The Survey Statistician, 2024, Vol. 89, 68–73.

EARLY CAREER SURVEY STATISTICIAN

The growth of new researchers in the era of new data sources

Veronica Ballerini[1] **and Lisa Braito**[2]

[1]University of Florence, Italy, veronica.ballerini@unifi.it
[2]University of Florence, Italy, lisa.braito@unifi.it

Abstract

In this article, we introduce a new section of The Survey Statistician, the "Early Career Survey Statistician." We report our personal experience in the field of survey statistics as early career researchers and, inspired by the recent workshop on methodologies for official statistics held in Rome last December, we make a digression into the new challenges in the era of innovative data.

Keywords: Early Career Researchers, Innovative data, Machine Learning, Nonprobability samples

1 Introduction

In the last issue of *The Survey Statistician*, Prof. Danny Pfefferman raised his concern about the involvement of "young" statisticians in the activities of the IASS. In the same article, he reported a suggestion from a reviewer that did not go unheeded: allocating a special section in TSS for young survey statisticians. Said and done, this is the first introductory number of a new section of TSS, the "Early Career Survey Statistician" (ECSS). The ECSS welcomes original research works of junior researchers, summaries of their research, review papers on survey statistics-related topics, reviews of events on survey statistics, and innovations introduced by junior researchers at the statistical offices. An "early career survey statistician" is a person with up to 5 years of employment or within the fifth year since the achievement of their Ph. D., who is researching in the field of survey statistics. To continue citing the recent article by Pfefferman in TSS, "the outcome of our [*survey statisticians'*] work affects directly so many applications and decision makings." In the era of novel data sources and huge data availability, this is truer than ever. Indeed, the research work of survey statisticians is instrumental to the proper secondary use of big or complex data and nonprobabilistic samples in general, and it is essential to contribute to making the estimates obtained reliable. Every time and in every context, change has always been embraced by the "youth"; we should also ride the change in the survey statistics field.

I (Veronica) met many early career researchers like myself in this year's events on survey statistics; the majority of us come from mixed backgrounds and our research crosses different statistical fields. Such a transdisciplinary attitude might be a positive aspect. Another of my main research interests is causal inference for clinical and observational studies, which is a broad field perceiving the opportunities offered by data integration and facing its challenges at the same time.

Lisa, who is a Ph. D. student in my own Department, and I have come across such challenges; for this reason, we are broadly reviewing the state of the art of what concerns new data sources and methods, and their opportunities. These have been topics also at the core of the "2nd Workshop on Methodologies for Official Statistics," held at the Italian National Statistical Institute (Istat) in Rome last December, from which we partially draw inspiration in this article. All presentations are available at https://www.istat.it/en/archivio/288564. We share some insights here, hoping to stimulate other early career researchers to get involved in such new survey statistics challenges and calling for contributions for the next issues of this new TSS section.

2 Challenges is Opportunities in Survey Statistics

If one were to identify the most significant drawback in the production of reliable statistics, it would be poor timeliness. Hence, the (almost) real-time production of "big" data is very attractive. However, it is important to quickly elaborate trustworthy data to prevent users from blindly relying on raw big data. Because "big" is not enough for reliable inference: "[...] classical statistical ideas continue to have a crucial role to play in keeping data analysis honest, efficient, and effective. [...] huge new computing resources do not put an end to the need for careful modelling, for honest assessment of uncertainty, or for good experimental design" (Lockhart, 2018). With the words of Monica Pratesi at the aforementioned workshop, "uncertainty is here to stay; so, we need inference".

Innovative data need to be processed and analyzed using innovative methods. Thus, there is a need for bridging traditional survey statistics with machine learning methods, which are more suitable to deal with such nontraditional data.

2.1 Innovative data for survey statistics

Satellite data, remote sensing, mobile network data, mobile sensor data, social media platforms, web scraping, Google trends, web surveys, and scanner data: all of these, and even more, belong to the class of "innovative data". As pointed out by Stefano Iacus in his master class on "Limits and challenges of incorporating innovative data in official statistics", their main competitive advantages are the fine geographic and temporal granularity, the profound timeliness, and the large coverage they can reach.

Let us consider the illustrative case brought forth by Iacus and his research conducted in collaboration with the European Comission, focusing on the complexities of migration as a phenomenon. Migration presents inherent characteristics that pose challenges when relying solely on traditional data sources for analysis. Surveys, for instance, suffer from high costs, infrequent data collection, and limited coverage, making them inadequate tools to comprehensively analyze this dynamic and multifaceted phenomenon. Given the rapid evolution of migration, the aforementioned limitations in timeliness and granularity associated with traditional sources, such as surveys, contribute to their diminished reliability. Moreover, the elusive nature of the target population further reduces the capture probability of the units within these surveys. To tackle these limitations, Iacus and his collaborators explored the combined use of innovative and traditional data. For insights on this topic see, e. g., Carammia, Iacus and Wilkin (2022) and Spyratos et al. (2020).

Another context where innovative data might be useful is what is called "data equity", namely the need for representative data and disaggregated statistics about those population groups that may be discriminated by the data production process. For instance, sometimes the linking procedure may introduce biases because linkage errors can disproportionately affect members of population subgroups due to, e. g., spelling and/or typing errors (National Academies of Sciences, Engineering, and Medicine, 2023). Alternative data sources are crucial in advancing data equity by identifying data gaps or misrepresentations. They contribute by offering insights into population subgroups that are often under-represented in traditional surveys, such as individuals experiencing homelessness or residing in institutions like nursing homes. Additionally, these sources facilitate the generation of statistics that are disaggregated by key characteristics like race, ethnicity, education, disability status, and other factors of interest. This inclusive approach helps address disparities and ensures a more comprehensive understanding of diverse demographic groups.

Despite the undeniable opportunities, the utilization of non-traditional data sources for statistical purposes also presents several challenges. Before methodological considerations, let us underline that a relevant matter pertains to the management and processing of these data. Notably, new data sources lack stability, as they may cease to exist over time based on the discretion of the private entities that possess them. Data quality, data management and processing are open research fields, especially for NSIs.

Foremost among the methodological challenges is the issue of data linkage, which is further complicated by concerns surrounding data protection and privacy. Then, transparency, in general, remains a significant concern when dealing with such data sources, as the data production process is often unclear, necessitating reverse engineering efforts to integrate this data coherently with traditional sources. Additionally, various types of biases must be carefully addressed, such as selection bias stemming from the nonprobabilistic nature of these data.

Traditionally, a nonprobability sample is a sample with an unknown participation mechanism and an unknown sampled population; nonprobability samples that have been deeply investigated in the last decades are not only web surveys, volunteer surveys, administrative data, but also probability samples that encounter issues such as very low nonresponse rates and nonignorable nonresponse. The issues related to the nonprobability samples are intrinsic in their definition. Nowadays, the set of nonprobability samples comprises also the realm of big data. Whereas the literature about the integration of probability and nonprobability samples in a traditional framework is rich (among others, see Wu (2022) and his master class presentation slides during the workshop in Rome), with emerging new data sources and reshaped views of traditional data sources, data integration and data harmonization have become a very broad area that calls for continued research. Survey statisticians have started making efforts to integrate traditional literature on combining probability samples and innovative data (Yang and Kim, 2020). On the one hand, when the nonprobability sample has a large sample size and a probability sample including the response variable is available, it is possible to exploit the auxiliary information in the big data to improve the efficiency of the estimators of interest (Kim and Tam, 2021; Yang and Ding, 2020). On the other hand, when the big data includes the response variable, research has been done in the direction of leveraging probability samples to correct for selection bias improving robust mass imputation methods; for instance, see Yang, Kim and Hwang (2021). Lastly, a problem that may arise is linked to the large availability of variables in the big data sample; in this case, irrelevant auxiliary variables can introduce large variability in the estimation. Research is moving towards variable selection approaches tackling data integration and estimation rather than prediction; see, e. g., Chen, Valliant and Elliott (2018) and Yang, Kim and Song (2019).

2.2 Machine learning in survey statistics

The paradigm shift occurring in the field of survey statistics involves not only data sources but also methods for inference. Within this context, the shortcomings (and opportunities) we previously discussed about innovative data are similarly applicable to the new methodologies. The final session of the workshop, organized in collaboration with IASS, discussed these topics in detail, focusing on machine learning methods (ML) in survey statistics. In the last decade, the discussion concerning ML in survey statistics has been a hot topic. As pointed out by Puts and Daas (2021), ML methods provide several advantages, such as better scalability, less sensitivity to outliers and erroneous data, and the ability to capture non-linear relationships. However, the authors also reflect upon challenges and limitations arising from the application of ML methods in survey statistics. Among the challenges, accessibility and clarity have to be taken into account; especially in the field of survey statistics, it is important to define and fully understand the process by which results are obtained. This is a problem for some ML algorithms that result in black boxes, and it touches on the topic of explainability and the development of explainable AI. Furthermore, we need to pay attention to accuracy and reliability. In the employment of ML methods, the spotlight is often far from the uncertainty assessment and the estimators' robustness itself; these concerns are at the core of survey and official statistics instead.

It is noteworthy that the use of ML methods in survey statistics is two-fold. On the one hand, there is the application of classical prediction tasks to problems of survey methodology (optimizing data collection, adaptive survey designs, predicting nonresponse break off in online web surveys) and survey statistics (imputation, classification, data integration, automatic coding, anomaly detection, forecasting), as also mentioned by IASS president Natalie Shlomo. The advantages of ML in this context have been already exploited by NSIs, especially for what concern imputation, data quality assurance, survey sampling, document classification issues, time series forecasting, topic modeling, sentiment analysis, and geospatial analysis. For examples of works on these topics, see Beck, Dumpert, Feuerhake (2018); Buskirk, Bear and Bareham (2018); Burkirk et al. (2018); Kern et al. (2023), Kern, Klausch and Kreuter (2019); see also UNECE (2021). However, the application of these methods is still conceived as "experimental" statistics, in the sense that they are not consistently integrated in the NSIs systems.

On the other hand, "ML in survey statistics" also refers to the development of novel intersections between ML and survey statistics methodologies. This area represents an ongoing research field with ample scope for exploration and innovation. Examples of research areas at this intersection are the issues of sampling the population to obtain representative training sets, using stratification in the context of ML, reducing spurious correlations and assessing causal relationships, correcting the bias caused by the ML model, dealing with concept drift (Puts and Daas, 2021). To have some insights on research studies going in this direction see Breidt and Opsomer (2017); Chen and Haziza (2019); Dagdoug, Goga, and Haziza (2023a, 2023b); see also Buskirk and Kirchner (2020).

3 Concluding remarks

Encouraging the active participation of early career statisticians in research and collaboration endeavors will not only nurture their professional growth but also foster the development of novel methodologies and approaches that address the evolving challenges and opportunities in survey statistics. The interdisciplinary background we may have could be a potential strength, not only in the research work per se, but also in the creation of synergies among statisticians, data scientists, computer scientists, and operational and applied researchers, which is crucial for advancing the field of survey statistics in the era of data-driven decision-making. Inspired by the topics of the recent workshop on methodologies for official statistics held in Rome last December, we overviewed some of the hot topics in modern survey statistics, namely the use of innovative data sources and the role of machine learning methods in this field, to stimulate other early career statisticians to contribute to this field.

References

Beck, M., Dumpert, F. and Feuerhake, J. (2018) Machine Learning in Official Statistics.
arXiv preprint, https://arxiv.org/abs/1812.10422; https://DOI:10.18356/9789210011143.

Breidt, F. J., and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32 (2) 190 - 205, May 2017. https://doi.org/10.1214/16-STS589

Buskirk, T. D., Bear, T., and Bareham, J. (2018). Machine made sampling designs: applying machine learning methods for generating stratified sampling designs. *Paper presented at the BigSurv18 Conference*, Barcelona, Spain (25-27 October 2018).

Buskirk, T. D., and Kirchner, A. (2021). Why machines matter for survey and social science researchers: Exploring applications of machine learning methods for design, data collection, and analysis. *Big data meets survey science: A collection of innovative methods*, Eds. Hill, C. A.; Biemer, P. P.; Buskirk, T. D.; Japc, L.; Kirshner, A.; Kolenikov, S.; Lyberg, L. E. John Wiley & Sons, 9-62. https://DOI:10.1002/9781118976357.

Buskirk, T. D., Kirchner, A., Eck, A., and Signorino, C. S. (2018). An introduction to machine learning methods for survey researchers. *Survey Practice*, 11(1).

Carammia, M., Iacus, S. M., and Wilkin, T. (2022). Forecasting asylum-related migration flows with machine learning and data at scale. *Scientific Reports*, 12(1), 1457. https://doi.org/10.1038/s41598-022-05241-8

Chen, J. K. T., Valliant, R., and Elliott, M. R. (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 117–144.

Chen, S., and Haziza, D. (2019). Recent developments in dealing with item non-response in surveys: A critical review. *International Statistical Review*, 87, S192-S218. https://doi.org/10.1111/insr.12305

Dagdoug, M., Goga, C. and Haziza, D. (2023a). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 542, 1234-1251.

Dagdoug, M., Goga, C. and Haziza, D. (2023b). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. *Journal of Survey Statistics and Methodology* 11, 141-188.

Kern, C., Eckman, S., Beck, J., Chew, R., Ma, B., and Kreuter, F. (2023). Annotation Sensitivity: Training Data Collection Methods Affect Model Performance. *arXiv preprint* arXiv:2311.14212, https://doi.org/10.48550/arXiv.2311.14212.

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey research methods*, 13(1), 73-93. https://doi.org/10.18148/srm/2019.v1i1.7395

Kim, J. K. and Tam, S. M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89(2), 382-401.
https://doi.org/10.1111/insr.12434

Lockhart, R. (2018). Special issue on big data and the statistical sciences: Guest editor's introduction. *The Canadian Journal of Statistics*, 46(1), 4-9. https://doi.org/10.1002/cjs.11350

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Eds. S. L. Lohr, D. H. Weinberg, K. Marton. The National Academies Press. https://doi.org/10.17226/26804

Pfeffermann, D. (2023). The IASS–50 Years of Activity. *The Survey Statistician*, 88, 72-74.

Puts, M. J. H. and Daas, P. J. H. (2021). Machine Learning from the Perspective of Official Statistics. *The Survey Statistician*, 84, 12-17.

Spyratos, S., Vespe, M., Natale, F., Iacus, S. M., and Santamaria, C. (2020). Explaining the travelling behaviour of migrants using Facebook audience estimates. *Plos ONE*, 15(9), e0238947.

UNECE (2021). *Machine Learning for Official Statistics*. United Nations publication, Geneva. https://unece.org/sites/default/files/2022-09/ECECESSTAT20216.pdf

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*. 48(2), 283-311. http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm

Yang, S. and Ding, P. (2019) Combining Multiple Observational Data Sources to Estimate Causal Effects, *Journal of the American Statistical Association*, 115:531, 1540-1554, doi:10.1080/01621459.2019.1609973

Yang, S., Kim, J. K., and Hwang, Y. (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47(1), 29-58.

Yang, S., Kim, J. K., and Song, R. (2019). Doubly robust inference when combining probability and nonprobability samples with high-dimensional data. *Journal of the Royal Statistical Society*, Series B, 82, 445–465.

Yang, S., and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 3, 625-650.