STATISTICS IN TRANSITION new series, August 2020
Special Issue, pp. 49–54, 10.21307/stattrans-2020-027
Submitted – 31.01.2020; paper accepted for published – 30.06.2020

Discussion of "Small area estimation: its evolution in five decades", by Malay Ghosh

David Newhouse[1]

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 (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

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

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 R2, 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[2]. 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 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

[2] Important exceptions are Pokhriyal and Jacques (2017) and Erciulescu et al (2018).

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.

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.

ELBERS, CHRIS, PETER LANJOUW, PHILLIPPE GEORGE LEITE, (2008). Brazil within Brazil: Testing the poverty map methodology in Minas Gerais, The World Bank.

ENGSTROM, RYAN, JONATHAN HERSH, DAVID NEWHOUSE, (2017). Poverty from space: Using high-resolution satellite imagery for estimating economic well-being. The World Bank.

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.

GONZÁLEZ-MANTEIGA, W., LOMBARDÍA, 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.

HAY, SIMON I., et al., (2009). A world malaria map: Plasmodium falciparum endemicity in 2007, PLoS medicine 6.3.

JEAN, NEAL, et al., (2016). Combining satellite imagery and machine learning to predict poverty, Science 353.6301, pp. 790–794.

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.

LANGE, S., UTZ JOHANN PAPE, PETER PÜTZ, (2018). Small area estimation of poverty under structural change, The World Bank.

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.*

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.

MOLINA, I., J. N. K. RAO, (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, 38.3, pp. 369–385.

MOLINA, I., MARHUENDA, Y., (2015). sae: An R package for small area estimation. *The R Journal*, 7(1), pp. 81–98.

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.

STEELE, JESSICA, E., et al., (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14.127, 20160690.

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.

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.

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.

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.

ZHAO, QINGHUA., (2006). User manual for POVMAP, World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ ManualPovMap. pdf.