

USING GEOSPATIAL DATA FOR SMALL AREA ESTIMATION



WORLD BANK GROUP

David Newhouse
JPSM Session on Small Area
Estimation

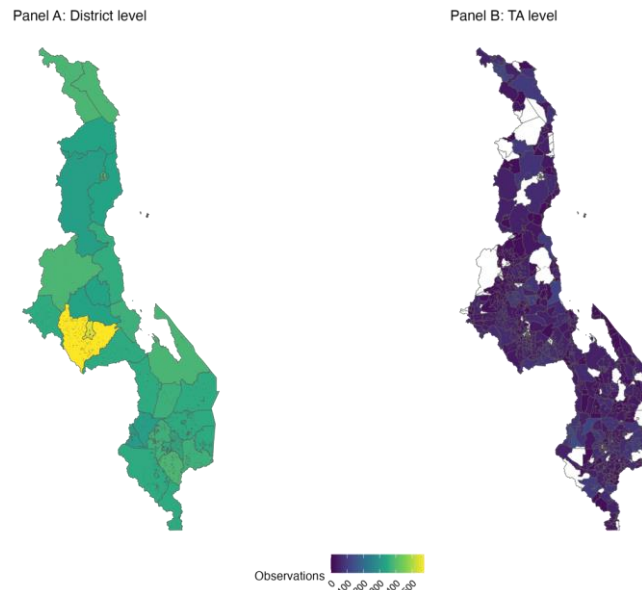
March 27, 2024

Outline of talk

1. Motivation for geospatial small area estimation
2. What types of geospatial data are available?
3. What do preliminary results show about the accuracy of geospatial small area estimates?
4. Existing and forthcoming tools for small area estimation using geospatial data

Motivation

- Important to monitor socioeconomic indicators including Sustainable Development Goals
 - Household probability sample surveys are often used to measure key socioeconomic outcomes
 - But surveys typically unable to generate reliable estimates for small areas
 - For example, not nearly enough households to estimate Traditional Authority (TA) level outcomes in Malawi using 2019 Integrated Household survey



Motivation

- Using Small Area Estimation to combine surveys with more geographically comprehensive data sources can help. SAE is:
 - Useful for targeting and evaluating interventions
 - Can generate estimates in areas not covered by survey
 - although these are often significantly less accurate than estimates for sampled areas
 - Can potentially assess and partially correct for selection bias in first stage of sample surveys
 - Can help generate more reliable estimates for small population subgroups

Auxiliary data

- Traditionally census or administrative data used, but geospatial data provide an intriguing alternative
 - Recent census data may not exist or may not be obtainable
 - Availability of geospatial indicators has exploded
 - Google earth engine and Microsoft planetary computer
- Geospatial data is:
 - Geographically comprehensive
 - Unlike mobile phone data, not subject to selection bias
 - Updated frequently
 - Very geographically granular,
 - Can often link surveys at EA level or small (admin-4) admin level
 - Can very occasionally link surveys at household level
 - Opens up new ways to use auxiliary data at sub-area level
 - Often publicly available, at least for many useful indicators
 - A second-best option when recent census data are unavailable

Partial list of satellites

Satellite	Frequency	Resolution	Availability
MODIS	Globe every 1 to 2 days	250 m	Public
LANDSAT	Globe every 8 days	30 m	Public
Sentinel 2	Globe every 10 days	10 m	Public
Planet	Globe every day	3-5 m	Private
Maxar	60% every month	0.5 m	Private

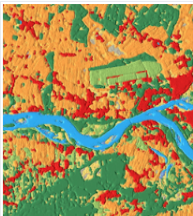
Example: Land cover data in Google Earth Engine

Earth Engine Data Catalog

Search English

HomeView all datasetsBrowse by tagsLandsatMODISSentinelPublisherCommunityAPI Docs

Dynamic World V1



Dataset Availability
2015-06-27T00:00:00Z–2024-03-11T14:47:50

Dataset Provider
[World Resources Institute Google](#)

Earth Engine Snippet

```
ee.ImageCollection("GOOGLE/DYNAMICWORLD/V1")
```

Tags
globalgooglelandcoverlandusesentinel2-derived

DescriptionBandsImage PropertiesTerms of UseCitationsDOIs

Resolution
10 meters

Bands

Name	Min	Max	Description
water	0	1	Estimated probability of complete coverage by water
trees	0	1	Estimated probability of complete coverage by trees
grass	0	1	Estimated probability of complete coverage by grass
flooded_vegetation	0	1	Estimated probability of complete coverage by flooded vegetation
crops	0	1	Estimated probability of complete coverage by crops
shrub_and_scrub	0	1	Estimated probability of complete coverage by shrub and scrub
built	0	1	Estimated probability of complete coverage by built
bare	0	1	Estimated probability of complete coverage by bare
snow_and_ice	0	1	Estimated probability of complete coverage by snow and ice
label	0	8	Index of the band with the highest estimated probability

Example: Building footprints data in Microsoft Planetary Computer

Microsoft | Planetary Computer [Explore](#) [Data Catalog](#) [Hub](#) [Applications](#) [Documentation](#) [Request access](#)

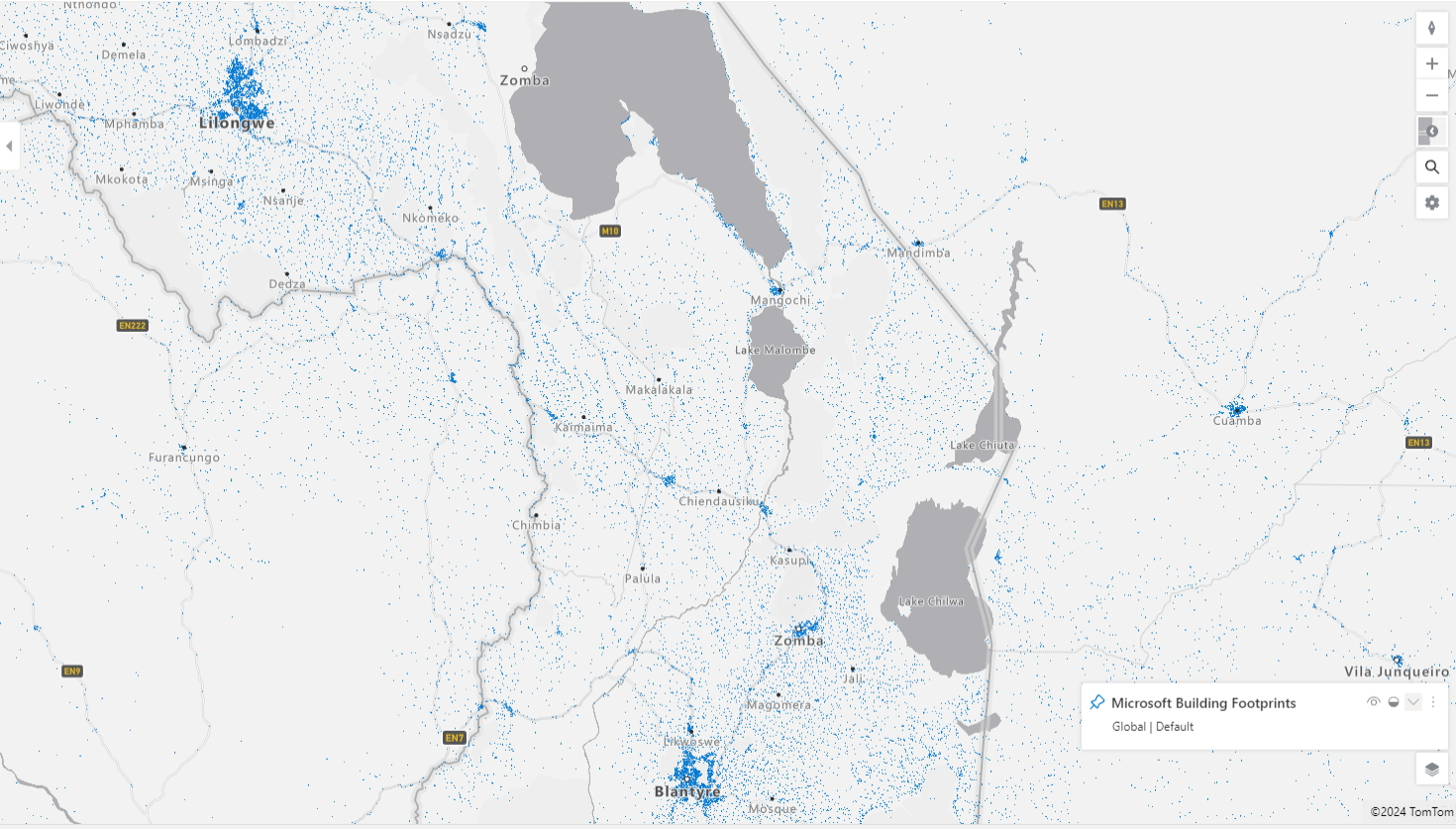
Explore datasets [Advanced](#) [Clear](#)

- Microsoft Building Footprints
- Global
- Default

Microsoft Building Footprints Showing 48 items that matched your filter.

- Building footprints**
07/06/2022 00:00:00 UTC
- Building footprints**
06/14/2022 00:00:00 UTC
- Building footprints**
10/06/2014 — 11/11/2016
- Building footprints**
10/06/2014 — 11/11/2016
- Building footprints**
10/06/2014 — 11/11/2016

[Explore results in the Hub](#)



Microsoft Building Footprints
Global | Default

©2024 TomTom

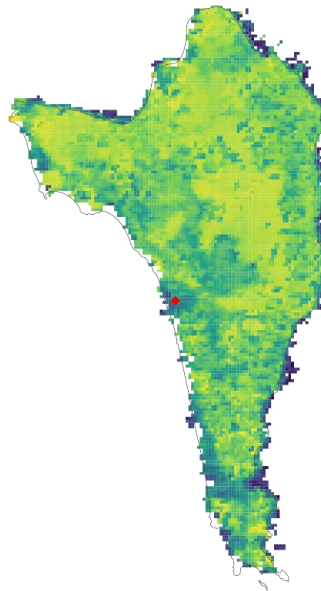
[Sitemap](#) [Contact Microsoft](#) [Privacy](#) [Terms of use](#) [Trademarks](#) [Safety & eco](#) [About our ads](#) [Consumer Health Privacy](#) [Your Privacy Choices](#) © Microsoft 2024

Example: NDVI from MODIS

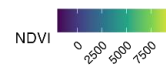
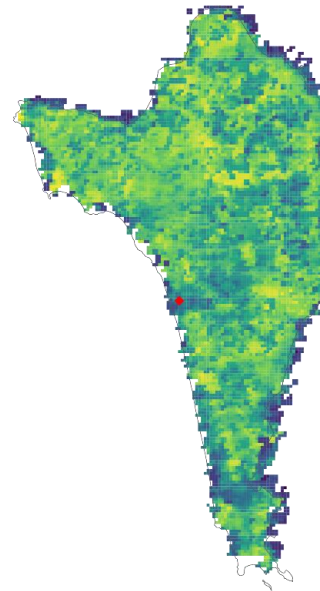
- MODIS images globe every 1 or 2 days with coarse resolution
- NDVI = Normalized Difference Vegetation Index, a measure of vegetation density
- Patterns change during the year

Phu Quoc, Vietnam (2022)

Panel A: January NDVI

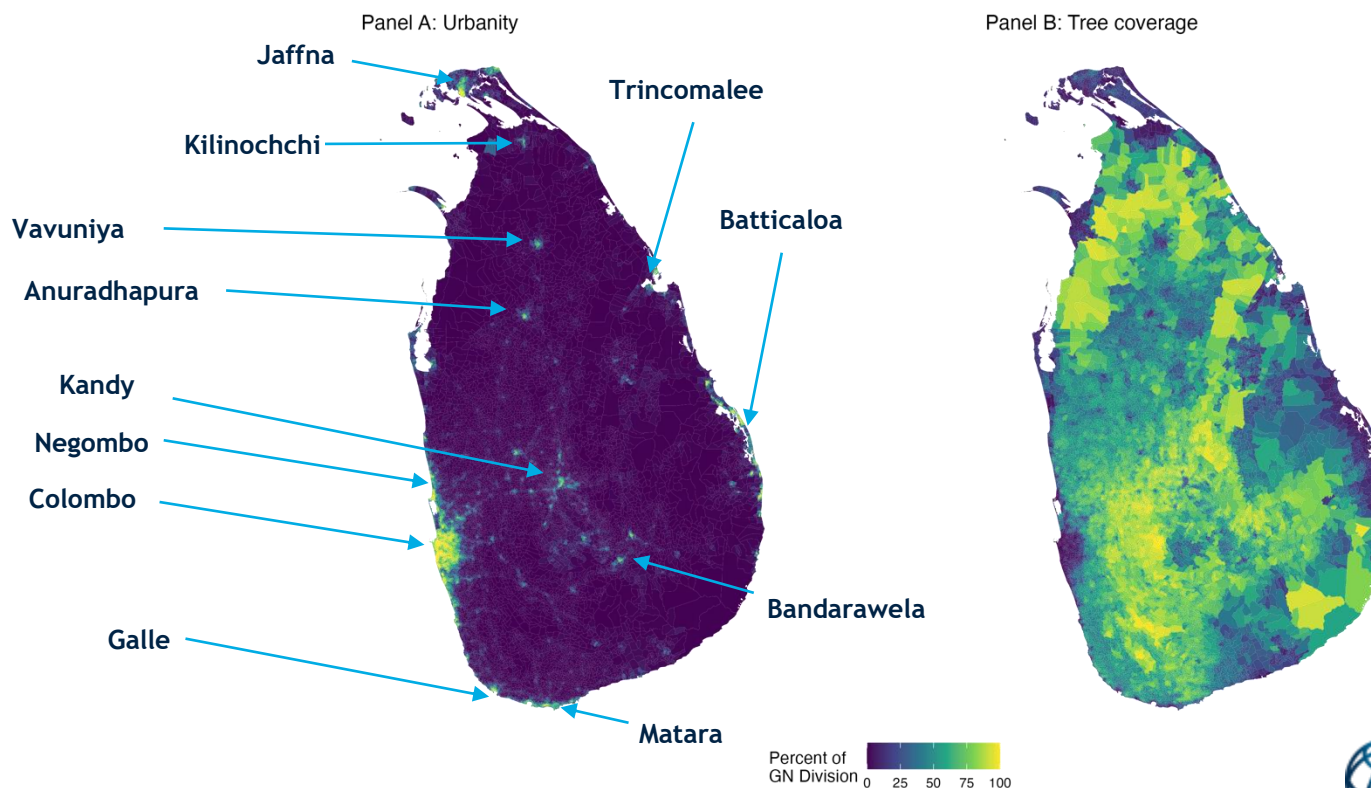


Panel B: July NDVI



Land classification percentages by GN Division in Sri Lanka

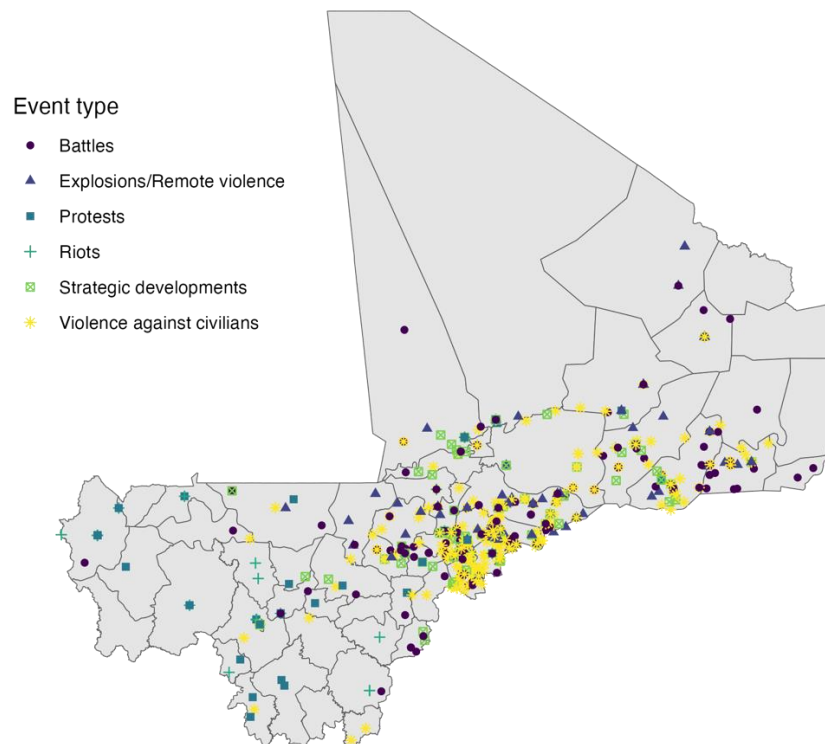
- Clear negative relationship between urbanity and tree coverage
- Urban measure identifies cities and towns



Armed Conflict Location and Events Data (ACLED)

Contains information on location of violent events

Violent events in Mali, 2019



More example indicators

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 (PDSI)	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 University 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

Example Indicators (continued)

1. Building footprints

1. Can be obtained from Google [open buildings](#), Worldpop (Africa only), Meta estimates of population density, [World Settlement footprint](#), Microsoft [Bing Maps](#) via planetary computer
2. Typically cross-sectional, not updated over time
3. Not all products available for all countries

2. Abstract features constructed from imagery

- Constructed by comparing pixels from cloud free image mosaics (now available on google earth engine)
- Predictive of building density, population density, poverty, and wealth
- MOSAICS (Rolf et al, 2021)
 - Though pre-packaged indicators not yet geographically comprehensive
- Sp.feas package in Python

3. Meta relative wealth index

- Predictions of principal components of asset index from Demographic and Health Surveys

Example Indicators (continued)

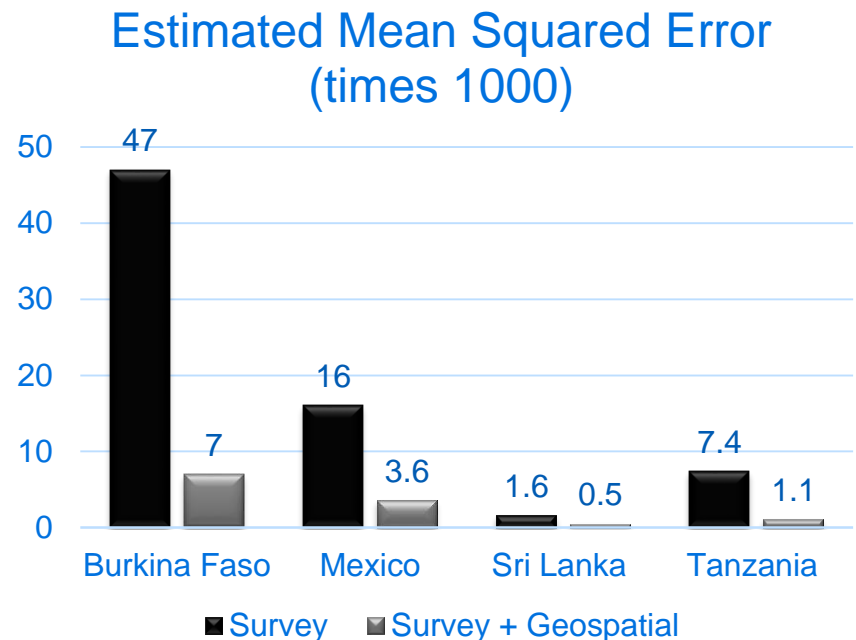
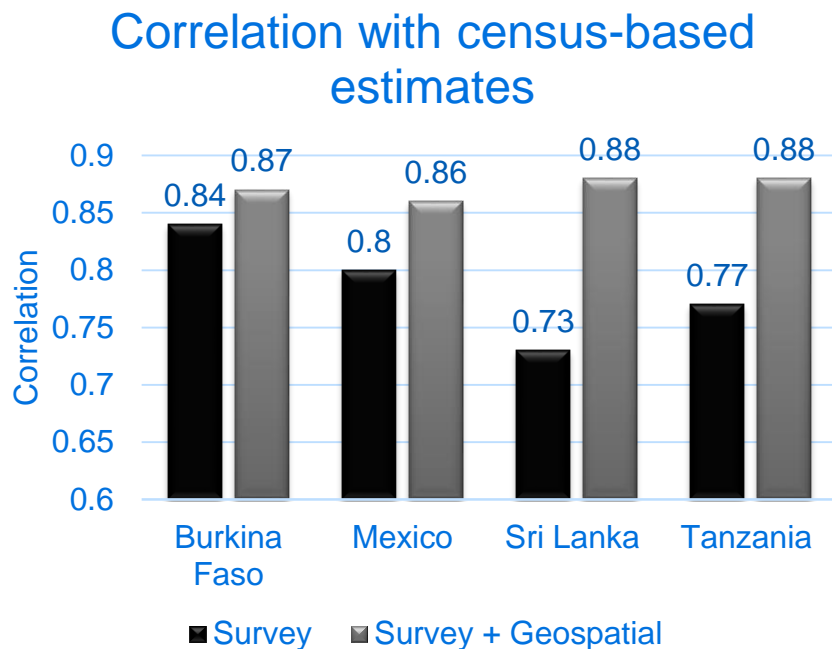
- **Open Streetmap**
 - Public open-source “Wikipedia” collection of roads, buildings and amenities
 - Useful for creating measures of accessibility
 - Concerns about incomplete mapping particularly in rural areas (Barrington-Leigh and Millard-Ball, 2017)
 - Quality of building footprint data is highly variable (Biljecki et al, 2023)
 - Can be supplemented by proprietary information (i.e. google maps)
- **Proprietary indicators**
 - Cars/trucks
 - Roof types
 - Dynamic building footprints – Pace and characteristics of new construction
 - Crop type, yield estimates
 - Information from “internet of things” – car locations, cell phone pings

Geospatial indicators are very predictive of population density

- In Sri Lanka, out of sample R^2 of 0.75 when predicting population density with publicly available indicators (Engstrom et al, 2020)
 - Increases to 0.83 when using proprietary indicators
- In DRC, out of sample R^2 0.79 for out-of-sample predictions of population totals at the microcensus-cluster level (Boo et al, 2022)
- Population density is correlated to many important socioeconomic outcomes
 - Including poverty and wealth (Page and Pande, 2018, Casteneda et al, 2018)
- SAE reduces sampling error at the expense of introducing model error
 - Does this improve accuracy relative to direct estimates?
 - Depends on sample, outcome, predictive power of auxiliary data

Geospatial small area estimates tend to be more accurate than direct estimates for poverty measures

- This enables substantial increases in precision while maintaining coverage



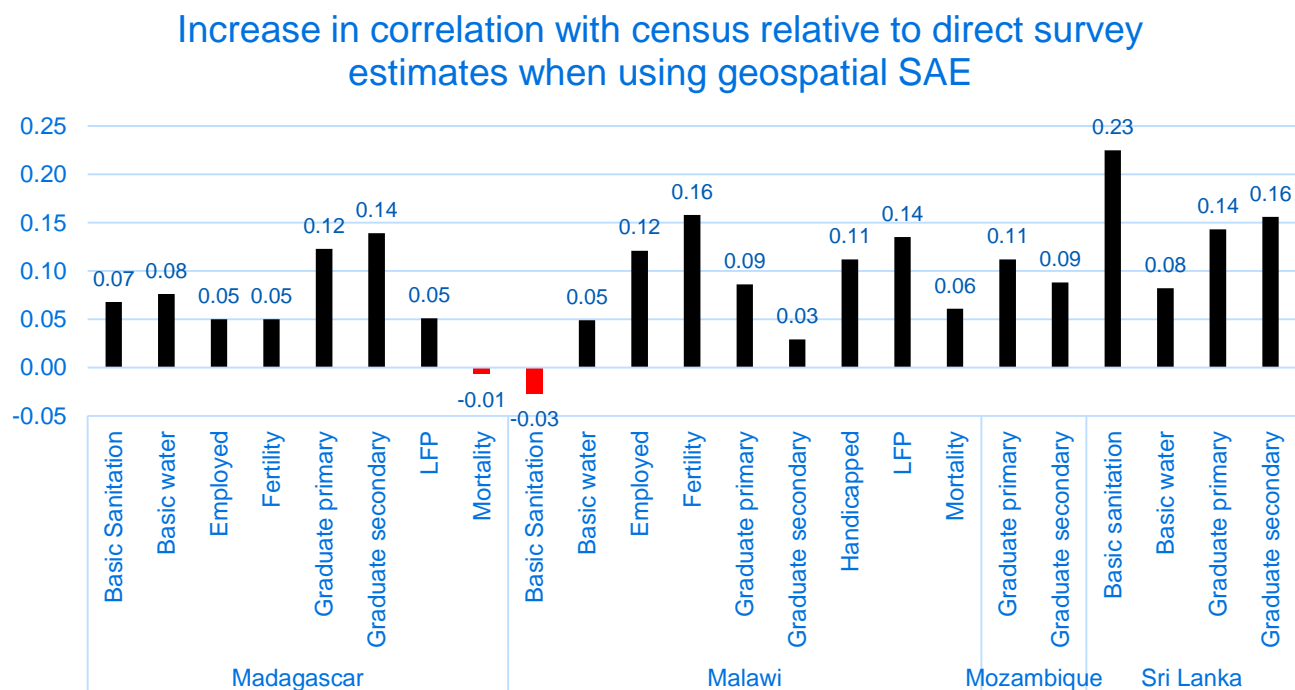
Sources: Masaki et al (2022), Newhouse et al (2023), Edochie et al (Forthcoming)

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

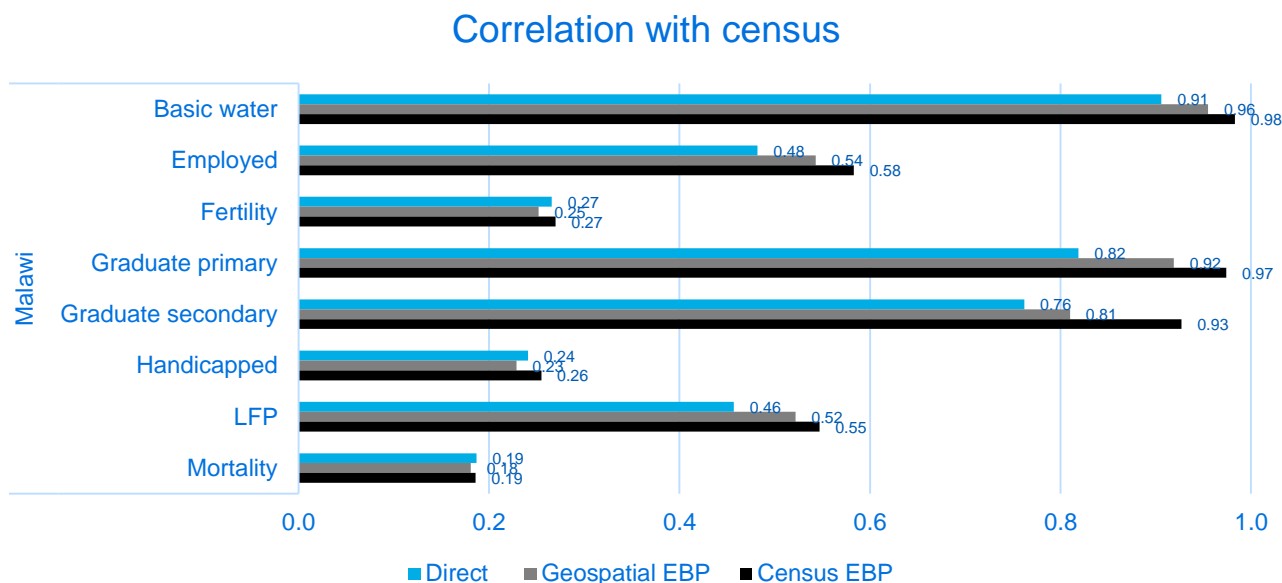
Can we better understand how much geospatial SAE helps without a census?



Note: Correlations taken over sampled areas only, represent average over 100 simulated samples from census data

Some outcomes are less well-suited for SAE in general

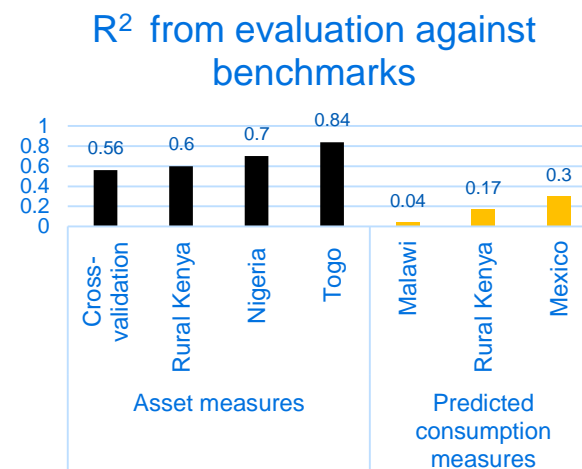
- Fertility, handicapped, mortality difficult to estimate
 - Both with geospatial data used so far and census auxiliary data
- Rare events that are difficult to estimate accurately using sample surveys
 - Accurate estimates need an informative sample
- Important research agenda to develop diagnostics to identify these cases
- Geospatial data are a partial and potentially useful substitute for census data for water, employed, attainment, and labor force participation



Note: Correlations taken over sampled areas only, represent average over 100 simulated samples from census data

Model performance depends critically on the outcome

- For example, Meta relative wealth index is much more accurate when measuring asset poverty than consumption poverty
 - Wealth used as outcome because of data availability
 - Consumption or income is standard measure for poverty measurement



Sources: Chi et al (2022), Gualavisi and Newhouse (2023), Newhouse et al (2022)

- Illustrates importance of democratizing SAE
 - Providing access to geospatial data and tools to national statistics offices (NSOs) and others with sensitive survey data
 - NSOs can model outcomes without having to share geolocations or unit-record census data

Tools to facilitate small area estimation

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 in development, not yet released:

1. Options to further speed up computation
 - Calculate expected value of headcount and mean instead of monte-carlo simulations
 - Compute subset of indicators
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

Conclusions

- Rapid recent advances in publicly geospatial auxiliary data
 - Clearly useful in some important cases
 - Population, poverty, wealth, labor force participation, ag?, Others?
 - Second-best alternative to recent available census data
 - Not well-suited for all relevant outcomes (but neither is census data)
 - Stock of publicly available geospatial indicators should continue to improve
 - More geographically comprehensive
 - New indicators wishlist: Changes in building footprints, crop type, crop yield, cars/trucks
 - Potential of Synthetic Aperture Radar to circumvent cloud cover issues
- Important to better understand when SAE (with geospatial or census data) improves on direct estimates
 - Requires minimum amount of signal in training sample and predictive power of auxiliary data
 - How to measure this without a census benchmark?

Conclusions

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

Table of Contents

I.	Introduction.....	3
II.	Using geospatial data for small area estimation: The broader context	5
III.	Geospatial data availability.....	14
A.	Geospatial Indicators.....	14
B.	Linking geospatial data to survey data	24
IV.	Geospatial SAE methods.....	26
A.	Alternative models	27
B.	Selecting predictors and evaluating assumptions	37
C.	Estimating Point Estimates and Uncertainty	39
D.	Transformations.....	41
E.	Validation methods.....	45
V.	Skills and tools needed to incorporate geospatial data into small area estimation	54
a.	Shapefiles	54
b.	Rasters	55
c.	Projections	57
d.	R packages.....	58
e.	Conducting Small Area Estimation in R.....	61
f.	Example code	63
VI.	Conclusion	63

References

- Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PloS one*, 12(8), e0180698.
- Biljecki, F., Chow, Y. S., & Lee, K. (2023). Quality of crowdsourced geospatial building information: A global assessment of OpenStreetMap attributes. *Building and Environment*, 237, 110295.
- Boo, G., Darin, E., Leasure, D. R., Dooley, C. A., Chamberlain, H. R., Lázár, A. N., ... & Tatem, A. J. (2022). High-resolution population estimation using household survey data and building footprints. *Nature communications*, 13(1), 1330.
- Castañeda, A., Doan, D., Newhouse, D., Nguyen, M. C., Uematsu, H., & Azevedo, J. P. (2018). A new profile of the global poor. *World Development*, 101, 250-267.
- Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3), e2113658119.
- Edochie, I., D. Newhouse, T. Schmid, and N. Wurz, "Povmap: Extension to the emdi package for small area estimation", available at Comprehensive R Archive Network
- Edochie, I., D. Newhouse, T. Schmid, N. Tzavidis, E. Foster, A. Ouedraogo, A. Sanoh, and A. Savadogo (forthcoming), "Small area Estimates of Poverty in Four West African countries", mimeo
- 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.
- Gualavisi, M., & Newhouse, D. L. (2022). Integrating Survey and Geospatial Data to Identify the Poor and Vulnerable, Policy Research Working Paper no. 10257.
- 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.
- 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.
- Page, L., & Pande, R. (2018). Ending global poverty: Why money isn't enough. *Journal of Economic Perspectives*, 32(4), 173-200.
- Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., ... & Hsiang, S. (2021). A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1), 4392.