

BAYESIAN METHODS FOR SMALL AREA ESTIMATION USING SPATIO-TEMPORAL MODELS

V. Gómez-Rubio^{1*}, N. Best¹ and S. Richardson¹

¹ Imperial College London, U.K. (*v.gomezrubio@imperial.ac.uk)

Introduction

National statistical bureaus often provide estimates of different small area indicators (e.g., unemployment, average income) at different geographical levels which have been computed using different methods. Spatio-temporal models, for example, take into account different geographic and temporal structures of the data in order to improve estimation. The purpose is to borrow strength from contiguous regions and observations close in time, because they will share similar patterns. By using different spatial and temporal structures it is possible to investigate the pattern of the data and to choose the best among different sets of models.

Bayesian spatio-temporal models

Using a direct estimator of the area mean (\hat{Y}_{it}) and its sampling variance ($\hat{\sigma}_{ij}^2$), both obtained from survey data. In addition, other auxiliary information can be available as area means (\bar{X}_{it}) from additional sources. We propose the following Bayesian model to improve Small Area Estimation of the area mean values in area i at time t :

$$\begin{aligned} \hat{Y}_{it} | \alpha, \beta, \sigma_{it}^2 &\sim N(\alpha + \bar{X}_{it}\beta + v_i + w_t + u_{it}, \sigma_{it}^2) \\ v_i | v_{-i}, \sigma_v^2 &\sim CAR(\sigma_v^2) & \sigma_v^2 &\sim Ga^{-1}(0.5, 0.0005) \\ w_t | w_{-t}, \sigma_w^2 &\sim CAR(\sigma_w^2) & \sigma_w^2 &\sim Ga^{-1}(0.5, 0.0005) \\ u_{it} | \sigma_u^2 &\sim N(0, \sigma_u^2) & \sigma_u^2 &\sim Ga^{-1}(0.5, 0.0005) \\ f(\alpha, \beta) &\propto 1 \end{aligned} \quad (1)$$

CAR stands for Conditional Autoregressive specification, which is used to model the spatial (v) and temporal (w) effects. This means the value of v_i (w_t) depends only on its neighbours v_{-i} (w_{-t}) and it is normally distributed with mean the average value on v_{-i} and variance $\sigma_v^2/|v_{-i}|$, where $|v_{-i}|$ denotes the number of neighbours.

Swedish LOUISE Register

The Swedish population register contains different socioeconomic variables at the individual and household level. Our data includes data from year 1992 to 1999 at the municipality level.

Equivalised Income per household

We have used different SAE methods to provide an estimate of the average *Equivalised Income per household* from 1992 until 1999 using the following covariates: # persons in household, # pers. employed in household, and education, age and sex of head of household. The covariates are available as area means or area proportions in the case of binary variables such as education and sex.

Survey design

In the analysis carried out in EURAREA (2004), the households have been sampled without replacement at the municipality level, taking a sample size of 1% of the total number of households in each municipality. We have reduced the sample size to 1/1000 the total number of households in order to compare the area level models to the combined models.

Results (Area Level Model)

Covariate	R. Effects (u)	Spatial ($v + u$)	Temporal ($w + u$)	Full ($v + w + u$)	Direct est.
Intercep	1211 (9.764)	1214 (7.527)	1211 (9.689)	1215 (7.289)	
# persons.	-34.85 (7.163)	-33.01 (8.840)	-35.02 (9.259)	-27.9 (12.640)	
# empl.	48.05 (5.082)	38.81 (7.365)	56.08 (6.299)	54.57 (10.030)	
ed. head	48.0 (6.274)	46.6 (7.122)	41.69 (6.518)	34.74 (9.011)	
age head	31.47 (4.724)	35.66 (4.835)	23.74 (5.672)	18.16 (7.159)	
sex head	-18.33 (6.662)	-12.96 (9.078)	-21.23 (6.893)	-19.57 (9.881)	
AEMSE	12044.1	10775.2	11412.2	10179.5	56260.4
AREMSE	7.480	6.834	6.985	6.361	41.511
DIC (p_D)	30850 (682)	30480 (663)	30770 (658.7)	30410 (641.7)	

The quality of the models has been assessed using the true values of the area means, which are known in this case. The Average Empirical MSE (AEMSE) and the Average Relative Empirical MSE (AREMSE) have been used to assess how close are the estimated values to the true ones, whilst the Deviance Information Criterion DIC have been used to choose the best model.

Acknowledgements

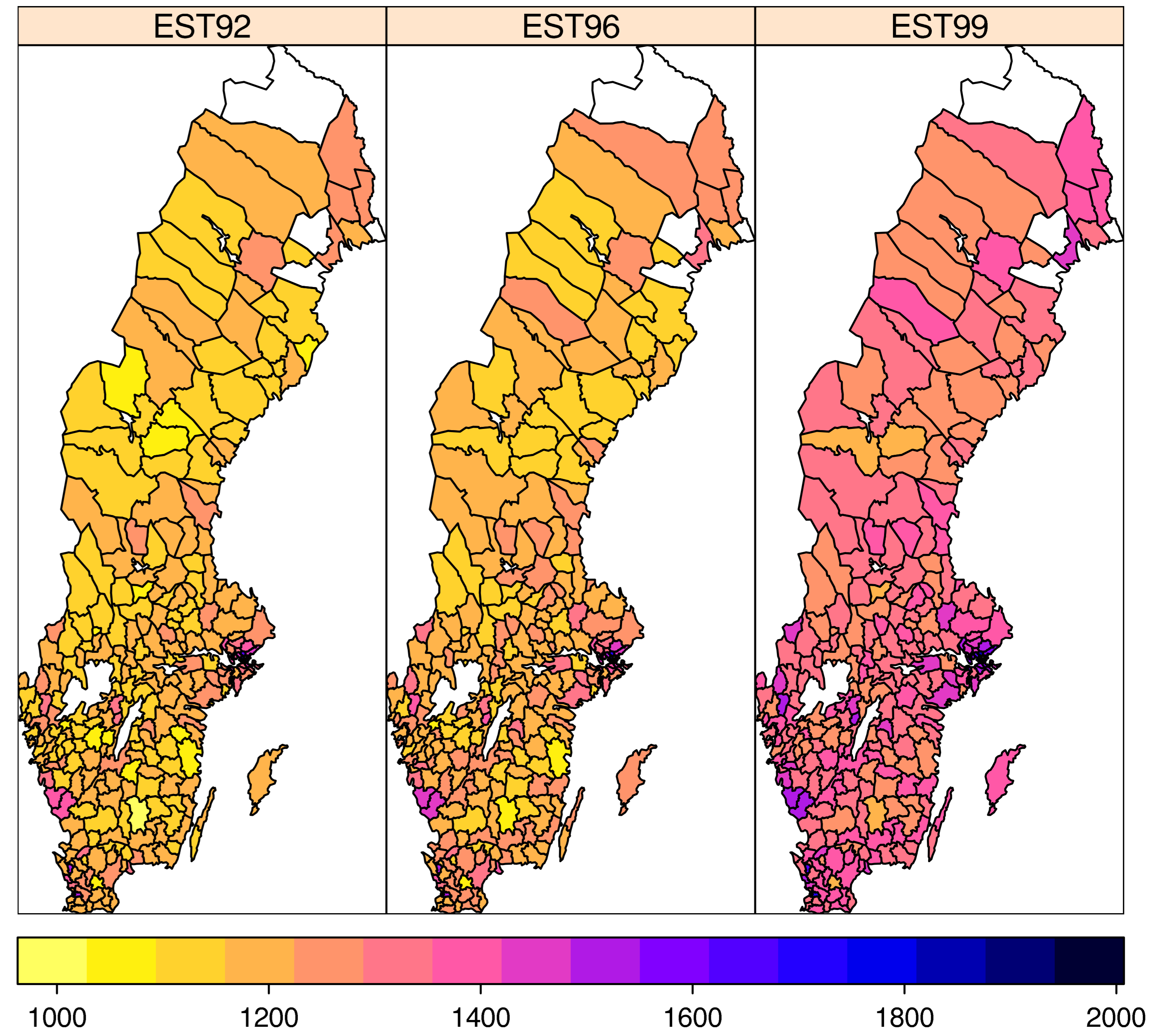
This work has been supported by BIAS Project, funded by Economic and Social Research Council, UK.

References

- * EURAREA Consortium (2004). *Project Reference Volume*. Published by the EURAREA Consortium. URL: <http://www.statistics.gov.uk/eurarea/>
- * V Gómez-Rubio, N Best and S Richardson. *A comparison of different methods for small area estimation*. In preparation.
- * L Knorr-Held (2000). *Bayesian modeling of inseparable space-time variation in disease risk*. *Statistics in Medicine* **19**: 2555-67.
- * JNK Rao (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- * S Richardson, JJ Abellán and N Best (2006). *Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK)*. *Stat. Methods Med. Res.* **15**: 385-407

Estimated average equivalised income per household

The following figure shows the estimated values of the average income using the Full model, which is the one with the lowest AEMSE, AREMSE and DIC.



Combining individual and aggregated data

In order to provide more accurate estimates of the coefficients of the covariates in the model, we propose the use of the following model using *individual* data (y_{ijt}):

$$\begin{aligned} y_{ijt} | \alpha, \beta, \sigma_{it}^2 &\sim N(\alpha + x_{ijt}\beta + v_i + w_t + u_{it}, \sigma_{it}^2) \\ \hat{Y}_{it} &= \hat{\alpha} + \bar{X}_{ij}\hat{\beta} + \hat{v}_i + \hat{w}_t + \hat{u}_{it} \\ v_i | v_{-i}, \sigma_v^2 &\sim CAR(\sigma_v^2) & \sigma_v^2 &\sim Ga^{-1}(0.5, 0.0005) \\ w_t | w_{-t}, \sigma_w^2 &\sim CAR(\sigma_w^2) & \sigma_w^2 &\sim Ga^{-1}(0.5, 0.0005) \\ u_{it} | \sigma_u^2 &\sim N(0, \sigma_u^2) & \sigma_u^2 &\sim Ga^{-1}(0.5, 0.0005) \\ f(\alpha, \beta) &\propto 1 \\ \log(\sigma_{it}^2) | \sigma^2 &\sim N(0, \sigma^2) & \sigma^2 &\sim Ga^{-1}(a, b) \end{aligned}$$

Survey design

We have used the same sample as in the area level analysis. This assures that results are comparable (in terms of AEMSE and AREMSE) and this sample size is computationally feasible when estimating the models using individual information.

Results (Combined Level Model)

Covariate	R. Effects (u)	Spatial ($v + u$)	Temporal ($w + u$)	Full ($v + w + u$)	Direct est.
Intercep	1223 (3.530)	1219 (3.302)	1224 (3.476)	1220 (3.223)	
# persons.	-114.9 (3.548)	-114.3 (3.515)	-114.5 (3.502)	-114.3 (3.600)	
# empl.	224.6 (3.466)	224.6 (3.398)	224.8 (3.421)	225.4 (3.401)	
ed. head	78.33 (2.746)	75.51 (2.785)	78.31 (2.793)	75.03 (2.758)	
age head	174.5 (2.729)	174.9 (2.754)	173.9 (2.669)	174.3 (2.675)	
sex head	53.47 (2.619)	53.98 (2.575)	53.25 (2.597)	53.92 (2.560)	
AEMSE	14083.6	10057	12783.2	8296.2	56260.4
AREMSE	8.291	5.979	7.390	4.776	41.511
DIC (p_D)	426200 (2842)	426100 (2719)	426200 (2815)	426000 (2658)	

Discussion

By combining individual information from direct surveys and aggregated data from other sources it is possible to improve the quality of small area indicators. In our case, estimating the parameters and effects of the model using individual data and then using aggregated information to provide estimates of the average income per household shows clear better results than the area level models. We have found that the three criteria based on the AEMSE, AREMSE and DIC agree on the choice of the *best* model.

Further work

In a purely spatial setting, Gómez-Rubio et al. (in prep.) show how Bayesian models provide better estimates than a wide range of other SAE methods (EURAREA, 2004; Rao, 2003). These models will be extended to the study of proportions in order to estimate, for example, the rate of unemployment. We also expect to assess the impact of the sample size on the quality of the estimates. Finally, we hope to extend this models to the case in which the sample is restricted to a few areas and not all the municipalities in Sweden.