Small Area Estimation in the Era of Machine Learning and Alternative Data Sources: Opportunities, Challenges and Outlook

Nikos Tzavidis*

*University of Southampton, Southampton, UK

1 Introduction

Advances in machine (statistical) learning algorithms along with the availability of large datasets from alternative sources have led to the production of small area estimates globally and at refined spatial scales. Within the space of a few years, some analysts have moved away from an overcautious use of models for survey estimation to using machine learning algorithms. Private firms and research organizations are publishing small area-type estimates using machine learning methods and data from several sources. The following example helps to set the context. Meta recently developed a methodology to produce, among other estimates, global estimates of average wealth at 2.4km2 resolution (Chi et al., 2022). The methodology uses household wealth data collected from surveys in 56 low and middle income countries. Data from alternative sources which include remote sensing data, mobile phone data, topographic maps, and privacy-protecting connectivity data from Facebook are processed to create quantitative features at disaggregated spatial scale (village level). These quantitative features act as predictors of wealth. Using spatial markers, the average asset-based wealth of each village is matched to quantitative features from the alternative data sources to create the ground truth data (data used to estimate/train models/algorithms). In the final step, the data are used to train a supervised machine learning algorithm to predict the relative wealth for each populated 2.4km2 grid cell in 135 low and middle income countries, including in grid cells with no ground truth data (out-of-sample grid cells).

In many cases, new small area-type estimates are being produced outside the mainstream statistical literature and do not acknowledge important methodological and applied work in the small area estimation literature over the past 30 years. For example, Meta's methodology does not account for the fact that the average asset-based wealth of each village is an estimate itself and, therefore, is affected by sampling error. In the small area literature, area-level models have been proposed to account for this (Fay & Herriot, 1979). It is also not clear how estimates produced with new methods compare to estimates produced with industry standard methods and whether they would pass quality control checks we use in survey and official statistics, for example, precision thresholds as quantified by uncertainty measures. And yet, estimates and methodology of the type described above attract significant interest in applied research.

Section 1 in Meng (2018) is entitled *Prologue: Paradise gained or lost?*. Meng (2018) points out that *"It is generally true that we, as statisticians, lament the increasing loss of principled statistical methodology.* However, as he puts it, *"... the more we lament how our nutritious recipes are increasingly being ignored, the more fast food is being produced, consumed and even celebrated as the cuisine of a coming age. Indeed, some of our most seasoned chefs are working tirelessly to preserve our time-honored culinary skills, while others are preparing themselves for the game of speed cooking. Yet others need a daily nightcap to ease the nightmare of the forever lost statistical paradise, even before it actually arrives."* As statisticians working in survey and official statistics, we should neither lament nor be complacent to ignore or endorse new methods impetuously. By doing so, we run the risk that important research becomes outdated and irrelevant. Meng (2018) suggests a way forward. *"...I see a paradise, or even paradises, gained if there is a sufficient number of us who can engage in what we have advertised to be the hallmark of our discipline, that is, principled thinking and methodology development for dealing with uncertainty. Fast food will always exist because of the demand... But this is the very reason that we need more people to work on understanding and warning about the ingredients that make fast food (methods) harmful; to study how to reduce the harm without unduly affecting their appeal; and to supply healthier and tastier meals (more principled and efficient methods) that are affordable (applicable) by the general public (users). This is how I see paradises arising."*

The introductory section in Meng (2018) captures the spirit in which I set out to write this article. New and rediscovered algorithmic tools and data offer opportunities sufficient to support a period of exciting research, but also pose significant challenges. The way new small area estimates are presented and certain methodological choices raise several questions. Notable differences between industry standard and new methods can be classified into the following broad themes: (a) the types of models/algorithms used and their specification including the spatial scale at which estimates are produced, (b) uncertainty quantification, including evaluation and communication of the uncertainty estimates, and (c) the use of alternative sources of data and how these are integrated with survey data. I discuss each of these themes and highlight areas where, in my view, further research is needed. What I will be advocating is that we must critically appraise the use of new methods and data, compare these to industry standard methods, and use theoretical and empirical evidence to influence the direction of research and applications. The discussion cannot be exhaustive, but I hope it will generate an interesting debate and inspire ideas for research.

2 Models and algorithms

Linear models (area-level or unit-level) with random effects to account for unobserved area heterogeneity are the industry standard for small area estimation (Battese et al., 1988; Molina & Rao, 2010). An important part of the small area estimation literature has focused on strategies to mitigate the impact of model misspecification. This includes the use of data-driven transformations to ensure the validity of assumptions for the model error terms (Rojas-Perilla et al., 2020), model selection under data-driven transformations (Lee et al., 2023), models with flexible distributional assumptions (Graf et al., 2019), outlier-robust methods (Chambers & Tzavidis,

2006; Jiongo et al., 2013; Chambers et al., 2014), and models that relax linearity using, for example, splines (Opsomer et al., 2008). This line of research shows that statisticians strive to achieve robustness of model-based inference against departures from the model assumptions. This is important if model-based methods are to be adopted in practice. Machine learning, on the other hand, brings several advantages, including automatic model selection, possible robustness to model misspecification, and the ability to capture complex relationships between predictors, thus relaxing the assumption of linearity. At the same time, these advantages present risks that can lead to what may be considered an offhand approach to using machine learning.

Applications of machine learning to small area estimation do not explicitly recognize the importance of the data structure and how this should inform the specification of the algorithm. An important element of small area estimation is how we measure unobserved heterogeneity. This is typically done by using area random effects. Using the principle of parsimony, if predictors explain the heterogeneity among areas sufficiently, using random effects is not necessary (Datta et al., 2011). However, failing to include random effects when needed can lead to biased estimates. Capturing complex relationships between predictors, automatic model selection, and access to large sets of data may reduce the need to model unobserved heterogeneity, but this does not justify assuming the absence of unobserved heterogeneity a priori. Using random effects is a popular but not the only approach to quantifying unobserved heterogeneity. Including additional features as predictors, using contextual variables, and nonparametric approaches to measuring unobserved heterogeneity in the spirit of Chambers & Tzavidis (2006) are alternative approaches to the same end. Krennmair et al. (2024) study extensions of random forests for small area estimation of general parameters. Using real data, the authors illustrate the advantages of modeling unobserved heterogeneity, the importance of using data transformations, and the impact of debiasing the random forest fit before using this for small area estimation. These are topics of active research. Tuning a machine learning algorithm is also an important step that should not be ignored. Understanding the impact of tuning on the small area estimates is important and requires further empirical work. Despite some evidence for the robustness of machine learning methods to model misspecification, recent research (Krennmair et al., 2024) shows that a black-box approach to using machine learning carries risks and should be avoided. Simply showing that a set of point estimates is reasonably well correlated with external validation datasets is not sufficient because it ignores how precise the external estimates are (Corral et al., 2025). Of course, this assumes that we know how to measure precision. We return to this point below when we discuss uncertainty quantification.

The choice of spatial scale at which estimates are produced is also worth noting. Producing estimates at 2.4km2 resolution globally, as in Meta's case, requires extrapolating the use of the model/algorithm where no ground-truth data exist. Doing so with linear models is straightforward, too, but the fact that we can does not mean that we should. Out-of-sample estimation is risky even at a higher geographic level where only a few areas may be out-of-sample. So, is extrapolation justified because of the ability of machine learning algorithms to capture complex relationships and the assumption that there is no unobserved heterogeneity? Producing estimates at 2.4km2 resolution creates aesthetically appealing maps. However, it also raises the

question about the utility of such estimates, given the anticipated high error of the estimates at this level and the fact that policy implementation usually operates at higher levels of geography. Assessing the ability of machine learning algorithms to produce useful estimates at low spatial scales and validating machine learning-based estimates is an area of active research.

3 Estimating and communicating uncertainty

Estimating the uncertainty of small area estimates, commonly quantified by the mean squared error (MSE), is an integral part of small area estimation. Estimating the MSE using analytic and bootstrap or jackknife methods has been the focus of important theoretical and applied work for more than thirty years (Prasad & Rao, 1990; Hall & Maiti, 2006; Jiang et al., 2002; Chambers et al., 2011; Chambers & Chandra, 2013). This is in contrast to the use of machine learning methods for small area estimation. Even when the estimation of uncertainty is discussed (Chi et al., 2022), the method used is not clearly described. In Chi et al. (2022) the error is estimated by using a regression model (or variations of this) on residuals. However, it is not clear what kind of uncertainty the estimated error is measuring. The proposed method is also not compared to industry-standard MSE measures. Chi et al. (2022) mention that producing error estimates is important for policy decisions. However, producing estimates at 2.4km2 resolution shows that the true focus is on the first-order problem of prediction/point estimation. Using uncertainty measures at this scale pays lip service to the norm of quantifying uncertainty. This is because we know that with high probability, at this scale (and with so many out-of-sample grid cells) uncertainty is expected to be large enough to make the use of estimates for practical purposes meaningless. Although uncertainty quantification is an integral part of statistical quality assurance frameworks, it appears that this may be treated by some scientific fields as a second-order problem.

Estimating the uncertainty of the estimates produced with machine learning methods is a topic of current research interest. Krennmair et al. (2024) apply block-bootstrap schemes for MSE estimation with random forests. Using conformal methods to generate distribution-free prediction intervals when applying machine learning methods to complex survey data (Wieczorek, 2023) is an interesting area for SAE research. Understanding how we can use the recently proposed framework by Zhang et al. (2024) in the context of small area estimation is also an interesting research area. Quantifying uncertainty must remain at the core of statistical work. Considering the advances in algorithmic methods for prediction/estimation, by failing to understand and promote the importance of using uncertainty measures, we risk that this important part becomes obsolete. Therefore, more work is needed to understand and promote the utility of uncertainty estimates for applied work.

4 Alternative data sources and integration with survey data

Model-based and model-assisted small area estimation is commonly implemented with the aid of auxiliary variables from population data, for example censuses. Increasing demand for estimates of non-linear statistics means that we must increasingly rely on population micro-data. Relying

on census data is restrictive. Censuses are less frequent in many countries in the global south and are usually only conducted every ten years. Using population data in the intercensal period therefore relies on the strong assumption that the distribution of census variables has not changed over this period. For example, in current research that uses census microdata as predictors in small area models, we find that using outdated census data in countries in the global south can lead to an overestimation of poverty rates in urban areas. This can be attributed to significant progress in household conditions in urban areas, which is not captured by outdated census data.

The increasing availability, improved quality, frequency, and coverage of data from alternative sources, for example, from remote sensing, mobile, and social networks raises the question about the role that such data sources can play in small area estimation. As mentioned above, Meta has already developed a methodology that relies on data from alternative sources as predictors in machine learning algorithms. Among alternative data sources, remote sensing data have gained popularity. Advances in the processing, global coverage, frequency, and free access of remote sensing data have created renewed interest in their use in predictive models. This is in contrast to data, for example, from mobile networks that are not easily accessible. However, using data from alternative sources is not without problems. By definition, zonal statistics (descriptive statistics summarizing the values of geospatial variables) are computed at an aggregate spatial scale e.g., a grid cell, or village as in Meta's case. Therefore, zonal statistics are contextual predictors and act only as proxies of household characteristics. For this reason, models that use only geospatial data as predictors are expected to be less predictive than models that use household characteristics. Is the reduced predictive power enough to produce estimates of acceptable precision? Current research shows that geospatial data can be of added value for small area estimation (Edochie et al., 2024) especially when considering that relying on outdated census data carries risks due to the possible changes in the distribution of census predictors in the intercensal period.

The predictive ability of geospatial data depends on the type of outcome we are interested in predicting and the context of the country. The spatial scale to be used for integrating geospatial and survey data, and model specification (area-level, unit-level or unit-context) is a topic of current research. Although geospatial data have been successfully used in several applications in countries in the global south, we cannot assume that the same geospatial variables will be equally predictive in other countries. The types of geospatial data and their utility in countries that are data-rich are the focus of current research. Data-rich settings also offer the opportunity to compare geospatial-based estimates against what are considered to be "gold standard" estimates. This line of research that explores the best uses of available data can inform data collection and survey design more generally, leading to a better allocation of available resources. Another compelling reason for using geospatial and other alternative data is that they offer a natural approach to updating the estimates in off-census years. However, if analysts are keen to continue using census data, SPREE-type methods (Isidro et al., 2016; Zhang & Chambers, 2004) offer an alternative approach to intercensal updating. Current research focuses on different estimation strategies in the off-census years. Last but not least, developing open-source software to automatically download, process, and integrate geospatial and survey data will minimize the steep learning curve of working with GIS software. This is another area of current research focus.

5 Conclusions and research outlook

Machine learning methods and access to large data are creating a hype that can overstate their ability to produce small area estimates of superior quality and added value to those produced by industry standard methods. New methods offer opportunities for small area estimation (and for survey and official statistics more generally) research and practice that should not be ignored. However, the compound effect of using new algorithms and data sources on the quality of estimates is not well understood. In this section, I summarize some areas that in my view would benefit from further theoretical and empirical research.

More research is generally needed on the specification of machine learning algorithms. Assuming a priori the absence of unobserved heterogeneity is a strong assumption. How to allow for unobserved heterogeneity in machine learning algorithms is a topic of current research with particular emphasis on (a) parametric and nonparametric methods and (b) how splitting criteria, e.g., in random forests can be used for this purpose. Evaluating the impact of the tuning of machine learning algorithms on the estimates is a further topic of research, as is finding an appropriate spatial scale at which useful estimates can be produced. The latter topic is of particular interest because there seems to be a misconception in parts of the literature that the lower the spatial scale, the more useful the estimates are. This line of thinking is, in my view, problematic because it ignores the uncertainty of the estimates. Measuring the uncertainty of estimates produced with machine learning algorithms is an important area where more theoretical and empirical research is needed. Understanding whether and how existing MSE estimators can be adapted for use with machine learning methods is a topic of current interest. The use of conformal methods to generate distribution-free prediction intervals (Wieczorek, 2023), using the framework proposed by Zhang et al. (2024), and finding effective ways to communicate the utility of uncertainty measures for applied research are all important research areas. More research is also needed on how to integrate survey data and data from alternative data sources. The role of geospatial (and other) data for estimation in off-census years and how this compares to updating methods that use census data is an area of current applied research interest. Finally, transferring best practice from countries with limited data sources to data-rich countries is also important. This can help us to consider how to use data from both alternative (new) and administrative sources to rethink the design of major surveys.

Acknowledgments: The author disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the Data and Evidence to End Extreme Poverty (DEEP) research programme. DEEP is a consortium of the Universities of Cornell, Copenhagen, and Southampton led by Oxford Policy Management, in partnership with the World Bank - Development Data Group and funded by the UK Foreign, Commonwealth and Development Office. The work is also supported by the UKRI-ESRC strategic research grant ES/X014150/1 for "Survey data collection methods collaboration: securing the future of social surveys", known as Survey Futures. Survey Futures is a collaboration

of twelve organizations, benefitting from additional support from the Office for National Statistics and the ESRC National Centre for Research Methods. Further information can be found at www.surveyfutures.net.

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), 28–36. doi: 10.1080/01621459.1988.10478561

Chambers, R., & Chandra, H. (2013). A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, *22*(2), 452–470. doi: 10.1080/10618600.2012.681216

Chambers, R., Chandra, H., Salvati, N., & Tzavidis, N. (2014). Outlier robust small area estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76 (1)*, 47-69. doi: 10.1111/rssb.12019

Chambers, R., Chandra, H., & Tzavidis, N. (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. *Survey Methodology*, *2*(37), 153–170. Retrieved from `https://www150.statcan.gc.ca/n1/pub/12-001-x/index-eng.htm`

Chambers, R., & Tzavidis, N. (2006). M-quantile models for small area estimation. *Biometrika*, *93*(2), 255-268. doi: 10.1093/biomet/93.2.255

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, E., J. (2022). Microestimates of wealth for all low and middle income countries. *PNAS*, *119*(3), 381–399. doi: 10.1073/pnas.2113658119

Corral, P., Henderson, H., & Segovia, S. (2025). Poverty mapping in the age of machine learning. *Journal of Development Economics*, *172*, 1033–77. doi: 10.1016/j.jdeveco.2024.103377

Datta, G., Hall, P., & Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, *106*(493), 362–374. doi: 10.1198/jasa.2011.tm10036

Edochie, I., Newhouse, D., Tzavidis, N., Schmid, T., Foster, E., Hernandez, A. L., ... Savadogo, A. (2024). Small area estimation of poverty in four west african countries. *Journal of Official Statistics*. doi: 10.1177/0282423X241284890

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366), 269-277. doi: 10.1080/01621459.1979.10482505

Graf, M., Marín, J. M., & Molina, I. (2019). A generalized mixed model for skewed distributions applied to small area estimation. *Test*, *28*(2), 565–597. doi: 10.1007/s11749-018-0594-2

Hall, P., & Maiti, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(2), 221–238. doi: 10.1111/j.1467-9868.2006.00541.x

Isidro, M., Haslett, S., & Jones, G. (2016). Extended structure preserving estimation (espree) for updating small area estimates of poverty. *Annals of Applied Statistics*, *10*(1), 451–76. doi: 10.1214/15-AOAS900

Jiang, J., Lahiri, P., & Wan, S.-M. (2002). A unified jackknife theory for empirical best prediction with m-estimation. *Annals of Statistics*, *6*(30), 1782–1810. doi: 10.1214/aos/1043351257

Jiongo, V. D., Haziza, D., & Duchesne, P. (2013, 07). Controlling the bias of robust small-area estimators. *Biometrika*, *100*(4), 843-858. doi: 10.1093/biomet/ast030

Krennmair, P., Würz, N., Schmid, T., & Tzavidis, N. (2024). Random forests and mixed effects random forests for small area estimation of general parameters: A poverty mapping case study in Mozambique. *Submitted*.

Lee, Y., Rojas-Perilla, N., Runge, M., & Schmid, T. (2023). Variable selection using conditional AIC for linear mixed models with data-driven transformations. *Statistics and Computing*, *33*(27). doi: 10.1007/s11222-022-10198-9

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. *The Annals of Applied Statistics*, *12*(2), 685–726. doi: 10.1214/18-AOAS1161SF

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, *38*(3), 369–385. doi: 10.1002/cjs.10051

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G., & Breidt, F. J. (2008). Non-parametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *70*(1), 265–286. doi: 10.1111/j.1467-9868.2007.00635.x

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, *85*(409), 163–171. doi: 10.1080/01621459.1990.10475320

Rojas-Perilla, N., Pannier, S., Schmid, T., & Tzavidis, N. (2020). Data-driven transformations in small area estimation. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(1), 121–148. doi: 10.1111/rssa.12488

Wieczorek, J. (2023). Design-based conformal prediction. *Survey Methodology*, *49*(2), 443–473. Retrieved from `https://www150.statcan.gc.ca/n1/pub/12-001-x/index-eng.htm`

Zhang, L.-C., & Chambers, R. (2004). Small area estimates for cross-classifications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(2), 479–496. doi: `10.1111/j.1369-7412.2004.05266.x`

Zhang, L.-C., Sanguiao-Sande, L., & Lee, D. (2024). Design-based predictive inference. *Journal of Official Statistics*, *42*, 404–432. doi: 10.1177/0282423X241277719