IMPROVING ESTIMATES OF MEAN WELFARE AND UNCERTAINTY IN DEVELOPING COUNTRIES

WORLD BANK GROUP

Josh Merfeld
David Newhouse

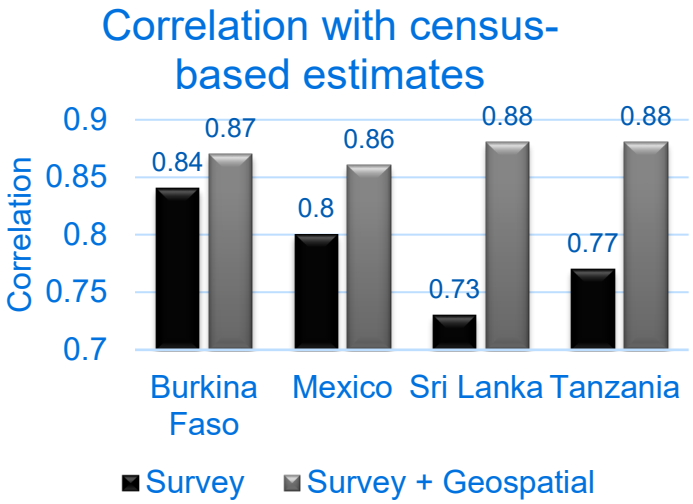
June 5, 2023

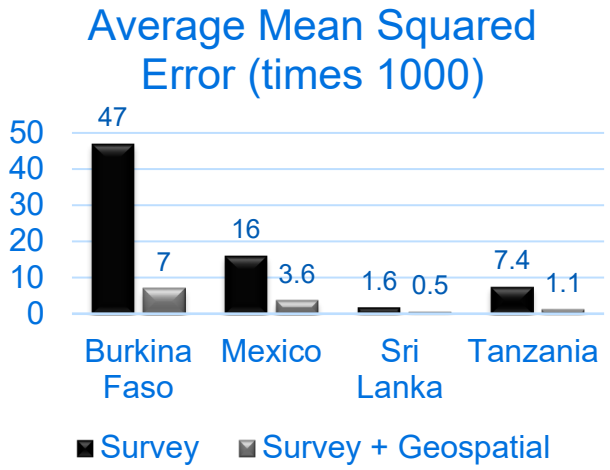
Funding provided by the Knowledge for Change Program's phase-IV programmatic research project "Understanding Trends in Sub-National Differences in Economic Well-Being in Low and Middle- Income Countries"

Background

1. Geospatial data are predictive of wealth, welfare, and poverty in a variety of contexts
   - Jean et al, 2016, Burke et al, 2021, Mcbride et al, 2021, Yeh et al, 2020, Chi et al, 2022, Khachiyan et al, 2022, Engstrom et al, 2022, many others.
   - Accuracy varies greatly depending on context, indicator, training data, estimation method, evaluation method, and benchmark evaluation data

2. Adding geospatial data significantly improves on survey data for small area poverty estimates





Source: Edochie et al (forthcoming), Masaki et al (2022), Newhouse et al (2022)

WORLD BANK GROUP

Which prediction method generates the most accurate point and uncertainty estimates?

1. This paper evaluates four candidate prediction methods for combining geospatial and survey data to generate small area estimates of mean asset index for target area level

1. Three tree-based machine learning methods and commonly used linear mixed model
   - Separately for in and out-of-sample areas
   - We do not yet compare with convolutional neural networks trained directly to imagery, this is planned for future
   - Focus on predictive accuracy, but parsimony and interpretability also matter (Efron, 2020)

- Evaluate against census data in four countries

- Evaluate a proposed residual bootstrap procedure for estimating uncertainty using ML estimators that takes spatial correlation within target areas into account
   - Fills a gap in the literature on how to estimate uncertainty effectively when using machine learning methods

Candidate prediction method 1: Empirical Best Predictor (EBP) linear mixed model

- Household welfare modeled as a function of village characteristics with area-level conditional random effect
  - Battese, Harter, and Fuller 1988, Jiang and Lahiri 2006, Molina and Rao, 2010, Masaki et al (2022)

$$\mathrm{G}(Y_{rash}) = X_{ras}\beta_1 + \bar{X}_{ra}\beta_2 + D_r\beta_3 + \nu_{ra} + \epsilon_{rash}$$

$\mathrm{G}(Y_{rash})$ is transformed welfare for household h in sub-area s, area a, region r, X_{ras} are geospatial indicators, D_r are regional dummies

- Area effect ν_{ra} is conditioned on survey data
  - Empirical Bayesian framework, survey prior updated by prediction
- Use model to simulate welfare repeatedly and calculate mean or poverty rate
  - Use parametric bootstrap approach to estimate precision
- Implemented in EMDI package in R, new povmap package coming soon

WORLD BANK GROUP

Candidate method 2: Cubist regression

Generates "model trees" (Kuhn and Johnson, 2013, Wang and Witten, 1996, Quinlan, 2014) to predict sub-area poverty rates. Grows a decision tree for which each terminal node contain linear regression model.

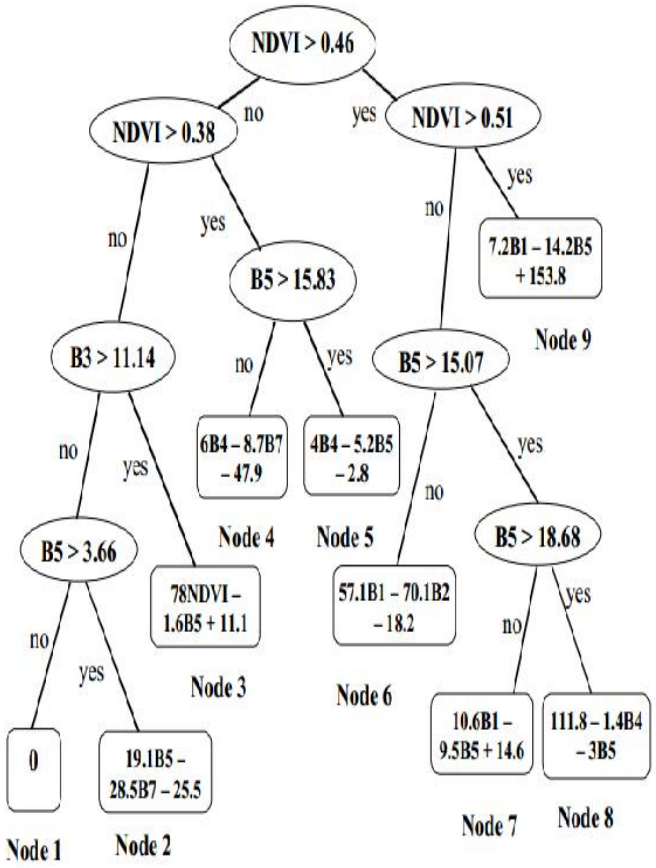
Can estimate "committees" which incorporates boosting that repeatedly predict residuals

Hyperparameters, including number of rules and committees, tuned through cross-validation or selected manually

Model is easy to understand when selecting no committees and small number of rules

Predict sub-area poverty rates using Cubist package in R and aggregate to areas

Sample cubist model for predicting tree canopy



WORLD BANK GROUP

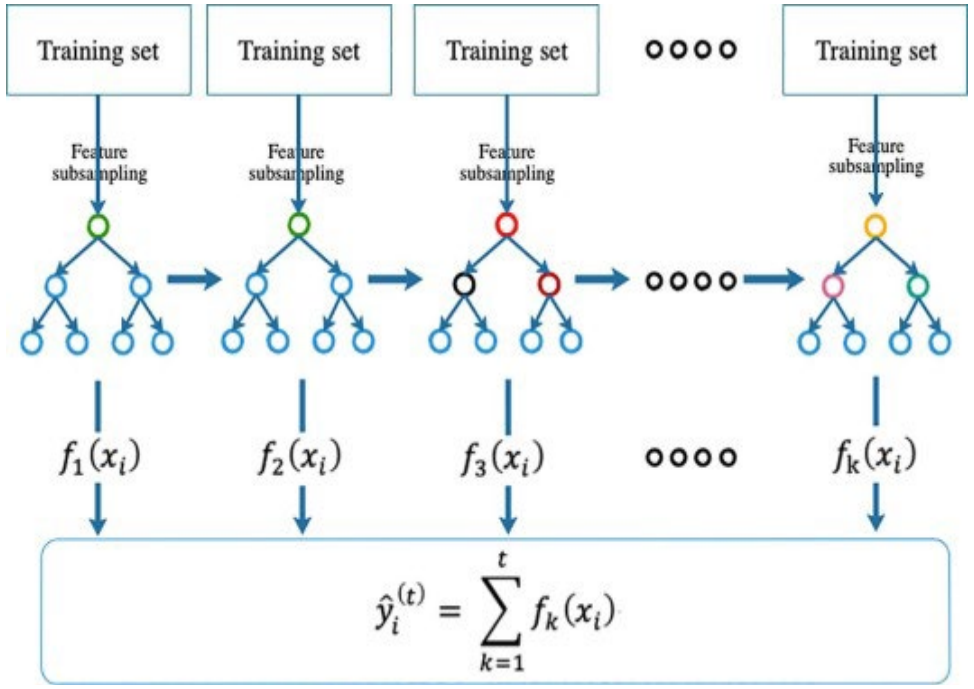
Candidate method 3: Extreme Gradient Boosting (XGboost)

Very popular method in machine learning (Chen and Guestrin 2016)

Develops a set of regression forests $f_k(x_i)$ that iteratively predict residuals from previous estimate

Hyperparameters tuned through cross-validation

Predict sub-area poverty rates using XGboost package in R and aggregate to areas



WORLD BANK GROUP

Candidate method 4: Boosted Regression Forests (BRF)

- Derived from Generalized Random Forests (Athey, Tibsherani, and Wager 2019)
- Like XGboost, grows sequence of regression forests that iteratively predict residuals
- Grows "Honest" trees
  - Use different subsamples of the data to grow trees and to estimate the values of leaves
  - Gives desirable theoretical properties of consistency and asymptotic normality
  - At small cost in predictive performance in most settings
- Also uses slightly different splitting rule and different hyperparameters than XGboost
  - Tuned automatically through survey cross-validation
- Predict sub-area poverty rates and aggregate to areas using GRF package in R

Estimating uncertainty

For linear EBP models, use standard parametric bootstrap approach

- Butar and Lahiri (2003), Gonzalez-Manteiga et al (2008)

For tree-based machine learning methods, use random effect block bootstrap

- Chambers and Chandra (2013), Krennmair and Schmid (2022)
- Accounts for correlation across sub-areas within areas when estimating uncertainty
- Assumes independent errors across areas

1. Calculate sub-area residuals and area (a) residuals from predictions

$$\hat{e}_{sa} = \hat{y}_{sa}^{ML} - \hat{y}_{sa}^{Direct} \text{ and } \hat{e}_{a} = \hat{y}_{a}^{ML} - \hat{y}_{a}^{Direct}$$

2. Sample $\hat{e}_{sa}$ and $\hat{e}_{a}$ with replacement to obtain $\tilde{e}_{a}^{k}$ and $\tilde{e}_{sa}^{k}$ for replication k

3. Use bootstrapped residuals to simulate sub-area wealth index or poverty

$$\tilde{y}_{sa}^{k} = \hat{y}_{sa}^{ML} + \tilde{e}_{a}^{k} + \tilde{e}_{sa}^{k}$$

4. Aggregate to target area to obtain $\tilde{y}_{a}^{k}$

5. Repeat steps 2-4 100 times, take 5th and 95th percentile of distribution to obtain estimated confidence interval

WORLD BANK GROUP

Uses geolocated census data from four countries

Can be linked to geospatial data using sub-area identifiers and corresponding shapefile

	Madagascar	Malawi	Mozambique	Sri Lanka
Target area	Commune	TA	Locality	DS Division
Number	1515	420	1258	331
Sub-area	Fokontany	EA	Bairro	GN Division
Count	14,412	18,700	65,707	13,984
Share of population	100%	20% Extract	100%	100%
Households	5 mn	0.8 mn	6 mn	4.8 mn
Year	2017	2018	2017	2012

Geospatial features

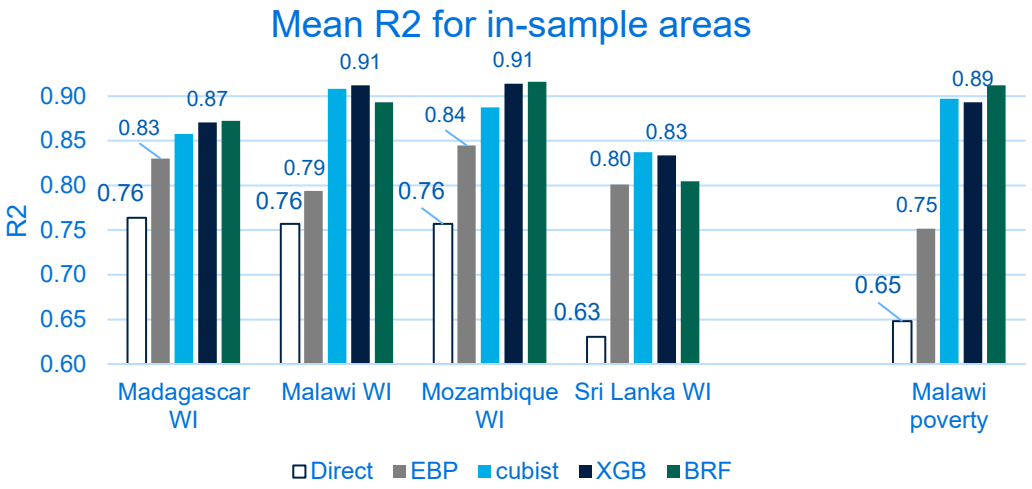
Basic list, mostly publicly available from Google Earth Engine

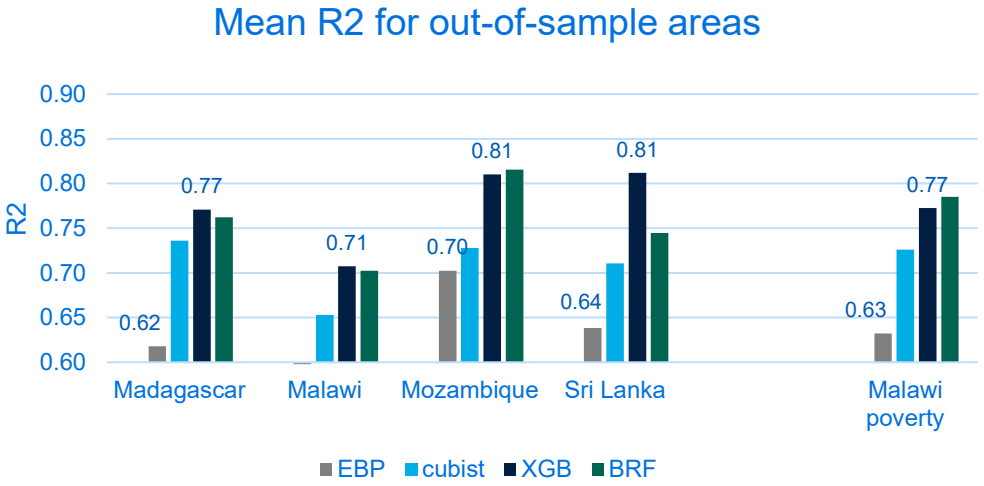
Indicator	Source
Population	Worldpop
Precipitation	TerraClimate
Temperature	TerraClimate
Nightlights	VIIRS
Land Cover	Copernicus
Elevation	Conservation Science Partners
NDVI	MODIS
Pollution measures	Sentinel 5P/Copernicus
Distance to key cities	Constructed by authors

Evaluation strategy

1. Estimate wealth index in census data for four countries using PCA
   - And poverty rates in Malawi based on predicted per capita consumption
2. Calculate area-level benchmark reference estimates using full census
3. Draw two-stage sample from the census
   - First stage draws 500 sub-areas using probability proportional to population size
   - 8 households per sub-area in second stage, N=4000 households
   - This is a realistic sample but further robustness checks would be useful
   - Generate small area estimates for target areas by combining sample with geospatial indicators
4. Repeat step 3 100 times
5. Compare with full census
   - Accuracy: R^2 vs census benchmark
   - Coverage rate: Share of areas for which confidence interval contains true value

Results on accuracy

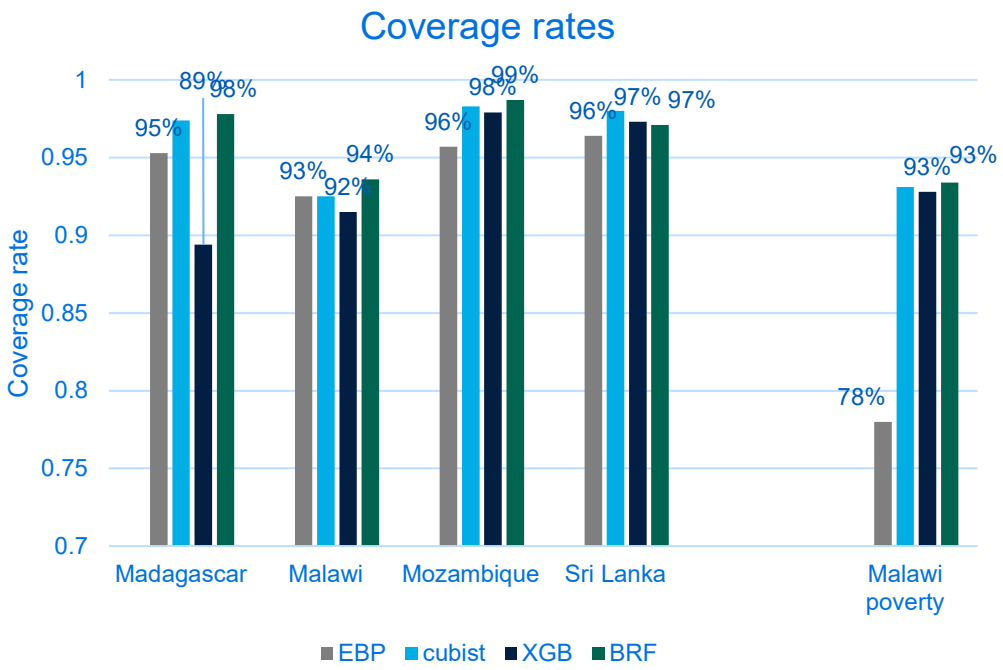




- Estimates highly accurate for in-sample areas
- XGboost usually most accurate
  - BRF usually close
- In-sample much better than out of-sample
- XGboost and BRF much more accurate than EBP and moderately more accurate than Cubist in out-of-sample areas
- Outside Sri Lanka, little difference between XGboost and BRF

WORLD BANK GROUP

Results on coverage



- Residual bootstrap gives good uncertainty estimates, overall coverage rate > 90%
- BRF coverage rates higher than XGboost in Madagascar and Malawi
- EBP significantly underestimates coverage for poverty
  - Because model ignores uncertainty in estimated variance components $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$ in parametric bootstrap procedure

WORLD BANK GROUP

Conclusions

1. Further evidence that geospatial small area estimation works well, especially for sampled areas
   - Estimates are less accurate for non-sampled areas, but still very good
   - More due to differences between sampled and non-sampled areas (like less population) than being out of the sample per se (analysis in paper)
2. XGboost and BRF outperform linear mixed models in terms of accuracy
   - Especially out of sample
   - But in-sample too, despite presence of conditional random effect in linear mixed model
   - May not hold in a smaller sample of sub-areas
3. Cubist not quite as good as XGboost and BRF
   - Limiting to three rules and no committees worsens performance but still beats EBP
4. Random effect block bootstrap works well to estimate uncertainty
   - while accounting for within-area correlation

WORLD BANK GROUP

Implications and future work

1. Data fusion can improve accuracy a lot in realistic settings

2. XGboost and BRF are attractive methods for small area estimation in these settings
   - Tree-based ML methods are more robust to outliers and more accurate than linear models
   - Users need to balance additional accuracy against loss in parsimony and interpretability
   - XGboost and BRF could further benefit from adding conditional random effect, building on mixed effect random forests model (Krennmair and Schmid, 2022)
     - But conditional random effects appear to be a minor factor when applying tree-based machine learning techniques to a sufficiently large sample
     - Even simple cubist regression with 3 rules and no committees usually outperforms EBP

3. Residual block bootstrap works well
   - Simple and effective way to incorporate spatial correlation into ML estimation

4. Important agenda for further work
   - Compare with CNN-based estimates
   - Better understand how accuracy depends on sample size and structure

WORLD BANK GROUP

Thank you!

References

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized Random Forests. *The Annals of Statistics*, *47*(2), 1148-1178.

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.

Burke, M., Driscoll, A., Lobell, D. B., & Ermon, S. (2021). Using satellite imagery to understand and promote sustainable development. Science, 371(6535), eabe8628.

Butar, F. B., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*(1-2), 63-76.

Chambers, R., & Chandra, H. (2013). A random effect block bootstrap for clustered data. Journal of Computational and Graphical Statistics, 22(2), 452-470.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

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.

Efron, B. (2020). Prediction, estimation, and attribution. *International Statistical Review*, *88*, S28-S59.

Engstrom, R., Hersh, J., & Newhouse, D. (2022). Poverty from space: Using high resolution satellite imagery for estimating economic well-being. *The World Bank Economic Review*, *36*(2), 382-412.

Edochie. I., D. Newhouse, T. Schmid, N. Tzavidis, E. Foster, A. Ouedraogo, A. Sanoh, A. Savadogo, forthcoming, “Small area Estimates of Poverty in Four West African countries”

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), 443-462.

WORLD BANK GROUP

References

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, *15*, 1-96.

Khachiyan, A., Thomas, A., Zhou, H., Hanson, G., Cloninger, A., Rosing, T., & Khandelwal, A. K. (2022). Using Neural Networks to Predict Microspatial Economic Growth. *American Economic Review: Insights*, *4*(4), 491-506.

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

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).

McBride, L., Barrett, C.B., Browne, C., Hu, L., Liu, Y., Matteson, D.S., Sun, Y. and Wen, J., 2022. Predicting poverty and malnutrition for targeting, mapping, monitoring, and early warning. Applied Economic Perspectives and Policy, 44(2), pp.879-892.

Molina, I., & Rao, J. N. K. (2010). Small area estimation of poverty indicators. Canadian Journal of statistics, 38(3), 369-385.

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, World Bank Policy Research Paper 10175

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Wang, Y., & Witten, I. H. (1996). Induction of model trees for predicting continuous classes.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., ... & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature communications*, *11*(1), 2583.

WORLD BANK GROUP