Small Area Estimation of Poverty Indicators using Interval Censored Income Data

Paul Walter, Marcus Groß, Timo Schmid & Nikos Tzavidis
Freie Universität Berlin & University of Southampton

1. Motivation

- In order to fight poverty, it is essential to have knowledge about its spatial distribution.
- Small area estimation (SAE) methods enable the estimation of poverty indicators at a geographical level where direct estimation is either not possible, due to a lack of sample size or very imprecise (Rao & Molina , 2015).
- One commonly used SAE method is the empirical best predictor (EBP) (Molina, 2010).
- Estimation becomes imprecise, when due to confidentially or other reasons the dependent variable in the underlying mixed model, such as income, is censored to particular intervals.
- To get more precise estimates, two methodologies, one based on the expectation maximization algorithm (EM) (Dempster et al., 1977) (Stewart, 1983) and one based on the stochastic expectation maximization (SEM) algorithm are introduced (Caleux, 1985).

How do the proposed methods assist in improving the precision of small area prediction when the dependent variable is censored to particular intervals?

2. The EBP Approach (Molina & Rao, 2010)

Nested error linear regression model (1)

$$y_{ij} = \mathbf{x}_{ij}^T\beta + u_i + e_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, D,$$
$$u_i \overset{iid}{\sim} N(0, \sigma_u^2), \text{ the random area-specific effects} \qquad (1)$$
$$e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \text{ the unit-level error terms}$$

where y_{ij} is unknown and only observed to fall into a certain interval (A_{k-1}, A_k) on a continuous scale and k_{ij} $(1 \leq k_{ij} \leq K)$ is the indicator into which of the intervals y_{ij} falls.

- Use sample data to estimate β, σ_u, σ_e, u_i
- Generate $u_i^* \sim N(0, \hat{\sigma}_u^2)$ & $e_{ij}^* \sim N(0, \hat{\sigma}_e^2)$

Micro-simulating a synthetic population:

- Generate a synthetic population under the model a large number of times each time estimating the target parameter.
- Linear and non-linear poverty indicators can be computed.

Arbitrary data example

ID	Unobserved y_{ij}	Observed interval	k_{ij}	x_{ij}
1	1251	$[1000, 1500)$	3	861
2	498	$[0, 500)$	1	311
3	898	$[500, 1000)$	2	557
4	753	$[500, 1000)$	2	421
5	325	$[0, 500)$	1	306
.	.	.	.	
.	.	.	.	
.	.	.	.	
N	5212	$[4000, 8000)$	8	1350

3. Methodology

- Reconstructing the distribution of the unknown y_{ij} is necessary to estimate the parameters of model (1).
- From Bayes theorem it follows that $f(y_{ij}|x_{ij}, k_{ij}) \propto f(k_{ij}|y_{ij}, x_{ij})f(y_{ij}|x_{ij})$ with

$$f(k_{ij}|y_{ij}, x_{ij}) = \begin{cases} 1, & \text{if } A_{k-1} \leq y_{ij} \leq A_k \\ 0, & \text{else} \end{cases} \quad \text{and } f(y_{ij}|x_{ij}) \sim N(x_{ij}^T\beta + u_i, \sigma_e^2).$$

4. Estimation and Computational Details (EM and SEM Algorithm)

1. Estimate $\hat{\theta} = (\hat{\beta}, \hat{u}_i, \hat{\sigma}_e^2)$ from model (1) using the midpoints of the intervals as a substitute for the unknown y_{ij}.
2. Generate pseudo samples to reconstruct the distribution of the unknown y_{ij}:
   - **EM:** Estimate $E[I(A_{k-1} \leq y_{ij} \leq A_k) \times \pi(y_{ij}|x_{ij})]$, the expected value of a two sided truncated normal distributed variable as pseudo $\tilde{y}_{ij}$:
   $$\tilde{y}_{ij} = E[I(A_{k-1} \leq y_{ij} \leq A_k) \times \pi(y_{ij}|x_{ij})] = (x_{ij}^T\hat{\beta} + \hat{u}_j) + \hat{\sigma}_e \frac{\phi(Z_{k-1}) - \phi(Z_k)}{\Phi(Z_k) - \Phi(Z_{k-1})},$$
   obtaining $(\tilde{y}_{ij}, x_{ij})$ for $j = 1, \dots n_i$ and $i = 1, \dots, D$. The conditional variance is given by the variance of a two sided truncated normal distributed variable as
   $$Var(y_{ij}|x_{ij}, k_{ij}, u_i) = \hat{\sigma}_e^2 \underbrace{\left\{\left[\frac{Z_{k-1}\phi(Z_{k-1}) - Z_k\phi(Z_k)}{\Phi(Z_k) - \Phi(Z_{k-1})}\right] - \left[\frac{\phi(Z_{k-1}) - \phi(Z_{k-1})}{\Phi(Z_k) - \Phi(Z_{k-1})}\right]^2\right\}}_{:=s_{ij}}$$
   with $Z_k = (A_k - (x_{ij}^T\hat{\beta} + \hat{u}_i))/\hat{\sigma}_e$.
   - **SEM:** Sample from the conditional distribution $\pi(y_{ij}|x_{ij})$ by drawing randomly from $N(x_{ij}^T\hat{\beta} + \hat{u}_i, \hat{\sigma}_e^2)$ within the given interval $A_{k-1} \leq y_{ij} \leq, A_k$ obtaining $(\tilde{y}_{ij}, x_{ij})$ for $j = 1, \dots n_i$ and $i = 1, \dots, D$.
3. Re-estimate the vector $\hat{\theta}$ from model (1) by using the pseudo sample $(\tilde{y}_{ij}, x_{ij})$ obtained in step 2. The variance $\hat{\sigma}_e^2$ is given by:
   - **EM**:
   $$\hat{\sigma}_e^2 = \frac{\sum_{j=1}^{n_i}\sum_{i=1}^{D}(\tilde{y}_{ij} - (x_{ij}^T\hat{\beta} + \hat{u}_i))^2}{\sum_{j=1}^{n_i}\sum_{i=1}^{D}(1 - s_{ij})}$$
   - **SEM**:
   $$\hat{\sigma}_e^2 = \frac{\sum_{j=1}^{n_i}\sum_{i=1}^{D}(\tilde{y}_{ij} - (x_{ij}^T\hat{\beta} + \hat{u}_i))^2}{(N - 1)}$$
4. Number of iterations:
   - **EM:** Iterate steps 2.-3. until convergence.
   - **SEM:** Iterate steps 2.-3. $B + M$ times, with B burn-in iterations and M additional iterations.
5. Final parameter estimation:
   - **EM:** Obtain $\hat{\theta}$ from the last iteration step.
   - **SEM:** Discard the burn-in iterations and estimate $\hat{\theta}$ by averaging the obtained M estimates.

4. Simulation Setup

- Assume a finite population U of size $N = 40000$, partitioned into $D = 40$ regions $U_1, U_2, \dots, U_D$ of sizes $N_i = 1000$
- Let n_i be the sample size in region i with $n_i = 20$ and $\sum_{i=1}^{D} n_i = 800$
- 500 samples were randomly drawn from the following scenario:

$$y_{ij} = 1300 - 1x_{ij} + u_i + e_{ij}, \quad x_{ij} \sim GB2(5.0, 800, 0.4, 0.5),$$
$$u_i \overset{iid}{\sim} N(0, 1 \times 10^4), \quad e_{ij} \overset{iid}{\sim} N(0, 1.5 \times 10^5), \quad j = 1, \dots, n_i, \quad i = 1, \dots, D.$$

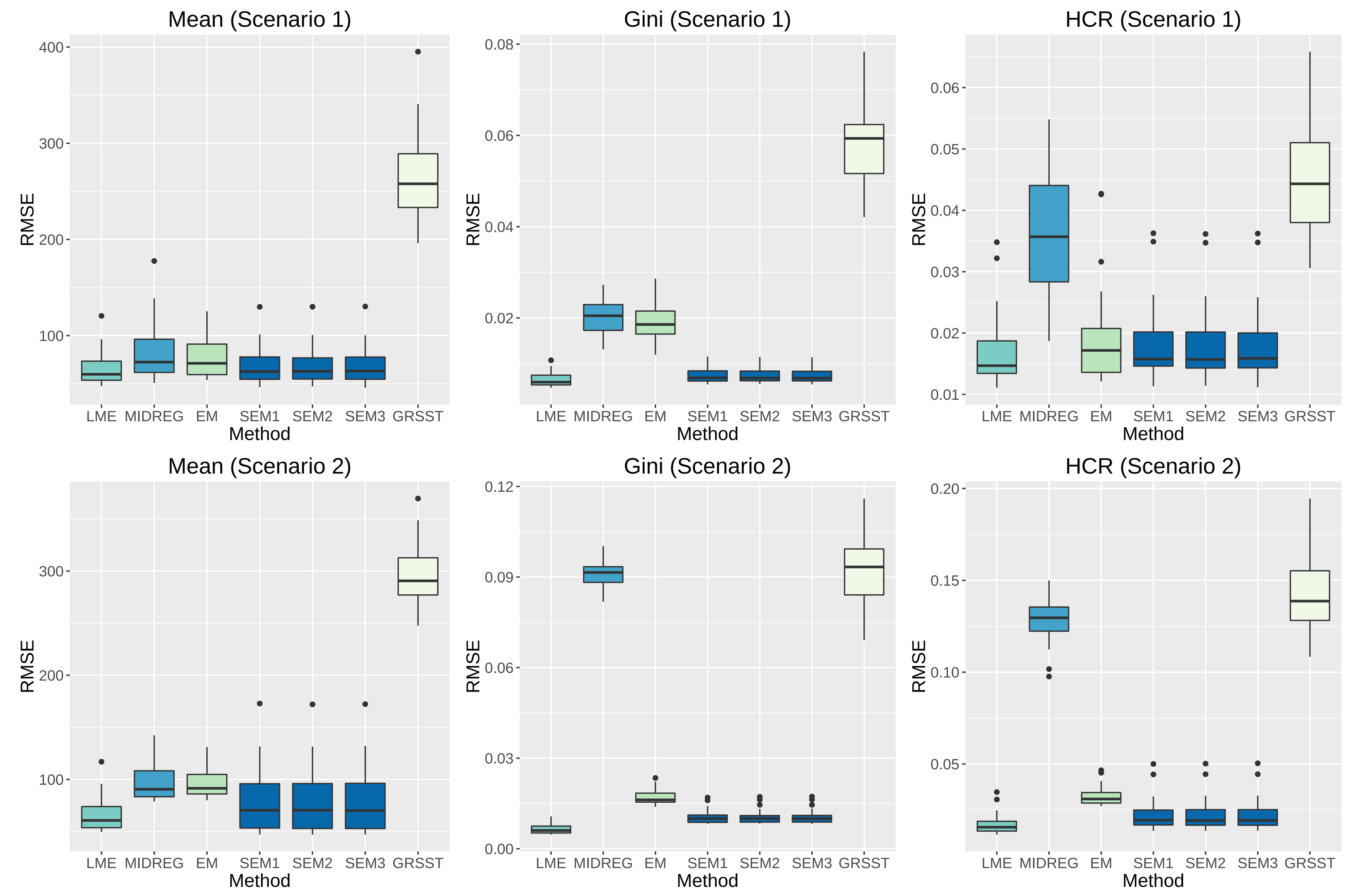
Interval censoring of Y_{ij}									
Intervals scenario 1	$[0, 500)$	$[500, 1000)$	$[1000, 1500)$	$[1500, 2000)$	$[2000, 2500)$	$[2500, 3000)$	$[3000, 4000)$	$[4000, 8000)$	$[8000, Inf)$
Frequencies scenario 1	34	705	4715	11908	11975	6191	3300	1050	122
Intervals scenario 2	$[0, 2000)$	$[2000, 4000)$	$[4000, 6000)$	$[6000, 8000)$	$[8000, Inf)$				
Frequencies scenario 2	17362	21466	893	157	122				

5. Simulation Results

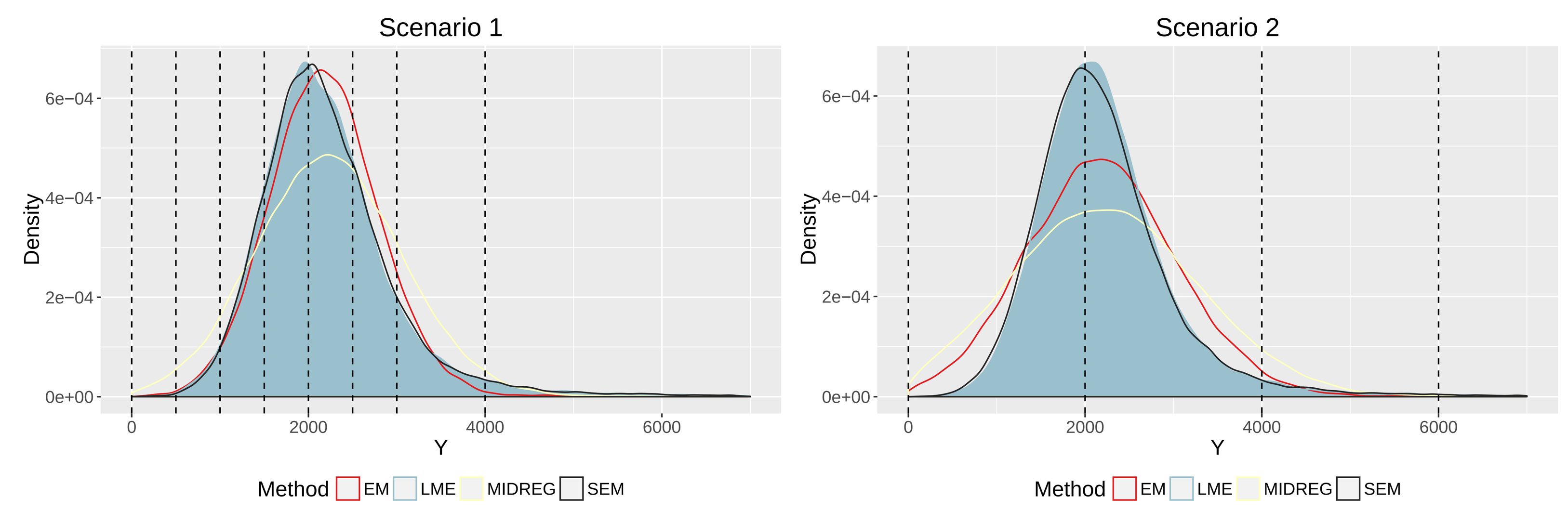
The following methods were applied for parameter estimation of model (1):

- LME - Estimate the model parameters with the true y_{ij} to evaluate the performance of the estimation methods relying on the interval censored y_{ij}.
- MIDREG - Estimation based on the interval midpoints as proxy for the unknown y_{ij}.
- EM - Estimation based on the generated pseudo $\tilde{y}_{ij}$.
- SEM - Estimation based on the drawn pseudo $\tilde{y}_{ij}$, with 100, 200 and 400 iterations (SEM1, SEM2 and SEM3).
- GRSST - Direct estimation using the GRSST estimator with $B = 5$ and $M = 20$ (Groß et al. 2016).

Performance of the EBPs



Density plots of $\hat{y}$ from a particular simulation run



6. Discussion and Outlook

- Simulation results show that the use of the SEM algorithm increases the accuracy, in terms of RMSE, in the EBPs.
- The amount of accuracy gained depends strongly on the number of intervals. However, the SEM algorithm still outperforms the other methods in the presence of many intervals.
- A high number of iterations (e.g. 200 or 400) does not improve the precision any further.
- **Further research:** How can possible violations of model assumptions be detected? How can transformations be applied for handling non-normally distributed error terms?

References

[1] Rao, J.N.K. & Molina, I. (2015), Small area estimation. John Wiley & Sons.
[2] Foster, J., Greer, J. & Thorbecke, E. (1984), A class of decomposable poverty measures. Econometrica, 52(3):761-766.
[3] Molina, I. & Rao, J.N.K. (2010), Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), 369-385.
[4] Caleux, G. & Dieboldt, J. (1985), The sem algorithm: a probalistic teacher algorithm derived from the em algorithm for the mixture problem. Computational Statistics Quarterly, 2:73-82.
[5] Stewart, M. B. (1983), On least square estimation when the dependent variable is grouped. Review of Economic Studies, 50(4):737-753.
[6] Dempster, A., Laird, N., & Rubin, D. (1977), Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1 - 38.
[7] Groß, M., Rendtel, U., Schmid, T., Schmon, S.& Tzavidis, N. (2016), Estimating the density of ethnic minorities and aged people in Berlin: multivariate kernel density estimation applied to sensitive georeferenced administrative data protected via measurement error. Journal of the Royal Statistical Society. Series A.

For further information

Paul Walter, Marcus Groß, Timo Schmid
paul.walter@fu-berlin.de, marcus.gross@fu-berlin.de
timo.schmid@fu-berlin.de
Department of Economics
Garystr. 21 D-14195 Berlin

Nikos Tzavidis
N.TZAVIDIS@soton.ac.uk
Department of Social Statistics & Demography
SO17 1BJ Southampton