Modeling Approaches to a Spatio-Temporal Small Area Estimation

Ulrike Schach

Department of Statistics, University of Dortmund, 44221 Dortmund, Germany

uschach@statistik.uni-dortmund.de

Abstract

The distribution of cancer mortality in Germany is collected in two different data sets, one with a high spatial resolution but aggregated data over time, the other with yearly data on a coarse spatial scale. This is due to privacy protection laws, as the data become nearly individual when analyzing rare cancer types or strata of age groups. The aim of this paper is to present a modeling approach to estimate the missing data from the given spatial and temporal marginals. Parameters of spatial and temporal auto-correlation, dispersion, and temporal trend parameters are estimated simultaneously within the Bayesian model, using MCMC techniques based on the Metropolis Hastings algorithm.

Keywords: Spatio-temporal marginals, small area estimation, CAR, Bayes model, Metropolis-Hastings algorithm

1 Introduction

1.1 Small Area Estimation

The general idea of a small area estimation in its original sense is an intrapolation of information collected on a larger spatial scale to local areas within the study region. Additional variables with a high correlation to the variable of interest can be used to improve the estimation, see Rao and Yu (1994). For the purpose of this paper, the small area estimation is necessary as the frequency and distribution of cancer mortality in Germany is published in two different types of resolution. One data set has a high spatial resolution, but aggregated data over time, the other one is based on yearly data, but consists of aggregated data over space. According to Becker et al.(1984), pp.3-4, this mode of data presentation is required by privacy protection laws and tabulation procedures. When regarding rare cancer types or further subdivision by age group the time by location cell frequencies become too small. As knowledge about this data is desirable, however, the estimation of the missing data will be performed in this analysis. The small area estimation in this context uses spatial and temporal dependence structures, based on a Bayesian hierarchical modeling approach. The underlying size of the population at risk is available with the highest spatial and temporal resolution.

1.2 Specific Modifications

As described above, the original idea of small area estimation uses covariables to break up the given marginal data into site-specific data for the small area. This can be performed with a regression approach, by combining a number of additional variables. However, usually the data are taken from cross sectional studies or census data. Thus they include no temporal dimension. Furthermore, even for the spatial aspect, the covariables are considered to include the total spatial variation. In this context, the idea of a Bayesian small area estimation is used to break up given marginals into data for the small area. For this model, the spatial as well as the temporal dimension are important features, and they are modeled as complex dependencies within the data.

1.3 Data Sets

The spatio-temporal small area estimation will be illustrated using data on stomach cancer mortality among men in Germany. As described above, the estimation is performed by combining two types of marginal data sets. Data set I has a coarse spatial structure of the 30 regions ("Regierungsbezirk") of former West Germany, but displays yearly data from 1976 to 1990. Data set II consists of stomach cancer mortality figures with the high spatial resolution of 327 districts ("Landkreis") within the study area, but temporaly aggregated over five year periods from 1976-1980, 1981-1985, and 1986-1990. We will use the following notation:

D_{ti}	number of cancer cases at time t in district i (unknown)
N_{ti}	population size at time t in district i
r_{ti}	raw mortality rate at time t in district i, $r_{ti} = (D_{ti}/N_{ti}) * 100,000$

where $i = 1, \ldots, I$ denotes the districts within the study region, and $t = 1, \ldots, T$ represents time points for the analysis. Using the notation introduced above, the aim of the small area estimation is to obtain $\hat{D}_{ti}$, using the given marginals $D_{.i}$ and $D_{t.}$, where a dot indicates summation over the dotted index. Additionally, the underlying population size N_{ti} is known and available for the analysis. The region of Braunschweig, located in Lower Saxony in the center of Germany, will be used to illustrate the spatio-temporal small area estimation. Braunschweig has been chosen, as it is a region with a reasonable number of districts. Thus, we consider the following table, with the given marginals. The aim is to fill up the missing data, modeled as unknown parameters.

year \ distr.	$i=1$	$i=2$	$\dots$	$i=I$	Σ
t=1					$D_{1.}$
t=2					$D_{2.}$
$\vdots$					$\vdots$
t=T					$D_{T.}$
Σ	$D_{.1}$	$D_{.2}$	$\dots$	$D_{.I}$	$D_{..}$

Table 1: Marginal data for the small area estimation.

For the study area of Braunschweig, we have 11 districts (I=11) and consider five years (T=5) from 1986 to 1990.

2 Spatio-temporal small area estimation

2.1 Spatial dependence structures

The spatial structure of the observed area is based on an irregular lattice structure. We imply stochastic dependence of neighboring sites. Two districts are considered to be neighbors, if they share a common border. We model the spatial dependence as a Markov type departure from independence, where the time series definition of a Markov dependence is transferred to spatial data as described by Cressie (1993), pp.402-410. This means that the observation in district i is dependent on its neighboring sites, denoted by $\{-i\}$. However, it is independent of the remaining sites on the lattice, given the values at the neighboring sites. Markov processes of this type are said to have a conditional autoregressive (CAR) structure. As we consider spatio-temporal dependencies, we model the spatial dependence via the temporal dependence. This approach differs considerably from the purely spatial CAR structure and will be explained in section 2.2. The resulting neighborhood dependencies of the study region of Braunschweig are indicated through the graph in figure 1.

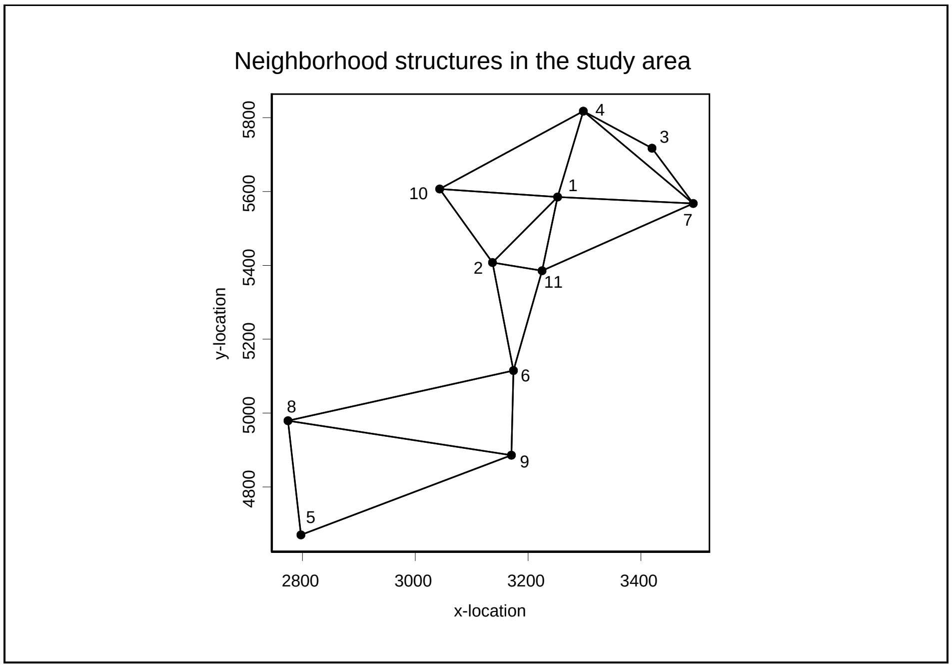

Figure 1: Plot of the neighborhood structures within the study area.

Each of the eleven points displays the corresponding district within the region of Braunschweig, and two points have been connected, if they are neighbors, i.e. if they share a common border.

2.2 Spatio-temporal model

As described above, a major aim of our method is to estimate the missing parameters of cancer mortality, using the given marginals. Additional to the estimation of the unknown cell frequencies, we use a hierarchical Bayesian model that simultaneously estimates parameters of spatial and temporal autocorrelation, dispersion, and temporal trend parameters, as described by Schach (2000). Based on the idea of conditional independence, given neighboring sites in space and time, we introduce the following three-stage hierarchical spatio-temporal model.

The multivariate process of the rates starts at the first time point $t = 1$ with an overall level μ_1 and the site-specific variation b_i, where for reasons of identifiability the b_i's are constrained to sum to zero. The distribution of the vector of the b_i is multivariate Normal. It is assigned a non-informative covariance matrix, to keep it as general as possible. On the first stage we begin with the specification of the prior distributions.

Stage 1: Prior distributions

$$\begin{aligned}
\text{b.init} &\sim \text{MVN}(\gamma, V) \\
\gamma &\sim \text{MVN}(0, U),\ U = \text{unity matrix} \\
V &\sim \text{Wishart}(U, I) \\
b & \quad\ \text{restricted} \\
\mu_t &\sim \text{N}(\bar{\mu}_t, 1000),\ \bar{\mu}_t = D_{t.}/N_{t.} \\
\alpha &\sim \text{N}(0, 0.0001) \\
\beta &\sim \text{N}(0, 0.0001)
\end{aligned}$$

The CAR idea will be implemented in this spatio-temporal context in a modified form. Instead of assigning a conditional autoregressive dependence structure to the vector of spatial random effects b.init at time $t = 1$ directly, we are assigning it a non informative multivariate Gaussian prior distribution. We expect that the spatial dependence or heterogeneity between neighboring sites arises through the model assumption and the data. Vector γ and the precision matrix V are hyperparameters, necessary for the multivariate normal distribution of the vector b.init. The resulting vector b is restricted with a sum-to-zero constraint on b.init to assure identifiability at time $t = 1$, see also Besag and Kooperberg (1995). The overall level μ_t is assigned a non informative Gaussian prior, dependent on t.

Given all rates up to $t-1$, we assume that the rate in district i at time t depends only on the rate in district i at time $t-1$ and on the mean of the neighboring sites of i at time $t-1$. The spatial and temporal dependence is modeled in a way that the differences of the regressor variables of the overall mean at time $t-1$ have a linear effect on the outcome of the rates at time t. Here the regression coefficient for the temporal dependence is α and for the spatial dependence β, respectively. Thus, we arrive at the equations for the rates as presented in stage two of the model. The frequencies of cancer deaths themselves are modeled as Poisson variables of the rates multiplied by the underlying population size as parameters. α and β are parameters of temporal and spatial autocorrelation, which are assigned non informative Gaussian prior distributions. They are estimated simultaneously, along with parameters of dispersion and temporal trend.

Stage 2: Estimation for the small area

Functional relation:

$$\begin{aligned} t = 1 : r_{1i} &= \mu_1 + b_i \\ t > 1 : r_{ti} &= \mu_t + \alpha \left(r_{t-1,i} - \mu_{t-1}\right) + \beta \left(\bar{r}_{t-1,-i} - \mu_{t-1}\right) \\ \lambda_{ti} &= r_{ti}\, N_{ti} \end{aligned}$$

$$\hat{D}_{ti} \mid \lambda_{ti} \sim \mathrm{Poi}(\lambda_{ti})$$

On the third stage of the model we use an indirect adjustment of the sum of the estimated mortality figures to the observed marginals. As we have to account for spatial and temporal marginals, the adjustment is two-dimensional.

Stage 3: Indirect adjustment

$$\begin{aligned} D_{t.} &\sim \mathrm{N}(\hat{D}_{t.}, 1000) \\ D_{.i} &\sim \mathrm{N}(\hat{D}_{.i}, 1000) \end{aligned}$$

This computational trick avoids to assign given data (i.e. marginals) to a sum of estimated parameters, which is cumbersome in this type of Bayesian framework. The parameter estimation is invariant under the change of the order of the two adjustments.

3 Results

3.1 Parameter estimation for the Bayesian model

Before we begin with the presentation of the estimated parameters of the model for the study region of Braunschweig for the years from 1986 to 1990, we will take a look at model diagnostics and convergence, according to Brooks and Gelman (1998). Due to the complexity of the model, a Metropolis Hastings algorithm has been used for the generation of the Markov chains, as illustrated by Brooks (1998) and Chen et al.(2000). We have chosen a burn-in period of 4,000 iteration steps and 8,000 additional recorded updates keeping each 20th iteration for the estimation. As proposed by Neal (1998) we have allowed for over-relaxation. Figure 2 shows the satisfactory acceptance rates of the sampler. We have used the WinBUGS software for the simulation according to Spiegelhalter et al.(2000).

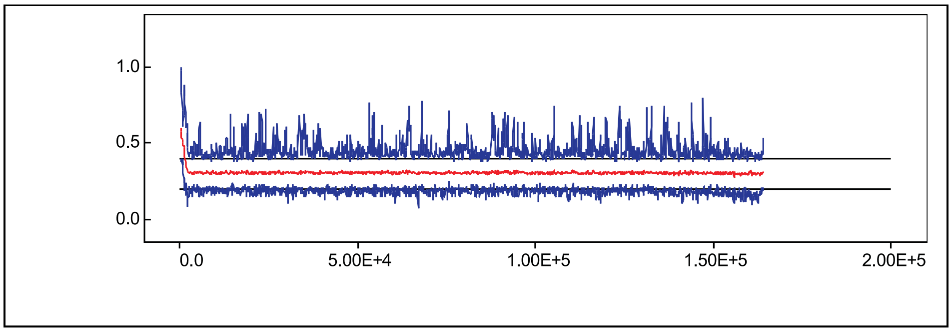

Figure 2: Metropolis Hastings acceptance rates.

Considering the traces, e.g. for the overall level at time $t = 5$ we can see low autocorrelation of the chain in figure 3.

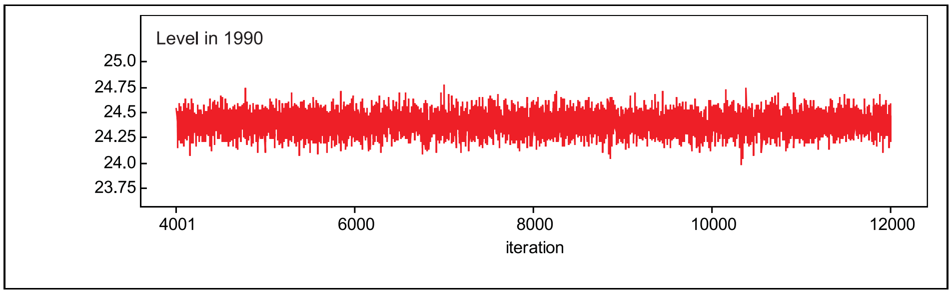

Figure 3: Trace for the level in 1990.

Having checked the model diagnostics, we can look at the resulting parameter estimates of the mortality rates per 100,000 persons at risk. Remembering table 1, it has been the aim to fill the missing cell frequencies. The given yearly data had only a coarse spatial structure of single figures for the whole region. After the application of the spatio-temporal small area estimation, we obtain a spatial structure on the basis of the districts for every year, where the aggregated data over five years has been split up into yearly data. To demonstrate the parameter estimates, the resulting rates per 100,000 inhabitants at risk are displayed in figure 4.

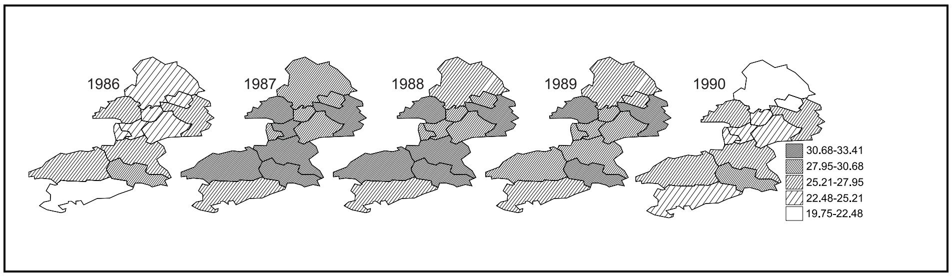

Figure 4: Spatial and temporal resolution of the rates for the study area for the years from 1986-1990.

It is worth mentioning that the parameters of spatial and temporal autocorrelation, dispersion, and temporal trend are simultaneously estimated within the model.

3.2 Proportional Partition

An elementary way of partitioning the $D_{t.}$ into $\{\hat{D}_{ti}, i = 1, \ldots, I\}$ consists of splitting $D_{t.}$ according to corresponding population sizes $\{N_{ti}, i = 1, \ldots, I\}$. This can be justified by a Binomial model with equal rates in all districts. We can use this Binomial model to calculate confidence intervals for the estimated number of cancer deaths by standard methods, proportional to the size of the underlying population at risk. Figure 5 shows the results for the proportional partition and for our more realistic Bayesian approach.

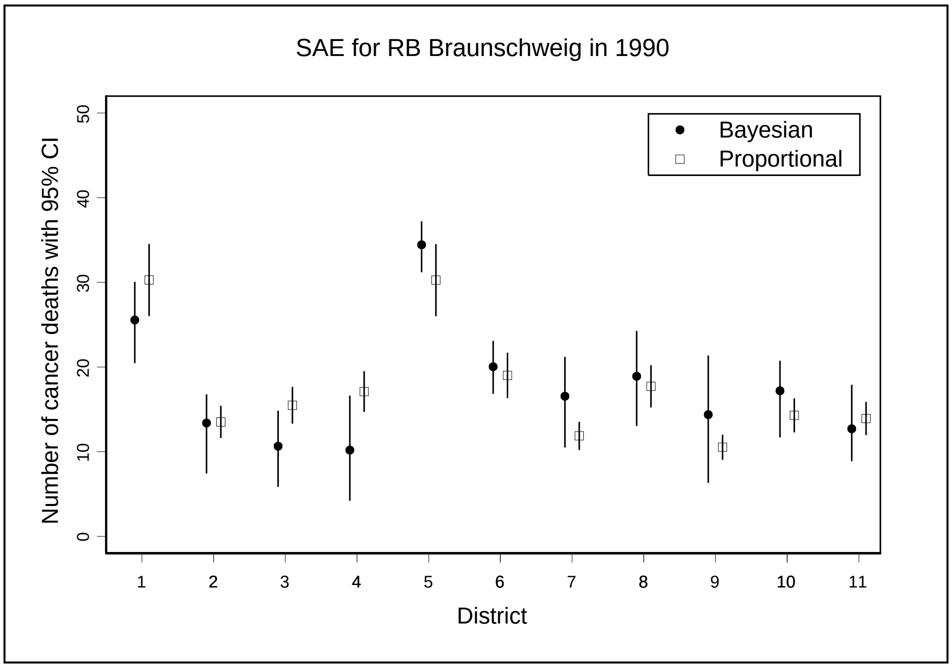

Figure 5: Cancer death estimates and confidence intervals of the Bayestian approach and the Binomial model in 1990.

4 Discussion

The idea of small area estimation has been applied to combine two different kinds of marginals, in order to obtain data with the highest spatial and temporal resolution, based on underlying population figures. The spatio-temporal small area estimation has been performed with an approach that accounts for spatial and spatio-temporal dependencies within the data. When comparing the resulting parameter estimates $\hat{D}_{ti}$ for the small area with those obtained using a proportional partition, the parameter estimates are nearly identical. However, the resulting confidence intervals of the Bayesian model are larger but more realistic, as they include explicitly the spatial and temporal dependence structures. When simply regarding the proportional partition, one will be mislead by the seeming accuracy. Due to the idea of the algorithm to estimate a comparably large number of parameters out of a relatively small number of data, the resulting parameter estimates are strongly dependent. That is why the chains of the Metropolis-Hastings algorithm take longer burn-in periods to reach stationarity of the posterior distribution and a good mixing. Iteration times are prolonged due to the complexity of the model.

When additional covariables are to be included in the model, the proportional partition is no longer valid. Our model can be extended to several temporal blocks, as well as to clusters of regions, as described by Knorr-Held and Raßer (2000) by increasing the number of spatial neighbors. So far, we have explained the procedure with raw mortality rates without an adjustment for age group towards an age-specific standardization. The model can also be refined in that direction. The method is well applicable to the analysis of multi-directional trends within different study regions, for the analysis of temporal trends within small spatial units, and it can easily be extended to age groups and additional covariables. Future research has to be undertaken in the direction of goodness of fit of the model, as proposed by McDonald et al. (1999) . A properly designed simulation study would be an appropriate measure of variability and reproducability of the estimation.

Acknowledgements

This research was supported by the German Research Council (DFG) through the Graduate College and the Collaborative Research Centre at the University of Dortmund (SFB 475): Reduction of complexity for multivariate data structures.

References

[1] Becker, N.; Frentzel-Beyme, R. and Wagner, G. (1984): Atlas of Cancer Mortality in the Federal Republic of Germany. 2nd edition. Springer-Verlag, Berlin

[2] Besag, J. and Kooperberg, C. (1995): On conditional and intrinsic autoregressions. *Biometrika 82, Vol. 4, 733–746.*

[3] Brooks, S. (1998): Markov chain Monte Carlo method and its application. *The Statistician 47, Vol. 1, 69–100.*

[4] Brooks, S. and Gelman, A. (1998): General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics 7, Vol. 4, 434–455.*

[5] Chen, M.; Shao, Q. and Ibrahim, J. (2000): Monte Carlo Methods in Bayesian Computation. Springer, New York

[6] Cressie, N. A. (1993): Statistics for Spatial Data. Wiley, New York

[7] Knorr-Held, L. and Raßer, G. (2000): Bayesian Detection of Clusters and Discontinuities in Disease Maps. *Biometrics, Vol. 56, 13–21.*

[8] McDonald, J. W.; Smith, P. W. and Forster, J. J. (1999): Exact Goodness of Fit of Log-Linear Models for Rates. *Biometrics, Vol. 55, 620–624.*

[9] Neal, R. (1998): Suppressing random walks in Markov Chain Monte Carlo using ordered overrelaxation. In: *Learning in Graphical Models (M. Jordan, ed.). Kluwer Academic Publishers, Dordrecht, 205–230.*

[10] Rao, J.N. and Yu, M. (1994): Small-area Estimation by Combining Time-series and Cross-sectional Data. *The Canadian Journal of Statistics, Vol. 22, 511–528.*

[11] Schach, U. (2000): Spatio-temporal models on the basis of innovation processes and application to cancer mortality data. *Technical Report, Vol. 16/2000, SFB 475, Univ. of Dortmund*

[12] Spiegelhalter, D.; Thomas, A. and Best, N. (2000): WinBUGS Version 1.3 User Manual. Institute of Public Health, Cambridge and Imperial College School of Medicine, London