SMALL AREA ESTIMATES OF MONETARY POVERTY USING SATELLITE DATA: EVIDENCE FROM MEXICO

WORLD BANK GROUP

David Newhouse
Anusha Ramakrishnan
Tom Swartz
Joshua Merfeld
Partha Lahiri

With key contributions from Boris Babenko and Jonathon Hersh.

April, 2022

Integrating survey data with big data

- Recent advances in the availability of “big data” from satellites and cell phones (World Bank, 2021; Burke, 2021).
  - Many predictive geospatial indicators derived from satellite imagery and crowd-sourcing applications are freely available
  - Can help fill spatial gaps in surveys and reduce sampling error

- Growing body of innovative research using geospatial or other big data to predict poverty
  - Jean et al, 2017; Yeh et al,2020; Masaki et al, 2020; Browne et al 2021, Chi et al, 2021; Engstrom et al, 2021, Aiken et al, 2021
  - Estimates typically generated and evaluated at cluster level, using cross-validation
  - Results demonstrate that geospatial data predicts poverty reasonably well

- Also applied to population, health, and agriculture
  - Gething et al., 2016; Golding et al., 2017, Erciulescu et al, 2019, Wardrop et al., 2018; etc.

Integrating survey data with big data

- Most of this literature linking poverty and big data does not use “small area estimation” techniques
  - Steele et al (2017), Pokhriyal and Jacques (2017), Masaki et al (2020), Steele et al (2021) are notable exceptions

- Literature currently uses a wide variety of methodologies
  - Different prediction methods: Pure prediction (machine learning), Bayesian, and Empirical Best Predictor methods
  - Models at different levels: Household level, village level, target area level
  - Different indicators: Wealth index vs poverty rates
  - Different auxiliary data: Well-defined geospatial features, features extracted from machine learning predictions, CDR data, etc.

WORLD BANK GROUP

Key points

1. Geospatial small area estimation improves significantly on direct survey estimates and should be applied more frequently
2. Household level models are slightly more accurate than village level models and much more precise than area level models
3. It is important to use either Empirical Best Predictor or Bayesian methods
    - In-sample predictions are much more accurate and precise than out-of-sample predictions
    - Samples should seek to cover as many target areas as possible

WORLD BANK GROUP

This is actually an old idea

- Battese, Harter, and Fuller (1988): Used Empirical Best Predictor model to combine satellite data and survey data to estimate areas under soybean and corn cultivation in 12 Iowa counties

- Seminal paper in the sae statistics literature, 989 cites in Google Scholar
    - First to apply empirical best predictors to unit-level models
    - Subsequently extended by Molina and Rao (2010) to handle non-linear indicators such as poverty rates

- But to our knowledge this method was never used or applied with other geospatial data until recently (Masaki et al, 2020)

Empirical Best Predictor method has many advantages

1. Effectively integrates survey data and auxiliary data
    - Treats survey as prior, updated by predictions using auxiliary data
    - Simpler than pure Bayesian methods, does not require specifying a prior distribution
    - Assumes normality, but not an issue with proper transformation of data

2. Theory is well-known and accepted in statistical community
    - i.e., used by Mexican NSO with census data for official poverty estimates

3. Relies on linear regression framework that is more transparent than other machine learning methods
    - LASSO is a simple form of machine learning that integrates well with this framework at little cost
    - Other machine learning methods like random forest may predict better but theory is still new (Krenmair and Schmid, 2021)

4. Can be implemented using "off-the-shelf software" relatively easily

5. Moderately underestimates uncertainty but additional refinements could fix that

Testing geospatial data in Mexico: Four main research questions

1. How much more accurate and precise are small area estimates of municipal poverty rates in Mexico, obtained by combining survey data with geospatial indicators, than direct estimates from survey data?

2. How does the accuracy and precision of municipal poverty estimates differ for sampled and non-sampled municipalities?

3. Are small area estimates using survey and geospatial data more or less accurate than older small area estimates generated using a household census?

WORLD BANK GROUP

Main research questions

4. How do estimates vary across three different types of small area estimation models?
    - Household model: predict transformed household per capita income using AGEB and municipal variables
    - Sub-area model: Predict AGEB poverty rates using AGEB and municipal variables
    - Area-level model: Predict municipal poverty rates using municipal variables

- Mexican AGEBs are like a US block group
    - Urban AGEBs contain ~1500 people
    - About 60,000 AGEBs and 2,500 municipalities in Mexico

WORLD BANK GROUP

Survey and evaluation data

1. MCS-ENIGH 2014 survey
   - Contains 58,125 households, 75% urban
   - Covers 892 (out of 2,433) municipalities = target areas
   - Contains AGEB-level identifiers
   - Source of official poverty estimates for urban/rural areas of each state

2. Evaluate against official 2015 municipal poverty estimates
   - Derived by Mexican government using MCS-ENIGH 2014 survey and 2015 intercensus
   - 2015 intercensus contains 5.8 million households
   - Used Empirical Best Predictor Model using 2014 survey data
     - Model based on demographic, labor, housing quality variables at individual, household and municipal level
     - Divided 32 states into 6 groups, separate model for each group
     - High R^2s of models predicting log per capita income, between 0.52 and 0.57

3. Compare with official 2010 municipal poverty estimates
   - Estimates based on 2010 survey and census data containing household, demographic, labor, housing quality at individual, household and municipal level
   - Useful to compare accuracy of geospatial estimates to older traditional poverty map

Auxiliary data

- Derived by Orbital Insight, inc. using proprietary algorithms applied to imagery from Planet, Inc. (3 to 5 m resolution)

1. Land classification
    - Proprietary convolutional neural network assigns probability to each pixel of 6 classes: Building, road, water, grassland, forest, and background (all others)

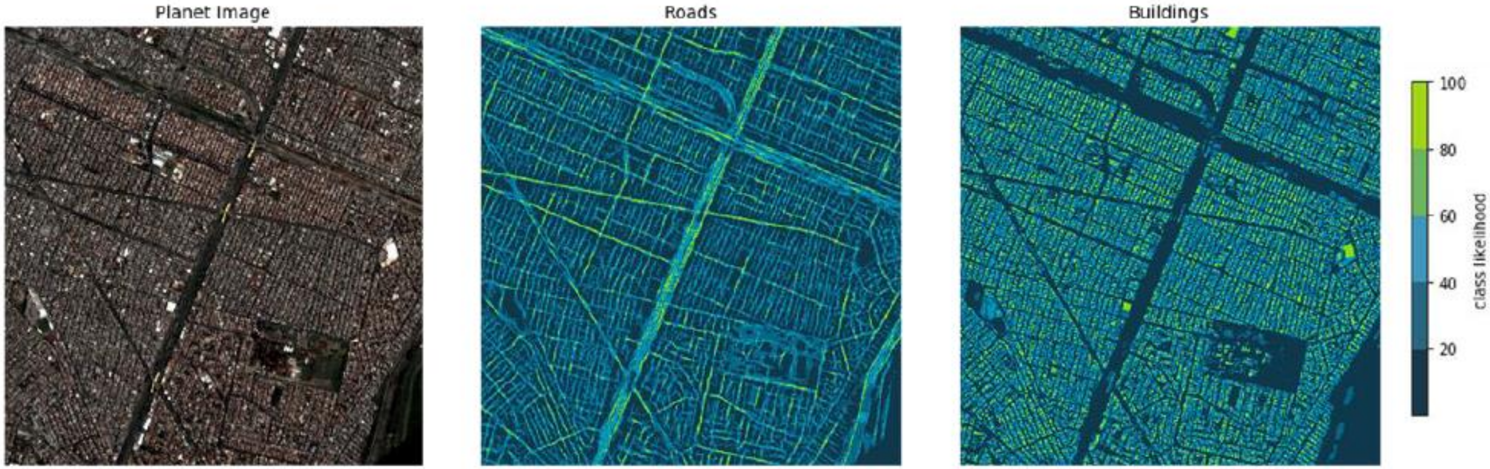

Figure A3: Example Pixel-level Land-use Results, Mexico City (Satellite image (c) 2017, Planet)

Auxiliary data

2. Train CNN to predict moderate and extreme poverty rate directly
    - Divided Mexico into rectangular tiles of about 750 sq meters each, roughly 2.6 mn tiles total
    - Assigned a tile equal to AGEB estimated poverty rate from household survey if tile intersected with sampled AGEB
    - Used Googlenet architecture with fine-tuning from imagenet (Babenko et al, 2017)
    - Aggregated predictions up to AGEB level, weighing by area of intersection between tile and AGEB.

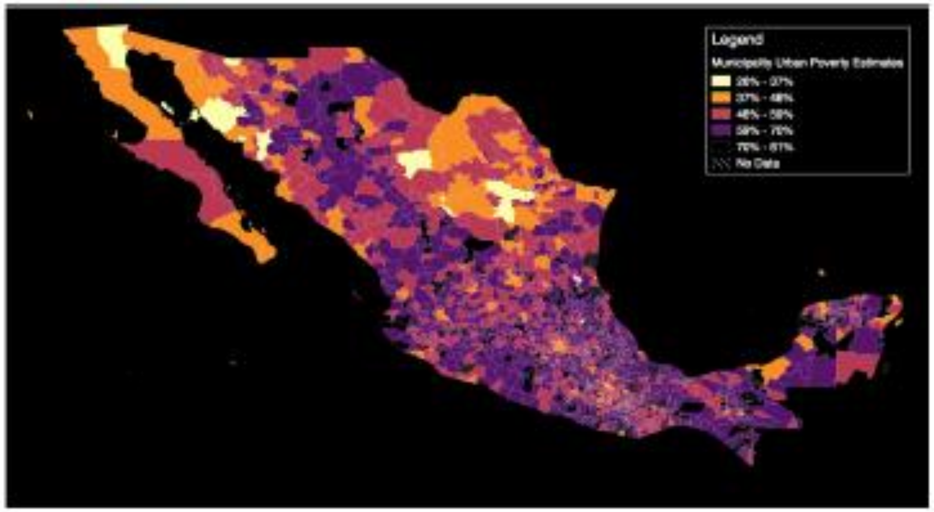

Figure 2: Poverty Estimates, Urban Municipalities

WORLD BANK GROUP

Household model

Signs of coefficients make sense

Auxiliary variables - AGEB average	Normalized per capita income
CNN Predicted percent extremely poor	-0.34
CNN Predicted percent not poor	0.79***
Percent building	0.66***
Percent forest	-0.25***
Auxiliary variables - Municipal average	
CNN Predicted extreme poverty rate	-0.97***
Percent building	0.03
Percent grass	-0.55***
Percent of population rural	-0.24***
Constant	-0.07
16 State Dummies	Yes
Number of observations	57,660
Adjusted R2	0.13

5 main findings

1. Combining satellite indicators with household survey data significantly improves accuracy and greatly improves precision compared to using survey data alone.
   - In the preferred specification, correlation with the benchmark official estimates rises from 0.8 to 0.86 when using small area estimates
   - Median coefficient of variation cut in half - 19.8 for small area estimates vs 38.5 for survey estimates

2. Household-level model moderately underestimates uncertainty
   - For household model, coverage rate is 77 percent for in-sample municipalities and 83 percent of out of sample municipalities
   - Moderately lower than the 86 percent for sampled municipalities when using appropriate (Horvitz-Thompson approximation) variance estimator.
   - Median CV rises to 25 if the mean squared error estimates are adjusted to maintain 86 percent coverage, still much less than 38.5 for direct estimates
   - After adjustment, improvement in precision roughly equivalent to increasing sample size by factor of 2.4, at very low cost

5 main findings

3. Predictions are more accurate and much more precise for sampled municipalities than non-sampled municipalities
   - Correlation with official estimates is 0.7 for non-sampled municipalities vs 0.86 for sampled municipalities
   - Median CV is 33.9 for non-sampled municipalities vs 19.8 for sampled municipalities.

4. Household model outperforms sub-area and area-level models in this context
   - Estimates from household model are more precise and accurate than sub-area and area model estimates in sampled municipalities
   - In non-sampled municipalities, household model estimates are at least as accurate as sub-area or area models

5. Geospatial small area estimates are significantly less accurate than 2010 estimates based on household unit-record census data
   - Geospatial poverty maps are a second-best solution when recent census data is not available
   - Need more research to better understand when to rely on old census poverty maps and when to update with geospatial estimates

Poverty predictions for municipalities: Mean poverty and precision

	Sampled municipalities			Non-sampled municipalities		
	Mean Poverty (pre-calibration)	Mean MSE	Median CV	Mean poverty (pre-calibration)	Mean MSE	Median CV
Direct survey estimates						
Horvitz-Thompson approximation	0.282	155.8	38.5	N/A	N/A	N/A
Small Area estimates						
Household model	0.281	35.8	19.8	0.355	150.3	33.9
Sub-area model	0.282	101.5	35.6	0.365	306.2	47.3
Area-level model	0.227	64.7	28.1	0.271	158.1	37.5
Official 2010 estimates	0.266	N/A	N/A	0.459	N/A	N/A
Official 2015 estimates	0.298	N/A	N/A	0.426	N/A	N/A

WORLD BANK GROUP

Poverty predictions for municipalities: Correlation and accuracy

	Sampled municipalities				Non-Sampled municipalities		
	Corr	RMSD	Coverage Rate		Corr	RMSD	Coverage Rate
Direct Survey Estimates (H-T)	**0.800**	0.126	0.856		N/A	N/A	N/A
Household model	**0.862**	0.094	0.769		0.701	0.181	0.825
Sub-Area model	0.834	0.103	0.910		0.696	0.183	0.941
Area-level model	0.796	0.110	0.824		0.662	0.198	0.801
Official 2010 estimates	**0.912**	0.083	N/A		0.904	0.109	N/A

Robustness check: Simulations with municipal covariates

- Do repeated simulations using intercensus data
  - Use municipal level predictors only because AGEB level identifiers not publicly available in intercensus
  - Correlation between estimates and benchmark higher than before
    - Because sample is drawn from population used to construct benchmark
  - Household model estimates equally accurate in-sample and more accurate out of sample.

Average over 100 Simulations	RMSD	Correlation
Sampled municipalities		
Direct survey estimates (H-T)	0.294	0.926
Household model	0.272	0.941
Area-level model	0.274	0.937
Intercensus benchmark		
Non-sampled municipalities		
Household model	0.380	0.803
Area-level model	0.405	0.749
Intercensus benchmark		

Lessons learned

1. Small area estimation with geospatial data improves accuracy and greatly improves precision of small area estimates of monetary poverty
    - Expands the production possibility frontier between granularity and precision for survey data
    - At low cost because publicly available geospatial indicators predict poverty reasonably well

2. Household model appears to do better than sub-area and area level models in this context
    - More accurate and more precise, especially for sampled areas
    - Information on welfare levels is richer than poverty status
    - Functional form more amenable to poverty estimation
    - Offers more flexibility in calculating different statistics like Ginis and poverty gaps

Lessons learned

3. Using Bayesian or Empirical Bayesian methods is crucial
   - Greatly increases precision and significantly increases accuracy compared to unconditional predictions

4. Optimal survey design changes in presence of free, predictive, big data
   - Surveys should cover all target admin areas if possible
   - Potential gains in accuracy and efficiency to expanding size of second stage of surveys, to improve machine learning for prediction

5. These techniques can be applied to improve survey data at relatively low cost
   - Working on software to facilitate access to free geospatial indicators and application of small area estimation methods

Remaining questions

1. In what circumstances are geospatial poverty maps better than older census-based maps?
    - Can we tell from survey data based on how fast regional poverty patterns are changing?

2. Geospatial features
    - Are there better and/or less expensive geospatial features? Other sources of big data?

3. Machine learning
    - Are there better methods of model selection?
    - Can Bayesian or Empirical Bayesian methods be combined with fancier machine learning methods like random forests and extreme gradient boosting?

4. Other indicators
    - Can method accurately predict other poverty and inequality measures besides headcount like Gini coefficients or Poverty Gap?
    - What method and auxiliary data is best for small area estimates of inequality?

Thank you!

References

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.

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., ... & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. PloS one, 16(9), e0255519.

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

Carter, G. M., & Rolph, J. E. (1974). Empirical Bayes methods applied to estimating fire alarm probabilities. Journal of the American Statistical Association, 69(348), 880-885.

Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2021). Micro-Estimates of Wealth for all Low-and Middle-Income Countries, forthcoming in Proceedings of the National Academiy of Science

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. Scientific American, 236(5), 119-127.

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

Engstrom, R., Hersh, J. S., & Newhouse, D. (2021) Poverty from space: using high-resolution satellite imagery for estimating economic well-being, forthcoming in World Bank Economic Review

References

Fay III, R. E., & Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366a), 269-277.

Ghosh, M. (2020). Small area estimation: its evolution in five decades. Statistics in Transition. New Series, 21(4), 1-22.

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.

Halbmeier, C., Kreutzmann, A. K., Schmid, T., & Schröder, C. (2019). The fayherriot command for estimating small-area indicators. The Stata Journal, 19(3), 626-644.

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), 1-96.

Kreutzmann, A. K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., & Tzavidis, N. (2019). The R package emdi for the estimation and mapping of regional disaggregated indicators. Journal of Statistical Software, 91(7).

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of multivariate analysis, 101(4), 882-892.

References

Masaki, T., Newhouse, D., Silwal, A. R., Bedada, A., & Engstrom, R. (2020). Small area estimation of non-monetary poverty with geospatial data.

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

Peterson, R. A., & Cavanaugh, J. E. (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. Journal of Applied Statistics.

Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences*, *114*(46), E9783-E9792.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, *14*(127), 20160690.

Tarozzi, A., & Deaton, A. (2009). Using census and survey data to estimate poverty and inequality for small areas. *The review of economics and statistics*, *91*(4), 773-792.

World Bank. *World Development Report 2021: Data for Better Lives*. The World Bank, 2021.

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), 1-11.

WORLD BANK GROUP