NEW DEVELOPMENTS IN SMALL AREA ESTIMATION: A PRACTITIONER'S PERSPECTIVE

David Newhouse
Senior Economist
Development Economics Data Group

February 28, 2024

Small area estimation

1. Allows surveys to “borrow strength” from richer auxiliary data

2. Improves estimates of socioeconomic indicators by making them more granular and precise
    - Can improve monitoring and evaluation
    - Can improve targeting of policy interventions
    - Can assess potential bias in surveys

3. How to do it? Practitioners choose
    1. Outcomes of interest
    2. Auxiliary data
    3. A statistical method
    4. A model
    5. A software package to generate estimates and diagnostics

Small area estimation

- Lots of choices, some methodological issues are being worked through
  - Best to be cautious about strong assertions
  - Performance of different methods depends on underlying data

- Important research agenda to guide practice, useful for researchers to talk to practitioners
  - Important to evaluate and interpret evidence carefully

1. Outcomes

1. Small area estimation well-suited for indicators that are
    1. Useful for policymakers
    2. Expensive to collect from respondents
    3. Can be predicted using auxiliary data

2. Most small area estimation has traditionally focused on poverty, but can be applied to other indicators
    - Labor outcomes, human capital outcomes, political views, etc.

2. Auxiliary data

1. Recent census or administrative data preferred
   - But not always available

2. Geospatial indicators provide a feasible alternative in many settings
   - Public indicators are exceptionally good at picking up spatial variation in population density (Wardrop et al, 2108, Engstrom et al, 2020)
   - These are correlated with lots of indicators of interest (population, poverty, wealth, connectivity, labor force participation, educational attainment)
   - Quality and availability of publicly available indicators derived from imagery is improving rapidly
   - Raw imagery also available
   - Proprietary imagery has more potential but largely unexplored.
   - Imagery generally not as informative as current census or administrative data, can suffer from measurement error (cloud cover)

3. Other forms of big data such as CDR are intriguing but may suffer from selection bias

Example geospatial data

Variable	Source	Resolution	Year
Population structure	WorldPop	100 m	2018
Population density	WorldPop	100 m	2018
Temperature	TerraClimate	4 km	2018
Palmer Draught Severity Index (PSDI)	TerraClimate	4 km	2018
Distance to OSM major roads	WorldPop, Open Streetmap	100 m	2016
Radiance of night-time lights	VIIRS	500 m	2018
Net primary production	FAO Remote Sensing for Water Productivity (WaPOR) 2.1	240 m	2018
Rainfall	Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS)	5.5 km	2018
Elevation	NASA's SRTM Digital Elevation (3 arc seconds spatial resolution)	30 m	2018
Cellphone tower count	The OpenCell ID project	1 km	2022
Years since change to impervious surface	Tsinghua via Google Earth Engine	30 m	2018
Building count	Worldpop	100 m	2018
Coefficient of variation on buildings	Worldpop	100 m	2018
Land cover classifications	Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3	100 m	2018

2. Auxiliary data

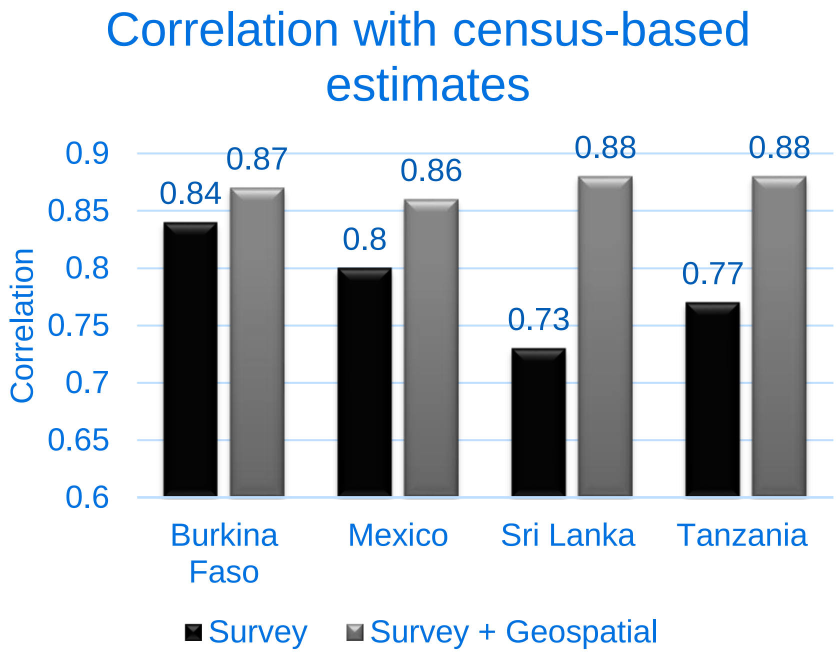
1. Should we use geospatial data or old census data? Depends on:
    - Age of census and how quickly regional patterns in outcome have changed
    - The nature of the sample
        - How many EAs, how many households per EA, geographic stratification
    - How informative the auxiliary data census variables and geospatial variables are for the indicator

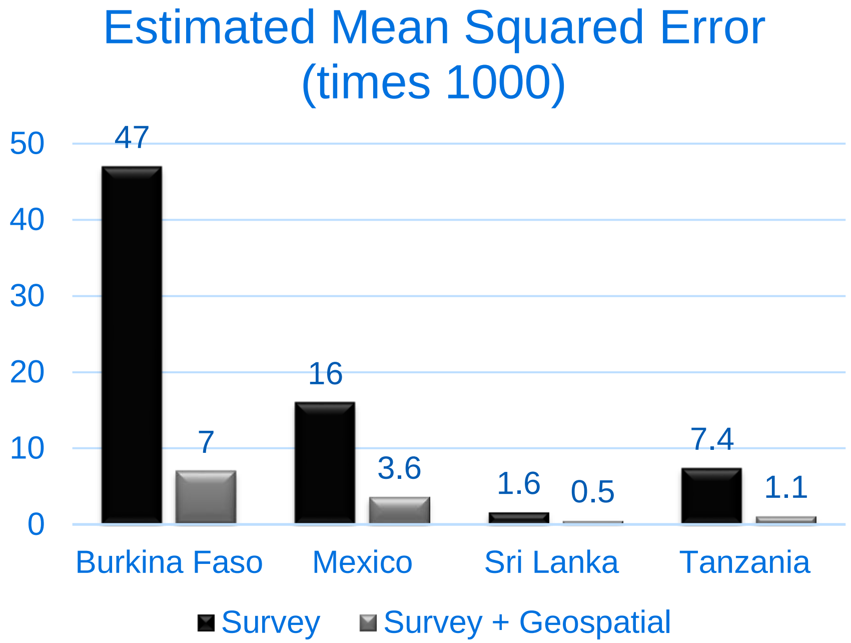
2. Evidence from multiple contexts that small area estimates of monetary and non-monetary poverty with geospatial indicators can perform well
    - Because of a systematic relationship between poverty and population density

3. 5-year old census data gives slightly more accurate predictions than current geospatial indicators in Mexico (Newhouse et al, 2022)
    - More empirical evidence on this in different contexts would be useful, evidence from Mozambique is coming soon.
    - At some point censuses become so old that it is better to use current geospatial data

Small area estimation of poverty tends to improve on direct estimates

Relatively small improvements in correlations can enable to large improvements in MSE





Sources: Masaki et al (2022), Newhouse et al (2023), Edochie et al (Forthcoming)
Notes: Results based on actual household survey data. Survey estimates are direct estimates, survey + geospatial are EBP estimates using a linear mixed model.

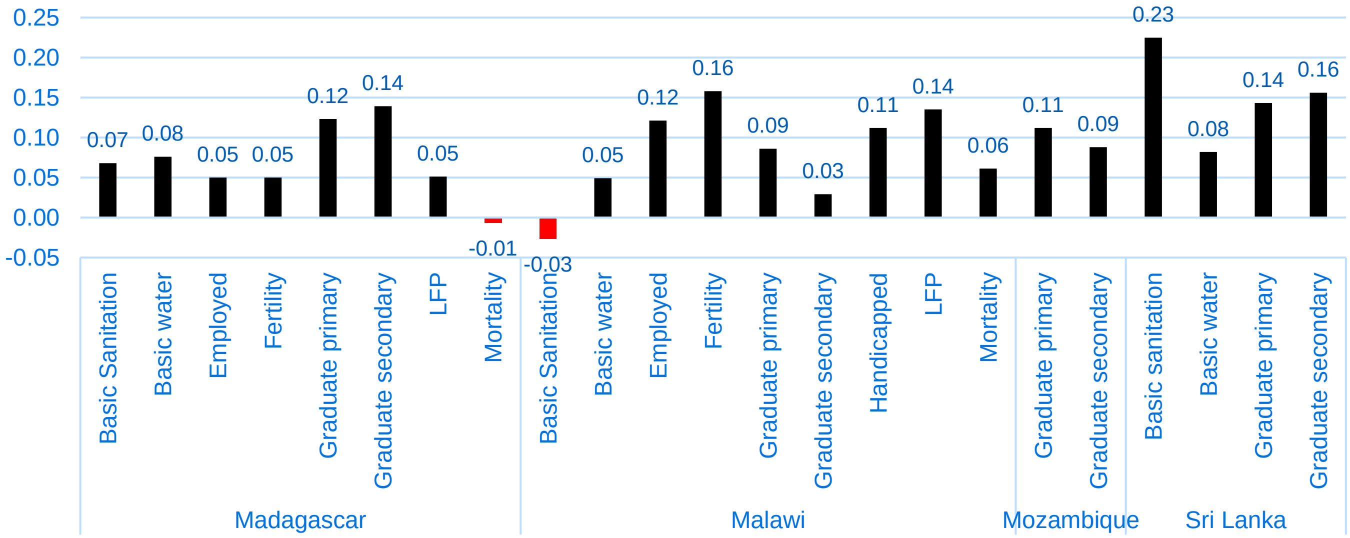
Human capital indicators show potential but performance varies

Preliminary results suggest that SAE estimates improve on direct estimates of human capital indicators in 17 out of 19 cases, worsen accuracy in 2

Can we better understand how much geospatial SAE helps without a census?

Small increases in correlation are valuable in terms of increasing efficiency

Increase in correlation with census relative to direct survey estimates when using geospatial SAE



Note: Preliminary results based on 100 simulations of survey data

3. Statistical method

1. Many candidates, divided along 3 dimensions:
   A. Bayesian vs. Frequentist
   B. For frequentist models, Purely synthetic vs. empirical Bayesian
   C. Parametric vs. Non-parametric (tree-based)

2. Bayesian methods work well but frequentist is preferred
   A. Close theoretical connection between Bayesian approach and parametric bootstrap (Hirose and Lahiri, 2021)
   B. Practitioners often are more comfortable with frequentist methods
   C. Adjusted maximum likelihood can avoid issues in variance component estimation (i.e. Li and Lahiri 2010)
   D. Software applying EBP with parametric bootstrap arguably easier to understand and use

3. Statistical method

1. Frequentist method

	Synthetic	Empirical Bayesian
Parametric model	ELL, M-quantile	EBP
Tree-based model	Machine learning (XGboost)	Mixed Effect Random Forests (MERF)

- Das and Haslett (2019) compare ELL, M-quantile, and EBP and stress that the methods are based on different sets of assumptions and "no method is uniformly best"
    - Evidence so far suggests that this is generally true, although EBP and Xgboost tend to do well, at least with geospatial data (Merfeld and Newhouse, 2023)
    - MERF does not do as well as other options when we have tested it so far
    - But future work could improve it or add gradient boosting with conditional random effects

Transformations

1. Empirical Bayesian methods assume normal error terms
   - Literature has explored non-normal error terms but not yet implemented in software
   - Transformations are important to make normality assumptions more palatable
   - Logs traditionally used for welfare/poverty estimation:

$$\ln Y_{rah} = X_{rah}\beta_1 + \bar{X}_{ra}\beta_2 + D_r\beta_3 + \nu_a + \epsilon_{rah}$$

$\ln Y_{rah}$ = log per capita consumption of household h within target area a and region r.
X_{rah} = Household characteristics present in auxiliary (census) data, such as assets and demographic characteristics
$\bar{X}_{ra}$ = Household characteristics present in auxiliary (census) data, such as assets and demographic characteristics, aggregated to area level
D_r = regional dummy variables
$\nu_a = E\left[\nu_a | Y_{ragh}\right] + v_a^*$ is a conditional random effect, conditioned on survey data Y_{ragh}
$v_a^* \sim N(0, \sigma_v^2(1 - \gamma_a))$, $\epsilon_{rah} \sim N(\sigma_\varepsilon^2)$, $\gamma_a = \frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_\varepsilon^2 \sum_h w_h^2}{\left(\sum_h w_h\right)^2}}$, w_h are sample weights

Transformations

- Other transformations are available and probably preferable
    - Povmap R package supports:
        - Log-shift: $\ln(Y_{rah} + k)$, k chosen to make distribution close to normal
        - Box-Cox
        - Dual
        - Arcsin for proportions
        - Rank-order: Y_{rah} transplanted into normal distribution depending on welfare rank, can be back-transformed with approximations
    - More research on relative merits of different transformations in different settings would be worthwhile

4. Selecting Models

- Model selection can be manual or using an automated algorithm (LASSO or Stepwise)

- LASSO (Least Absolute Selection and Shrinkage Operator) generally works well
    - Prevents overfitting by essentially equalizing in out of sample R2 using cross-validation or other methods
    - Not path dependent like stepwise
    - Packages exist to make this straightforward
    - Still need to check the model to make sure coefficients make sense

Should we include household-level variables in the model?

- Baseline model:

$$\ln Y_{rah} = X_{rah}\beta_1 + \bar{X}_{ra}\beta_2 + D_r\beta_3 + \nu_a + \epsilon_{rah}$$

R=region, a=target area, h = household
Y=household per capita consumption, X=predictors, D = Dummy variables

- How important are household-level variables? Could we estimate a “unit-context model”

$$\ln Y_{rash} = \bar{X}_{ras}\beta_1 + \bar{X}_{ra}\beta_2 + D_r\beta_3 + \nu_a + \epsilon_{rah}$$

Where s is a “sub-area” like a census block or village, below the target area.

Or a sub-area model

$$\hat{P}_{ras} = \bar{X}_{ras}\beta_1 + \bar{X}_{ra}\beta_2 + D_r\beta_3 + \nu_a + \epsilon_{rah}$$

Where $\hat{P}_{ras}$is the estimated poverty rate in the sub-area

Should we use household level variables?

1. Almost always good to include contextual auxiliary variables in model
    - Reduces variance of random effect

2. If household census data is old, it could be better to omit household level variables and only use aggregates (Lange, Putz, and Pape, 2022)

- Why? Usually aggregates can linked directly between surveys and censuses through geographic identifiers, while household characteristics cannot be

- Using household characteristics in model requires assuming that survey and census variables follow the same distribution. This worsens “age bias” when census is old.

- Dropping household-level variables often has minor impacts on accuracy

Geospatial auxiliary data

- Geospatial data is also an option. If household census data is old, it may be better to use current geospatial data than old household census data

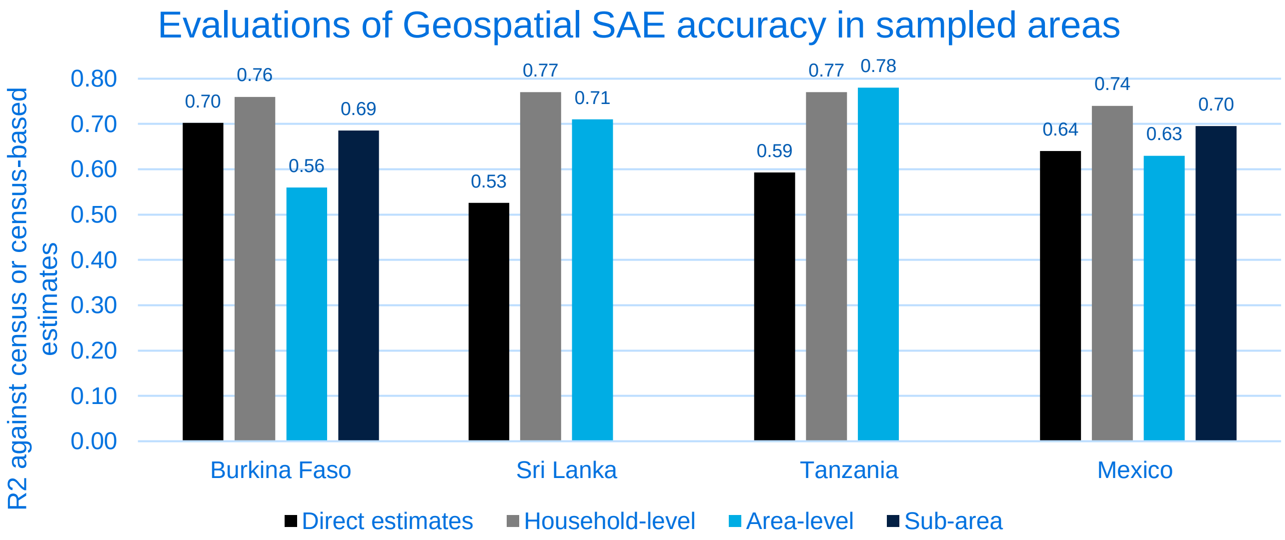
Geospatial data usually only available at aggregate level (villages or small grids), requires unit-context or sub-area level model

3. Auxiliary data should be used at lowest level of aggregation possible
    - When sub-area auxiliary data is available, unit-context or sub-area models are better options than area level models
    - Helps better protect against selection bias in sample data
    - In addition, area-level models require survey estimates of variance at small area, which are usually imprecise
        - Variance smoothing can help
    - Sub-area and unit-context models use identical auxiliary data
        - But may be minor advantage to modeling transformed continuous welfare measure rather than poverty rates

What does the evidence say?

- Unit-context models either more accurate (Burkina Faso, Sri Lanka, Mexico) or as accurate (Tanzania) compared to area-level models
  - Hypothesis: U-C has bigger advantage when there is selection bias in the sample
  - U-C model can better "fill in gaps" in missing EAs than area-level models i

- Estimates for sampled areas much more accurate than results for non-sampled areas (results not shown)



What about tree-based machine learning?

Often does slightly better than linear EBP for poverty and wealth due to more flexible modeling
Better suited for cases where interactions and non-linearities matter

Two important downsides:

- No random effect. Sample data used only to calibrate model, not as direct input into estimate
- Model difficult to explain and parameters difficult to communicate. Model consists of multiple decision trees with many parameters

One proposed method combines random effects with tree-based model

(Krennmair and Schmid, 2022). Jury is still out on when it does well. More research needed.

Could do more to experiment with interactions in linear EBP models

Benchmarking

- Ensures that population-weighted average of small area estimates match direct survey estimates at higher “regional” level
  - Direct estimates at that level are reliable and publishable
  - Attractive to national statistical offices to ensure consistency between small area estimates and survey estimates at higher level
  - Either slightly helps or worsen accuracy depending on data, usually not a big deal

- Simple ratio-benchmarking generally works ok in the cross-section (Pfeffermann et al, 2014)
  - Multiply all estimates within a region by a fixed constant to ensure agreement
  - Benchmarking ratios are useful diagnostic information about model accuracy
  - Can be useful for bias reduction when using transformations
  - Simple ratio benchmarking can theoretically estimate poverty rates > 100% in very poor areas
  - In these cases one can benchmark the “non-poverty rate” instead.

Povmap package

R Povmap package: An Extension of the EMDI package

Version 1.0 available on CRAN, version 2.0 in development and available on github

Povmap/EMDI is designed to make SAE easy for practitioners:

1. Unit-level, unit-context models, area-level models of means and headcounts
2. Calculates point estimates and MSE estimates
3. Include options for sample and population weights
4. Automates many choices for transformations, including “adaptive transformations”)
5. Automates benchmarking to survey-based estimates at higher level
    - Both internal and external benchmarking
6. Options to parallelize across multiple cores for increased speed
7. Integrates useful code for diagnostics, reporting, and output
8. Integrates nicely with Stata
9. Excellent documentation in three vignettes

Povmap package

Version 2.0 features coming:

1. Options to further speed up computation

- Calculate expected value of headcount and mean instead of monte-carlo simulations
- Use data.table package for data processing
- Compute subset of indicators
- Drop duplicates from auxiliary data when estimating unit-context models

2. Support for “twofold models” (Marhuenda et al 2018) with area and sub-area random effects

3. Support for “ELL” models (Elbers, Lanjouw, and Lanjouw, 2003)

4. Support for Machine Learning models (extreme gradient boosting) with standard errors

5. Consolidated documentation

Geolink package

Software to facilitate linking publicly available geospatial indicators to survey data

Working prototype for rainfall and night-time lights. More indicators and documentation currently being added. Expected release fall 2024

Conclusion: Recommendations for good practice

- Survey data should try to cover all target areas

- Auxiliary data: If old census data is available, can use both old census and geospatial data
  - Model selection procedure can decide which variables to use
  - Important research agenda to experiment with improved indicators

- Method
  - Use one-fold EBP model unless there is a good reason not to
  - Such as concerns about non-linearities and interactions or outliers

- Model:
  - Unit-context models are a useful option when unit-level census data is old or unavailable
    - Including household variables can worsen age bias when census is old
    - Predicting well across areas matters much more than predicting well within areas.

Conclusion: Recommendations for good practice

- Be wary of general claims about statistical methods
  - Accuracy of all methods depends on underlying data
  - Parameters of model-based simulations can be manipulated,
    - Design-based simulations are more informative but may not consider all relevant cases
  - Most simulations testing methods do not explicitly consider cases where data is biased
    - For example, biased census data due to age
    - Or biased survey data due to selection bias (informative sampling) or measurement error
- Consider using R povmap package for small area estimation due to flexibility, ease of use, features, and pace of development

Helpful resources

- UN toolkit on small area estimation available at: https://unstats.un.org/wiki/display/SAE4SDG/

- World Bank guidelines on poverty mapping available at https://www.worldbank.org/en/events/2023/02/07/guidelines-to-small-area-estimation-for-poverty-mapping

- Consultation draft of “Small Area Estimation with Geospatial Data: A Primer” available at https://unstats.un.org/iswgh

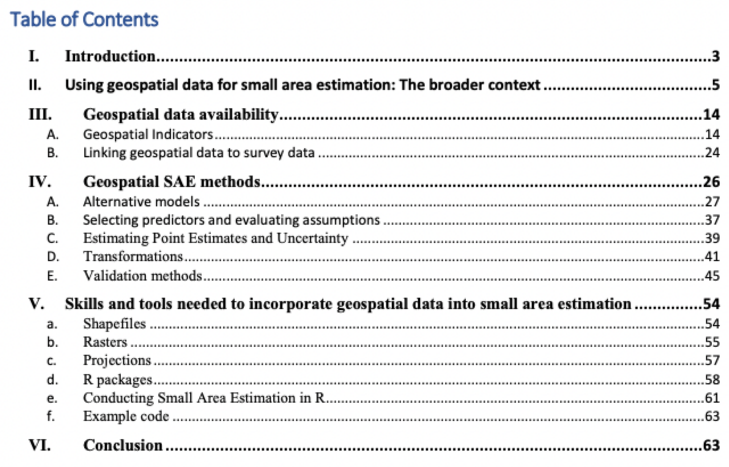


Table of Contents

Thank you!

WORLD BANK GROUP

References

Das, S., & Haslett, S. (2019). A comparison of methods for poverty estimation in developing countries. International Statistical Review, 87(2), 368-392.

Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. Econometrica, 71(1), 355-364.

Engstrom, R., Newhouse, D., & Soundararajan, V. (2020). Estimating small-area population density in Sri Lanka using surveys and Geo-spatial data. *PloS one*, *15*(8), e0237063.

Guadarrama, M., Molina, I., & Rao, J. N. K. (2016). A comparison of small area estimation methods for poverty mapping. Statistics in Transition new series, 17(1), 41-66.

Hirose, M. Y., & Lahiri, P. (2021). Multi-Goal Prior Selection: A Way to Reconcile Bayesian and Classical Approaches for Random Effects Models. *Journal of the American Statistical Association*, *116*(535), 1487-1497.

Krennmair, P., & Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. Journal of the Royal Statistical Society Series C: Applied Statistics, 71(5), 1865-1894.

Lange, S., Pape, U. J., & Pütz, P. (2022). Small area estimation of poverty under structural change. *Review of Income and Wealth*, *68*, S264-S281.

Marhuenda, Y., Molina, I., Morales, D., & Rao, J. N. K. (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), 1111-1136.

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2022). Small area estimation of non-monetary poverty with geospatial data. *Statistical Journal of the IAOS*, *38*(3), 1035-1051.

Merfeld, J. D., & Newhouse, D. (2023). Improving Estimates of Mean Welfare and Uncertainty in Developing Countries (No. 10348). The World Bank.

Newhouse, D., Merfeld, J., Ramakrishnan, A. P., Swartz, T., & Lahiri, P. (2022). Small Area Estimation of Monetary Poverty in Mexico using Satellite Imagery and Machine Learning.

Newhouse, D. (2023). Small Area Estimation of Poverty and Wealth Using Geospatial Data: What have We Learned So Far?. *Calcutta Statistical Association Bulletin*, 00080683231198591.

Pfeffermann, D., Sikov, A., & Tiller, R. (2014). Single-and two-stage cross-sectional and time series benchmarking procedures for small area estimation. *Test*, *23*, 631-666.

Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., ... & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. *Proceedings of the National Academy of Sciences*, *115*(14), 3529-3537.