STATISTICS IN TRANSITION new series, June 2023
Vol. 24, No. 3, pp. 27–30, https://doi.org/10.59170/stattrans-2023-031
Received – 19.05.2023; accepted – 31.05.2023

Comments on „Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day" by Graham Kalton

Risto Lehtonen[1]

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 micro-level 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 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

[1] University of Helsinki, Finland. E-mail: risto.lehtonen@helsinki.fi.

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.

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, pp. 941–963.

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.

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.

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

Rancourt, E., (2018). *Admin-First as a statistical paradigm for Canadian official statistics: Meaning, challenges and opportunities*. Proceedings of Statistics Canada Symposium 2018.

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. *Jpn J Stat Data Sci*, 3, pp. 625–650.

Zhang, L.-C., (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1), pp. 41–63.

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, (In Press).

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.