SMALL AREA ESTIMATION WITH GEOSPATIAL DATA

WORLD BANK GROUP

Takaaki Masaki, David Newhouse, Ani Rudra Silwal,
Adane Bedada, Ryan Engstrom

April 21, 2020

Structure of presentation

1. Motivation
2. Methods
3. Validation procedure
4. Results
5. Conclusion

WORLD BANK GROUP

1. Motivation

- Rapidly growing body of evidence that satellite indicators are strongly correlated with subnational poverty, welfare, and population density
  - i.e. Jean et al (2016), Steele et al (2017), Engstrom et al (2017), Head et al (2017), Pokhriyal and Jacques (2017), Engstrom et al (2019), others
  - Indicators and imagery increasingly available and accessible

- Satellite imagery attractive because coverage is universal
  - But indicators are difficult to interpret
    - I.e. Night-time lights reflects a mix of urbanization, highways, and economic growth
  - and satellite indicators rarely available at household level

1. Motivation

- Combining survey and satellite data can give more recent, precise and granular measures of household socioeconomic characteristics
  - How much does satellite data boost precision of survey estimates?
    - *Precision ≠ out-of-sample predictive accuracy*
  - Which method is best for combining survey with satellite data for small area estimation?

- Lack of validation exercises testing different methods for combining household data with satellite auxiliary data for small area estimation

2. Two types of Empirical Best models

- Use Empirical Best models
  - Workhorse for small area estimation
  - Seems to give similar results to Hierarchical Bayesian models (Guadarrama et al, 2016)

- Area level models
  - Fay and Herriot (1979) and many subsequent variants
  - Poverty rate is modeled as a linear function of predictors
  - F-H estimates are weighted average of direct sample estimates and predictions from area-level linear model.
    - Gives more weight to predictions if they are precise relative to sample
  - Estimated using FHSAE command in Stata

WORLD BANK GROUP

2. Two types of Empirical Best models

- Household-level nested error EB models
  - Household welfare modeled as a function of village characteristics with area random effect
    - Area effect conditioned on area mean of sample residuals
    - Use parametric bootstrap approach to estimate precision
  - Battese, Harter, and Fuller (1988), Elbers Lanjouw and Lanjouw (2003), Molina and Rao (2010)
  - Variants implemented in both Stata and R
    - Stata SAE command (Zhao, 2006, Van der Weide, 2014, Nguyen et al 2017). Updated in September 2019 (v2) and February 2020 (v3)
    - R EMDI package (Kreutzmann et al 2016) and R SAE package (Molina and Marhuenda 2015)
  - We use a modified version of R EMDI package that allows for weights in model estimation and aggregation of simulation results

WORLD BANK GROUP

3. Validation procedure

- Use the 2012 census for Sri Lanka and Tanzania
  - Sri Lanka: ≈5 million households, 331 Divisional Secretariat Divisions (DSDs), about ≈14,000 Grama Niladhari Divisions (GNDs)
  - Tanzania: ≈9 million households, 159 districts, about ≈15,000 villages
- Surveys representative for 25 TZA regions and 25 LKA districts
- Aim to generate similarly precise estimates for areas of interest
  - 159 districts for Tanzania; 331 Subdistricts (DSDs) for Sri Lanka
- Construct "non-monetary welfare index" using first principal component
  - Include household size, education, age, dependency ratio, sex, sector of work (e.g., agriculture, livestock, fishing), disability, cash transfer beneficiary
  - Define "non-monetary poverty" as falling below 4th percentile in Sri Lanka and 20th percentile in Tanzania

Validate using synthetic sample from census

- Designed to mimic actual household survey
  - Select same set of EAs as the 2016 HBS in Sri Lanka and 2018 HBS in Tanzania
  - Randomly select same number of households as in HBS in each EA.
  - Use household weights for each EA from sample survey.

- Synthetic sample, like actual budget survey, contain observations from all small areas (DSDs for Sri Lanka; districts for Tanzania)
  - This is critical for generating precise predictions in each area

WORLD BANK GROUP

Criteria for evaluating methods

- Precision
  - Average standard error and coefficient of variation
  - Should see significant reduction compared with survey estimate
- Correlation with “truth” from census
- Coverage rate
  - Share of DSDs or districts in which the estimated 95% confidence interval for welfare index deprivation contains actual census value
  - Indicates whether estimated confidence intervals are accurate
  - Target is 95% coverage rate
- Mean poverty rate
  - Prefer methods that can accurately predict national poverty rate before benchmarking

Transform dependent variable

- Welfare typically follows a right-skewed distribution
- Transformation helps make the distribution more normal
  - EBP models require normality assumptions for error terms
- Standard procedures to transform welfare for SAE models:
  - Natural log
  - Log-shift
  - Box-Cox
- These do not necessarily ensure normality at tails
- We therefore implement ordered quantile normalization (Cavanaugh and Peterson, 2019)
  - Normalizes based on household ranking of non-monetary welfare
  - Guarantees normal welfare distribution in absence of ties
  - Implemented in bestnorm R package

Spatial data for welfare prediction

- Variety of geospatial data considered to estimate welfare
- Calculate zonal statistics at village and area levels
- Construct census of households linked to zonal statistics
  - No household-level characteristics in model or auxiliary data
- Use post-lasso to select model to predict household welfare as a function of village and area zonal statistics

WORLD BANK GROUP

Using spatial data to predict welfare

Candidate spatial Variables Used to Predict Welfare

Sri Lanka	Tanzania
• **Urbanization**: Nighttime light, global human settlement layer (GHSL) • **Agro-climactic**: Rainfall, elevation, slope forest cover change • **Spatial features**: Fourier Transform (FT), Gabor, Histogram of Oriented Gradients (HOG), Lacunarity (LAC), Line Support Regions (LSR), Normalized Difference Vegetation Index (NDVI), PanTex, Structural Feature Sets (SFS)	• **Urbanization**: Nighttime light, global human settlement layer (GHSL), global urban footprint, population, Agglomeration Index, building footprints • **Agro-climactic**: Rainfall and temperature, elevation, Köppen-Geiger climate classification, crop yield, net primary productivity (NPP) and Normalized Difference Vegetation Index (NDVI) • **Market access index:** Weighted sum of city population reachable within one hour of driving • **Natural disaster risks**: Estimated size of GDP and population exposed to flood and drought events

Model selection

Sri Lanka Model	Summary Statistics Tanzania
Number of spatial variables: 15 Conditional R^2: 0.296	Number of spatial variables: 16 Conditional R^2: 0.323

R^2 represents share of variation in household welfare index explained by village and district remote sensing indicators, with the random effect conditioned on the area means of sample residuals. Number of variables excludes district-fixed effects or provincial fixed-effects and rural dummy variable.

WORLD BANK GROUP

4. Results: Bias and precision

Mean and Precision	Sri Lankan subdistricts			Tanzanian districts		
	Mean poverty	Mean SE	Mean CV	Mean poverty	Mean SE	Mean CV
Direct survey estimates						
Horvitz-Thompson approximation	4.0	2.9	66.6	19.7	7.3	43.4
EA-Clustered standard errors	4.0	2.8	59.4	19.7	6.6	38.1
Area level Fay-Herriot model	3.8	2.5	59.4	16.8	5.6	34.2
Household-level EBP model	3.9	1.5	30.4	20.6	3.2	20.1

Notes: Mean poverty refers to the unweighted mean of the predictions across areas. Mean SE and CV refers to the unweighted mean of the Standard Error and Coefficient of Variation across areas.

WORLD BANK GROUP

Results: Accuracy and coverage

Mean and Precision	Sri Lankan subdistricts			Tanzanian districts		
	Rho	Mean Abs Error	Cov Rate	Rho	Mean Abs. Error	Cov Rate
Direct survey estimates						
Horvitz-Thompson approximation	74.4	2.8	81.6	77.0	6.6	85.5
EA-Clustered standard errors	74.4	2.8	77.0	77.0	6.6	76.1
Area level Fay-Herriot model	83.9	1.9	89.4	88.3	4.1	91.8
Household-level EBP model	88.8	1.6	83.7	87.6	4.1	73.6

Notes: rho refers to unweighted correlation across areas between estimated headcount poverty and true headcount poverty calculated from the census. Mean absolute error is the average magnitude of the discrepancy between the estimates and the census. Coverage rate is the percent of areas for which the estimated confidence interval contains the census value.

WORLD BANK GROUP

Summary of Key Results

- Household nested error model achieves large efficiency gain compared to direct estimates
  - Sri Lanka: CV falls from 67 to 30.
  - Tanzania: CV falls from 43 to 20.
  - Roughly equivalent to quadrupling the sample size in both cases
- Robust to different methods of variable selection (e.g., stepwise)
  - Though overfitting reduces accuracy of estimates
- Coverage rates underestimated but reasonable
  - 84 percent in Sri Lanka and 74 percent in Tanzania
  - Comparable to coverage rates for standard estimates using sample and clustering on enumeration areas

WORLD BANK GROUP

Household level EBP model more precise than F-H after rescaling standard errors to achieve common 95% coverage

Precision after rescaling SEs to achieve 95% coverage	Sri Lankan subdistricts			Tanzanian districts		
	Scale factor	Mean SE	Mean CV	Scale factor	Mean SE	Mean CV
Horvitz-Thompson direct estimate with rescaled SE	N/A	N/A	N/A	2.5	18.9	108
Fay-Ferriot model with rescaled SE	1.69	4.3	100.4	1.2	6.8	41.1
Household-level EBP model with rescaled SE	1.54	2.4	47.6	1.85	5.7	32.3

5. Conclusions

- Household-level EBP model performs well
  - Significantly more accurate in Sri Lanka, slightly less accurate than F-H in Tanzania
  - More efficient than F-H after rescaling SEs to achieve 95% coverage
  - Efficiency gain relative to direct estimates comparable to quadrupling sample size
    - Satellite-augmented estimates for Tanzanian districts roughly as precise as survey estimates for region

- Standard errors are underestimated
  - Due to assumptions of independent disturbances within area and non-stochastic estimated variance components
  - Consistent with existing literature (i.e. Hall and Maiti 2006)
  - But coverage rates are reasonable and comparable to direct survey estimates clustered on PSU

WORLD BANK GROUP

Conclusions

- Normalizing transformation works well
- Benchmarking slightly overestimates standard errors, but this helps mitigates downward bias from EBP model
- Future research agenda
  - Finish paper (incorporate design-based simulation?)
  - Push forward on developing software that estimates sub-area models ala Torabi and Rao (2015)
  - Test monetary poverty predictions in Mexico
  - Incorporate bestnorm transformation into R EMDI package
  - Incorporate adjustment for intra-hh correlation when conditioning on sample mean residuals
  - Estimate standard errors properly when benchmarking
- Operationalization agenda
  - Tanzania (complete), Ethiopia, Somalia, Borno State, others?
  - Post-crisis phone monitoring surveys?

References

- Battese, George E., Rachel M. Harter, and Wayne A. Fuller. "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association* 83.401 (1988): 28-36.
- Das, Sumonkanti, and Ray Chambers. "Robust mean-squared error estimation for poverty estimates based on the method of Elbers, Lanjouw and Lanjouw." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.4 (2017): 1137-1161.
- Demombynes, G., Elbers, C., Lanjouw, J. O., & Lanjouw, P. (2007). *How good a map? Putting small area estimation to the test*. The World Bank.
- Diallo, Mamadou S., and J. N. K. Rao. "Small area estimation of complex parameters under unit-level models with skew-normal errors." *Scandinavian Journal of Statistics* 45.4 (2018): 1092-1116.
- Elbers, Chris, Jean O. Lanjouw, and Peter Lanjouw. "Micro–level estimation of poverty and inequality." *Econometrica* 71.1 (2003): 355-364.
- Elbers, Chris, Peter Lanjouw, and Phillippe George Leite. *Brazil within Brazil: Testing the poverty map methodology in Minas Gerais*. The World Bank, 2008.
- Engstrom, Ryan, Jonathan Hersh, and David Newhouse. "Poverty from space: Using high-resolution satellite imagery for estimating economic well-being." (2017).
- Engstrom, Ryan, David Newhouse, and Vidhya Soundararajan. *Estimating Small Area Population Density Using Survey Data and Satellite Imagery: An Application to Sri Lanka*. The World Bank, 2019.

References

- Fay III, Robert E., and Roger A. Herriot. "Estimates of income for small places: an application of James-Stein procedures to census data." *Journal of the American Statistical Association* 74.366a (1979): 269-277.
- González-Manteiga, Wenceslao, et al. "Bootstrap mean squared error of a small-area EBLUP." *Journal of Statistical Computation and Simulation* 78.5 (2008): 443-462.
- Guadarrama, María, Isabel Molina, and J. N. K. Rao. "A comparison of small area estimation methods for poverty mapping." *Statistics in Transition new series* 1.17 (2016): 41-66.
- Hall, Peter, and Tapabrata Maiti. "On parametric bootstrap methods for small area prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2 (2006): 221-238.
- Head, Andrew, et al. "Can human development be measured with satellite imagery?." *ICTD*. 2017.
- Jean, Neal, et al. "Combining satellite imagery and machine learning to predict poverty." *Science* 353.6301 (2016): 790-794.
- Kreutzmann, Ann-Kristin, et al. "The R package emdi for the estimation and mapping of regional disaggregated indicators." *Journal of Statistical Software* (2018).
- Hall, Peter, and Tapabrata Maiti. "On parametric bootstrap methods for small area prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68.2 (2006): 221-238.
- Jiang, Jiming, and Partha Lahiri. "Mixed model prediction and small area estimation." *Test* 15.1 (2006):
- Lange, Simon, Utz Johann Pape, and Peter Pütz. *Small area estimation of poverty under structural change*. The World Bank, 2018.

References

- Molina, Isabel, and J. N. K. Rao. "Small area estimation of poverty indicators." *Canadian Journal of Statistics* 38.3 (2010): 369-385.
- Molina, Isabel, and Yolanda Marhuenda. "sae: An R package for small area estimation." *The R Journal* 7.1 (2015): 81-98.
- Nguyen, Minh C., et al. *Small Area Estimation: An extended ELL approach*. mimeo, 2017
.
- Peterson, Ryan A., and Joseph E. Cavanaugh. "Ordered quantile normalization: a semiparametric transformation built for the cross-validation era." *Journal of Applied Statistics* (2019): 1-16.
- Pokhriyal, Neeti, and Damien Christophe Jacques. "Combining disparate data sources for improved poverty prediction and mapping." *Proceedings of the National Academy of Sciences* 114.46 (2017): E9783-E9792.
- Torabi, Mahmoud, and J. N. K. Rao. "On small area estimation under a sub-area level model." *Journal of Multivariate Analysis* 127 (2014): 36-55.
- Tzavidis, Nikos, et al. "From start to finish: a framework for the production of small area official statistics." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.4 (2018): 927-979.
- Van der Weide, Roy. *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, 2014.
- Zhao, Qinghua. "User manual for povmap." *World Bank. http://siteresources. worldbank. org/INTPGI/Resources/342674-1092157888460/Zhao_ ManualPovMap. pdf* (2006).

ANNEX

WB tools differ from EBP tools in important ways

	Stata ELL		Modified R EMDI
	Version 2	Version 3	
Heteroscedasticity correction	Yes	Yes	No
Type of bootstrap	Traditional Non-Parametric clustered bootstrap with varying sample composition	Parametric bootstrap with constant sample composition	Parametric bootstrap with constant sample composition
Model fitting method	Henderson method 3 with non-parametric bootstrap	Henderson method 3 with non-parametric bootstrap	Maximum likelihood estimation
Empirical Best estimation	Optional	Required	Required

Model selection

- Baseline method is post lasso
  - Avoids setting arbitrary threshold for stepwise regression
  - Use "plugin" lasso estimator
    - Described in detail in Statacorp (2019) and Belloni and Chernuzhukov (2011)
    - Allows for heteroskedastic residuals
    - Tends to give slightly more parsimonious models than standard out of sample cross-validation lasso

- Variable pool consists of:
  - Sri Lanka: All remote sensing indicators at GND and DSD levels plus district dummies
  - Tanzania: All remote sensing indicators at village and district levels plus provincial dummies

WORLD BANK GROUP

Regression of log household welfare on lasso-selected variables

Sri Lanka		Tanzania	
Variable	Coefficient	**Variable**	Coefficient
Sub-area variables		***Sub-area variables***	
1990 built-up area	-0.30	Sum of night time lights	0.03*
2014 built-up area	1.82***	Mean population (GHSL)	0.01
Standard deviation of rain	0.00	Minimum agglomeration index	-0.01*
Rain Z score, Q2	0.05*	Mean std. dev. of size of 5 nearest buildings	-0.01
Rain Z-score squared, Q4	-0.45	Sum of mean size of 5 nearest buildings	0.30***
Fourier Transform mean, scale 7	0.08***	Number of buildings within 100 m	0.05**
Line Support Region mean, Scale 7	0.47**	% area never built-up, 1975-2015 (GHSL)	-0.25
		% of areas built up from 1990 to 2000 (GHSL)	0.13
Area variables		% of areas built up from 1975 to 1990 (GHSL)	1.47***
1990 built-up area	2.47***	% of area built-up (GUF)	0.04
Standard deviation of rain	1.06***	Precipitation in 2014	0.00
Rain Z score squared, Q3	-0.49	% Humid tropical rainforest (Kloppen clasification)	5.49***
Rain Z-score squared, Q4	-0.46	Standard deviation of NDVI	56.67***
Gabor standard deviation, scale 5	-0.83***		
Histogram of Ordered Gradients standard deviation, Scale 5	-1.74***	***Area variables***	
Line Support Region mean, Scale 7	-0.69**	Mean of night time lights	0.15
Structural feature sets mean, scale 7	1.00***	Minimum of night time lights	0.48
		Maximum economist costs of drought	0.00

Notes: Super-area and rural/urban dummies are not reported for brevity.

WORLD BANK GROUP

Tanzania factor loadings

Variable	Scoring Coefficients
Literate	0.45
Ever attended school	0.45
Head age	-0.2
Household size	-0.22
Dependency ratio	-0.41
Male head	0.12
Non-agricultural work	0.29
Non-livestock work	0.25
Non-fishing work	0.01
No disability	0.07
Cash transfer beneficiary	0.13

WORLD BANK GROUP

Sri Lanka factor loadings

Variable	Scoring Coefficients	Variable	Scoring Coefficients
Household size	0.09	Head education category	
Dependency ratio	0	No schooling	-0.11
Children 0-14 and 65+	0.04	Up to grade 5	-0.17
Children 0-14 Only	-0.05	Grade 5-10	-0.13
Gender ratio	0.04	O or A level	0.29
Household education		College Degree or higher	0.13
No schooling	-0.07		
Up to grade 5	-0.14	Age of head in years	0.04
Grade 5-10	-0.31		
O or A level	0.27	Head employment status	
College Degree or higher	0.16	Unemployed	-0.0005
Household Assets		Public sector	0.15
House	0.04	Private sector	-0.6
Computer	0.27	Out of labor force	-0.04
Landphone	0.23	Male head	0.11
TV	0.28	Head marital status	0
Housing characteristics		Unmarried	-0.04
Roof	0.14	Married	0.12
Private Toilet	0.18	Widowed	-0.1
Wall	0.19	Divorced	-0.06
Waste disposal	0.15		
Safe water	0.13		
Main cooking fuel is wood	-0.23		
Electricity for light	0.27		

5. Discussion: Why does EBP produce downwardly biased SEs?

- Failure to account for design effects when incorporating sample into model estimates. The model is:

(1) $G(Y_i) = \beta_1 X_{sa} + \beta_2 X_a + \beta_3 X_r + \eta_a \mid \overline{e_a} + \varepsilon_i$

(2) $\eta_a \sim N\left(\overline{e_a}, \sigma^2\eta(1-\gamma)\right)$

(3) $\gamma = \frac{\sigma^2\eta}{\sigma^2\eta + \frac{\sigma^2\varepsilon}{N}}$

- Conditioning on $\overline{e_a}$ without adjusting for interhousehold correlation underestimates standard error of sample estimate

WORLD BANK GROUP