Using Alternative Spatial data sources for SAE in developing countries

Angela Luna Nikos Tzavidis Paul Smith
Southampton Statistical Sciences Research Institute
University of Southampton

Jessica Steele Kristine Nilsen
WorldPop
University of Southampton

Statistische Woche

Universität Trier, 10-13 September 2019

This project has received support from the European Union Horizon 2020 programme via INGRID (Grant 730998) and MAKSWELL (Grant 770643) projects.

Small area estimation problem

- Interest on estimation of social and demographic indicators at sub-national levels
- Estimates are often built using survey data. Insufficient sample sizes impede obtaining reliable estimates for small areas under the typical survey inference framework
- Area-specific indicator is seen as a realisation of a statistical model with common terms across areas. Use of auxiliary data sources and statistical models to *borrow strength* across areas
- Very active area of research. See Rao and Molina (2015); Pfeffermann (2013).

Alternative data sources

$$\text{Survey} \sim \text{Census} + \text{Administrative records} + \ldots$$

- Remotely Sensed (RS) data
- Mobile phone (CDR) data
- Web-scraped data, Social media data, ...

Possible uses as covariates, response, reference,...

See Marchetti et al. (2015); MAKSWELL Project, Work Package 2: van den Brakel, Buelens, et al. (2019) and van den Brakel, Smith, et al. (2019).

Use of RS data for SAE

Advantages

- Broadly available and frequently updated (*no one is left behind*). Low cost. Particularly useful in low-income countries where high quality survey, census and administrative data may be scarce. See:
  - US geological survey `https://earthexplorer.usgs.gov/`
  - European Space Agency `www.geoportal.org`
  - `http://trends.earth/docs/en/` (land coverage)
  - Open Street Map `http://extract.bbbike.org/`
  - WorldPop Research Project `http://www.worldpop.org.uk/`
  - ... (more at the end).
- Flexible definition of target geography

Use of RS data for SAE

Limitations

- Explanatory power? Unclear link between variables and outcome
- Potential for irregular coverage, e.g., due to atmospheric conditions
- Substantial pre-processing. Measurement error?
- Potential for uneven reference periods
- It is not clear in which situations a given aggregation strategy should be preferred
- Adequate use requires some degree of specialized knowledge

Modelling approaches

Statistical vs Algorithmic/Mapping approaches †.

Statistical modelling (SAE)

- Real observations of the phenomenon are required (survey)
- Sampling design is taken into account
- Model assessment (GOF) and area-specific uncertainty of estimates (MSE)
- Area-level models: Fay-Herriot: Frequentist/HB; can include spatial/temporal effects
- Relatively coarse geography. Administrative boundaries + survey design

† For a detailed discussion from a geospatial perspective see the report from the Task Team on Satellite Imagery and Geospatial data, UN (2017)

Modelling approaches (2)

Algorithmic/Mapping approaches

- Observations required (supervised methods)
- Generally, no consideration of sampling design
- GLMM's, classification trees, support vector machines,... Often inclusion of spatial effects as default
- Aim to very granular geographies
- Crossvalidation. MCR and MSEP. If Bayesian inference, uncertainty measured using posterior variances
- Examples:
  - Poverty measurement in Bangladesh; Steele et al. (2017)
  - Slum mapping in Casablanca, Morocco; Rhinane et al. (2011)

Presentation aims

Illustration of both approaches in a realistic set-up.
Poverty measurement in Bangladesh in the spirit of Steele et al. (2017). Wealth index $\sim$ RS covariates.

Aims:

- Identify common points and differences between both approaches
- Illustrate the use of standard packages for each approach (`sae`, `BRugs`, `inla`) for the fitting of spatial and non-spatial models
- Identify potential methodological issues

Poverty measurement in Bangladesh

Target: Average WI by Upazila (Level 3).

Survey data - DHS 2014

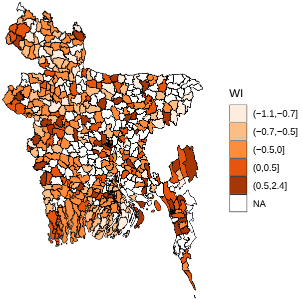
- Stratified 2-stage cluster design. At least one cluster selected in 365/508 (72%) Upazilas
- Response: WI computed via PCA
- $n = 17$K households. $\bar{n}_i = 34$ households

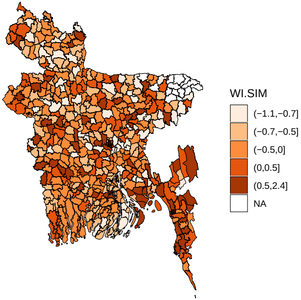
RS data:

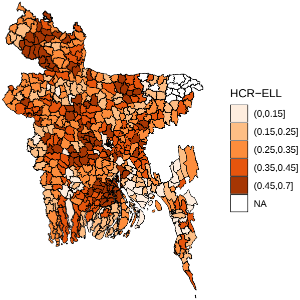
- 18 variables as starting point in Steele et al. (2017).
  - Night time lights (NOAA-US)
  - Elevation (CGIAR-CSI)
  - Accessibility to areas with more than 50K people (A global map of accessibility - European Commission Joint Research Centre)

Complete cases exercise

Imputation of direct estimates and their sampling variances in out-of sample areas.







Methods

Fay-Herriot model

$$\boxed{\hat{\theta}_i^d = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i + e_i}$$

$\hat{\theta}_i^d$ a direct estimator of θ_i; $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$.

From a frequentist perspective, σ_i^2 is assumed known (estimation + smoothing). An EBLUP for θ is given by

$$\hat{\theta}_i^{FH} = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} + \hat{u}_i = \hat{\gamma}_i \hat{\theta}_i^d + (1 - \hat{\gamma}_i) \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$$

with $\hat{\gamma}_i = \hat{\sigma}_u^2 / (\sigma_i^2 + \hat{\sigma}_u^2)$. Analytic MSE estimator by Prasad and Rao, 1990. Parametric bootstrap can also be used. Available in R package `sae`.

Methods

Fay-Herriot model

$$\boxed{\hat{\theta}_i^d = \boldsymbol{x}_i^T \boldsymbol{\beta} + u_i + e_i}$$

$\hat{\theta}_i^d$ a direct estimator of θ_i; $u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$.

From a Bayesian (HB) perspective:

$$\hat{\theta}_i^d | \theta_i \overset{ind}{\sim} N(\theta_i, \sigma_i^2)$$

$$\theta_i | \beta, \sigma_u^2 \overset{ind}{\sim} N(\boldsymbol{x}_i^T \boldsymbol{\beta}, \sigma_u^2)$$

Information on σ_i^2 can included via another level or an informative prior. See You and Chapman (2006).

The posterior mean and variance of θ_i are used for inference.

Available in R packages `BayesSAE`, `hbsae`. Can also use BUGS, JAGS, Stan, ...

Methods (2)

Fay-Herriot model - Spatial extensions

SAR

Assumes $\mathbf{v} = \rho \mathbf{W}\mathbf{v} + \mathbf{u}$, with $\mathbf{u} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I})$, therefore

$$\hat{\boldsymbol{\theta}}^d = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho \mathbf{W})^{-1}\mathbf{u} + \mathbf{e}$$

See Cressie (1993) and Pratesi and Nicola Salvati (2008).

- $\mathbf{W}$ is an adjacency matrix row-standardised to sum 1, leading to $\rho \in (-1, 1)$.
- ML/REML estimates obtained via `sae`. MSE analytical + bootstrap.

Methods (2)

Fay-Herriot model - Spatial extensions

iCAR

Besag, York, and Mollié (1991). It assumes

$$v_i | v_{j \neq i}, \sigma_u^2 \sim N\left(\frac{1}{n_i}\sum_{j \sim i} v_j,\ \frac{1}{n_i}\sigma_u^2\right) \tag{1}$$

which leads to $\mathbf{v} \sim N(\mathbf{0}, \sigma_u^2(\mathbf{D} - \mathbf{W}))$, with $\mathbf{D} = diag\,\{n_i\}$ and $\mathbf{W}$ an adjacency matrix. Notice that (1) doesn't define a proper distribution as $(\mathbf{D} - \mathbf{W})$ is non-invertible. The constraint $\sum_i v_i = 0$ is required to make the v_i identifiable.

- Computationally convenient as covariance matrix is sparse
- Available in R-INLA: latent specification `besag`, `bym`, and in BUGS: distribution `car.normal`.

Fitting non-spatial models

Complete cases. Point estimates and variances obtained using the sampling design. Smoothing of variance estimates using GVF.

$$\widehat{WI_i} = \beta_0 + \beta_1 \times ELEV + \beta_2 \times NL + \beta_3 \times ACC + u_i + e_i$$

$u_i \overset{iid}{\sim} N(0, \sigma_u^2)$; $e_i \overset{ind}{\sim} N(0, \sigma_i^2)$. R-INLA latent specification `iid`.

- M1 Standard FH model using `sae`. $\sigma_i^2 = \hat{\sigma}_i^2$ fixed
- M2 Standard Gaussian model in R-INLA. $\sigma_i^2 = \sigma_e^2$ unknown
- M3 R-INLA with $\sigma_{e_i}^2 = g_i \sigma_e^2$; $g_i = v_i / \bar{v}_i$ fixed.
  $\tau = 1/\sigma_e^2$; $\pi(\tau) \sim \text{Gamma}\left(\frac{\bar{n}_i - 1}{2} - 1\,,\ \frac{(\bar{n}_i - 1)\bar{v}_i}{2}\right)$.
- M4 HB using `BRugs` with $\pi(\tau_i)$ as in M3.

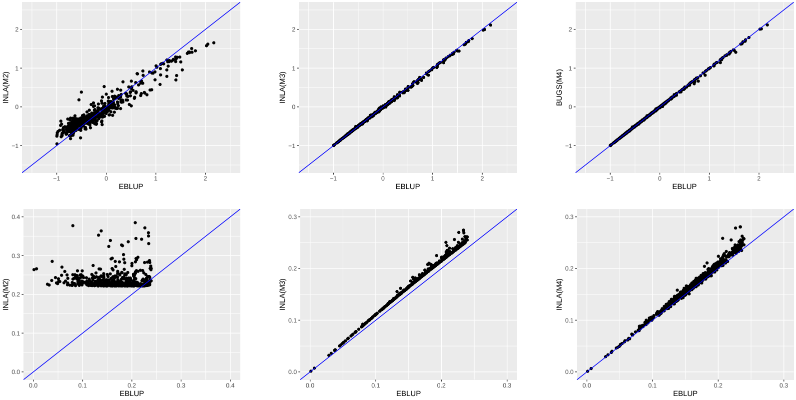
Results non-spatial models

- Small differences in the fixed effects
- Large differences in the variance decomposition. Using scaling to allow for heteroscedasticity nearly eliminates all differences

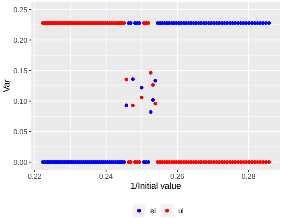
	M1	M2	M3	M4
$\hat{\beta}_0$	0.6662	0.6866	0.6604	0.6469
$\hat{\beta}_{elev}$	-0.0553	-0.0530	-0.0557	-0.0548
$\hat{\beta}_{nl}$	0.3137	0.3112	0.3141	0.3159
$\hat{\beta}_{acc}$	-0.0878	-0.0892	-0.0874	-0.0838
$\hat{\sigma}_e^2$	0.0362	0.1219	0.0423	0.0408
$\hat{\sigma}_u^2$	0.1889	0.1058	0.1800	0.1838

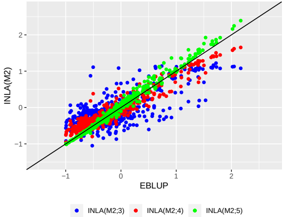
Results non-spatial models (2)

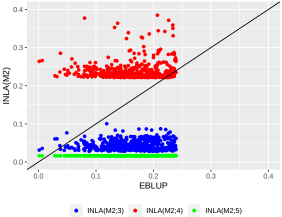
- Some impact on the point estimates
- Large Impact on uncertainty measures



Results non-spatial models (3)







Fitting spatial models

- M5 SAR correlation in `sae`. $\sigma_i^2 = \hat{\sigma}_i^2$ fixed
- M6 iCAR correlation in R-INLA, `besag` specification. $\sigma_i^2 = \sigma_e^2$ unknown
- M7 As M6, `bym` specification.
- M8 As M7, $\sigma_{e_i}^2 = g_i \sigma_e^2$; $g_i = v_i / \bar{v}_i$ fixed.
  $\tau = 1/\sigma_e^2;\ \pi(\tau) \sim \text{Gamma}\left(\frac{\bar{n}_i - 1}{2} - 1\,,\ \frac{(\bar{n}_i - 1)\bar{v}_i}{2}\right)$.
- M9 HB using `BRugs` with $\pi(\tau_i)$ as in M8.

Results spatial models

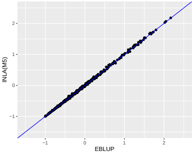
	M1	M5	M6	M7	M8	M9
$\hat{\beta}_0$	0.6662	0.6567	0.4208	0.4141	0.4147	0.3414
$\hat{\beta}_{elev}$	-0.0553	-0.0472	0.0688	0.0718	0.0716	0.0869
$\hat{\beta}_{nl}$	0.3137	0.2999	0.2714	0.2695	0.2695	0.2649
$\hat{\beta}_{acc}$	-0.0878	-0.0939	-0.1086	-0.1093	-0.1093	-0.1078
$\hat{\sigma}_e^2$	0.0362	0.0362	0.1875	0.0793	0.0378	0.0408
$\hat{\sigma}_u^2$	0.1889	0.1816	0.0432	0.1020	0.1480	0.1344
$\hat{\sigma}_v^2$				0.0448	0.0437	0.1144
ρ		0.2458				

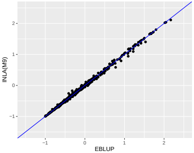
- Spatial confounding for the iCAR models. See Reich, Hodges, and Zadnik (2006), Prates, Assunção, Rodrigues, et al. (2019)

Results spatial models (2)

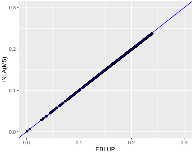
M5 (SAR - `sae`) and M9 (iCAR - `BRugs`):

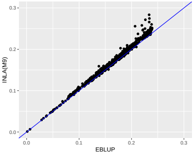
$\widehat{WI}_i$





SRMSE



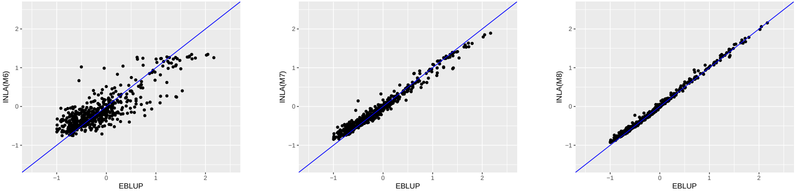


- Not relevant improvement by the inclusion of spatially correlated random effects.

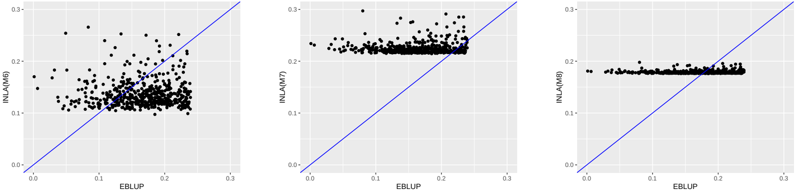
Results spatial models (3)

M6 (Besag), M7 (bym) and M8 (bym + scale) - `R-INLA`:

$\widehat{WI_i}$



SRMSE



Concluding remarks

- Ready to use software for approximate Bayesian inference offers interesting possibilities. However, some degree of specialized knowledge is necessary for its correct use.
- The use of spatially correlated effects did not improve substantially our estimates. Spatial confounding appears under iCAR. Negligible effects on the point/variance estimates.
- Although the proposed specification of models for R-INLA leads to reasonable point estimates, more work is necessary to identify a way to produce appropriate variance estimates.
- Including spatially correlated random effects may help to improve estimates in out-of sample areas. However, assessing whether the type of shrinkage induced by a particular model is reasonable for SAE purposes would be relevant. Gomez-Rubio et al. (2005), Saei and Chambers (2005), Chandra, N. Salvati, and Chambers (2007).

References

Besag, Julian, Jeremy York, and Annie Mollié (1991). "Bayesian image restoration, with two applications in spatial statistics". In: *Annals of the institute of statistical mathematics* 43.1, pp. 1–20.

Chandra, H., N. Salvati, and R. Chambers (2007). *Small area estimation for spatially correlated populations - A comparison of direct and indirect model-based methods*. Research report. University of Wollongong.

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Gomez-Rubio, R. et al. (2005). *Bayesian Statistics for Small Area Estimation*. Tech. rep. NCRM.

Marchetti, Stefano et al. (2015). "Small area model-based estimators using big data sources". In: *Journal of Official Statistics* 31.2, pp. 263–281.

Pfeffermann, Danny (2013). "New important developments in small area estimation". In: *Statistical Science* 28.1, pp. 40–68.

Prasad, N.G.N. and J.N.K. Rao (1990). "The estimation of the mean squared error of small-area estimators". In: *Journal of the American Statistical Association* 85.409, pp. 163–171.

Prates, Marcos Oliveira, Renato Martins Assunção, Erica Castilho Rodrigues, et al. (2019). "Alleviating Spatial Confounding for Areal Data Problems by Displacing the Geographical Centroids". In: *Bayesian Analysis* 14.2, pp. 623–647.

Pratesi, Monica and Nicola Salvati (2008). "Small area estimation: the EBLUP estimator based on spatially correlated random area effects". In: *Statistical methods and applications* 17.1, pp. 113–141.

References (cont.)

Rao, J.N.K. and Isabel Molina (2015). *Small area estimation*. 2nd. John Wiley & Sons.

Reich, Brian J, James S Hodges, and Vesna Zadnik (2006). “Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models”. In: *Biometrics* 62.4, pp. 1197–1206.

Rhinane, Hassan et al. (2011). “Detecting slums from SPOT data in Casablanca Morocco using an object based approach”. In: *Journal of Geographic Information System* 3.03, p. 217.

Saei, A and R. Chambers (2005). *Working paper M05/03: Empirical Best Linear Unbiased Prediction for out of sample areas*. Research report. Southampton Statistical Sciences Research Institute, University of Southampton.

Steele, Jessica E et al. (2017). “Mapping poverty using mobile phone and satellite data”. In: *Journal of The Royal Society Interface* 14.127, p. 20160690.

Van den Brakel, J.A., B. Buelens, et al. (2019). *Aspects of existing databases, traditional and non-traditional data sources and collection of good practices*. Work Package 2, deliverable 2.1. MAKSWELL Project.

Van den Brakel, J.A., P.A. Smith, et al. (2019). *Methodological aspects of using Big data*. Work Package 2, deliverable 2.2. MAKSWELL Project.

You, Yong and Beatrice Chapman (2006). “Small area estimation using area level models and estimated sampling variances”. In: *Survey Methodology* 32.1, p. 97.

- US geological survey https://earthexplorer.usgs.gov/
- European Space Agency www.geoportal.org
- http://trends.earth/docs/en/ (land coverage)
- Open Street Map http://extract.bbbike.org/
- WorldPop Research Project http://www.worldpop.org.uk/
- National Oceanic and Atmospheric Administration http://ngdc.noaa.gov/eog/viirs.html
- European Commission Joint Research Centre https://forobs.jrc.ec.europa.eu/products/gam/
- Center for International Earth Science Information Network(CIESIN) http://sedac.ciesin.columbia.edu/data/
- CGIAR Consortium for Spatial Information http://www.cgiar-csi.org/data
- WorldClim Global Climate Data http://www.worldclim.org
- European Commission Global Human Settlement Layer https://ghsl.jrc.ec.europa.eu/
- World Database on Protected Areas http://www.protectedplanet.net/
- Oak Ridge National Laboratory Land Coverage http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10024_1
- ETH Zurich International Conflict Research http://www.icr.ethz.ch/data/geoepr