The spatial distribution of poverty in Sri Lanka in 2016

Ani Silwal, Adane Bedada, Paul Corral, Dilhanie Deepawansa, Ryan Engstrom David Newhouse, Minh Nguyen

WORLD BANK GROUP

IARIW/World Bank Conference
Washington DC
November 7, 2019

WORLD BANK GROUP

Motivation

- High demand for more frequent granular poverty estimates
- Most common method for small area estimation relies on combining a survey with census data (Elbers, Lanjouw, and Lanjouw 2003)
  - Needs a new census to credible measure changes in small areas
- Growing body of work demonstrates strong correlations between features derived from satellite imagery with village-level poverty and population density
  - (i.e. Jean et al, 2016, Engstrom et al, 2017, Engstrom et al, 2018, Head et al, 2017, etc.)
- Can satellite imagery be used for small area estimation?

Motivation

- We compare different methods for obtaining small area estimates in Sri Lankan contexts
- Aim to generate estimates for 331 DS Divisions (subdistricts)
- Using auxiliary data from ~14,000 GN Divisions (villages)
  - Sample contains ~2500 GN Divisions covering 328 DS Divisions
- Can use either Bayesian models (Pokhriyal and Jacques, 2017) or Empirical Best Linear Unbiased Predictors (EBLUP) models
- Prefer EBLUPs because
  - Framework very similar to traditional ELL (Van der Weide, 2014)
  - Doesn't require specifying prior distribution
  - Huge statistics literature on EBLUPs (Morris, 1983, Jiang and Lahiri 2006, many others)

What methods could we use for small area estimation?

- 1. Area model: Traditional Fay-Herriot (Fay and Herriot 1979):
  - Simple EBLUP model at subdistrict level
  - Weighted average of direct survey estimates and model predictions
    - Predictions get more weight if they are more precise relative to sample estimate
  - Advantages
    - Simple, easy to understand, well-established literature
  - Disadvantages:
    - Does not utilize village level variation in satellite indicators
    - Underestimates standard errors by ignoring uncertainty in variance estimates
    - Assumes poverty rate is a linear function of auxiliary data

What methods could we use for SAE?

2. Unit level models
    - Regress household welfare on village and subdistrict level indicators with subdistrict random effect and household error term

- Random effect is conditioned on the sample
    - Differs from traditional ELL, which estimates random effect in survey model and applies it to census simulation
    - Average of sample residuals is used as a prior distribution for area random effect (assuming normality)
    - This reduces variance of area effect and adds precision

- Conditioning on sample is critical when auxiliary data varies at village level
    - Sample is now large compared with "effective size" of census
    - Sample contributes important variation to increase precision

What methods could we use for SAE?

- Predict log household per capita consumption using auxiliary data at village and district level
  - Simulate log household welfare in census by drawing from district random effect (conditioned on sample residuals) and household error term and adding to predicted log welfare
  - Exponentiate simulated log welfare and compare with poverty line
  - Repeat many (usually 100) times

Data: HIES and Census

- **Household Income and Expenditure Survey (HIES) 2016**
  - Nationally representative survey used to estimate the official poverty statistics in Sri Lanka (N=21,756 households).
  - The HIES 2016 sample includes households in 328 out of 331 DSDs and 2,491 out of 14,022 GNDs in the country.

- **Census of Population and Housing 2012**:
  - Includes 5.23 million households with 19.74 million individuals.
  - We compute GND and DSD-level means of all candidate variables

Satellite imagery: spatial features

Source of satellite data: cloud free mosaic of -2017-2018 Sentinel-2 imagery, collected every 5 days by Sentinel 2A and 2B satellites. Imagery is made publicly available by the European Space Agency. Resolution is 10m per pixel.

- **Fourier Transform** (FT): used to detect high or low frequency of lines.
- **Gabor Filter:** a linear Gaussian filter used for edge detection
- **Histogram of Oriented Gradients** (HOG): captures the orientation and magnitude of the shades of the image
- **Lacunarity** (LAC): describes the extent of gaps and holes in a texture
- **Line Support Regions** (LSR): characterize line attributes
- **Normalized Difference Vegetation Index** (NDVI), vegetation index that provides information about the health and amount of vegetation
- **PanTex**, which is a built-up presence index derived from the grey-level co-occurrence matrix (Pesaresi et al, 2008)
- **Structural Feature Sets** (SFS), which are statistical measures to extract the structural features of direction lines (Huang et al, 2007)

Results: Fay-Herriot model at subdistrict level

Method	Estimated poverty rate	Coefficient of Variation
Direct estimates	4.1	71.9
Fay-Herriot (Traditional)	5.2	60.0

Note: Direct estimates are computed using Horwitz-Thompson approximation, which allows for correlation among all households within district unlike the standard method in Stata

Model of log of consumption used for Unit level model (Beta model)
(Specification chosen from candidate variables using lasso)

Variable	Coefficient	T-stat
Share of households with highest education of Degree or higher in GN	1.22***	(7.34)
Share of households with highest education = Grades 6-9 in GN	-0.27***	(-4.71)
Percent of HHs in DS with access to pipe water within premises	0.00**	(2.10)
Share of households whose floors are made from perm/semi-perm materials in GN	0.13**	(2.00)
Share of household heads that are employed in public sector in GN	0.19**	(2.21)
Mean household size in GN	-0.15***	(-8.48)
Percent of HHs in GN with access to internet	0.00***	(5.73)
Share of households that own a TV in GN	0.25***	(4.33)
Percent of HHs in GN with access to pipe water within premises	0.00**	(2.16)
Mean night light intensity in 2016, DS	0.00*	(1.90)
Z-score of 2016-Q3 rainfall in GN	-0.10***	(-3.80)
Tree Cover - Gain, GN	0.95***	(3.93)
Mean slope of GN	-0.00***	(-2.79)
ndvi_sc7_std_gn	0.54*	(1.91)
pantex_sc3_min_mean_ds	-2.62**	(-2.00)
fourier_sc3_std_ds	0.01*	(1.73)
sfs_sc7_mean_gn	0.01***	(2.76)
Sector = Estate	-0.17***	(-7.16)
District==Ampara & sector==Urban	-0.21***	(-2.75)
District==Trincomalee & sector==Urban	-0.20**	(-2.49)
District==Kurunegala & sector==Rural	0.09**	(2.44)
District==Puttalam & sector==Rural	0.13***	(2.71)
District==Badulla & sector==Rural	-0.14***	(-3.34)
District==Kalutara & sector==Rural	0.09**	(2.23)
District==Galle & sector==Rural	0.09**	(2.29)
District==Matara & sector==Estate	-0.30***	(-2.93)
District = Batticaloa	-0.11*	(-1.95)
District = Anuradhapura	0.14***	(3.14)
District = Ratnapura	-0.15***	(-2.87)
District = Hambantota	0.18***	(3.04)
Constant $R^2 = 0.18$ N=21,571	9.21***	(82.17)

Results: Comparison of SAE methods

			Mean		Coefficient of Variation	
Method	Het. correction	Sampling weights	Consumption (Rs./mo/pc)	Poverty (%)	Consumption (Rs./mo/pc)	Poverty (%)
			Weighted mean		Unweighted mean	
Direct estimates			14,193	4.1	17.9	71.9
R - EBP	No	No	14,820	5.2	8.2	38.2
Stata - EBP	No	No	14,724	5.2	7.3	27.6
Stata - EBP	No	Yes	14,770	5.3	7.4	27.5
Stata - EBP	Yes	No	14,633	4.1	7.5	49.5
Stata - EBP	Yes	Yes	14,708	4.2	7.7	51.5

Note: Direct estimates are computed using Horwitz-Thompson approximation

Summary of key results

- R SAE gives major increases in precision relative to direct estimates and F-H
  - Cuts CV in half relative to direct estimates, equivalent to a four-fold increase in size of effective sample
  - CV for average consumption is 8.2%, <10% threshold used by NSO for district poverty estimates
- Heteroscedasticity correction reduces both bias and precision of poverty estimate
  - May substantially increase coverage of poverty estimate
- R SAE poverty estimates are less precise than Stata SAE for same specification
  - More work needed to understand why
- Weights have a minor effect on the results
- Small area estimation with remote sensing appears to work well when using an appropriate method but more testing is needed

Next steps

- Test models selected only from a pool of remote sensing variables
- Test properly specified subarea models (Torabi and Rao, 2014)
- If time, implement coverage test of "welfare index" using 2012 census