

# **Current challenges in survey research:**

## **Graham Kalton on probability and nonprobability sampling**

## **Malay Ghosh on small area estimation**

with comments

SCIENTIFIC EDITORS: WŁODZIMIERZ OKRASA, DOMINIK ROZKRUT



Statistical Research Papers  
Volume 5

SCIENTIFIC EDITORS:  
WŁODZIMIERZ OKRASA  
DOMINIK ROZKRUT

# Current challenges in survey research:

**Graham Kalton on probability and nonprobability sampling**

**Malay Ghosh on small area estimation**

with discussions by:

Julie Gershunskaya

Ying Han

Partha Lahiri

Risto Lehtonen

Yan Li

Isabel Molina

Ralf Münnich

David Newhouse

Danny Pfeffermann

J. N. K. Rao



**Papers originally published in and peer-reviewed for *Statistics in Transition new series*, vol. 21, 2020, 4, *Statistics in Transition new series* and *Statistics of Ukraine*, Joint Special Issue, vol. 24, 2023, 1, and the jubilee *Statistics in Transition new series* issue, vol. 24, 2023, 3**

**Scientific consultation**

Tomasz Żądło – PhD, DSc, Assoc. Prof., University of Economics in Katowice, Poland

**Coordination of the editorial process**

Statistics Poland, Statistical Products Department, Scientific Journals Division

**Editorial work, printing and binding**

Statistical Publishing Establishment

Publication available at: <https://srp.stat.gov.pl>

Quoting from the publication requires providing the source

Warszawa 2024

© Copyright by Główny Urząd Statystyczny, Polskie Towarzystwo Statystyczne and the authors, some rights reserved. CC BY-SA 4.0 licence 

**ISBN 978-83-67087-96-4**

**e-ISBN 978-83-67087-97-1**

# Contents

<b>Introduction</b> .....	<b>5</b>
<b>1. Probability and nonprobability sampling</b> .....	<b>7</b>
<b>Graham Kalton</b> Probability vs. nonprobability sampling: from the birth of survey sampling to the present day .....	<b>9</b>
<i>Comments</i>	
<b>Danny Pfeffermann</b> .....	<b>30</b>
<b>Risto Lehtonen</b> .....	<b>33</b>
<b>Julie Gershunskaya, Partha Lahiri</b> .....	<b>37</b>
<b>Ralf Münnich</b> .....	<b>45</b>
<b>Graham Kalton, Rejoinder</b> .....	<b>49</b>
<i>Post Scriptum</i>	
<b>Włodzimierz Okrasa, Dominik Rozkrut</b> Celebrating the 100th issue and the 30th anniversary .....	<b>53</b>
<b>2. Small area estimation</b> .....	<b>55</b>
<b>Malay Ghosh</b> Small area estimation: its evolution in five decades .....	<b>57</b>
<i>Comments</i>	
<b>Julie Gershunskaya</b> .....	<b>78</b>
<b>Ying Han</b> .....	<b>84</b>
<b>Yan Li</b> .....	<b>89</b>
<b>Isabel Molina</b> .....	<b>94</b>
<b>David Newhouse</b> .....	<b>99</b>
<b>Danny Pfeffermann</b> .....	<b>105</b>
<b>J. N. K. Rao</b> .....	<b>107</b>
<b>Malay Ghosh, Rejoinder</b> .....	<b>113</b>
<b>3. Reconstructing Ukraine's statistical system</b> .....	<b>123</b>
<b>Dominik Rozkrut, Włodzimierz Okrasa, Oleksandr H. Osaulenko, Misha V. Belkindas, Ronald L. Wasserstein</b> The post-conflict reconstruction of the statistical system in Ukraine. Key issues from an international perspective .....	<b>125</b>
<b>4. Appendix</b> .....	<b>137</b>
Scientific articles published in special issues of <i>Statistics in Transition new series</i> in 2015–2023	

Contents

<i>Statistics in Transition new series</i> and <i>Survey Methodology</i> , Small Area Estimation – Joint Issue Part 1 (post-SAE2014 papers), Volume 16, Number 4, December 2015 .....	<b>138</b>
<i>Statistics in Transition new series</i> and <i>Survey Methodology</i> , Small Area Estimation – Joint Issue Part 2 (post-SAE2014 papers), Volume 17, Number 1, March 2016 .....	<b>143</b>
<i>Statistics in Transition new series</i> , Statistical Data Integration – Special Issue , Volume 21, Number 4, August 2020 .....	<b>149</b>
<i>Statistics in Transition new series</i> and <i>Statistics of Ukraine</i> , A New Role for Statistics – Joint Special Issue, Volume 24, Number 1, February 2023 .....	<b>157</b>

# Introduction

The purpose of this publication is to commemorate the 100th issue of *Statistics in Transition new series* and the 30th anniversary of the launch of the journal. This coincidence provides an excellent opportunity to recall several previous publications that *Statistics in Transition new series* occasionally releases as Special Issues, focusing on specific topics of current research interest.

The book includes the following three articles from the recent special issues of the journal:

- Graham Kalton's paper entitled *Probability vs. nonprobability sampling: from the birth of survey sampling to the present day*, together with four discussion papers focusing on current research and future directions of development in the field;
- Malay Ghosh's paper entitled *Small area estimation: its evolution in five decades*, together with seven discussion papers;
- A multi-authored paper by Dominik Rozkrut, Włodzimierz Okrasa, Oleksandr H. Osaulenko, Misha V. Belkindas, and Ronald L. Wasserstein, entitled *The post-conflict reconstruction of the statistical system in Ukraine. Key issues from an international perspective*. This article is based on presentations given during a special session of the Federal Committee on Statistical Methodology 2022 Research and Policy Conference (Washington, 25–27 October 2022), entitled *A Marshall Plan for Reconstructing National Statistical Offices After Conflict: Practical Guidance from International Principles*.

In addition, the Appendix contains tables of contents and forewords (prefaces) for each of the selected special issues published in the years 2015–2023:

- *Small Area Estimation I. Statistics in Transition new series and Survey Methodology* Joint Issue, Vol. 16/4 (December 2015). Guest Editors: Raymond Chambers and Malay Ghosh;
- *Small Area Estimation II. Statistics in Transition new series and Survey Methodology* Joint Issue, Vol. 17/1 (March 2016). Guest Editors: Risto Lehtonen and Graham Kalton;
- *Statistical Data Integration*. Special Issue, Vol. 21/4 (August 2020). Guest Editor: Partha Lahiri;

- *A New Role for Statistics. Statistics in Transition new series and Statistics of Ukraine* Joint Special Issue, Vol. 24/1 (February 2023). Editors: Włodzimierz Okrasa and Oleksandr H. Osaulenko.

All the papers published in *Statistics in Transition new series* are available on an open-access basis at <https://sit.stat.gov.pl/archives>

## Scientific editors

Prof. Włodzimierz Okrasa



Editor-in-Chief  
*Statistics in Transition new series*

Dr Dominik Rozkrut



President  
*Statistics Poland*

# **1. Probability and nonprobability sampling**





# Probability vs. nonprobability sampling: from the birth of survey sampling to the present day<sup>1</sup>

**Abstract:** At the beginning of the 20th century, there was an active debate about random selection of units versus purposive selection of groups of units for survey samples. Neyman's (1934) paper tilted the balance strongly towards varieties of probability sampling combined with design-based inference, and most national statistical offices have adopted this method for their major surveys. However, nonprobability sampling has remained in widespread use in many areas of application, and over time there have been challenges to the Neyman paradigm. In recent years, the balance has tilted towards greater use of nonprobability sampling for several reasons, including: the growing imperfections and costs in applying probability sample designs; the emergence of the internet and other sources for obtaining survey data from very large samples at low cost and at high speed; and the current ability to apply advanced methods for calibrating nonprobability samples to conform to external population controls. This paper presents an overview of the history of the use of probability and nonprobability sampling from the birth of survey sampling at the time of A. N. Kizær (1895) to the present day.

**Key words:** Anders Kizær, Jerzy Neyman, representative sampling, quota sampling, hard-to-survey populations, model-dependent inference, internet surveys, big data, administrative records.

## 1. Introduction

This paper presents a selection of the major developments that have taken place over the years since social surveys were first introduced in the late 19th century. I restrict my coverage to surveys of households and persons and my focus is on the sampling methods used to conduct such surveys. Major changes have also taken place in modes of data collection, in questionnaire design, and in other aspects of survey research over the years, but these topics are outside the scope of this paper. My paper

---

<sup>a</sup> Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA.  
E-mail: gkalton@gmail.com. ORCID: <https://orcid.org/0000-0002-9685-2616>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 1–22.  
<https://doi.org/10.59170/stattrans-2023-029>.

on the more general theme of survey research over the past 60 years overlaps with this paper and gives greater coverage on some topics (Kalton, 2019).

The changes that have occurred in methods of survey sampling have arisen for many reasons, including developments in sampling theory, the continuing growth in computer power (that was non-existent for the first fifty years of survey research), new sampling frames, and the problems created by a broader and more challenging range of applications of social surveys that has occurred as the potential for survey research has been more fully recognized. While acknowledging these changes, it is noteworthy that many aspects of the sampling methods that have been superseded over time have remained relevant. Indeed, much of the current discussion of the use of nonprobability sampling and big data sources has roots in the early days of survey research.

Without attempting to date the origins of survey research, early applications of survey research for studying the social conditions of populations took off in the late 1800's. English examples include Charles Booth's large-scale survey of the social conditions of the population of London that was started in 1886, Seebohm Rowntree's survey of working-class poverty in York that was conducted a decade later, and Bowley's survey of working-class conditions in Reading in 1912, which he followed up with surveys in four other English towns (three of which were conducted by Burnett-Hurst under Bowley's direction). See Caradog Jones (1949) for the early surveys in England, Converse (2017) for an account of the history of survey research in the United States from its beginnings at the turn of the century through until 1960, and Stephan (1948) for a history of the use of sampling procedures dating back from earlier times through until the 1940's, primarily in the United States.

The London and York surveys were complete censuses of the surveys' target populations. As complete censuses, they were deemed statistically acceptable at the time; they were known as 'monographs' of their local communities. For the London survey, the target population was households with school-aged children, while for the York survey it was households that did not have servants (conducted only in streets that were likely to contain households without servants). Bowley had long argued for the use of sampling for such surveys, and he played a major role in its adoption (Aldrich, 2008). He used sampling for the first time in the five towns surveys, where systematic sampling was employed (Bowley, 1913), and he introduced the idea of measuring sampling errors for survey estimates.

As Kish (1995) notes, the emergence of the field of survey sampling can be dated from work led by the Norwegian statistician Anders Kiær, the first Director of Statistics Norway. Kiær developed a sampling method that he termed "representative sampling". Kiær's method of purposive sampling is worth reviewing both for the

procedures he devised to make a sample nationally ‘representative’ and for the reactions to the method from statisticians attending meetings of the International Statistical Institute (ISI) at the time. The next section provides a brief overview of these issues.

## 2. Kiær’s representative method of statistical surveys

Kiær’s sampling methodology is described in detail in his monograph *The Representative Method of Statistical Surveys*, first published in Norwegian in 1897 and republished in 1976 with an English translation (Kiær, 1976). The monograph provides a good deal of detail on the sample designs Kiær developed for two large-scale surveys—one on personal income and property (PIP) and the other on living conditions (LC) – as well as reporting the objections to his methods that he received when he presented them at ISI meetings. As distinct from the surveys of English towns cited above, Kiær aimed to produce survey estimates for the whole of Norway. For this purpose, he developed two-stage area sample designs for his surveys: at the first stage, he selected a “representative” sample of administrative districts (rural districts or counties, towns, and cities); at the second stage, he drew samples of people for each survey. The choice of the sampled first-stage units was carefully fashioned to give geographical spread and to achieve a good representation of the Norwegian population in terms of characteristics collected in the 1891 Population Census (e.g., age, marital status, occupation, urbanicity).

The sample for the PIP survey was defined as men aged 17, 22, 27, etc. who had names starting with certain letters, selected from 1891 census records that were being processed at the time, with a total sample size of around 11,400 men. The sample size for the LC survey was around 80,000 adults. The sample size to be obtained in each selected rural county was specified based on calculations from census data; within selected counties, the enumerators were instructed to follow certain routes and to select different types of houses, but otherwise they were left to make the selections. In the smaller towns, every 9th, 5th, or 3rd house was selected. An extra sampling stage was introduced in the largest towns. For example, the sample of houses in Oslo was selected within a sample of streets. Moreover, a higher proportion of the streets with larger populations was included in the sample, but this feature was counterbalanced by the selection of houses at a lower rate in the large streets.

The driving objective with Kiær’s approach was to produce a representative sample that would constitute a microcosm of the Norwegian population. He invented some intricate methods to attempt to achieve this objective. His purposive selection of first stage administrative units sometimes incorporated ideas of probability proportional to size sampling and subsampling at different rates in compensation, thereby avoiding an excessive sample concentration in a few large districts. Similarly,

his street sample in Oslo has the same feature. He also employed a simple 2:1 weighting adjustment to compensate for the smaller proportion of members of the rural population in the PIP survey. (Before the advent of computers, anything other than simple integer weighting adjustments would have been extremely difficult to routinely apply.)

Despite his thoughtful approach, Kiær encountered a great deal of criticism of his methods when he presented them to the ISI in 1895. The dominant criticism, however, was not of the representative method, *per se*, but rather of a sample-based enquiry rather than a complete enumeration. In the words of one strong critic, von Mayr: “We remain firm and say: no calculations when observations can be made”. Kiær also made presentations on the representative method at the 1897, 1901, and 1903 ISI sessions, at which they were subjected to similar criticisms, together with another one. At the 1903 session, von Bortkiewicz reported the results of a significance test he had conducted that found that Kiær’s representative samples were not truly representative. See Kruskal and Mosteller (1980) for a detailed account of the ISI sessions.

At the same time, Kiær expertise was under attack at home for the LC survey, which was conducted on behalf of a parliamentary labor commission to inform a very contentious social security act that would provide highly expensive disability insurance. A three-person “critique committee” was established to review the commission’s major recommendation and its statistical basis. One committee member, the actuary Jens Hjorth, was extremely critical of Kiær’s statistics, including the survey design, the representative sample design, and the analysis. The attacks on the statistics that Kiær’s produced for the commission were forceful, extensive, and widely debated. In the end, based on the results of some new surveys, Kiær admitted that he had initially seriously underestimated the extent of disability. After that time, representative sampling for large-scale surveys disappeared in Norway. Lie (2002) provides an informative account of the rise and fall of Kiær’s representative sampling method.

The ISI discussion of survey sampling fell into abeyance until 1924 when the ISI appointed a commission for studying the application of the representative method in statistics. By that time, the idea of a “partial investigation” was widely accepted. In its 1926 report (Jensen, 1926), the Commission concluded that a sample was acceptable if it was sufficiently representative of the whole. To satisfy this condition the sample could be produced either by random selection with equal probability or by purposive selection of groups with a representative overall sample. The report also recommended that the survey results should, wherever possible, be accompanied by an indication of the errors to which they are liable.

### 3. Neyman's seminal paper

In 1934, Neyman presented his classic paper comparing the methods of random and purposive selection to the Royal Statistical Society (Neyman, 1934). Covering more than the comparison, the paper contained a detailed discussion of a methodology for making inferences from random—or, more generally, probability—samples of finite populations, including providing a definition of a confidence interval in this context. He also critically examined the assumptions made when using data from a purposive sample to produce an accurate estimate of a population parameter.

He discussed the sample design of purposive selection of groups used by Gini and Galvani in selecting a sample of records from the already-processed Italian General Census of 1921 that was to be used as the basis for later analysis. For their sample, Gini and Galvani (1929) selected a sample of twenty-nine of the 214 districts in Italy, balanced on seven covariables (note that departs from Kær's stipulation that a large wide-spread sample of areas is needed). While the sample worked well for the averages of the control variables, it often failed to adequately represent the national population for other characteristics, and for the distributions of the control variables. These findings led them to raise questions about representative sampling.

Neyman's paper was a watershed for survey sampling, leading to widespread adoption of probability sampling, particularly by national statistical offices. It also led to the development of an extensive range of sampling methods and the associated theory applicable to a variety of practical survey problems, as described in the several texts on survey sampling that appeared in the 1950's. The many contributions of statisticians at the U.S. Census Bureau led by Morris Hansen are particularly noteworthy; see, for example, the two-volume text by Hansen, Hurwitz, and Madow (1953). Statisticians active in research on sample designs for agricultural surveys, such as Yates in England and Mahalanobis in India, also made important contributions to the advancement of the subject. The sampling text by Yates (1949) was among the first books on survey sampling methods. In 1950, Mahalanobis went on to establish and lead the famous socio-economic National Sample Survey (NSS) of India. An interesting feature of the NSS sample design was that the sample was composed of four replicate samples. The survey results were presented for each replicate separately as well as for the full sample, with the aim of communicating to readers an indication of the amount of sampling error in the survey estimates (see, for example, Mahalanobis, 1946). This was thus a forerunner of variance estimation using replication methods.

Note that perfect application of Neyman's design-based inference for probability sampling depends on:

- The availability of a sampling frame that provides complete coverage of the finite target population;

- A sample design that assigns known and non-zero selection probabilities to every element in the target population;
- Survey responses from every sampled unit; and
- The use of survey weights in the analysis to compensate for unequal selection probabilities.

Under these conditions (and assuming no response errors), survey estimates can be computed that are design-consistent estimates of the population parameters without the need to make any assumptions about the characteristics of the survey population. Model assumptions made about the population structure may be used to make the sample design more efficient or in the computation of the survey estimates, but the consistency of the survey estimates remains irrespective of the validity of the model. What the model assumptions do affect is the precision of the survey estimates. For example, in a stratified sample, if the sampling fraction in a stratum is set at a higher rate because the elements in a stratum are incorrectly modeled to be more variable, the (weighted) sample mean will still be unbiased, but it will be less precise than if the stratum element variance has been correctly modeled. Similarly, if a set of auxiliary variables  $\mathbf{X}$  is available for all population elements, and a function of the  $x$ 's,  $f(\mathbf{X})$ , is used as a working model to predict the survey variable  $y$ , then the finite population total may be estimated by

$$\hat{Y}_d = \Sigma_U \hat{f}(X_i) + \Sigma_s w_i e_i, \quad (1)$$

where  $\Sigma_U$  and  $\Sigma_s$  denote summations over the population and sample respectively,  $\hat{f}(X_i)$  denotes the model estimate of  $y_i$  using the sample estimates of the unknown model parameters,  $e_i = y_i - \hat{f}(X_i)$ , and the weight  $w_i$  is the inverse of element  $i$ 's selection probability. By including the weighted estimate of the population total of the  $e_i$ 's in this estimate,  $\hat{Y}_d$  is a consistent estimator of the population total  $Y$  irrespective of the suitability of the working model; the choice of working model affects only the precision of the estimate  $\hat{Y}_d$ . This estimator is model-assisted, using the terminology coined by Särndal, Swensson, and Wretman (1992), but it is not model-dependent. For simple random sampling, Cochran (1953) gave an early example of a model-assisted estimator with the ratio estimator  $\hat{Y} = (\bar{y}/\bar{x})X$ , where  $X$  denotes the population total for the auxiliary variable  $x$ . An additional, important, feature of design-based inference is that estimates of the variances of sample estimates can be computed from the sample itself.

While the lack of dependence of design-based inference on model assumptions is the major attraction of probability sampling, it needs to be acknowledged that probability sampling is rarely perfectly executed in practice. There are two main sources of imperfection: noncoverage and nonresponse. Noncoverage, which arises because

the sampling frame fails to include some elements of the target population, is widespread and its magnitude is often underrated. Area sampling is widely used in social surveys, selecting a probability sample of geographical areas, listing the households or dwelling units in the sampled areas, selecting a probability sample of households, and selecting either all or a probability sample of persons in those households. Even when the sample of areas provides complete geographical coverage, noncoverage arises often from incomplete listing of households or dwelling units within sampled areas, and from incomplete listing of persons within sampled households. Nonresponse occurs when a sampled element fails to provide acceptable responses to some or all the survey questions. In the early years of probability sampling, response rates were high, and these two sources of imperfection were treated as minor blemishes that received little attention. They were either ignored or treated by simple weighting adjustments (simple, in part because more complex adjustments were computationally infeasible at the time).

Probability sampling has two main drawbacks to be balanced against the theoretical attractions of design-based inference: cost and timeliness. The extra costs of probability sampling include the costs of tracking down sampled individuals, including repeat calls when the individual is not initially available. When area sampling is used, the sampling costs also include the costs of listing units within sampled areas. For similar reasons, collecting survey data from a probability sample takes longer, making the production of the survey estimates less timely. Timeliness is important for all surveys, but particularly for surveys where the results are highly time-dependent, such as political polls, surveys of outbreaks of certain infections, and surveys of areas that have experienced a recent disaster.

A variety of less rigorous sampling methods are used in an attempt to apply a probability sampling approach to address these drawbacks. However, since all these methods require modeling assumptions, none of them can be classified as probability sampling. For convenience, they are called ‘pseudo-probability’ methods in what follows. In the early days of design-based inference, the quasi-probability sampling method known as quota sampling was widely used in market research and in other applications. That method is described in Section 4. Three other quasi-probability sampling methods are described briefly in Section 5.

#### **4. Quota sampling**

To set the scene for the need for imposing quota controls on a sample of the general population, consider the infamous Literary Digest Poll of 1936. To forecast the outcome of the 1936 U.S. Presidential Election, the Literacy Digest mailed a questionnaire to a sample of ten million individuals selected from telephone directories, lists of automobile owners, and registered voters. The results obtained from the two mil-



lion respondents indicated a clear-cut victory for Alf Landon with 57 percent of the vote, whereas in fact Franklin Roosevelt won with 61 percent of the vote. The upper-class bias of the sample, and of the respondents within the sample, is a major part of the explanation of the discrepancy between these percentages. No weighting adjustments were employed to attempt to address the bias at the time. (Lohr and Brick, 2017, reweighted the sample using respondents' reports of their voting in the 1932 election, and these adjustments led to a correct prediction of the outcome, but the estimate of the vote for Roosevelt still fell far short of the actual vote.) This study serves to demonstrate that a large sample size does not necessarily yield good estimates. See Converse (2017) for more details.

Market researchers and pollsters developed the methods of quota sampling separately from the developments in probability sampling, with the aim of addressing the biases from uncontrolled sampling. There are various forms of quota sampling, with the essence of all of them being to control the types of persons to be interviewed. Interviewers are instructed to make their samples of respondents conform to specified quota controls by such characteristics as sex, age group, and employment status. The controls could be independent (e.g., so many men and so many women, so many persons over 35 and so many persons 35 years of age or less) or the numbers to be interviewed could be interrelated (e.g., so many men over 35, so many women over 35). Sudman (1966) describes a method of quota sampling for national face-to-face interview surveys that he termed "probability sampling with quotas". He employed the four quota control groups of men under 35, men 35 and older, employed women and unemployed women, with the control groups chosen to give appropriate representation to young men and employed women. See also Stephenson (1979). The interviewing field force would generally be distributed across the country in a balanced way, either in areas selected to be representative, along the lines employed by Kiær, or in areas selected by a probability sample design. Sometimes additional controls are imposed, for example specifying the routes the interviewers were to follow, with no more than one person sampled in any household. Quota controls can also be applied in telephone surveys, mall intercept surveys, internet surveys (see Section 6), and other types of survey.

Quota sampling has two main advantages over probability sampling: cost and timeliness. Quota sampling is less costly because interviewers do not need to chase up elusive sampled units and because it avoids the costs of sampling specific households or persons (often including the associated listing costs). For the same reasons, a quota sample can be speedily fielded, and the data collected more rapidly than with a probability sample.

Quota sampling is a form of nonprobability sampling that assumes that the respondents in a quota group are an equal probability sample of the population in that

group. Note that this assumption also assumes that nonrespondents in the group are missing at random; nonresponse occurs with quota sampling, in essence with respondents substituted for the nonrespondents. Studies that have been conducted to evaluate quota sampling have found that the results are often similar to those produced by probability sampling, but this is not always the case (see Moser and Stuart, 1953, also Moser and Kalton, 1971; Stephan and McCarthy, 1958). For further references on quota sampling, see Kruskal and Mosteller (1980).

*Random Route Sampling.* Random route, or random walk, sampling is another quasi-probability sampling method that avoids the cost of, and associated time involved with, the listing operation. There are various versions of this method, but each starts with a random selection of a starting household and the interviewers then follow specified rules for walking patterns to follow and selection methods to use for serially identifying the subsequent households. The method has often been used in Europe and it is used in the Expanded Programme of Immunization (EPI) sampling method described in Section 5. Bauer (2014, 2016) discusses the selection errors that can occur with random route sampling and demonstrates that the method does not produce an equal probability sample, as its users generally assume.

## **5. Pseudo-probability sample designs for “hard-to-survey populations”**

Recent years have seen a major increase in the use of social survey methods to study the characteristics of “hard-to-survey populations” (Tourangeau, Edwards, Johnson, Wolter, Bates, 2014). Such populations are of various types, but all comprise only a small proportion of the general population and a population for which there is no separate sampling frame. This section presents three examples of sample designs for such populations. The first example is an inexpensive method that has been very widely used for vaccination surveys of the extremely rare population of 1-year-old children. The other two examples describe methods for sampling rare populations where membership of that population is a sensitive characteristic.

### *a. The EPI sampling method*

For almost 50 years, the World Health Organization’s Expanded Programme on Immunization (EPI) has used simple, inexpensive, sample designs in developing countries for measuring childhood immunization at the district level. Many thousands of EPI surveys have been conducted over this period, and the sample design has evolved over time. The sample design is a two-stage sample of clusters of communities (e.g., villages, towns, health service districts) that are sampled with outdated measures of estimated population sizes, with samples of eligible children selected within selected communities. The standard overall sample size is small, with the

selection of 30 clusters and 7 children in each cluster. The design is often known as  $30 \times 7$  design. Except in smaller communities, no household listings are made. Instead, the interviewer goes to the center of the village, chooses a random direction by spinning a bottle on the ground, and counts the number of households in that direction to the edge of the community. The interviewer then chooses a random number (for instance, from the numbers on a banknote) to identify the first sampled household. The second sampled household is then the one closest to the first, and so on, sequentially until survey data are collected on seven eligible children. Levy and Lemeshow (2008, pp. 427-428) describe the EPI sampling methods and Bennett (1993) describes some of the modifications to the original method.

The US Centers for Disease Control and Prevention (CDC) recommends a probability  $30 \times 7$  sample design for its rapid needs assessment tool, the Community Assessment for Public Health Emergency Response (CASPER) program. In this case, the clusters are generally census blocks with counts of households obtained from the U.S. Census Bureau or by using a GIS program for use in the PPES selection of thirty clusters. The fieldworker counts or estimates the number of households in a sampled cluster, divides that number by seven to give the sampling interval for systematic sampling, proceeds to select the sample from a random starting point, selecting subsequent households using a serpentine walking procedure. A crude weighting adjustment is proposed for use in the data analysis. Details are provided by CDC (2019).

#### *b. Venue-based sampling*

Venue-based sampling (also known as location sampling, time-space sampling, center sampling, and intercept sampling) is used for sampling members of a rare population at places that they frequent. It is applicable for rare populations that visit certain locations. It can be used to survey nomadic populations and for sampling hidden rare populations where the membership of that population is a sensitive matter. The method requires the construction of a frame of locations and a decision on the overall time period for the survey, selecting a sample of location/time periods for data collection, and selecting all or a sample of members of the survey population visiting each sampled location in the sampled data collection time period (Kalton, 1991). Two issues of concern arise when sampling hidden populations. One relates to the population coverage provided by the frame of locations and the overall time period: What proportion of the population will fail to visit any of the locations in that time period? Another issue relates to the multiplicity problem: How to account for the variability in the numbers of visits made to any of the locations by different sample members during the overall time period? These numbers are needed for use in weighting to compensate for unequal selection probabilities, but they are un-

known. At best, they can be estimated by asking respondents questions about their general frequencies of visiting the locations. See MacKellar, Gallagher, Findlayson, Lansky, and Sullivan (2007) for a description of the sampling methods used for surveying men who have sex with men (MSM) in a number of metropolitan areas in the United States.

*c. Respondent driven sampling*

Respondent driven sampling (RDS) is a form of link-trace sampling that selects the sample based on the social networks that exist for some populations. RDS has become a popular method for sampling rare hidden populations that have this feature, such as injection drug users and sex workers. The method starts by identifying a small set of members of the population of interest, who serve as *seeds* for the subsequent sample. The seeds respond to the survey, including responding to a question asking how many members of the survey population they know. They are then asked to recruit a set number of members of that population for the survey, the *alters*. The alters then go through the same process, recruiting further sample members. Under idealized circumstances, Heckathorn (1997) has shown that RDS produces a probability sample. However, the many conditions required for this to apply will not hold in practice (Gile and Hancock, 2010).

## **6. Internet surveys**

Recruiting the sample via the internet is a relatively recent approach for conducting social research. This approach has become extremely popular and has led to several alternative methods. See, for example, Baker, Blumberg, Brick *et al.* (2010) for a review of these methods. Surveys based on internet sampling have the great attractions of obtaining responses from large samples at low cost and high speed. However, their nonprobability sampling methods raise concerns about potential biases in the survey estimates. Those without, or with limited, access to the internet are excluded from these surveys and the survey respondents are clearly not a representative sample of the general population.

One form of internet sampling, known as river sampling, attaches invitations to participate in a survey on a number of internet sites, usually with offers of some form of compensation. The biases in the sample selection process make the representativeness of the sample highly questionable. Questions also need to be raised about the honesty and thoughtfulness of the responses.

Another form of internet sampling employs an opt-in internet panel. (An opt-in internet panel is distinct from an internet panel that selects a household panel by probability sampling and then conducts many data collections from the panel over time, albeit typically with low response rates). Extremely large numbers of people are

recruited for opt-in internet panels to be available to be approached to respond to surveys over time, sometimes as one of a range of services they may be asked to provide, in exchange for a payment for their services. The panel members can then be selected for invitation to respond to a given survey based on their responses to the screening instrument used in their recruitment.

In some ways, these large-scale nonprobability internet surveys bring to mind the abysmal results obtained from the 1936 Literacy Digest Poll referred to early. However, there are two major differences from the uncontrolled sample in the Digest Poll. One is the attempt to select a representative quota sample in design with internet panels. The other is the use of weighting adjustments in the analysis to achieve the same purpose. Before around 1970, lacking today's computers, complex calibration weighting adjustments were infeasible, but now advanced adjustment methods have been developed and are readily employed for both probability samples (particularly those with low response rates) and for nonprobability samples. With river sampling, a limited number of variables can be collected as part of the data collection for use in calibrating the sample to known or estimated population characteristics. The data collected in the screening instrument for an on-line panel can provide a much greater range of variables that can be used in sample selection and in the application of complex calibration adjustments to make the weighted sample correspond to a wide range of external controls. Nevertheless, serious doubts will persist about whether external data are available for the key auxiliary calibration variables at the population level or for a probability sample of that population, and whether the responses to the on-line survey can be treated as equal to the responses from the external source. Thus, for any given survey estimate, there must be concerns about how representative the nonprobability sample members are of the general population within the controls imposed in design or weighting. There will inevitably remain some residual biases of unknown magnitude and, with large samples, these biases can have a dominant influence on the level of accuracy of the survey estimates (Meng, 2018; Kalton, 2021, pp. 136–137).

## **7. Model-dependent inference**

In 1976, Fred Smith—my late friend and colleague at the University of Southampton at that time—wrote a paper reviewing the foundations of survey sampling in which he raised the question of why finite population inference should be so different from inference in the rest of statistics. His view at the time was that ‘survey statisticians should accept their responsibility for providing stochastic models for finite populations in the same way as statisticians in the experimental sciences’ (Smith, 1976); he moderated his position in a subsequent paper (Smith, 1994). Smith (1976) and papers by Brewer (1963), Royall (e.g., 1970, 1976) and others led to a spirited and

longstanding debate about the choice between design-based (model-assisted) inference and model-dependent (or model-based) inference. I was a discussant of Fred's 1976 paper and I subsequently published two papers on the role of models in survey sampling inference, with a defense of design-based inference in most circumstances applicable in large-scale social surveys (Kalton, 1983, 2002). However, models are needed to deal with the sampling imperfections of noncoverage and nonresponse, and they are needed for subgroup analyses in which the sample sizes are not adequate to provide design-based estimators of adequate precision. With the large decline in response rates that has occurred since the 1970's, it is no longer possible for survey statisticians to treat nonresponse as a minor blemish that can be brushed under the carpet in using design-based inference. I will return to this point later.

The model-dependent approach has led to the development of the prediction approach to survey inference. With this approach, an estimate of the population total  $Y$  is given by

$$\hat{Y}_m = \sum_{i \in s}^n y_i + \sum_{i \notin s}^N \hat{f}(X_i) \quad (1)$$

where the first summation is over the observed values in the sample  $s$  of size  $n$  and the second summation is over the model predictions of the  $y$  values for the nonsampled elements in the population. For comparison with the model-assisted design-based estimator  $\hat{Y}_d$  in (1), the model-dependent estimator may be expressed as  $\hat{Y}_m = \sum_s e_i + \sum_U \hat{f}(X_i)$ . In practice, greater care is used to develop the model for  $\hat{Y}_m$  than is the case in developing the working model for  $\hat{Y}_d$ . If the same model is used,  $\hat{Y}_m$  likely has lower variance than  $\hat{Y}_d$ . However,  $\hat{Y}_m$  has a design bias if the model is mis-specified, as is always the case to some extent, and the magnitude of the bias is unknown. The texts by Valliant, Dorfman, and Royall (2000) and Chambers and Clark (2012) describe the prediction approach in detail. The first chapter of Valliant et al. (2000) provides a useful review of design-based and model-based inference and includes further references. Note that the equation for  $\hat{Y}_m$  does not include selection probabilities (except possibly for estimating the model parameters) and does not require a probability sample. However, as Valliant, Dorfman, and Royall (2000, pp. 19–22) argue, randomization has the benefit of giving some protection against imbalance in factors uncontrolled in the design.

In my experience, until recently the prediction approach has had limited utility for large-scale social surveys of households and persons for the following reasons:

1. As distinct from surveys of establishments, there are generally little, if any, data available from the sampling frame about every member of the target population for use in the prediction models. Although some countries maintain up-to-date population registers that contain a selection of individual characteristics, in many

countries area sampling is used, with frame construction for individuals or households being performed only in selected areas. In these latter countries, no frame data is available for all members of the target population.

2. Social surveys are multipurpose in nature. They collect survey data on many variables, often numbering in the hundreds, and these data are analyzed in many ways, producing thousands of estimates. As a rule, these surveys are primarily conducted to produce descriptive estimates of parameters of the survey's finite population. These estimates need to be produced rapidly and to be consistent with each other. (These days, analytic estimates are also often produced, mostly through secondary analyses—see section 7).
3. A large proportion of the variables collected in social surveys are categorical in nature. They often cannot be as well predicted from auxiliary data as is the case with some of the continuous variables collected in business surveys.

However, even with large-scale social surveys, model-dependent estimation has a role to play in the production of descriptive estimates for small subclasses for which the sample sizes are too small to yield design-based estimates of adequate precision. This situation occurs particularly when the subclasses are geographical-defined administrative areas. The growth of interest by policy makers and others in separate estimates for administrative districts of all sizes has led to the development of the subject known as *small area estimation*. For many years, small area estimates, which are obtained using model-dependent prediction methods, were viewed with considerable skepticism by design-based statisticians but they have now become widely accepted in many fields of application. Ghosh (2020) gives a history of the development of small area estimation over five decades and Rao and Molina (2015) give a detailed description of this large and growing field.

The theoretical developments in model-based inference have now become increasingly relevant for social surveys to address the sampling imperfections and limitations with probability samples, and for the analyses of nonprobability samples; the use of nonprobability sampling for social research has grown rapidly in recent years, in particular for internet surveys.

## **8. Analytic uses of survey data**

As computing power and software came into widespread use in the 1970's, survey data collected using complex sample designs were used, mostly in secondary analyses, to produce analytic statistics that studied the relationships between variables, often looking for causal connections. Initially, multiple regression was the main form of analysis, with interest directed to the magnitude of the regression coefficients. Many analysts argued that their interest in the results of these analyses was

not for the specific finite population surveyed, but rather as estimates of super-population parameters of universal generality, and that, with the “correct” model, aspects of the sample design were irrelevant. From this perspective, probability sampling of the finite population becomes irrelevant and, unless survey weights and clustering were important as predictor variables, their inclusion in the analysis in a standard design-based way serves only to lower the precision of the estimated regression coefficients. The counter position was that no model is totally correct and that the estimation of the population regression coefficients, often termed census parameters, using the survey weights provides a safer approach. There is extensive literature on this topic. See, for example, DuMouchel and Duncan (1983).

Over time, the use of regression methods with survey data has been extended to include a wide range of regression models and other multivariate analysis techniques such as categorical data analysis, multilevel modeling, and longitudinal analyses. It is outside the scope of this paper to describe the application of these methods with complex survey data. See Skinner, Holt, and Smith (1989), Chambers and Skinner (2003). Applications of a range of multivariate methods with complex survey data are well described in the texts by Korn and Graubard (1999) and Heeringa, West, and Berglund (2017).

## **9. Administrative records and big data**

A great deal of attention has been paid recently to the use of administrative records as an alternative source of research data. There are obvious serious issues of privacy and confidentiality to be addressed when government-maintained administrative data are used in this way. For this reason, this approach is particularly suited to researchers in government agencies. The approach has notable potential attractions in terms of cost and sample size, but it needs to be recognized that it has its limitations. For instance, what is the coverage of the frame of the records, especially regarding program enrollment versus eligibility? Do the records contain the data needed to measure the concepts as the researcher would like to define them? Are the record data measured consistently across the population, or are there differences in the procedures used in different administrative areas? Are the data measured consistently over time to enable time series data to be validly analyzed? How might changes in program rules affect temporal comparisons? How long is the period between data collection and the researcher’s access to an analyzable dataset? Do the records contain the full set of variables needed for the analyses? In many cases, a single set of administrative records does not contain all the variables needed for the analyses. In this situation, it may be possible to link two or more sets of records, but record linkage problems need to be overcome and greater issues of confidentiality must be addressed.



How accurate are the data recorded in the records? Survey researchers have devoted a great deal of effort to training a relatively small number of interviewers to ask and record respondents' answers in a standard way. The situation is different with administrative records. Charlie Cannell, my late friend and colleague at the University of Michigan's Survey Research Center, had the following quotation from Josiah Stamp (1880–1941) in a plaque on his office wall:

“The government are very keen on amassing statistics. They collect them, add them, raise them to the  $n$ th power, take the cube root and prepare wonderful diagrams. But you must never forget that every one of these figures comes in the first instance from the village watchman, who just puts down what he damn pleases.”

While not claiming that current administrative records are as bad as this quotation might suggest, those who use such records for statistical purposes should carefully assess their quality and the distortions to which they may be subjected. See the paper by Hand (2018) and the ensuing discussion for a detailed discussion of the advantages and limitations of administrative records for research purposes.

In addition to government-maintained administrative records, there are other sources of social research data. In some cases, nongovernment records, such as those maintained by private organizations, may contain relevant information. However, they are subject to similar quality concerns, and access to the records may be hard to obtain. There are also sources of big data that occur on a flow basis, such as from linking cell phones to their GPS locations. The cell phone locations can be used to provide information about commuter times and even about long-distance travel trips if the home location is identified. Another source of big data is from scrapings on the web. Google Flu Trends (GFT) is a well-known and cautionary example. By analyzing extremely large numbers of flu-related searches on the web, Google developed models to predict local flu outbreaks in real time, avoiding the inevitable delay with other data sources. However, the models have since been found to fail (Lazer, Kennedy, King, and Vespignani, 2014), which serves as a warning that the apparent attraction of very big datasets can be illusory. For another example, see Bradley, Kuriwaki, Isakov, Sejdinovic, Meng, and Flaxman (2021).

## 10. Concluding remarks

As illustrated in previous sections, the choice between purposive selection and probability sampling was a subject of debate in the early period of survey research. It was not until after Neyman's (1934) paper that probability sampling and design-based inference were established as the gold standard for large-scale surveys conducted by national statistical offices. With a perfectly executed probability sample and no response error, the analyst has the security of being able to report the survey findings

as being subject only to a measurable degree of sampling error, whereas with nonprobability sampling the analyst can always be challenged that a purposive sample is not representative of the population with respect to the variables of analytic interest.

The preeminence of probability sampling for government surveys in the years from 1940 to, say, 2010 was not universal. There are costs incurred with probability sampling and a probability sample takes more time to draw and data collection takes longer. As illustrated in earlier sections, failures to devise probability sampling methods that can be applied with acceptable cost and timeliness for certain populations has given rise to the development of shortcut methods that depart in varying degrees from rigorous probability sampling.

In the early days, the idea of a “representative sample” was restricted to a sample that was representative in its design, as was the case with Kiær’s designs. The use of weighting adjustments in the analysis to achieve representativeness was seldom considered. The failure of the Literacy Digest poll in predicting the result of the U.S. Presidential election made clear that an extremely large unrepresentative sample could, without weighting adjustments, yield bad results.

Over the years, the implementation of probability sampling in social surveys has been increasingly challenged in many—but not all—countries by a steady decline in the willingness of the public to participate in surveys. Despite greater efforts to encourage response, response rates have declined dramatically in recent years. In reaction, greater efforts have been made to compensate for nonresponse, with major advances in the techniques employed. While replication methods of variance estimation can be applied to reflect the effect of the use of these techniques on the precision of the survey estimates, their use results in lower precision. Furthermore, the nonresponse adjustment model cannot be assumed to be “correct,” and the extent of any remaining nonresponse bias cannot be assessed. With its current heavy reliance on nonresponse models, in many countries probability sampling with design-based inference no longer retains its status as the undisputed gold standard. Moreover, the current levels of nonresponse have led to a marked increase in the costs of conducting a survey with probability sampling, both because of the increase in the initial sample size needed to produce the required sample size and because of the increased efforts to counteract nonresponse. For example, in the U.S. random digit dialing (RDD) was widely used with telephone surveying in the later part of the last century and the early part of this one because of the cost-efficiency of this modality (particularly for surveying rare populations). However, response rates for RDD surveys have plummeted to a level as low as 10 to 20 percent, largely ruling out this form of sampling.

With the security of model-free probability sampling with design-based inference now a thing of the past, model-dependent methods appear to be taking on a major role in social statistics. Research on making valid inferences from nonprobability samples is ongoing (see, for example, Valliant, 2020). Models are increasingly used to analyze data from a combination of data sources, including survey data from probability and nonprobability samples, administrative records, and other sources of big data. Thus, there is much research currently underway on making inferences from combinations of probability and nonprobability samples and from probability samples and other data sources (Kim and Wang, 2019; Beaumont and Rao, 2021; Rao, 2021),

In summary, after a long period in which probability sampling methods have dominated, the current situation is in a state of flux. New methods involving nonprobability sampling, internet sampling, administrative records, and big data are under constant modification and development. Brackstone (1999) lists six aspects of data quality for a statistical agency that remain applicable: relevance (how well the data meet the needs of the clients); accuracy (including both bias and variance); timeliness (time between the reference point and the time of data availability); interpretability (availability of relevant metadata); and coherence (ability to bring the data into a broader framework, including over time). The new data collection methods need to be assessed against these measures and, furthermore, the extensive research on response errors that has been conducted in the past now needs to be applied with the new methods of data collection. This is an exciting and challenging time for survey methodologists.

## References

- Aldrich, J., (2008). Professor A.L. Bowley's theory of the representative method. (Discussion Papers in Economics and Econometrics, 801), University of Southampton. <https://eprints.soton.ac.uk/150493>.
- Baker, R., Blumberg, S.J., Brick, J.M., Couper, M P., Courtright, M., Dennis, J.M., Dillman, D., Frankel, M.R., Garland, G., Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, 74(4), pp. 711–781. <https://doi.org/10.1093/poq/nfq048>.
- Bauer J.J., (2014). Selection errors of random route samples. *Sociological Methods and Research*, 43(3), pp. 519–544. <https://doi.org/10.1177/0049124114521150>.
- Bauer J.J., (2016). Biases in random route surveys. *Journal of Survey Statistics and Methodology*, 4(2), pp. 263–287. <https://doi.org/10.1093/jssam/smw012>.
- Beaumont J-F., Rao, J.N.K., (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, pp. 11–22. [http://isi-iass.org/home/wp-content/uploads/Survey\\_Statistician\\_2021\\_January\\_N83\\_02.pdf](http://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2021_January_N83_02.pdf).

- Bennett, S., (1993). Cluster sampling to assess immunization: a critical appraisal. *Bulletin of the International Statistical Institute, 49<sup>th</sup> Session*, 55(2), pp. 21–35.
- Bowley, A.L., (1913). Working-class households in Reading. *Journal of the Royal Statistical Society*, 76(7), pp. 672–701. <https://doi.org/10.1111/j.2397-2335.1913.tb03071.x>.
- Bradley, V.C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., Flaxman, S., (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, 600(7890), pp. 695–700. <https://doi.org/10.1038/s41586-021-04198-4>.
- Brewer, K.R.W., (1963). Ratio estimation in finite populations: some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5(3), pp. 93–105. <https://doi.org/10.1111/j.1467-842X.1963.tb00288.x>.
- Caradog Jones, D., (1949). *Social Surveys*. Hutchinson's University Library, London.
- CDC, (2019). Community Assessment for Public Health Emergency Response (CASPER) Toolkit. 3<sup>rd</sup> ed., CDC, Atlanta. <https://www.cdc.gov/nceh/casper/>.
- Chambers, R., Clark, R., (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford University Press, Oxford. <https://doi.org/10.1093/acprof:oso/9780198566625.001.0001>.
- Chambers, R.L., Skinner, C.J., Eds., (2003). *Analysis of Survey Data*. Wiley, Chichester. <https://doi.org/10.1002/0470867205>.
- Cochran, W.G., (1953). *Sampling Techniques*. Wiley, New York.
- Converse, J.M., (2017). *Survey Research in the United States: Roots and Emergence 1890–1960*. Routledge, New York. <https://doi.org/10.4324/9781315130491>.
- DuMouchel, W.H., Duncan, G.J., (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, 78(383), pp. 535–543. <https://doi.org/10.1080/01621459.1983.10478006>.
- Ghosh, M., (2020). Small area estimation: its evolution in five decades (with discussion). *Statistics in Transition*, 21(4), pp. 1–67. <https://doi.org/10.21307/stattrans-2020-022>.
- Gile, K.J., Hancock, M.S., (2010). Respondent-driven sampling: an assessment of current methodology. *Sociological Methodology*, 40(1), pp. 285–327. <https://doi.org/10.1111/j.1467-9531.2010.01223.x>.
- Gini, C., Galvani, L., (1929). Di una applicazione del metodo rappresentativo. *Annali di Statistica*, 6(4), pp. 1–107.
- Hand, D.J., (2018). Statistical challenges of administrative and transaction data (with discussion). *Journal of the Royal Statistical Society, A*, 181(3), pp. 555–605. <https://doi.org/10.1111/rssa.12315>.
- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory. Volume I: Methods and Applications. Volume II: Theory*. Wiley, New York
- Heckathorn, D.D., (1997). Respondent-driven sampling: a new approach to the study of hidden populations. *Social Problems*, 44(2), pp. 174–199. <https://doi.org/10.2307/3096941>.
- Heeringa, S.G., West, B.T., Berglund, P.A., (2017). *Applied Survey Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. <https://doi.org/10.1201/9781315153278>.
- Jensen, A., (1926) The report on the representative method in statistics. *Bulletin of the International Statistical Institute*, 22, pp. 355–376. <https://gallica.bnf.fr/ark:/12148/bpt6k61602s/f13.item.r=The%20report%20on%20the%20representative%20method%20in%20statistics>.

- Kalton, G., (1983). Models in the practice of survey sampling. *International Statistical Review*, 51, pp. 175–188. <https://doi.org/10.2307/1402747>.
- Kalton, G., (1991). Sampling flows of mobile human populations. *Survey Methodology*, 17(2), pp. 183–194. <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1991002/article/14503-eng.pdf?st=n0qimAao>.
- Kalton, G., (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics*, 18, pp. 129–154. <https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/models-in-the-practice-of-survey-sampling-revisited.pdf>.
- Kalton, G., (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87 (S1), pp. S10–S30. <https://doi.org/10.1111/insr.12287>.
- Kalton, G., (2021). *Introduction to Survey Sampling*. 2<sup>nd</sup> ed. SAGE Publications, Thousand Oaks, California. <https://doi.org/10.4135/9781071909812>.
- Kiær, A.N., (1976). *The Representative Method of Statistical Surveys*. English translation, Statistisk Centralbyro, Oslo.
- Kim, J.K., Wang, Z., (2019). Sampling techniques for big data analysis. *International Statistical Review*, 87(S1), pp. S177–S191. <https://doi.org/10.1111/insr.12290>.
- Kish, L., (1995). The hundred years' war of survey sampling. *Statistics in Transition*, 2(5), pp.813–830.
- Korn, E.L., Graubard, B.I., (1999). *Analysis of Health Surveys*. Wiley, New York. <https://doi.org/10.1002/9781118032619>.
- Kruskal, W., Mosteller, F., (1980). Representative sampling, IV: The history of the concept in statistics, 1895–1939. *International Statistical Review*, 48(2), pp. 169–195. <https://doi.org/10.2307/1403151>.
- Lazer, D., Kennedy, R., King, G., Vespignani, A., (2014). The parable of Google flu: Traps in big data analysis. *Science*, 343(6176), pp. 1203–1205. <https://doi.org/10.1126/science.1248506>.
- Levy, P.S., Lemeshow, S., (2008). *Sampling of Populations. Methods and Applications*. 4<sup>th</sup> ed. Wiley, Hoboken, NJ. <http://dx.doi.org/10.1002/9780470374597>.
- Lie, E., (2002). The rise and fall of sampling surveys in Norway, 1875–1906. *Science in Context*, 15(3), pp. 385–409. <https://doi.org/10.1017/S0269889702000534>.
- Lohr, S.L., Brick, J.M., (2017). Roosevelt predicted to win: Revisiting the 1936 Literary Digest Poll. *Statistics, Politics, and Policy*, 8(1), pp. 65–84. <https://doi.org/10.1515/spp-2016-0006>.
- MacKellar, D.A., Gallagher, K.M., Findlayson, T., Sanchez, T., Lansky, A., Sullivan, P. S., (2007). Surveillance of HIV risk and prevention behaviors of men who have sex with men—a national application of venue-based, time-space sampling. *Public Health Reports*, 122(1), Supplement 1, pp. 39–47. <https://doi.org/10.1177/00333549071220S107>.
- Mahanalobis, P.C., (1946). Proceedings of a Meeting of the Royal Statistical Society Held on July 16th, 1946, Professor M. Greenwood, F.R.S., in the Chair. *Journal of the Royal Statistical Society*, 109(4), pp. 325–378. <https://doi.org/10.1111/j.2397-2335.1946.tb04685.x>.
- Meng, X-L., (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12(2), pp. 685–726. <https://doi.org/10.1214/18-AOAS1161SF>.
- Moser, C.A., Kalton, G., (1971). *Surveys Methods in Social Investigation*. 2<sup>nd</sup> ed. Heinemann, London.

- Moser, C.A., Stuart, A., (1953). An experimental study of quota sampling. *Journal of the Royal Statistical Society, A*, 116(4), pp. 349–405. <https://doi.org/10.2307/2343021>.
- Neyman, J., (1934). On two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), pp. 558–625. <https://doi.org/10.1111/j.2397-2335.1934.tb04184.x>.
- Rao, J.N.K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83(1), pp. 242–272. <https://doi.org/10.1007/s13571-020-00227-w>.
- Rao, J.N.K., Molina, I., (2015). *Small Area Estimation*. 2<sup>nd</sup> ed. Wiley, Hoboken, N. J. <https://doi.org/10.1002/9781118735855>.
- Royall, R.M., (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57(2), pp. 377–387. <https://doi.org/10.1093/biomet/57.2.377>.
- Royall, R.M., (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71(355), pp. 657–664. <https://doi.org/10.1080/01621459.1976.10481542>.
- Särndal, C.E., Swensson, B., Wretman, J., (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York. <https://doi.org/10.1007/978-1-4612-4378-6>.
- Skinner, C.J., Holt, D., Smith, T.M.F., Eds., (1989). *Analysis of Complex Surveys*. Wiley, Chichester.
- Smith, T.M.F., (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society, A*, 139(2), pp. 183–204. <https://doi.org/10.2307/2345174>.
- Smith, T.M.F., (1994). Sample surveys 1975–1990; an age of reconciliation? *International Statistical Review*, 62(1), pp. 5–34. <https://doi.org/10.2307/1403539>.
- Stephan, F.F., (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association*, 43(241), pp. 12–39. <https://doi.org/10.1080/01621459.1948.10483247>.
- Stephan, F.F., McCarthy P.J., (1958). *Sampling Opinions. An Analysis of Survey Procedures*. Wiley, New
- Stephenson, C.B., (1979). Probability sampling with quotas: An experiment. *Public Opinion Quarterly*, 43(4), pp. 477–497. <https://doi.org/10.1086/268545>.
- Sudman, S., (1966). Probability sampling with quotas. *Journal of the American Statistical Association*, 61(315), pp. 749–771. <https://doi.org/10.1080/01621459.1966.10480903>.
- Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M., Bates, N., Eds., (2014). *Hard-to-Survey Populations*. Cambridge University Press, Cambridge, U. K. <https://doi.org/10.1017/CBO9781139381635>.
- Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263. <https://doi.org/10.1093/jssam/smz003>.
- Valliant, R., Dorfman, A.H., Royall, R.M., (2000). *Finite Population Sampling and Inference. A Prediction Approach*. Wiley, New York.
- Yates, F., (1949). *Sampling Methods for Censuses and Surveys*. Griffen, London.

## **Comments on *Probability vs. nonprobability sampling: from the birth of survey sampling to the present day* by Graham Kalton<sup>1</sup>**

I like to congratulate Professor Kalton for writing this very constructive article on probability versus nonprobability sampling. I learned a lot from reading it. In what follows, I add a few comments on this topic.

1. Professor Kalton emphasizes the issue of representative samples. In my view, probability samples and obviously nonprobability samples are practically never representative, even if balanced in advance on certain control (covariate) variables. A major reason for this is nonresponse, which might be “not missing at random” (NMAR), in which case the response probabilities depend on the target study variable, even after conditioning on known covariates. However, even in the case of simple random sampling and complete response, the actual sample may not be representative with respect to the unknown study variables, simply because of the randomness of the sample selection, unless the sample size is sufficiently large. Clearly, this problem worsens when sampling with unequal probabilities. Classical design-based theory overcomes this problem by restricting the inference to the randomization distribution over all possible sample selections. Thus, an estimator of a population mean is unbiased if its average over all possible samples that could have been drawn equals the true population mean, but in practice, we only have one sample. The use of models does not solve this problem either. A good model has to account for the sampling probabilities and the model assumed for the population values, and the inference need to account for both stochastic processes. As illustrated in many articles, ignoring the sampling process

---

<sup>a</sup> Department of Statistics, Hebrew University, Jerusalem, Israel & Southampton Statistical Sciences Research Institute, University of Southampton, UK. E-mail: msdanny@mail.huji.ac.il; msdanny@soton.ac.uk.  
ORCID: <https://orcid.org/0000-0001-7573-2829>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 23–25.  
<https://doi.org/10.59170/stattrans-2023-030>.

when fitting models to the sample data results with biased estimators of the model parameters in the case of informative sampling, by which the sampling probabilities are correlated with the outcome variables, again after conditioning on the model covariates. See, e.g. Pfeffermann and Sverchkov (1999) for empirical illustrations. In the case of NMAR nonresponse, the model has to account also for the unknown response probabilities.

2. The problem of nonresponse is indeed troubling and requires the use of models in the case of NMAR nonresponse, even in the case of design-based inference. The use of a response model enables to adjust the base sampling weights by the inverse of the estimated response probabilities, viewed as a second stage of the sampling process. I should say though that unlike a common perception, the response model can be tested, by testing the model of the study variable holding for the responding units, which accounts for the sampling design and the response. See, e.g. Pfeffermann and Sikov (2011).
3. Professor Kalton discusses the pros and cons of internet surveys “standing on their own”. I like to add that internet surveys are often used as one, out of several possible modes of response. For example, a questionnaire is sent to all the sampled units. It encourages them to respond via the internet. Those who do not respond are approached by telephone. When no response is obtained, an interviewer is sent for a face-to-face interview.
4. A well-known problem with this procedure is of mode effects; different estimates obtained from the respondents to the different modes, either because of differences between the characteristics of respondents responding with the different modes, (selection effect), or because of responding differently by the same sampled unit, depending on the mode of response (measurement effect). Several approaches to deal with this problem have been proposed in the literature. See, e.g. De Leeuw et al. (2018) for a comprehensive review.
5. My last 2 comments refer to inference from nonprobability samples:
6. Denote by  $S_{NP}$  the nonprobability sample. Rivers (2007) proposes to deal with the possible non-representativeness of  $S_{NP}$  by the use of sample matching. (Rivers considers a Web sample as the nonprobability sample but here I extend the idea to a more general nonprobability sample.) The approach consists of using a probability (reference) sample  $S_R$  from the target population, drawn with probabilities  $\pi_k = Pr(k \in S_R)$ , and matching to every unit  $i \in S_R$  an element  $k \in S_{NP}$ , based on known auxiliary (matching) variables  $\mathbf{x}$ . Denote by  $S_M$  the matched sample. Suppose that it is desired to estimate a population total of a study variable  $Y$ , based on measurements  $\{\tilde{y}_j, j \in S_{NP}\}$ . Estimate,  $\hat{Y}_T = \sum_{j \in S_M} w_j \tilde{y}_j$ ;  $w_j = (1/\pi_j)$ . Clearly, the base sampling weights can be modified to account for nonresponse.



7. This is an intriguing approach, but its success depends on the existence of a reference probability sample  $S_R$ , which allows sufficiently close matching, and ignorability of membership in the nonprobability sample  $S_{NP}$ , conditional upon the matching variables. I do not know whether this approach is used in practice, but I think that it deserves further investigation, with proper modifications.
8. The last two decades have witnessed the rapid growing of data science. One of the facets of this growth is that some people are agitating that the existence of all sorts of “big data” and the new advanced technologies that have been developed to handle these data, will soon replace the use of sample surveys. In an article I published in 2015, I overviewed some of the problems with the use of big data for the production of official statistics but clearly, when such data sources are available, accessible and timely, they cannot and should not be ignored. Big data can be viewed as a big, nonprobability sample, which for all kinds of reasons is not representative of the target population, and relying just on them can yield biased inference. Integrating big data with surveys is a major issue for research. See, e.g. Kim and Zhonglei (2018) and Rao (2021) for possible approaches, with references to other studies.

*I conclude my discussion by congratulating Statistics in Transition for its 30<sup>th</sup> anniversary and the publication of its 100<sup>th</sup> issue. This is one of the best journals of its kind and I wish it to continue prospering in the coming years.*

## References

- De Leeuw, E.D., Suzer-Gurtekin, Z. and Hox, J., (2018). The Design and Implementation of Mixed Mode Surveys. In *Advances in Comparative Survey Methods*. Wiley, New York. <https://doi.org/10.1002/9781118884997.ch18>.
- Kim, J.K. and Zhonglei Wang, (2019). Sampling Techniques for Big Data. *International Statistical Review*, 87(S1), pp. 177–191. <https://doi.org/10.1111/insr.12290>.
- Pfeffermann, D., (2015). Methodological Issues and Challenges in the Production of Official Statistics. *The Journal of Survey Statistics and Methodology (JSSAM)*, 3(4), pp. 425–483. <https://doi.org/10.1093/jssam/smv035>.
- Pfeffermann, D. and Sverckov, M., (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya*, 61, pp. 166–186.
- Pfeffermann, D. and Sikov, A., (2011). Imputation and Estimation under Nonignorable Non-response in Household Surveys with Missing Covariate Information. *Journal of Official Statistics*, 27, pp. 181–209. <https://www.scb.se/contentassets/ff271eeeca694f47ae99b942de61df83/imputation-and-estimation-under-nonignorable-nonresponse-in-household-surveys-with-missing-covariate-information.pdf>.
- Rao, J.N.K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, 83, pp. 242–272.
- Rivers, D., (2007). *Sampling for Web Surveys*. Joint Statistical Meetings, Proceedings of the Section on Survey Research Methods.

## **Comments on *Probability vs. nonprobability sampling: from the birth of survey sampling to the present day* by Graham Kalton<sup>1</sup>**

I would like to congratulate Professor Graham Kalton for his significant and inspiring article entitled as “Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day”. The article provides an elegant overview of the history of survey sampling, covering the purposive approaches that dominated the sampling field in the early days but from the 1940s, at least in official statistics, were gradually replaced entirely by probability-based approaches. Today we may be facing a paradigm shift again, but the direction is the opposite. Non-probability-based approaches are becoming viable, if not the only option, in fields that are moving towards big data and other new data sources and new methodological approaches.

The country’s data infrastructure forms the basis of official statistics and opens up for me an important perspective on Kalton’s presentation. Both probability and non-probability sampling and inference can benefit from statistical data infrastructures that contain a rich selection of micro-level covariates drawn from a variety of administrative and other registers. Perhaps the best options are in countries where population data from register sources and sample data are linked for combined microlevel databases. However, the utility of model-based (prediction) approaches for large-scale social surveys of households and persons will be limited if unit-level data for population members is missing from the sampling frames, as pointed out by Prof. Kalton. This is an important point and I think it can be extended to design-based model-assisted approaches that use mixed models in particular.

Countries differ much in terms of infrastructures based on administrative data. For example, Constance Citro calls for a move to multiple data sources that

---

<sup>a</sup> University of Helsinki, Finland. E-mail: risto.lehtonen@helsinki.fi.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 27–30. <https://doi.org/10.59170/stattrans-2023-031>.

include administrative records and, increasingly, transaction and Internet-based data (Citro 2014). Eric Rancourt argues that Statistics Canada is facing the new data world by modernizing itself and embracing an admin-first (in the broadest sense) paradigm as a statistical paradigm for the agency (Rancourt 2018). According to the United Nations Economic Commission for Europe (UNECE) report on register-based statistics in the Nordic countries, Central Population Registers of Denmark, Finland, Norway and Sweden were established in the sixties, and for example a totally register-based census was first implemented in Denmark (1981) and next in Finland (1990) (UNECE 2007). The number of national statistical institutes that have adopted or are developing administrative data infrastructures is increasing, as also described in the UNECE report on the use of registers and administrative data for population and housing censuses (UNECE 2018). This development can enhance the use of methods that utilize modeling and individual-level population frame data for model-assisted or prediction-based estimation with probability-based or non-probability-based sample data sets and their combinations.

The situation is different in countries that do not have similar high-quality population registers as for example in the Nordic countries. A recent contribution by Dunne and Zhang (2023) provides one important methodological approach for such countries. The authors present an innovative system (the PECADO application) for population estimates compiled from administrative data only.

Today, in the Nordic countries, as Finland, a majority of official statistics are based on administrative register combinations. In Finland, official statistics are produced by 13 expert organisations in the field of public administration and is coordinated by Statistics Finland. Probability samples are mainly used for regular social surveys such as labour force surveys and special surveys, e.g. Time Use survey. In these surveys, the sample elements can be uniquely linked with the elements in the register databases that often contain a lot of important background data including demographic, regional, socio-economic, income, educational, labour force status, and other variables. Thus these data need not to be collected by direct data collection methods from the respondents, and measurement errors are avoided. In addition, these variables are also used for calibration and model-assisted estimation procedures.

As an example, let me describe briefly the sampling and estimation design of the Labour Force Survey (LFS) of Finland. According to the quality description, in most European countries the LFS is based on a sample of households, and all members of a sample household living at the same address are interviewed. Finland is one of the Nordic countries where LFS is based on sampling of individual persons. The sample of about 12,500 persons is drawn by stratified probability sampling from Statistics Finland's population database, which is based on the Central Population Register.

Auxiliary information from registers include gender, age, region and language and selected register variables on employment, completed education and degrees, and income from the Employment Service Statistics of the Ministry of Economic Affairs and Employment, Statistics Finland's Register of Completed Education and Degrees, and the Tax Administration's Incomes Register (Quality Description: Labour Force Survey, Statistics Finland 2022). Sample data are linked to data from the registry using unique ID keys that exist across all data sources and are used in estimation procedures, including nonresponse adjustments. My experience is that this type of data infrastructure can also provide an excellent sampling and auxiliary data platform for e.g. methodological research in survey statistics; see for example Lehtonen, Särndal and Veijanen (2003, 2005).

Data infrastructures based on integrated administrative and other registers should be based on appropriate statistical theory and methodology for quality assessment and control and quality improvement. Recent sources in the field are for example Zhang (2012), Zhang and Haraldsen (2022) and the book on register-based statistics by Anders Wallgren and Britt Wallgren (2014). Research in statistical data integration and data science methods relevant for official statistics also is extending. A recent source is Yang and Kim (2020).

Experiences show that data infrastructures for official statistic containing a wealth of micro-level information on the population and an option for integration of the various register and sample data sources provide a flexible and efficient framework for survey estimation with probability-based samples. For non-probability samples, the variables of interest are typically in the non-probability data source. Most current methods for valid inference require an auxiliary data source containing the same covariates as the non-probability sample. These data can be obtained from the statistical population register or, more commonly, from a probability sample from it (e.g. Kim, Park, Chen and Wu 2021; Wu 2022). It can be foreseen that although the golden age of probability sampling may be over, probability sampling and non-probability sampling are not in conflict, but can complement each other.

## References

- Citro, C. F., (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2), pp. 137–161. <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14128-eng.pdf?st=wp6JZiua>.
- Dunne, J. and Zhang, L.-C., (2023). A system of population estimates compiled from administrative data only. *Journal of the Royal Statistical Society Series A: Statistics in Society*. <https://doi.org/10.1093/jrsssa/qnad065>.
- Kim, J.-K., Park, S., Chen, Y. and Wu, C., (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(3), pp. 941–963. <https://doi.org/10.1111/rssa.12696>.

- Lehtonen, R., Särndal, C.-E. and Veijanen, A., (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology*, 29(1), pp. 33–44. [https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6605-eng.pdf?st=Cf8qNa\\_2](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6605-eng.pdf?st=Cf8qNa_2).
- Lehtonen, R., Särndal, C.-E. and Veijanen, A., (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition*, 7(3), pp. 649–673. [https://stat.gov.pl/cps/rde/xbcr/pts/PTS\\_sit\\_7\\_3.pdf](https://stat.gov.pl/cps/rde/xbcr/pts/PTS_sit_7_3.pdf).
- Quality Description: Labour force survey, Statistics Finland 2022, (2022). [https://www.tilastokeskus.fi/til/tyti/2022/01/tyti\\_2022\\_01\\_2022-02-22\\_laa\\_001\\_en.html](https://www.tilastokeskus.fi/til/tyti/2022/01/tyti_2022_01_2022-02-22_laa_001_en.html)
- Rancourt, E., (2018). *Admin-First as a statistical paradigm for Canadian official statistics: Meaning, challenges and opportunities*. Proceedings of Statistics Canada Symposium 2018. <https://www.statcan.gc.ca/en/conferences/symposium2018/program>.
- United Nations Economic Commission for Europe, (2007). *Register-based statistics in the Nordic countries: Review of best practices with focus on population and social statistics*. United Nations, New York. <https://digitallibrary.un.org/record/609979?ln=en>
- UNECE, (2018). *Guidelines on the use of registers and administrative data for population and housing censuses*. United Nations, New York and Geneva. <https://unece.org/guidelines-use-registers-and-administrative-data-population-and-housing-censuses-0>
- Yang, S. and Kim, J.K., (2020). Statistical data integration in survey sampling: a review. *Japanese Journal of Statistics and Data Science Publishing model*, 3(2), pp. 625–650. <https://doi.org/10.1007/s42081-020-00093-w>.
- Zhang, L.-C., (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1), pp. 41–63. <https://doi.org/10.1111/j.1467-9574.2011.00508.x>.
- Zhang, L.-C. and Haraldsen, G., (2022). Secure big data collection and processing: framework, means and opportunities. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(4), pp. 1541–1559.
- Wallgren, A. and Wallgren, B., (2014). *Register-Based Statistics: Administrative Data for Statistical Purposes*. Second edition. Wiley.
- Wu, C., (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, 48(2), pp. 283–311. <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00002-eng.pdf?st=otYKlz2q>.

## **Discussion of *Probability vs. nonprobability sampling: from the birth of survey sampling to the present day* by Graham Kalton<sup>1</sup>**

In this excellent overview of the history of probability and nonprobability sampling from the end of the nineteenth century to the present day, Professor Graham Kalton outlines the essence of past endeavors that helped to define philosophical approaches and stimulate the development of survey sampling methodologies. From the beginning, there was an understanding that a sample should, in some ways, resemble the population under study. In Kiær's ideas of "representative sampling" and Neyman's invention of probability-based approach, the prime concern of survey sampling has been to properly plan for representing characteristics of the finite population. Post-stratification and other calibration methods were developed for the same important goal of better representation.

Professor Kalton's paper underscores growing interest in the use of nonprobability surveys. With recent proliferation of computers and the internet, wealth of data becomes available to researchers. However, "opportunistic" information collected with present-day capabilities usually is not purposely planned or controlled by survey statisticians. No matter how big such a nonprobability sample could be, it may inaccurately reflect the finite population of interest, thus presenting a substantial risk of an estimation bias.

Below, we discuss several recent papers that propose ways to incorporate non-probability surveys to produce estimates for both large and small areas. Specifically, we will consider two situations often encountered in practice. In the first situation,

---

<sup>a</sup> 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>.

<sup>b</sup> University of Maryland, College Park, MD 20742. USA. E-mail: plahiri@umd.edu.

ORCID: <https://orcid.org/0000-0002-7103-545X>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 31–37.

<https://doi.org/10.59170/stattrans-2023-032>.

© Julie Gershunskaya, Partha Lahiri. Article available under the CC BY-SA 4.0 licence 

a nonprobability sample contains the outcome variable of interest, and the main task is to reduce the selection bias with the help of a reference probability sample that does not contain the outcome variable of interest. In the second situation, a probability sample contains the outcome variable of interest, but there is little or no sample available to produce granular level estimates. For such a small area estimation problem, we consider a case when we have access to a large nonprobability sample that does not contain the outcome variable but contains some related auxiliary variables also present in the probability sample. In both situations, researchers have discussed statistical data integration techniques in which a reference probability sample is combined with a nonprobability sample in an effort to overcome deficiencies associated with both probability and nonprobability samples.

One way to account for the selection bias of a nonprobability sample is by estimating the sample inclusion probabilities, given available covariates. Then, the inverse values of estimated inclusion probabilities are used, in a similar manner as the usual probability sample selection weights, to obtain estimates of target quantities. Several approaches to estimation of nonprobability sample inclusion probabilities (or propensity scores) have been considered in the literature. Recent papers by Chen et al. (2020), Wang et al. (2021), and Savitsky et al. (2022) propose ways to estimate these probabilities based on combining nonprobability and probability samples. Kim J. and K. Morikawa (2023) propose an empirical likelihood based approach under a different setting. To save space, we will not discuss their approach. We now review three statistical data integration methods.

The approaches concern with the estimation of probabilities  $\pi_{ci}(x_i) = P\{c_i = 1|x_i\}$  to be included into the nonprobability sample  $S_c$ , for units  $i = 1, \dots, n_c$ , where  $c_i$  is the inclusion indicator of unit  $i$  taking on the value of 1 if unit  $i$  is included into the nonprobability sample, and 0 otherwise;  $x_i$  is a vector of known covariates for unit  $i$ ;  $n_c$  is the total number of units in sample  $S_c$ . The problem, of course, is that we cannot estimate  $\pi_{ci}$  based on the set of units in nonprobability sample  $S_c$  alone, because  $c_i = 1$  for all  $i$  in  $S_c$ . The probabilities are estimated by combining set  $S_c$  with a probability sample  $S_r$ . Due to its role in this approach, the probability sample here is also called “reference sample”.

Assuming both nonprobability and probability samples are selected from the same finite population  $P$ , Chen et al. (2020) write a log-likelihood, over units in  $P$ , for the Bernoulli variable  $c_i$ :

$$\ell_1(\boldsymbol{\theta}) = \sum_{i \in S_c} \log \left[ \frac{\pi_{ci}(x_i, \boldsymbol{\theta})}{1 - \pi_{ci}(x_i, \boldsymbol{\theta})} \right] + \sum_{i \in P} \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})]. \quad (1)$$

where  $\boldsymbol{\theta}$  is the parameter vector in a logistic regression model for  $P$   $\pi_{ci}$ .

Since finite population units are not observed, Chen et al. (2020) employ a clever trick and re-group the sum in (1) by presenting it as a sum of two parts: part 1 involves the sum over the nonprobability sample units and part 2 is the sum over the whole finite population:

$$\ell_1(\boldsymbol{\theta}) = \sum_{i \in S_c} \log \left[ \frac{\pi_{ci}(x_i, \boldsymbol{\theta})}{1 - \pi_{ci}(x_i, \boldsymbol{\theta})} \right] + \sum_{i \in S_r} w_{ri} \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})], \quad (2)$$

Units in part 1 of the log-likelihood in (2) are observed; for part 2, Chen et al. (2020) employ the pseudo-likelihood approach by replacing the sum over the finite population with its probability sample based estimate:

$$\ell_1(\boldsymbol{\theta}) = \sum_{i \in P} \{c_i \log[\pi_{ci}(\boldsymbol{\theta})] + (1 - c_i) \log [1 - \pi_{ci}(x_i, \boldsymbol{\theta})]\}, \quad (3)$$

where weights  $w_{ri} = 1/\pi_{ri}$  are inverse values of the reference sample inclusion probabilities  $\pi_{ri}$ . Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

One shortcoming of the Chen et al. (2020) approach is that their Bernoulli likelihood is formulated with respect to an unobserved indicator variable. Although the regrouping employed in (2) helps to find a solution, results obtained by Wang et al. (2021) indicate that it is relatively inefficient, especially when the nonprobability sample size is much larger than the probability sample size.

Wang et al. (2021) formulate their likelihood for an *observed* indicator variable and thus their method is different from the approach of Chen et al. (2020). To elaborate, Wang et al. (2021) introduce an imaginary construct consisting of two parts: they *stack* together non- probability sample  $S_c$  (part 1) and finite population  $P$  (part 2). Since nonprobability sample units belong to the finite population, they appear in the stacked set twice. Let indicator variable  $\delta_i = 1$  if unit  $i$  belongs to part 1, and  $\delta_i = 0$  if  $i$  belongs to part 2 of the stacked set; the probabilities of being in part 1 of the stacked set are denoted by  $\pi_{\delta_i}(x_i) = P\{\delta_i = 1|x_i\}$ . Wang et al. (2021) assume the following Bernoulli likelihood for observed variable  $\delta_i$ :

$$\ell_2(\tilde{\boldsymbol{\theta}}) = \sum_{i \in S_c} \log [\pi_{\delta_i}(x_i, \tilde{\boldsymbol{\theta}})] + \sum_{i \in P} \log [1 - \pi_{\delta_i}(x_i, \tilde{\boldsymbol{\theta}})], \quad (4)$$

Where  $\tilde{\boldsymbol{\theta}}$  is the parameter vector in a logistic regression model for  $\pi_{\delta_i}$ . Since the finite population is not available, they apply the following pseudo-likelihood approach:



$$\hat{\varrho}_2(\tilde{\theta}) = \sum_{i \in S_c} \log [\pi_{\delta i}(x_i, \tilde{\theta})] + \sum_{i \in S_r} w_{ri} \log [1 - \pi_{\delta i}(x_i, \tilde{\theta})]. \quad (5)$$

Existing ready-to-use software can be used to obtain estimates of  $\pi_{\delta i}$ . However, the actual goal is to find probabilities  $\pi_{\delta i}$  rather than probabilities  $\pi_{\delta i}$ . Wang et al. (2021) propose a two-step approach, where at the second step, they find  $\pi_{\delta i}$  by employing the following identity:

$$\pi_{\delta i} = \frac{\pi_{ci}}{1 + \pi_{ci}} \quad (6)$$

Savitsky et al. (2022) use an exact likelihood for the estimation of inclusion probabilities  $\pi_{\delta i}$ , rather than a pseudo-likelihood based estimation. They propose to stack together nonprobability,  $S_c$ , and probability,  $S_r$ , samples. In this stacked set,  $S$ , indicator variable  $z_i$  takes the value of 1 if unit  $i$  belongs to the nonprobability sample (part 1), and 0 if unit  $i$  belongs to the probability sample (part 2). In this construction, if there is an overlap between the two samples,  $S_c$  and  $S_r$ , then the overlapping units are included into stacked set  $S$  twice: once as a part of the nonprobability sample (with  $z_i = 1$ ) and once as a part of the reference probability sample (with  $z_i = 0$ ). We do not need to know which units overlap or whether there are any overlapping units. The authors use first principles to prove the following relationship between probabilities  $\pi_{zi}(x_i) = P\{z_i = 1|x_i\}$  of being in part 1 of the stacked set and the sample inclusion probabilities  $\pi_{ci}$  and  $\pi_{ri}$ :

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ri} + \pi_{ci}} \quad (7)$$

A similar expression (7) was derived by Elliott (2009) and Elliott and Valliant (2017) under the assumption of non-overlapping nonprobability and probability samples. The derivation given in Savitsky et al. (2022) does not require this assumption.

To obtain estimates of  $\pi_{ci}$  from the combined sample, Beresovsky (2019) proposed to parameterize probabilities  $\pi_{ci} = \pi_{ci}(\pi_i, \theta)$ , as in Chen et al. (2020), and employ identity (7) to present  $\pi_{zi}$  as a composite function of  $\theta$ ; that is,  $\pi_{zi} = \pi_{zi}(\pi_{ci}(x_i, \theta)) = \pi_{ci}(x_i, \theta) / (\pi_{ri} + \pi_{ci}(x_i, \theta))$ .

The log-likelihood for observed Bernoulli variable  $z_i$  is given by

$$\ell_3(\theta) = \sum_{i \in S_c} \log [\pi_{zi}(\pi_{ci}(x_i, \theta))] + \sum_{i \in S_r} \log [1 - \pi_{zi}(\pi_{ci}(x_i, \theta))]. \quad (8)$$

Since the log-likelihood *implicitly* includes a logistic regression model formulation for probabilities  $\pi_{ci}$ , Beresovsky (2019) labeled the proposed approach Implicit Logistic Regression (ILR). For the maximum likelihood estimation (MLE), the score equations are obtained from (8) by taking the derivatives, with respect to  $\theta$ , of the composite function  $\pi_{zi} = \pi_{zi}(\pi_{ci}(\theta))$ . This way, the estimates of  $\pi_{ci}$  are obtained directly from (8) in a single step. Savitsky et al. (2022) parameterized the likelihood, as in (8), and used the Bayesian estimation technique to fit the model.

Note that to implement the ILR approach, the reference sample inclusion probabilities  $\pi_{ri}$  have to be known for all units in the combined set. This is not a limitation for many probability surveys. As discussed in Elliott and Valliant (2017), if probabilities  $\pi_{ri}$  cannot be determined exactly for units in the nonprobability sample, they can be estimated using a regression model. Savitsky et al. (2022) used Bayesian computations to simultaneously estimate  $\pi_{ri}$  and  $\pi_{ci}$  for nonprobability sample units, given available covariates  $x_i$ .

It must be noted that the estimation method of Wang et al. (2021) can be similarly modified to avoid the two-step estimation procedure: a logistic regression model could be formulated for inclusion probabilities  $\pi_{ci}$ , while probabilities  $\pi_{si}$  in (6) could be viewed as a composite function,  $\pi_{\delta i} = \pi_{\delta i}(\pi_{ci}(x_i, \theta)) = \pi_{ci}(x_i, \theta) / (1 + \pi_{ci}(x_i, \theta))$ . This approach is expected to be more efficient. Moreover, it avoids  $\pi_{ci}$  estimates greater than 1 that could occur when the estimation is performed in two steps. Once modified this way, preliminary simulations indicate that Wang et al. (2021) formulation would produce more efficient estimates than the Chen et al. (2020) counterpart, unless in a rare situation where the whole finite population rather than only a reference sample is available.

Simulations show that the exact likelihood method based on formulation of Savitsky et al. (2022) and Beresovsky (2019) performs better than the pseudo-likelihood based alternatives. In the usual situation where the reference probability sample fraction is small, the relative benefits of the exact likelihood approach are even more pronounced.

The existence of a well-designed probability reference sample plays a crucial role in the efforts to reduce the selection bias of a nonprobability sample. Importantly, an ongoing research indicates that the quality of estimates of the nonprobability sample inclusion probabilities is better if there is a good overlap in domains constructed using covariates from both samples. This observation harks back to problems appearing in traditional poststratification methods and to the notion of “representative sampling”. Since survey practitioners usually do not have control over the planning or collection of the emerging multitude of nonrandom opportunistic samples, efforts should be directed to developing and maintaining comprehensive probability samples that include sets of good quality covariates. Beaumont et al. (2023) proposed several model

selection methods in application of the modeling nonprobability sample inclusion probabilities.

We now turn our attention to the second data integration situation involving small area estimation, a topic Professor Kalton touched on. This is a problem of great interest for making public policies, fund allocation and regional planning. Small area estimation programs already exist in some national statistical organizations such as the Small Area Income and Poverty Estimates (SAIPE) program of the US Census Bureau (Bell et al., 2016) and Chilean government system (Casas-Cordero Valencia et al., 2016.) The importance placed in the United Nations Sustainable Development Goals (SDG) for disaggregated level statistics is expected to increase the demand for such programs in various national statistical offices worldwide. Standard small area estimation methods generally use statistical models (e.g., mixed models) that combine probability sample data with administrative or census data containing auxiliary variables correlated with the outcome variable of interest. For a review of different small area models and methods, see Jiang and Lahiri (2006), Rao and Molina (2015), Ghosh (2020), and others.

A key to success in small area estimation is to find relevant auxiliary variables not only in the probability sample survey but also in the supplementary big databases. Use of a big probability or nonprobability sample survey could be useful here as surveys typically contain a large number of auxiliary variables that are also available in the probability sample survey. In the context of small area estimation, Sen and Lahiri (2023) considered a statistical data integration technique in which a small probability survey containing the outcome variable of interest is statistically linked with a much bigger probability sample, which does not contain the outcome variable but contains many auxiliary variables also present in the smaller sample. They essentially fitted a mixed model to the smaller probability sample that connects the outcome variable to a set of auxiliary variables and then imputed the outcome variable for all units of the bigger probability sample using the fitted model and auxiliary variables. Finally, they suggested to produce small area estimates using survey weights and imputed values of the outcome variable contained in the bigger probability sample survey. As discussed in their paper, such a method can be used even if the bigger sample is a nonprobability survey using weights constructed by methods such as the ones described earlier.

The development of new approaches demonstrates how the methods of survey estimation continue to evolve by taking into the future the best from fruitful theoretical and methodological developments of the past. As Professor Kalton highlights, we will increasingly encounter data sources that are not produced by standard probability sample designs. Statisticians will find ways to respond to new challenges, as is reflected in the following amusing quote:

...D.J. Finney once wrote about the statistician whose client comes in and says, “Here is my mountain of trash. Find the gems that lie therein”. Finney’s advice was to not throw him out of the office but to attempt to find out what he considers “gems”. After all, if the trained statistician does not help, he will find someone who will. (source: David Salsburg, ASA Connect Discussion)

Of course, nonprobability samples should not be viewed as a “mountain of trash”. Indeed, they can contain a lot of relevant information for producing necessary estimates. It is just that one needs to explore different innovative ways to use information contained in nonprobability samples. In the United States federal statistical system, the need to innovate for combining information from multiple sources has been emphasized in the National Academies of Sciences and Medicine (2017) report on Innovations in Federal Statistics. As discussed, statisticians have been already engaged in suggesting new ideas, such as statistical data integration, to extract information out of multiple non-traditional databases. In coming years, statisticians will be increasingly occupied with finding solutions for obtaining useful information from non-traditional data sources. This is indeed an exciting time for survey statisticians.

## References

- Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K., (2023). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada’s crowdsourcing data. *Survey Methodology* (accepted in 2023 and expected to appear in 2024).
- Bell, W. R., Basel, W. W., and Maples, J. J., (2016). *An overview of the US Census Bureau’s small area income and poverty estimates program*, pp. 349–378. Wiley Online Library.
- Beresovsky, V., (2019). On application of a response propensity model to estimation from web samples. In ResearchGate.
- Casas-Cordero Valencia, Encina, C., J., and Lahiri, P., (2016). *Poverty mapping for the Chilean Comunas*, pp. 379–404. Wiley Online Library. <https://doi.org/10.1002/9781118814963.ch20>.
- Chen, Y., Li, P., and Wu, C., (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532), pp. 2011–2021. <https://doi.org/10.1080/01621459.2019.1677241>.
- Elliott, M. R., (2009). Combining data from probability and non-probability samples using pseudo weights. *Survey Practice*, 2(6), pp. 813–845. <https://doi.org/10.29115/SP-2009-0025>.
- Elliott, M. R., Valliant, R., (2017). Inference for Nonprobability Samples. *Statistical Science*, 32(2), pp. 249–264. <https://doi.org/10.1214/16-STS598>.
- Ghosh, M., (2020). Small area estimation: Its evolution in five decades. *Statistics in Transition new series, Special Issue on Statistical Data Integration*, 21(4), pp. 1–67. <https://doi.org/10.21307/stattrans-2020-022>.
- Jiang, J., Lahiri, P., (2006). Mixed model prediction and small area estimation, editor’s invited discussion paper. *Test*, 15(1), pp. 1–96.

- Kim J., Morikawa, K., (2023). An empirical likelihood approach to reduce selection bias in voluntary samples. *Calcutta Statistical Association Bulletin*, 75(1). <https://doi.org/10.1177/00080683231186488>.
- National Academies of Sciences, E. 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>.
- Rao, J. N. K., Molina I., (2015). *Small Area Estimation, 2nd Edition*. Wiley. <https://doi.org/10.1002/9781118735855>.
- Savitsky, T. D., Williams, M. R., Gershunskaya, J., Beresovsky, V., and Johnson, N. G., (2022). *Methods for combining probability and nonprobability samples under unknown overlaps*. <https://doi.org/10.48550/arXiv.2208.14541>.
- Sen, A., Lahiri P., (2023). *Estimation of finite population proportions for small areas: a statistical data integration approach*. <https://doi.org/10.48550/arXiv.2305.12336>.
- Wang, L., Valliant, R., and Li, Y., (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(4), pp. 5237–5250. <https://doi.org/10.1002/sim.9122>.

## **Discussion of *Probability vs. nonprobability sampling: from the birth of survey sampling to the present day* by Graham Kalton<sup>1</sup>**

Let me first thank Dr. Kalton for his amazing historical review of the development of survey sampling from its origin, contrasting purposive sampling, until now, where some elements of purposive sampling in terms of web or big data seem to supersede the well-elaborated theory of survey statistics. Shall the message be that we do not need any sampling courses at universities anymore, that official statistics should turn to modelling using data with unknown data generating processes, or actually even be substituted by (commercial) *data krakens*? Hardly so! Graham Kalton emphasises a modern thinking about the use of these new data sources which may also have some advantages and he urges future research on data integration methods using (very) different kinds of data while strongly taking quality aspects into account.

Within the last decade, we could observe many new uses of classical data like administrative data and new types of data stemming from internet sources or technical measurement processes such as satellite, mobile phone or scanner data. Already the availability of these new data leads to a huge increase in developing new methodologies and uses. Indeed, official statistics also forced research on new data types, such as scanner data or web-scraped data and others. In Europe, these statistics are often called experimental statistics to emphasise that these statistics cannot (yet) be evaluated using the classical quality concepts, as, e.g. proposed within the European Statistics Code of Practice (<https://ec.europa.eu/eurostat/web/quality/european-quality-standards/european-statistics-code-of-practice>). Some examples can be drawn from [https://www.destatis.de/EN/Service/EXDAT/\\_node.html](https://www.destatis.de/EN/Service/EXDAT/_node.html) or <https://ec.europa.eu/eurostat/web/experimental-statistics>.

---

<sup>a</sup> Department of Economics, Economic and Social Statistics, Trier University, Germany.  
E-mail: [muennich@uni-trier.de](mailto:muennich@uni-trier.de). ORCID: <https://orcid.org/0000-0001-8285-5667>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 39–41.  
<https://doi.org/10.59170/stattrans-2023-033>.

During the Covid crisis, and especially in light of the political discussion in Germany, however, one could observe little understanding of data quality and statistics. Timeliness – with its urge of getting data and producing statistics immediately – often lead to the use of available (infection) data, which certainly were influenced by unknown biases. The impact of statistics on these available data in terms of evidence-based policy could hardly be understood at the time, but still legal processes like lockdowns were initiated. To state this message more strongly: whenever a legislation process is involved, and especially so if a direct impact on society is the outcome, we must make sure that high quality requirements on data gathering and statistical methodology are set as well as met. High quality typically cannot be achieved with low costs. England was one of the few very good examples during the pandemic, since they were setting up a special Covid survey to better understand the pandemic and to provide adequate and reliable information.

Certainly, this example already shows some critical aspects in data gathering and data quality. Dr. Kalton was emphasising timeliness and accuracy as very important goals of data quality. For sure, these are of utmost importance! However, in practice, both quality principles suffer from budget constraints and cost controls. This directly leads to two questions: Do modern data help to provide more timely and accurate statistics at lower costs? Is there, in case of conflicts, an *ultimate* quality principle?

The first question is already answered by Dr. Kalton. Of course, modern web or big data can help to gather information quickly. Interesting approaches are of course the use of satellite or scanner data. With electronic cash systems, price changes could be tracked much faster than via the use of survey data. However, one always has to understand the advantages as well as the disadvantages of these data generation processes, and one must be able to measure the quality of the output.

Let me briefly sketch one current German debate which, in my view, perfectly fits into this discussion. In the past years, more and more internet surveys were preferred to data from traditional market and opinion research. This immediately led to a discussion on the quality of the outcomes. And certainly, timeliness, accuracy, and costs played an important role within this discussion. The two major arguments were the following: internet surveys suffer from unknown biases. Classical surveys, in the meantime, have to consider response rates considerably below 20%. Under these conditions, most likely both areas have to consider statistical models with strong assumptions to at least reduce possible biases induced by either web surveys or non-response. In my view, one important question has not been raised yet. What is the aim of the survey?

The ultimate aim that necessitates data collection in the first place is of crucial importance for evaluating the importance of the different quality principles. In case one is interested in getting information on current public opinion, probably time-

liness and costs are more important than high accuracy. However, in evidence-based policy making, and especially when information for legislative action is needed, I must stress that accuracy must always be considered to be the major principle. This is even more important when large budgets or financial equalization schemes are involved. Additionally, in these cases one must also be able to measure the quality of the outcome of the statistics. This is still a major drawback of using web or big data. And to stress this point, in legislation processes, I strongly urge to involve independent official statistics with its transparent data production process.

With this discussion, I do not want to be misunderstood. Modern data and modern statistical methods are important. And the direction of research, as Dr. Kalton pointed out, will be complex modelling and data integration. Also administrative, register, and related data are important and can provide very good information. However, with all these data, we always have to understand their quality and we should be able to measure the quality of the resulting statistics. Especially in the context of big data, quality measurement may have to be enhanced (cf. Münnich and Articus, 2022, and the citations therein).

Sampling itself may also follow new directions. Classical sampling optimization may be adequately applied in more special cases that allow focusing on specific goals, e.g. the design optimization in the German Censuses 2011 and 2022 (see Münnich et al., 2012, and Burgard, Münnich, and Rupp, 2020). However, likely robustness of methods against assumptions has to be incorporated in design optimization. On the other hand, data integration, multi-source environments, geo-spatial modelling, small area estimation and other modern methods may yield new ideas and directions in sampling theory and application. One example may be sampling from big data sources to reduce complexity.

Despite the mentioned new directions, many ideas have been well-known for a long time. In data analytics, we differentiate between descriptive, predictive, and prescriptive aims. Data that were gathered to describe a state of a system cannot be used to analyse interventions on the system. Indeed, we need the right data and not just merely available data. In conclusion, the exact purpose of the statistics under consideration plays an extremely important role for the selection of data and the priority of the different quality principles.



## References

- Burgard, J. P., Münnich, R., & Rupp, M., (2020). Qualitätszielfunktionen für stark variierende Gemeindegrößen im Zensus 2021. *ASTA Wirtschafts-und Sozialstatistisches Archiv*, 14(1), pp. 5–65. With discussion. <https://doi.org/10.1007/s11943-019-00256-6>.
- Münnich, R., Articus, C., (2022). Big Data und Qualität – ist viel gleich gut? Pp 85–101. In Wawrzyniak, B., Herter, M. (Ed.), *Neue Dimensionen in Data Science*. Wichmann.
- Münnich, R., Gabler, S., Ganninger, M., Burgard, J. P., Kolb, J.-P., (2012). *Stichprobenoptimierung und Schätzung im Zensus 2011*. Destatis: Wiesbaden, Statistik und Wissenschaft, Vol. 21, [https://www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEMonografie\\_derivate\\_00000209/1030821129004.pdf](https://www.statistischebibliothek.de/mir/servlets/MCRFileNodeServlet/DEMonografie_derivate_00000209/1030821129004.pdf).

## Rejoinder<sup>1</sup>

I should like to thank the discussants for their kind remarks, for their valuable comments on the present state and future directions of the field, and for the many references they cite. Since I have no disagreements with them, I will confine my rejoinder to a few issues that their contributions have surfaced for me.

I will start by rectifying an oversight in my treatment of the early history of survey research and survey sampling: Carl-Erik Särndal has reminded me of the major developments that occurred in Russia during the early years. The impetus for these developments was the need for local self-government units known as *zemstva* to collect data about their populations for administrative purposes. Initially such data were collected with 100% enumerations, but around 1875 sample surveys were introduced for cost savings. The survey procedures were coordinated across *zemstva* and a number of sampling methods were evaluated with input from theoretical statisticians. These statisticians made a number of important contributions, including an impressive early text (1924) entitled *The Foundations of the Theory of the Sampling Method* by A. G. Kowalsky. Although Russian statisticians were at the frontiers of developments in survey sampling until the late 1920's, their contributions were not fully recognized outside Russia. For example, Tschuprow (1923) and Kowalsky in his 1924 text both derived the optimum allocation formula for stratified sampling a decade before Neyman did so in his famous 1934 paper (after learning of Tschuprow's paper, Neyman (1952) recognized Tschuprow's priority for the results). Mespoulet (2002), Zarkovic (1956), Zarkovic (1962), and Seneta (1985) provide further details about early survey research and research on survey sampling in Russia.

Danny Pfeffermann has pointed out that probability samples are almost never representative because of nonresponse—and I would add noncoverage—that is not mis-

---

<sup>a</sup> Joint Program in Survey Methodology, University of Maryland, College Park, MD, USA.  
E-mail: gkalton@gmail.com. ORCID: <https://orcid.org/0000-0002-9685-2616>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. 43–45.  
<https://doi.org/10.59170/stattrans-2023-034>.

sing completely at random (NMAR or MCAR). Moreover, I do not think the non-response should be viewed as missing at random (MAR), that is MCAR after conditioning on known covariates. Using standard weighting adjustments based on known covariates will not make the sample representative. My favorite quotation from George Box is “Essentially, all models are wrong, but some are useful”. Nonresponse adjustments should be viewed from this perspective as useful but not perfect. Another George Box quotation: “Statisticians, like artists, have the bad habit of falling in love with their models.” But there is a difference: artists have artistic license to paint over a model’s blemishes whereas statisticians should attempt to identify and repair the blemishes.

Risto Lehtonen points out the considerable attractions of population registers, as have existed for some time in several Scandinavian countries and are in development elsewhere. Such registers can be viewed as surveys with 100% samples, and the quality of their data should be assessed accordingly: What is their actual coverage? How up-to-date are they? How accurate are the data they contain?

Risto’s discussion of population registers also reminded me of a point that I should have addressed more fully: there is a wide variation in the data infrastructure for social research across countries. For example, most developing countries are not in a position to use administrative records or the internet. They rely on probability sample surveys to satisfy their data needs. Fortunately, they have not yet experienced the severe declines in response rates that are so harmful to surveys in most high-income countries.

Julie Gershunskaya and Partha Lahiri address two important current areas of research. One is the research on how to employ a probability sample to reduce the bias in estimates from a nonprobability sample, making use of auxiliary variables collected in both samples. The auxiliary variables aim to capture the key variables that are predictors of membership in the nonprobability sample. Challenges to be addressed with this approach include identifying the key variables; dealing with the fact that some response categories that occur frequently in the probability sample are very sparsely represented in the nonprobability sample; and concerns about the equivalence of the responses to the key variables obtained in the two samples that use different modes of collection. The results from this approach should be viewed with caution. However, recalling George Box’s quotation above, imperfect models can be useful. Julie and Partha rightly say that the aim of these models is to reduce, not eliminate, bias. The question to be asked is how to assess whether the models have reduced bias to an acceptable level.

The second area that Julie and Partha address is small area estimation. I should have written more about this methodology whose use has now become so widespread. My first practical exposure to small area estimation occurred in the late 1990’s, when

I chaired a National Academy of Sciences' panel that was asked to advise about the quality of the small area estimates of the numbers of poor school-aged children that were being developed in the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. The central issue was whether the estimates, which were produced for 3,000 counties and 14,000 school districts, were appropriate and sufficiently reliable to be used in allocating very large sums of money directly to school districts. At that time, this was a novel application of small area estimates, and subject to considerable questioning. After extensive evaluation of the area level models by both the Panel and the Census Bureau (Citro and Kalton, 2000), the Panel concluded that the small area estimates were "fit for use" for the purpose of this fund allocation, despite a recognition of substantial errors in the individual estimates. The Panel was influenced by the fact that the legislation stipulated that the funds should be distributed directly to the school districts and that, even though the small area estimates were not ideal, they were the best available. I was persuaded by my experience on the Panel that, with strong predictors and careful model development and testing, small area estimation methods have an important role to play in responding to policy makers' increasing demands for local area estimates.

Ralf Münnich emphasizes the importance of assessing the overall quality of statistical estimates in the light of the uses of the estimates. As he notes, timeliness is often in conflict with accuracy. In some situations, timeliness may be paramount, and accuracy may suffer. However, one must guard against the risk that accuracy is so low that the resulting estimates are misleading. Estimates based on big data sources or even large surveys conducted with an overriding emphasis on speed may, because of their sample sizes, appear to be well-grounded but that may well be illusory.

It is often argued that although individual estimates may be subject to serious biases, these biases will cancel out for differences between estimates, either between subgroups of the sample or across time. While the underlying model for that argument often appears reasonable, the assumptions underpinning it need to be carefully assessed in each case.

Ralf also points out the importance of cost constraints. When the cost constraints severely limit a study to a very small sample size, it may be preferable to forego the extra costs involved in selecting and fielding a probability sample, in favor of a quasi-probability sample or a nonprobability sample design. As Kish (1965, p. 29) notes: "Probability sampling is not a dogma, but a strategy, especially for large numbers".

Finally, Ralf and other discussants have pointed out the attractions of data integration. I also see these attractions, but I think that the challenges of mode effects arising from different data sources should not be underestimated.

In conclusion, I congratulate *Statistics in Transition* on celebrating its 30<sup>th</sup> anniversary. It plays a distinct and important role among statistics journals. With the major

changes in statistical methodology taking place in official statistics and in social research, it has a bright future for the contributions it can make.

## References

- Citro, C. F., Kalton, G. Eds., (2000). *Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. National Academy Press, Washington D.C.
- Kish, L., (1965). *Survey Sampling*. Wiley, New York.
- Mespoulet, M., (2002). From typical areas to random sampling: sampling methods in Russia from 1875 to 1930. *Science in Context*, 15(3), pp. 411–425. <https://doi.org/10.1017/S0269889702000546>.
- Neyman, J., (1952). Recognition of priority. *Journal of the Royal Statistical Society, A*, 115(4), 602.
- Seneta, E., (1985). A sketch of the history of survey sampling in Russia. *Journal of the Royal Statistical Society, A*, 148(2), pp. 118–125. <https://doi.org/10.2307/2981944>.
- Tschuprow, A. A., (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron*, 2(4), pp. 646–683.
- Zarkovic, S. S., (1956). Note on the history of sampling methods in Russia. *Journal of the Royal Statistical Society, A*, 119(3), pp. 336–338. <https://doi.org/10.2307/2342741>.
- Zarkovic, S. S., (1962). A supplement to “Note on the history of sampling methods in Russia”. *Journal of the Royal Statistical Society, A*, 125(4), pp. 580–582. <https://doi.org/10.2307/2982615>.

# Post Scriptum<sup>1</sup>

Włodzimierz Okrasa  
Dominik Rozkrut

## Celebrating the 100th issue and the 30th anniversary

This volume contains a selection of articles submitted for publication in the issue printed as the hundredth in 30 years of publishing *Statistics in Transition*.

With a sense of deep gratitude and the highest appreciation we would like to thank, both personally and on behalf of all the editorial bodies, Professor Graham Kalton for preparing his Invited Paper entitled *Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day*. Dr. Kalton is a long-time friend of our journal, and he serves as a member of our Editorial Board. The issues discussed in Dr. Kalton's paper are particularly appropriate at this time, as major changes are taking place in survey research methods and in sources of official statistics. The paper and the discussion pieces should therefore be of interest to members of the international statistician community and to members of national statistical offices.

Despite the relatively short time for reactions, we are grateful to five eminent experts, four of whom are associated with *SiTns*, for preparing four discussion pieces related to the paper. The authors of the four discussions are Professor Danny Pfeffermann, Dr. Julie Gershunskaya and Professor Partha Lahiri, Professor Risto Lehtonen, and Professor Ralf Münnich. Each of the discussions provides insightful observations supplementing some of the issues picked out from those discussed by Graham Kalton. They share concerns about the current challenges to probability sampling and design-based inference primarily caused by the serious declines in response rates, especially in high-income countries. They point to the possibilities of using alternative modalities (administrative data, big data, internet data, scientific data, etc.) for data collection that can supplement or replace probability samples. They describe the considerable body of research that is in progress to enable

---

<sup>1</sup> The article was published as Preface in *Statistics in Transition new series*, vol. 24, 2023, 3, pp. I.

these alternative data sources to produce valid population estimates from the nonprobability samples associated with the modalities, to the data integration methods that are being developed to combine the data obtained from different sources.

The first issue appeared in July 1993, and for the next fifteen years it was a semi-annual publication. In 2007, the title of the journal was slightly changed to *Statistics in Transition new series* and it became a quarterly publication. To celebrate the historical significance of these milestones, we dedicate the first part of this issue to them, opening it with a specially prepared Invitation Paper, along with four discussion pieces of the issues raised in that paper.

An addendum to this section contains a paper by Professor Jacek Wesołowski entitled *Rotation schemes and Chebyshev polynomials*, as being inspired in a way by the Invited Paper, and as an indication of other types of effects that it may have as well. It is noteworthy that as our journal celebrates its 30th anniversary, the journal's name *Statistics in Transition* well reflects the radical changes in the methodology of survey statistics and official statistics that are currently underway, as indicated in the Invited Paper and the discussions in this section.

Włodzimierz Okrasa  
Editor-in-Chief  
*Statistics in Transition new series*

Dominik Rozkrut  
President  
Statistics Poland

## **2. Small area estimation**





## Small area estimation: its evolution in five decades<sup>1</sup>

**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.

---

<sup>a</sup> Department of Statistics, University of Florida, Gainesville, FL, USA. E-mail: ghoshm@stat.ufl.edu.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 1–22.  
<https://doi.org/10.21307/stattrans-2020-022>.

## 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 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 Erculescu, 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 postcensal 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  $y_1, \dots, y_m$  are the direct estimates of average income for  $m$  areas with population sizes  $N_1, \dots, N_m$ , 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  $Corr(y_i, x_i) = .70$ . For nonsampled counties, regression synthetic estimates were  $\hat{M}_i = .52 + .74x_i$ .

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  $j$  ( $j = 1, \dots, G$ ). 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  $j$ th unit in the  $i$ th area;  $j = 1, N_i$ ;  $i = 1, \dots, m$ . Let  $x_{ij}$  denote some auxiliary characteristic for the  $j$ th unit in the  $i$ th 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  $y_{ij}$ ,  $j = 1, \dots, n_i$  with corresponding auxiliary variables  $x_{ij}$ ,  $j = 1, \dots, n_i$ . Let  $\sum_{j=1}^{n_i} x_{ij} / n_i \bar{x}_i$  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 to  $\bar{x}_i$  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  $y_{ij} \overset{ind}{\sim} (bx_{ij}, \sigma^2 x_{ij})$ . Best linear unbiased 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:  $y_i = \theta_i + e_i, i = 1, \dots, m$  and Linking Model:  $\theta_i = x_i^T b + u_i, i = 1, \dots, m$ . The target is estimation of the  $\theta_i, i = 1, \dots, m$ . 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 nonidentifiability 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  $y = (y_1, \dots, y_m)^T$ ;  $\theta = (\theta_1, \dots, \theta_m)^T$ ;  $e = (e_1, \dots, e_m)^T$ ;  $u = (u_1, \dots, u_m)^T$ ;  $X^T = (x_1, \dots, x_m)$ ;  $b = (b_1, \dots, b_p)^T$ .

We assume  $X$  has rank  $p (< m)$ . In vector notations, we write  $y = \theta + e$  and  $\theta = X_b + u$ .

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

It is possible to give an alternative Bayesian formulation of the Fay-Herriot model. Let  $y_i \mid \theta_i \stackrel{\text{ind}}{\sim} N(\theta_i; D_i)$ ;  $\theta_i \mid b \stackrel{\text{ind}}{\sim} N(x_i^T b, A)$ . Then the Bayes estimator of  $\theta_i$  is  $(1 - B_i)y_i + B_i x_i^T \tilde{b}$ , where  $B_i = D_i / (A + D_i)$ . If instead we put a uniform( $R_p$ ) prior for  $b$ , the Bayes estimator of  $\theta_i$  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)  $\tilde{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$  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  $\tilde{b}_L = (X^T X)^{-1} X^T y$ .

Then  $E \|y - x \hat{b}_L\|^2 = (m - p)A + \sum_{i=1}^m D_i (1 - r_i)$ ,  $r_i = x_i^T (x^T x)^{-1} x_i$ ,  $i = 1, \dots, m$ .

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  $\theta$  is  $\hat{\theta}_i^{EB} = (1 - \hat{B}_i^L) y_i + \hat{B}_i^L x_i^T \tilde{b}(\hat{A}_L)$ , where

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



Prasad and Rao also found an approximation to the mean squared error (Bayes risk) of their EBLUP or EB estimators. Under the subjective prior  $\theta_i \overset{\text{ind}}{\sim} N(x_i^T b, A)$ , the Bayes estimator of  $\theta_i$  is  $\hat{\theta}_i^B = (1 - B_i)y_i + B_i x_i^T b, B_i = D_i(A + D_i)$ . Also, write  $\hat{\theta}_i^{EB}(A) = (1 - B_i)y_i + B_i x_i^T \tilde{b}(A)$ . Then  $E(\hat{\theta}_i^{EB} - \theta_i)^2 = E(\hat{\theta}_i^B - \theta_i)^2 + E(\hat{\theta}_i^{EB}(A) - \hat{\theta}_i^B)^2 + E(\hat{\theta}_i^{EB} - \hat{\theta}_i^{EB}(A))^2$

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  $E(\hat{\theta}_i^{EB}(A) - \hat{\theta}_i^B)^2 = B_i^2 x_i^T (X^T V^{-1} X)^{-1} x_i = g_{2i}(A)$ , 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  $2B^2(D_i + A)^{-1} A^2 \sum_{i=1}^m (1 - B_i)^2 / m^2 = g_{3i}(A)$ , 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 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.

$$\text{Average Relative Absolute Bias} = (51)^{-1} \sum_{i=1}^{51} |e_i - e_{i,TR}| / e_{i,TR}$$

$$\text{Average Squared Relative Bias} = (51)^{-1} \sum_{i=1}^{51} (e_i - e_{i,TR})^2 / e_{i,TR}^2$$

$$\text{Average Absolute Bias} = (51)^{-1} \sum_{i=1}^{51} |e_i - e_{i,TR}|$$

$$\text{Average Squared Deviation} = (51)^{-1} \sum_{i=1}^{51} (e_i - e_{i,TR})^2$$

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  $2D^2 (A + D_i)^{-3} [\sum_{i=1}^m (A + D_i)^{-2}]^{-1}$ , 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

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

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  $D_1 = \dots = D_m$

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). Pfeiffermann 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  $f(y_i|\theta_i) = \exp[y_i\theta_i - \psi(\theta_i) + h(y_i)]$ ,  $i = 1, \dots, m$ . An example is the Bernoulli ( $p_i$ ) where  $\theta_i = \text{logit}(p_i) = \log(p_i/(1 - p_i))$  and Poisson( $\lambda_i$ ) where  $\theta_i = \log(\lambda_i)$ . One models the  $\theta_i$  as independent  $N(x_i^T b, A)$  and proceeds. Alternately, use beta priors for the  $p_i$  and gamma priors for the  $\lambda_i$ .

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  $y_i \stackrel{ind}{\sim} (\theta_i, D_i)$ . 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  $b$  and  $A$  are referred to as observed best predictive estimators. The corresponding estimators of the  $\theta_i$  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  $i$ th local area has  $N_i$  units with a sample of size  $n_i$ . We denote the sampled observations by  $y_{i1} \dots, y_{in_i}$ ,  $i = 1, \dots, m$ . Consider the model

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

The  $u_i$ 's and  $e_{ij}$ 's are mutually independent with the  $u_i$  iid  $(0, \sigma^2)$ , 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  $j$ th segment in the  $i$ th county;  $x_{ij} = (1, x_{ij1}, x_{ij2})^T$ , where  $x_{ij1}$  denotes the no. of pixels classified as corn for the  $j$ th segment in the  $i$ th county and  $x_{ij2}$  denotes the no. of pixels classified as soybean for the  $j$ th segment in the  $i$ th 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  $\bar{Y}_i = N^{-1} \sum_{j=1}^{N_i} y_{ij}$ , 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  $j$ th business firm in the  $i$ th census division in Canada and  $x_{ij} = (1, x_{ij})^T$ , where  $x_{ij}$  is the gross business income of the  $j$ th business firm in the  $i$ th 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  $y_i = (y_{i1}, \dots, y_{iN_i})^T$ ,  $X_i = (X_{i1}, \dots, X_{iN_i})^T$ ,  $e_i = (e_{i1}, \dots, e_{iN_i})^T$ ,  $i = 1, \dots, m$ . Thus, the model is rewritten as

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

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  $\bar{X}_i = N_i^{-1} \sum_{j=1}^{N_i} x_{ij}$ ,  $1, \dots, m$ .

For known  $\sigma_u^2$  and  $\sigma_e^2$ , 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  $\sigma_u^2$  and  $\sigma_e^2$ . The EBLUP of  $\bar{X}_i^T b + u_i$  is now found by substituting these estimates of  $\sigma_u^2$  and  $\sigma_e^2$  in the BLUP formula. Estimation of  $\sigma_e^2$  is based on the moment identity

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

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

$$E \left[ \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^T \tilde{b} \right]^2 = (n-p)\sigma_c^2 + \sigma_u^2 \left[ m - \sum_{i=1}^m n_i^2 x_i^T (x^T x)^{-1} \bar{x}_i \right],$$

where  $\hat{b} = (X^T X)^{-1} X^T y$ ,  $y = (y_1^T \dots, y_m^T)^T$ . If this results in a negative estimator of  $\sigma_u^2$ , 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  $\sigma_e^2$  and  $\sigma_u^2$  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  $\sigma_u^2$  and  $\sigma_e^2$ , it suffices to assign a prior to  $\lambda$ . This is what was proposed in Datta and Ghosh (1991), who assigned a Gamma prior to  $\lambda$ . The Bayesian approach of Datta and Ghosh (1991) did also accommodate 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  $\theta_i$  is the  $i$ th 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  $j$ th area. The direct estimate for  $\theta_T$  is  $\sum_{i=1}^m w_i \hat{\theta}_i$ . Also, let  $\hat{\theta}_i$  denote an estimator of  $\theta_i$  based on a certain model. Then  $\sum_{i=1}^m w_i \hat{\theta}_i$  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

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

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  $\hat{\theta}_j^G$  some generic model-based estimator of  $\theta_j$ .

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  $\sum_{i=1}^m \phi_j E(e_j - \theta_j)^2$  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 The resulting estimator of  $\theta_i$  is

$$\hat{\theta}_i^{WFO} = \hat{\theta}_i^{BLOP} + \lambda_i \left( \sum_{j=1}^m w_j \hat{\theta}_j - \sum_{j=1}^m w_j \hat{\theta}_j^{BLOP} \right)$$

where  $\lambda_i = w_i \phi_i^{-1} / \left( \sum_{j=1}^m w_j \hat{\theta}_j - \sum_{j=1}^m w_j \hat{\theta}_j^B \right)$

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  $\lambda_i$ . 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  $MSE(\hat{\theta}_i^{DB}) = MSE(\hat{\theta}_i^{EB}) + g_4(A) + o(m^{-1})$ , where  $MSE(\hat{\theta}_i^{EB})$  is the same as the one given in Prasad and Rao (1990), and

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

We may recall that  $B_i = \frac{D_i}{A+D_i}$ ,  $X^T = (x_1, \dots, x_m)$  and  $V = \text{Diag}(A + D_1, \dots, A + D_m)$  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 Bench-marked 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  $\delta_i$  and the  $u_i$  are independent with  $\delta_i$  iid Bernoulli( $\gamma$ ) and  $u_i$  iid  $N(0, \sigma^2)$ .

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  $\lambda_i^2$  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).

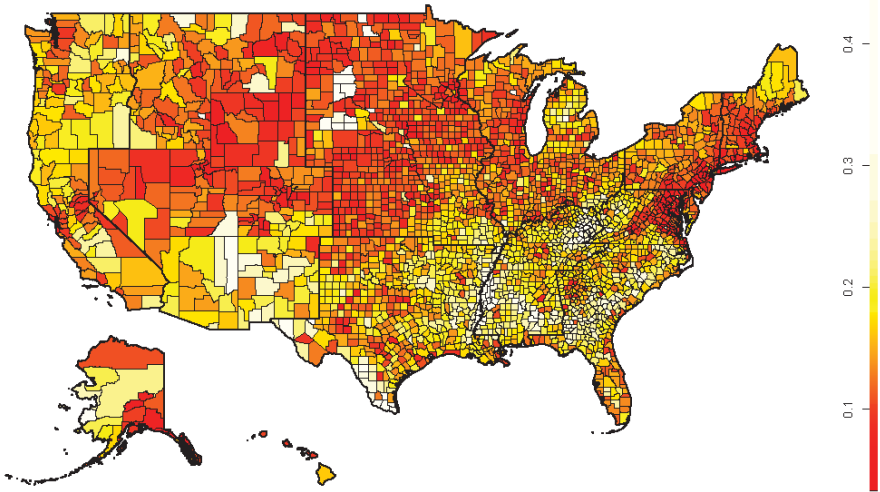
As an example, consider estimation of 5-year (2007–2011) county-level overall poverty ratios in the US. There are 3,141 counties in the data set. The covariates are foodstamp participation rates. The map given in Figure 1 gives the poverty ratios for all the counties of US. Some salient findings from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

**Figure 1.** Map of posterior means of  $\theta$ 's.



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 distributions, 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 transformation 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)

$z_i | \theta_i \stackrel{ind}{\sim} N(\theta_i, D_i)$  and (ii)  $\theta_i \stackrel{ind}{\sim} N(x_i^T \beta, A)$ .  $\theta_i$  has the  $N(\hat{\theta}_i^B, D_i(1 - B_i))$  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  $E(\theta_i | z_i) = \exp[\hat{\theta}_i^B + (1/2)D_i(1 - B_i)]$ .

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

A third example is the Poisson model for count data. There  $y_i \stackrel{ind}{\sim} \text{Poisson}(\lambda_i)$ . Then one models  $z_i = y_i^{1/2}$  as independent  $N(\theta_i, 1/4)$  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 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(1), pp. 119–143. <https://doi.org/10.5705/ss.2011.048>.
- Armagan, A., Dunson, D. B., Lee, J., and Bajwa, W. U., (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika*, 100(4), pp. 1011–1018. <https://doi.org/10.1093/biomet/ast028>.
- 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(401), pp. 28–36. <https://doi.org/10.2307/2288915>.
- Bell, W. R., Datta, G. S., and Ghosh, M., (2013). Benchmarking small area estimators. *Biometrika*, 100, pp. 189–202. <https://doi.org/10.1093/biomet/ass063>.
- 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. <https://doi.org/10.1002/9781118814963.ch19>.
- 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. <https://doi.org/10.1093/jssam/smt041>.
- Bivariate hierarchical Bayesian model for estimating cropland cash rental rates at the county level. *Survey Methodology*, in press.
- Erciulescu, A., Berg, E., Cecere, W., and Ghosh, M. (2019). A bivariate hierarchical Bayesian model for estimating cropland cash rental rates at the county level. *Survey Methodology*, 45(2), 199–216. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2019002/article/00002-eng.pdf>.
- Booth, J. G., Hobert, J., (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93(441), pp. 262–272. <https://doi.org/10.1080/01621459.1998.10474107>.
- Butar, F. B., Lahiri, P., (2003). On measures of uncertainty of empirical Bayes small area estimators. *Journal of Statistical Planning and Inference*, 112(1–2), pp. 63–76. [https://doi.org/10.1016/S0378-3758\(02\)00323-3](https://doi.org/10.1016/S0378-3758(02)00323-3).
- Carvalho, C. M., Polson, N. G., Scott, J. G., (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2), pp. 465–480. <https://doi.org/10.1093/biomet/asq017>.
- 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. <http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000585.pdf>.
- Das, K., Jiang, J., Rao, J. N. K., (2004). Mean squared error of empirical predictor. *Annals of Statistics*, 32(2), pp. 818–840. <https://doi.org/10.1214/009053604000000201>.

- Datta, G. S., Ghosh, M., (1991). Bayesian prediction in linear models: applications to small area estimation. *The Annals of Statistics*, 19(4), pp. 1748–1770. <https://doi.org/10.1214/aos/11176348369>.
- 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(2), pp. 613–627. <https://www3.stat.sinica.edu.tw/statistica/oldpdf/A10n214.pdf>.
- 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(1), pp. 183–196. <https://doi.org/10.1093/biomet/92.1.183>.
- Datta, G. S., Ghosh, M., Steorts, R., and Maples, J. J., (2011). Bayesian benchmarking with applications to small area estimation. *TEST*, 20(3), pp. 574–588. <https://doi.org/10.1007/s11749-010-0218-y>.
- 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(493), pp. 362–374. <https://doi.org/10.1198/jasa.2011.tm10036>.
- Datta, G. S., Mandal, A., (2015). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, 110(512), pp. 1735–1744. <https://doi.org/10.1080/01621459.2015.1016526>.
- Elbers, C., Lanjouw, J. O., and Lanjouw, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), pp. 355–364. <https://doi.org/10.1111/1468-0262.00399>.
- Erciulescu, A. L., Franco, C., and Lahiri, P., (2021). Use of administrative records in small area estimation. In *Administrative Records for Survey Methodology*, pp. 231–267. Eds. P. Chun M. D. Larsen, G. Durrant and J. P. Reiter. Wiley, New York. <https://doi.org/10.1002/9781119272076.ch10>.
- 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(366), pp. 269–277. <https://doi.org/10.1080/01621459.1979.10482505>.
- 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*, 9(1), pp. 55–93. <https://doi.org/10.1214/ss/1177010647>.
- 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(411), pp. 273–282. <https://doi.org/10.1080/01621459.1998.10474108>.
- Ghosh, M., Steorts, R., (2013). Two-stage Bayesian benchmarking as applied to small area estimation. *TEST*, 2(4), pp. 670–687. <https://doi.org/10.1007/s11749-013-0338-2>.
- Ghosh, M., Kubokawa, T., and Kawakubo, Y., (2015). Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation. *Biometrika*, 102(3), pp. 647–659. <https://doi.org/10.1093/biomet/asv010>.

- Ghosh, M., Myung, J., and Moura, F. A. S., (2018). Robust Bayesian small area estimation. *Survey Methodology*, 44(1), pp. 101–115. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2018001/article/54959-eng.pdf>.
- Gonzalez, M. E., Hoza, C., (1978). Small area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association*, 73(261), pp. 7–15. <https://doi.org/10.1080/01621459.1978.10479991>.
- Griffin, J. E., Brown, P. J., (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1), pp. 171–188. <https://doi.org/10.1214/10-BA507>.
- 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(366a), pp. 405–410. <https://doi.org/10.1080/01621459.1979.10482527>.
- Jiang, J., Lahiri, P., (2001). Empirical best prediction of small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, 53(2), pp. 217–243. <https://doi.org/10.1023/A:1012410420337>.
- Jiang, J., Lahiri, P., and Wan, S-M., (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics*, 30(6), pp. 1782–1810. <https://doi.org/10.1214/aos/1043351257>.
- Jiang, J., Lahiri, P., (2006). Mixed model prediction and small area estimation (with discussion). *TEST*, 5(1), pp. 1–96. <https://doi.org/10.1007/BF02595419>.
- Jiang, J., Nguyen, T., and Rao, J. S., (2011). Best predictive small area estimation. *Journal of the American Statistical Association*, 106(494), pp. 732–745. <https://doi.org/10.1198/jasa.2011.tm10221>.
- 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(388), pp. 853–862. <https://doi.org/10.1080/01621459.1984.10477102>.
- Lahiri, P., Rao, J. N. K., (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90(430), pp. 758–766. <https://doi.org/10.1080/01621459.1995.10476570>.
- Lahiri, P., Pramanik, S., (2019). Evaluation of synthetic small area estimators using design-based methods. *Austrian Journal of Statistics*, 48(4), pp. 43–57. <https://doi.org/10.17713/ajs.v48i4.790>.
- 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. <https://doi.org/10.1080/01621459.1984.10478062>.
- 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(23), pp. 3189–3200. [https://doi.org/10.1002/\(SICI\)1097-0258\(19991215\)18:23<3189::AID-SIM309>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1097-0258(19991215)18:23<3189::AID-SIM309>3.0.CO;2-C).
- Molina, I., Rao, J. N. K., (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 8(3), pp. 369–385. <https://doi.org/10.1002/cjs.10051>.
- 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 area effects. *Survey Methodology*, 41(1), 1–19. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14161-eng.pdf>.

- Morris, C. N., (1983). Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association*, 78(381), pp. 47–55. <https://doi.org/10.1080/01621459.1983.10477920>.
- Pfeffermann, D., (2002). Small area estimation: new developments and direction. *International Statistical Review*, 70(1), pp. 125–143. <https://doi.org/10.1111/j.1751-5823.2002.tb00352.x>.
- Pfeffermann, D., (2013). New important developments in small area estimation. *Statistical Science*, 28(1), pp. 40–68. <https://doi.org/10.1214/12-STS395>.
- Pfeffermann, D., Tiller, R. B., (2005). Bootstrap approximation of prediction MSE for state-space models with estimated parameters. *Journal of Time Series Analysis*, 26(6), pp. 893–916. <https://doi.org/10.1111/j.1467-9892.2005.00448.x>.
- Polson, N. G., Scott, J. G., (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9, pp. 501–538. <https://doi.org/10.1093/acprof:oso/9780199694587.003.0017>.
- 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(409), pp. 163–171. <https://doi.org/10.1080/01621459.1990.10475320>.
- Raghunathan, T. E., (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, 88(424), pp. 1444–1448. <https://doi.org/10.1080/01621459.1993.10476431>.
- 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(4), pp. 523–526. <https://doi.org/10.21307/stattrans-2015-029>.
- Rao, J. N. K., Molina, I., (2015). *Small Area Estimation*, 2nd Edition. Wiley, New Jersey. <https://doi.org/10.1002/9781118735855>.
- Scott, J. G., Berger, J. O., (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5), pp. 2587–2619. <https://doi.org/10.1214/10-AOS792>.
- Slud, E. V., Maiti, T., (2006). Mean squared error estimation in transformed Fay-Herriot models. *Journal of the Royal Statistical Society, B*, 68(2), pp. 239–257. <https://doi.org/10.1111/j.1467-9868.2006.00542.x>.
- 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(524), pp. 1476–1489. <https://doi.org/10.1080/01621459.2017.1419135>.
- Wang, J., Fuller, W. A., and Qu, Y., (2008). Small area estimation under a restriction. *Survey Methodology*, 34, pp. 29–36. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2008001/article/10619-eng.pdf>.
- 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. <https://doi.org/10.1016/j.jmva.2013.10.012>.
- 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(1), pp. 217–229. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2013001/article/11830-eng.pdf>.

## Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

### 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

---

<sup>a</sup> 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>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 23–29.  
<https://doi.org/10.21307/stattrans-2020-023>.

“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.

## 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  $M_0$  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 is

$$y_i | \theta_i \stackrel{ind}{\sim} N(\theta_i, D_i) \quad (1)$$

$$\theta_i | \mathbf{b} \stackrel{ind}{\sim} N(x_i^T \mathbf{b}, A) \quad (2)$$

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 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  $\widehat{D}_i$  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 Sugawawa et al. (2017). Assume in domain  $i = 1, \dots, m$ , the following model M1 holds for pair  $(y_i, \widehat{D}_i)$

$$y_i | \theta_i, D_i \stackrel{ind}{\sim} N(\theta_i, D_i), \quad \theta_i | \mathbf{b}, A \stackrel{ind}{\sim} N(x_i^T \mathbf{b}, A), \quad (3)$$

$$\widehat{D}_i | D_i \stackrel{ind}{\sim} G\left(\frac{n_i-1}{2}, \frac{n_i-1}{2D_i}\right), \quad D_i | \gamma \stackrel{ind}{\sim} IG(a_i, c_i \gamma), \quad (4)$$

where (3) is the usual FH model for the point estimate and (4) describes a companion model for observed variance  $\widehat{D}_i$  (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;  $\gamma$  is an unknown parameter;  $a_i$  and  $c_i$  are positive known constants, Sugawawa 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):

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

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

In this model, we assume the existence of  $K$  latent clusters having cluster-specific intercepts  $b_{0k}$ ,  $k = 1, \dots, K$  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  $\boldsymbol{\gamma}_k$ ; is a vector of covariates for the variance model for area  $i$ ; 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 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(4), pp. 1053–1063. <https://www3.stat.sinica.edu.tw/statistica/oldpdf/A7n416.pdf>.
- 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(2), pp. 173–187. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2012002/article/11756-eng.pdf>.
- 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*, 8(2), pp. 181–205, <https://doi.org/10.1093/jssam/smz004>.
- Hansen, M. H., (1987). Some History and Reminiscences on Survey Sampling. *Statistical Science*, 2(2), pp. 180–190. <https://doi.org/10.1214/ss/1177013352>.
- Liu, B., Lahiri, P., Kalton, G., (2014). Hierarchical Bayes modelling of survey-weighted small area proportions. *Survey Methodology*, 40(1), pp. 1–13. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2014001/article/14030-eng.pdf>.
- Mahalanobis, P. C., (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109(4), 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(3), pp. 775–790. <https://doi.org/10.1111/sjos.12061>.
- 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(1), pp. 150–167. <https://doi.org/10.1111/sjos.12246>.
- You, Y., Chapman, B., (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32(1), pp. 97–103. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2006001/article/9263-eng.pdf>.

## **Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>**

### **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

---

<sup>a</sup> Gallup, Inc, USA. E-mail: [ying\\_han@gallup.com](mailto:ying_han@gallup.com). ORCID: <https://orcid.org/0000-0003-0082-5654>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 30–34. <https://doi.org/10.21307/stattrans-2020-024>.

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).

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 over-emphasized. 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. <https://ro.uow.edu.au/eispapers/762/>.
- Fellegi, I., Sunter, A., (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 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(S1), pp. S139–S157. <https://doi.org/10.1111/insr.12295>.
- Jiang, J., Lahiri, P., (2006). Mixed model prediction and small area estimation. *Test*, 15(1), pp. 1–96. <https://doi.org/10.1007/BF02595419>.
- 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. <https://doi.org/10.1214/aos/1043351257>.
- Kim, J., Chambers, R., (2012). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, 56(9), pp. 2756–2770. <https://doi.org/10.1016/j.csda.2012.02.026>.
- 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. <https://doi.org/10.1198/016214504000001277>.
- Lohr, S. L., Rao, K., (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika*, 96(2), pp. 457–468. <https://doi.org/10.1093/biomet/asp003>.
- 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. <https://doi.org/10.1080/01621459.1965.10480846>.
- Pfefferman, D., (2013). New important developments in small area estimation. *Statistical Science*, 28(1), pp. 40–68. <https://doi.org/10.1214/12-STS395>.
- Rao, J., Molina, I., (2015). *Small Area Estimation*. Wiley, second edition. <https://doi.org/10.1002/9781118735855>.
- 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. <https://doi.org/10.1111/anzs.12052>.

## Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

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

---

<sup>a</sup> 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>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 35–39. <https://doi.org/10.21307/stattrans-2020-025>.

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, Mallows’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 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 Suntornclost, 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*:  $AIC = 2k + n \cdot \log(\frac{SSE_k}{n})$ , and
- 4) *BIC*:  $BIC = k \cdot \log n + n \cdot \log(\frac{SSE_k}{n})$ ,

where

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

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:**

$$y_i = \theta_i + e_i \text{ and } \theta_i = x_{ij}^T \beta + v_i, \tag{1}$$

where  $\theta_i$  is the unobserved true mean for small area  $i$ ;  $y_i$  is the survey-weighted estimate of  $\theta_i$ ;  $v_i$  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  $\bar{D}_w = \frac{\sum_{i=1}^m (1-h_{ii}) D_i}{m-k}$ ,  $h_{ii} = x_i^T (X^T X)^{-1} x_i$ , 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  $j = 1, \dots, n_i$  in area  $i = 1, \dots, m$ , 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  $v_i \sim iid N(0, \sigma_v^2)$  is independent of  $e_{ij} \sim iid N(0, \sigma_e^2)$ . Define  $\sigma^2 = \sigma_e^2 + \sigma_v^2$ . The outcome in unit  $j$  of area  $i$  is denoted by  $y_{ij}$ , and  $x_{ij} = (1, x_{ij1}, x_{ij2}, \dots, x_{ijk})$  is a vector of length  $k+1$  with the values of the covariates  $x_1, x_2, \dots, x_k$  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  $\begin{pmatrix} \frac{1}{2} 1_{n_i}^T \\ P_i \end{pmatrix}$  is orthogonal for  $i = 1, 2, \dots, m$ , and transformed the data by

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

The transformed model can then be written as:

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

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:**

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

Compared to the unit-level nested error regression model (2), an additional error term  $\gamma_{ij} \sim iid N(0, \sigma_\gamma^2)$  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  $\theta_{ij}$ , given by

$$\theta_i^{LL} = x_i^{LL} \beta + u_i^{LL}, \tag{5}$$

which is similar to Model (3) but with unobserved response  $\theta_i^{LL}$ . 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  $\theta_i^{LL}$  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 new series* 21(4), pp. 68-83. <https://doi.org/10.21307/stattrans-2020-031>.
- Han, B., (2013). Conditional Akaike information criterion in the Fay-Herriot model. *Statistical Methodology*, 11, pp. 53–67. <https://doi.org/10.1016/j.stamet.2012.09.002>.
- Jiang, J., Thuan, N., (2015). *The Fence Methods*. World Scientific Publishing Co. Pte. Ltd., Singapore. <https://doi.org/10.1142/9116>.
- Lahiri, P. ed., (2001). *Model Selection*. Beachwood, OH: Lecture Notes–Monograph Series, Institute of Mathematical Statistics. <https://doi.org/10.1214/lnms/1215540957>.
- Lahiri, P., Suntornchost, J., (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, 77(2), pp. 312–320. <https://doi.org/10.1007/s13571-015-0096-0>.
- 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), pp. 302–371. <https://doi.org/10.1007/s13571-018-0161-6>.
- Meza, J. L., Lahiri, P., (2005). A note on the Cp statistic under the nested error regression model. *Survey Methodology*, 31(1), pp. 105–109. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2005001/article/8094-eng.pdf>.
- Rao, J. N. K., Molina, I., (2015). *Small Area Estimation*, 2nd Edition. Hoboken: Wiley. <https://doi.org/10.1002/9781118735855>.

## Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

### 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 World 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 Battese 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

---

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

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 40–44.  
<https://doi.org/10.21307/stattrans-2020-026>.

inspired by the literature on multiple imputation rather than by the small area estimation literature.

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 tracts). 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 World 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 led 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(401), pp. 28–36. <https://doi.org/10.1080/01621459.1988.10478561>.
- Corral, P., Molina, I., Nguyen, M., (2020). *Pull your small area estimates up by the bootstraps*. World Bank Policy Research Working Paper 9256. <https://openknowledge.worldbank.org/server/api/core/bitstreams/a28589a8-6a1a-5a42-a8b8-26e1f7d60f60/content>.
- Diallo, M., RAO, J., (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, 45(4), pp. 1092–1116. <https://doi.org/10.1111/sjos.12336>.
- Elbers, C., Lanjouw, J. O., Lanjouw, P., (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71(1), pp. 355–364. <https://doi.org/10.1111/1468-0262.00399>.
- 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(1), pp. 105–120. <http://isas.org.in/isa/volume/vol66/issue1/09-CaterinaFerretti.pdf>.
- González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D., San-Tamaría, L., (2008). Bootstrap mean squared error of a small-area eblup. *Journal of Statistical Computation and Simulation*, 78(5), pp. 443–462. <https://doi.org/10.1080/00949650601141811>.
- 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(2), pp. 565–597. <https://doi.org/10.1007/s11749-018-0594-2>.

- 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. <https://doi.org/10.1016/j.csda.2017.11.007>.
- 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, pp. 1897–1904. <http://www.asasrms.org/Proceedings/y2003/Files/JSM2003-000108.pdf>.
- 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(4), pp. 1111–1136. <https://doi.org/10.1111/rssa.12306>.
- Molina, I., (2019). *Desagregación de datos en encuestas de hogares: metodologías de estimación en áreas pequeñas*. Seriesdela Comisión Económicapara América Latinayel Caribe (CEPAL), Estudios Estadísticos LC/TS.2018/82/Rev.1, CEPAL. <https://repositorio.cepal.org/server/api/core/bitstreams/5792f51b-c686-4624-9673-6bf6f6fa0d9d/content>.
- Molina, I., Marhuenda, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), pp. 81–98. <https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf>.
- 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(2), pp. 852–885. <https://doi.org/10.1214/13-AOAS702>.
- Molina, I., Rao, J. N. K., (2010). Small Area Estimation of Poverty Indicators. *The Canadian Journal of Statistics*, 38(3), pp. 369–385. <https://doi.org/10.1002/cjs.10051>.
- 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(1–2), pp. 131–147. [https://doi.org/10.1016/S0378-3758\(98\)00211-0](https://doi.org/10.1016/S0378-3758(98)00211-0).
- 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. <https://doi.org/10.1596/1813-9450-7028>.
- 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(3), pp. 431–439. <https://doi.org/10.2307/3316146>.

## Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

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

---

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

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 45–50.  
<https://doi.org/10.21307/stattrans-2020-027>.



(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 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 R<sup>2</sup>, 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<sup>2</sup>. 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

---

<sup>2</sup> Important exceptions are Pokhriyal and Jacques (2017) and Erciulescu et al (2018).

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 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. <https://doi.org/10.1126/science.aac4420>.

- 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. <https://doi.org/10.1111/1468-0262.00399>.
- Elbers, Chris, Peter Lanjouw, Phillippe George Leite, (2008). *Brazil within Brazil: Testing the poverty map methodology in Minas Gerais*. The World Bank. <https://doi.org/10.1596/1813-9450-4513>.
- Engstrom, Ryan, Jonathan Hersh, David Newhouse, (2017). *Poverty from space: Using high-resolution satellite imagery for estimating economic well-being*. The World Bank. <https://doi.org/10.1596/1813-9450-8284>.
- 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. <https://doi.org/10.1596/1813-9450-8776>.
- González-Manteiga, W., Lombardia, 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. <https://doi.org/10.1080/00949650601141811>.
- Hay, Simon, I., et al., (2009). A world malaria map: *Plasmodium falciparum* endemicity in 2007. *PLoS medicine*, 6(10). <https://doi.org/10.1371/annotation/a7ab5bb8-c3bb-4f01-aa34-65cc53af065d>.
- Jean, Neal, et al., (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), pp. 790–794. <https://doi.org/10.1126/science.aaf7894>.
- 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). <https://doi.org/10.3390/rs9090931>.
- Lange, S., Utz Johann Pape, Peter Pütz, (2018). *Small area estimation of poverty under structural change*. The World Bank. <https://doi.org/10.1596/1813-9450-8472>.
- 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*, 102(1), 202–219. <https://doi.org/10.1093/ajae/aaz051>.
- 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. <https://doi.org/10.1111/rssa.12306>.
- Molina, I., J. N. K. Rao, (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38(3), pp. 369–385. <https://doi.org/10.1002/cjs.10051>.
- Molina, I., Marhuenda, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), pp. 81–98. <https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf>.
- 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. <https://doi.org/10.1073/pnas.1700319114>.
- Steele, Jessica, E., et al., (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127), 20160690. <https://doi.org/10.1098/rsif.2016.0690>.
- 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. <https://doi.org/10.1016/j.jmva.2014.02.001>.



- 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. <https://doi.org/10.1596/1813-9450-7028>.
- 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. <https://doi.org/10.1073/pnas.1715305115>.
- 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. <https://doi.org/10.1073/pnas.1812969116>.
- Zhao, Qinghua., (2006). *User manual for POVMAP*. World Bank. <https://documents1.worldbank.org/curated/en/529951468135927356/pdf/765140WP0User000Box374382B00PUBLIC0.pdf>.

## Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

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 non-response is often informative, implying inevitably the same problem in at least some of the areas. NMAR nonresponse need to be handled properly, irrespectively

---

<sup>a</sup> 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>.

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 51–52. <https://doi.org/10.21307/stattrans-2020-028>.



- of the method of inference, whether design- or model-based; following the frequentist or the Bayesian approach.
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.

# Discussion of *Small area estimation: its evolution in five decades* by Malay Ghosh<sup>1</sup>

## 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

---

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

<sup>1</sup> The article was published in *Statistics in Transition new series*, vol. 21, 2020, 4, pp. 51–52.  
<https://doi.org/10.21307/stattrans-2020-029>.

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) 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 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 is used.

Fay and Herriot (1979) proposed a different moment estimator of by solving two equations iteratively.

The moment estimators of 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  $\theta$ . 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 lead to smaller average MSE than the AML-based EBLUPs when  $\theta$  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  $\theta$  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  $|\text{est.} - \text{truth}|/\text{truth}$ . 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 relative 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 (2021) 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 (Guadarrama 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(1), pp. 137–157. <https://doi.org/10.21307/stattrans-2020-008>.
- Guadarrama, M., Molina, I., Rao, J. N. K., (2016). A comparison of small area estimation methods for poverty mapping. *Statistics in Transition new series*, 17(1), pp. 41–66. <https://doi.org/10.21307/stattrans-2016-005>.
- 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*. [https://doi.org/10.1007/978-1-4612-2944-5\\_6](https://doi.org/10.1007/978-1-4612-2944-5_6).
- 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(384), pp. 776–793. <https://doi.org/10.1080/01621459.1983.10477018>.
- Hidirolou, M. A., Beaumont, J-F., Yung, W., (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, 45(1), pp. 101–126. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2019001/article/00009-eng.pdf>.
- Jiang, J., Rao, J. S., (2020). Robust small area estimation: An overview. *Annual Reviews*, 7, pp. 337–360. <https://doi.org/10.1146/annurev-statistics-031219-041212>.
- Kalton, G., (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review*, 87(S1), pp. S10–S30. <https://doi.org/10.1111/insr.12287>.
- Li, H., Lahiri, P., (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101(4), pp. 882–892. <https://doi.org/10.1016/j.jmva.2009.10.009>.
- Molina, I., Marhuenda, Y., (2015). Sae: An R package for Small Area Estimation. *The R Journal of Statistics*, 7(1), pp. 81–98. <https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf>.
- 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(1), pp. 1–19. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14161-eng.pdf>.



- Pfeffermann, D., Ben-Hur, D., (2018). Estimation of randomization mean squared error in small area estimation. *International Statistical Review*, 87(S1), pp. S31–S49. <https://doi.org/10.1111/insr.12289>.
- Rao, J. N. K., (2003). *Small Area Estimation*. Hoboken, NJ: Wiley. <https://doi.org/10.1002/0471722189>.
- Rao, J. N. K., (2021). On making valid inferences by integrating data from surveys and other sources. *Sankhya, Series B*, 83(1), pp. 242–272. <https://doi.org/10.1007/s13571-020-00227-w>.
- Rao, J. N. K., Rubin-Bleuer, S., Estevao, V. M., (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, 44(2), pp. 151–166. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2018002/article/54958-eng.pdf>.
- Sinha, S. K., Rao, J. N. K., (2009). Robust small area estimation. *Canadian Journal of Statistics*, 37(3), pp. 381–399. <https://doi.org/10.1002/cjs.10029>.
- 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(4), pp. 927–979. <https://doi.org/10.1111/rssa.12364>.

Malay Ghosh<sup>1</sup>

## Rejoinder

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. Gershunskaya 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  $y_i, \check{D}_i$  when possible, must be undertaken. Unfortunately, without the availability of microdata, especially for secondary users of surveys, modeling the 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

---

<sup>1</sup> Department of Statistics, University of Florida, Gainesville, FL. USA. E-mail: ghoshm@stat.ufl.edu.

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  $a_i$ ,  $a_i$  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  $\gamma$ , which enhances complexity.

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  $\pi D_i \propto D_i^{-1/2}(1 + D_i)^2$ , 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, 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.

Finally, 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 impossible 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. I 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 Regression Shrinkage) can provide a very direct variable selection and estimation in one stroke.

For simplicity 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  $\beta$  is given by

$$\check{\beta}_{LASSO} = argmin_{\beta} \left[ \|Y - X\beta\|^2 + \lambda \sum_j |\beta_j| \right]$$

where  $\lambda$  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 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 transformation 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.

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.

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?

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.

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.

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. Pfeiffermann. 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(3), pp. 761–766. <https://doi.org/10.2307/1913475>.
- Gelman, A., (2006). Prior distributions for variance parameters in hierarchical models. (Comment on article by Browne and Draper). *Bayesian Analysis*, 1(3), pp. 515–553. <https://doi.org/10.1214/06-BA117A>.
- Hall, P., Maiti, T., (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, B*, 68(2), pp. 221–238. <https://doi.org/10.1111/j.1467-9868.2006.00541.x>.

- Hirose, M. Y., Lahiri, P., (2018). Estimating variance of random effects to solve multiple problems simultaneously. *The Annals of Statistics*, 46(4), pp. 1721–1741. <https://doi.org/10.1214/17-AOS1600>.
- Tong, X., Xu, X., Ghosh, M., Ghosh, P., (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, 80(2), pp. 215–246. <https://doi.org/10.1007/s13171-017-0118-2>.
- 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. <https://doi.org/10.1016/j.jmva.2013.10.012>.

### **3. Reconstructing Ukraine's statistical system**



Dominik Rozkrut<sup>a</sup>  
Włodzimierz Okrasa<sup>b</sup>  
Oleksandr H. Osaulenko<sup>c</sup>  
Misha V. Belkindas<sup>d</sup>  
Ronald L. Wasserstein<sup>e</sup>

# The post-conflict reconstruction of the statistical system in Ukraine<sup>1</sup>

## Key issues from an international perspective<sup>2</sup>

### Introduction

Statistics has accompanied the social forms of human civilization since its inception, reflecting also conflicts and wars. Statistics acts as a beacon, especially in turbulent times, capturing the most important aspects of reality, while helping decision-makers navigate key choices in the face of adversity of a radically changing situation. To this end, statisticians of a war-affected country make every effort by adapting the way statistics work to overcome methodological and organizational obstacles in everyday professional work, including innovative development of research instruments to

---

<sup>a</sup> Statistics Poland, Poland. E-mail: D.Rozkrut@stat.gov.pl. ORCID: <https://orcid.org/0000-0002-0949-8605>.

<sup>b</sup> University of Cardinal Stefan Wyszyński in Warsaw and Statistics Poland, Poland.

E-mail: w.okrasa@stat.gov.pl. ORCID: <https://orcid.org/0000-0001-6443-480X>.

<sup>c</sup> National Academy of Statistics and Accounting and Audit, Kyiv, Ukraine.

E-mail: O.Osaulenko@nasoa.edu.ua. ORCID: <https://orcid.org/0000-0002-7100-7176>.

<sup>d</sup> International Association of Official Statistics (IAOS) and ODW Consulting, USA.

E-mail: MishaBelkin-das@opendatawatch.com.

<sup>e</sup> American Statistical Association, USA. E-mail: ron.wasserstein@amstat.org.

<sup>1</sup> The article was published in *Statistics in Transition* new series and *Statistics of Ukraine*, vol. 24, 2023, 1, pp. 1–12. <https://doi.org/10.59170/stattrans-2023-001>.

<sup>2</sup> Based on the presentations given by the panelists at the session *Marshall Plan for Reconstructing National Statistical Offices After Conflict: Practical Guidance from International Principles. The role of statistical societies*: Misha Belkindas; Ronald Wasserstein; Włodzimierz Okrasa and Dominik Rozkrut. The session was organized by Jennifer Park, Committee on National Statistics (CNSTAT); it was chaired by Dominik Rozkrut and commented by Albert Kroese (International Monetary Fund). It took place during the Federal Committee on Statistical Methodology/FCSM-2022 Research and Policy Conference, October 25–27, Washington D.C.

© D. Rozkrut, W. Okrasa, O. H. Osau-lenko, M. V. Belkindas, R. L. Wasserstein. Article available under the

substitute the destroyed or unavailable ones. Historical records indicate that the first statistical tables began to appear in Sumer, Egypt, ancient China, Babylon, and Assyria. Statistics continues its role with increasing scope and importance through centuries, with especially hard time during the Second World War, when conducting statistical research was prohibited in the German-occupied countries. [However, the compilation of statistics in some countries subjected to the most hostile occupation was conducted, including Poland, where the census was carried out in 1941, by the Underground State]. Currently, we are witnessing how Ukraine gives examples of heroism also in the sphere of official statistics, striving to fulfill its mission of constantly informing state institutions and society despite the extraordinary wartime challenges.

Remarkably, the National Statistical Office of Ukraine has continued to operate since the beginning of the Russian-Ukrainian war. This resilience is a testimony to the essential nature of objective, accurate, reliable, and timely national statistics to inform policy-making, and the steadfastness of the national statisticians behind the numbers. The healthy functioning of a national statistical office has implications for its relationships with bi-lateral and multi-lateral agreements with donor countries and organizations, and therefore, for the security of its country. This year marks the anniversary of the Fundamental Principles of Official Statistics (FPOS, 1992), championed by many esteemed thought leaders, including Józef Oleński (former President of Statistics Poland), Jean-Louis Bodin (INSEE, France), and Katherine Wallman (former, U.S. Chief Statistician and chair, UNECE CES), within the context of the United Nations Economic Commission for Europe Conference of European Statisticians as a way to support the production of national statistics among countries transitioning from centrally planned economies to market economies.

The Fundamental Principles of Official Statistics subsequently was endorsed by the highest body of the UN, the General Assembly (2014). There have been additional efforts to develop aspirational and practical guidance for national statistical offices to strengthen capacity. The European Statistics Code of Practice (2005), U.S. Office of Management and Budget's Statistical Policy Directive 1 (2014) (now embedded in the Evidence Act), OECD's Recommendation of the Council on Good Statistical Practice (2015), American Statistical Association's Ethical Guidance for Statistical Practice (2022) are but a few. There have also been efforts to develop implementation guidance for these principles; notably, FPOS (2011, 2015, and 2020).

After the Second World War, the Marshall Plan was implemented to assist in the reconstruction and strengthening of nation states affected by conflict. Similarly, panelists of the FCSM-2022 session hosted by Jennifer Park, Committee on National Statistics (CNSTAT), discussed a set of the following issues:

- What would a Marshall Plan for national statistics in Ukraine look like?

- How can the parameters of the FPOS and other aspirational guidance inform practical steps of such a plan?
- What roles could various entities take to implement such a plan?
- What elements are essential in the short term? Over the longer term?

In the panel discussion summarized below, representatives of the wide spectrum of international statistical community addressed key aspects of the above questions, taking into account the current situation of statistical institutions and the circumstances in which Ukrainian statisticians try to fulfill their tasks in the conditions of war. The vast majority of the problems and challenges – along with practical ways to deal with them – are presented in the articles by Ukrainian statisticians that make up this issue. Additional information was provided by Oleksandr Osaulenko and his colleagues for the purpose of panel discussion.

## Summary of the panel discussion

As an introduction to the session, its co-organizer (with J. Park of CNSTAT) and chairman, Dominik Rozkrut, characterized briefly the situation of Ukraine's state and society including information on the influx of immigrants from the war zone to Poland. He also quoted some results of the household budget surveys (recently conducted by Statistics Poland) concerning the scope and types of assistance provided to the immigrants. The extent of involvement of Polish households in various forms of help to refugees – such as hosting, food, clothes, other in-kind and in-cash assistance – seems impressive: in total, 78 percent of households, i.e. members of every three out of four dwelling units (about 11.5 out of 15.3 million) participated in one or other types of such assistance.

The international official statistics community has been filled with discussions on timeliness vs granularity of statistics for years. However, events such as a global pandemic or war become a practical test of methodological advancement and organisational solutions, a real test of the organisation's agility and readiness to meet the sudden information needs of societies. The full-scale war in Ukraine became a unique challenge for Polish statistics. The sudden influx of war refugees, a powerful economic shock through drastic increases in energy prices and disruption of value chains posed a severe challenge to Statistics Poland. This challenge is connected both with the need to provide the latest information about the rapidly changing socio-economic processes, but also with the need to anticipate future needs after the end of the war.

Statistics Poland, as was the case with the pandemic, showed a quick response. As early as April 2022, i.e. just after the outbreak of war, additional questions were introduced in the surveys currently carried out or planned, addressed to households, enterprises, social economy entities and non-profit organisations, local government units, and entities operating accommodation facilities. Even earlier (from mid-March 2022),



work was started on a new pilot study addressing refugees at reception points, aimed at characterizing people fleeing the war from Ukraine to Poland. At the same time, Statistics Poland was preparing a plan to use data on Ukrainian citizens residing in Poland from public administration registers and information systems. To provide a legal basis for accessing data from information systems and official registers of public administration, Statistics Poland actively joined the work on the Act and amendments to the Act of March 12, 2022, on assistance to Ukrainian citizens in connection with an armed conflict in the territory of that country as well as preparing the regulation of the Council of Ministers amending the Statistical Work Program for 2022. As a result of these activities, Statistics Poland obtained access to newly established or amended administrative registers, including data on citizens of Ukraine who have come to the territory of the Republic of Poland, social security numbers, education, social assistance, social security, and healthcare insurance. The activities' scope was extensive and covered both the rapid development of research methodology and the enormity of organisational activities related to the implementation of large-scale mass research. As a result, a wide range of statistical research results were quickly obtained.

The statistics of aid and support for Ukraine can serve as the best example here. The invasion on February 24 this year caused, out of concern for their life and health, millions of people in Ukraine decided to leave their country and seek shelter outside its borders, mainly in Poland. Significantly extended social surveys showed that 70.2% of households in Poland granted support to the inhabitants of Ukraine from February 24, 2022, to the end of the first half of 2022, and social economy entities declared 8.0 million recipients of support. From February 24 to March 31, 2022, 28.8 thousand social economy entities (29.6%), including 28.6 thousand non-profit organisations (29.8%) and 0.2 thousand cooperatives (16.9%) (social economy sector), took additional measures to benefit those in need in connection with the war on the territory of Ukraine providing those in need with material support with an estimated value of PLN 511 million and financial support of PLN 140 million. Of 28.8 thousand social economy entities involved, 98.1% conducted activities in Poland and 7.8% in Ukraine. Natural persons were the primary recipients of aid provided by social entities in connection with the war in Ukraine. They were supported by 67.1% of non-profit organisations and 99.1% of cooperatives declaring their commitment to helping. One of the most basic forms of aid in connection with hostilities in Ukraine was a donation in kind (64.2% of non-profit organisations and 37.9% of cooperatives). The estimated value of in-kind support provided by non-profit organisations amounted to PLN 509.3 million, and in the case of cooperatives, PLN 1.8 million (PLN 511.1 million in total).

These are only very brief examples of timely and granular statistics provided based on extending the information scope of the surveys. In total, 17 surveys were expanded

to include new questions used as direct data sources in surveys, of which 16 were modified immediately following the start of the invasion (in spring 2022). Similarly, the same happened in the case of economic surveys, where a focus was put on providing evidence to assess the economic impacts of the war, primarily through monitoring of business tendencies, employment, Inflation, and financial results.

A fine example of an ad-hoc study is a refugee health study created from scratch, designed and implemented jointly with the World Health Organisation. Statistics Poland conducted a pilot study in April and May 2022 among refugees from Ukraine who stayed at reception points in the Podkarpackie region. It was then followed up by a regular survey conducted from June to August 2022. The study covered the way and place of crossing the border; characteristics of people crossing the border due to citizenship, sex, age, and education; planned place/country of stay; intention to work in Poland, taking advantage of education, intent to return to Ukraine after the end of hostilities, health care needs in Poland, access to health care, information on vaccination against COVID-19 and vaccinations for childhood diseases, mental health, the health needs of refugees, and information on their health status in the context of WHO's planning of future assistance for this group of people.

To sum up, the scope of the conducted research and the results obtained were unprecedented internationally. The actions taken and the results obtained were the subject of many international discussions, often indicated as examples of good practices for other countries in the future. In 2023, the UN Statistical Commission introduced information on developing the refugee health survey methodology, which will be further developed jointly by Statistics Poland and WHO to benefit the international community.

Along with mass relocations, also within the territory of Ukraine, institutions and statistical research centers in several areas are emptying, making it impossible to provide data on a regular basis. However, the Ukrainian system of official statistics continues to function and perform its main functions on a scale that can be achieved in wartime conditions only due to the involvement of its devoted staff in headquarters and regions.

### **Main features of official statistics in Ukraine<sup>3</sup>**

The regulatory framework of the state statistical system's functioning is based on the Law of Ukraine "On Official Statistics", issued by the Verkhovna Rada (the Parliament) to be entered into force on January 1, 2023. The law harmonizes the national statistical system with European principles and standards to make it able to produce high quality

---

<sup>3</sup> This section is based on the presentation by W. Okrasa and O. H. Osaulenko "Statistics in troubled times – the case of Ukraine".

statistical information about the economic, social, demographic and environmental situation in Ukraine and its regions. The law is based on the provisions of Regulation (EC) No. 223/2009 of the European Parliament and the Council dated 11.03.2009, which in turn is the basic document within the framework of the implementation of the EU Statistical Compendium and the provisions of the Generic Law on Official Statistics. The Law contains the main provisions of the European Statistics *Code of Practice*.

The State Statistics Service/SSS of Ukraine is a central executive body in the field of statistics – its activities are guided and coordinated by the Cabinet of Ministers of Ukraine. The UA SSS also ensures the development and implementation of state policy in the field of statistics, its offices and staff: 27 regional offices; 6455 employees. According to Ukrainian authorities, national statistical system is reformed and modernized in accordance with the EU/Eurostat principles: 27 (35.1%) of the state statistical observation centers fully meet the requirements of the EU Compendium; 50 (64.9%) of the state statistical observation centers partially meet the requirements of the EU Compendium. The UA State Statistics Service (SSS) strives to fully implement the EU Statistical Requirements Compendium.

## **The challenges of war – the voice of Ukrainian statisticians**

To illustrate the difficulties faced by Ukrainian statisticians, let us quote excerpts from some of the articles contained in this collection:

- *The inability to conduct national statistical surveys makes it difficult to estimate the size of the population due to being limited to existing sources: data from mobile operators, data from administrative registers, and from a special population sample survey, (Volodymyr Sarioglo, Maryna Ogay).*
- *Despite the current extreme situation (...), the CPI must be compiled on an ongoing basis – this is done using Big Data, especially direct cash data, expanding the sample size and improving its design while reducing the burden on respondents and obtaining more reliable transaction price data by incorporating real-time information on household expenditure, (Tetiana Kobylynska, Iryna Legan, Olena Motuzka).*
- *In order to assure operation of the official statistics in Ukraine (under the Martial Law) the involvement of alternative data sources, including Big Data, is necessary. These data should be introduced in parallel or in mix with conventional data sources, to fill the gaps in conventional data due to the war. [Ukraine has an extensive network of private digital services: e.g. Monobank, express delivery “Nova poshta”; mobile phones, social networks, Google analytics, etc. have to be considered too, (Olha Kuzmenko, Hanna Yarovenko, Larysa Perkhun).*

- *The war in Ukraine affects all forms of international economic relations, highlighting the problem of asymmetric economic interdependence in the green transition to climate neutrality, accompanied by raw materials, energy and food crises. The question arises how to minimize the impact of the crisis on the environment as part of getting rid of the carbon footprint of the past (Russian) energy model towards building a sustainable circular ecosystem in Ukraine, (Olga Vasyechko).*
- *The war in Ukraine forced auditors to tackle new challenges due to new risks emerging that need to be recognized, systematized, and treated accordingly – including identification of persons involved in terrorist activities and the proliferation of weapons of mass destruction – while complying with the legal requirements concerning both factors associated with military aggression against Ukraine and those involving compliance with International Standards on Auditing, (Tetyana Chala, Oleksiy Korepanov, Iuliia Lazebnyk, Daryna Chernenko, Georgii Korepanov).*
- *The assessment of the scale and effects of forced external migration of Ukrainians as a result of Russian aggression – based on the data of the State Border Guard – shows that “military emigrants” are, in general, people with higher education than the national average, mainly women who easily adapt to life abroad, especially in Poland (due to the minimal linguistic and cultural differences), (Oleg Krekhivskiy, Olena Salikhova).*

Among the hardships caused by the war, statisticians feel the following the most:

- lack of effective sampling frames and data sources;
  - o production of official statistics continues using administrative data – this allows the assessment of key macroindicators like Ukraine’s GDP, and to publish statistical information on foreign trade in goods, etc.,
  - o regional authorities continue to register prices at the points of sale of goods, which allowed to continue producing the Consumer Price Index in Ukraine as a whole, and by regions,
- respondents are legally deprived of the obligation to provide data during Martial Law;
  - o however, respondents continue to provide primary data as part of a voluntary activity – reporting rate is over 65%,
- regional offices located in temporarily occupied territories or near the military zones may perform their functions only partially, or not at all;
  - o in order to ensure the continuity of the production of official statistics, a back-up system for the collection and processing of data has been established, according to which, for a regional office that is temporarily unable to perform certain statistical tasks, such tasks are delegated to be performed by another office (located in a safer place),

- surveys of household living conditions and demographic data production have been suspended;
- employees of state statistical offices in several regions were forced to migrate (to other regions or abroad);
- frequent air alarms force employees to spend a lot of time in shelters.

## **Research works, trainings, infrastructure base and technology**

Given the extremely difficult circumstances, the tasks performed by Ukrainian statistical institutions during the past 12 months can be considered impressive, embracing 5,292,492 respondent reports processed and the numbers of products based on the statistical research (observations): 19,072 statistical information/reports – 320 Open Data sets – 3,792 press releases – 3,177 data collections – 125,223 users of the “Respondent Account” service – 139 visits to the “Search by USREOU code” service – 180,934 completed international questionnaires based on the results of the state statistical observations – 2,800,936 visits on the SSS official website.

However, drastic cuts in funding resulted in a significant reduction in the scope of the program implemented this year. To be more specific, in 2021, UAH 805,000 (approx. U\$ 22,000) was allocated to two research projects: (i) methodology of conducting sample surveys of the population: “Statistics of income and living conditions in the European Union EU-SILC” (USD 10,000) and (ii) methodology for conducting an integrated survey of short-term enterprise statistics (USD 12,000). In 2022, due to budget constraints, the expenditure on the implementation of these two scientific research works was cancelled. The draft budget for 2023 does not provide for expenses for the implementation of these two projects.

Training and retraining programs cover approximately 1,000 employees per year. In 2021, 269 people were retrained; in 2022 only 15. In 2022, the cost of studying statistics students is UAH 8,150,000 (U\$220,270). Nevertheless, in 2022 there were 110 students, slightly less than in 2021 (120 students).

The infrastructure base and technologies are in a deplorable state. There are practically no sources for the renovation of technical equipment – the last time the fleet of servers and computers was renewed in 2014. The challenges of martial law require an increase in the share of field work as remote work. At the same time, there are practically no laptops, etc. at the headquarters of the UA State Statistical Service and its regional branches. The most urgently needed assistance should include modern technologies for collecting data and creating analytical databases, including “alternative” new sources of information (Big Data, analysis of satellite images, smart statistics etc.)

## Issues raised by panelists and the views expressed

The presentations referred below concerned on the problem of the type and scope of aid for Ukraine from the two complementary points of view – the national organization that engages in international projects, which is American Statistical Association (ASA), by Ron Wasserstein, Executive Director of ASA, and the international organization (IAOS), by Misha Belkindas, President of IAOS – towards establishing needed support and coordinate cooperation between national and international offices.

Focusing his presentation on *The role of statistical societies*, Ron Wasserstein<sup>4</sup> summarized what professional associations do and why it matters to the NSOs, using activities of ASA and some other societies as examples. He concluded some ideas on how associations can help.

Professional societies, such as ASA, conduct a wide spectrum of activities, each of them can provide a platform for arranging for a respective support to UASSS, and eventually other NSOs. These range from facilitating scientific gathering – conferences, networking opportunities – and collaboration in the form of workshops, colloquia interest groups, and knowledge dissemination through meetings and journals to statistical capacity building, including technical training, leadership and communication skills, accreditation/certification, and support education of future.

The indication of this type of activity is also consistent with the results of a survey conducted by ASA among its members, asking them about things they consider fundamental in the activities of a professional society: over 90% of respondents selected meetings and publications. In addition to meetings, professional societies hold smaller gatherings that bring people together to discuss research and methodological interests.

Also, many societies have local or regional groups (ASA calls them “chapters” and has 75 of them) that facilitate gatherings of statisticians. And many have groups organized around statistical topics of interest. Chapters and interest groups (called “Sections” at ASA) often function like smaller versions of the organization, offering their own meetings and networking opportunities, having a newsletter, and so on.

Professional societies serve the extraordinarily important function of disseminating knowledge through meetings and journals. For NSOs, meetings might well be the place where research and methodology can be discussed in an audience of peers in and out of government.

Professional societies serve the function of providing skills to members that they need beyond their formal education. Members provide technical training from the beginning to advanced level through seminars, webinars, workshops, etc. And there

---

<sup>4</sup> *Standard disclaimer:* ASA gives the author time and opportunity to speak at events like this. However, the views expressed here are those of the author and should not be construed as an official statement or position of ASA –R.WJ].

are many non-technical skills the statistician needs that NSS's can provide as well, and often there are few options for getting such training. As one example worth mentioning here is the ASA project which provides training for individuals interested in serving as experts witnesses in the court system.

Another cluster of envisaged forms of possible assistance concerns setting standards and promoting ethical practice (develop and disseminate guidelines for ethical practice); advocating for the profession and for sound practice, and developing relationships with like societies elsewhere.

He also pointed out four key themes for which this is relevant to the NSOs, especially in the context of their creation or reconstruction: a knowledgeable base of support – an independent voice – a source of skilled workers – a source for international connections.

In conclusion, Ron Wasserstein suggested three channels through which statistical societies can help one another: (i) share structures and governance, (ii) share expertise, (iii) share resources. The NSO's staffs need to engage in research, develop as professionals, meet with other statisticians (in government and out), and so on.

### **But there are other reasons why NSS's bring value to NSO's.**

Taking the perspective of an international organization, IAOS, its President, Misha Belkindas, concentrated his presentation, *Building/Redeveloping National Statistical Offices after Conflict*, around the intertwined fundamental issues:

- A. Creation of an international coalition – how to establish it and whom to approach?
- B. What an international coalition should/can do in the case of Ukraine?
- C. What activities are currently going on over there?

First of all, creation of an international coalition should start with identification of potential donors led most likely by the World Bank or a regional development bank, and involve others partners, such as IMF, UNSD, OECD, EU, UN regional commissions, UN specialized agencies (ILO, FAO, others). Also included should be interested countries providing funds and technical assistance (TA), private sector, providers of funds and TA, and international NGOs, such as ISI, IAOS and others. Such a type of endeavors is not unprecedented. The TA project to the countries of the former Soviet Union in the 1990s, with a broad coalition under joint leadership of IMF, WB, UNSD, OECD and Eurostat, can serve as an example. Even more relevant in this context seem to be the projects on strengthening statistical systems of Armenia during the war with Azerbaijan, or in the former Yugoslavia – the case of Bosnia and Herzegovina.

Other projects implemented under the banner of the World Bank encompass a multi-country lending facility STATCAP (started in 2004) with loans to Burkina Faso and Ukraine, followed by loans with grant elements to many countries in all the continents. And the World Bank large lending projects in the Kyrgyz Republic, Tajikistan

and Uzbekistan, also regional lending programs in Africa, including such countries like war-torn Somalia. The Inter-American Development Bank has also organized similar types of projects.

In addition to the needs covered by activities of international organizations with appropriate funds, there are several areas requiring assistance for which NGOs may be suitable implementers. However, this would have to be preceded by the creation of appropriate executive structures enabling NGOs to perform their tasks at the central and regional levels.

Such an international coalition, diversified in its interests and abilities to provide need-adjusted assistance, can help with data collection, processing and production of statistical outputs, including problems related to (i) design a data collection mechanism – surveys with imperfect sampling frames, usage of administrative data, other data, other means of data capture, and (ii) procurement and installation of means/lines for data transmission, storage, manipulation and publication, and (iii) design and start of implementation of HR policies with an emphasis on training and retraining of NSO staff; (iv) initiate the development of young cadre, if needed, develop curriculum for local universities, or in neighboring countries; and (v) train data users – policy makers, journalists, civil society. It would be important to also draft, with assistance of international agencies, a new law on statistics which adheres to international standards, and create a political environment for adherence to Fundamental Principles of Official Statistics.

Taking into account the specificity of the disrupted Ukrainian SSS, and priorities for its reconstruction in accordance with the new Law on Statistics (which the Parliament passed recently), it should be mentioned that the Ukraine Government is familiar with a large-scale international aid: the first institutional building loan to the Government was approved by the World Bank in the early 1990s (with a USD 9 million funding for the NSO). In addition, there was a large-scale international Technical Assistance program TACIS, which Ukraine was one of the recipients. The second loan in the amount of USD 32 million was approved in early 2004 and addressed institutional building, HR, data collection, processing, development of specialized software, etc.

The new project will most likely include the following activities towards rebuilding/refurbishing the SSS and its regional offices: purchase of a large amount of IT equipment (servers, PCs, etc.) – attract new staff to the statistical service – launch a large staff training and retraining program – support local universities and the National Academy of Statistical Education in providing equipment, developing a curriculum and providing trainers.

As regards current activities, IAOS received a request from the Institute of Economic Forecasting of the Ukrainian Academy of Sciences, which was tasked by the



Government to develop methods for the calculation of the damage caused by the war. They want to start from the damage done to the agricultural sector, in particular small-scale farming. IAOS has so far approached the ASA Statisticians Without Borders and received a positive answer. FAO agreed to render assistance by including Ukraine in their AGRIS program. World Bank cannot finance Ukraine from their Trust Fund 50x2030 as the country is not eligible – however, IAOS will continue to help on this and will try to obtain sources from a Trust Fund, at least in-kind.

## Conclusions

Ukrainians, fighting for the preservation of their state, need – apart from military means of defense – reliable information for the state, its institutions and people, for now and for the fundamental reconstruction of their country in the near future.

The above presentations and discussion provide an illustration of the type and amount of work that is expected on this line based on documentation of disruptions in data production and the general functioning of the state statistical service. The voice of Ukrainian experts is the leading for designing an effective strategy of internationally coordinated activities – including statistical capacity building at each level of the state statistical system' units.

The manuscripts submitted by Ukrainian statisticians for this Special Issue prepared jointly by *Statistics in Transition new series* and *Statystyka Ukraina* reflect the type and scale of the problems and challenges faced by Ukrainian statisticians in the conditions of war.

Reconstruction, modernization and strengthening of the national information infrastructure – with state statistical service as its institutional backbone – should become the goal of various initiatives and missions of the international community of statisticians. Starting with showing the areas of destruction and related needs – identified together with experts from Ukraine – is a project that requires more extensive work.

## 4. Appendix

Scientific articles published in the following special issues of *Statistics in Transition new series* in 2015–2023:

1. *Statistics in Transition new series* and *Survey Methodology* Small Area Estimation – Joint Issue Part 1 (post-SAE2014 papers) Volume 16, Number 4, December 2015 <https://sit.stat.gov.pl/Issue/42>.
2. *Statistics in Transition new series* and *Survey Methodology* Small Area Estimation – Joint Issue Part 2 (post-SAE2014 papers) Volume 17, Number 1, March 2016 <https://sit.stat.gov.pl/Issue/41>.
3. *Statistics in Transition new series* Statistical Data Integration – Special Issue Volume 21, Number 4, August 2020 <https://sit.stat.gov.pl/Issue/22>.
4. *Statistics in Transition new series* and *Statistics of Ukraine* A New Role for Statistics: Joint Special Issue Volume 24, Number 1, February 2023 <https://sit.stat.gov.pl/Issue/62>.

**1. *Statistics in Transition new series and Survey Methodology*  
Small Area Estimation – Joint Issue Part 1  
Volume 16, Number 4, December 2015, pp. 485–488<sup>1</sup>**

**From the Guest Editors (Part 1)**

The first part of this Joint Issue of *Statistics in Transition* and *Survey Methodology* includes eight articles. These two issues have been split according to which guest editors have been looking after the articles. They are not necessarily sequenced according to the themes that appeared in the original conference programme.

The first six contributions in this thematic issue of SIT and SMJ represent articles that are firmly methodological in their perspective. The first paper, by J.N.K. Rao provides a unifying perspective for the remaining five contributions. In this review paper, Rao highlights important new developments in SAE since the publication of his encyclopedic 2003 book. As he notes in his abstract, much of this new methodological development has focused on addressing the practical issues that arise when model-based SAE methods are applied in practice. An important dichotomy in this regard follows from the nature of the available data for SAE. Historically, such data have been area level aggregates of one form or another, typically direct sample-based estimates. Issues addressed in Rao's paper then include the choice of appropriate weights for these aggregates as well as methods for dealing with the not uncommon situation where there is a negligible area level variance component in the basic area-level model (the so-called Fay-Herriot model) used to smooth these aggregates across the areas, or where this smoothing model is necessarily non-linear, reflecting a GLM for the underlying survey variable. Issues associated with estimation of both unconditional as well as conditional MSEs of these model-based estimators are also discussed. In the second half of his paper, Rao switches his attention to SAE where unit level data from the small areas of interest are available. This is a fast-growing set of applications, reflecting new capabilities in data collection. Here, the focus is on sample weighting and benchmarking as important requirements for users interested in design consistency of SAE outputs, together with important new developments in dealing with outliers in the survey data, applications to poverty mapping and dealing with informative sampling methods. Model selection and checking is extremely important in the unit level case, and the paper briefly describes some new developments in this regard.

The next three papers in this issue focus on a new methodology for area level SAE. The first, by Bonnery, Cheng, Ha and Lahiri, notes that users of SAE outputs typically

---

<sup>1</sup> Numbering of pages in the original publication.



require more than just estimates of area averages, and are often interested in small area distributions as well as rankings across small areas. In this context, these authors develop a triple goal SAE methodology for US state level unemployment, with estimates structured so that they are simultaneously efficient for estimation of area level average unemployment as well as the empirical distribution of area level unemployment, while also staying as close as possible to the actual ranking of the real small area means. An interesting idea that is discussed in this paper is the fact that in practice it is not just one area average that is of interest, but an “ensemble” of such averages corresponding to the area-level distribution of a characteristic of interest. This immediately leads to a corresponding ensemble of models, which these authors fit using a Bayesian MCMC approach.

The general theme of the usefulness of incorporating time series information in SAE solution is repeated in the paper by van den Brakel and Buelens. Here, though the attention is directed towards appropriate model specification when the estimation must be carried out at regular intervals, using data from repeated surveys and practical considerations rule out survey-specific model optimisation. An approach to covariate selection for small area survey estimates obtained from a repeated survey under a Fay-Herriot specification is defined, with the model specification carried out simultaneously over a number of “editions” of the survey while being constrained to be the same for each edition. The final model is chosen by minimising the average conditional AIC over all the editions, with the small area estimates at each time period computed using a Hierarchical Bayes approach.

The next paper, by Karlberg, switches gears and considers SAE under a unit level model. In particular, in this paper Karlberg addresses two of the difficult issues that arise when the available unit level data are non-negative values drawn from an economic population, as would be the case for a business survey. These conditions often lead to a highly right-skewed distribution of the sample data values, with outliers a not uncommon feature, together with the presence of excess zeros. Both of these data characteristics are not conducive to SAE based on the industry standard linear mixed model for unit level data. Instead, Karlberg combines a log scale linear mixed model for the strictly positive data (to deal with their high skewness) and a logistic model for the presence of zero values (a hurdle model) in order to define a specification for the zero-inflated observed data. Simulation results for SAE based on this approach are promising, but application to a real business survey data set turns out to be disappointing, reflecting the very complex nature of such data. Clearly further research is needed for SAE in business surveys.

The fifth paper, by Franco and Bell, shows how the Fay-Herriot approach can be extended to where the underlying averages are derived from binary survey variables, so that the basic area-level model can be specified as linear on a logit scale. This

model is then combined with time series of aggregates from the small areas, allowing for information to be “borrowed” across both time and space. An application to improving county-level poverty estimates in the SAIPE programme of the US Bureau of the Census is used to demonstrate the efficiency gains of the approach.

The sixth paper, by Luna, Zhang, Whitworth and Piller, represents a fundamental departure from the random area effect-based SAE models that underpin the previous papers. Here, the underlying data consist of historical counts, represented by an out-of-date census (or register)-based cross-tabulation of interest, where one of the dimensions of the tabulation is the area identifier, as well as up-to-date information on margins of the cross-tabulation derived from a current survey. Such data are naturally modelled using a log-linear specification, and the authors consider the use of a generalized SPREE approach to recover the current cross-tabulation. Alternative GSPREE models with increasingly complex interaction structure are investigated and applied to estimation of population counts within ethnic group in small areas in the United Kingdom. Interestingly, these authors report that for these data more complex model specifications do not necessarily lead to improvement in the resulting survey estimates, essentially because the sparse nature of the available data does not allow these more complex models to be adequately fitted.

The last two contributions focus on small area education. Small area estimation is gaining increasing popularity among survey statisticians, economists, sociologists and many others. Unfortunately, small area courses are offered only in a handful of universities and that too just as an elective. However, there is a definite need for small area teaching, and the papers by Burgard and Münnich as well as Golata have addressed this very important issue. The paper by Burgard and Münnich has hit the mark very directly. What the paper emphasizes is that rather than giving a series of lectures on the different small area techniques and the associated theory behind them, it is more important to combine the theory with actual simulations. In this way, students can have hands on experience of the subject as well as are able to make a comparison of the different small area methods which they have learnt. Like Burgard and Münnich, Golata also appreciates very well the need for small area education. To this end, she conducted a survey with participants from both the academics and National Statistical Institutes. Her objective went beyond questions on small area teaching, and enquired several related pertinent questions such as risks encountered in applying SAE as well as important sources on SAE developments. The results of her survey are listed in a series of tables and graphs to provide the reader with a better understanding of the state of the art.

Several persons (in addition to the Editor and Guest Editors) have served as reviewers of papers published in this thematic issue of the journal: we would like to thank all the authors for taking the time to turn their SAE 2014 presentations into

the interesting and thought provoking papers published here. We acknowledge the efforts of Giovanna Ranalli, Nicola Salvati, Hukum Chandra and Timo Schmid, who helped review the first six papers: their encouraging and productive comments directly contributed to their obvious quality.

Raymond Chambers and Malay Ghosh  
Guest Editors

**1. Statistics in Transition new series and Survey Methodology  
Small Area Estimation – Joint Issue Part 1  
Volume 16, Number 4, December 2015**

Post-conference papers: Small Area Estimation conference, Poznań, 3rd-5th September, 2014

**The issue consists of the following articles:**

J. N. K. Rao, *Inferential issues in model-based small area estimation: some new developments* <https://doi.org/10.59170/stattrans-2015-026>.

Daniel Bonnéry, Yang Cheng, Neung Soo Ha, Partha Lahiri, *Triple-goal estimation of unemployment rates for U.S. states using the U.S. Current Population Survey data* <https://doi.org/10.59170/stattrans-2015-027>.

Jan A. van den Brakel, Bart Buelens, *Covariate selection for small area estimation in repeated sample surveys* <https://doi.org/10.59170/stattrans-2015-028>.

Forough Karlberg, *Small area estimation for skewed data in the presence of zeroes* <https://doi.org/10.59170/stattrans-2015-029>.

Carolina Franco, William R. Bell, *Borrowing information over time in binomial/logit normal models for small area estimation* <https://doi.org/10.59170/stattrans-2015-030>.

Angela Luna, Li-Chun Zhang, Alison Whitworth, Kirsten Piller, *Small area estimates of the population distribution by ethnic group in England: a proposal using structure preserving estimators* <https://doi.org/10.59170/stattrans-2015-031>.

Jan Pablo Burgard, Ralf Münnich – *SAE teaching using simulations* <https://doi.org/10.59170/stattrans-2015-033>.

Elżbieta Gołata, *SAE education challenges to academics and NSI* <https://doi.org/10.59170/stattrans-2015-034>.

Raymond Chambers, Malay Ghosh, *From the Guest Editors (Part 1)*.

**2. Statistics in Transition new series and Survey Methodology**  
**Small Area Estimation – Joint Issue Part 2**  
**Volume 17, Number 1, March 2016, pp. 3–6<sup>2</sup>**

**From the Guest Editors (Part 2)**

The second part of this Joint Issue of Statistics in Transition and Survey Methodology includes seven articles. These two issues have been split according to which guest editors have been looking after the articles. They are not necessarily sequenced according to the themes that appeared in the original Conference programme.

The first paper, by Erciulescu and Fuller, presents a small area procedure where the mean and variance of an auxiliary variable are subject to estimation error. They consider fixed and random specifications for these auxiliary variables. Their study was motivated by a situation where the sample used for small area estimation was a subsample of a larger survey. The larger survey furnished estimates of the distribution of the auxiliary variables. They demonstrate that efficiency gains associated with the random specification for the auxiliary variable measured with an error can be obtained. They propose a parametric bootstrap procedure for the mean squared error of the predictor based on a logit model. The resulting bootstrap procedure has a smaller bootstrap error than a classical double bootstrap procedure with the same number of samples.

The second paper, by Münnich, Burgard, Gabler, Ganninger and Kolb, develops a sampling design that can support accurate estimation for the 2011 German Census. In contrast to carrying out a classical census, a register-assisted census, using population register data and an additional sample, was implemented. The main objective of the census was to produce the total population counts at fairly low levels of geography. Ralf Münnich et al. provide an overview of how the sampling design recommendations were set up to fulfill legal requirements and to guarantee an optimal, yet flexible, source of information. Small area methods, as well as traditional methods, were used to produce these counts. Empirical results of the small area estimation are presented.

The next three papers present developments in small area estimation methodology and practical application in various fields of empirical research and statistics production, including poverty research and fisheries statistics. The first paper, by Guadarrama, Molina and J. N. K. Rao, provides a review on methods for the estimation of poverty indicators for small areas, including design-based direct estimation and a number of model-based small area estimation methods: the Fay-Herriot area level model,

<sup>2</sup> Numbering of pages in the original publication.





the World Bank poverty mapping method (the ELL method) and three Bayesian variants previously published by the authors. These are the empirical best/Bayes (EB) and hierarchical Bayes (HB) methods and a Census EB method providing an extension of the EB method. While the Fay-Herriot method employs area-level data, the other methods require unit-level auxiliary information. The ELL, EB, Census EB and HB methods rely on statistical data infrastructures where access to unit-level records of population units taken for example from administrative registers and population censuses is available for research and statistics production. This option is becoming frequently met in an increasing number of countries and much of current small area research is conducted under this assumption. The list of advantages and disadvantages, reported for each of the methods, appears helpful for practitioners facing the challenge of choosing a small area method for a particular estimation task. Statistical properties (bias and accuracy) of methods are assessed empirically by model-based simulation experiments with unit-level synthetic data following a nested error model, throwing further light on the methodological summaries of the methods. Extensive simulation scenarios of varying complexity include informative sampling and a nested error model with outliers; these scenarios in particular are important for practical purposes. For practical application, it is important that also situations are considered where some of the underlying assumptions of the methods do not hold, which is often the case in practice. The conclusions drawn by the authors on the relative performance of the methods are useful for researchers and practitioners.

Because of its applicability in various data infrastructures, the Fay-Herriot model has been widely used in small area estimation purposes all over the world and new developments are often needed to extend the method for practical situations at hand. A robust hierarchical Bayesian approach for the Fay-Herriot area-level model is presented in the second paper, written by Chakraborty, Datta and Mandal. The starting point is the authors' observation on a possible poor performance of the standard Fay-Herriot area-level model in the presence of outliers. The new method is aimed for cases where extreme values are met for some of the random effects of small area means, causing problems in the standard Fay-Herriot procedure under normality assumptions of the random effects. The authors propose a two-component normal mixture model, which is based on noninformative priors on the model variance parameters, regression coefficients and the mixing probability. The method is aimed as an alternative to a scale mixture of normal distributions with known mixing distribution for the random effects. The authors apply their method to real data of US Census Bureau for poverty rate estimation at county level. The results indicate that probabilities of having large random effects are expected to be low for most areas but can be large for some areas, thus calling for attention to handle the

possible heterogeneity of the data. Simulation studies based on artificially generated data are conducted to assess the performance of the proposed method against the standard Fay-Herriot model. In the first set of experiments, the authors verify the robustness of the proposed method to outliers in the cases considered. In further simulations, the authors show that their method tends to perform better than the Fay-Herriot method when the possibility of presence of outliers is high, and performs similarly in situations where outliers are not expected. In their concluding notes the authors provide a useful discussion on the possible causes of exceptionally large random effects for certain areas, calling for a careful specification of the linking model and the choice of the explanatory (auxiliary) variables.

The third paper, by Hernandez-Stumpfhauser, Breidt and Opsomer, provides a refinement of the Fay-Herriot approach for a particular small area estimation problem. The authors consider a practical problem of developing a new weighting procedure for a regular fisheries survey in the United States on recreational fishing in saltwater. For the estimation of the recreational catch, fishing catch per trip is estimated from one survey and the number of fishing trips from another survey. Data from these two surveys are combined to estimate recreational fishing catch in 17 US states. For weighting procedure, estimates are needed for the fraction of fishermen who leave the fishing site during a prespecified time interval on a selected day. The distribution of daily departure times is needed within spatio-temporal domains subdivided by mode of fishing. Direct estimates could be obtained but they are not sufficient because of a large number of estimation domains, causing very small (even zero) domain sample sizes. The authors develop a small area estimation solution based on the Fay-Herriot approach. More specifically, the authors show that with a certain hierarchical model formulation that is slightly more complex as the standard mixed model, fast and accurate model selection procedure based on variational/Laplace approximation to the posterior distribution can be implemented for the particular estimation problem considered. Even if the underlying linear mixed model can be complex involving fixed and random effects for the states, waves and fishing modes and interaction terms, the method can serve as a cost-effective alternative to the computationally more demanding MCMC sampler. By empirical comparison of MCMC and the proposed variational/Laplace approaches using real data, the authors show that the results are essentially identical, thus motivating the use of the method in practice.

The production of small area statistics by national statistical agencies and international statistical institutes is becoming more and more important for societal planning and evaluation and the allocation of public funds to regional areas and other population subgroups. In the next paper, Kordos presents a personal view on the development of certain aspects of small area estimation methodology and practice

in the context of official statistics. The author first summarizes the main approaches in small area estimation with some historical remarks. He continues by discussing the important issue of the use of administrative records in official statistics production and as auxiliary information in the construction of estimators for various regional indicators. The author presents a summary of international conferences on small area estimation organized in past years, covering a period from 1985. Further, he presents a review of selected international small area estimation programs and research projects on small area estimation. A special property of these research activities is that they are conducted in cooperation with research communities on small area estimation and actors whose responsibility is in the production of official small area statistics. The interaction has proven fruitful in motivating ongoing research and development in small area estimation methodology and for boosting the implementation of methods in regular official statistics production. This aspect might well be taken as the main message of the paper by Kordos.

In many national statistical institutes, the design-based approach has offered the prevailing paradigm in official statistics production for decades. Good reasons are the ability of the approach to provide estimates having favorable statistical properties such as design-unbiasedness, which is often appreciated by the clients, and the availability of powerful statistical procedures and tools that use effectively the auxiliary information supplied in various forms. Calibration techniques and generalized regression estimation are examples of such methods. While relative standard errors of design-based estimates can be sufficiently small for population domains whose sample size is large, this is not necessarily the case for small domains. It is in this field of action where model-based small area estimation is challenging the design-based approach. In the final paper, Hidirolou and Estevao present an empirical assessment of selected design-based methods against some existing model-based area estimation small methods, considered at Statistic Canada. Traditional design-based estimators include the Horvitz-Thompson estimator, two variants of calibration estimators and a modified regression estimator. A synthetic estimator and the standard EBLUP and its variant called pseudo-EBLUP represent model-based methods. The relative performance of the methods is assessed in design-based simulation experiments, where in addition to "ideal" conditions also misspecified models are considered. The relative performance of the methods differs depending on whether the model holds or not. Of the traditional design-based estimators, the domain-specific calibration estimator and the modified regression estimator indicate the best efficiency. The model-based small area estimators tend to outperform the design-based methods in efficiency, especially for small domains. As expected, the model-based methods can suffer from large design bias in cases where the model is misspecified.

Several persons (in addition to the Editor and Guest Editors) have served as reviewers of papers published in this thematic issue of the journal. We acknowledge the efforts of F. Jay Breidt, Isabel Molina, Domingo Morales, Ari Veijanen, Mamadou Diallo and Jon Rao: their encouraging and productive comments directly contributed to the quality of the papers.

Risto Lehtonen and Graham Kalton  
Guest Editors

**2. *Statistics in Transition new series and Survey Methodology*  
Small Area Estimation – Joint Issue Part 2  
Volume 17, Number 1, March 2016**

Post-conference papers: Small Area Estimation conference, Poznań, 3rd-5th September, 2014.

**The issue consists of the following articles:**

Andreea L. Erciulescu, Wayne A. Fuller, *Small area prediction under alternative model specifications* <https://doi.org/10.59170/stattrans-2016-001>.

Ralf Münnich, Jan Pablo Burgard, Siegfried Gabler, Matthias Ganninger, Jan-Philipp Kolb, *Small area estimation in the German Census 2011* <https://doi.org/10.59170/stattrans-2016-002>.

María Guadarrama, Isabel Molina, J. N. K. Rao, *A comparison of small area estimation methods for poverty mapping* <https://doi.org/10.59170/stattrans-2016-003>.

Adrijo Chakraborty, Gauri Sankar Datta, Abhyuday Mandal, *A two-component normal mixture alternative to the Fay-Herriot model* <https://doi.org/10.59170/stattrans-2016-004>.

Daniel Hernandez-Stumpfhauser, F. Jay Breidt, Jean D. Opsomer, *Variational approximations for selecting hierarchical models of circular data in a small area estimation application* <https://doi.org/10.59170/stattrans-2016-005>.

Jan Kordos, *Development of small area estimation in official statistics* <https://doi.org/10.59170/stattrans-2016-006>.

Michael A. Hidiroglou, Victor M. Estevao, *A comparison of small area and calibration estimators via simulation* <https://doi.org/10.59170/stattrans-2016-007>.

Graham Kalton, Risto Lehtonen, *From the Guest Editors (Part 2)*.

**3. *Statistics in Transition new series*  
 Statistical Data Integration – Special Issue  
 Volume 21, Number 4, August 2020, pp. III–VI<sup>3</sup>**

**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.

<sup>3</sup> Numbering of pages in the original publication.



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.

Włodzimierz Okrasa

Editor

### 3. *Statistics in Transition new series* **Statistical Data Integration – Special Issue** **Volume 21, Number 4, August 2020, pp. II<sup>4</sup>**

#### **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 compiled 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

---

<sup>4</sup> Numbering of pages in the original publication.



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 Włodzimierz 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 emphasis 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](http://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 *SiTns*. 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 Suntornc host 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 multi-purpose 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-in-Chief

**3. Statistics in Transition new series**  
**Statistical Data Integration – Special Issue**  
**Volume 21, Number 4, August 2020**

**The issue consists of the following articles:**

**Invited paper:** Malay Ghosh, *Small area estimation: its evolution in five decades*  
<https://doi.org/10.21307/stattrans-2020-022>.

Julie Gershunskaya, *Discussion* <https://doi.org/10.21307/stattrans-2020-023>.

Ying Han, *Discussion* <https://doi.org/10.21307/stattrans-2020-024>.

Isabel Molina, *Discussion* <https://doi.org/10.21307/stattrans-2020-026>.

David Newhouse, *Discussion* <https://doi.org/10.21307/stattrans-2020-027>.

Danny Pfeffermann, *Discussion* <https://doi.org/10.21307/stattrans-2020-028>.

J. N. K. Rao, *Discussion* <https://doi.org/10.21307/stattrans-2020-029>.

Malay Ghosh, *Rejoinder* <https://doi.org/10.21307/stattrans-2020-030>.

**Papers dealing with Small Area Estimation**

Song Cai, J. N. K. Rao, Laura Dumitrescu, Golshid Chatrchi, *Effective transformation-based variable selection under two-fold subarea models in small area estimation*  
<https://doi.org/10.21307/stattrans-2020-031>.

Andre Felipe Azevedo Neves, Denise Britz do Nascimento Silva, Fernando Antonio da Silva Moura, *Skew normal small area time models for the Brazilian annual service sector survey* <https://doi.org/10.21307/stattrans-2020-032>.

### **Papers dealing with advances in probabilistic record linkage and analysis of linked data**

Loredana Di Consiglio, Tiziana Tuoto, *A comparison of area level and unit level small area models in the presence of linkage errors* <https://doi.org/10.21307/stattrans-2020-033>.

Sabyasachi Bera, Snigdhanu Chatterjee, *High dimensional, robust, unsupervised record linkage* <https://doi.org/10.21307/stattrans-2020-034>.

### **Papers dealing with statistical methods for longitudinal data, merged data and data fusion**

Takumi Saegusa, *Confidence bands for a distribution function with merged data from multiple sources* <https://doi.org/10.21307/stattrans-2020-035>.

Xuze Zhang, Saumyadipta Pyne, Benjamin Kedem, *Model selection in radon data fusion* <https://doi.org/10.21307/stattrans-2020-036>.

Daniel Bonn ery, Yang Cheng, Partha Lahiri, *An evaluation of design-based properties of different composite estimators* <https://doi.org/10.21307/stattrans-2020-037>.

### **Papers dealing with synthetic data for microsimulations, disclosure avoidance and multi-purpose inference**

Jan Pablo Burgard, Hanna Dieckmann, Joscha Krause, Hariolf Merkle, Ralf Munnich, Kristina M. Neufang, Simon Schmaus, *A generic business process model for conducting microsimulation studies* <https://doi.org/10.21307/stattrans-2020-038>.

M. Jahangir Alam, Benoit Dostie, Jorg Drechsler, Lars Vilhuber, *Applying data synthesis for longitudinal business data across three countries* <https://doi.org/10.21307/stattrans-2020-039>.

Partha Lahiri, Jiraphan Suntornchost, *A general Bayesian approach to meet different inferential goals in poverty research for small areas* <https://doi.org/10.21307/stattrans-2020-040>.

Włodzimierz Okrasa, From the *Editor*

Partha Lahiri, *Preface*

***Statistics in Transition new series and Statistics of Ukraine***  
**A New Role for Statistics: Joint Special Issue**  
**Volume 24, Number 1, February 2023, p. I<sup>5</sup>**

## **Preface**

This volume is the result of the need of the moment – similarly felt by both of us, the undersigned – to meet the demand of the international statistical community for first-hand knowledge of the multiple consequences of the war in Ukraine for the functioning of the national statistical system as well as for statistics as a discipline, and as “statistics without borders”. On behalf of the editorial offices and scientific boards and committees of the *Statistics in Transition new series* and *Statistics of Ukraine*, about half a year ago we invited researchers and practitioners to submit manuscripts to a joint Special Issue devoted to statistical data production in wartime conditions.

From the descriptions of situations in which the national statistical system functions, including evidence on organizational and methodological problems and challenges, presented in this volume, a vision for a new role of statistics and statisticians emerges as important participants in ongoing processes.

We present this volume to the Readers in the hope that it will shed light on these issues and draw attention to those that require immediate attention and reflection by members of the international statistical community.

Włodzimierz Okrasa  
Editor-in-Chief  
*Statistics in Transition new series*

Oleksandr H. Osaulenko  
Editor-in-Chief  
*Statistics of Ukraine*

---

<sup>5</sup> Numbering of pages in the original publication.

#### **4. *Statistics in Transition new series and Statistics of Ukraine* A New Role for Statistics: Joint Special Issue Volume 24, Number 1, February 2023, pp. IX–XIV<sup>6</sup>**

##### **From the Editors**

The tragic events currently taking place in Ukraine have affected all aspects of life and activity, in private and public spheres, including an unspeakably difficult situation of the state statistics services.

The presented to the readers issue, entitled *A New Role for Statistics*, is the product of the jointly undertaken task by *Statistics in Transition new series* and *Statistics of Ukraine* to showing some of the enormity of problems experienced by the statisticians of Ukraine and the ways they are dealing with them. As first-hand accounts, articles by Ukrainian statisticians also provide information about the disruptions and types of assistance expected as well. A large part of possible reactions from the international community of statisticians has already been preliminarily identified and addressed in the opening of this volume basing on summary of the presentations that panelists representing various types of institutions and organizations gave at the session devoted to these issues within the last FCSM2022 conference (see *The post-conflict reconstruction of the statistical system in Ukraine...*).

This issue contains 15 papers, which focus on functioning of statistical system in war conditions demonstrating the role of statistics in documenting the effects of Russian aggression on the economy and society of the invaded country. Particular emphasis is put on the humanitarian crisis and the degradation of people's well-being, and on challenges faced by statisticians along with new tasks and approaches to overcome them.

This Joint Special Issue gives us also the opportunity to express our appreciation and thanks for all our contributors: authors, reviewers and all the participants of the editorial process.

This Joint Special Issue starts with the paper *Problems relating to the statistical research of the national market of logistics services in war conditions* by Nataliia Hrynychak, Olha Yatsenko, Olena Bulatova, and Olena Ptashchenko. The article discusses the theoretical principles of statistical research with regard to the national market of logistics services during wartime, and identifies the main structural changes that occurred due to the hostilities faced by the country. The authors determine the main factors influencing the functioning of the analysed market during war, as the statistical study of these factors is considered necessary for the transformation and

---

<sup>6</sup> Numbering of pages in the original publication.

© Włodzimierz Okrasa, Oleksandr H. Osaulenko. Article available under the CC BY-SA 4.0 licence 

development of logistics services. According to the results of the evaluation, analysis and structuring of relevant indicators and factors affecting the development of the logistics services market, their priority is determined according to the type of logistics services, which makes it possible to identify new opportunities for development both at the micro- and at the macro level.

The article entitled *Using Big Data by Ukrainian official statistics when martial law applies: problems and solutions* by Oleksandr H. Osaulenko and Olena Horobets focuses on issues of the secure operation of official statistics in Ukraine during the application of martial law. The level of digitalisation in Ukraine as the basis for using Big Data was analysed by the proposed indices of internetisation, social progress and digital transformation, and several problems (methodological, legal, financial, and managerial) were identified as vital for statistical offices on their way to the implementation of Big Data in statistical processes. proposals concern tools for Big Data processing, The authors discuss the proposals such as Data Hypercube as a way for presenting Big Data for their visualisation, applications of Web scraping in estimating the consumer prices index, analyses of labour and real estate markets, and the applications of specialised software for the collection, processing and analysis of Big Data sets.

Nataliia Reznikova, Iryna Zvarych, Roman Zvarych, and Ivashchenko Oksana in their paper *The impact of the Russian-Ukrainian war on the green transition and the energy crisis: Ukrainian scenario of circular economy development* analyse how to minimise the impact of the energy crisis on the environment as one of the ways of getting rid of carbon footprints resulting from the growth of the russian energy and building a circular sustainable ecosystem in Ukraine. The paper determines the impact that the war has on the practice of applying resource nationalism associated with a wide variety of modern global problems. It also identifies the dominant diversification tendencies in the EU in terms of the circularity of the economy. The proposed concept of a global inclusive circular economy can be considered as a complex multidimensional system, whose main components are based on the economic, sociological, environmental and circular aspects of life.

The next article *A statistical study of climate change in Ukraine under martial law* by Tetiana Kobylinska, Iryna Legan, and Olena Motuzka presents the development of theoretical and methodological foundations of statistical research in the field of national environmental and economic accounting, which forms the basis for the development of indicators of climate change under martial law and shapes the adaptation to these changes. The paper studies issues of producing ecological information relating to Ukraine according to statistical data, and describes the main problems which arise during the construction of national environmental accounts were characterised. The article identified the key factors which influence to the largest extent the quality of



statistical data and calculations, and which are necessary for the transformation and development of the statistical estimation of climate change under Russian military aggression.

Olha Lubenchenko, Svitlana Shulga, and Halyna Pavlova discuss *Method of auditing in conditions of martial law*. The authors consider methodical recommendations on the actions of auditors during martial law that relate to such stages of the audit as the preparatory phase, the planning phase, the task implementation and the final phase. Under martial law, new risks are emerging, systematized by the authors and related to the identification of persons involved in terrorist activities and the proliferation of weapons of mass destruction. The paper has been developed also to assess ethical threats in the light of martial law. The war in Ukraine has forced auditors to tackle new challenges in complying with the latest legal requirements for identifying those involved in military aggression against Ukraine, on the one hand, and requiring careful compliance with International Standards on Auditing.

In the next manuscript entitled *Current challenges related to the consumer price index (CPI) in Ukraine* Olga Vasyechko analyses how to contribute to the maintenance and compilation of the consumer price index (CPI) in the current extreme situation caused by the Russian military aggression against Ukraine. The interaction between the ideal and conditional concepts of the index and their practical implementation is considered as a potential source of compilation improvement. The author argues that the main factor of the modern criticism of the CPI is the systematic deviation of the practical form of the index from its theoretical foundations. The revision of the paradigm of primary data sources allows for a significant reduction in the methodological and organizational limitations imposed by the extreme conditions of Russia's military aggression against Ukraine. In the conditions caused by the war, this kind of information allows regular estimates of the consumer price index for a large number of goods without the loss of quality, and control the structure of consumption both in general and by region, and opens prospects for reducing discrepancies between conventional concept of the CPI, its ideal concepts and their practical application.

Volodymyr Sarioglo and Maryna Ogay's article presents *Approach to population estimation in Ukraine using mobile operators' data* discussing the task of developing effective approaches to estimating the population size using data from existing sources, in particular the data of mobile operators regarding the number, location and mobility of subscribers. The article highlights the results of a study on the use of data from mobile operators, data from administrative registers, and the results of a special population sample survey on the use of mobile communication for the purpose of estimating the population. It also provides the results of experimental calculations of the population size in Ukraine as a whole and in particular regions. The developed approaches can be used to assess and monitor the number and location of the population

of Ukraine, provided the availability and proper preparation of data of mobile operators, the availability of administrative records containing information about the population, the availability of sample surveys, in particular on the peculiar use of mobile communications by the population.

Taisiia Bondaruk, Liudmyla Momotiuk, and Iryna Zaichko focus on *Budgetary policy of Ukraine in time of challenges and its impact on financial security*. The aim of the study is to deepen the theoretical and methodological foundations of the creation and implementation of budgetary policy in Ukraine, evaluation of its impact on the financial security in time of challenges. The study uses methods of comparative analysis, grouping in the process of evaluating the current state of budgetary policy indicators, methods of normalization and standardization of data, modelling, and graphical analysis of data for normalizing the financial security indicators and determining the dynamics of financial security components. The materials and reports containing statistical data from the Ministry of Finance of Ukraine and the State Statistics Service of Ukraine served as the basis of the study. It was determined that the components of the state's financial security in the face of martial law and pandemic do not take into account the impact of budgetary policy. Therefore, in the course of comprehensive integrated assessment of the financial security of the state, additional indicators were proposed.

The paper by Tetyana Chala, Oleksiy Korepanov, Juliia Lazebnyk, Daryna Chernenko, and Georgii Korepanov deals with *Statistical modelling and forecasting of wheat and meslin export from Ukraine using the singular spectra analysis*. The article presents the problems related to the functioning of the worldwide market of wheat and meslin. The structure of wheat export by Ukrainian regions is analysed in comparison with the total export. The localisation coefficient is applied to measure the regional unevenness of the distribution of wheat export volumes and the total export by regions of the country. The modelling and forecasting of the volumes and prices of export of wheat and meslin from Ukraine are based on Singular Spectrum Analysis. The study particularly focuses on the individual components of time series, such as trend, annual, semi-annual, four-month, three-month seasonal components. The reliability of the forecast is confirmed by the calculation of the MAPE forecast error and Henry Theil's inequality coefficient. The article proposes an algorithm for calculating the relative indicators of the structure for the individual components of the reconstructed time series, identified through the singular spectral analysis.

The next article prepared by Halyna Holubova *A comparative analysis of the principal component method and parallel analysis in working with official statistical data* describes the basic conceptual approaches to the definition of principle components. Moreover, the methodological principles of selecting the main components are presented. A comparative analysis of the eigenvalues was performed by means of two methods: the Kaiser criterion and the parallel Horn analysis on the example of several

data sets. The study shows that the method of parallel analysis produces more valid results with actual data sets. The author believes that the main advantage of Parallel analysis is its ability to model the process of selecting the required number of main components by determining the point at which they cannot be distinguished from those generated by simulated noise. The Parallel analysis method uses multiple data simulations to overcome the problem of random errors. This method assumes that the components of real data must have greater eigenvalues than the parallel components derived from simulated data which have the same sample size and design, variance and number of variables.

Oleg Krekhivskiy and Olena Salikhova in their manuscript consider *A new industrial strategy for Europe – new indicators of the results of its implementation*. The paper discusses the experiences resulting from EU's adoption and implementation of a wide variety of policy measures in response to the COVID-19 crisis. These measures included stimulating the relocation and expansion of manufacturing to reduce vulnerability, depending on imports, ensuring the stability and development of industrial production. The study proposes and tests a new approach to assessing the consequences of relocation policies aimed at developing the local production potential, increasing the value added by activity, and expanding the share of local value added in industry exports. The manuscript focuses on the formation of statistical analysis tools for assessing the changes of the specialisation and identifying the country's comparative advantages. The authors propose new indicators: RSP – coefficient of Revealed Specialisation of Production, CAVA – coefficient of Comparative Advantage in Value Added by Activity and EVA – coefficient of Comparative Advantages in the Domestic Value Added Exports.

The paper entitled *Assessing the maturity of the current global system for combating financial and cyber fraud* by Olha Kuzmenko, Hanna Yarovenko, and Larysa Perkhun assesses the maturity of systems for counteracting financial and cyber fraud with the view of their future integration at global-level. The calculations made by the authors were based on indicators for 76 countries, which characterized each country's level of cybersecurity and its ability to combat financial fraud in 2018. The authors conducted a bifurcation analysis of the maturity of current global system for combating financial and cyber fraud and produced its phase portraits. It was found to be mature ("Government Efficiency Index – Ease of Doing Business" and "Ease of Doing Business – Crime Index") and insufficient mature ("Government Efficiency Index – Crime Index"), with the components' imbalance indicating high system's sensitivity to react on changes. The constructed 'Equilibrium States' phase portraits showed non-equilibrium phase portraits of the 'saddle' type. The obtained results made it possible to identify determinants of a global integrated system's instability to combat financial and cyber fraud.

Ella Libanova and Oleksii Pozniak in their paper *War-driven wave of Ukrainian emigration to Europe: an attempt to evaluate the scale and consequences (the view of Ukrainian researchers)* evaluate the scale and consequences of the emigration of Ukrainians triggered by the military aggression of the Russian Federation. The paper also attempts to determine the composition of the refugees. According to the estimation of the Ptukha Institute for Demography and Social Studies of the National Academy of Sciences of Ukraine based on the data from the State Border Guard Service, the number of ‘refugees from the war in Ukraine’ reached 3 million as of the end of June 2022. The potential amount of irreversible migration losses, depending on the military and economic factors, ranges from 600–700 thousand to 5–5.5 million people. Considering the fact that approximately 3 million Ukrainians had already been staying (working) abroad before 2022, the war is likely to result in a demographic catastrophe for Ukraine, whose demographic potential has been utterly exhausted.

The article prepared by Maryna Puhachova and Oleksandr Gladun entitled *Using electronic registries to study the COVID-19 pandemic and its consequences* analyses systems of electronic information resources (registers and databases) in the field of the healthcare in different countries. These systems provide information to support the treatment of patients, and also also accumulate large amounts of statistics, thus enabling their qualitative operational analysis. The authors summarise information on the use of electronic registers and databases to create an information base for the study of the COVID-19 pandemic and its consequences in different countries, and formulate proposals for the improvement of electronic health systems in Ukraine. On the basis they propose a list of electronic registers that can significantly improve the analysis of both, the course and the consequences of the coronavirus disease.

Deepika Rajoriya and Diwakar Shukla’s manuscript *Under military war weapon support the economic bond level estimation using generalized Petersen graph with imputation* presents a sample based estimation methodology for estimating the mean economic bond value among countries involved in the military support or business. The problem is derived from current Russia-Ukraine war situation. A node sampling procedure is proposed whose bias, mean-squared error and other properties are derived. Results are supported with empirical studies. Findings are compared with particular cases and confidence intervals are used as a basic tool of comparison. Pattern imputation is used together with a new proposal of CI-Imputation method who has been proved useful for filling the missing value, specially when secret economic support data from involved countries found missing.

Włodzimierz Okrasa  
Editor-in-Chief  
*Statistics in Transition new series*

Oleksandr H. Osaulenko  
Editor-in-Chief  
*Statistics of Ukraine*

**4. Statistics in Transition new series and Statistics of Ukraine  
A New Role for Statistics – Joint Special Issue  
Volume 24, Number 1, February 2023**

**The issue consists of the following articles:**

Dominik Rozkrut, Włodzimierz Okrasa, Oleksandr H. Osaulenko, Misha V. Belkindas, Ronald L. Wasserstein, *The Post-Conflict Reconstruction of the Statistical System in Ukraine. Key Issues from an International Perspective* <https://doi.org/10.59170/stattrans-2023-001>.

Nataliia Hrynychak, Olha Yatsenko, Olena Bulatova, Olena Ptashchenko, *Problems relating to the statistical research of the national market of logistics services in war conditions* <https://doi.org/10.59170/stattrans-2023-002>.

Oleksandr H. Osaulenko, Olena Horobets, *Using Big Data by Ukrainian official statistics when martial law applies: problems and solutions* <https://doi.org/10.59170/stattrans-2023-003>.

Nataliia Reznikova, Iryna Zvarych, Roman Zvarych, Oksana Ivashchenko, *The impact of the Russian-Ukrainian war on the green transition and the energy crisis: Ukrainian scenario of circular economy development* <https://doi.org/10.59170/stattrans-2023-004>.

Tetiana Kobylinska, Iryna Legan, Olena Motuzka, *A statistical study of climate change in Ukraine under martial law* <https://doi.org/10.59170/stattrans-2023-005>.

Olha Lubenchenko, Svitlana Shulga, Halyna Pavlova, *Method of auditing in conditions of martial law* <https://doi.org/10.59170/stattrans-2023-006>.

Olga Vasyechko, *Current challenges related to the consumer price index (CPI) in Ukraine* <https://doi.org/10.59170/stattrans-2023-007>.

Volodymyr Sarioglo, Maryna Ogay, *Approach to population estimation in Ukraine using mobile operators' data* <https://doi.org/10.59170/stattrans-2023-008>.

Taisiia Bondaruk, Liudmyla Momotiuk, Iryna Zaichko, *Budgetary policy of Ukraine in time of challenges and its impact on financial security* <https://doi.org/10.59170/stattrans-2023-009>.

Tetyana Chala, Oleksiy Korepanov, Iuliia Lazebnyk, Daryna Chernenko, Georgii Korepanov, *Statistical modelling and forecasting of wheat and meslin export from Ukraine using the singular spectral analysis* <https://doi.org/10.59170/stattrans-2023-010>.

Halyna Holubova, *A comparative analysis of the principal component method and parallel analysis in working with official statistical data* <https://doi.org/10.59170/stattrans-2023-011>.

Oleg Krekhivskiy, Olena Salikhova, *A new industrial strategy for Europe – new indicators of the results of its implementation* <https://doi.org/10.59170/stattrans-2023-012>.

Olha Kuzmenko, Hanna Yarovenko, Larysa Perkhun, *Assessing the maturity of the current global system for combating financial and cyber fraud* <https://doi.org/10.59170/stattrans-2023-013>.

Ella Libanova, Oleksii Pozniak, *War-driven wave of Ukrainian emigration to Europe: an attempt to evaluate the scale and consequences (the view of Ukrainian researchers)* <https://doi.org/10.59170/stattrans-2023-014>.

Maryna Puhachova, Oleksandr Gladun, *Using electronic registries to study the COVID-19 pandemic and its consequences* <https://doi.org/10.59170/stattrans-2023-015>.

Deepika Rajoriya, Diwakar Shukla, *Under military war weapon support the economic bond level estimation using generalized Petersen graph with imputation* <https://doi.org/10.59170/stattrans-2023-016>.

Włodzimierz Okrasa, Oleksandr H. Osaulenko, *Preface*

Włodzimierz Okrasa, Oleksandr H. Osaulenko, *From the Editors*

## Information about the series

**Statistical Research Papers** is a series of scientific monographs published by Statistics Poland, presenting works from the field of statistics and related sciences which have significantly contributed to the development of the world's science, written in English.

Statistics Poland is a publisher ranked on the list of publishers issuing peer-reviewed scientific monographs, compiled by the Minister of Science and Higher Education of the Republic of Poland. More information is available at Statistics Poland Research Portal: <https://research.stat.gov.pl>.

The portal also presents information on scientific journals published by Statistics Poland, as well as information about selected upcoming international events which will be attended by the representatives of the Polish official statistics.

This publication's aim is to celebrate the 100th issue of *Statistics in Transition new series* and the 30th anniversary of the launch of the journal. The book features several 'special issues' of *Statistics in Transition new series* focusing on specific topics of current research interest released in 2015–2023. The Appendix supplements the publication by providing tables of contents and forewords for each of the selected special issues.

*This book is an excellent resource for researchers, practitioners, and policymakers interested in the latest developments in sample survey methodology.*

Prof. Tomasz Żądło

University of Economics in Katowice, Poland



The digital version of the monograph is available at  
[srp.stat.gov.pl](http://srp.stat.gov.pl)

ISBN 978-83-67087-96-4  
e-ISBN 978-83-67087-97-1

Free copy