Exploring geographic variation using small area estimation

Nikos Tzavidis
(University of Southampton)[1]

RSS Social Statistics Section
London, June 27 2016

[1] Joint work with Timo Schmid, Angela Luna,Li-Chun Zhang, Natalia Rojas

The SAE problem

Aim: Estimate finite population linear & non-linear parameters e.g. averages, medians, percentiles.

User requirements for more disaggregated estimates have been increasing in the past 10 years or so. Now we need estimates for many small areas:

- Geographic areas: municipalities, districts, neighbourhoods,...
- Domains: combinations of factors e.g. Age, Ethnicity, Labour Force status,...by area.

For inference to work well, s needs to be big enough.

- Areas with 2, 3 observations?
- Areas with no observations at all?

SAE addresses the problem of small domain/area sample sizes.

Three stages

Stage I. Specification

1. Specify user needs.
2. Specify a set of target indicators to be estimated and a target geography/set of domains.

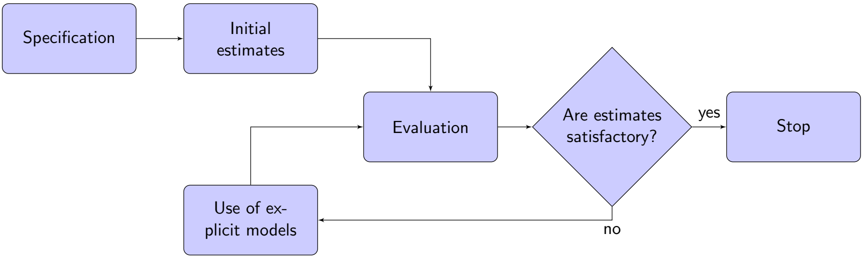
Stage II. Analysis/Adaptation

3. Initial estimates.
4. Use of explicit models.

Stage III. Evaluation

5. MSE estimation.
6. Model and Design based evaluation.
7. Further evaluation tasks.

Three stages



Stage I. Specification

Target geography

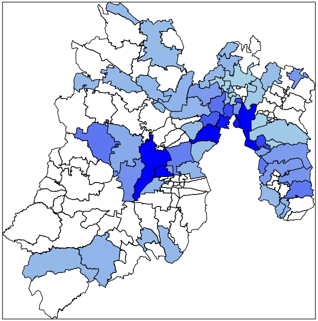
A chosen level of geography should provide meaningful (background of the problem) and useful (data availability) estimates.

Follow in decreasing level of aggregation and avoid the temptation of getting unrealistically low.

- SAE is a prediction problem. Access to good auxiliary data is, in most cases, crucial.
- Survey, Census, Administrative data can be used for modelling and evaluation purposes.
- Consider the coverage of the sources in relation to the target geography.

Example 1: Estimating non-linear indicators in the State of Mexico (EDOMEX)

Stage I. Specification



- Estimate income related indicators for municipalities.
- Geography is fixed, defined by the user.
- 125 municipalities in State of Mexico. Only 58 are included in the survey. For the municipalities in the sample, the average sample size is 47 households.
- Definition of geography determines use of SAE methods.

Stage II. Analysis/Adaptation

3. Initial estimates.
4. Use of explicit models.

Stage II. Analysis/Adaptation

3. Initial estimates

Produce a triplet of estimates (direct, synthetic, composite) for each area at the given level of geography:

- Direct: uses only-domain specific data, e.g., $\hat{\bar{Y}}_k^D = \bar{X}_k \hat{\beta}_k$.
- Synthetic: borrows information from other areas/domains, e.g., $\hat{\bar{Y}}_k^S = \bar{X}_k \hat{\beta}$.
- Composite: it is a convex combination of a Direct and a Synthetic estimators, e.g., $\hat{\bar{Y}}_k^C = \phi \hat{\bar{Y}}_k^D + (1 - \phi) \hat{\bar{Y}}_k^S$.

Unlikely these estimators to produce estimates with acceptable coefficients of variation (CVs).

Stage II. Analysis/Adaptation

4. Use of explicit models

General considerations

- Access to microdata? Unit-level or Area-level models.
- Continuous responses: start with Linear Models.
- Discrete responses: start with Generalized Linear Models.
- Unexplained area heterogeneity: Mixed Models.
- Out of sample areas? Synthetic estimators.

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

Some non-linear Income-based indicators

- FGT measures (Foster et al.,1984))

$$FGT(\alpha, t) = \sum_{i=1}^{N} \left(\frac{t-y_i}{t}\right)^{\alpha} \mathbb{1}(y_i \leq t),$$

$\alpha = 0$ - Head Count Ratio; $\alpha = 1$ - Poverty Gap.

- Gini coefficient

$$Gini = \frac{N+1}{N} - \frac{2\sum_{i=1}^{N}(N+1-i)y_{(i)}}{N\sum_{i=1}^{N} y_{(i)}}.$$

- Quintile Share Ratio

$$QSR_{80/20} = \frac{\sum_{i=1}^{N}[y_i \mathbb{1}(y_i > q_{0.8})]}{\sum_{i=1}^{N}[y_i \mathbb{1}(y_i \leq q_{0.2})]}.$$

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

SAE methodologies for complex Income-based indicators

- The World Bank Approach (Elbers et al., 2003).
- The EBP Approach (Molina & Rao, 2010, CJS).
- The M-Quantile Approach (Marchetti et al., 2012 ; Chambers & Tzavidis, 2006, Biometrika).
- EBP based on normal mixtures (Elbers & Van der Weidel, 2014; Lahiri and Gershunskaya, 2011).
- MvQ methods based on Asymmetric Laplace distribution (Tzavidis et al., 2015).

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

The EBP Method (under normality)

Point of departure: Unit-level Mixed effects model.

$$y_{ik} = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + u_k + \epsilon_{ik}, u_k \sim N(0, \sigma_u^2); \epsilon_{ik} \sim N(0, \sigma_e^2).$$

Summary of the Method

- Use sample data to estimate β, σ_u^2, σ_ϵ^2, γ_k.
- Generate $u_k^* \sim \ N(0, \hat{\sigma}_u^2(1-\gamma_k))$ and $\epsilon_{ik}^* \sim \ N(0, \hat{\sigma}_\epsilon^2)$,

$$y_{ik}^* = \mathbf{x}_{ik}^T\hat{\boldsymbol{\beta}} + \hat{u}_k + u_k^* + \epsilon_{ik}^*$$

- Calculate the indicator of interest using the y_{ik}^*.

Micro-simulation of a synthetic population. Repeat the process L times.

Stage II. Analysis/Adaptation

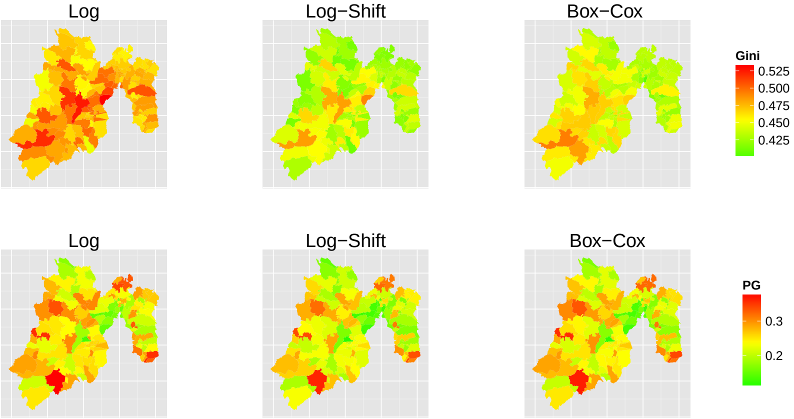
4. Use of explicit models - Adaptation

If residual diagnostics indicate violation of model assumptions, Adapt the model.

- Explore the use of transformations. Deciding on appropriate transformations is not straightforward, but offers a possible avenue for improving the model.
- Use robust methods as an alternative to transformations (Chambers & Tzavidis, 2006; Ghosh et al., 2008; Sinha & Rao, 2009; Chambers et al., 2014; Dongmo Jiongo et al., 2013).
- Use non-parametric models (Opsomer et al., 2006; Ugarte et al., 2009)
- Elaborate the random effects structure e.g. include spatial structures (Pratesi & Salvati, 2008; Schmid & Münnich, 2014).
- Consider extensions to two-fold models (Morales et al., 2015).

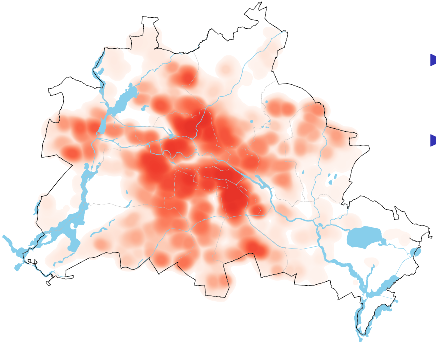
Stage II. Analysis/Adaptation

4. Use of explicit models and Adaptation. SAE in EDOMEX



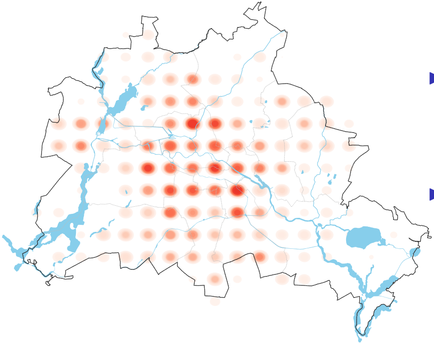
Choice of transformation possibly important for parameters involving the whole distribution. Gini more sensitive than PG

Example 2: Estimating population densities in the presence of measurement error in geo-coordinates Groß, Rendtel, Schmid, Schmon, Tzavidis (2016) Journal of the Royal Statistical Society A



- Estimate area-specific ethnic and age densities in Berlin
- Berlin register data publicly available but geo-coordinates aggregated at 447 urban planning areas - **Density structure is not preserved**

Example 2: Estimating population densities in the presence of measurement error in geo-coordinates Groß, Rendtel, Schmid, Schmon, Tzavidis (2016) Journal of the Royal Statistical Society A



- **Solution**: Treat aggregation of geo-coordinates as a measurement error process
- Reverse measurement error, derive precise density estimates at flexible levels of geography.

Rounding and kernel density estimation

Measurement error model & estimation

- True, unknown, values $X_i = (X_{i1}, X_{i2})$ given the rounded values $W_i = (W_{i1}, W_{i2})$ are distributed in a rectangle with W_i in its center,
$$[W_{i1} - \frac{1}{2}\mathrm{r}, W_{i1} + \frac{1}{2}\mathrm{r}] \times [W_{i2} - \frac{1}{2}\mathrm{r}, W_{i2} + \frac{1}{2}\mathrm{r}],$$
r denotes the rounding parameter.
- Can be seen as a measurement error model with uniformly distributed measurement error $U_i = (U_{i1}, U_{i2})$, $U_{i1}, U_{i2} \sim Unif(-\frac{1}{2}r, \frac{1}{2}r)$ and U_{i1}, U_{i2} independent of W_{i1} and W_{i2} such that,
$$X_{i1} = W_{i1} + U_{i1}, \ i = 1, 2, .., n$$
$$X_{i2} = W_{i2} + U_{i2}, \ i = 1, 2, .., n.$$

Measurement error model & Estimation

From Bayes theorem follows that

$$\pi(X|W) \propto \pi(W|X)\pi(X)$$

- $\pi(W|X)$ (measurement error model) is defined by a product of Dirac distributions

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in [W_{i1} - \frac{1}{2}r, W_{i1} + \frac{1}{2}r] \times [W_{i2} - \frac{1}{2}r, W_{i2} + \frac{1}{2}r] \\ 0 & \text{else.} \end{cases}$$

- $\pi(X) = \prod_{i=1}^{n} f(X_i)$ is initially unknown, we propose an iterative procedure.

- Estimation via a stochastic Expectation?Maximization
  - E-step: Draw samples from $\pi(X_i|W_i)$ creating a pseudosample of X in each iteration as a replacement of the E-step
  - M-step: Apply kernel density estimationto the pseudo-sample
  - Iterate E and M steps until convergence

Measurement error model & Estimation

From Bayes theorem follows that

$$\pi(X|W) \propto \pi(W|X)\pi(X)$$

- $\pi(W|X)$ (measurement error model) is defined by a product of Dirac distributions

$$\pi(W_i|X_i) = \begin{cases} 1 & \text{for } X_i \in [W_{i1} - \frac{1}{2}r, W_{i1} + \frac{1}{2}r] \times [W_{i2} - \frac{1}{2}r, W_{i2} + \frac{1}{2}r] \\ 0 & \text{else.} \end{cases}$$

- $\pi(X) = \prod_{i=1}^{n} f(X_i)$ is initially unknown, we propose an iterative procedure.

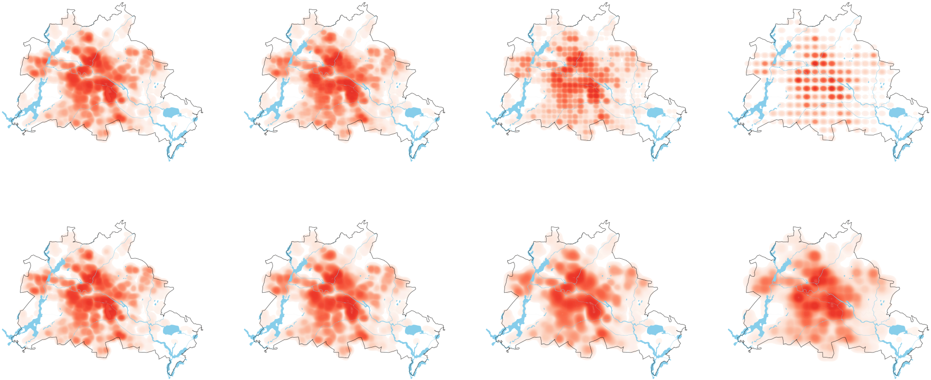
- Estimation via a stochastic Expectation?Maximization
  - E-step: Draw samples from $\pi(X_i|W_i)$ creating a pseudosample of X in each iteration as a replacement of the E-step
  - M-step: Apply kernel density estimationto the pseudo-sample
  - Iterate E and M steps until convergence

Application: Data Sources

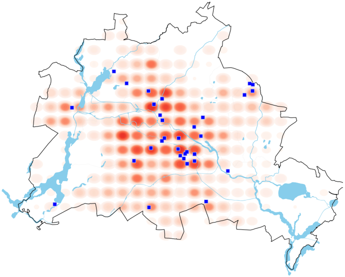
- The data contains all **308,754 Berlin household addresses** on the 31st of December 2012 with the **exact geo-coded coordinates** subject to different degrees of rounding errors.
- Registration at the local residents' office is compulsory in Germany and is carried out by the federal state authorities.
- One of the scenarios we explore is **rounding by using grids of size 2000 meters by 2000 meters** that approximately correspond to the LOR geography.
- The original data includes the total number of residents at their principal residence and the number of persons according to **some key demographic characteristics**:
    - Ethnic background (Ethnic)
    - Age (Age over 60).

Density of population: Ethnic minority background

Naive (top panel) and SEM estimators (bottom panel) with rounding step sizes of 0 (left), 500, 1250 and 2000 m (right).



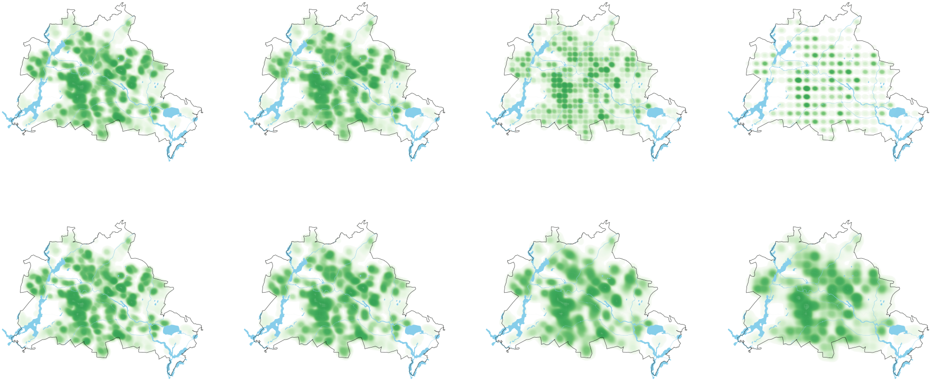
Advisory services for ethnic minorities



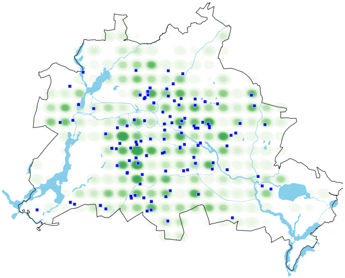
Ethnic background for rounding step size of 2000 m. Blue points indicate migrant advisory centres in Berlin.

Density of population: Aged 60 and above

Naive (top panel) and SEM estimators (bottom panel) with rounding step sizes of 0 (left), 500, 1250 and 2000 m (right).



Care for the elderly



Age above 60 for rounding step size of 2000 m. Blue points indicate retirement houses in Berlin.

Innovations in SAE Methodologies

NCRM Innovation project funded by ESRC

WP 1: Innovations in statistical methodologies

1 Model specification and data transformations
   1.1 Scaled power transformations
   1.2 Optimal values and ML/REML estimation
   1.3 Sensitivity analysis
2 Semi/non-parametric methods for continuous and discrete outcomes
   2.1 Semi-parametric estimation of distribution functions
   2.2 Robust prediction of random effects via discrete mixtures
   2.3 Robust SAE methods for discrete outcomes
   2.4 Semi-parametric estimation for discrete outcomes
3 Developing novel measures of uncertainty

Innovations in SAE Methodologies

WP 2: SAE using Indirect Survey Calibration (ISC) / Spatial Microsimulation

1. Model specification
   1.1 Model (Benchmark) selection
   1.2 Donor pool selection
2. ISC algorithms
   2.1 Impact of weight range restrictions on estimate quality
   2.2 Benchmark relaxation strategies
   2.3 Integerisation
3. Estimating uncertainty

WPs 3 - Bridging the gaps between WPs 1 and 2

1. Spatial variability of Census covariates
2. Statistical theorisation and translation
3. Full empirical performance evaluation of the methods across WPs 1 & 2

The Team

University of Southampton

- Nikos Tzavidis
- Li-Chun Zhang
- Yves Berger
- Graham Moon

University of Liverpool

- Paul Williamson

University of Sheffield

- Adam Whitworth

University of Exeter

- Karyn Morrissey

University of Portsmouth

- Liz Twigg

International Experts & Stakeholders

Free University Berlin

- Timo Schmid

University Technology Sydney

- James Brown

University of Wollongong

- Ray Chambers

Australian National University

- Steve Haslett

National & International Organisations

- UK Office for National Statistics
- Welsh Assembly Government
- Mexican National Council for the Evaluation of Social Development Policy (CONEVAL)
- National Statistics Office of Brazil