Let C be the number of comunas covered by the model and R be the number of MCMC samples after burn-in. Let $\theta_{c;r}$ and $\sigma_{T;r}$ denote the rth MCMC draw of θ_c and σ_T, respectively $(r = 1,\ldots,R)$. We define the $C \times R$, matrix $\tilde{Q}^s_\alpha = (\tilde{Q}^s_{(c,r);\alpha})$, where the (c,r) entry is defined as

$$\tilde{Q}^s_{(c,r);\alpha} \equiv \tilde{Q}^s_{c;\alpha}(\theta_{c;r},\sigma_{T;r}).$$

This matrix $\tilde{\mathcal{Q}}_{\alpha}^{s}$ provides samples generated from the posterior distribution of $\{\tilde{Q}_{c,\alpha},\, c = 1,\ldots,C\}$ and so is adequate for solving a variety of inferential problems in a Bayesian way. We now elaborate on the following three different inferential problems:

(1) Point estimation of an indicator of interest and the associated measure of uncertainty: This is the focus of current poverty mapping research in both classical and Bayesian approaches. Under the squared error loss function, the Bayes estimate of $Q_{c;\alpha}$ for comuna c and the associated measure of uncertainty are the posterior mean and posterior standard deviation of $\tilde{Q}_{c;\alpha} \equiv \tilde{Q}_{c;\alpha}(\theta_c, \sigma_T)$, respectively. These can be approximated by the average and standard deviation over the columns of $\tilde{\mathcal{Q}}_{\alpha}^{s}$, respectively, for the row c, which corresponds to comuna c.

(2) Identification of comunas that are not in conformity with a given standard of a poverty indicator: In this inferential problem, the goal is to flag a comuna for which the true poverty indicator (e.g., poverty rate) exceeds a pre-specified standard, say a. In this case, point estimates, whether direct estimates or posterior means, do not give any idea about the quality of flagging a comuna not meeting the given standard. A reasonable Bayesian solution for this inferential problem is to flag comuna c for not meeting the given standard if the posterior probability $P(\tilde{Q}_{c;\alpha} > a|\text{data})$ is greater than a specified cutoff, say 0.5. This posterior probability for comuna c can be easily approximated by the proportion of columns of $\tilde{\mathcal{Q}}_{c;\alpha}^{s}$ exceeding the threshold for row c. If the posterior distribution of $\tilde{Q}_{c;\alpha}$ is approximately normal, then one can alternatively use the posterior mean and posterior standard deviation to approximate the posterior probability. However, such an approximation may not perform well in many situations.

(3) Identification of the worst (best) comuna, i.e., the comuna with the maximum (minimum) value of the poverty indicator: A common solution is to identify the comuna with the maximum (minimum) point estimate of the indicator. Evidently, the use of direct point estimates would be quite misleading since such a method may identify a small comuna as being the worst (best) in terms of the indicator, even though it is not, simply because of high variability in the direct estimates. The Bayesian point estimates (posterior means) are definitely better than the direct estimates as they have generally less variability. However, the use of posterior means alone does not provide any quality measure associated with the identification of the worst (best) comuna. Even the use of posterior means along with posterior standard deviations does not help either as posterior standard deviations relate to the individual areas. A reasonable Bayesian solution in this case would be to compare the posterior probabilities $P(\tilde{Q}_{c;\alpha} \geq \tilde{Q}_{k;\alpha}\ \forall k|\text{data})$ for different comunas and select the worst (best) comuna for which this posterior probability is the maximum (minimum). Thus, along with the identification of the worst (best) comuna, we also obtain these posterior probabilities suggesting a quality of the identification of worst (best) comuna. We can use $\tilde{\mathcal{Q}}_{\alpha}^{s}$ matrix to approximate these posterior probabilities. For row c and column r of $\tilde{\mathcal{Q}}_{\alpha}^{s}$ corresponding to comuna c and MCMC replicate r, respectively, we can create a binary variable indicating if

the comuna is the worst (best) among all comunas. The posterior probability for this comuna $P(\tilde{Q}_{c;\alpha} \geq \tilde{Q}_{k;\alpha}\ \forall k|\text{data})$ can then be approximated by the average of these binary observations over R columns.

2.5. Numerical Results

As mentioned in the introduction, a number of researchers focused on the problem of estimation and its measure on uncertainty. While our general approach can address this problem, we choose to illustrate the general Bayesian approach for the relatively understudied inferential problems related to the identification of areas with extreme poverty (e.g., the second and the third inferential problems mentioned in Section 2.4). The data analysis presented in this section is based on the hierarchical model stated in Section 2.3 implemented on CASEN 2009 data for a given region containing 54 comunas and comuna level auxiliary variables listed in Section 2.1 We illustrate our methodology for poverty rates ($\alpha = 0$) and poverty gaps ($\alpha = 1$), two important poverty measures in the FGT class of poverty indices. After 10,000 burn-in, we generate 54×10000 matrix $\tilde{Q}^s_{c;\alpha}$ for $\alpha = 0$ (corresponding to poverty rate index) and $\alpha = 1$ (poverty gap index). We checked the convergence of MCMC convergence using the potential scale reduction factor introduced by Gelman and Rubin (1992).

Numerical results are shown in Tables 1-4. We carry out the data analysis using WinBugs-R interface. Table 1 addresses the second inferential goal, i.e., flagging the comunas that do not meet certain pre-specified standard for poverty rate. Table 2 is similar to Table 1 except that this is for the poverty gap measure. We use three different standards based on three different multipliers $(1.10, 1.25$ and $1.50)$ of the regional direct estimate of the respective measure. These standards are for illustration only and our approach can use any other standards that are deemed reasonable. We need a cutoff for these posterior probabilities in order to flag comunas that do not meet the given standard. To illustrate our approach, we use 0.5 as the cutoff. In other words, a comuna is deemed out of the range with respect to the pre-specified standard if the posterior probability is more than 0.5. Comunas 33 and 13 do not meet all three standards for both poverty rate and poverty gap measures. Other comunas meet the more liberal standard (1.5 times the regional poverty measure) with respect to both poverty rate and poverty gap measures. In contrast, when the standard is very conservative (1.1 times the regional poverty measure) all the comunas are not in conformity with the given standard. For a moderate standard (1.25 times the regional poverty measure), comunas 33, 13, 22, 18, 2, 6, 45, 16, 30 do not satisfy the standard in terms of poverty rate measure. The comunas 21, 5, 17 and 15 are added to the list when we consider the poverty gap measure. The standard and the cut-off to be used are subjective, but the Bayesian approach with different standard and cutoff combinations should give policy makers some useful guidance in making certain policy decisions. In order to save space in Table 1 (Table 2), we report results for the 29 out of 54 comunas in the region with highest posterior probabilities of poverty rate (poverty gap) exceeding the most conservative threshold.

Table 1: The posterior probabilities that poverty rate for a comuna exceeds three different thresholds; $Q_{r;0}$ is direct estimate of regional poverty rate. The table presents results for the 29 comunas (out of 54 comunas in the region) with the highest $P(\widetilde{Q}_{c;0} > 1.10Q_{r;0}|\text{data})$

Comuna	$P(\widetilde{Q}_{c;0} > 1.10Q_{r;0}\|\text{data})$	$P(\widetilde{Q}_{c;0} > 1.25Q_{r;0}\|\text{data})$	$P(\widetilde{Q}_{c;0} > 1.50Q_{r;0}\|\text{data})$
33	1.0000	0.9995	0.6172
13	1.0000	0.9988	0.5636
22	0.9952	0.7962	0.0314
18	0.9904	0.6996	0.0100
2	0.9834	0.4939	0.0005
6	0.9809	0.5331	0.0006
45	0.9786	0.5755	0.0032
16	0.9721	0.5157	0.0015
30	0.9662	0.5086	0.0024
21	0.9362	0.3925	0.0013
5	0.9356	0.3878	0.0010
17	0.9258	0.3840	0.0012
15	0.9185	0.3643	0.0015
25	0.8822	0.2524	0.0002
43	0.8755	0.2266	0.0000
38	0.8612	0.2209	0.0003
27	0.8466	0.2139	0.0002
26	0.8425	0.3259	0.0022
51	0.8365	0.2941	0.0009
24	0.7835	0.1216	0.0000
29	0.7111	0.0995	0.0000
28	0.7030	0.1085	0.0000
31	0.6771	0.0700	0.0000
35	0.6694	0.1018	0.0000
36	0.6404	0.0731	0.0000
41	0.6142	0.0591	0.0001
37	0.6041	0.0775	0.0000
7	0.5705	0.0386	0.0000
47	0.5179	0.0417	0.0000

Table 2: Posterior probabilities that poverty gap for a given comuna exceeds three different thresholds; $Q_{r,1}$ is direct estimate of regional poverty gap. The table presents results for the 29 comunas (out of 54 comunas in the region) with the highest $P(\tilde{Q}_{c;1} > 1.10Q_{r,1}|\text{data})$ values.

Comuna	$P(\widetilde{Q}_{c;1} > 1.10Q_{r,1}\|\text{data})$	$P(\widetilde{Q}_{c;1} > 1.25Q_{r,1}\|\text{data})$	$P(\widetilde{Q}_{c;1} > 1.50Q_{r,1}\|\text{data})$
33	1.0000	0.9998	0.9266
13	1.0000	0.9994	0.9060
22	0.9966	0.9143	0.2635
18	0.9918	0.8327	0.1195
2	0.9893	0.7516	0.0395
45	0.9827	0.7577	0.0781
6	0.9824	0.7174	0.0300
16	0.9792	0.7189	0.0490
30	0.9693	0.6764	0.0489
21	0.9467	0.5871	0.0320
5	0.9420	0.5656	0.0240
17	0.9339	0.5592	0.0292
15	0.9333	0.5631	0.0337
38	0.9001	0.4329	0.0081
25	0.8923	0.4203	0.0079
43	0.8802	0.3812	0.0070
26	0.8751	0.5310	0.0657
27	0.8745	0.3970	0.0104
51	0.8540	0.4497	0.0305
24	0.8223	0.2674	0.0018
29	0.7700	0.2401	0.0026
28	0.7441	0.2390	0.0026
31	0.7321	0.1848	0.0005
35	0.6924	0.2091	0.0026
36	0.6671	0.1631	0.0006
37	0.6399	0.1772	0.0021
7	0.6376	0.1243	0.0002
41	0.6355	0.1365	0.0003
47	0.5586	0.1095	0.0003

Table 3 displays approximations (by MCMC) to the posterior probabilities of a comuna being the worst (Prob.Max) in terms of both poverty rate and poverty gap measures. According to the Prob.Max criterion, comuna 33 stands out as the worst comuna in terms of both poverty rate and poverty gap measures. Table 4 displays approximations (by MCMC) to the posterior probabilities of a comuna being the best (Prob.Min) in terms of both poverty rate and poverty gap measures. According to Prob.Min criterion, comuna 8 emerges as the best comuna in terms of both poverty rate and poverty gap measures. These probabilities are also giving us a good sense of confidence we can place on our decision, which is not possible with poverty rate and poverty gap estimates alone. Tables 3 and 4 do not report results for comunas with negligible posterior probabilities.

Table 3: Posterior probability that poverty rate or poverty gap for a given comuna is the maximum (Prob.Max). The table does not include comunas with negligible posterior probabilities.

Comuna	Prob.Max	
	Poverty Rate	Poverty Gap
33	0.5126	0.5246
13	0.4496	0.4301
22	0.0169	0.0215
18	0.0051	0.0044
45	0.0025	0.0031
17	0.0021	0.0021
26	0.0021	0.0042
30	0.0017	0.0017
21	0.0013	0.0016
15	0.0010	0.0011
16	0.0009	0.0012
51	0.0008	0.0009
6	0.0007	0.0006
5	0.0006	0.0006
2	0.0005	0.0009
27	0.0005	0.0004
38	0.0005	0.0006
25	0.0003	0.0001
35	0.0001	0.0002
41	0.0001	0.0000
43	0.0001	0.0000
28	0.0000	0.0001

Table 4: Posterior probability that poverty rate or poverty gap for a given comuna is the minimum (Prob.Min). The table does not include comunas with negligible posterior probabilities.

Comuna	Prob.Min	
	Poverty Rate	Poverty Gap
8	0.5310	0.5161
1	0.3929	0.3945
42	0.0240	0.0268
48	0.0186	0.0237
12	0.0121	0.0139
4	0.0075	0.0089
34	0.0057	0.0079
3	0.0052	0.0047
14	0.0009	0.0011
10	0.0009	0.0005
23	0.0008	0.0012
44	0.0002	0.0004
46	0.0001	0.0002
40	0.0001	0.0001

3. Concluding Remarks

We point out inappropriateness of using point estimates for all inferential purposes and propose a general Bayesian approach to solve different inferential problems in the context of poverty mapping. The proposed approach provides not only an action relevant to the inferential problem but also a way to assess the quality of such action. To make the methodology user-friendly one can store the Q_α^s matrix of size $C \times R$, where C is the number of comunas and R is the number of MCMC replications. This way the users do not need to know how to generate this matrix, which requires knowledge of advanced Bayesian computing. Once the user has access to this generated matrix, he/she can easily carry out a variety of statistical analysis such as the ones presented in the paper with greater ease. While we illustrate the approach for the FGT poverty indices, the approach is general and can deal with other important indices such as the ones given in sustainable development goals. We have taken one working model to illustrate the approach, but the approach is general and can be applied to other models that are deemed appropriate in other projects.

4. Acknowledgments

We thank Editor-in-Chief Professor Wlodzimierz Okrasa and an anonymous referee for reading an earlier version of the article carefully and offering a number of constructive suggestions, which led to a significant improvement of our article. The first author's research was partially supported by the U.S. National Science Foundation grant SES-1758808.

REFERENCES

BELL, W. R., BASEL, W. W., MAPLES, J. J., (2016). An overview of the U.S. Census Bureaus Small Area Income and Poverty Estimates Program. In M. Pratesi (Ed.) *Analysis of poverty data by small area estimation* (pp. 349-377). West Sussex: Wiley & Sons, Inc.

CASAS-CORDERO VALENCIA, C. ENCINA, J., LAHIRI, P., (2016). Poverty mapping in Chilean comunas. In M. Pratesi (Ed.) *Analysis of Poverty Data by Small Area Estimation* (pp. 379–403). West Sussex: Wiley & Sons, Inc.

CHATTERJEE, S., LAHIRI, P., LI, H., (2008). Parametric bootstrap approximation to the distribution of EBLUP and related prediction intervals in linear mixed models. *The Annals of Statistics*, 36, 3, pp. 1221–1245.

CLAYTON, D., KALDOR, J., (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, pp. 671–681.

EFRON, B., MORRIS, C., (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70, pp. 311–319.

ELBERS, C., LANJOUW, J. O., LANJOUW, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, pp. 355–364.

FAY, R. E., HARRIOT, R., (1979). Estimation of income from small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, pp. 269–277.

FOSTER, J., GREER, J., THORBECKE, E., (1984). A class of decomposable poverty measures. *Econometrica*, 52, pp. 761–766.

FRONCO, C., BELL, W. R., (2015). Borrowing information over time in Binomial/Logit normal models for small area estimation. *Joint Special Issue of Statistics in Transition and Survey Methodology,* 16, 4, pp. 563–584.

GELMAN, A., (2015). 3 New priors you can't do without, for coefficients and variance parameters in multilevel regression. *Statistical Modeling, Causal Inference, and Social Science blog, 7 Nov.* http://andrewgelman.com/2015/11/07/priors-for-coefficients-and-variance-parameters-in-multilevel-regression/

GELMAN, A., PRICE, P. N., (1999). All maps of parameter estimates are misleading. *Statistics in Medicine*, 18, pp. 3221–3234.

GELMAN, A., RUBIN, D. B., (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7,4, pp. 457–511.

GHOSH, M., (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association*, 87, 418, pp. 533–540.

JIANG, J., LAHIRI, P., (2006). Mixed model prediction and small area estimation. *TEST*, 15, pp. 1–96.

JONES, H. E., SPIEGELHALTER, D. J., (2011). The identification of unusual health-care providers from a hierarchical model. *American Statistician*, 65, pp. 154–163, DOI: 10.1198/tast.2011.10190.

LAHIRI, P., (1990), "Adjusted" Bayes and empirical Bayes estimation in finite population sampling, *Sankhya*, B, 52, pp. 50–66.

LANGFORD, I. H., LEWIS, T., (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, Ser. A*, 161, pp. 121–160.

LI, H. and LAHIRI, P., (2010). Adjusted maximum method for solving small area estimation problems, *Journal of Multivariate Analysis*, 101, pp. 882–892.

LOUIS, T. A., (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association*, 79, 386, pp. 393–398.

MARSHALL, R. J., (1991). Mapping disease and mortality rates using empirical Bayes estimators. *Appllied Statistics*, 40, pp. 283–294.

MOLINA, I., NANDRAM, B., RAO, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical Bayes approach. *The Annals of Applied Statistics*, Vol. 8, No. 2, pp. 852–885.

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38, pp. 369–385.

MOLINA, A., RICHARDSON, S., (1991). Empirical Bayes estimates of cancer mortality rates using spatial models. *Statistics in Medicine*, 10, pp. 95–112.

MORRIS, C. N., CHRISTIANSEN, C. L., (1996). Hierarchical models for ranking and for identifying extremes, with applications, in Bayesian statistics 5, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F.M. Smith, Oxford: Oxford University Press, pp. 277–296.

NANDRAM, B., SEDRANSK, J., PICKLE, L. W., (2000). Bayesian analysis and mapping of mortality rates for chronic obstructive pulmonary disease. *Journal of American Statistical Association*, 95, pp. 1110–1118.

PFEFFERMANN, D., (2013). New important developments in small area estimation. *Statistical Science*, 28, pp. 40–68.

PRASAD, N. G. N., RAO, J. N. K., (1990). The estimation of the mean squared error of small-area estimators. *Journal of American Statistical Association*, 85, pp. 163–171.

RAO, J.N.K. and MOLINA, A., (2015). *Small area estimation.* Wiley.

RAVALLION, M., (1994). *Poverty comparisons.* Chur, Switzerland: Harwood Academic Press.

SHEN, W., LOUIS, T. A., (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of Royal Statistical Society Series B*, 60,Part 2, pp. 455–471.

TSUTUKAWA, R. K., SHOOP, G. L., MARIENFELD. C. J., (1985). Empirical Bayes estimation of cancer mortality rates. *Statistics in Medicine*, 4, pp. 201–212.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 254–255

About Guest Co-Editors

Guest Editor-in-Chief

Lahiri Partha is a Professor of Survey Methodology and Mathematics at the University of Maryland, College Park, and an Adjunct Research Professor at the Institute of Social Research, University of Michigan, Ann Arbor. His areas of research interest include data linkages, Bayesian statistics, survey sampling and small-area estimation. Dr. Lahiri served on the editorial board of a number of international journals such as the Journal of the American Statistical Association and Survey Methodology. He also served on many advisory committees, including the U.S. Census Advisory committee and the U.S. National Academy panel. Over the years, Dr. Lahiri has advised various local and international organizations, such as the United Nations Development Program, the World Bank, and the Gallup Organization. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics, and an elected member of the International Statistical Institute.

Guest Co-Editors

Beaumont Jean-François presently serves as a Chief Researcher in the International Cooperation and Methodology Innovation Centre at Statistics Canada, where he has worked for the last 24 years. He received his Master's degree in Statistics from Laval University, Canada, after completing a baccaluareate in Actuarial Science. His main current research topics of interest are: statistical data integration, small area estimation, robust estimation for finite populations and bootstrap variance estimation for sample surveys. He is currently an associate editor of *Survey Methodology and Metron*. Earlier he served as interim editor of *Survey Methodology*.

Chaudhuri Sanjay is currently an Associate Professor in the Department of Statistics and Applied Probability at the National University of Singapore. He received his PhD in Statistics from the University of Washington, Seattle, after completing both the undergraduate and postgraduate degrees at the Indian Statistical Institute in Calcutta, India. His current research interests include the development and analysis of statistical methodologies, especially when based on empirical likelihood, analysis of complex survey data, order restricted inference, graphical Markov models, survival

analysis and application of statistics to real-life problems. Sanjay is a current member of the editorial boards of the *Journal of American Statistical Association*, *Journal of Non-parametric Statistics*, *Journal of Statistical Planning and Inference*, *Sankhya series B*, *Stat*, *Statistics and Applications* and *Statistics in Transition new series*.

Drechsler Jörg is a Distinguished Researcher in the Department for Statistical Methods at the Institute for Employment Research in Nürnberg, Germany. He received his PhD in Social Science from the University in Bamberg and his DSc in Statistics from the Ludwig-Maximilians University in Munich. He is also an Adjunct Associate Professor in the Joint Program in Survey Methodology at the University of Maryland College Park, and an Honorary Professor at the University of Mannheim, Germany. His main research interests are data confidentiality and non-response in surveys. He served on the editorial board of the *Journal of Official Statistics*, *Journal of Privacy and Confidentiality*, *Journal of Survey Statistics and Methodology*, *Journal of the Royal Statistical Society*, Series A, and *Wirtschafts- und Sozialstatistisches Archiv*.

Larsen Michael D. is a Professor and a Department chair at Saint Michael's College, Colchester, Vermont, USA. He received a Ph.D. in Statistics from Harvard University. His areas of interest include record linkage, survey sampling, missing data, modelling of complex data, and teaching statistics and probability. He has been a faculty member at six institutions with a tenure at three. He has collaborated with the U.S. Census Bureau, National Cancer Institute, and the National Center for Health Statistics on record linkage. He serves as a statistical consultant for companies and government agencies, including the U.S. Veterans Administration, the Internal Revenue Service, and the Department of Labor. He is a Fellow of the American Statistical Association and an elected member of the International Statistical Institute. He served as president of the Washington Statistical Society and executive editor of *Chance* journal.

Szymkowiak Marcin is currently an Assistant Professor in the Department of Statistics at Poznań University of Economics and Business (Institute of Informatics and Quantitative Economics). He is also a consultant of the Centre for Small Area Estimation at the Statistical Office in Poznań. He has many years of experience in the field of statistical data analysis, including official statistics. Dr. Szymkowiak specializes in small area estimation, methods of dealing with non-response (imputation and calibration), survey sampling, statistical methods of data integration (probabilistic record linking, statistical matching), and multivariate data analysis. He has participated in many domestic and international projects in cooperation with the Central Statistical Office, the World Bank and Eurostat.

STATISTICS IN TRANSITION new series, Special Issue, August 2020
Vol. 21, No. 4, pp. 256–261

About the Authors

Alam Jahangir M. is an Assistant Professor in the Economics Department at Truman State University in the United States. His main areas of research interest are: applied macroeconomics, international trade and economic growth and development. Within these areas, his special focus is on firm productivity, allocation of resources, cross-country income and productivity differences. As a faculty member at Truman, he has taught several courses related to his research interest.

Bera Sabyasachi is a graduate student in the School of Statistics at the University of Minnesota.

Burgard Jan Pablo is an Associate Professor at the Trier University in the field of statistics, econometrics and data science. His main research interests are: methods for regional estimation, handling of errors in variables in statistical models and the development of tailor-made quantitative solutions for interdisciplinary empirical problems.

Bonnery Daniel is an Assistant Researcher in the Epidemiology and Modelling group, Department of Plant Sciences, University of Cambridge. Previously, he worked at the INSEE, Ensai, University of Toulouse I Capitole, and at the University of Maryland. His research interest is, broadly, accounting for the selection process in statistics.

Cai Song is an Associate Professor in the School of Mathematics and Statistics at Carleton University, Canada. His research interests include empirical likelihood inference, asymptotic statistical methods, small area estimation, modern survey methodology and statistical computing for big and/or high-dimensional data.

Chatrchi Golshid is a mathematical statistician at Statistics Canada. She received her MSc (2012) and PhD (2019) degrees in Statistics from Carleton University. She is currently working in the Social Statistics Methods Division as a Senior Methodologist. Her research interests include small area estimation and sampling.

Chatterjee Snigdhansu is a Professor in the School of Statistics and the Director ofthe Institute for Research in Statistics and its Applications (IRSA, http://irsa.stat.umn.edu/) at the University of Minnesota. His research interests include statistical foundations of data science and machine learning, high dimensional data geometry, Bayesian statistics, resampling methods, and applications of statistics, artificial intelligence and machine learning in multiple domains.

Cheng Yang is a Senior Mathematical Statistician at the Substance Abuse and Mental Health Services Administration (SAMHSA). He is also an Adjunct Professor of Statistics at the George Washington University. Earlier, he served as a Lead Scientist for the Current Population Survey (CPS), American Time Use Survey (ATUS) and Housing Vacancy Survey (HVS) at the U.S. Census Bureau. His research interests include statistical modelling, survey methodology, small area estimation, labour force statistics and health statistics.

Di Consiglio Loredana, a PhD in Statistical Methods, is a senior researcher at the Italian National Institute, Directorate for Methodology and Statistical Process Design, and currently a project manager for methods of estimation of integrated population census and system of statistical registers. Her main activities and research interest relate to sampling, small area estimation, population size estimation, statistical methods for multi-source statistics and big data for official statistics. She has been involved in numerous international projects and task forces in the fields of small area estimation and use of multi-sources.

Dieckmann Hanna is a Research Associate at the Economic and Social Statistics Department at Trier University. Her research interests include spatial microsimulation and linking micro and macro models.

Dostie Benoit is a Full Professor at the Department of Applied Economics of HEC Montréal. He is also the academic director of the Québec inter-University Centre for Social Statistics (QICSS) and a member of the board of the Canadian Research Data Center Network (CRDCN). He received his PhD in economics from Cornell University in 2001, and is a Fellow at the IZA and CIRANO. His research interests include statistical models for linked employer-employee data, duration models, returns to human capital, firm-sponsored training, productivity, turnover and labour reallocation.

Drechsler Jörg - see co-editors.

Dumitrescu Laura is a Lecturer in the School of Mathematics and Statistics at Victoria University of Wellington. Her main areas of interest include asymptotic inference, complex sampling surveys, small area estimation, longitudinal data analysis and measurement errors.

Gershunskaya Julie is a Mathematical Statistician with the Statistical Methods Staff of the Office of Employment and Unemployment Statistics at the U.S. Bureau of Labor Statistics. Her main areas of interest include robust small area estimation and treatment of influential observations, with the application to the U.S. Current Employment Statistics Program.

Ghosh Malay is a Distinguished Professor of Statistics at the University of Florida. His current research interests include small area estimation, Bayesian variable selection and multiple testing, machine learning and general Bayesian theory. He is

a Fellow of the Institute of Mathematical Statistics and of the American Statistical Association. Dr. Ghosh has supervised over 60 PhD students. He is the recipient of the Jerzy Neyman Medal of the Polish Statistical Society in 2012, the Lifetime Achievement Award of the International Statistical Association in 2017, the Small Area Estimation award in 2019, and more recently, in 2020, the Samuel S. Wilks Memorial Award of the American Statistical Association.

Han Ying has worked as a Statistician at Gallup, Inc., a management consulting company based in Washington, D.C. since 2018, when she received her doctorate degree in Mathematical Statistics from University Maryland, College Park. She provides her expertise in sampling designs and weighting to collect reliable survey data internationally, through phone or face-to-face interviews, and applies statistical methods to answer complex research questions. Her main areas of interest include sampling methodology, record linkage and small area estimation.

Kedem Benjamin is a Professor in the Department of Mathematics, where he is the Director of the Statistics Program, and an affiliate faculty member of the Institute of Systems Research, University of Maryland, College Park. His main areas of interest are time series analysis, semiparametric statistical inference and data fusion, including repeated, out of sample fusion in the estimation of very small tail probabilities, using computer generated data.

Krause Joscha is a post-doctoral Researcher at the Economic and Social Statistics Department at Trier University. His research interests are: statistical modelling, regularized regression analysis, and robust and computational statistics with applications in the fields of social medicine and epidemiology.

Lahiri Partha - see co-editors.

Li Yan is a Professor in the Joint Program in Survey Methodology (JPSM) and in the Department of Epidemiology and Biostatistics at the University of Maryland, College Park, and an Adjunct Research Professor of the Institute of Social Research, University of Michigan, Ann Arbor. A primary research goal of Dr. Li is to combine her background in computer sciences, genetics, and survey methodology with her research experience in cancer epidemiology and genetics, to develop statistical methods that would enhance designing and analyzing complex samples in social and biomedical settings. She is particularly interested in the area of nonprobability sample design/analysis, health disparity, statistical genetics involving statistical inferences using case-control, cohort and cross-sectional studies, and surveys with complex survey designs

Merkle Hariolf is a Research Associate at the Economic and Social Statistics Department at Trier University. His main fields of interest are synthetic datasets and spatial microsimulations.

Molina Isabel is an Associate Professor at the Department of Statistics at Universidad Carlos III de Madrid, where she is also a Deputy Director. Her main areas of interest include small area estimation, especially as applied to poverty mapping, linear and generalized linear mixed models, survey methodology and resampling techniques (bootstrap). Currently, she is a member of two editorial boards: *Survey Methodology* and *Journal of Survey Statistics and Methodology*. She is the co-author of the 2nd edition of the *Small Area Estimation* published by Wiley.

Moura Fernando is a Full Professor at the Statistics Department, Federal University of Rio de Janeiro (UFRJ). Currently he has served as head of Statistics Department of UFRJ. He is an elected member of the International Statistical Institute (ISI). He is also an associate editor of the *International Journal REVSTAT*. He completed his PhD in Statistics at the University of Southampton and has an MSc and a BSc in Statistics and Actuary Science. His main areas of interest include small area estimation, survey methods, Bayesian statistics, statistical modelling and statistical methods applied to actuary science.

Münnich Ralf is the Chair of the Economic and Social Statistics Department at Trier University, and the president of the Research Institute for Official and Survey Statistics. His main research interests are: survey statistics, sampling designs, variance estimation and data quality in complex surveys, Monte Carlo and computer-intensive statistical methods, small area statistics, statistical indicators and statistical foundations of microsimulations.

Neufang Kristina M. is a Research Associate at the Economic and Social Statistics Department at Trier University. Her research interests are dynamic microsimulations and multisource estimation.

Neves André is an Economist in the Brazilian Institute of Geography and Statistics (IBGE). He worked at two structural economic surveys: the Annual Service Survey (PAS) and Annual Trade Survey (PAC). He has currently been involved in the organization of the Monthly Service Survey. He graduated from the National School of Statistical Sciences (ENCE), and his Master's thesis featured population studies and social surveys. Presently, he is a PhD student at ENCE. His main areas of interest are economic statistics and small domain estimation.

Newhouse David is a Senior Economist in the World Bank's poverty and equity global practice, and a fellow at the Institute of Labor Economics. In his scientific work, he has focused on the economic measurement and analysis related to poverty and jobs in developing countries. He currently co-leads the Sub-Saharan Africa team for statistical development, the expansion of the Bank's Global Monitoring Database of harmonised household survey data, and efforts to incorporate satellite imagery into poverty measurement. He holds a PhD in Economics from Cornell University,

and has published over twenty articles in respected field journals on labour, poverty, health and education in developing countries.

Pfeffermann Danny is a Professor of Statistics at the University of Southampton, UK, and Professor Emeritus at the Hebrew University of Jerusalem, Israel. As of 2013, he has been the Government Statistician and Director General of the Central Bureau of Statistics in Israel. He is a past President of the Israel Statistical Society and a past President of the International Association of Survey Statisticians (IASS). For the last 20 years, he has also served as a consultant for the US Bureau of Labor Statistics. Danny Pfeffermann is a fellow of the American Statistical Association and an elected member of the International Statistical Institute (ISI). He received a BA degree in Mathematics and Statistics and MA and PhD degrees in Statistics from the Hebrew University of Jerusalem. He is a recipient of numerous prestigious awards, including Waksberg award in 2011, the West Medal by the Royal Statistical Society in 2017, the Julius Shiskin Memorial Award for Economic Statistics in 2018, and the SAE 2018 Award for his distinguished contribution to the SAE methodology and the advancement of Official Statistics in the Central Bureau of Statistics in Israel.

Pyne Saumyadipta is the Scientific Director of the Public Health Dynamics Laboratory at University of Pittsburgh, USA. His research interests include big data systems in life sciences and health informatics, data fusion, high-dimensional data simulation and modelling to study heterogeneity in populations. He is the President of Health Analytics Network and an Associate Editor of the *Statistics and Applications* journal. Formerly, he has held the P.C. Mahalanobis Chair and was at the same time professor and head of bioinformatics at the C. R. Rao Advanced Institute of Mathematics, Statistics and Computer Science. He co-edited 'Big Data Analytics', published by Springer in 2016, and a 2-volume 'Handbook of Statistics: Disease Modelling and Public Health', published by Elsevier in 2017.

Rao J. N. K. is a Distinguished Research Professor at Carleton University, Ottawa, Canada. In recent years, his research focused on small area estimation (SAE), which resulted in the publication of two books, Rao (2003) and Rao and Molina (2015),through Wiley publishing house. He received the First Award for outstanding contributions to SAE at SAE (2017) conference in Paris. He is also the recipient of the Waksberg Award (2005) for his contributions to survey sampling methodology. He also received honorary doctorates from University of Waterloo (2008) and the Catholic University of Sacred Heart, Italy (2013).

Saegusa Takumi is an Assistant Professor at the Department of Mathematics, University of Maryland, College Park. His main areas of interest include complex sampling in public health, empirical process theory and semiparametric inference.

Silva Denise is a Principal Researcher at the National School of Statistical Sciences (ENCE) from the Brazilian Institute of Geography and Statistics (IBGE). She is also

the president of the International Association of Survey Statisticians (IASS), has been vice-president of the Inter-American Statistical Institute (IASI), and is an elected member of the International Statistical Institute (ISI). She is an associate editor of *International Statistical Review*, *Statistical Journal* of the IAOS and *Revista Brasileira de Estatística*. She has completed her PhD in Statistics at the University of Southampton and has an MSc and a BSc in Statistics. Her main areas of interest are survey methods, statistical modelling for social sciences, small area estimation and time series analysis.

Schmaus Simon is a Research Associate at the Economic and Social Statistics Department at Trier University. His research areas include dynamic microsimulation models and synthetic data generation methods.

Suntornchost Jiraphan is an Assistant Professor at the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand. Her main areas of interest include various fields in probability theory and statistics, such as statistical inference, statistical distributions, survival analysis, small area estimation and time series analysis.

Tuoto Tiziana is the Project Manager for data integration methods at the Italian National Institute of Statistics, Directorate for Methodology and Statistical Process Design. Her main research interests are: survey methodology, statistical methods for data integration, record linkage, non-sampling errors, population size estimation, statistical methods for multi-source statistics and big data for official statistics. She has been involved in many internationally-founded projects in the field of data integration aimed at improving censuses in low-income countries and at wider utilisation of big data in official statistics. She collaborates with several international organizations, providing lectures and on-the-job training. She serves as a reviewer in several international peer-reviewed journals.

Vilhuber Lars is the Executive Director of the Labor Dynamics Institute at Cornell University and a Senior Research Associate in the Economics Department of that University. He is also data editor for the American Economic Association, co-chair of Innovations in Data and Experiments for Action (IDEA) Initiative at J-PAL, and managing editor of the *Journal of Privacy and Confidentiality*. Lars Vilhuber moreover chairs the Scientific Committee of the French Centre d'accès sécurisé aux données, is the incoming chair of the American Statistical Association's Committee on Privacy and Confidentiality, and a member of the board of the Canadian Research Data Center Network (CRDCN).

Zhang Xuze is a PhD Candidate in Mathematical Statistics at the Department of Mathematics, University of Maryland, College Park. He is under the scientific guidance of Professor Benjamin Kedem, and his research interest is semiparametric inference in data fusion.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page*: http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).