Small area estimation for population counts in the German Census 2011

Ralf Münnich* Jan Pablo Burgard* Martin Vogt*

Abstract
In modern survey applications, National Statistical Institutes have always the pressure to reduce costs. This task plays an important role in the next European Census round 2010/11, where some countries try to employ a register-based Census which may help to reduce costs by far. Small area estimation methods are expected to allow for high-level results for small areas and domains under the given budget constraints. The present article focuses on the estimation of higher educated people as example for a variable which is not overserved in the register. Based on the classical basic unit and area-level models, binomial mixed-models will be elaborated. Finally, spatial small area models will be assessed. The entire study will be accompanied by a Monte-Carlo study which will foster comparisons of all models within a realistic framework.

Key Words: Census, small area estimation, binomial mixed-models, spatial modeling, census design

1. Introduction

In the next European Census round 2010/11 in some European countries a new approach for conducting Censuses will be adopted. In order to reduce costs and the response burden for the citizens countries like Germany and Switzerland will apply a register-based census. In addition to evaluating the population register a sample will be drawn.

In the German case, the sample is used for two different goals. On the one hand, register errors, i.e. over- and undercounts, are estimated and further used for correcting the register counts to population counts. On the other hand a large set of variables like education and employment variables which are not covered by register data will be directly estimated from the sample while using the registers as auxiliary information. The latter is similar to the American Census long-form questionnaire (cf. `http://2010.census.gov/2010census/pdf/2010ACSnotebook.pdf`). Further details on the German Census 2011 can be found at `http://www.zensus2011.de/Statistik-Portal/en/Zensus/`.

A major problem when conducting a Census based on samples arises in gaining accurate information on smaller regions such as communities or even subclasses of the population by content. Given a sample size of approximately 10% of the population may lead to relatively small sample sizes in small sub-groups which leads to unacceptably high standard errors when applying classical estimators such as the Horvitz-Thompson estimator or the generalized regression estimator (GREG). One strategy to overcome these estimation problems is to apply small area estimators which may help to gain efficiency in small sub-groups by borrowing strength from the information on the entire sample (cf. Rao, 2003, or Jiang and Lahiri, 2006). Outcomes of a comparative simulation study can be drawn from Magg, Münnich

*University of Trier, Faculty IV, Economics Economics and Social Statistics Department, D-54296 Trier, Germany, e-mail: {muennich,jpburgard,vogt4502}@uni-trier.de

and Schäfer (2006), Münnich and Magg (2006) or Münnich, Gabler and Ganninger (2007).

The above studies show that small area estimators may help to reduce the relative root mean square error of census estimates considerably. However, one can also see that some area estimates suffer from biases. This is mainly due to applying standard area and unit-level small area estimators (cf. EURAREA standard estimators in Münnich et al., 2004) to data which may violate some assumptions such as normality of error terms.

In the present study we compare different small area estimators based on normal and binomial mixed-models. The study is conducted using different size levels of area information in order to evaluate the effect of area size variations. Further, the integration of spatial information was applied to the standard Fay-Herriot estimator.

Tthe following Section presents the different models of interest. After giving an overview of the German Census models for small area statistics including normal, binomial, and Poisson models are sketched. Finally, a spatial extension of the Fay-Herriot model is given. Chapter three presents the design and outcome of the Monte-Carlo study. Finally, the main results are summarized.

2. The models of interest

2.1 The German Census

Let Y denote the variable of interest which in our study is the higher education, a variable which is not presented in the population register. X is a set of auxiliary variables that can be drawn from the population register such that prediction methods can be applied on the full set of register information.

In order to adequately apply the prediction approach, the register errors had to be incorporated carefully in the setup of the models. The sampling units are addresses. Hence, the dependent variable consists of count data and here is the amount of people within an address having an higher education level. The independent variables for the same address are taken out of the register. Applying normal model a straight forward approach can be applied using the counts as continuous variables.

In binomial models this straight forward routine may lead to problems in cases where more people with higher education are living in an address than the total amount of registered people. Standard naive modeling would yield $n_a > N_a$ where N_a would have been determined from the register. This, however, did not occur in the study. Ignoring these addresses which might appear only in small addresses would have lead to slightly biased estimates.

The alternative is to model directly on the sample and using the sample estimates on the register. Here one would expect a frame error due to the differences in the amounts of over- and undercounts. Corrections for frame errors have not been considered so far and will be implemented in future studies. This convenient case may not lead to the assumption that these errors may be neglected. In cases where the variable of interest shows larger amounts of outcomes one would expect considerable biases. Deeper investigations of this effect, however, were out of the scope of this study.

2.2 Normal and binomial mixed-models

As the German universe is of non-negligible size the choice of modeling had to consider the computational effort whilst maintaining the accuracy of the estimates. The predictive models which was applied in this paper uses three parts, estimation, prediction, and aggregation. First, estimates are calculated to link the dependent variable to the covariates. Then a prediction was performed while applying the model to the register covariates. These prediction were finally aggregated in the areas of interests.

The classical standard unit and area-level estimators generally are built upon the assumption of normal distribution of the model. In the studies cited above they proofed to perform reasonably good also for count data. In this study we aim improving the accuracy of the estimates using other more sensible distributional assumptions like the Poisson or the binomial. The following models were used:

$$\begin{aligned} y_{ij} &= x_{ij}\beta + v_i + e_{ij} \\ v_i &\sim \mathrm{N}(0, \sigma_v^2) \\ e_{ij} &\sim \mathrm{N}(0, \sigma_e^2) \end{aligned}$$

as the normal model and

$$\begin{aligned} y_{ij} &\sim \mathrm{Bin}(n_{ij}, \rho_{ij}) \\ \mathrm{logit}(\rho_{ij}) &= x_{ij}\beta + v_i + e_{ij} \\ v_i &\sim \mathrm{N}(0, \sigma_v^2) \\ e_{ij} &\sim \mathrm{N}(0, \sigma_e^2) \end{aligned}$$

as a binomial model. The Poisson model

$$\begin{aligned} y_{ij} &\sim \mathrm{Po}(_{ij}) \\ \log(\rho_{ij}) &= X_{ij}\beta + v_j + \varepsilon_{ij} \\ v_i &\sim \mathrm{N}(0, \sigma_v^2) \\ \varepsilon_{ij} &\sim \mathrm{N}(0, \sigma_\varepsilon^2) \end{aligned}$$

generally showed peculiarities in the estimates which may be due to the shape of the count data which showed not to follow a Poisson distribution.

2.3 Spatial models

In general, in household datasets spatial patters can be observed. To allow for spatial correlations in small area models, the independence assumption of the random effects v_i is substituted by the popular conditional autoregressive (CAR) structure in the version proposed by Banerjee, Carlin, and Gelfand (2004):

$$v_i|\{v_l : l \neq i\} \sim N\left(p \sum_{l \in A_i} \frac{Q_{i,l}}{\sum_{j=1}^{n} Q_{i,j}} v_l \ , \ \frac{\sigma_v^2}{\sum_{j=1}^{n} Q_{i,j}} \right) \ ,$$

where A_i denotes a set of neighboring areas of the i-th area, $\{Q_{i,l}\}$ are known constants satisfying $Q_{il} = Q_{li}$ and p , σ_v^2 is the unknown parameter vector. Sun, Tsutakawa, and He (2001) proved the propriety for the general linear mixed model which includes the above model. In this application a nearest neighbor structure is assumed and σ_e^2 is assumed to be known. The model has been implemented using WinBUGS from R.

3. Simulation Study

3.1 Design of the study

The Monte-Carlo simulation study was performed on a SUN cluster at the University of Trier. The core tools have been programmed under R using the package nlme. The setup of the study is a standard Monte-Carl set-up for design-based simulations in survey statistics (cf. Münnich, Gabler, and Ganninger, 2007).

The data used in the simulations for this paper are partially synthetic data. The core of the dataset is a subset of the German population register. It contains 24 mostly fully represented districts from 4 federal states. The size of the population in total is approximately 6.2 mio. which are distributed in nearly 1 mio. addresses. 284 communities can be found within the districts. The variable provided by the registers are gender, age, nationality, marital status, time of residence at the actual place, and residence status. As mentioned before the target variable is the number of higher educated people which can be translated to ISCED level 5.

The areas of interest for which estimates are to be determined were split into two different tasks. One task is related to community level information. The communities were selected such that a wide variety of community sizes were guaranteed with some very small communities and some large towns. The second task was elaborated to consider more homogeneous sizes of areas, the so-called sampling points. The sampling points were built in order to better define areas in which separate samples can be conducted.

Table 1: Definition of the sampling points

SMP	Sampling Point Type	Size
0	Fraction of a community	at least 200,000
1	Community with at least 10,000 inhabitants	at least 10,000
2	Aggregations of communities within a union community which have together at least 10,000 inhabitants	at least 10,000
3	Aggregations of communities that were not include in one of the above types per district	at least 0

The colors in Table 1 are corresponding to the color use in the following graphs. The distribution of the target variable in the sampling points and communities is given in Figure 1. On average approximately 10-15% higher educated people can be found in the given sampling points and communities.

In addition to the register variables further variables were added synthetically, e.g. educational achievement, training qualification, and employment status. These synthetic variables were generated considering the structures of the German microcensus sample. A detailed description of this procedure can be drawn from Kolb (2008). In order to cope with the register error problem a vector of over- and undercounts was generated as a 4-level generalized mixed-effects model according to Burgard (2009). Simulation studies showed that the choice of the over and undercount model may have a strong impact on small area estimates (see Münnich et al., 2008). The latter report includes details on other over- and undercount models and the general design of the study.

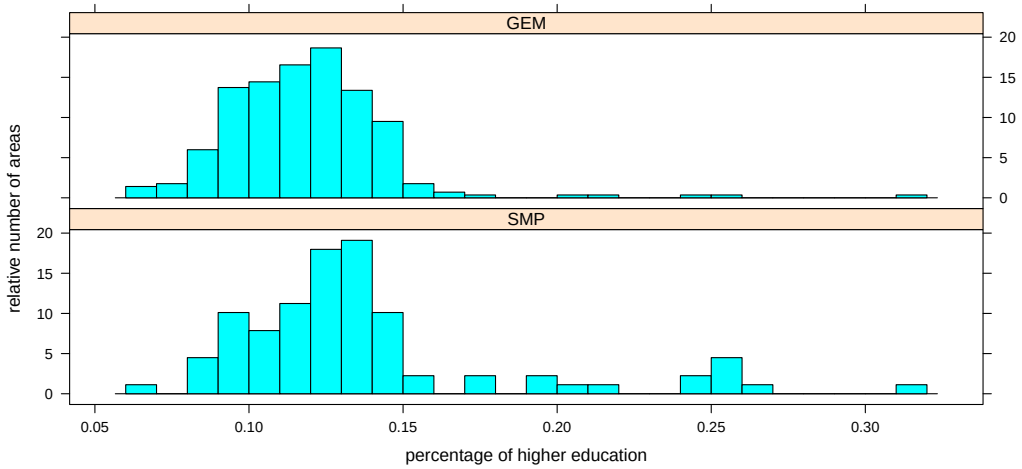

Figure 1: Percentage of higher education by areas with respect to sampling points and communities

3.2 Sampling design

One major problem applying optimal sampling strategies in the Census context is the fact that accurate estimates have to be produced on different levels of aggregation. In order to allow for optimizing the design on all communities or sampling points in parallel, Münnich, Gabler, and Ganninger (2008) introduced a multidimensional optimal allocation problem on sampling point level as a multicriteria decision problem minimizing the maximal relative mean squared error of the sampling point estimates using the combined regression estimator as a benchmark estimator or any p-norm of the relativ mean squared errors of the sampling point estimates under given constrains, e.g. budget constraints. Details can be found in Münnich, Gabler, and Ganninger (2009).

Within the given study three different sampling designs were used, simple random sampling (SRS) within sampling points, stratified random sampling (StrRS) within sampling points with address size classes as strata and optimal allocation. In all cases 10% of the population was sampled. For SRS, the optimality was achieved using a standard 2-norm with respect to the population shares and is denoted by SRS_Opt. StrRS_Opt is a straight forward extension to stratified random sampling. The main focus of accuracy goals for the German Census estimates is laid on large communities with 10,000 and more inhabitants or on district estimates. Hence optimality may be achieved while reducing the samples sizes in small communities down to 2 in order to still allow for variance estimation. This extreme design is denoted by StrRS_OptSpec and a variant of the previous stratified design.

3.3 Estimates on sampling points and communities

One target of the study is to compare the estimates on different levels of area sizes, namely sampling points and communities. The main focus within the following graphs is to evaluate the different size classes of the areas of interest. Table 2 presents the size classes of the communities using the colors in the following graphs.

Table 2: Size classes of the communities in the dataset

COM	Community Type
0	Communities with less than 1,001 inhabitants
1	Communities with more than 1,000 and less than 10,001 inhabitants
2	Communities with more than 10,000 and less than 100,001 inhabitants
3	Communities with more than 100,000 and less than 1,000,001 inhabitants
4	Communities with more than 1,000,000 inhabitants

As a general result the sampling point estimates seem to produce a larger dispersion of the estimates while producing a smaller range of biases in comparison to the community estimates. This seems to result from the fact that within sampling points in many cases several heterogeneous communities were aggregated which in the case of non-sampled areas may lead to a higher variability of the estimates. The sampling point, however, generally yield smaller relative root mean squared errors which is due to the more homogeneous sizes of the areas.

The following two graphs present the relative root mean squared error (RRMSE), the relative bias (RBias), and the relative dispersion (RDisp) of the area-specific estimates in columns. The rows contain the three given designs. The four estimators are based on either sampling point estimates (SMP) or community estimates (COM), the normal (Nor) or binomial model (Bin), as well as the two estimation cases. S indicates the naiv estimates and AS the alternative estimates on sample based information. The kernel density curve depicts the distribution of the measures of all areas. The vertical lines give precise measures split into the size classes defined before.

Figure 2 shows that the RRMSE of the binomial models is considerably better than for the normal model for all three designs. This results from better biases and dispersions. The effect of taking either AS or S has a high impact mainly on the bias of the normal model. The difference is by far smaller in the binomial model. In this case also little effects can be seen when comparing SRS and StrRS. The Spec designs, however, show a severe impact on all measures and models in sampling point types two and three as expected due to the much smaller sample sizes in these sampling points. Amazingly for the normal model SRS ourperforms StrRS. This may result from the fact that the optimal allocation prefers drawing larger addresses which leads to biased estimates. Again, this effect is much smaller in binomial models.

Figure 3 in contrast to Figure 2 shows the results on community level. As expected due to the much higher variability of the area sizes, the estimates are by far less accurate than on sampling point level. Major impact, of course, can be found in small communities. Again, the normal alternative estimator, denoted by A, yields very high biases which results in a poor performance in terms of RRMSE. The differences between binomial and normal models is considerably smaller than in the previous case. The big differences occur within the community size classes rather than between the models and estimators.

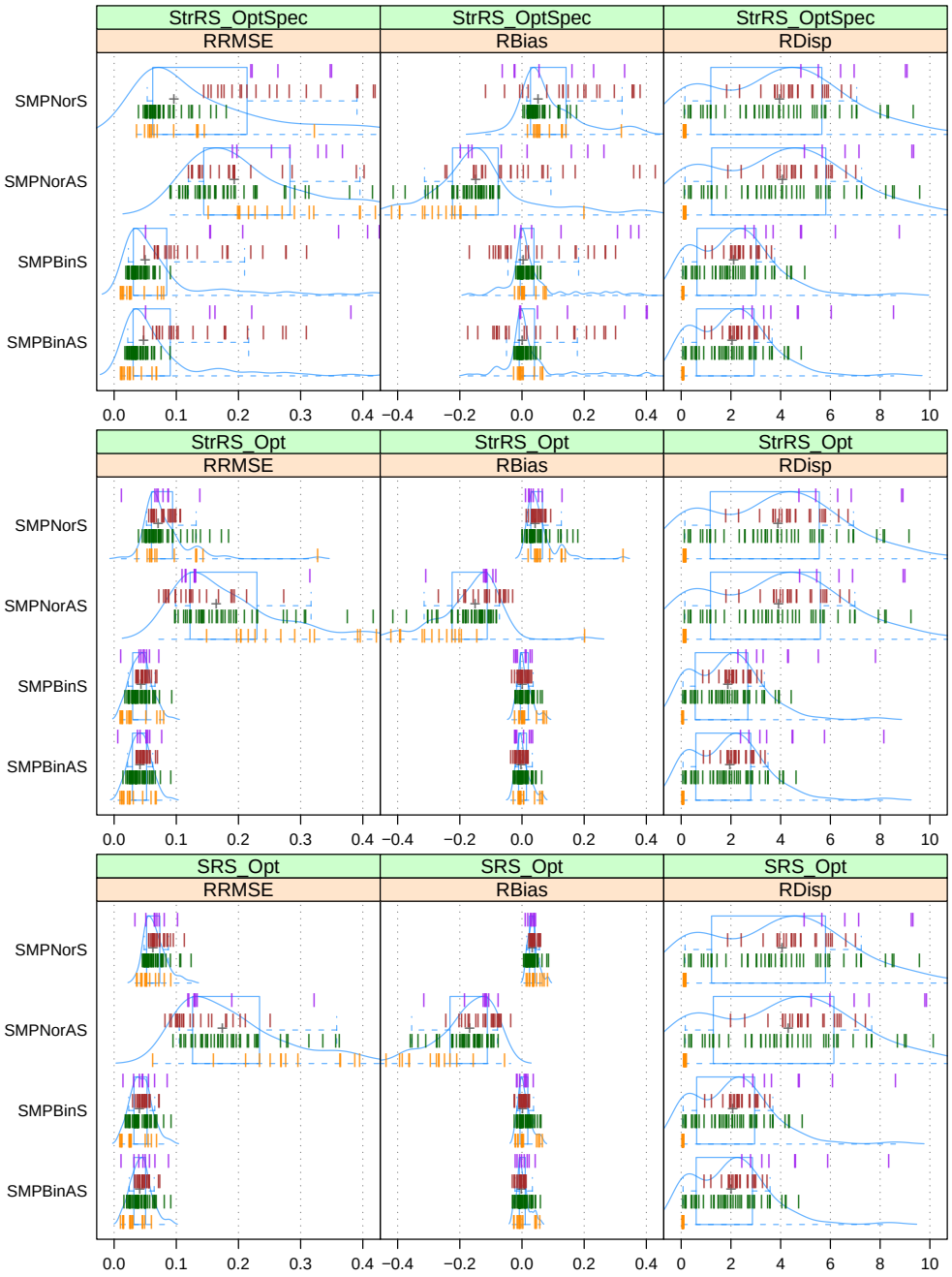

Figure 2: Comparison of estimators for the three designs and three measures on sampling point level

The relative poor performance of the sampling point based estimates with regards to the relative dispersion results from the sampling point specific random effect. This effect could be omitted by introducing a community based crossed effect.

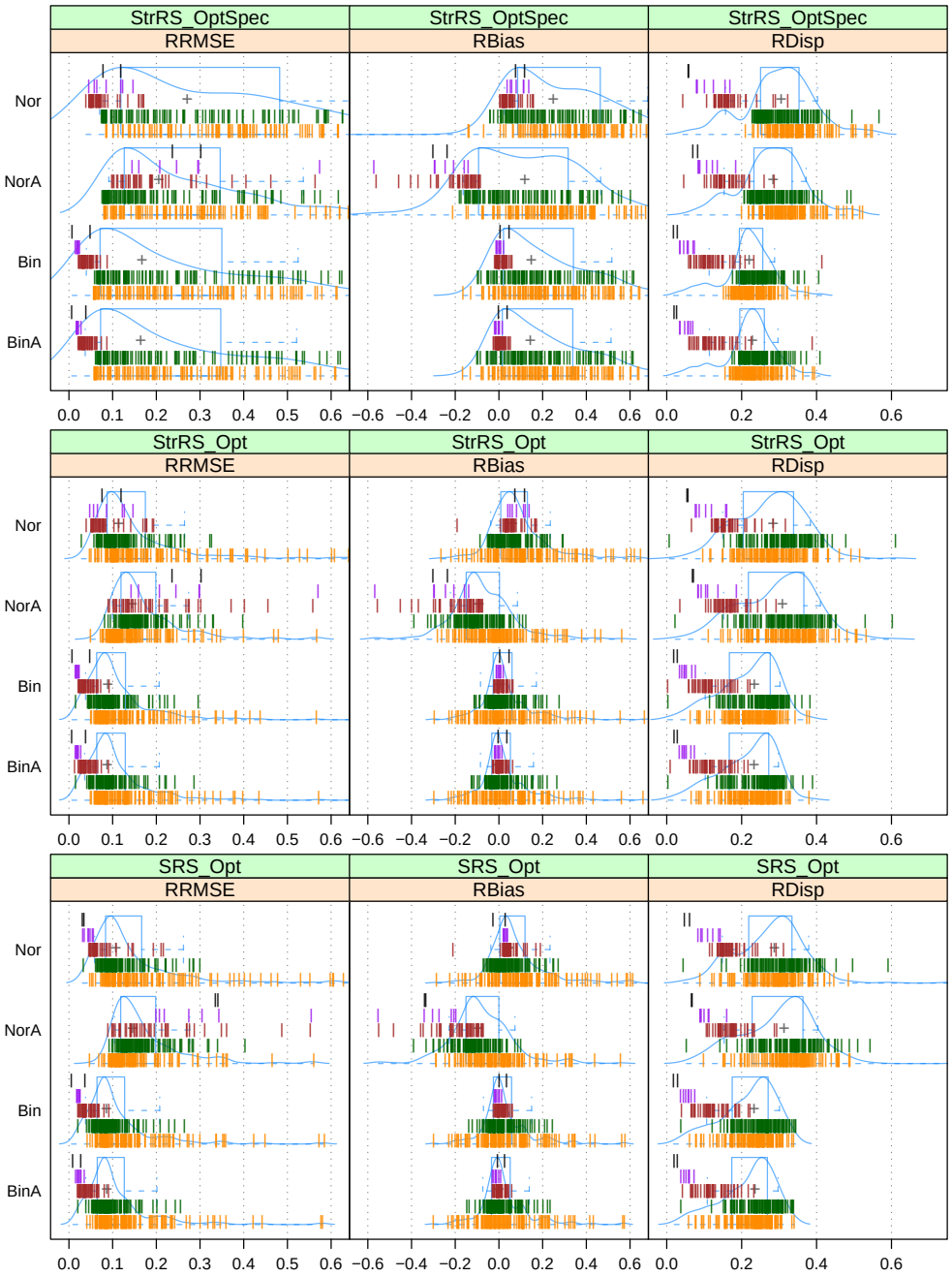

Figure 3: Comparison of estimators for the three designs and three measures on community level

In many studies which introduce spatial correlation considerably improvements of the estimates are found when applying the spatial structure. Serious objections against spatial modeling arise from the idea that the computational burden may spoil the gain in efficiency especially in cases with little or no spatial correlation.

As can be drawn from Figure 4 the Fay-Herriot and the spatial Fay-Herriot estimators perform similarly. The slightly higher relative dispersion in the spatial Fay-Herriot results from the more complex structure of the variance term. This

effect is intensified in the presence of unsampled areas in combination with the underlying missing spatial correlation.

Since almost no spatial correlation is given in the present example for higher education one shall point out that the loss in efficiency due to the more complex estimation process is remarkably small. The possible gains in efficiency can be drawn from Vogt and Lahiri (2010) and Vogt and Münnich (2009).

The gain in efficiency of the other four estimators is mainly due to the usage of the much more precise information in the estimation process. Nevertheless, the difference to the Fay-Herriot models seems much less than expected.

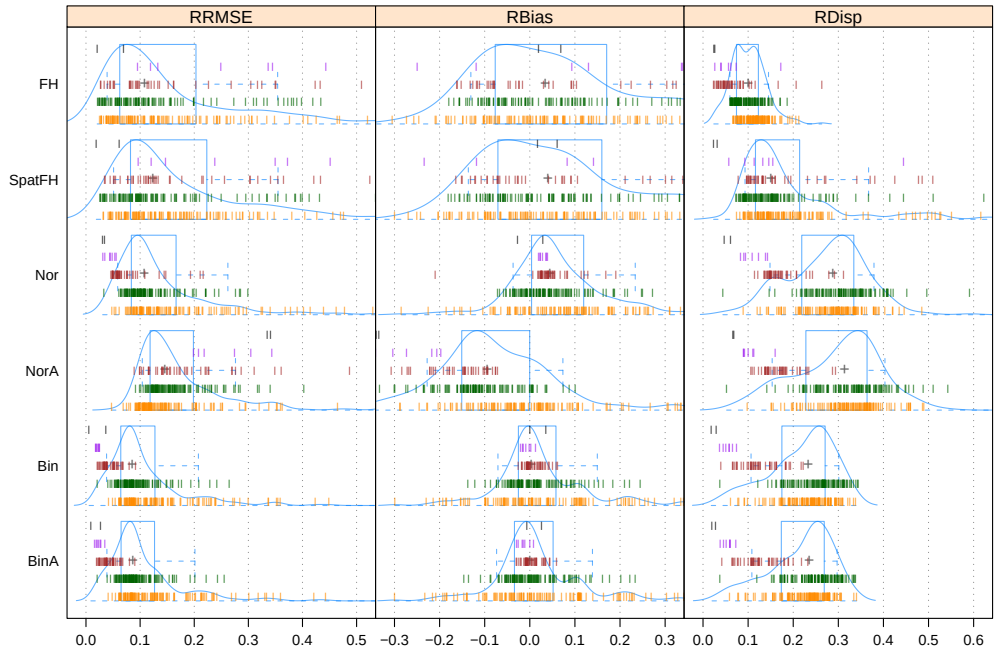

Figure 4: Comparison of spatial and non-spatial Fay-Herriot estimators to the normal and binomial small area estimators on community level

4. Summary and outlook

Within the present study we could see that binomial modeling helps to improve the accuracy of standard normal modeling in small area statistics. The Poisson model was not feasible and needs further investigation. Normal models may be applied with some care in order to avoid disadvantageous situations. Spatial modeling may lead to improvements of the estimates in cases where spatial correlation is available. At least they did not show considerable losses in efficiency in the present example which does not contain any spatial information.

In all cases we could see that sophisticated designs may lead to problems when applying small area models. This was also observed in other studies referenced in this paper. This may lead to the conclusion that it is worth to reduce variation of design weights in the design stage in order to allow for accurate small area modeling. Indeed, this was proposed in Münnich, Gabler, and Ganninger (2009).

Acknowledgements

The paper was presented at the Joint Statistical Meeting 2009 in the invited session on *Recent Advances in Small-area Statistics* organized by Professor Ansu Chatterjee, Univesrity of Minnesota. The authors thank Ansu Chatterjee and Professor Partha Lahiri, University of Maryland and JPSM, for the invitation to this session and having the opportunity for presenting this research.

The research was conducted in connection with the German Census sampling project funded by the German Ministry of Inner Affairs and the German Federal Statistical Office.

REFERENCES

Banerjee, S., Carlin, B., and Gelfand, A.E. (2004), "Hierarchical Modeling and Analysis for Spatial Data", Boca Raton: Chapman & Hall/CRC.

Burgard, J.P. (2009), "Erstellung von Karteileichen- und Fehlbestandsmodellen durch Multilevel-Modelle", unpublished diploma thesis, University of Trier.

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

Kolb, J.-P. (2008), "Die Erzeugung von synthetischen Populationen als Basis zur Mikrosimulation", unpublished diploma thesis, University of Trier.

Magg, K., Münnich, R., Schäfer, J. (2006), "Small Area Estimation beim Zensus 2011", `http://www.forschungsdatenzentrum.de/publikationen/veroeffentlichungen/fdz_beitraege_zu_den_nutzerkonferenzen_band_I.pdf`.

Münnich, R., Gabler, S., Ganninger, M. (2007), "Some Remarks on the Register-based Census 2010/2011 in Germany", Proceedings of the meeting on "Innovative Methodologies for Censuses in the New Millenium", Southampton, 2007, `http://www.s3ri.soton.ac.uk/isi2007/papers/Paper03.pdf`.

Münnich, R., Gabler, S., Ganninger, M. (2008), "Zensus 2011 – Projekt zur methodischen Grundlagenforschung", presentation at the meeting of the German Census Commission, Wiesbaden, 5. June 2008.

Münnich, R., Gabler, S., Ganninger, M., Burgard, J.P., Kolb, J.-P. (2009), "Stichprobenverfahren und Allokation des Stichprobenumfangs für den Zensus 2011, unpublished research report for the German Census Sampling Project.

Münnich, R., Magg, K. (2006), "Design und Schätzqualität im registergestützten Zensus, Ergebnisse einer Monte- Carlo-Studie". In: Faulbaum, F., Wolf, C. (eds.), Stichprobenqualität in Bevölkerungsfragen. Tagungsberichte, Band 12. Informationszentrum Sozialwissenschaft, Bonn, 111 – 137.

Münnich, R., Magg, K., Sõstra, K., Schmidt, K., Wiegert, R. (2004), Workpackage 10: Variance Estimation for Small Area Estimates: Deliverables 10.1 and 10.2. URL `http://www.dacseis.de`. - IST-2000-26057-DACSEIS Reports.

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

Sun, D., Tsutakawa, K., and He, Z. (2001), "Propriety of posteriors with improper priors in hierarchical linear mixed models", Statistica Sinica, 11, 77 – 96.

Vogt, M., Lahiri, P. (2010), "Modelling in Small Area Estimation", in submission.

Vogt, M., Münnich, R. (2009), "On the existence of a posterior distribution for spatial mixed models with binomial responses", *Metron*, 67 (2), 199 – 207.