

## Preface to the special issue on integrating data from multiple sources for production of official statistics

Authors	Fosen,Johan; Holmberg,Anders; Jansson,Ingegerd; Krapavickaitė,Danutė et al
Published in	Journal of Official Statistics
DOI	<a href="https://doi.org/10.1177/0282423X251332063">10.1177/0282423X251332063</a>
Publication Date	2025-06
Document Version	publishersversion
Link	<a href="https://research.tilburguniversity.edu/en/publications/e9051148-0c6a-4eba-869b-d7a38f3c7834">https://research.tilburguniversity.edu/en/publications/e9051148-0c6a-4eba-869b-d7a38f3c7834</a>
Citation	Fosen, J, Holmberg, A, Jansson, I, Krapavickaitė, D & de Waal, T 2025, 'Preface to the special issue on integrating data from multiple sources for production of official statistics', Journal of Official Statistics, vol. 41, no. 2, pp. 539-546. <a href="https://doi.org/10.1177/0282423X251332063">https://doi.org/10.1177/0282423X251332063</a>
Download Date	2026-05-17 12:36:45
Rights	<p>General rights</p> <p>Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.</p> <ul style="list-style-type: none"> <li>- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.</li> <li>- You may not further distribute the material or use it for any profit-making activity or commercial gain</li> <li>- You may freely distribute the URL identifying the publication in the public portal"</li> </ul> <p>Take down policy</p> <p>If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.</p>

# Preface to the Special Issue on Integrating Data from Multiple Sources for Production of Official Statistics

Journal of Official Statistics

2025, Vol. 41 (2) 539–546

© The Author(s) 2025

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/0282423X251332063

[journals.sagepub.com/home/jof](https://journals.sagepub.com/home/jof)

Johan Fosen<sup>1</sup>, Anders Holmberg<sup>2</sup>,  
Ingegerd Jansson<sup>3</sup> , Danutė Krapavickaitė<sup>4</sup>   
and Ton de Waal<sup>5,6</sup> 

## Keywords

non-probability data, selection bias, multisource statistics, administrative data, new data sources

## 1. Introduction

Societal changes in recent years have spurred the interest in using multiple data sources for producing official statistics. There are several reasons for this. Combining data sources gives the opportunity to produce more timely and detailed statistics, and to respond quickly to unexpected or sudden events. Other drivers are the possibility to reduce data collection and processing costs, and to reduce response burden (De Waal et al. 2020). However, there are also challenges. In this special issue, we focus on some of them.

Traditionally, official statistics producers have used survey samples, possibly aided by administrative data for example in the design of surveys or in estimation.

<sup>1</sup>Statistics Norway, Oslo, Norway

<sup>2</sup>Methodology & Data Science Division, Australian Bureau of Statistics, Belconnen, ACT, Australia

<sup>3</sup>Statistics Sweden, Solna, Sweden

<sup>4</sup>Lithuanian Statistical Society, Vilnius, Lithuania

<sup>5</sup>Statistics Netherlands, The Hague, The Netherlands

<sup>6</sup>Tilburg University, Tilburg, The Netherlands

## Corresponding author:

Ingegerd Jansson, Statistics Sweden, Solna strandväg 86, Solna SE-17154, Sweden.

Email: [ingegerd.jansson@scb.se](mailto:ingegerd.jansson@scb.se)



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial

use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Using more than one data source is not new, but the number of administrative data sources is constantly increasing, as is the number of other types of data sources that face similar challenges. With the growing availability of data, the interest in taking on the challenges is also growing at national statistical institutes (Ascari et al. 2020; HLG-MOS 2017).

Administrative data, and other types of data sources not based on probability surveys, suffer from problems with representativeness if the intended target population is not covered by the data source (see e.g., Elliott and Valliant 2017; Kim and Tam 2021; Lohr and Raghunathan 2017, or Bakker et al. 2015). At the same time, response rates in sample surveys are decreasing to alarmingly low levels, causing increased challenges in adjusting for nonresponse. Combining available data sources for production of high-quality official statistics is seen as a promising way to solve these problems.

Multisource statistics or data integration encompasses different methodological areas, such as acquiring data, editing data, linking data, statistical matching, producing estimates, and describing or evaluating data quality. In this issue, most of the articles focus on combining data sources to estimate or minimize error and bias due to selection or under- or over-coverage, while data collection and data analysis are also discussed.

## **2. Non-Probability Sampling and Probability Sampling**

With probability sampling, any population unit has a known positive probability of being included in the sample, using a random mechanism for sampling. Its scientific ground was laid down by Tschuprow (1923) and Neyman (1934). The statistical properties have been widely studied, also when auxiliary variables from administrative registers are used in sample design and estimation. The introduction of model assisted survey sampling in Särndal et al. (1992) and the development thereafter with new types of estimators and designs illustrate this well (see e.g., Deville and Särndal 1992; Deville and Tillé 2004).

Administrative registers are increasingly replacing probability samples as the basis for official statistics. These registers are collected for administrative purposes and not for statistical purposes. Further, when used as the basis for official statistics, several administrative registers often need to be integrated. When two administrative registers essentially include the same units, the two registers can be linked at micro-level (unit level) to produce a data source with a larger set of variables. On the other hand, when the administrative registers include different units, these registers can be merged to increase the number of units. In both cases, micro-integration is usually needed to reconcile/harmonize the variables from the different administrative registers.

Administrative sources are fully enumerated samples from the viewpoint of its custodian, being typically a governmental agency. This is also the case for other types of data sources that have emerged and become of interest for official statistics production in recent years, such as road sensor data, position data from mobile network operators, transaction data from private vendors, or private data collected from smartphones. The target population, from the data holders' perspective, of

the source typically differs from the official statistics' target population. Thus, the data sources are non-probability samples: The set of units in such a source is a selection from the official statistics' target population, and the selection mechanism is whether or not the unit falls under the primary target population of the data custodian. Hence, the statistical concept of a frame population and its alignment to the target population (the coverage) is not the focus of the data custodian. The selection mechanism is then not controlled and usually non-random.

In addition, there are other non-probabilistic selection mechanisms corresponding to, for example, the non-response selection mechanism in a probability sample. When using a non-probability data source as the main source for official statistics, all these selection mechanisms are put together since they are all non-random. This non-randomness is a source for selection bias, with respect to the official statistics' target population.

Compared with non-probability samples, probability samples have well-known statistical properties and the advantage that the sampling error in ideal circumstances can be controlled. However, nonresponse in surveys can cause selection bias. Auxiliary information from other sources has for many years been used to try and reduce the bias, for example, by post-stratification or other calibration methods. The bias-reducing ability of such weighting methods is limited, and with increasing non-response, the selection bias challenge of random samples is increasing.

The variables of the data sources described above are not defined for the purpose of official statistics production and might not exactly adapt to the target variables of the official statistics (Zhang 2012). This is a source of measurement error which can be reduced by careful micro-integration in which variables are harmonized, typically after first linking two or more data sources. For the probability-based sample surveys on the other hand, measurement errors can be minimized by taking advantage of the opportunity to carefully design the measurement instrument through questionnaire design.

Non-probability data sources, although more susceptible to selection bias and measurement error than probability samples, have the advantage of lower data collection cost and not adding to further response burden since the data have already been collected. Not least, official statistics based on administrative registers, can be aggregated at small areas, as opposed to statistics based on probability samples where the sampling variance generally becomes too large if the sample size within an area is too small.

### **3. In This Issue**

To produce official statistics based on integrated data and with non-probability sources, methodological challenges need to be solved. Contributions to this important work are provided in this special issue. The articles in the issue go through all non-probability data processing stages. The issue starts with articles devoted to non-probability data collection and their linkage and correction for under-coverage and over-coverage. It is followed by parameter estimation using probability and non-probability data. This topic attracts attention in four of eight articles. Finally, the issue presents an analysis of the quality of the linear regression model based on

a non-probability sample data set with omitted regressors. In all articles a probability sample plays some balancing role.

Schouten et al. (2025) discuss the use of smart surveys and evaluate their usefulness in three use cases. A smart survey employs one or more smart features and some form of interaction with respondents, to put the data derived from smart features into context. Public online data, sensors, and data donation are examples of smart features that are handled with the help of the respondent. Data collection via smart devices, such as smartphones, can potentially improve measurement of variables and representation of the target population, but they also bring new challenges. With smart surveys comes a need for new logistics, new expertise, and new robust infrastructure that satisfy privacy requirements. The authors discuss consequences of hybrid data collection, as well as methodology to reduce errors in smart surveys. They give practical criteria to help decide if hybrid data collection has a positive business case and if further investments in systems for data collection, processing, and integration are justified. The criteria are applied to three case studies, the measurement of mobility, consumption, and physical activities.

A challenge in data integration occurs when two datasets due to privacy preservation cannot be linked directly. An example is in international trade where some of the export information is only known to the exporting country, whereas some of the corresponding import information is only known to the importing country. A linkage at micro level of the import and export information is needed, but a direct linkage cannot be done due to the need for privacy preservation. Privacy is preserved when linking the two datasets using a private set intersection protocol, and in this setting the article by Dasyuva et al. (2025) proposes a method for estimating population parameters while considering the errors occurring from both false positive and false negative linkages.

The Italian Permanent census data is an administrative register suffering from both under- and over-coverage error. The article by Ballerini et al. (2025) considers one sample designed to assess the under-coverage of the administrative source, and another sample designed to assess the over-coverage. A Bayesian analysis is used to estimate the population counts taking into account the coverage error. The resulting count estimates can be calculated at a detailed level, and the method also provides uncertainty estimates.

The selection bias of a non-probability sample is usually unknown and difficult to estimate. In the article by Villalobos Aliste et al. (2025), a model for this selection bias is proposed. The model makes it possible to estimate the expected mean squared errors for an estimator based on the probability sample and an estimator based on the non-probability sample. These expected mean squared errors are then used to construct a composite estimator based on both samples. The paper applies this method to estimate domain proportions for categorical variables.

Sometimes a non-probability data set may be considered as random. Analysis of the assumptions may lead to the approximation of the non-probability data formation mechanism by pseudo-probabilistic sampling methods. Čiginas et al. (2025) study a non-probability sample approximated by the Poisson pseudo-sampling design. The total of the binary study variable in the non-probability sample is estimated using the inverse probability weighted (IPW) estimator with estimated

propensity scores. It may be biased. Its variance is estimated taking into account randomness of the non-probability sample, by bootstrap and by other known variance estimators. If the study variable is available also in a probability sample, the total is estimated by the design-based estimator and finally the linear combination of both estimators is applied. This decreases the bias of the estimator of the total and stabilizes the variance estimator, becoming more robust to model misspecification of the propensity scores.

Non-probability samples are often selective and unweighted estimators based on these samples are therefore often biased. A frequently used approach to correct for this selection bias is to determine pseudo-weights for the units of the non-probability sample and to use a weighted estimator to estimate parameters of interest, such as a population mean. Several methods to construct pseudo-weights have been proposed in the literature. In many of these methods, different models can be used, for example models for inclusion in the non-probability sample. In practice, it is very difficult to determine which method and which associated model(s) yield the best set of pseudo-weights. In the article by Liu et al. (2025), performance measures are proposed to determine the best (or at least a good) method and associated models for pseudo-weights.

Non-probability samples can often be obtained from commercial vendors. However, very little is known about the quality of these non-probability samples. The article by Murray-Watters et al. (2025) first investigates whether the quality of non-probability samples is consistent across sample vendors. This does not appear to be the case, which means that there is a risk that a purchased non-probability sample is of poor quality and estimates based on that sample are highly biased. This is an important issue for anyone considering purchasing a non-probability sample from a commercial vendor. To reduce the risk of highly biased estimates, the authors propose to average estimates based on non-probability samples from multiple vendors. They investigate several approaches for averaging estimates based on these samples. The proposed approaches were evaluated in a real-world application using one probability sample from the German Internet Panel and eight non-probability samples from different vendors.

Omitted variable bias is a well-known problem in regression analysis. If relevant variables are not included in the data sets available to researchers, bias will occur in the estimated coefficients of the model. Linked administrative data sets are increasingly used and offer large sample sizes and higher precision in key variables, but they often lack relevant variables. In their article, Du et al. (2025) compare different cross-sectional and panel data structures, containing administrative and survey data. The authors provide a formal framework for understanding estimation bias due to the omission of important variables, as well as estimation bias arising from the use of imperfect proxy variables. The framework is applied to German administrative labor market data that are linked to survey data.

#### **4. Future Directions**

This issue highlights problems that arise when integrating data sources and using non-probability data, presenting examples of their solution and inviting survey

statisticians to search for new solutions of the by-standing statistical problems. Below, we summarize and generalize the main proposals for future work proposed by the authors of the articles.

New data sources include not only smart data derived by machines measuring and recording the events in the physical world, but also loosely structured data from social networks, business process-mediated data like business transactions, reference tables and relationships, bank records, and many other cases. Framework development and applications are needed aiding integrated use of survey data, administrative data, smart data, social networks and machine generated data, and artificial intelligence and machine learning methods for parameter estimation in finite population and uncertainty quantification.

Conditions for successful usage of different data sources need to be created, such as development of a linkage strategy using a trusted execution environment for record linkage which avoids ethical pitfalls and preserves privacy and confidentiality. Fully automating the design of the linkage strategy with machine learning techniques, within the trusted execution environment, would be useful. Statistically sound resampling techniques for linked data should be developed. Progress in this area would enable correct statistical inferences.

Methods to estimate under-coverage and over-coverage in administrative data sources, using data of probability samples, are proposed in this issue but statistical methods should be developed, which, using probability survey data, may adjust also for other kinds of non-sampling errors, such as misclassification, measurement errors, and item nonresponse.

Models can be constructed for estimating bias of estimators obtained in non-probability samples when estimating finite population domain proportions, taking into account unbiased estimators from a simple random sample. Future research may introduce models for estimation of bias in non-probability samples in other cases, such as when the study variable is continuous, when the estimated parameter may be linear and nonlinear, for complex sampling design of the probability sample, for informative non-probability sample, or when dependent parameter estimators in both samples make the study complicated.

Non-probability samples may under certain assumptions be treated as pseudo-probability samples and integrated with the probability sample at one time point or over time when estimating nonlinear parameters, taking into account randomness of the non-probability sample and non-sampling errors in both samples.

Performance measures for weights taken from other applications and used in non-probability samples are important. The range of these measures could be extended to continuous univariate and multivariate study variables, or linear and non-linear parameters of estimation. When non-probability sample data come from different vendors, researchers have to estimate their quality by some quality indicator. The researcher may in addition use a model to estimate the bias in the estimates from the non-probability sample or try to apply some performance measure for quality assessment of the weights.

Utilizing panel data reveals the presence of omitted variable bias in cross-sectional results, indicating that panel analysis captures relevant unobservable

components more effectively than an expanded set of regressors at a single time point. The findings suggest that data producers should consider making panel data based on administrative sources available and restrict surveys to variables that are required for the analysis itself but not available from administrative sources.


The issue does not include any articles integrating data over spatial or hierarchical dimensions. This situation arises when one data source is more granular than others and when integration subsequently happens on the least granular level, potentially causing multilevel methodological questions. We foresee future work on such questions as well as studies of the more general question when solutions based on integrated data (given that they satisfy quality requirements) are cost-effective or not, compared to relying on a single data source.


Data integration and the use of non-probability data are important topics for production of official statistics. We hope that this issue of JOS will advance the knowledge in the field and stimulate ideas for further development of official statistics.


## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iDs

Ingegerd Jansson  <https://orcid.org/0009-0002-8853-0542>

Danutė Krapavickaitė  <https://orcid.org/0000-0002-7638-6697>

Ton de Waal  <https://orcid.org/0000-0003-1515-3790>

## References

- Ascari, G., K. Blix, G. Brancato, T. Burg, A. McCourt, A. van Delden, D. Krapavickaitė, N. Ploug, S. Soltus, P. Stoltze, and T. de Waal. “Quality of Multisource Statistics – The KOMUSO Project.” *The Survey Statistician* 81: 36–51.
- Ballerini, V., M. Di Zio, B. Liseo, and S. Toti. 2025. “Combining Administrative and Survey Data to Correct Coverage Errors in Register-Based Statistics: A Bayesian Approach.” *Journal of Official Statistics* 41 (2): 598–618. DOI: <https://doi.org/10.1177/0282423X241312739>.
- Bakker, B. F. M., P. G. M. Van der Heijden, and S. Scholtus. 2015. “Preface.” *Journal of Official Statistics* 31 (3): 349–55. DOI: <https://doi.org/10.1515/jos-2015-0021>.
- Čiginas, A., D. Krapavickaitė, and V. Nekrašaitė-Liegė. 2025. “Evaluating the Impact of a Non-Probability Sample-Based Estimator in a Linear Combination with an Estimator from a Probability Sample.” *Journal of Official Statistics* 41 (2): 649–74. DOI: <https://doi.org/10.1177/0282423X251331346>.
- Dasylyva, A., M. De Cubellis, F. De Fausti, and L. Franssen. 2025. “Linking Trade Data from Different National Statistical Offices Through a Private Set Intersection.” *Journal of Official Statistics* 41 (2): 569–97. DOI: <https://doi.org/10.1177/0282423X251329407>.

- Deville, J.-C., and C.-E. Särndal. 1992. "Calibration Estimators in Survey Sampling." *Journal of the American Statistical Association* 87 (418): 376–82. DOI: <https://doi.org/10.1080/01621459.1992.10475217>.
- Deville, J.-C., and Y. Tillé. 2004. "Efficient Balanced Sampling: The Cube Method." *Biometrika* 91 (4): 893–912. DOI: <https://doi.org/10.1080/01621459.1992.10475217>.
- De Waal, T., A. Van Delden, and S. Scholtus. 2020. "Multi-Source Statistics: Basic Situations and Methods." *International Statistical Review* 88 (1): 203–28. DOI: <https://doi.org/10.1111/insr.12352>.
- Du, S., R. A. Wilke, and P. Homrighausen. 2025. "On Omitted Variables, Proxies, and Unobserved Effects in Empirical Regression Analysis." *Journal of Official Statistics* 41 (2): 725–44. DOI: <https://doi.org/10.1177/0282423X241312644>.
- Elliott, M. R., and R. Valliant. 2017. "Inference for Nonprobability Samples." *Statistical Science* 32 (2): 249–64. DOI: <https://doi.org/10.1214/16-STS598>.
- HLG-MOS. 2017. "A Guide to Data Integration for Official Statistics. Version 2.0." HLG-MOS Guide to Data Integration for Official Statistics.pdf.
- Kim, J.-K., and S.-M. Tam. 2021. "Data Integration by Combining Big Data and Survey Sample Data for Finite Population Inference." *International Statistical Review* 89 (2): 382–401. DOI: <https://doi.org/10.1111/insr.12434>.
- Liu, A.-C., S. Scholtus, K. Van Deun, and T. De Waal. 2025. "Performance Measures for Sample Selection Bias Correction by Weighting." *Journal of Official Statistics* 41 (2): 675–99. DOI: <https://doi.org/10.1177/0282423X251318463>.
- Lohr, S. L., and T. E. Raghunathan. 2017. "Combining Survey Data with Other Data Sources." *Statistical Science* 32 (2): 293–312. DOI: <https://doi.org/10.1214/16-STS584>.
- Murray-Watters, A., S. Zins, J. W. Sakshaug, and C. Cornesse. 2025. "Averaging Non-Probability Online Surveys to Avoid Maximal Estimation Error." *Journal of Official Statistics* 41 (2): 700–24. DOI: <https://doi.org/10.1177/0282423X241312775>.
- Neyman, J. 1934. "On the 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): 558–606. DOI: <https://doi.org/10.2307/2342192>.
- Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. Springer-Verlag Publishing. DOI: <https://doi.org/10.1007/978-1-4612-4378-6>.
- Schouten, B., P. Lugtig, and A. Luiten. 2025. "Can Smart Surveys Have a Positive Business Case? An Evaluation on Three Case Studies." *Journal of Official Statistics* 41 (2): 547–68. DOI: <https://doi.org/10.1177/0282423X251321634>.
- Tschuprow, A. A. 1923. "On the Mathematical Expectation of the Moments of Frequency Distributions in the Case of Correlated Observations." *Metron* 2 (4): 646–83.
- Villalobos Aliste, S. F., S. Scholtus, and T. De Waal. 2025. "Combining Probability and Nonprobability Samples on an Aggregated Level." *Journal of Official Statistics* 41 (2): 619–48. DOI: <https://doi.org/10.1177/0282423X241293751>.
- Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-Based Statistics and Data Integration." *Statistica Neerlandica* 66 (1): 41–63. DOI: <https://doi.org/10.1111/j.1467-9574.2011.00508.x>.

Received: March 15, 2025

Accepted: March 18, 2025