# Spatial robust small area estimation applied on business data

Münnich, Ralf
University of Trier

Schmid, Timo
University of Trier

Zimmermann, Thomas
University of Trier

**Abstract**

Small area methods are now widely used to provide reliable information about quantities of interest at a disaggregated level, especially when small sample sizes may lead to inaccurate estimates using classical procedures. Their applicability in the case of business data, however, is not straightforward, since business data are often characterized by outliers violating the normality assumptions present in the standard models. One way to overcome these peculiarities is the application of robust methods. Therefore, a short overview of the recently used robust small area methods is given and a spatial extension of the robust EBLUP estimator and its mean squared error estimation that we use within the BLUE-ETS project, is presented. Moreover, some of the most efficient sampling techniques may lead to large variations of the design weights which cause problems for statistical modelling. Two approaches of how to deal with this issue are design-consistent pseudo-EBLUP estimators and box-constraint optimisation techniques restricting the variation of the survey weights. We investigate the performance of robust and design-consistent model-based estimators by means of a design-based simulation study.

## 1 Introduction

Economic and political decision processes are increasingly based on specific indicators and other statistical information. Nowadays the necessity of developing regional indicator values or disaggregated values is evident in order to allow for regional or group-specific comparisons. Surveys which shall deliver the necessary information for these indicators, however, are generally constructed for larger areas, e.g. countries or NUTS2 domains. Hence, sample information on levels such as NUTS3 and below is rarely available so that classical estimates lead to high variances of the estimates.

Applying small area estimation methods may lead to highly improved accuracy of the estimates of interest. Especially in business statistics outliers in connection with small sample sizes lead to severe problems while applying standard models which are based on normal assumptions due to the high sensitivity of the model estimates towards these influential units. One way to overcome these peculiarities is the application of robust methods. Two such robust small area methods are the robust EBLUP estimator and the robust M-quantile approach. But in business data, spatial dependencies often occur, so there is a need to enhance these models, which is already done for the M-quantile approach. In this talk we present an overview of the recently used robust small area methods and present a spatial extension of the robust EBLUP estimator and its MSE estimation.

Another problem of many model-based small area estimators is their validity with respect to the underlying sampling design. Most model-based estimators are not design-unbiased unless the design is self-weighting within the areas. Self-weighting sampling designs, however, might not be efficient in the presence of influential outliers as it is often the case in business statistics. This is due to the fact that designs attempting to minimise the sampling variance may be characterized by highly different survey weights which pose problems for statistical modelling.

We compare the performance of spatial robust estimators with other small area estimators by means of a design-based simulation study. Employing a design-based methodology enables us to analyse the strengths and weaknesses of the estimators within a realistic, real-life setting. Our design-based simulation study is carried out on synthetic population data based on Italian business data.

# 2 Spatial Robust Small Area Methods

We start from the general linear mixed model, which is given by

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{v} + \boldsymbol{e}, \tag{1}$$

where the vectors $\boldsymbol{v}$ and $\boldsymbol{e}$ are independently normally distributed with means 0 and covariance matrices $G$ and $R$, respectively, depending on a vector of variance parameters $\boldsymbol{\theta}$. $X$ is the known matrix with auxiliary variables and $Z$ denotes the design matrix for the random effects $\boldsymbol{v}$. Furthermore, the variance-covariance matrix of the variable of interest $\boldsymbol{y}$ is obtained via $V = R + ZGZ^T$.
Following Rao (2003) or Jiang and Lahiri (2006), the best linear unbiased predictor (EBLUP) for a linear combination $\mu$ of the regression coefficient $\boldsymbol{\beta}$ and the random effect $\boldsymbol{v}$ in model (1) is given by

$$\hat{\mu}(\hat{\boldsymbol{\theta}}) = \boldsymbol{l^T}\hat{\boldsymbol{\beta}} + \boldsymbol{m^T}\hat{\boldsymbol{v}}, \tag{2}$$

where $\boldsymbol{l}$ and $\boldsymbol{m}$ are specific vectors and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$.
From a robust viewpoint, the EBLUP can be very sensitive to outliers or skewed distributions. Therefore, Sinha and Rao (2009) substituted the estimators $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{v}}$ in (2) by robust alternatives $\hat{\boldsymbol{\beta}}^{\psi}$, $\hat{\boldsymbol{\theta}}^{\psi}$ and $\hat{\boldsymbol{v}}^{\psi}$, leading to the robust EBLUP (REBLUP) of $\mu$

$$\hat{\mu}_{\psi}(\hat{\boldsymbol{\theta}}^{\psi}) = \boldsymbol{l^T}\hat{\boldsymbol{\beta}}^{\psi} + \boldsymbol{m^T}\hat{\boldsymbol{v}}^{\psi}. \tag{3}$$

Due to the complex form of the REBLUP estimators and the lack of knowledge of the underlying distributions of the random effects $\boldsymbol{v}$ and the error term $\boldsymbol{e}$, the corresponding estimators of the MSE cannot be obtained in any closed form. Thus, Sinha and Rao (2009) suggest a parametric bootstrap based on the method of Hall and Maiti (2006).
An alternative approach to outlier robust small area estimation is the M-quantile regression-based method introduced by Chambers and Tzavidis (2006). The M-quantile regression is based on quantile regression without specifying the random effects to explain between small area variation unlike the general linear mixed model (1). Detailed information can be also found in Tzavidis et al. (2010) and Salvati et al. (2011). Furthermore, we define a spatial extension of the robust EBLUP estimator given by (3). Therefore, we have to extend the general linear mixed model (1) to allow for spatial correlated area effects $\boldsymbol{v}$. We focus in this paper similar to Pratesi and Salvati (2009) and Salvati et al. (2011)

on the simultaneously autoregressive models (SAR). The spatial correlated random effect with covariance matrix

$$G = \sigma_u^2 \left( (I - pW)(I - pW^T) \right)^{-1} , \tag{4}$$

where the parameter $p$ denotes the spatial autoregressive parameter and W is the proximity matrix, can be directly incorporated into model (1). The matrix W describes the neighborhood structure between the small area and $p$ defines the strength of the spatial dependencies among the random effects.

Similar to SINHA and RAO (2009), we maximize the density of $\boldsymbol{y}$, but now with respect to $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $p$ leading to spatial robust ML-estimators of $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $p$ by solving the spatial robust ML-equations

$$\begin{aligned} X^T V^{-1} U^{\frac{1}{2}} \psi(r) &= 0 \\ \psi^T(r) U^{\frac{1}{2}} V^{-1} \frac{\partial V}{\partial \theta_l} V^{-1} U^{\frac{1}{2}} \psi(r) - \operatorname{tr}(V^{-1} \frac{\partial V}{\partial \theta_l} K) &= 0 \\ \psi^T(r) U^{\frac{1}{2}} V^{-1} \frac{\partial V}{\partial p} V^{-1} U^{\frac{1}{2}} \psi(r) - \operatorname{tr}(V^{-1} \frac{\partial V}{\partial p} K) &= 0, \end{aligned} \tag{5}$$

where $r = U^{-\frac{1}{2}}(y - X\boldsymbol{\beta})$ and $U$ is a diagonal matrix with diagonal elements equal to the diagonal elements of the matrix $V$. $K$ is also a diagonal matrix and $\psi$ is an influence function, e.g. the Huber function. Afterwards, we use $\hat{\boldsymbol{\beta}}^{\psi,sp}$, $\hat{\boldsymbol{\theta}}^{\psi,sp}$ and $\hat{p}$ for the estimation of the spatial robust area effects $\boldsymbol{v}$ by plugging in the spatial robust estimates in Fellner's equation (FELLNER, 1986) and solve this equation with an iterative algorithm and set $\hat{\boldsymbol{v}}^{\psi,sp}$ as our spatial robust area effects estimator. The spatial robust estimates $\hat{\boldsymbol{\beta}}^{\psi,sp}$, $\hat{\boldsymbol{\theta}}^{\psi,sp}$, $\hat{\boldsymbol{v}}^{\psi,sp}$ and $\hat{p}$ are then used to estimate (2), referred as the spatial robust EBLUP (SREBLUP) of $\mu$, as

$$\hat{\mu}_{\psi,sp}(\hat{\boldsymbol{\theta}}^{\psi,sp}) = \boldsymbol{l^T} \hat{\boldsymbol{\beta}}^{\psi,sp} + \boldsymbol{m^T} \hat{\boldsymbol{v}}^{\psi,sp}. \tag{6}$$

Further information can be found in SCHMID and MÜNNICH (2011b).

For the MSE estimation of the SREBLUP (6) of $\mu$, we adopt a parametric bootstrap method based on the spatial robust estimators following Rao and Sinha (2009). Detailed derivations of this bootstrap method is available in SCHMID and MÜNNICH (2011a).

# 3 Design issues

The sampling techniques used to produce business statistics differ much from those employed in other parts of official statistics. This is mainly due to the enormous impact of a small number of observations on the statistics of interest, e.g. total turnovers. Attempts to minimise the sampling variance could lead to large variations of the survey weights. The use of highly different survey weights may create major problems in statistical modelling as pointed out by GELMAN (2007). The literature on how to deal with this issue has evolved along two lines: to incorporate survey weights in the model or to construct designs which avoid large discrepancies between the design weights. In the field of small area estimation the former approach of including survey weights in the model has been studied extensively, see YOU and RAO (2002). A simulation study conducted by MÜNNICH and BURGARD (2011) showed that using design-consistent model-based estimators reduces the negative impact of highly different weights. Another approach to overcome the problem is the box-constrained optimisation proposed by Gabler et al. (2010) for the case of stratified sampling with optimal allocation, allowing differences between the weights only within certain boundaries. A comparison of the performance of different algorithms implementing these constraints is given by Münnich et al. (2011).

# 4 Simulation study

The aim of our study is to analyse the influence of selected sampling designs on small area estimators. We compare the performance of design-based estimators, taking the sampling design into account, and model-based estimators, which do not always consider the sampling design. Our simulation study is based on business data from the BLUE-ETS project that we took as a starting point for creating a larger artificial population. This population is taken from the Italian business register, whose entries have been extended to cover the cases of non-sampled businesses. The sampling designs used in our study are simple random sampling without replacement (SRS), stratified sampling with proportional allocation (PROP), stratified sampling with equal allocation (EQ), stratified sampling with optimal optimal allocation (OPT), stratified sampling with optimal allocation restricted by box constraints (BOX) and sampling with probability proportional to size ($\pi$PS). We use these designs to compare robust estimators as introduced in section 2 with the design-consistent estimators mentioned in section 3.

# Acknowledgements

The research was conducted within the BLUE-ETS research project which is funded by the European Commission within the 7th Framework Programme. For more information on the project, we refer to the project page http://www.blue-ets.eu.

# References

**Chambers, R.** and **Tzavidis, N.** (**2006**): *M-Quantile Models for Small Area Estimation.* Biometrika, 93 (2), pp. 255–268.

**Fellner, W. H.** (**1986**): *Robust Estimation of Variance Components.* Technometrics, 28 (1), pp. 51–60.

**Gabler, S.**, **Ganninger, M.** and **Münnich, R.** (**2010**): *Optimal allocation of the sample size to strata under box constraints.* Metrika.
URL http://dx.doi.org/10.1007/s00184-010-0319-3

**Gelman, A.** (**2007**): *Struggles with Survey Weighting and Regression Modeling.* Statistical Science, 22, pp. 153–164.

**Hall, P.** and **Maiti, T.** (**2006**): *On Parametric Bootstrap Methods for Small Area Prediction.* Journal of the Royal Statistical Society, 68 (2), pp. 221–238.

**Jiang, J.** and **Lahiri, P.** (**2006**): *Mixed Model Prediction and Small Area Estimation.* Test, 15 (1), pp. 1–96.

**Münnich, R.** and **Burgard, J. P.** (**2011**): *On the Influence of Sampling Design on Small Area Estimates.* Journal of the Indian Society of Agricultural Statistics, to appear.

**Münnich, R.**, **Sachs, E.** and **Wagner, M.** (**2011**): *A comparison of box constraint optimization algorithms for optimal allocation in stratified random sampling.* Submitted.

**Pratesi, M.** and **Salvati, N.** (**2009**): *Small area estimation in the presence of correlated random area effects.* Journal of Official Statistics, 25, pp. 37–53.

**Rao, J. N. K.** (**2003**): Small Area Estimation. New York: John Wiley and Sons.

**Salvati, N.**, **Tzavidis, N.**, **Pratesi, M.** and **Chambers, R.** (**2011**): *Small Area Estimation Via M-quantile Geographically Weighted Regression.* Forthcoming in TEST.

**Schmid, T.** and **Münnich, R.** (**2011**a): *Parametric bootstrap mean squared error estimation for the spatial robust EBLUP.* Proceedings of the Second ITACOSM Conference, Survey Research Methods and Applications, pp. 211–214.

**Schmid, T.** and **Münnich, R.** (**2011**b): *Spatial robust small area estimation.* Submitted.

**Sinha, S. K.** and **Rao, J. N. K.** (**2009**): *Robust Small Area Estimation.* The Canadian Journal of Statistics, 37 (3), pp. 381–399.

**Tzavidis, N.**, **Marchetti, S.** and **Chambers, R.** (**2010**): *Robust estimation of small area means and quantiles.* Australian and New Zealand Journal of Statistics, 52 (2), pp. 167–186.

**You, Y.** and **Rao, J. N. K.** (**2002**): *A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights.* The Canadian Journal of Statistics, 30, pp. 431–439.