Technische Universität München

Zentrum Mathematik

Hierarchical Binary Spatial Regression Models with Cluster Effects

Sergiy Prokopenko

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. C. Klüppelberg
Prüfer der Dissertation:	1. Univ.-Prof. C. Czado, Ph.D.
	2. Univ.-Prof. Dr. K. Ickstadt
	Universität Dortmund

Die Dissertation wurde am 02.02.2004 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 11.05.2004 angenommen.

Acknowledgement

My sincere thanks go to my supervisor Prof. Dr. Claudia Czado for her accurate support during my PhD studies. I would also like to thank Dr. Thomas W. Zängler for providing the data set under consideration. Further I express my thanks to Dipl.Math Kathleen Ehrlich, whose Diploma thesis helped me to select the starting set of covariates.

Abstract

This work is motivated by a mobility study conducted in the city of Munich, Germany. The variable of interest is a binary response, which indicates whether public transport has been utilized or not. One of the central questions is to identify areas of low/high utilization of public transport after adjusting for explanatory factors such as trip, individual and household attributes. The goal of this thesis is to develop flexible statistical models for a binary response with covariate, spatial and cluster effects. One approach for modeling spatial effects are Markov Random Fields (MRF). A modification of a class of MRF models introduced by Pettitt, Weir, and Hart (2002) is developed in this work. This modification has the desirable property to contain the intrinsic MRF in the limit and still allows for fast and efficient spatial parameter updates in Markov Chain Monte Carlo (MCMC) algorithms. In addition to spatial effects, cluster effects are taken into consideration. Group and individual approaches for modeling these effects are suggested. The first one models heterogeneity between clusters, while the second one models heterogeneity within clusters. An unidentifiability problem occurring in the second case is solved. For hierarchical spatial binary regression model with individual cluster effects two MCMC algorithms for parameter estimation are developed. The first one is based on a direct evaluation of the likelihood. The second one is based on the representation of binary responses with Gaussian latent variables through a threshold mechanism, which is particularly useful for probit models. Extensive simulations are conducted to investigate the finite sample performance of the MCMC algorithms developed. They demonstrate satisfactory behaviour. Finally the proposed model classes are applied to the mobility study.

Contents

Chapter 1

Introduction

This work has been motivated by a German mobility study investigating the usage of public transport options. The variable of interest was a binary response, whether public transport has been utilized or not. One central question of the investigators is to identify areas of low/high utilization of public transport after adjusting for explanatory factors such as trip, individual and household attributes. Therefore the goal is to develop flexible statistical models for a binary response with covariate, spatial and cluster effects. There are a great number of statistical models in the literature which incorporate covariates together with spatial information. In the context of general additive models, the simplest possibility to account for spatial information would be to use an additional nominal covariate indicating the region if there are multiple responses per region. But such an approach does not give a model for spatial dependence . This property is especially desired if the data volume is not large with respect to the number of covariates. In this case the assumption of a spatial structure (such as spatial smoothness) is especially helpful to be used as additional prior information.

There are two general approaches to incorporate spatial effects in a model. The first one is appropriate for data collected at specified point locations, while the other one uses data regions. The first approach is known as *generalized linear kriging* (see for example Diggle, Tawn, and Moyeed 1998). It is based on generalized linear mixed models (Breslow and Clayton 1993), where spatial random effects are modeled as realizations of a stationary Gaussian process with zero mean and a parameterized covariance structure. For binary data this approach models the success probability p_i as follows:

$$p_i = \mathbf{E}(Y_i|\mathbf{x}_i, b_i) = h(\eta_i) \quad \text{and} \quad \eta_i = \mathbf{x}_i'\boldsymbol{\alpha} + b_i \;, \quad i = 1, \ldots, n \;, \tag{1.1}$$

where $\mathbf{x}_i$ is the design vector of the random variable Y_i and $b_i, i = 1, \ldots, n$, are realizations of a zero mean stationary Gaussian process $\mathbf{b}$ at the locations of the Y_i's. The parameterization of the covariance structure by a covariance parameter $\boldsymbol{\delta}$ is usually

based on distances between the observed locations. Even in the case of normal responses $Y_i, i = 1, \ldots, n$, maximizing the likelihood over $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ becomes analytically intractable as soon as independence of the spatial effects $b_i, i = 1, \ldots, n$, cannot be assumed. One general approach therefore is to maximize the reduced log-likelihood $l(\mathbf{Y}; \hat{\boldsymbol{\alpha}}(\boldsymbol{\delta}), \boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$, where $\hat{\boldsymbol{\alpha}}(\boldsymbol{\delta})$ is the maximum likelihood estimate of $\boldsymbol{\alpha}$ for fixed $\boldsymbol{\delta}$, and profile over $\boldsymbol{\delta}$. But such estimation is computationally expensive for large data sets. For arbitrary responses parameter estimation is carried out by Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling (see Diggle, Tawn, and Moyeed 1998). For large data sets the updating of the covariance parameter $\boldsymbol{\delta}$ is difficult, since it requires to compute the determinant and inverse of a large dimensional variance-covariance matrix at each iteration. Heagerty and Lele (1998) remark (p.1104) that this step is computationally prohibitive already for sample sizes larger than 500. To overcome this problem they assume local independence between spatial effects which have a distance longer than some fixed value R. Heagerty and Lele (1998) use this idea for an iterative approach to determine the local conditional posterior mode of the spatial effect for the prediction at a new location. In contrast to Diggle, Tawn, and Moyeed (1998), Heagerty and Lele (1998) estimate spatial effects $b_i, i = 1, \ldots, n$ using a composite likelihood approach. Gelfand, Ravishanker, and Ecker (2000), which analyze a binary kriging model for the probit link function $h(\cdot)$ in (1.1), propose to apply MCMC with a suitably selected *importance sampling density*. They note that their method replaces a $n \times n$ matrix inversion with sampling from an n-dimensional normal, which for large values of n can be carried out much faster using a Cholesky decomposition. Their approach also does not need to compute the determinant of the variance-covariance matrix. It allows to determine the posterior distribution of the regression parameter $\boldsymbol{\alpha}$ and the covariance parameter $\boldsymbol{\delta}$, but the posterior distribution for the spatial effects $b_i,\ i = 1, \ldots, n$ cannot be calculated this way.

The other approach to incorporate a spatial model is appropriate when spatial effects are associated with data regions. These do not need to be on a regular lattice. The model equation is similar as in (1.1), but now data are assumed to be aggregated over regions and spatial effects are individual for each region instead for each observation, as before. Therefore the linear predictors are modeled as

$$\eta_i = \mathbf{x}_i' \boldsymbol{\alpha} + b_{j(i)}, \quad i = 1, \ldots, n,\ j = 1, \ldots, J\ ,$$

where J denotes the number of regions and $j(i)$ indicates the region associated with the i^{th} observation. The spatial effects $b_j,\ j = 1, \ldots, J$, are modeled as a realization from some Gaussian *Markov random field* (MRF) (Besag and Green 1993). Gaussian MRF's are also a zero mean Gaussian process. The name *Gaussian conditional autoregression* (Gaussian CAR) is also used, since such a distribution is typically given through its full conditionals. This last fact allows fast individual updating of $J << n$ spatial effects in

a Gibbs sampler scheme. More precisely, the distribution of the spatial effect b_j given all the other spatial effects depends only on the spatial effects of the neighbors of the j^{th} region. Therefore this approach requires some spatial neighborhood structure. This modeling is appropriate for our mobility application, since data are aggregated over postal codes of Munich, Germany. It is natural to consider two postal codes as neighbors if they have a joint border. In contrast to the stationary Gaussian process used in kriging, in Gaussian CAR models the explicit form of its precision matrix (inverse covariance matrix) is available. Therefore we do not need to compute the inverse of the variance-covariance matrix when updating the covariance parameter $\boldsymbol{\delta}$. Moreover this precision matrix is usually sparse, which allows to compute its determinant much faster, as in the kriging approach. Further, Pettitt, Weir, and Hart (2002) use this fact and propose a specific dependence structure which provides even an analytical computation of its determinant. The next difference to stationary Gaussian processes, used in kriging, is that some Gaussian CAR models possess an improper joint density. The simplest example is the intrinsic CAR model (Besag and Green 1993), whose precision matrix is only semi positive. Fahrmeir and Lang (2000) use improper intrinsic CAR models as a prior for a Bayesian semi parametric regression model for multi categorical time-space data, while Knorr-Held and Rue (2000) applied intrinsic CAR priors for Poisson models used in disease mapping. For our application we study more advanced proper Gaussian CAR models with a parameterized correlation matrix. In particular, we develop a modification of the Pettitt's CAR model, which includes in the limit a specific intrinsic CAR model. The modification we propose still has all nice properties of the Pettitt, Weir, and Hart (2002) CAR models: proper joint distributions, a similar interpretation of parameters, the same conditional correlations and more important allows for fast computation of the determinant of the precision matrix, providing fast Gibbs sampling. Gaussian CAR models will be considered in more detail in Section 2.

A principally different, well-known approach, also developed for spatial regression binary data over the region lattice is the *auto logistic regression model*. Huffer and Wu (1998), which use this method for the analysis of the distribution of plant species in Florida, U.S.A., propose to extend the auto logistic modeling of the success probability for each species by incorporating a fixed effect term $\mathbf{x}_i'\boldsymbol{\alpha}$:

$$\log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\alpha} + \gamma y_i^*, \quad y_i^* := \sum_{j:i\sim j} y_j \ ,$$

where "$i \sim j$" indicates that sites i and j are neighbors. They work with a regular rectangular lattice and one-observation-per-site data. But in spite of this simplicity Huffer and Wu (1998) note that exact MLE is not tractable, except when the number of sites is quite small, while two other estimation methods, namely the coding method (Besag

(1974)) and the maximum pseudo-likelihood method (Besag (1975)), seem to be not sufficiently efficient. For their application Huffer and Wu (1998) investigate a MCMC MLE approach, which produces the likelihood function via Monte Carlo simulations. Note they do not give any idea, how to take into consideration possible interactions between different species. For the Gaussian CAR approach Pettitt, Weir, and Hart (2002) solve this problem by modeling the correlation between several Gaussian CAR models for each species applied to tree biodiversity data. Also Carlin and Banerjee (2002) develop this approach for multiple cancer survival data with 3 types of cancer.

We close our short overview on spatial modeling for binary data by mentioning a *non parametric binary regression* approach, which was proposed by Kelsall and Diggle (1998) for the analysis of spatial variation in risk of disease. The idea is to model logits through spatially dependent intensities $\lambda_1(x)$ (cases) and $\lambda_2(x)$ (controls), where $x \in A \subset \mathbb{R}^2$ is the response point location:

$$p(x) := P(Y_i|X_i = x) = \frac{q_1\lambda_1(x)}{q_1\lambda_1(x) + q_2\lambda_2(x)} \quad \Rightarrow \quad \text{logit}(p(x)) = \log\left(\frac{\lambda_1(x)}{\lambda_2(x)}\right) + c \quad (1.2)$$

Kelsall and Diggle (1998) consider non parametric kernel estimators $\hat{p}_h(x)$ for $p(x)$, where h denotes the corresponding bandwidth. Cross-validation is used to optimize h. While in the Gaussian CAR approach we test the significance of the spatial effects $b_j,\ j = 1, \ldots, J$, Kelsall and Diggle (1998) construct tolerance contours, which indicate for each x whether $\hat{p}_h(x)$ is consistent with the proportional hazard assumption, which is given by the null hypothesis $H_0 : \frac{\lambda_1(x)}{\lambda_2(x)} = \text{const}$. These are determined by generating m times new data which are consistent with H_0 but otherwise similar in distribution to the original data. Finally, they construct new estimates $\hat{p}_h^{sim}(x)$ for each of the m generating data sets. The authors note that, since covariates are not included in (1.2), the collection of a stratified sample of controls can be very difficult, particularly when the number of covariates is large. Therefore they extend their model by including fixed effects $\mathbf{u}^t\boldsymbol{\alpha}$:

$$\text{logit}(p(x, u)) = \mathbf{u}^t\boldsymbol{\alpha} + g(x) \quad (1.3)$$

This extension allows to collect a simple random sample of controls and to take into account covariates by modeling their effects within (1.3). The authors however note that kernel regression estimation based on Model (1.3) is substantially more computer intensive, so that the simpler kernel regression method, based on (1.2) will sometimes be preferable.

In addition to spatial effects we extend our modeling of the linear predictor η_i by cluster random effects. It allows us to obtain more flexible models and to take into account possible overdispersion caused by unobserved heterogeneity. Examples of clusters are age groups or the household types (single, couple, with children etc.). We consider

two approaches, namely group and individual cluster effects. The first one, which models heterogeneity between clusters, implies the usual idea of having the same random effect within a cluster. These random effects are usually assumed to be some realization from the multivariate normal $N_K(0, \sigma_c^2 I_K)$ with usually unknown cluster variance σ_c^2. Here K denotes the number of clusters and I_K stands for the K-dimensional identity matrix. A different choice of the variance-covariance matrix is possible.

The second approach allows for heterogeneity within a cluster, i.e. we model cluster effects within a cluster as independent normally distributed random variables with zero mean and a cluster specific variance. Therefore we have to estimate K cluster specific variances instead of K cluster effects as before. Afterwards we will show how an unidentifiability problem occurring in the second case can be overcome. For this hierarchical spatial binary regression model with individual cluster effects we develop two estimating MCMC algorithms. The first one is useful if the likelihood of the data, given covariates and unknown parameters, can be easily computed as for binary logistic models. Markov Chains are then generated using Metropolis-Hastings steps. The second approach, which is particularly useful for probit model is based on latent variables, where the observed binary responses are generated through a threshold mechanism. For latent Gaussian variables this leads to binary probit models (see for example Albert and Chib 1993). For MCMC inference, Gaussian latent variables are considered as unknown additional "parameters" and are generated with the other parameters in a Gibbs sampling scheme. We note that block updating for the regression parameters $\boldsymbol{\alpha}$ and the spatial parameters $\mathbf{b}$ is available in this estimating algorithm. Therefore we achieve considerably better mixing than in the direct algorithm, where parameters are updated individually. This allows us to reduce the number of iterations in the corresponding Markov chains. Further, this method reduces the number of parameters, which require a numerically more complicated Metropolis-Hastings step to only one.

In connection with cluster modeling we mention briefly also another approach developed by Kuhnert, Mengersen, and Smith (2002). They use clustering of observations with similar design vectors. For forecasting they take the average response value over the corresponding cluster, if responses are continuous, while in the discrete case the most frequent value in the associated cluster is chosen. The determination of clusters is carried out using *reversible jump MCMC*. A major advantage of this approach is that complex covariate interactions can be accommodated. Kuhnert, Mengersen, and Smith (2002) note that this method can be especially useful for small binary data sets, where data is located in sparse areas across the study region and the observed success probability is small. In this case the objects under investigation may be better described by smooth representations of their covariates instead of being described solely by their geographical coordinates. For

our large binary mobility data with many covariates and larger success probabilities, however, the computational effort which is required in this approach and known difficulties with the change point problem in reversible jump block MCMC make this approach not attractive. Besides this the authors seem to assume that all covariates are metric.

The renainder of the thesis is organized as follows. In Section 2 we discuss spatial effects modeling using Gaussian CAR processes. We propose some modification of the Gaussian CAR models developed by Pettitt, Weir, and Hart (2002), which allows to achieve an intrinsic CAR model in the limit. Using this modification we develop in Section 3 a hierarchical spatial binary regression model with group cluster effects, while in Section 4 we present individual cluster effects modeling. In Section 5 we present the results of a comprehensive simulation study investigating the performance of the MCMC algorithms developed in Sections 3 and 4 in small samples. In Section 6 we apply our approach to data from a German mobility study. Finally Section 7 gives a discussion and presents an outlook for further research. The appendix contains in addition to proofs of some results presented in Chapter 2, an introduction to generalized linear models and MCMC methods.

Chapter 2

Modeling of Spatial Effects Using Gaussian CAR Models

The most popular kind of Markov random fields (MRF) are Gaussian MRF's (Besag and Green 1993), or Gaussian conditional autoregressive processes (Gaussian CAR) (see for detailed discussion Pettitt, Weir, and Hart 2002).

Let the random vector $\mathbf{b} \in R^J$ represent a Gaussian CAR process. Then its distribution is defined through its full conditionals as follows:

$$b_j|\mathbf{b_{-j}} \sim N\left(\mu_j + \sum_{j' \neq j} c_{jj'}(b_{j'} - \mu_{j'}),\, \kappa_j\right), \quad j = 1, \ldots, J\,,$$

where $\mathbf{b_{-j}} = (b_1, \ldots, b_{j-1}, b_{j+1}, \ldots, b_J)^t$. Here $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

In particular, a Gaussian CAR with zero-mean has the following full conditionals:

$$b_j|\mathbf{b_{-j}} \sim N\left(\sum_{j' \neq j} c_{jj'} b_{j'},\, \kappa_j\right)\,. \tag{2.1}$$

It is known (Besag and Green 1993), that the joint distribution of the vector $\mathbf{b}$ has the form

$$\mathbf{b} \sim N_J\left(0,\, (I_J - C)^{-1} M\right), \tag{2.2}$$

where $C = (c_{jj'})$ with $c_{jj} = 0$, $j = 1, \ldots, J$, and $M = \mathrm{diag}(\kappa_1, \ldots, \kappa_J)$. Here $N_J(\mu, \Sigma)$ denotes a J-dimensional normal distribution with mean vector μ and covariance matrix Σ. The precision matrix $Q = M^{-1}(I_J - C)$ consists of the elements $q_{jj'} = -\frac{c_{jj'}}{\kappa_j}$ for $j \neq j'$ and $q_{jj} = \frac{1}{\kappa_j}$. Requiring symmetry of the precision matrix gives the first restriction on the values $c_{jj'}$ and κ_j:

$$\frac{c_{jj'}}{\kappa_j} = \frac{c_{j'j}}{\kappa_{j'}}\,. \tag{2.3}$$

Due to its clear smoothing interpretation (see Example 1 below) and simple parameterization, Gaussian CARs are popular for modeling spatial effects, when data are aggregated in regions. The components of the vector $\mathbf{b}$ represent in this case regions, which have a known (often not regular) neighborhood structure. Gaussian CARs can also be used as prior distributions for spatial effects in Bayesian inference. Below we present some examples of Gaussian CARs. Further we will always assume that the neighborhood structure occurring in an Gaussian CAR has no isolated regions or groups of regions.

Example 1: The first, probably the most common example is the *intrinsic Gaussian CAR* (Besag and Green 1993). The distribution of the vector $\mathbf{b}$ is given as

$$b_j|\mathbf{b_{-j}} \sim N(\overline{b_j}, \frac{\tau^2}{N_j}), \quad j = 1, \ldots, J, \quad \text{and} \quad \overline{b_j} = \frac{\sum_{j\sim j'} b_{j'}}{N_j}, \tag{2.4}$$

where $N_j =$ # of neighbors of the region j, and "$j \sim j'$" denotes contiguous regions. In particular, we have $j \not\sim j$.

The smoothing effect for each b_j, $\quad j = 1, \ldots, J$, given its neighborhood is represented here clearly by its conditional mean $\overline{b_j}$, equal to the average value over the neighborhood, and its variance $\frac{\tau^2}{N_j}$, which decreases, if the number of neighbors increases. Note, that in contrast to the conditional mean of b_j given the remaining components, the conditional variance of the b_j depends on the number of its neighbors (which can be interpreted like the density of the regions in this area), but not on their (weighted) values. In order to get probably a more realistic smoothing effect, the components $b_{j'}$, $j' \neq j$ in (2.4) can be weighted, proportional to the joint border length, or inverse proportional to the distances between its centers. For example, if $\omega_{jj'}$ denotes the joint border length between the regions j and j' and $\omega_{j+} = \sum_{j'} \omega_{jj'}$ is the length of the whole border of region j, then the distribution of the vector $\mathbf{b}$ can be written through its full conditionals, as

$$b_j|\mathbf{b_{-j}} \sim N(\overline{b_j}, \frac{\tau^2}{\omega_{j+}}), \quad j = 1, \ldots, J \quad \text{and} \quad \overline{b_j} = \frac{\sum_{j\sim j'} \omega_{jj'} b_{j'}}{\omega_{j+}}. \tag{2.5}$$

Since $\omega_{jj'} = \omega_{j'j}$, Condition (2.3) is satisfied. According to (2.4) and (2.2) the precision matrix of $\mathbf{b}$ is equal to $\frac{Q_0}{\tau^2}$, where

$$Q_0 = (q_{jj'}) = \begin{cases} N_j, & \text{if } j = j' \\ -1, & \text{if } j \sim j' \\ 0, & \text{otherwise} \end{cases}. \tag{2.6}$$

We will show that this matrix is positive semi-definite with $\text{rank}(Q_0) = J - 1$, therefore $\mathbf{b}$ has an improper density. More precisely, we will prove that the joint distribution of the vector $\mathbf{b}$ can be presented in the following form:

$$\begin{cases} b_J & \sim & const \\ \mathbf{b_{-J}}|b_J & \sim & N_{J-1}\left(\mathbf{1}_{J-1}\, b_J,\, \tau^2 Q_{11}^{-1}\right),\ Q_{11} - \text{ positive definite} \end{cases}, \tag{2.7}$$

where $\mathbf{1}_{J-1} = (1, \ldots, 1)^t \in \mathbb{R}^{J-1}$ and the matrix $Q_{11} \in \mathbb{R}^{(J-1)\times(J-1)}$ is a (positive definite) sub matrix of Q_0 in which the last row and column have been deleted, i.e.

$$Q_0 = \begin{pmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{pmatrix} .$$

Here we denote $Q_{12} = (q_{1J}, \ldots, q_{J-1,J})^t$ and $Q_{22} = q_{JJ}$. The proof of (2.7) is given in the Appendix. Note that (2.7) does not represent a singular multivariate normal distribution.

Example 2: Pettitt, Weir, and Hart (2002) use a particular Gaussian CAR, where

$$b_j | b_{j'}, j \neq j' \sim N\left(\frac{\phi}{1+|\phi|N_j} \sum_{j\sim j'} b_{j'} \, , \, \frac{\tau^2}{1+|\phi|N_j} \right) . \tag{2.8}$$

Here the parameter ϕ measures the strength of the spatial dependency. There is no spatial dependency, if $\phi = 0$. Since maximum likelihood estimation is intractable for this model MCMC methods have been used to estimate ϕ and τ^2. It was shown (Pettitt, Weir, and Hart 2002) that a fast and simple update of ϕ for a Gibbs Step given the vector $\mathbf{b}$ and τ^2 is available. Note that in contrast to the intrinsic CAR, the joint distribution of $\mathbf{b}$ based on conditionals specified in (2.8) is a proper distribution, which leads to a proper posterior when used as a prior distribution. This will circumvent any problems in the Gibbs sampler arising from using an improper prior.

Example 3: For our analysis we introduce now a modified Pettitt's CAR. The distribution for $\mathbf{b}$ is given via its full conditionals as follows:

$$b_j | b_{j'}, j \neq j' \sim N\left(\frac{\phi}{1+|\phi|N_j} \sum_{j\sim j'} b_{j'} \, , \, \frac{(1+|\phi|)\tau^2}{1+|\phi|N_j} \right) . \tag{2.9}$$

This distribution differs from Pettitt's CAR (2.8) by the additional term $1 + |\phi|$ in the numerator of the conditional variance. This modification allows us to have the intrinsic CAR (2.4) in the limit, when $\phi \to \infty$. Further, this model has the same behavior as Pettitt's CAR (2.8) when ϕ goes to zero (no spatial dependency). Finally, all partial correlations between b_j and b_i given all the other sites are also the same, as in (2.8).

Note again that in contrast to its limiting case (2.4), the joint distribution of the modified Pettitt's CAR (2.9) is still proper (namely multivariate normal) for any $\phi \in \mathbb{R}$. The asymptotic behavior of the modified Pettitt's CAR if $\phi \to +\infty$ can be presented in the following simple form, which is similar to (2.7):

$$\left\{ \begin{array}{rcl} b_J & \sim & N\left(0, \, \sigma_J^2(\phi)\right) \quad \text{with } \lim_{\phi\to+\infty} \frac{\sigma_J^2(\phi)}{\tau^2 \frac{\phi}{J}} = 1 \\ \mathbf{b_{-J}} | b_J & \sim & N_{J-1}\left(\mu_J(\phi)\, b_J, \, \tau^2 \Sigma_{11}^{m.P}(\phi)\right) \quad \text{with} \\ & & \lim_{\phi\to+\infty} \mu_J(\phi) = \mathbf{1}_{J-1} \text{ and } \lim_{\phi\to+\infty} \Sigma_{11}^{m.P}(\phi) = Q_{11}^{-1} \quad . \end{array} \right. \tag{2.10}$$

The proof of (2.10) is provided in the Appendix.

In the modified Pettitt's model we can also achieve a simple update for ϕ. To indicate the dependency on ϕ we write now $Q^{m.P}(\phi)$ for the precision matrix of the modified Pettitt's model (2.9). In particular, the corresponding likelihood function $[\mathbf{b}|\phi, \tau^2]$ is given by

$$\left(\frac{1}{2\pi\tau^2}\right)^{\frac{J}{2}} |Q^{m.P}(\phi)|^{1/2} \exp\left(-\frac{1}{2\tau^2}\mathbf{b}'Q^{m.P}(\phi)\mathbf{b}\right) \tag{2.11}$$

Therefore each update of ϕ requires the computation of the determinant of $Q^{m.P}(\phi)$. Since the latter is not analytically tractable for general sets of irregularly spaced sites, we need some efficient fast numerical technique for calculating it. With the reparametrization $\psi = \frac{\phi}{1+|\phi|}$ we can apply the same procedure as in Pettitt, Weir, and Hart (2002). More precisely, if we define the diagonal matrix

$$D = \text{diag}(N_1 - 1, \ldots, N_J - 1) \quad \text{and} \quad \Gamma = (\gamma_{jj'})_{j,j'=1,\ldots,J} = \begin{cases} 1, & \text{if } j \sim j' \\ 0, & \text{if } j \not\sim j',\ j = j' \end{cases},$$

then $Q^{m.P}(\phi)$ can be written in the form

$$Q^{m.P}(\psi) = I_J + |\psi|D - \psi\Gamma = \begin{cases} I_J - \psi(\Gamma - D), & \text{if } \psi > 0 \\ I_J, & \text{if } \psi = 0 \\ I_J - \psi(\Gamma + D), & \text{if } \psi < 0\,. \end{cases}$$

If $(\lambda_1, \ldots, \lambda_J)$ are the eigenvalues of $\Gamma - D$ and $(\nu_1, \ldots, \nu_J)$ are the eigenvalues of $\Gamma + D$, then the determinant of $Q^{m.P}(\psi)$ is equal to

$$|Q^{m.P}(\psi)| = \begin{cases} \prod_j(1 - \psi\lambda_j), & \text{if } \psi > 0 \\ 1, & \text{if } \psi = 0 \\ \prod_j(1 - \psi\nu_j), & \text{if } \psi < 0 \quad . \end{cases} \tag{2.12}$$

and can be computed quickly for any value of ψ. This result shows a noticeable advantage over the general kriging model mentioned in the introduction, since the Gibbs step for the dependency parameter δ in this model is not analytically tractable for irregular neighborhood structures in general.

Chapter 3

Spatial Binary Regression with Group Cluster Effects

3.1 Formulation of the Model

In this section we will present a first model for the data from the mobility study. As written before, our response represents a binary vector $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ with

$$Y_i = \begin{cases} 1 & \text{if trip i used automobile} \\ 0 & \text{if trip i used public transport} \end{cases}, \quad i = 1, \cdots, n. \tag{3.1}$$

We assume, that the Y_i's follow a Bernoulli distribution with the success probabilities p_i and that the Y_i given p_i are independent for $i = 1, \ldots, n$. We model p_i through their logits as follows:

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i)}}_{\textit{random cluster effect}}. \tag{3.2}$$

Here the design vector $\mathbf{x_i}$ multiplied with the regression parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ represents the fixed effects. With the vector $\mathbf{b} = (b_1, \ldots, b_J)$ we attempt to take into consideration possible random spatial effects. As sites we take $J = 74$ postal code areas of the city of Munich. Therefore, the index $j(i)$ denotes the postal code of residence of the person who takes trip i. In order to be able to take into account possible spatial smoothness we assume, that the b_j's arise from the modified Pettitt's CAR (2.9).

To model heterogeneity between the clusters we allow also for random cluster effects, which are represented by the vector $\mathbf{c} = (c_1, \ldots, c_M)$. We assume that each of the M clusters (say age groups or household types) induces a group specific random effect, which we denote by $c_m, m = 1, \ldots, M$, respectively. Therefore we speak of group cluster effects. The index $m(i)$ denotes the cluster of trip i. Finally, we assume that $c_m \sim N(0, \sigma_c^2)$ *i.i.d.*

for $m = 1, \ldots, M$. This completes the description of Model (3.2). Figure 3.1 illustrates the hierarchical model structure of (3.2). Note that the likelihood of the response vector $\mathbf{Y}$

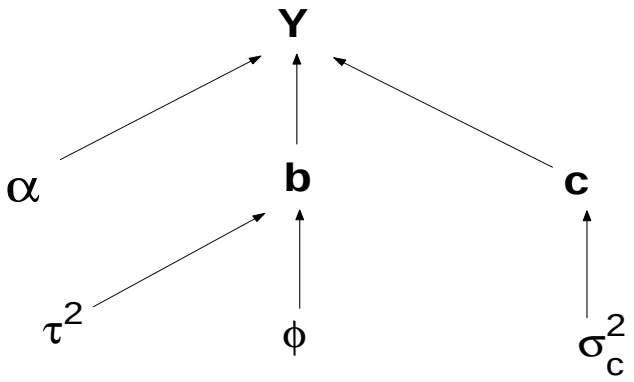

Figure 3.1: Hierarchical Model Structure for Model (3.2)

is therefore proportional to

$$[\mathbf{Y}|\ \boldsymbol{\alpha},\ \mathbf{b},\ \mathbf{c}] \propto \prod_{i=1}^{n} \frac{\exp(Y_i(\mathbf{x_i}^t\boldsymbol{\alpha} + b_{j(i)} + c_{m(i)}))}{1 + \exp(\mathbf{x_i}^t\boldsymbol{\alpha} + b_{j(i)} + c_{m(i)})} \ .$$

Finally, we remark that Model (3.2) is similar to a family of semi parametric models for multi categorical time-space data (with time- instead cluster effects) as discussed in Fahrmeir and Lang (2000).

3.2 Bayesian Inference Using MCMC Methods

Markov Chain Monte Carlo (MCMC) methods allow us to draw an arbitrary large number of joint samples from the posterior distribution $[\boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}, \phi, \tau^2, \sigma_c^2|\mathbf{Y}]$ approximately. With these samples we can easily make inference for these parameters using for example estimated posterior means or density estimates of the marginal posterior. Readers unfamiliar with MCMC methods can for example consult Casella and George (1992) for an introduction to the Gibbs sampler and Gilks, Richardson, and Spiegelhalter (1996) for applications of MCMC methods. We denote further the density of a random variable X by $[X]$ and the conditional density of X given Y by $[X|Y]$.

For the Bayesian approach, we assume independent prior distributions for the fixed effect $\boldsymbol{\alpha}$, the spatial parameters $\mathbf{b}$ given their dependence parameter ϕ and variance scalar τ^2, the cluster parameters $\mathbf{c}$ given their random variance σ_c^2 and the hyperparameters ϕ, τ^2, σ_c^2 respectively. In particular, we assume

$$[\boldsymbol{\alpha},\, \mathbf{b},\, \mathbf{c},\, \phi, \tau^2, \sigma_c^2] \;=\; [\boldsymbol{\alpha}] \times [\mathbf{b}|\phi, \tau^2] \times [\phi] \times [\tau^2] \times [\mathbf{c}|\sigma_c^2] \times [\sigma_c^2]\ .$$

3.2.1 Regression Parameter Update

The regression parameters $\boldsymbol{\alpha}$ can be updated jointly or individually which we discuss now. The fact that the full conditional distribution of $\boldsymbol{\alpha}$ is proportional to the product of the likelihood $[\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}]$ and the prior $[\boldsymbol{\alpha}]$ produces the following reduction for the joint regression parameter update:

$$\begin{aligned} [\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{b}, \mathbf{c}, \phi, \tau^2, \sigma_c^2] &= [\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{b}, \mathbf{c}] \propto [\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}] \times [\boldsymbol{\alpha}] \\ &\propto \prod_{i=1}^{n} \frac{\exp(Y_i(\mathbf{x_i}^t\boldsymbol{\alpha}))}{1+\exp(\mathbf{x_i}^t\boldsymbol{\alpha}+b_{j(i)}+c_{m(i)})}\,[\boldsymbol{\alpha}]\ . \end{aligned} \tag{3.3}$$

Similarly, the full conditional for an individual update of $\alpha_l,\ l = 1, \ldots, p$, has the following form:

$$[\alpha_l|\mathbf{Y}, \boldsymbol{\alpha}_{-l}, \mathbf{b}, \mathbf{c}, \phi, \tau^2, \sigma_c^2] \propto \prod_{i=1}^{n} \frac{\exp(Y_i(x_{il}\alpha_l))}{1+\exp(\mathbf{x_i}^t\boldsymbol{\alpha}+b_{j(i)}+c_{m(i)})}\,[\alpha_l]\ , \tag{3.4}$$

where $\boldsymbol{\alpha}_{-l} = (\alpha_1, \ldots, \alpha_{l-1}, \alpha_{l+1}, \ldots, \alpha_p)^t$. We used individual updates based on (3.4), since a good proposal density in the needed *Metropolis-Hastings (MH) step* is difficult to determine. As prior for $\alpha_l,\ l = 1, \ldots, p$ a normal distribution with zero-mean and large standard deviation was taken. Since both (3.3) and (3.4) do not represent any standard distribution, a direct Gibbs Step is therefore not available and an MH-Step is needed. As proposal density in the r^{th} iteration for α_l^r a normal density with the mean equal to the previous value α_l^{r-1} and a fixed value for the standard error was chosen. The value for this standard error was found using pilot runs. In particular, we used pilot runs to determine standard error values which give an acceptance rate between 30-60% (as proposed in

Bennett, Racine-Poon, and Wakefield (1996) or Besag, Green, Higdon, and Mengersen (1995)). The pilot runs also served as "burn in" phase.

3.2.2 Spatial Parameter Update

For the prior density of the vector $\mathbf{b}$ we use the modified Pettitt's conditional autoregression (2.9), which was introduced in Section 2. Since the computational effort for the joint update of $\mathbf{b}$ seems to be rapidly increasing with the dimension of $\mathbf{b}$, we use individual updates here as well. For these the full conditional densities are proportional to:

$$\begin{aligned}&[b_j|\mathbf{Y},\mathbf{b}_{-j},\boldsymbol{\alpha},\mathbf{c},\phi,\tau^2,\sigma_c^2] \propto [\mathbf{Y}|\boldsymbol{\alpha},\mathbf{b},\mathbf{c}]\times[b_j|\mathbf{b}_{-j},\boldsymbol{\alpha},\phi,\tau^2]\\ &\propto \prod_{i:j(i)=j}\frac{\exp(Y_i\,b_j)}{1+\exp(\mathbf{x_i}^t\boldsymbol{\alpha}+b_j+c_{m(i)})}\;\exp\left\{-\frac{1+|\phi|\,N_j}{2\,(1+|\phi|)\tau^2}\left(b_j-\frac{\phi}{1+|\phi|\,N_j}\sum_{j\sim j'}b_{j'}\right)^2\right\}\,,\end{aligned}\tag{3.5}$$

where $\mathbf{b}_{-j}=(b_1,\ldots,b_{j-1},b_{j+1},\ldots,b_J)^t$. Therefore, we also require an MH-Step for the spatial effects $\mathbf{b}$. As proposal density for b_j a similar proposal density as in the regression parameter update was used. We carried out a similar procedure with the pilot runs in order to find a good proposal standard error for each spatial parameter $b_j, j=1,\ldots,J$.

3.2.3 Spatial Dependence Parameter Update

The full conditional density for ϕ is given by

$$[\phi|\mathbf{Y},\boldsymbol{\alpha},\mathbf{b},\mathbf{c},\tau^2,\sigma_c^2]=[\phi|\mathbf{b},\tau^2]\propto[\mathbf{b}|\phi,\tau^2]\times[\phi]\,.\tag{3.6}$$

Since $\mathbf{b}|\phi,\tau^2 \sim N_J\left(0,\tau^2Q^{m.P}(\phi)^{-1}\right)$, the determinant of the matrix $Q^{m.P}(\phi)$ must be calculated for each $\phi-$iteration. We use the reparametrization $\psi=\frac{\phi}{1+|\phi|}$. Since once the eigenvalues of $\Gamma-D$ and $\Gamma+D$ are known, we can write the determinant of $Q^{m.P}(\phi)$ analytically as a function of ψ as in (2.12). Since $\psi\in(-1,1)$, it is reasonable to take a uniform distributed prior for ψ on $(-1,1)$. Such a prior choice corresponds to a heavy-tailed prior for ϕ, namely a Pareto distribution with the density $\sim\frac{1}{(1+|\phi|)^2}$. This density has no finite moments , but it is unimodal and symmetric with mode at 0. To generalize the link between the priors for ψ and ϕ we note that the prior for ψ proportional to $\frac{1}{(1-\psi)^{1-a}},\ \psi\in[0,1)$ corresponds to a prior for ϕ proportional to $\frac{1}{(1+\phi)^{1+a}},\ \phi\in[0,+\infty)$. It is a proper prior for $a>0$.

The last term in the conditional $[\phi|\mathbf{b},\tau^2]$ which depends on ψ is proportional to $\exp(-\frac{1}{2\tau^2}\mathbf{b}'Q^{m.P}(\psi)\mathbf{b})$. It is easy to see, that $\mathbf{b}'Q^{m.P}(\psi)\mathbf{b}$ can be written as a simple expression of ψ, namely

$$\mathbf{b}'Q^{m.P}(\psi)\mathbf{b}=\begin{cases}-\psi\mathbf{b}'(\Gamma-D)\mathbf{b}+\text{const}, & \text{if }\psi>0\\ \text{const}, & \text{if }\psi=0\\ -\psi\mathbf{b}'(\Gamma+D)\mathbf{b}+\text{const}, & \text{if }\psi<0\end{cases},$$

and therefore can be also calculated fast in each iteration.

As proposal distribution for the r^{th} iteration ψ^r we also take a normal density with mean equal to the previous iteration ψ^{r-1}, but now in contrast to the parameters $\boldsymbol{\alpha}$ and $\mathbf{b}$ truncated to the interval $(-1, 1)$. To determine a good proposal standard error we use a similar procedure with pilot runs as before. In addition, if the proposal standard error in the pilot runs is larger as 12, we exchange the normal distribution with a uniform distribution on the interval (-1,1) to reduce complexity.

3.2.4 Spatial Variance Parameter Update

For this update, an inverse gamma prior for τ^2 is used with density given by

$$[\tau^2] = \frac{1}{b_\tau^{(a_\tau)}\Gamma(a_\tau)(\tau^2)^{a_\tau+1}} \exp\left(-\frac{1}{b_\tau \tau^2}\right) , \tag{3.7}$$

where $a_\tau > 0$ and $b_\tau > 0$ are known hyperparameters. We denote this prior by $\tau^2 \sim\ IG(a_\tau, b_\tau)$. Since the full conditional for τ^2 depends solely on $\mathbf{b}$ and ϕ we can write

$$[\tau^2|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}, \phi, \sigma_c^2] = [\tau^2|\mathbf{b}, \phi] \propto [\mathbf{b}|\phi, \tau^2] \times [\tau^2] .$$

Substituting the expressions (2.11) and (3.7) into the last expression yields that the conditional distribution of $[\tau^2|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}, \phi, \sigma_c^2]$, is again $IG(a_\tau^*, b_\tau^*)$ with

$$a_\tau^* = a_\tau + \frac{J}{2} \text{ and } b_\tau^* = \left\{\frac{1}{b_\tau} + \frac{\mathbf{b}'\, Q(\phi)\, \mathbf{b}}{2}\right\}^{-1} .$$

For a less informative but proper prior choice, hyper parameters a_τ and b_τ are chosen in such a way that $IG(a_\tau, b_\tau)$ is widely dispersed. When a flat improper prior for τ^2 is used (as we have chosen) , the posterior $[\tau^2|\mathbf{b}, \phi]$ is $IG(a_\tau^*, b_\tau^*)$ with

$$a_\tau^* = \frac{J}{2} - 1 \text{ and } b_\tau^* = \left\{\frac{1}{2}\, \mathbf{b}'\, Q(\phi)\, \mathbf{b}\right\}^{-1} .$$

3.2.5 Cluster Parameter Update

As discussed before in the model description the cluster effect $\mathbf{c} = (c_1, \ldots, c_m)$ is taken as a random effect with prior $c_m \sim N(0, \sigma_c^2)$ for each cluster $m = 1, \ldots, M,\ i.i.d.$ Both joint and individual full conditionals for cluster parameters given the data $\mathbf{Y}$ and other parameters do not belong to any standard distribution. Therefore to update the cluster effects we also need a MH-step. As with $\mathbf{b}$ and $\boldsymbol{\alpha}$-updates, in order to avoid small acceptance rates we use individual updates here. The individual full conditionals can be written as follows

$$\begin{aligned} [c_m|\mathbf{Y}, \mathbf{c}_{-m}, \boldsymbol{\alpha}, \mathbf{b}, \phi, \tau^2, \sigma_c^2] &\propto [\mathbf{Y}|\boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}] \times [c_m|\sigma_c^2] \\ &\propto \prod_{i:m(i)=m} \frac{\exp(Y_i\, c_m)}{1+\exp(\mathbf{x_i}^t \boldsymbol{\alpha} + b_{j(i)} + c_m)}\ \exp\left\{-\tfrac{1}{2\,\sigma_c^2}\, c_m^2\right\} , \end{aligned} \tag{3.8}$$

where $\mathbf{c}_{-m} = (c_1, \ldots, c_{m-1}, c_{m+1}, \ldots, c_M)^t$. Again as proposal density for the r^{th} iteration c_m^r we chose a normal density with mean equal to c_m^{r-1} from the previous iteration. The proposal standard deviation was determined as for the $\boldsymbol{\alpha}$ or $\mathbf{b}$ parameters in pilot runs.

3.2.6 Cluster Variance Parameter Update

The full conditional density of the cluster variance parameter σ_c^2 has a similar form as the spatial variance parameter τ^2, namely:

$$[\sigma_c^2|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}, \phi, \tau^2] = [\sigma_c^2|\mathbf{c}] \propto [\mathbf{c}|\sigma_c^2] \times [\sigma_c^2] \propto \frac{1}{(\sigma_c^2)^{M/2}} \exp\left\{-\frac{1}{2\sigma_c^2}\,\mathbf{c}'\mathbf{c}\right\} [\sigma_c^2]\,. \qquad (3.9)$$

A direct Gibbs step is available by choosing as prior density an inverse gamma prior (see (3.7)) or an improper prior for σ_c^2. In particular, if $\sigma_c^2 \sim IG(a_c, b_c)$, then $\sigma_c^2|\mathbf{Y}, \boldsymbol{\alpha}, \mathbf{b}, \mathbf{c}, \phi, \tau^2 \sim IG(a_c^*, b_c^*)$ with

$$a_c^* = a_c + \frac{M}{2} \quad \text{and} \quad b_c^* = \left\{\frac{1}{b_c} + \frac{\mathbf{c}'\mathbf{c}}{2}\right\}^{-1}\,.$$

For an improper prior it follows that the full conditional density for σ_c^2 is a $IG(a_c^*, b_c^*)$ density with

$$a_c^* = \frac{M}{2} - 1 \quad \text{and} \quad b_c^* = \left\{\frac{1}{2}\,\mathbf{c}'\mathbf{c}\right\}^{-1}\,.$$

This density has a finite expectation for $M \geq 5$ and a finite variance for $M \geq 7$.

Chapter 4

Spatial Binary Regression with Individual Cluster Effects

4.1 Formulation of the Model

In this chapter we introduce for our binary transport data (3.1) a more advanced model where individual cluster effects are modeled by a normal distribution with fixed variance inside each cluster. In particular we assume for $i = 1, \ldots, n$:

$$\begin{aligned} &Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with} \\ &\theta_i := \log\left(\tfrac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i),k(i)}}_{\textit{random cluster effect}}, \end{aligned} \tag{4.1}$$

where for fixed $m = 1, \ldots, M$, $c_{m,k} \sim N(0, \sigma_m^2)$, $k = 1, \ldots, K_m$, *i.i.d.* As in Model (3.2), M denotes the number of clusters and $m(i)$ denotes the cluster of trip i. K_m stands for the number of trips, which belong to cluster m (i.e. $K_1 + \ldots + K_M = n$) and the index $k(i)$ gives the number of trip i in its cluster. The modeling of the fixed effects $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^t$ and the spatial effects $\mathbf{b} = (b_1, \ldots, b_J)^t$ remains as before. In contrast to (3.2), the cluster effects are now not the same for each trip in cluster m, namely c_m, but random realizations $c_{m,k},\ k = 1, \ldots, K_m$ from the same cluster distribution $N(0, \sigma_m^2)$. So we denote now the vector of cluster effects as $\mathbf{c} = (c_{1,1}, \ldots, c_{1,K_1}, \cdots, c_{M,1}, \ldots, c_{M,K_M})^t$. This allows for modeling heterogeneity within each cluster. In Model (4.1) we have to estimate in addition to the parameters $\boldsymbol{\alpha}, \mathbf{b}$ the cluster effect variances $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_M^2)^t$ instead of the cluster effects $\mathbf{c} = (c_1, \ldots, c_M)^t$ and their variance σ_c^2 for Model (3.2).

One problem with Model (4.1) is that even without an intercept term α_0 the model is unidentifiable. To understand the nature of the unidentifiability we first substitute in (4.1)

the logit link function with the probit link function, i.e. we assume for $i = 1, \ldots, n$:

$$\begin{aligned} &Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with} \\ &p_i = \mathbf{P}\{Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, c_{m(i),k(i)}\} = \Phi(\mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)} + c_{m(i),k(i)}), \end{aligned} \tag{4.2}$$

where $\Phi(\cdot)$ is the standard normal distribution function. We can interpret this representation of the success probabilities of binomial distributed values $Y_i,\ i = 1, \ldots, n$ through normally distributed latent variables $Z_i,\ i = 1, \ldots, n$ as follows

$$\begin{aligned} Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, c_{m(i),k(i)} \ &\Leftrightarrow\ Z_i \leq 0, \quad \text{where} \\ Z_i = -(\mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)} + c_{m(i),k(i)}) + \epsilon_i \ &, \epsilon_i \sim N(0,1)\ i.i.d. \end{aligned} \tag{4.3}$$

This is a similar representation as discussed in Albert and Chib (1993) for binary probit models. Since $c_{m(i),k(i)} \sim N(0, \sigma^2_{m(i)})$, we can also characterize the conditional distribution of Y_i given $x_i, \boldsymbol{\alpha}, b_{j(i)}, \sigma^2_{m(i)}$ by

$$\begin{aligned} Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, \sigma^2_{m(i)} \ &\Leftrightarrow\ Z_i \leq 0, \quad \text{where} \\ Z_i = -\eta_i + \epsilon_i^*,\ &\epsilon_i^* \sim N(0, 1 + \sigma^2_{m(i)}) \text{ independent and } \eta_i = \mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)}\ . \end{aligned} \tag{4.4}$$

Therefore we have for $i = 1, \ldots, n$

$$\mathbf{P}\{Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, \sigma^2_{m(i)}\} = \mathbf{P}\{Z_i \leq 0|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, \sigma^2_{m(i)}\} = \Phi\left(\frac{\mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)}}{\sqrt{1 + \sigma^2_{m(i)}}}\right). \tag{4.5}$$

Equation (4.5) shows that the parameters $\boldsymbol{\alpha}, \mathbf{b}$ and $\boldsymbol{\sigma^2}$ are not jointly identifiable in Model (4.2), since it is invariant with respect to the parameter vectors
$\left\{k \times (\boldsymbol{\alpha}^t, \mathbf{b}^t, \sqrt{1+\sigma_1^2}, \ldots, \sqrt{1+\sigma_M^2})^t,\ k \in \mathbb{R}\right\}$. If we define now

$$\boldsymbol{\alpha}' := \frac{\boldsymbol{\alpha}}{\sqrt{1+\sigma_1^2}}, \quad \mathbf{b}' := \frac{\mathbf{b}}{\sqrt{1+\sigma_1^2}}, \quad \sigma_m'^2 := \frac{1+\sigma_m^2}{1+\sigma_1^2},\ m = 2, \ldots, M,\ \sigma_1'^2 = 1, \tag{4.6}$$

then the marginal distributions (4.5) of $Y_i|\mathbf{x_i}, \boldsymbol{\alpha}, b_{j(i)}, \sigma^2_{m(i)}$ from Model (4.2) will coincide with the marginal distributions from the following model:

$$\begin{aligned} &Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with} \\ &p_i = \mathbf{P}\{Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}', b'_{j(i)}, \sigma'^2_{m(i)}\} = \begin{cases} \Phi\left(\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}\right) & \text{if } m(i) = 1 \\ \Phi\left(\frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right) & \text{if } m(i) = 2, \ldots, M\ . \end{cases} \end{aligned} \tag{4.7}$$

Using (4.4) it follows, that also the joint distribution of $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ in both Models (4.2) and (4.7) are equal, since

$$\begin{aligned} \mathbf{P}(\mathbf{Y} = \mathbf{y} \in \{0,1\}^n|\boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\sigma^2}) &= \mathbf{P}(\mathbf{Z} \in \prod_{i:Y_i=1}(-\infty, 0] \times \prod_{i:Y_i=0}(0, +\infty)|\boldsymbol{\alpha}, \mathbf{b}, \boldsymbol{\sigma^2}) \\ &\overset{\text{independence in(4.4)}}{=} \prod_{i:Y_i=1} \mathbf{P}(Z_i \in (-\infty, 0]|\boldsymbol{\alpha}, b_{j(i)}, \sigma_{m(i)}) \times \prod_{i:Y_i=0} \mathbf{P}(Z_i \in (0, +\infty)|\boldsymbol{\alpha}, b_{j(i)}, \sigma_{m(i)}) \\ &\overset{(4.5)}{=} \prod_{i:Y_i=1} \Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha} + b_{j(i)}}{\sqrt{1+\sigma^2_{m(i)}}}\right) \times \prod_{i:Y_i=0}\left(1 - \Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha} + b_{j(i)}}{\sqrt{1+\sigma^2_{m(i)}}}\right)\right) \overset{(4.7)}{=} \mathbf{P}(\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2\prime}})\ , \end{aligned}$$

where $\mathbf{Z} = (Z_1, \ldots, Z_n)^t$. Therefore Model (4.7) is an equivalent reparametrization of Model (4.2). But this representation (4.7) has one parameter less and is therefore identifiable. We applied the MCMC algorithm developed for Model (4.7) also to a simulated data set from the primary Model (4.2), since the joint distribution of (4.2) and (4.7) are equivalent. In this case the Gibbs sampler converged around the transformed true parameter values, namely $\boldsymbol{\alpha}' = \frac{\boldsymbol{\alpha}}{\sqrt{1+\sigma_1^2}}, \mathbf{b}' = \frac{\mathbf{b}}{\sqrt{1+\sigma_1^2}}$ and $\boldsymbol{\sigma^{2\prime}} = \frac{\boldsymbol{\sigma^2}}{\sqrt{1+\sigma_1^2}}$. Transformed hyperparameters of the spatial effect $\mathbf{b}$ are equal to $\tau^{2'} = \frac{\tau^2}{1+\sigma_1^2}$ and $\phi' = \phi$. So we showed that data arising from the quite natural but unfortunately unidentifiable Model (4.2) can be fitted by its identifiable reparametrization (4.7). Further, Model (4.7) can also be represented using normal latent variables:

$$\begin{aligned} &Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}', b'_{j(i)}, \sigma^{'2}_{m(i)} \;\Leftrightarrow\; Z'_i \leq 0, \quad \text{where} \\ &\quad Z'_i = -\eta'_i + \epsilon'_i,\ \epsilon'_i \sim N(0, \sigma^{'2}_{m(i)})\ ind. \quad \text{and} \quad \eta'_i = \mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}\ . \end{aligned} \tag{4.8}$$

As will be shown in this chapter, the representation of success probabilities $p_i, i = 1, \ldots, n$ via latent $Z_i, i = 1, \ldots, n$, allows us in the probit case to reduce significantly the number of variables, which need MH-updates in the MCMC algorithm. MH-updates are especially computational expensive for probit models because of the numerical complexity of the computations for $\Phi(\cdot)$ in the tails of the distribution.

Note that large values of cluster parameter $\sigma_m^{2'}$ indicate the large heterogeneity within cluster m. Such interpretation of cluster parameters follows from the reparametrization (4.6).

Another possibility to obtain identifiability in (4.2) would be still to use terms $c_{m(i),k(i)}$ (as in (4.2) and in (4.3)), but now with known variance $\sigma_1^{'2}$ (for example equal to 1, or equal to 0, i.e. $c_{m(i),k(i)} = 0$ for $m = 1$). But such an alternative has the disadvantage that not for all $\boldsymbol{\sigma^2}$-values from the model (4.2) exist corresponding values $\boldsymbol{\sigma^{2\prime}}$ to make the identifiable model equivalent to (4.2). So we can not avoid some additional limiting conditions for the vector $\boldsymbol{\sigma^2}$ from (4.2) in this case. For example, by setting in (4.2) $\sigma_1^{'2} = 1$, the other parameters $\boldsymbol{\alpha}', \mathbf{b}'$ and $\sigma_m^{2'}$, $m = 2, \ldots, M$, are determined from the following system, which would provide equivalence between (4.2) and this approach:

$$\begin{aligned} m = 1: &\quad \frac{\mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)}}{\sqrt{1+\sigma_1^2}} = \frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sqrt{1+1}} \\ m = 2, \ldots, M: &\quad \frac{\mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)}}{\sqrt{1+\sigma^2_{m(i)}}} = \frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sqrt{1+\sigma^{2'}_{m(i)}}}\ . \end{aligned}$$

From here it follows immediately that $\sigma_m^{2'} = 2\frac{1+\sigma_m^2}{1+\sigma_1^2} - 1,\ m = 2, \ldots, M$. However this means for specific values of σ_1^2 we might have $\sigma_m^{2'} < 0$, which violates the non negativity of a variance parameter. Similar requirements are needed to ensure positive variances in the case when $c_{1,k(i)} = 0$ ($\Leftrightarrow \sigma_1^2 = 0$), since then we need to define them as $\sigma_m^{2'} = \frac{1+\sigma_m^2}{1+\sigma_1^2} - 1,\ m = 2, \ldots, M$, which gives positive values only if $\sigma_1^2 = \min\limits_{m=1,\ldots,M}(\sigma_m^2)$.

The above discussion about probit model helps us also to understand the unidentifiability of our Logit Model (4.1), since the behavior of both probit and logit link functions is quite similar and becomes significantly different only in the tails. So we use the same idea as for the probit case to construct an identifiable logit model, which is approximately equivalent in distribution to (4.1). In particular we assume for $i = 1, \ldots, n$

$$\begin{aligned} &Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with} \\ &\log\left(\tfrac{p_i}{1-p_i}\right) = \begin{cases} \mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)} & \text{if } m(i) = 1 \\ \frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}} & \text{if } m(i) = 2, \ldots, M \end{cases}, \end{aligned} \tag{4.9}$$

where $p_i = \mathbf{P}\{Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}', b'_{j(i)}, \sigma'^2_{m(i)}\}$ and $\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2'}} := (\sigma_1^{2'}, \ldots, \sigma_M^{2'})^t$ are defined as in (4.6). The corresponding simulation study (see Figure 5.10) shows the convergence of the Gibbs sampler, which also indicates identifiability of Model (4.9).

Because the original model (4.1) is not identifiable we expect a MCMC algorithm based on (4.1), i.e. Markov chains $\{\boldsymbol{\alpha}^r\}, \{\mathbf{b}^r\}, \{\boldsymbol{\sigma^{2}}^r\}, \{\tau^{2r}\}$, $r = 1, 2, \ldots$ may not converge. However, with the parameter transformation given in (4.6) we achieve the identifiable Model (4.9) and therefore the transformed chains

$$\left\{\frac{\boldsymbol{\alpha}^r}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\mathbf{b}^r}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\boldsymbol{\sigma^{2}}^r}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\tau^{2r}}{1+\sigma_1^{2r}}\right\}, \{\phi^r\}, \quad r = 1, 2, \ldots \tag{4.10}$$

converge. The simulation study confirms this statement. These chains can be also used for the parameter estimation of $\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2'}}, \tau^{2'}, \phi'$ from the identifiable Model (4.9) as an alternative to constructing a MCMC algorithm directly based on Model (4.9). Generally, such alternative versus the common approach to simulate Markov Chains only from the identifiably reparametrized model might be preferable if the parameter updates in the unidentifiable model are simpler than in its identifiable analog.

We model our binary transport data (3.1) by both the Logit (4.9) and the Probit (4.7) Models. In the probit case we develop a Gibbs sampler procedure based on latent variable representation (4.8). An alternative MCMC algorithm uses characterization (4.7). In the next section we develop MCMC algorithms based on (4.9), (4.7) and (4.8), respectively.

4.2 Bayesian Inference for Hierarchical Spatial Binary Regression Models with Individual Cluster Effects

As pointed out before, we can not carry out Bayesian inference for both the primary Logit (4.1) or Probit (4.2) Hierarchical Spatial Binary Regression Models with Individual Cluster Effects, since they are unidentifiable. So for a Bayesian inference of data from (4.1) we use the approximately equivalent identifiable model (4.9) and similarly for data from (4.2) we use the equivalent identifiable model (4.7) or its latent variable representation (4.8).

4.2.1 Bayesian Inference for Logit Model (4.9)

We begin with the logit case (4.9) since its MCMC algorithm for the Bayesian inference is rather similar to the one described already in Section 3.2. Its hierarchical model structure is presented in Figure 4.1. From (4.9) it follows that the likelihood of the response vector

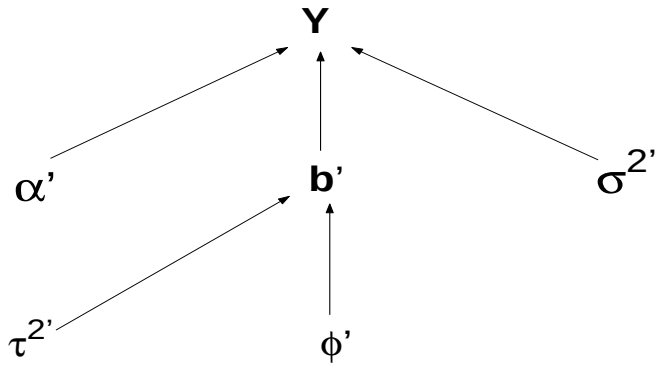

Figure 4.1: Hierarchical Model Structure for Model (4.9)

$\mathbf{Y}$ is proportional to

$$[\mathbf{Y}|\ \boldsymbol{\alpha}',\ \mathbf{b}',\ \boldsymbol{\sigma}'] \propto \prod_{i=1}^{n} \frac{\exp(Y_i \frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}})}{1 + \exp(Y_i \frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}})},$$

where $\boldsymbol{\sigma}' := (1, \sigma'_2, \ldots, \sigma'_M)^t,\ \sigma'_m := \sqrt{\sigma^{2'}_m},\ m = 2, \ldots, M$. We again assume independent prior distributions for the fixed effect $\boldsymbol{\alpha}'$, the spatial parameters $\mathbf{b}'$ given their dependence parameter ϕ' and the variance scalar $\tau^{2'}$ and the cluster parameters $\boldsymbol{\sigma}'$. Finally we

assume independence between the hyperparameters ϕ' and $\tau^{2'}$. Therefore the joint prior distribution is given by

$$[\boldsymbol{\alpha}',\, \mathbf{b}',\, \boldsymbol{\sigma}',\, \phi', \tau^{2'}] \;=\; [\boldsymbol{\alpha}'] \times [\mathbf{b}'|\phi', \tau^{2'}] \times [\phi'] \times [\tau^{2'}] \times [\boldsymbol{\sigma}'] \,.$$

The MCMC update procedure for the parameters $\boldsymbol{\alpha}'$ and $\mathbf{b}'$ remains the same as in Section 3.2, with the full conditionals replaced by

$$\begin{aligned}[\alpha_l'|\mathbf{Y}, \boldsymbol{\alpha}'_{-l}, \mathbf{b}', \boldsymbol{\sigma}', \phi', \tau^{2'}] \;&= [\alpha_l'|\mathbf{Y}, \boldsymbol{\alpha}'_{-l}, \mathbf{b}', \boldsymbol{\sigma}'] \\ &\propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}']\, [\alpha_l'] \quad \propto \prod_{i=1}^{n} \frac{\exp(Y_i \frac{x_{il}\alpha_l'}{\sigma'_{m(i)}})}{1+\exp(Y_i \frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}})}\, [\alpha_l']\end{aligned}$$

and

$$\begin{aligned}&[b_j'|\mathbf{Y}, \mathbf{b}'_{-j}, \boldsymbol{\alpha}', \boldsymbol{\sigma}', \phi', \tau^{2'}] \propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}']\, [b_j'|\mathbf{b}'_{-j}, \phi', \tau^{2'}] \\ &\propto \prod_{i:j(i)=j} \frac{\exp(Y_i \frac{b_j'}{\sigma'_{m(i)}})}{1+\exp(\frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b_j'}{\sigma'_{m(i)}})} \exp\left\{ -\frac{1+|\phi'|\, N_j}{2\,(1+|\phi'|)\tau^{2'}} \left(b_j' - \frac{\phi'}{1+|\phi'|\, N_j} \sum_{j\sim j'} b_{j'}' \right)^2 \right\}.\end{aligned}$$

Since the full conditionals of the spatial hyperparameters $\tau^{2'}$ and ϕ' given the data and other parameters depend only on the spatial effects $\mathbf{b}'$, their MCMC updates have the same form as described in Section 3.2. We update the parameter $\boldsymbol{\sigma}'$ in a similar way as the other main parameters $\boldsymbol{\alpha}'$ and $\mathbf{b}'$. The full conditional of $\sigma_m',\ m = 2, \ldots, M$, can be written as

$$\begin{aligned}[\sigma_m'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}'_{-m}, \phi', \tau^{2'}] \;&= [\sigma_m'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}'_{-m}] \\ &\propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}']\, [\sigma_m'] \quad \propto \prod_{i:m(i)=m}^{n} \frac{\exp(Y_i \frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b'_{j(i)}}{\sigma_m'})}{1+\exp(Y_i \frac{\mathbf{x_i}^t \boldsymbol{\alpha}' + b'_{j(i)}}{\sigma_m'})}\, [\sigma_m'] \,.\end{aligned}$$

It is reasonable to take a prior for $\sigma_m',\ m = 2, \ldots, M$, which is distributed around 1. Such a choice is clear from the reparametrization (4.6). According to (4.6) large deviations from 1 for some $\sigma_m',\ m = 2, \ldots, M$, correspond to large values for some $\sigma_m^2,\ m = 1, \ldots, M$, in the primary model (4.1), what would correspond to insignificance of the regression and spatial effects in these clusters. In our study we use a normal distribution $N(1, 1)$ truncated on the interval $[0.2, +\infty)$ as prior for $\sigma_m',\ m = 2, \ldots, M$.

4.2.2 Bayesian Inference for Probit Model (4.7)

We continue this section with the Bayes inference for the probit model. We consider first its explicit presentation (4.7), since its hierarchical structure is the same, as for the logit model (see Figure 4.1). The sampling is carried out for the same parameters, namely for

$\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}'$ and the hyperparameters $\phi', \tau^{2'}$. As joint prior distribution we assume the same prior as in the logit case (4.9). The only one difference is that now the likelihood of the response vector $\mathbf{Y}$ can be expressed using the probit function instead of the logit one:

$$[\mathbf{Y}|\ \boldsymbol{\alpha}',\ \mathbf{b}',\ \boldsymbol{\sigma}'] \propto \prod_{i:Y_i=1}^{n} \Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right) \times \prod_{i:Y_i=0}^{n} \left(1 - \Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right)\right) .$$

Therefore the full conditionals for $\boldsymbol{\alpha}', \mathbf{b}'$ and $\boldsymbol{\sigma}'$ are given by

$$\begin{aligned}
[\alpha'_l|\mathbf{Y}, \boldsymbol{\alpha}'_{-l}, \mathbf{b}', \boldsymbol{\sigma}', \phi', \tau^{2'}] &= [\alpha'_l|\mathbf{Y}, \boldsymbol{\alpha}'_{-l}, \mathbf{b}', \boldsymbol{\sigma}'] \propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}']\ [\alpha'_l] \\
&\propto \prod_{i:Y_i=1}^{n} \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right) \times \prod_{i:Y_i=0}^{n} \left(1 - \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right)\right) \times [\alpha'_l] \ ,
\end{aligned}$$

$$\begin{aligned}
[b'_j|\mathbf{Y}, \mathbf{b}'_{-j}, \boldsymbol{\alpha}', \boldsymbol{\sigma}^{\mathbf{2}'}, \phi', \tau^{2'}] &\propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}]\ [b'_j|\mathbf{b}'_{-j}, \phi', \tau^{2'}] \\
&\propto \prod_{\substack{i:\\ Y_i=1}}^{\substack{j(i)=j \\ n}} \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_j}{\sigma'_{m(i)}}\right) \times \prod_{\substack{i:\\ Y_i=0}}^{\substack{j(i)=j \\ n}} \left(1 - \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_j}{\sigma'_{m(i)}}\right)\right) \\
&\times \exp\left\{-\tfrac{1+|\phi'|\,N_j}{2\,(1+|\phi'|)\tau^{2'}} \left(b'_j - \tfrac{\phi'}{1+|\phi'|\,N_j} \sum_{j\sim j'} b'_{j'}\right)^2\right\}
\end{aligned}$$

and

$$\begin{aligned}
[\sigma'_m|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}{}_{-m}, \phi', \tau^{2'}] &= [\sigma'_m|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}{}_{-m}] \propto [\mathbf{Y}|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}]\ [\sigma'_m] \\
&\propto \prod_{\substack{i:\\ Y_i=1}}^{\substack{m(i)=m \\ n}} \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right) \times \prod_{\substack{i:\\ Y_i=0}}^{\substack{m(i)=m \\ n}} \left(1 - \Phi\left(\tfrac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right)\right) \times [\sigma'_m]
\end{aligned}$$

respectively. The full conditionals for the spatial hyperparameters ϕ' and $\tau^{2'}$ remain the same, as for the Logit Model (4.9), since they do not depend on the likelihood $[\mathbf{Y}|\ \boldsymbol{\alpha}',\ \mathbf{b}',\ \boldsymbol{\sigma}^{\mathbf{2}'}]$.

As before, the MH-step is needed for the parameters $\mathbf{b}', \boldsymbol{\alpha}', \boldsymbol{\sigma}^{\mathbf{2}'}$ and ϕ'. But in contrast to the logit case the MH-update is much more computational expensive for probit models. It is slow and less precise because of the numerical complexity of the computations for the $\Phi(\cdot)$ values. The second problem arises by the computation of $\Phi(\cdot)$ for large values, since this distribution function converges to 1 much faster, than the heavy-tailed logit distribution. For example, MATLAB is not able to distinguish $(1 - \Phi(x))$ from 0 already for $x > 8$, which corresponds to a 100%-relative error. As will be shown below, such great errors are not tolerable for the MH-procedure. To overcome this problem arising in the MCMC-inference for the explicit representation (4.7) we need efficient MH-steps for parameters, whose full conditionals include $\Phi(\cdot)$-terms, namely for $\mathbf{b}', \boldsymbol{\alpha}'$ and $\boldsymbol{\sigma}^{\mathbf{2}'}$. Let us consider the acceptance probability $a(\cdot, \cdot)$ in the MH-step for the regression parameter $\boldsymbol{\alpha}'$. Let $\pi(\alpha'_l)$ denote the full conditional of α'_l, $l = 1, \ldots, p$, and let $\alpha_l^{'p}$ the proposal value for α'_l and $\boldsymbol{\alpha}^{'p} := (\alpha'_1, \ldots, \alpha'_{l-1}, \alpha_l^{'p}, \alpha'_{l+1}, \ldots, \alpha'_p)^t$. Under the condition, that the transition

probability is homogeneous, the acceptance probability $a(\alpha'_l, \alpha_l^{'p})$ is given by:

$$a(\alpha'_l, \alpha_l^{'p}) = \min(1, \tfrac{\pi(\alpha_l^{'p})}{\pi(\alpha'_l)}),$$

where

$$\begin{aligned}
\ln \tfrac{\pi(\alpha_l^{'p})}{\pi(\alpha'_l)} &= \sum_{i:Y_i=1} \ln \frac{\Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}'+b'_{j(i)}}{\sigma'_{m(i)}} + \frac{x_{il}(\alpha_l^{'p}-\alpha'_l)}{\sigma'_{m(i)}}\right)}{\Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}'+b'_{j(i)}}{\sigma'_{m(i)}}\right)} \\
&\quad + \sum_{i:Y_i=0} \ln \frac{1-\Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}'+b'_{j(i)}}{\sigma'_{m(i)}} + \frac{x_{il}(\alpha_l^{'p}-\alpha'_l)}{\sigma'_{m(i)}}\right)}{1-\Phi\left(\frac{\mathbf{x_i}^t\boldsymbol{\alpha}'+b'_{j(i)}}{\sigma'_{m(i)}}\right)} + \ln \tfrac{[\alpha_l^{'p}]}{[\alpha'_l]} \\
&= \sum_{i:Y_i=1} \ln \tfrac{\Phi(\nu_i^p)}{\Phi(\nu_i)} + \sum_{i:Y_i=0} \ln \tfrac{1-\Phi(\nu_i^p)}{1-\Phi(\nu_i)} + \ln \tfrac{[\alpha_l^{'p}]}{[\alpha'_l]} \ .
\end{aligned}$$

Here we have set

$$\nu_i := \frac{\mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}} \quad \text{and} \quad \nu_i^p := \frac{\mathbf{x_i}^t\boldsymbol{\alpha}^{'p} + b'_{j(i)}}{\sigma'_{m(i)}} = \nu_i + \frac{x_{il}(\alpha_l^{'p} - \alpha'_l)}{\sigma'_{m(i)}} \ .$$

Therefore ratios like $\frac{\Phi(\nu_i^p)}{\Phi(\nu_i)}$ or $\frac{1-\Phi(\nu_i^p)}{1-\Phi(\nu_i)}$ must be calculated with a small relative error. Since the computation of the function $\Phi(\cdot)$ for small to medium sized arguments is quite precise, the simplest way to accelerate it would be to split the real axis into a sufficient number of intervals N and to approximate $\Phi(x)$ with one of the corresponding N values of Φ. These would be computed once at the beginning. For large absolute values of $\nu_i, \nu_i^p > 7$ we use the following approximations:

$$\lim_{x\to+\infty} \frac{1-\Phi(x)}{\frac{\phi(x)}{x}} = 1 \quad \text{and} \quad \lim_{x\to-\infty} \frac{\Phi(x)}{\frac{\phi(x)}{|x|}} = 1 \ , \tag{4.11}$$

which can be proved by using the theorem of L'Hospital. In Figure 4.2 we show the convergence picture for $x \geq 1$, where MATLAB is able to compute a non zero value for $\Phi(x)$ and $\phi(x) = \Phi'(x)$. With this approximation the logarithm $\ln\left(\frac{1-\Phi(\nu_i^p)}{1-\Phi(\nu_i)}\right)$ for the acceptance probability $a(\alpha'_l, \alpha_l^{'p})$ requires for $\nu_i, \nu_i^p > 7$ the simple form $\left(\frac{\nu_i^2}{2} - \frac{\nu_i^{p2}}{2}\right) + \ln \frac{\nu_i}{\nu_i^p}$, so we do not need to compute the function $\Phi(\cdot)$ for large positive arguments.

When we investigated the behavior of MATLAB for negative values, we were surprised to find out that MATLAB is able to compute function $\Phi(x)$ also for large negative values of x, namely for $x > -37$. Figure 4.3 shows that this computation is precise, since for the function $\phi(x)/|x|$ MATLAB provides a precise calculation in this interval. Therefore for values between $7 < x < 37$ we can calculate $\Phi(x)$ more precisely by $\Phi(x) = 1 - \Phi(-x)$, which is however slower as the approximative calculation (4.11).

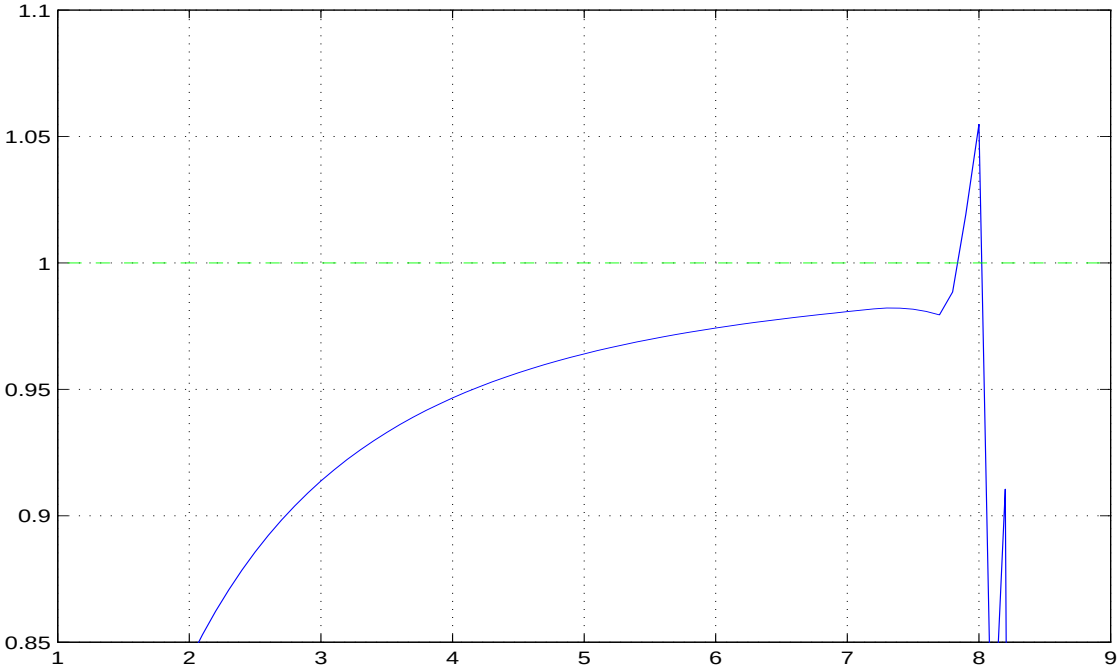

Figure 4.2: Values Calculated by MATLAB for the Ratio $\frac{1-\Phi(x)}{\frac{\phi(x)}{x}}$ for $x \geq 1$

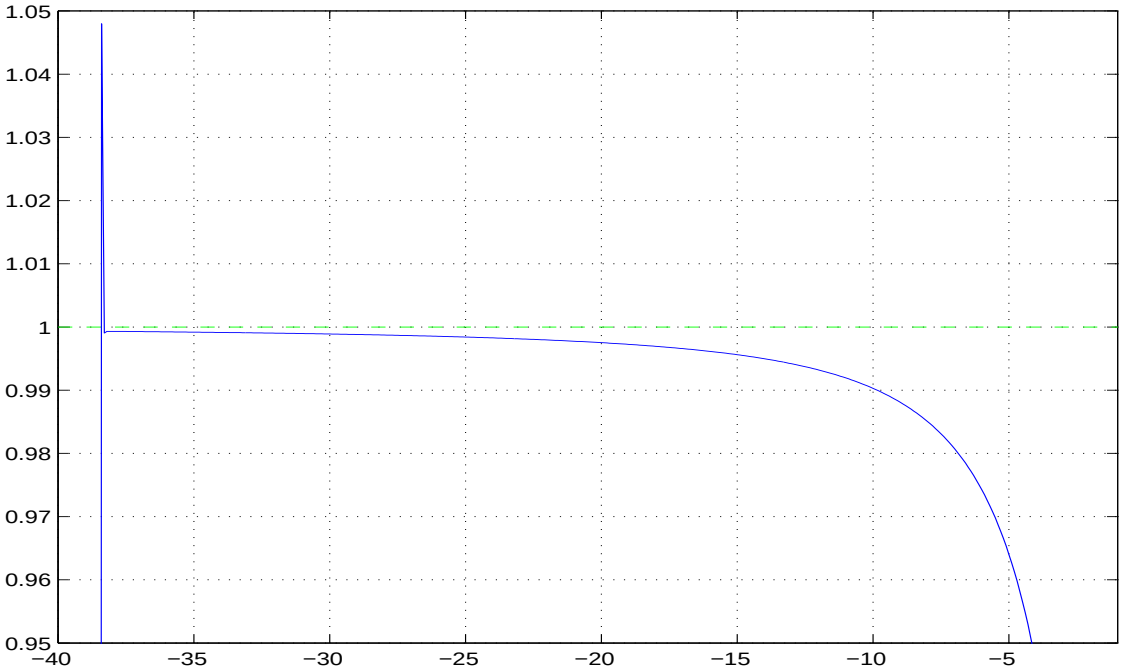

Figure 4.3: Values Calculated by MATLAB for the Ratio $\frac{\Phi(x)}{\frac{\phi(x)}{|x|}}$ for $x \leq -5$

4.2.3 Bayesian Inference for Probit Model (4.7) Based on Representation (4.8)

We consider finally the Bayes inference for the Probit Model (4.7) based on its latent variable representation (4.8). In the hierarchical structure of (4.8) (see Figure 4.4) the

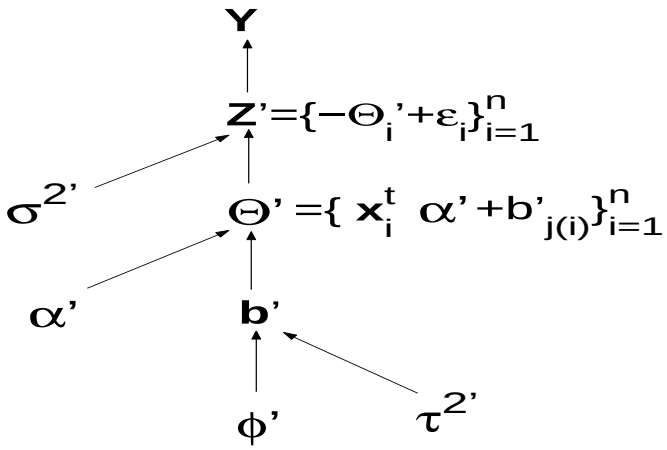

Figure 4.4: Hierarchical Model Structure for Model (4.8)

latent variables Z_i''s bridge the data $\mathbf{Y}$ and the model parameters $\boldsymbol{\alpha}', \mathbf{b}'$ and $\boldsymbol{\sigma}^{\mathbf{2}'}$. The full conditionals of these parameters do not depend on the binary vector $\mathbf{Y}$ given the vector $\mathbf{Z}' = (Z_1', \ldots, Z_n')^t$ and therefore represent some standard distributions, which do not need computational expensive MH-steps. Moreover, in contrast to the previous models joint updates for the parameter vectors $\boldsymbol{\alpha}'$ and $\mathbf{b}'$ are available here. Further, the full conditional of the latent vector $\mathbf{Z}'$ given the data $\mathbf{Y}$ and all the parameters has also a simple form, which is suitable for the direct joint updating. More precisely, since the latent variables Z_i''s are conditionally independent given $\eta_i', i = 1, \ldots, n$, we can immediately reduce the joint update of $[\mathbf{Z}'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}, \phi', \tau^{2'}]$ to the individual updates of $[Z_i'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}, \phi', \tau^{2'}]$ for $i = 1, \ldots, n$. Each of these univariate conditional distributions is equivalent to $[Z_i'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}]$, since given $\mathbf{b}'$ the information contained in ϕ' and $\tau^{2'}$ has no influence on $\mathbf{Z}'$. Moreover, we have $[Z_i'|\mathbf{Y}, \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}'}] = [Z_i'|Y_i, \boldsymbol{\alpha}', b'_{j(i)}, \sigma^{2'}_{m(i)}], i = 1, \ldots, n$, due again to the conditional independence. It is easy to see that these distributions are univariate truncated normal with mean $-\mathbf{x_i}^t\boldsymbol{\alpha}' - b'_{j(i)}$ and variance $\sigma^{2'}_{m(i)}$. The truncation interval is $(-\infty, 0]$ (or $[0, \infty)$) when $Y_i = 1$ (or $Y_i = 0$). We use rejection sampling for the generation of truncated univariate normal random variables in the numerical implementation as proposed by Robert (1995).

Now let us proceed with the parameter updates. To present the joint updates for the parameter vectors $\boldsymbol{\alpha}'$ and $\mathbf{b}'$ we need some matrix notations. So we define a design

matrix X of $p \times n$ dimension as $X = (\mathbf{x_1}, \ldots, \mathbf{x_n})^t$ and assume a full column rank. Further we define the matrix $B = (b_{ij})$ as a $n \times J$ spatial incidence matrix with $b_{ij} = 1$, if $j(i) = j$ and $b_{ij} = 0$, if $j(i) \neq j$. This implies that $B \cdot \mathbf{b'} = (b'_{j(1)}, \ldots, b'_{j(n)})^t$. Finally let $\Sigma := \mathrm{cov}(\mathbf{Z'}|\boldsymbol{\alpha'}, \mathbf{b'}, \boldsymbol{\sigma^{2'}}) = \mathrm{diag}(\sigma^{2'}_{m(1)}, \ldots, \sigma^{2'}_{m(n)})$.

Let us consider first the regression parameter $\boldsymbol{\alpha'}$ update. Under prior $N_p(\boldsymbol{\mu}_0, \Sigma_0)$ for $\boldsymbol{\alpha'}$ we immediately obtain, that its full conditional is given by

$$
\begin{aligned}
[\boldsymbol{\alpha'}|\mathbf{Z}, \mathbf{b'}, \boldsymbol{\sigma^{2'}}] &\propto [\mathbf{Z'}|\boldsymbol{\alpha'}, \mathbf{b'}, \boldsymbol{\sigma^{2'}}]\,[\boldsymbol{\alpha'}] \\
&\propto \exp\left\{-\tfrac{1}{2}(\mathbf{Z'} + X^t\boldsymbol{\alpha'} + B\mathbf{b'})^t\Sigma^{-1}(\mathbf{Z'} + X^t\boldsymbol{\alpha'} + B\mathbf{b'})\right\} \times [\boldsymbol{\alpha'}] \\
&\propto \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\alpha}'^t X\Sigma^{-1}X^t\boldsymbol{\alpha'} + 2\boldsymbol{\alpha}'^t X\Sigma^{-1}(\mathbf{Z'} + B\mathbf{b'}) + (\boldsymbol{\alpha'} - \mu_0)^t\Sigma_0^{-1}(\boldsymbol{\alpha'} - \mu_0)\right)\right\} \\
&\propto \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\alpha}'^t \underbrace{\left(X\Sigma^{-1}X^t + \Sigma_0^{-1}\right)}_{\Sigma^{-1}_{\boldsymbol{\alpha'}}}\boldsymbol{\alpha'} + 2\boldsymbol{\alpha}'^t \underbrace{\left(X\Sigma^{-1}(\mathbf{Z} + B\mathbf{b'}) - \Sigma_0^{-1}\boldsymbol{\mu}_0\right)}_{-\Sigma^{-1}_{\boldsymbol{\alpha'}}\boldsymbol{\mu}_{\boldsymbol{\alpha'}}}\right)\right\}.
\end{aligned}
$$

Therefore $\boldsymbol{\alpha'}|\mathbf{Z'}, \mathbf{b'}, \boldsymbol{\sigma^{2'}} \sim N_p(\boldsymbol{\mu}_{\boldsymbol{\alpha'}}, \Sigma_{\boldsymbol{\alpha'}})$ with

$$
\begin{aligned}
\Sigma_{\boldsymbol{\alpha'}} &= \left(X\Sigma^{-1}X^t + \Sigma_0^{-1}\right)^{-1} \quad \text{and} \\
\boldsymbol{\mu}_{\boldsymbol{\alpha'}} &= -\Sigma_{\boldsymbol{\alpha'}}\left(X\Sigma^{-1}(\mathbf{Z'} + B\mathbf{b'}) - \Sigma_0^{-1}\boldsymbol{\mu}_0\right).
\end{aligned}
\tag{4.12}
$$

For a flat improper prior of $\boldsymbol{\alpha'}$ (4.12) can be simplified by replacing the parameters $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\Sigma_0^{-1} = 0$, which gives also a proper distribution.

For the spatial parameter vector $\mathbf{b'}$ its joint full conditional $[\mathbf{b'}|\mathbf{Y}, \mathbf{Z'}, \boldsymbol{\alpha'}, \boldsymbol{\sigma^{2'}}, \phi', \tau^{2'}] = [\mathbf{b'}|\mathbf{Z'}, \boldsymbol{\alpha'}, \boldsymbol{\sigma^{2'}}, \phi', \tau^{2'}]$ can be found in a similar way as for $\boldsymbol{\alpha'}$. Under the J-variate normal $N_J\left(\mathbf{0}, \left(Q^{m.P}(\phi')\right)^{-1}\right)$ prior for $[\mathbf{b'}|\phi', \tau^{2'}]$, the full conditional $[\mathbf{b'}|\mathbf{Z'}, \boldsymbol{\alpha'}, \boldsymbol{\sigma^{2'}}, \phi', \tau^{2'}] \propto [\mathbf{Z'}|\boldsymbol{\alpha'}, \mathbf{b'}, \boldsymbol{\sigma^{2'}}]\,[\mathbf{b'}|\phi', \tau^{2'}]$ is also J-variate normal $N_J(\boldsymbol{\mu}_{\mathbf{b'}}, \Sigma_{\mathbf{b'}})$ with

$$
\begin{aligned}
\Sigma_{\mathbf{b'}} &= \left(B^t\Sigma^{-1}B + \tfrac{1}{\tau^{2'}}Q^{m.P}(\phi')\right)^{-1} \quad \text{and} \\
\boldsymbol{\mu}_{\mathbf{b'}} &= -\Sigma_{\mathbf{b'}}B^t\Sigma^{-1}(\mathbf{Z'} + X^t\boldsymbol{\alpha'}).
\end{aligned}
\tag{4.13}
$$

Note that in (4.13) for each update we need to invert the $J \times J$-dimensional precision matrix of $\mathbf{b'}$, what may be computationally very expensive, if the number of regions J is large. Since the band structure of the precision matrix $\Sigma^{-1}_{\mathbf{b'}} = B^t\Sigma^{-1}B + \frac{1}{\tau^{2'}}Q^{m.P}(\phi')$ coincides with the band structure of Q (note that $B^t\Sigma^{-1}B$ is a diagonal matrix) and is therefore known and stable, it might be more efficient to simulate a J-variate normal $N_J(\boldsymbol{\mu}_{\mathbf{b'}}, \Sigma_{\mathbf{b'}})$ vector using the *Cholesky decomposition* of the matrix $\Sigma_{\mathbf{b'}}$, which requires an order J^2 operation. It is significantly cheaper for large J, than the inversion of the matrix $\Sigma_{\mathbf{b'}}$ (an order J^3 operation) and more stable (see Gelfand, Ravishanker, and Ecker (2000), p.380). For a reduction of the computational effort it may be reasonable to reorder the vector $\mathbf{b'}$ in a way, which provides minimal bandwidth for its precision matrix. One method to facilitate this is the Cuthill-McKee Algorithm (George and Liu 1981). It is implemented

in MATLAB by the function *symrcm*. More precisely, the simulation algorithm using Cholesky's decomposition is therefore given as follows:

1. Find once at the beginning the order of $\mathbf{b}'$, which provides minimal bandwidth for its reordered prior precision matrix $Q_{re}^{m.P}$ and therefore for the reordered posterior precision matrix $\Sigma_{\mathbf{b}'_{re}}^{-1}$. Note that the subscript $_{re}$ indicates that the corresponding matrix is reordered.

2. Determine the Cholesky decomposition $L'L = \Sigma_{\mathbf{b}'_{re}}^{-1}$, where L is an upper triangular matrix.

3. Generate a J-variate normal $N_J\left(0, \Sigma_{\mathbf{b}'_{re}}\right)$ random vector Y by solving $LY = Z$, where $Z \sim N_J(0, I_J)$. Note that this equation can be solved fast by the Gauss method, since the matrix L is triangular. Therefore we do not need to invert the matrix L.

4. Solve equation $L\eta = -[B^t\Sigma^{-1}(\mathbf{Z}' + X^t\boldsymbol{\alpha}')]_{re}$ in a similar way as in Step 3.

5. Solve equation $L'\boldsymbol{\mu}_{\mathbf{b}'_{re}} = \eta$.

6. Finally determine the vector $\mathbf{b}_{re} = \boldsymbol{\mu}_{\mathbf{b}'_{re}} + Y$ which is a realization of a J-variate normal $N_J\left(\boldsymbol{\mu}_{\mathbf{b}'_{re}}, \Sigma_{\mathbf{b}'_{re}}\right)$.

7. Reorder the components of $\mathbf{b}_{re}$ back to achieve $\mathbf{b}$.

In contrast to Model (4.9) we update here the cluster variance parameters $\sigma_m^{2'}$, $m = 2, \ldots, M$, instead of σ'_m, $m = 2, \ldots, M$, since it can be updated directly under a suitable choice of prior. More precisely, the individual full conditionals for $\sigma_m^{2'}$, $m = 2, \ldots, M$, are given by

$$\begin{aligned}[\sigma_m^{2'}|\mathbf{Y}, \mathbf{Z}', \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2'}}_{-m}, \phi', \tau^{2'}] &= [\sigma_m^{2'}|\mathbf{Z}', \boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2'}}_{-m}] \\ \propto [\mathbf{Z}'|\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma^{2'}}]\,[\sigma_m^{2'}] &\propto \frac{1}{(\sigma_m^{2'})^{\frac{K_m}{2}}}\ \exp\left\{-\frac{1}{2\sigma_m^{2'}}\sum_{i:m(i)=m}(Z'_i + \mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)})^2\right\} \times [\sigma_m^{2'}],\end{aligned} \tag{4.14}$$

where $\boldsymbol{\sigma^{2'}}_{-m}$=$(1, \sigma_2^{2'}, \ldots, \sigma_{m-1}^{2'}, \sigma_{m+1}^{2'}, \sigma_M^{2'})^t$. If $[\sigma_m^{2'}]$ is $IG(a_0, b_0)$ (see (3.7)), we immediately obtain in (4.14) the $IG(a_{\sigma'}, b_{\sigma'})$ density function, up to a constant, where

$$\begin{aligned} a_{\sigma'} &= \tfrac{K_m}{2} + a_0 \quad \text{and} \\ b_{\sigma'} &= \left\{\frac{1}{b_0} + \frac{\sum_{i:m(i)=m}(Z'_i + \mathbf{x_i}^t\boldsymbol{\alpha}' + b'_{j(i)})^2}{2}\right\}^{-1}. \end{aligned} \tag{4.15}$$

As mentioned before in this section, it is reasonable to choose a prior for $\sigma_m^{2'}, m = 2, \ldots, M$, which is distributed around 1. Therefore we took $[\sigma_m^{2'}] \sim IG(3, 0.5)$ for $m = 2, \ldots, M$. This gives $\mathbf{E}(\sigma_m^{2'}) = \mathbf{Var}(\sigma_m^{2'}) = 1$.

We note finally, that the update of the spatial hyperparameters ϕ' and $\tau^{2'}$ remains the same as described in Section 3.2 for Models (3.2) and (4.9). So we have only one parameter, namely ϕ', that needs a MH-step.

Chapter 5

Simulation Studies

To investigate the performance of the MCMC algorithms for the proposed models (3.2), (4.9) and (4.7) we conducted several simulation studies. We chose one fixed parameter setting per model. In order to get an idea about random variability, for each model we simulated 4 data sets, which provided us 4 posterior estimates for the corresponding parameter vector. The simulated data sets have a similar sample size and structure as the mobility data. Further purposes of these simulation studies are to evaluate (1) the autocorrelation structure of the generated Markov Chains to assess the mixing of the MCMC algorithm, and (2) the length of burn-in and other convergence issues. Below we present the simulation study for each of the three models (3.2), (4.9) and (4.7).

5.1 Study 1: Hierarchical Spatial Binary Regression with Group Cluster Effects

5.1.1 Setup

The simulation study based on the Logit Model (3.2) has the following mean structure:

$$\Theta_i := \log\left(\frac{p_i}{1-p_i}\right) = x_{1i}\alpha_1 + x_{2i}\alpha_2 + b_{j(i)} + c_{m(i)}, \quad i = 1,\ldots,n,\ j = 1,\ldots,J,\ m = 1,\ldots,M.$$

We simulated $n = 2100$ binary responses residing in $J = 70$ regions arranged on a 7×10 regular lattice (i.e. 30 observations per region) and in $M = 5$ clusters (i.e. 420 observations per cluster) so, that each cluster is represented in each region with $30/5 = 6$ responses. The number of responses, the number of regions and clusters approximately corresponds to our real mobility data. The regression effect is simulated identically for each of the 4 data sets. More precisely, we chose x_{i1} as categorical covariate with possible values 0 or 1 (so that in each cluster×region cell both values are represented with $6/2 = 3$

observations) and x_{i2} as continuous covariate taking cycled integer values between 1 and 23, i.e. $x_{12} = 1, x_{22} = 2, \ldots, x_{23,2} = 23, x_{24,2} = 1, \ldots$. With this choice we achieved a good data mixing inside both regions and clusters. The true values for the regression parameters were taken as $\alpha_1 = -1$ and $\alpha_2 = 0.05$. A graphical representation of the regression effect $x_{1i}\alpha_1 + x_{2i}\alpha_2$ with regard to each cluster×region cell is given in Figure 5.1. We model the

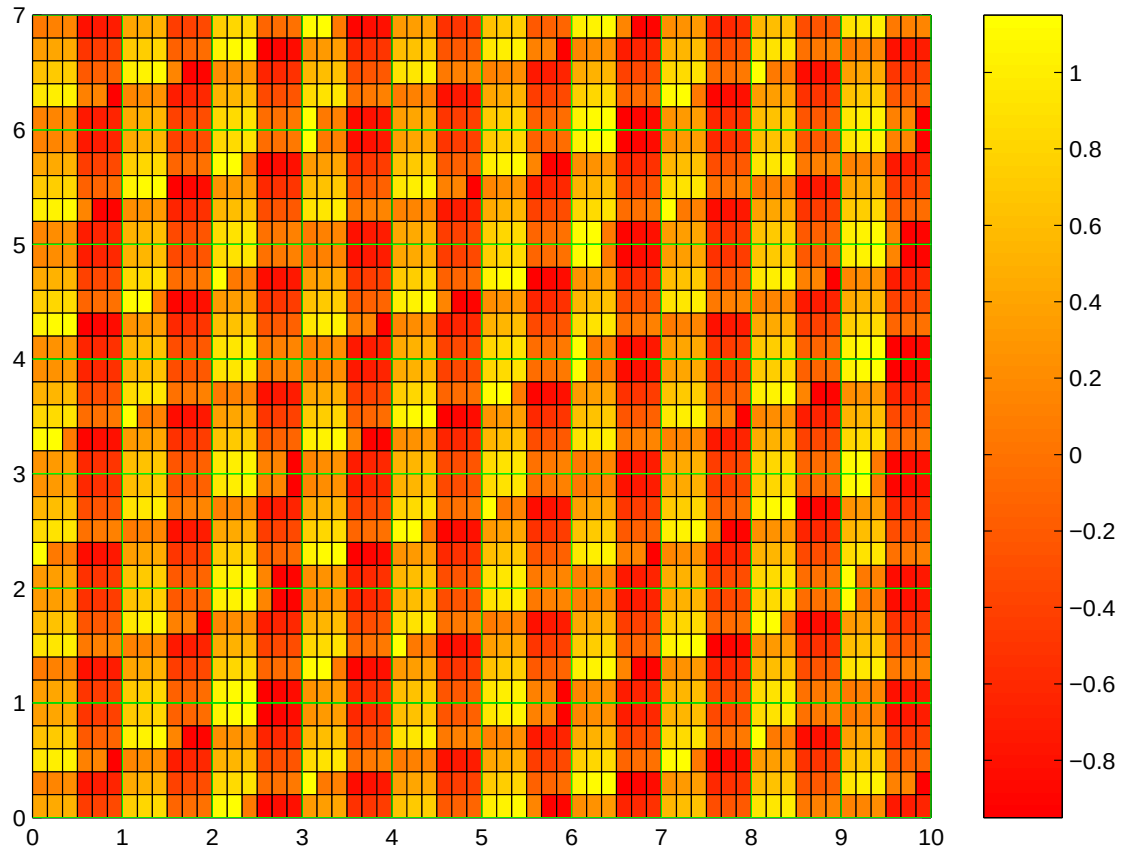

Figure 5.1: Regression Effects for Data Sets Based on Model (3.2)

spatial effects $b'_j,\ j = 1, \ldots, 70$, from the modified Pettitt's Model (2.9). For the spatial hyperparameters the values $\phi = 25$ and $\tau^2 = 0.64$ were chosen. With such a relatively large value for ϕ a strong spatial smoothing effect is presented, since (2.9) has according to (2.7) the intrinsic Gaussian CAR (2.4) in the limit, when $\phi \to \infty$. We chose τ^2 in such a way that the range of the observed spatial effects is somewhat similar to the range of the regression effects. For the neighborhood structure we chose the regular lattice 7×10 with a first order neighborhood dependence, i.e. the neighborhood of each region consists from the 2-4 regions which have a joint border. The simulated spatial effects for each of the 4 Data Sets are presented in Figure 5.2. Finally, for each data set we simulated group

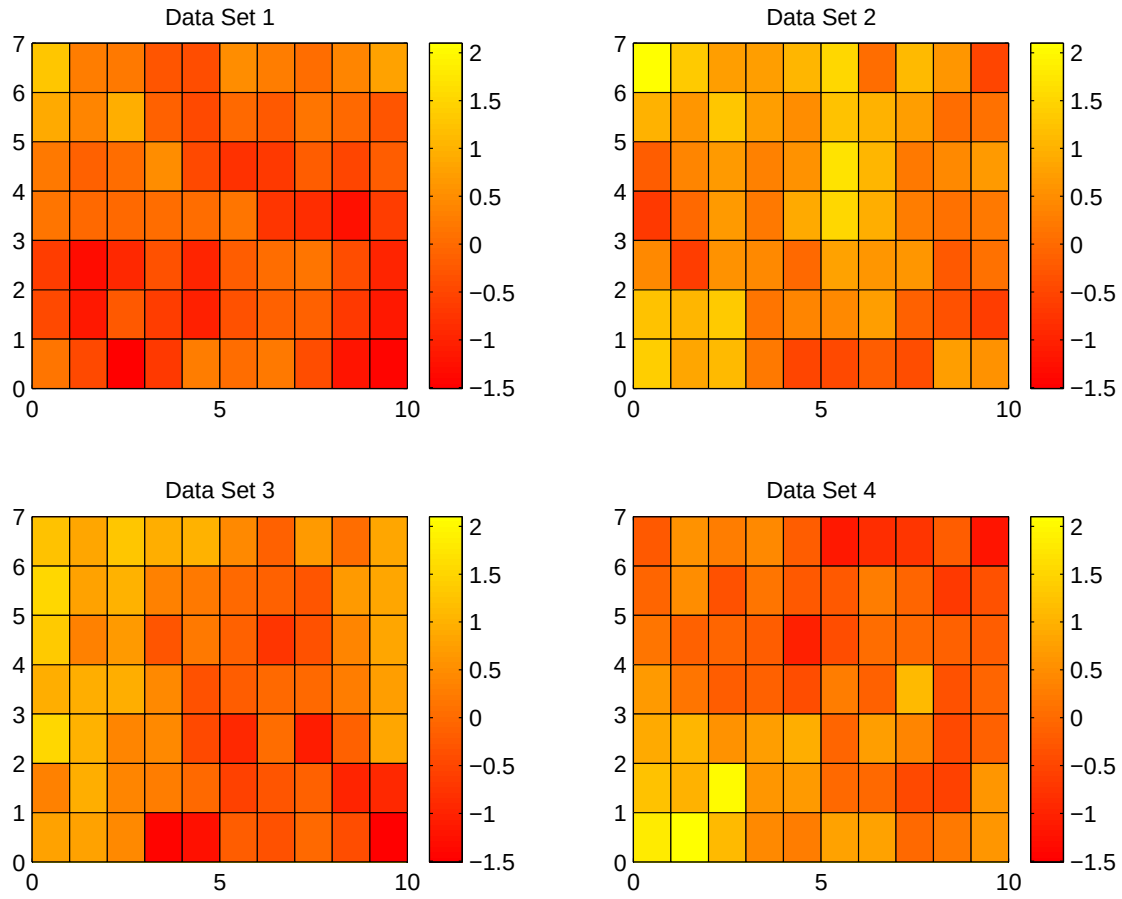

Figure 5.2: Simulated Spatial Effects for each of the 4 Data Sets Based on Model (3.2)

cluster effects $\mathbf{c} \sim N_5(0, \sigma_c^2)$ with $\sigma_c^2 = 1$.

Using these hyperparameter values we first simulated the spatial effects $\mathbf{b}$ and the cluster effects $\mathbf{c}$ for each data set and finally the binary responses using Model (3.2). The resulting data sets are presented in Figure 5.3, where black corresponds to $Y = 0$, while white to $Y = 1$.

Finally we used the following priors: for the regression parameters we chose $\alpha_1 \sim N(0, 100^2)$ and $\alpha_2 \sim N(0, 10^2)$ reflecting a diffuse prior choice. According to Model (3.2) the conditional prior for the spatial effects $b_j|\mathbf{b}_{-j}, \tau^2, \phi, \ j = 1, \ldots, J$, is given by the modified Pettitt's CAR (2.9) and the conditional prior for the cluster effects c_m given σ_c^2 is normal $N(0, \sigma_c^2)$. For the variance hyperparameters τ^2, σ_c^2 we chose flat priors, while for $\psi = \frac{\phi}{1+\phi}$, $\psi \in [0, 1)$ we chose as prior density $[\psi] \sim \frac{1}{(1-\psi)^{1-a}}$ with $a = 0.5$. This corresponds to a Pareto distribution for $\phi = \frac{\psi}{1-\psi}$, $\phi \in [0, +\infty)$, namely $[\phi] \sim \frac{1}{(1+\phi)^{1+a}}$.

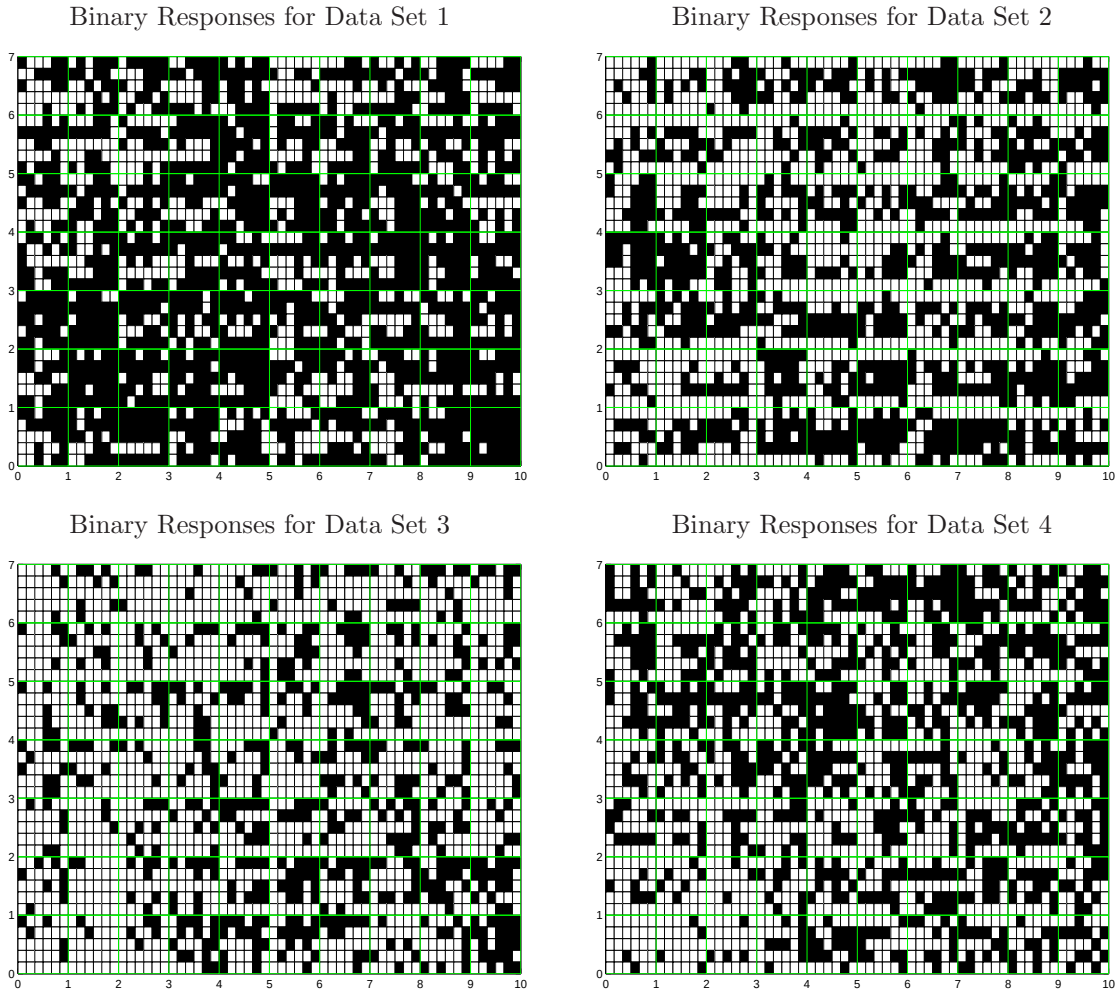

Figure 5.3: Simulated Binary Responses for each of the 4 Data Sets Based on Model (3.2)

5.1.2 Results

MCMC calculations were implemented in MATLAB. The MCMC algorithm described in Section 3.2 was run for 50,000 iterations with every 50th iteration recorded. These 50,000 iterations needed approximately 15 hours on a SUN workstation with CPU 600MHz and RAM 2.56GB. As "burn in" phase served 10 pilot runs with 300 iterations per each pilot run, which we used to determine optimal proposal standard error values for the MH-step (see Section 3.2). We achieved acceptance rates for the MH steps between $30\% - 60\%$ for each parameter. The resulting trace plots (not shown) show that such a length of "burn in" phase is enough. The autocorrelation plots (not shown) indicate, that the autocorrelations between recorded iterations are below 0.1. Figure 5.4 shows marginal posterior density estimates of the parameters $\alpha_0, \alpha_1, \psi = \frac{\phi}{1+\phi}, \tau^2$ and σ_c^2 from the 4 data sets, where the vertical fat dashed lines correspond to the true parameter value. For each density curve

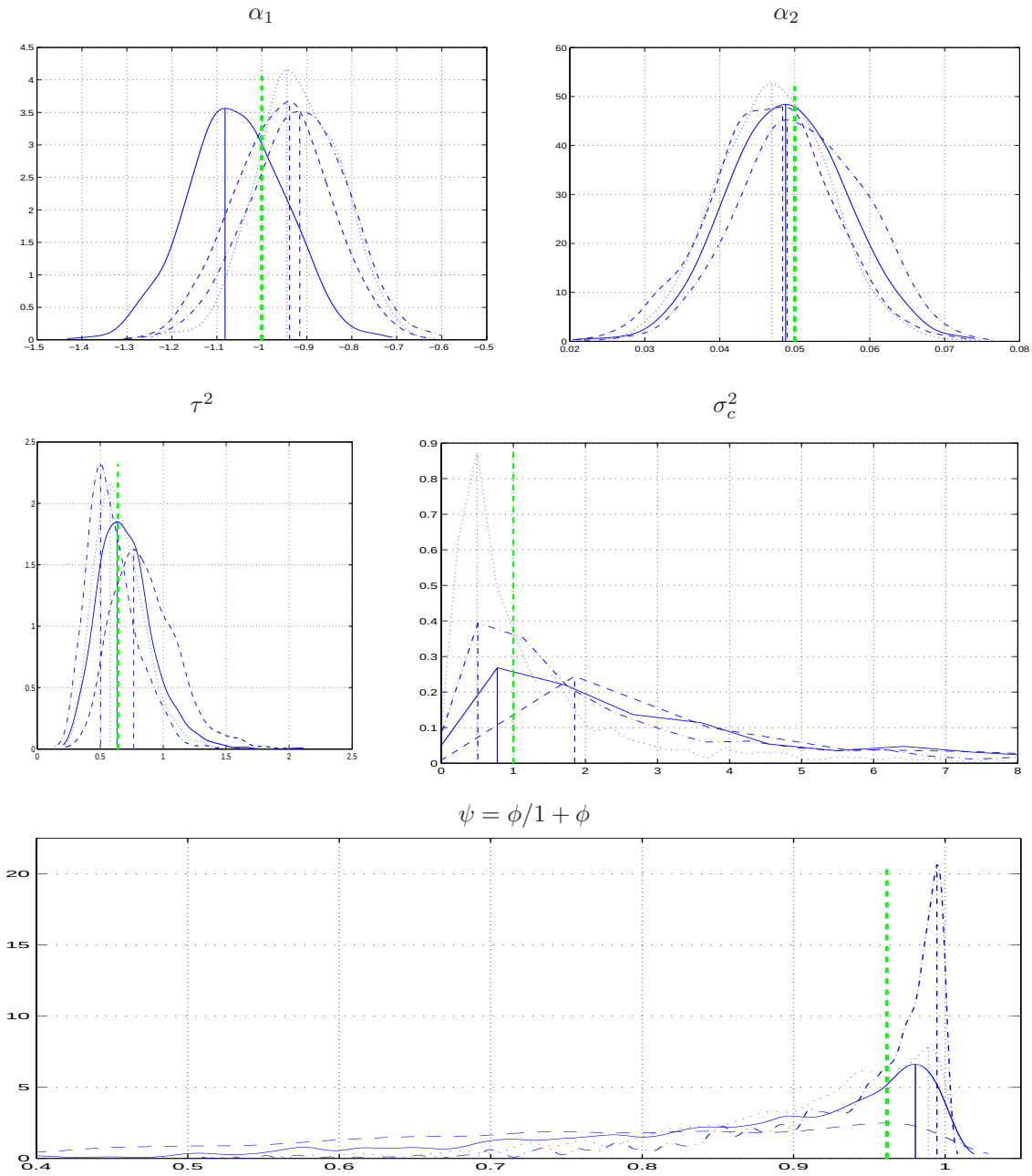

Figure 5.4: Estimated Marginal Posterior Densities for Parameters $\alpha_1, \alpha_2, \tau^2, \psi, \sigma_c^2$ in Model (3.2) (solid for Data Set 1, dashed for Data Set 2, dash-dot for Data Set 3, dotted for Data Set 4)

its mode is also marked by a thin vertical line. From this we see that in all four cases the true values are well inside 90% credible intervals. As mentioned before we chose for the parameter ψ a proper prior $[\psi] \sim \frac{1}{(1-\psi)^{1-a}}$ with $a = 0.5$. One can see, that with such a prior the parameter ψ is often overestimated in Model (3.2). If ψ is close to 1, this effect can cause estimation for $\phi = \frac{\psi}{1-\psi}$ with a large deviation from the true value. However other simulation studies show that when using $a = 1$(i.e $\psi \sim Uni[0, 1]$) the parameter ψ is underestimated. We note that another simulation study, conducted using an improper prior choice with $a = -1(\Leftrightarrow [\phi] \propto 1)$ indicates that in this case the posterior for ψ (and for ϕ) is improper . Next we consider the estimated marginal posteriors for the spatial and cluster effects. In Figure 5.5 the estimated posterior densities for Data Set 1 show that the posterior mode estimates of the spatial and cluster effects are quite precise. The same result is visible in comparison maps (see Figure 5.6) which compare true effects with their estimated posterior modes. In addition to the graphical checks provided by Figure 5.4, 5.5 and 5.6 we also calculated absolute and relative errors between the true parameter values and their posterior mode estimators. Table 5.1 contains these estimates.

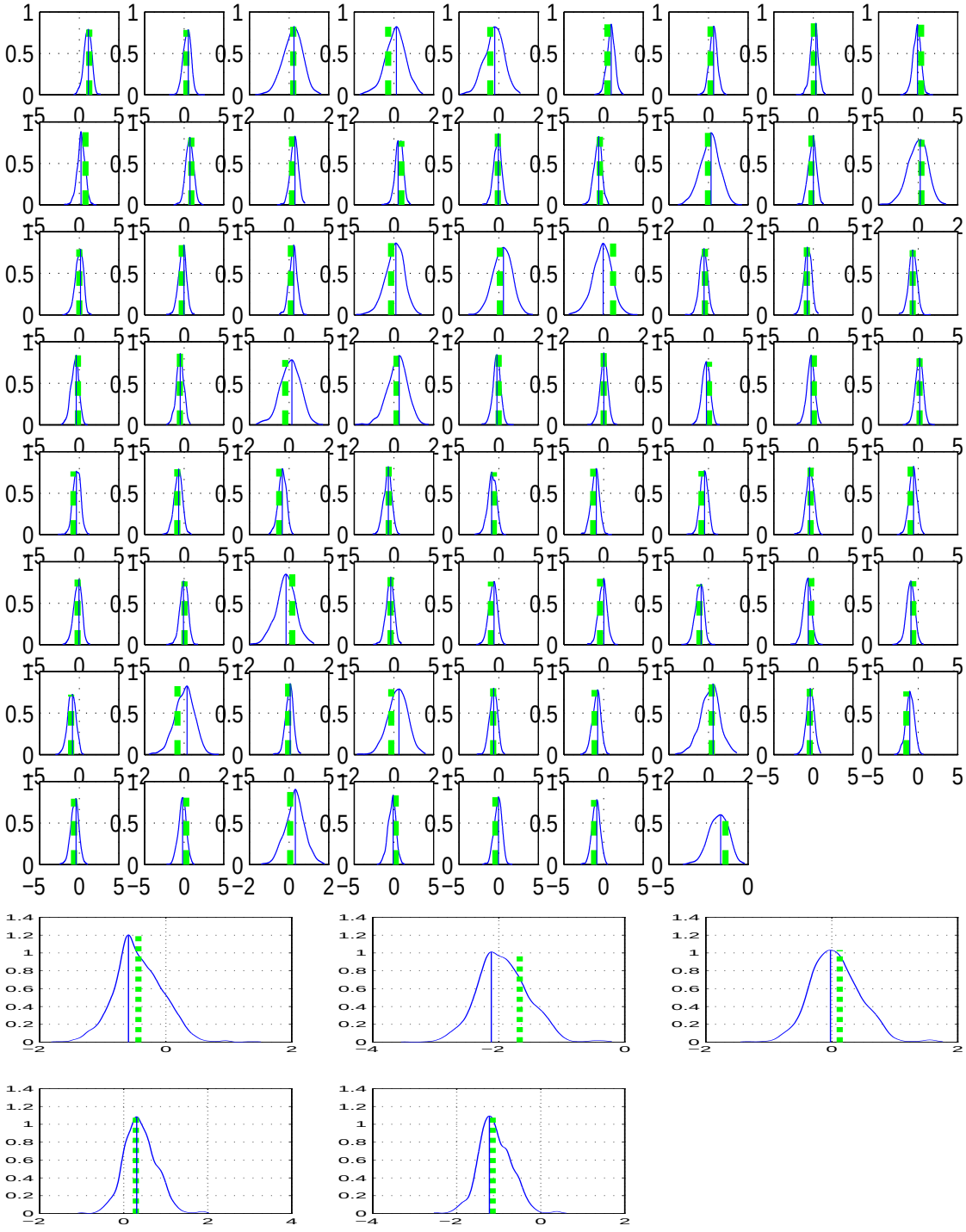

Figure 5.5: Estimated Marginal Posterior Densities for 70 spatial (above) and 5 cluster (below) Effects in Data Set 1 Based on Model (3.2) (Solid Line = Estimated Posterior Mode, Dashed Line = True Parameter)

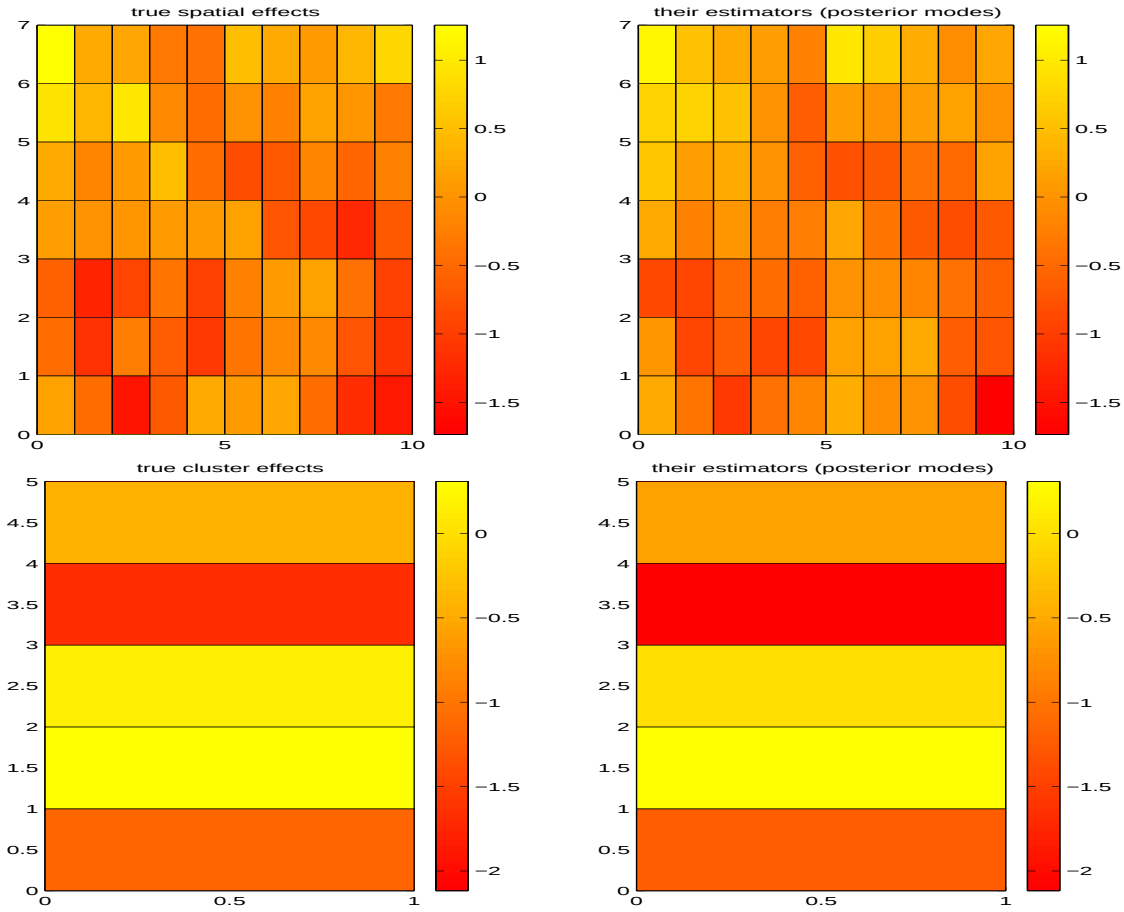

Figure 5.6: Comparison Maps for 70 Spatial (top row) and 5 Cluster (bottom row) Effects in Data Set 1 Based on Model (3.2)

	True	5%-**Quantile**	**Est.Post Mode**	**Est.Post Median**	**Est.Post Mean**	95%-**Quantile**	**True−Mode**	**Relative Error** $\left\lvert\frac{(\textbf{True}-\textbf{Mode})}{\textbf{True}}\right\rvert$
Data Set 1								
α_1	−1.0000	−1.2392	−1.0819	−1.0591	−1.0568	−0.8839	0.0819	0.0819
α_2	0.0500	0.0364	0.0488	0.0491	0.0490	0.0619	0.0012	0.0244
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}$	1.0012		1.0830	1.0603	1.0580		0.0819	0.0818*
$\|\|\cdot\|\|$ for **b**	5.2381		4.6541	4.8364	4.9151		2.4531	0.4683*
c_1	−0.4326	−1.0347	−0.5922	−0.4630	−0.4252	0.2222	0.1596	0.3690
c_2	−1.6656	−2.5648	−2.1176	−1.9465	−1.9244	−1.2286	0.4520	0.2714
c_3	0.1253	−0.5228	−0.0195	0.0353	0.0623	0.7317	0.1448	1.1554
c_4	0.2877	−0.2107	0.3083	0.3671	0.3973	1.0251	−0.0206	0.0716
c_5	−1.1465	−1.6771	−1.2237	−1.1191	−1.0819	−0.4321	0.0772	0.0674
$\|\|\cdot\|\|$ for **c**	2.0914		2.5353	2.3220	2.2839		0.5071	0.2425*
τ^2	0.6400	0.3851	0.6332	0.6774	0.7057	1.1005	0.0068	0.0106
ψ	0.9615	0.5727	0.9803	0.8951	0.8519	0.9950	−0.0188	0.0196
var_c	1.0000	0.5932	0.7772	2.4284	5.9258	17.7751	0.2228	0.2228
Data Set 2								
α_1	−1.0000	−1.1371	−0.9383	−0.9528	−0.9575	−0.7858	−0.0617	0.0617
α_2	0.0500	0.0326	0.0484	0.0467	0.0467	0.0596	0.0016	0.0324
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}$	1.0012		0.9395	0.9539	0.9587		0.0617	0.0617*
$\|\|\cdot\|\|$ for **b**	6.6018		4.5767	4.6486	4.6890		4.8098	0.7286*
c_1	1.1415	1.0685	1.5498	1.5502	1.5553	2.0777	−0.4083	0.3577
c_2	−0.9175	−1.0517	−0.4864	−0.5463	−0.5581	−0.0760	−0.4311	0.4699
c_3	−1.0401	−0.9967	−0.4214	−0.5083	−0.5188	−0.0509	−0.6187	0.5949
c_4	−1.2882	−1.2861	−0.7756	−0.7898	−0.7917	−0.3108	−0.5126	0.3979
c_5	1.1644	1.1840	1.6358	1.6753	1.6842	2.2172	−0.4714	0.4048
$\|\|\cdot\|\|$ for **c**	2.4984		2.4684	2.5279	2.5422		1.1047	0.4422*
τ^2	0.6400	0.5054	0.7659	0.8230	0.8551	1.3133	−0.1259	0.1967
ψ	0.9615	0.3862	0.9618	0.7636	0.7392	0.9833	−0.0002	0.0002
var_c	1.0000	0.8386	1.8509	2.6766	5.9309	16.4689	−0.8509	0.8509
Data Set 3								
α_1	−1.0000	−1.1005	−0.9158	−0.9164	−0.9197	−0.7504	−0.0842	0.0842
α_2	0.0500	0.0375	0.0490	0.0511	0.0511	0.0644	0.0010	0.0201
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}$	1.0012		0.9172	0.9178	0.9211		0.0842	0.0841*
$\|\|\cdot\|\|$ for **b**	6.1033		5.8560	6.1513	6.3071		3.6958	0.6055*
c_1	−0.1896	−1.5921	−0.8693	−0.6745	−0.6932	0.1149	0.6796	3.5840
c_2	0.6754	−0.6138	0.1555	0.2820	0.2782	1.0897	0.5199	0.7697
c_3	1.2071	−0.0494	0.8304	0.7961	0.7900	1.6017	0.3767	0.3120
c_4	0.9442	−0.4101	0.3841	0.4868	0.4736	1.2865	0.5600	0.5932
c_5	1.7136	0.4154	1.0956	1.2512	1.2577	2.0647	0.6180	0.3606
$\|\|\cdot\|\|$ for **c**	2.4036		1.6785	1.7236	1.7286		1.2528	0.5212*
τ^2	0.6400	0.3207	0.5030	0.5590	0.5952	0.9687	0.1370	0.2141
ψ	0.9615	0.7573	0.9946	0.9653	0.9326	0.9987	−0.0331	0.0344
var_c	1.0000	0.3984	0.5081	1.6045	3.7443	11.3068	0.4919	0.4919
Data Set 4								
α_1	−1.0000	−1.0673	−0.9442	−0.9215	−0.9179	−0.7614	−0.0558	0.0558
α_2	0.0500	0.0353	0.0469	0.0470	0.0472	0.0587	0.0031	0.0610
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}$	1.0012		0.9454	0.9227	0.9191		0.0558	0.0558*
$\|\|\cdot\|\|$ for **b**	5.7473		5.1443	5.1835	5.2333		2.3325	0.4058*
c_1	0.4062	−0.0196	0.5260	0.5513	0.5493	1.0740	−0.1198	0.2950
c_2	−0.6616	−1.2388	−0.6671	−0.6319	−0.6375	−0.1058	0.0055	0.0084
c_3	0.4101	−0.2891	0.3051	0.3018	0.2891	0.8295	0.1050	0.2560
c_4	−0.3256	−1.0141	−0.3933	−0.4094	−0.4280	0.1191	0.0677	0.2081
c_5	0.5966	0.2256	0.8217	0.8188	0.8065	1.3864	−0.2251	0.3774
$\|\|\cdot\|\|$ for **c**	1.1103		1.2825	1.2776	1.2749		0.2840	0.2558*
τ^2	0.6400	0.3940	0.5788	0.6322	0.6612	1.0193	0.0612	0.0956
ψ	0.9615	0.7148	0.9890	0.9267	0.8985	0.9963	−0.0274	0.0285
var_c	1.0000	0.2200	0.4981	0.8332	2.0182	6.3734	0.5019	0.5019

* value indicates $\frac{\|\text{True} - \text{Estimator}\|}{\|\text{True}\|}$, where $\|\|\cdot\|\|$ is the Euclidian norm.

Table 5.1: Estimated Posterior Mode, Median, Mean and Quantiles for Simulation Study 1 Based on Model (3.2)

5.2 Study 2: Hierarchical Spatial Binary Regression with Individual Cluster Effects using Model (4.9)

5.2.1 Setup

The simulation study based on the Logit Model (4.9) has the following mean structure:

$$\Theta_i' := \log\left(\frac{p_i}{1-p_i}\right) = \frac{x_{1i}\alpha_1' + x_{2i}\alpha_2' + b'_{j(i)}}{\sigma'_{m(i)}}, \quad i = 1,\ldots,n,\ j = 1,\ldots,J,\ m = 1,\ldots,M.$$

As for the previous simulation study we again chose $\alpha_1' = -1, \alpha_2' = 0.05$, $\tau^{2'} = 0.64, \phi' = 25$. As true values for the cluster parameters $\sigma'_m,\ m = 2,\ldots,M$, we take 4 samples from a $Uni[0.75, 1.25]$ distribution. According to Model (4.9) we set $\sigma_1' = 1$. In this way we got the following true cluster parameter values: $\boldsymbol{\sigma}' = (1, 1.2251, 0.8656, 1.0534, 0.9930)^t$. We generated 4 data sets with $n = 2100$ binary responses using the same $J = 70$ regions and $M = 5$ clusters. The modeling of the fixed regression effects $x_{1i}\alpha_1' + x_{2i}\alpha_2',\ i = 1,\ldots,n$, and the spatial effects $b_{j(i)},\ i = 1,\ldots,n$, remains as before. The simulated four data sets for Model (4.9) are presented in the Figure 5.7. Prior choices for $\boldsymbol{\alpha}', \mathbf{b}', \tau^{2'}, \psi'$ remain the same, while for the prior distribution of the cluster parameters $\sigma'_m,\ m = 1,\ldots,M$, we used $N(1,1)$ truncated on the interval $[0.2, +\infty]$.

5.2.2 Results

The corresponding MCMC algorithm is described in Section 4.2.1. Results are presented in the same way as for simulation Study 1, i.e. Figure 5.8 gives posterior density estimates of all parameters while Figure 5.9 presents spatial comparison maps and Table 5.2 presents estimated posterior location measures and quantiles. Figure 5.10 presents the time plots based on Data Set 1. Its convergence indicates identifiability of Model (4.9). In contrast to Simulation Study 1 we note that for this model ψ' is no longer overestimated (see Figure 5.8).

Figure 5.7: Simulated Binary Responses for each of the 4 Data Sets Based on Model (4.9)

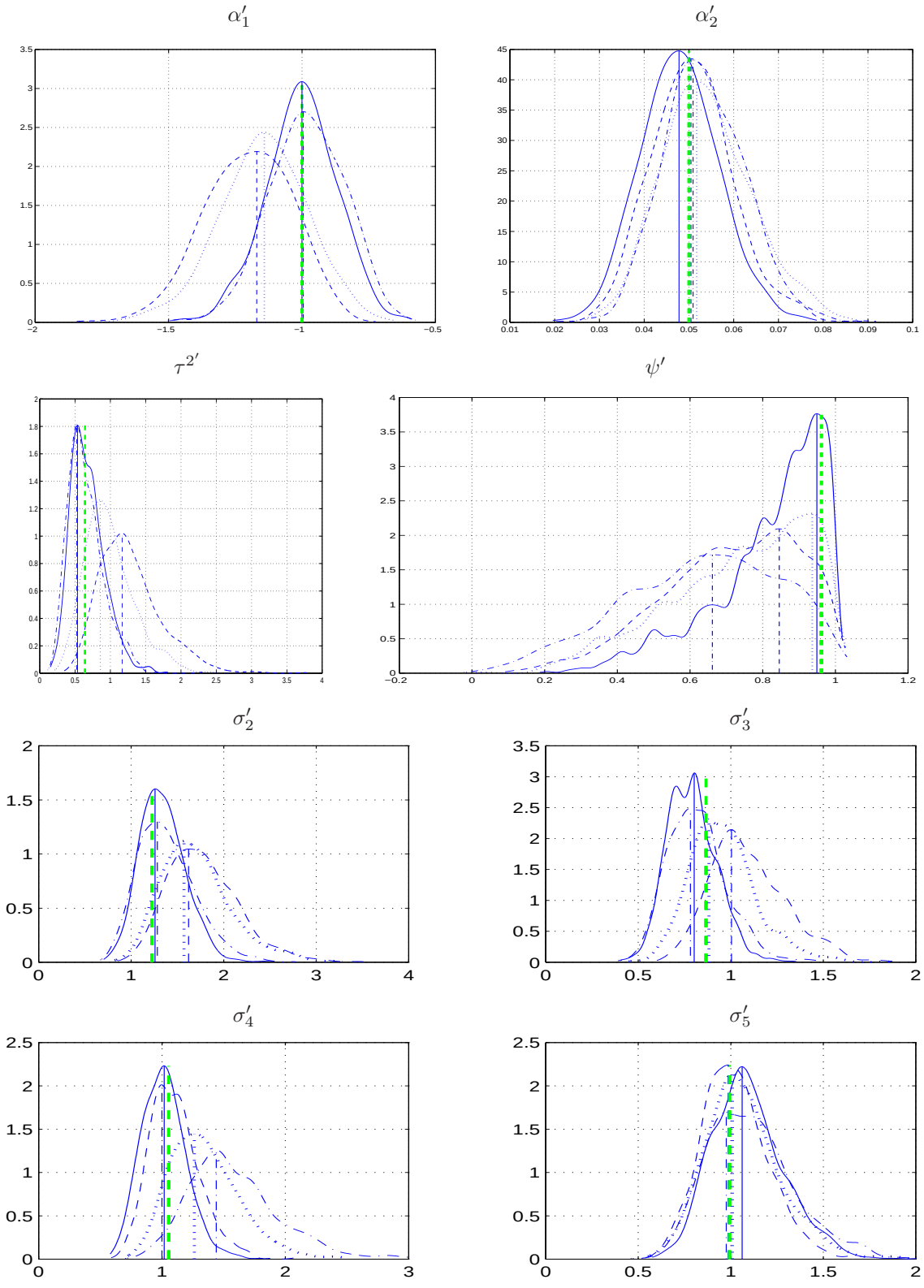

Figure 5.8: Estimated Marginal Posterior Densities for Parameters $\boldsymbol{\alpha}', \tau^{2'}, \psi', \boldsymbol{\sigma}'$ in Model (4.9) (solid for Data Set 1, dashed for Data Set 2, dash-dot for Data Set 3, dotted for Data Set 4)

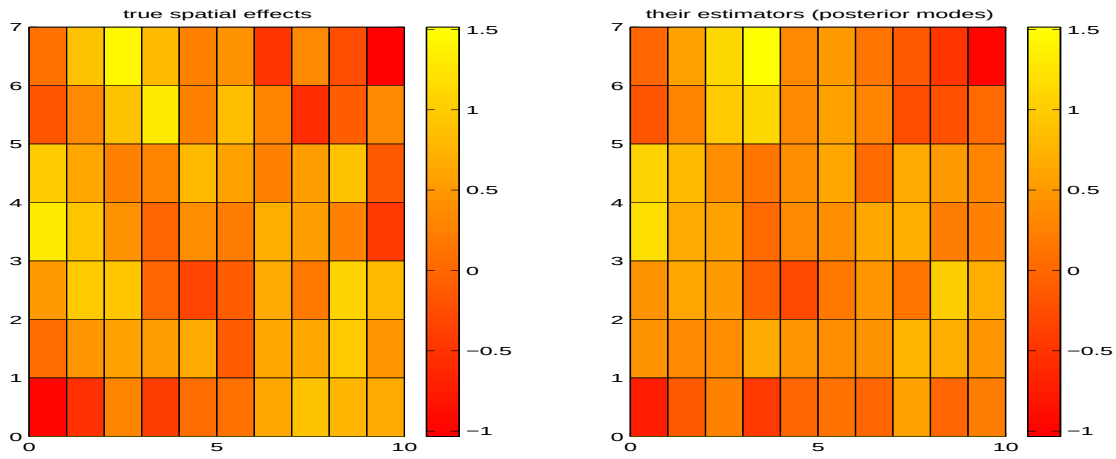

Figure 5.9: Comparison Maps for 70 Spatial Effects in Data Set 1 Based on Model (4.9)

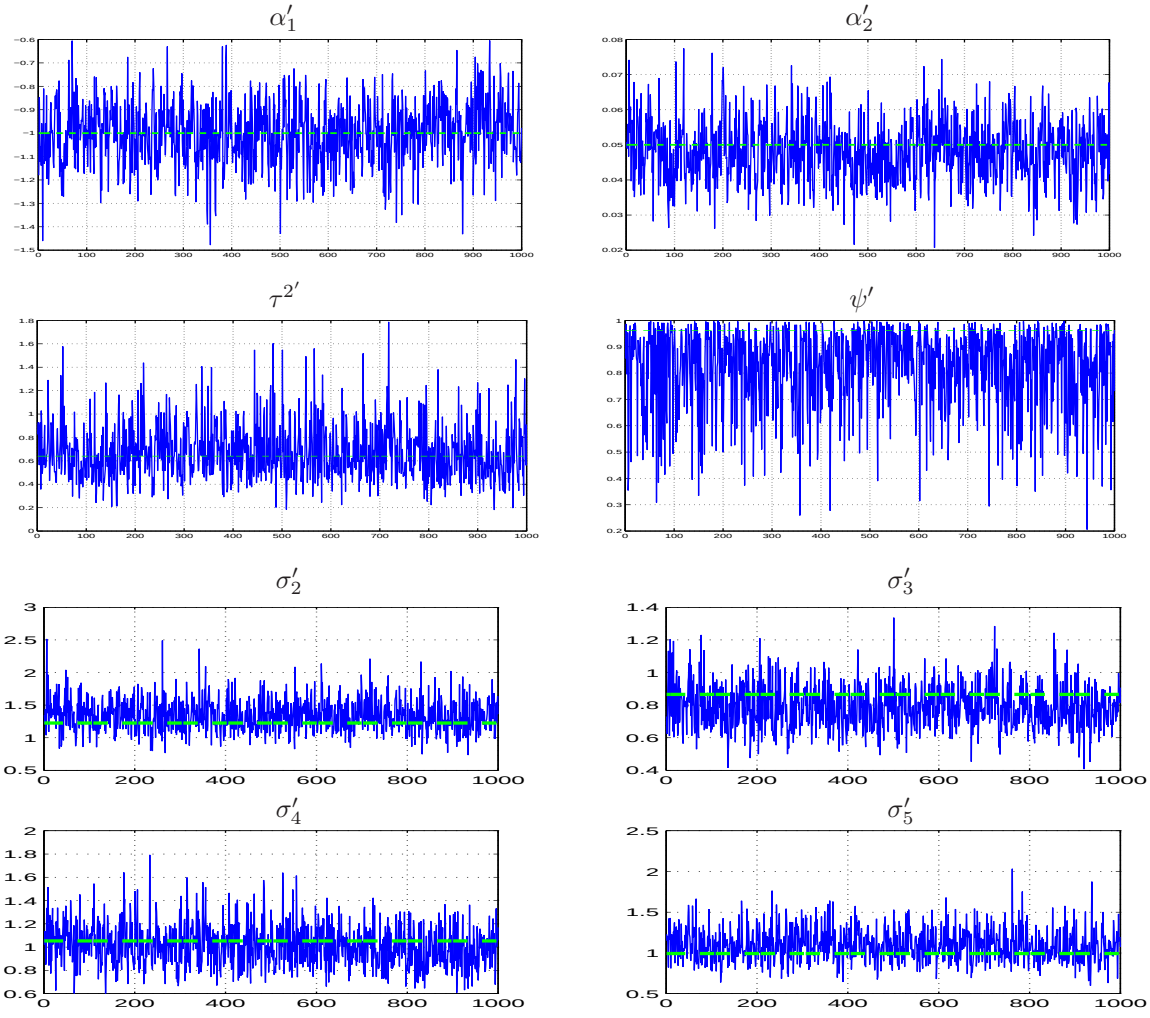

Figure 5.10: Trace Plots for Parameters $\boldsymbol{\alpha}', \tau^{2'}, \psi', \boldsymbol{\sigma}'$ Based on Data Set 1 in Model (4.9) (Dashed Line = True Parameter)

	True	5%-Quantile	Est.Post Mode	Est.Post Median	Est.Post Mean	95%-Quantile	True−Mode	Relative Error $\left\lvert\frac{(\text{True}-\text{Mode})}{\text{True}}\right\rvert$
Data Set 1								
α_1'	−1.0000	−1.2366	−1.0003	−1.0000	−1.0016	−0.7854	0.0003	0.0003
α_2'	0.0500	0.0342	0.0478	0.0478	0.0480	0.0630	0.0022	0.0443
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012		1.0014	1.0011	1.0028		0.0022	0.0022*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	5.3139		4.6435	4.7960	4.8697		2.4015	0.4519*
σ_2'	1.2251	0.9848	1.2580	1.3263	1.3420	1.7624	−0.0329	0.0269
σ_3'	0.8656	0.5932	0.8025	0.7834	0.7893	1.0220	0.0630	0.0728
σ_4'	1.0534	0.7465	1.0175	1.0099	1.0153	1.3157	0.0360	0.0341
σ_5'	0.9930	0.7989	1.0611	1.0738	1.0937	1.4535	−0.0682	0.0686
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}'_{-1}$	2.0846		2.0947	2.1321	2.1567		0.1049	0.0503*
$\tau^{2'}$	0.6400	0.3453	0.5351	0.6437	0.6781	1.1154	0.1049	0.1639
ψ'	0.9615	0.4915	0.9491	0.8594	0.8153	0.9887	0.0124	0.0129
Data Set 2								
α_1'	−1.0000	−1.4775	−1.1698	−1.1946	−1.2021	−0.9461	0.1698	0.1698
α_2'	0.0500	0.0365	0.0504	0.0503	0.0507	0.0665	−0.0004	0.0090
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012		1.1709	1.1956	1.2032		0.1698	0.1696*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	6.1657		5.7982	5.9475	6.0431		3.1543	0.5116*
σ_2'	1.2251	1.2328	1.6214	1.7361	1.7782	2.4898	−0.3963	0.3235
σ_3'	0.8656	0.7960	1.0037	1.0616	1.0948	1.4829	−0.1381	0.1596
σ_4'	1.0534	0.8149	0.9997	1.0850	1.1033	1.4690	0.0537	0.0510
σ_5'	0.9930	0.7474	0.9861	1.0075	1.0235	1.3331	0.0069	0.0070
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}'_{-1}$	2.0846		2.3681	2.5166	2.5739		0.4232	0.2030*
$\tau^{2'}$	0.6400	0.6952	1.1680	1.2076	1.2795	2.1002	−0.5280	0.8249
ψ'	0.9615	0.3756	0.8456	0.7312	0.7115	0.9711	0.1160	0.1206
Data Set 3								
α_1'	−1.0000	−1.2170	−0.9953	−0.9826	−0.9812	−0.7546	−0.0047	0.0047
α_2'	0.0500	0.0392	0.0509	0.0528	0.0535	0.0690	−0.0009	0.0183
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012		0.9966	0.9840	0.9826		0.0048	0.0048*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	4.6634		3.7925	3.9088	3.9925		2.7244	0.5842*
σ_2'	1.2251	0.9397	1.2818	1.3483	1.3851	1.9530	−0.0567	0.0463
σ_3'	0.8656	0.5861	0.7820	0.8084	0.8177	1.0977	0.0835	0.0965
σ_4'	1.0534	1.0675	1.4402	1.5200	1.5902	2.2758	−0.3868	0.3672
σ_5'	0.9930	0.7799	0.9758	1.0849	1.1053	1.5388	0.0172	0.0173
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}'_{-1}$	2.0846		2.2981	2.4411	2.5174		0.4002	0.1920*
$\tau^{2'}$	0.6400	0.3132	0.5232	0.6030	0.6518	1.1466	0.1168	0.1825
ψ'	0.9615	0.2399	0.6614	0.6483	0.6302	0.9480	0.3001	0.3121
Data Set 4								
α_1'	−1.0000	−1.4284	−1.1414	−1.1427	−1.1456	−0.8690	0.1414	0.1414
α_2'	0.0500	0.0389	0.0517	0.0531	0.0540	0.0725	−0.0017	0.0347
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012		1.1426	1.1440	1.1469		0.1414	0.1413*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	5.1904		5.1195	5.2180	5.3150		2.4346	0.4691*
σ_2'	1.2251	1.1819	1.5708	1.6527	1.7049	2.4287	−0.3457	0.2822
σ_3'	0.8656	0.6944	0.8817	0.9423	0.9604	1.2887	−0.0161	0.0187
σ_4'	1.0534	0.9662	1.2622	1.3162	1.3477	1.8574	−0.2088	0.1982
σ_5'	0.9930	0.7771	1.0095	1.0483	1.0695	1.4400	−0.0165	0.0166
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}'_{-1}$	2.0846		2.4201	2.5398	2.6056		0.4045	0.1941*
$\tau^{2'}$	0.6400	0.5069	0.8575	0.9385	0.9887	1.6938	−0.2175	0.3399
ψ'	0.9615	0.3522	0.9365	0.7647	0.7336	0.9783	0.0250	0.0260

* value indicates $\frac{\|\text{True} - \text{Estimator}\|}{\|\text{True}\|}$, where $\|\cdot\|$ is the Euclidian norm.

Table 5.2: Estimated Posterior Mode, Median, Mean and Quantiles for Simulation Study 2 Based on Model (4.9)

5.3 Study 3: Hierarchical Spatial Binary Regression with Individual Cluster Effects using Model (4.7)

5.3.1 Setup

The only difference between this model and the previous Model (4.9) is that now we use a probit link instead of a logit one as in Model (4.9). Therefore the generation of the data sets for this model (see Figure 5.11) is similar and uses the same parameter values

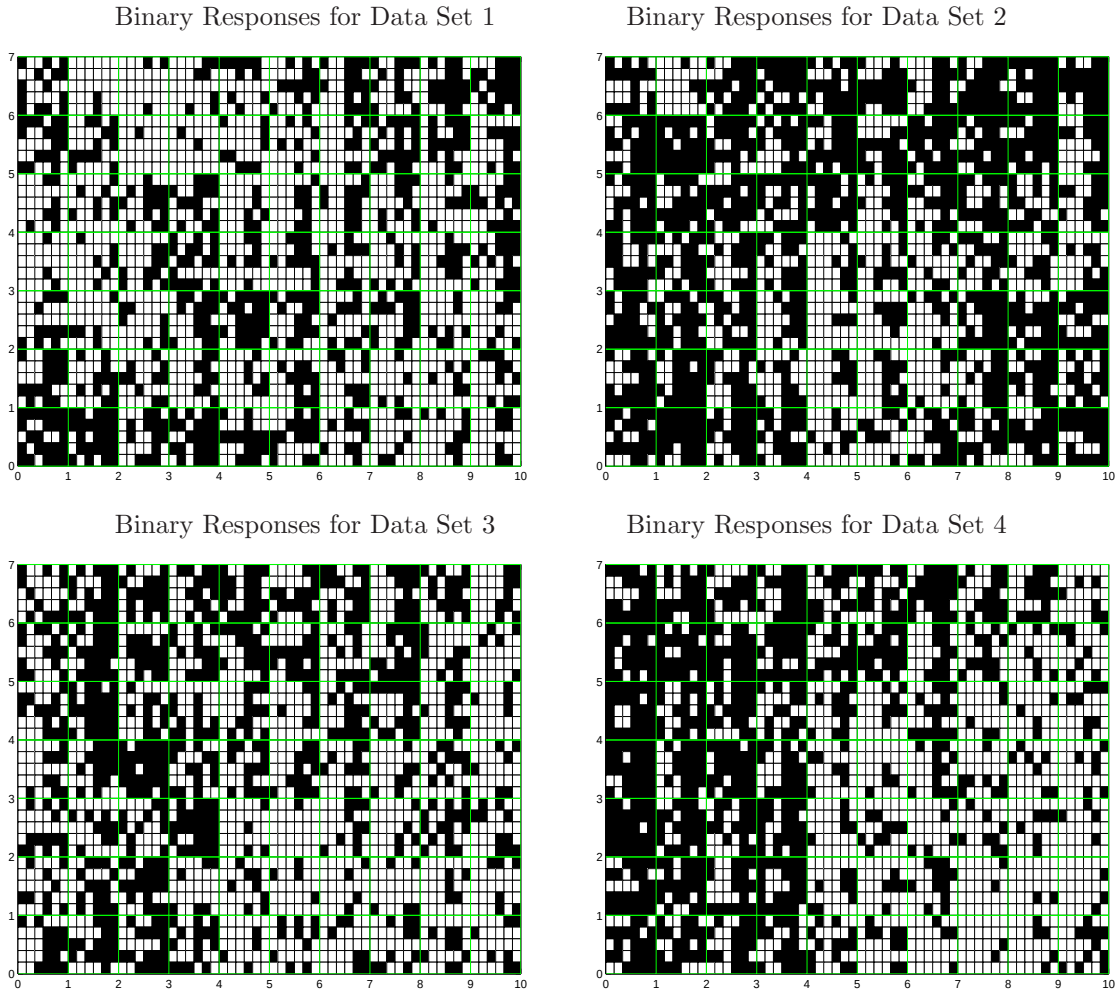

Figure 5.11: Simulated Binary Responses for each of the 4 Data Sets Based on Model (4.7)

$\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}', \tau^{2'}, \phi'$ as before. However, the MCMC algorithm is based now on the utilization of the latent variables $Z_i', \ i = 1, \ldots, n$ (for details see Section 4.2.3) which allow a direct Gibbs step for all parameters except ψ'. Moreover, now in contrast to Model (4.9) a joint

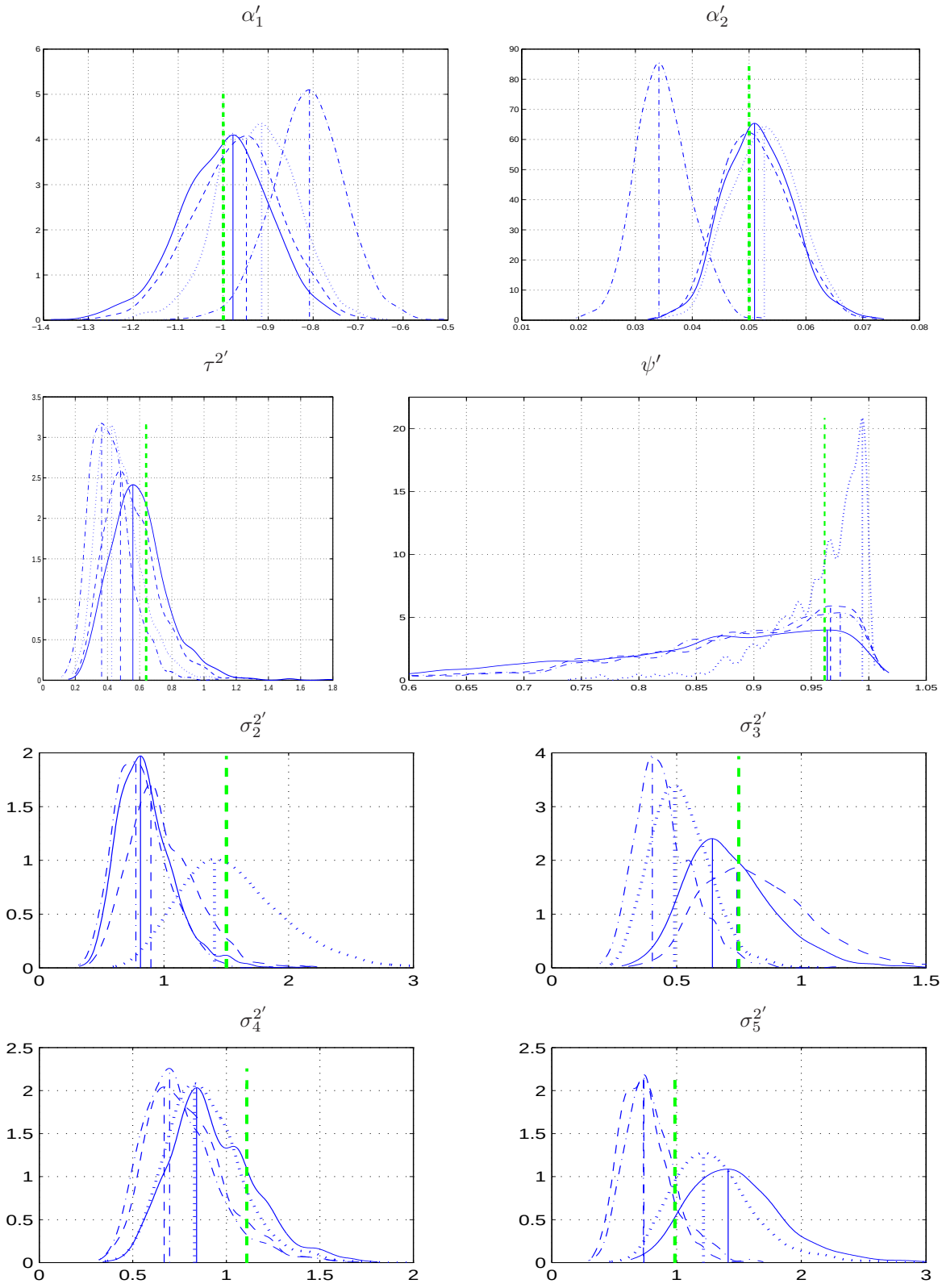

Figure 5.12: Estimated Marginal Posterior Densities for Parameters $\boldsymbol{\alpha'}, \tau^{2'}, \psi', \boldsymbol{\sigma^{2'}}$ in Model (4.7) (solid for Data Set 1, dashed for Data Set 2, dash-dot for Data Set 3, dotted for Data Set 4)

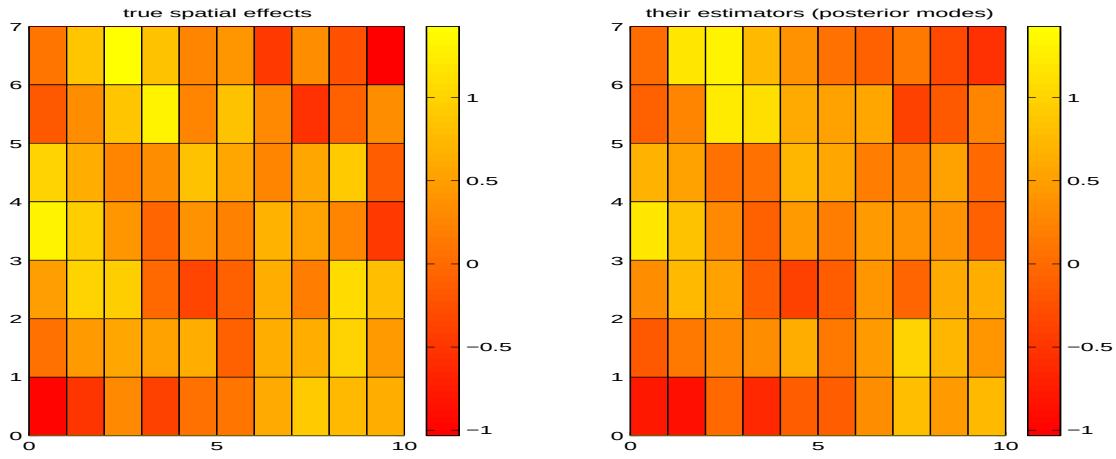

Figure 5.13: Comparison Maps for 70 Spatial Effects in Data Set 1 Based on Model (4.7)

update for $\boldsymbol{\alpha}'$ and $\mathbf{b}'$ is possible. One further difference is that in this model we update the parameter $\boldsymbol{\sigma}^{\mathbf{2}'}$ instead of $\boldsymbol{\sigma}'$ as before. As a prior for the parameter $\sigma_m^{2'},\ m = 2, \ldots, M$, we use an $IG(3, 0.5)$ distribution (with $\mathbf{E}(\sigma_m^{2'}) = \mathbf{Var}(\sigma_m^{2'}) = 1$), while the other parameters have the same priors as before.

5.3.2 Results

Since the corresponding MCMC algorithm requires double computation time compared to the previous simulation studies we run 25,000 instead of 50,000 iterations with every 25th iteration recorded. However the corresponding autocorrelation plots (not shown here) indicate even better mixing as by the previous models. This effect is due to joint updates. Unfortunately, the precision of the estimators in this study is not so good as before. This holds especially for the variance parameters $\sigma_m^{2'},\ m = 2, \ldots, M$. Also for the regression parameter $\boldsymbol{\alpha}'$ the estimator based on the third data set has relatively large deviations from the corresponding true values.

	True	5%-Quantile	Est.Post Mode	Est.Post Median	Est.Post Mean	95%-Quantile	True−Mode	Relative Error $\left\lvert\frac{(\text{True}-\text{Mode})}{\text{True}}\right\rvert$
Data Set 1								
α_1'	−1.0000	−1.1680	−0.9783	−0.9900	−0.9982	−0.8459	−0.0217	0.0217
α_2'	0.0500	0.0426	0.0510	0.0514	0.0516	0.0616	−0.0010	0.0194
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012	0	0.9796	0.9913	0.9995	0	0.0217	0.0217*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	5.3139	0	4.5940	4.6784	4.7389	0	1.9146	0.3603*
$\sigma_2^{2'}$	1.5008	0.5538	0.8115	0.8279	0.8595	1.2995	0.6892	0.4593
$\sigma_3^{2'}$	0.7492	0.4678	0.6434	0.6887	0.7168	1.0448	0.1058	0.1413
$\sigma_4^{2'}$	1.1097	0.5934	0.8410	0.8925	0.9275	1.3238	0.2687	0.2422
$\sigma_5^{2'}$	0.9860	0.9511	1.4125	1.4498	1.4852	2.1802	−0.4264	0.4325
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}^{\mathbf{2}'}$	2.2399	0	1.9429	2.0144	2.0781	0	0.8604	0.3841*
$\tau^{2'}$	0.6400	0.3454	0.5578	0.5766	0.5962	0.9323	0.0822	0.1284
ψ'	0.9615	0.5661	0.9638	0.8697	0.8362	0.9891	−0.0023	0.0024
Data Set 2								
α_1'	−1.0000	−1.1355	−0.9484	−0.9658	−0.9697	−0.8081	−0.0516	0.0516
α_2'	0.0500	0.0420	0.0499	0.0508	0.0512	0.0618	0.0001	0.0021
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012	0	0.9497	0.9671	0.9710	0	0.0516	0.0515*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	5.5595	0	4.7861	4.8455	4.8880	0	2.2909	0.4121*
$\sigma_2^{2'}$	1.5008	0.6130	0.8951	0.9433	0.9887	1.4696	0.6057	0.4036
$\sigma_3^{2'}$	0.7492	0.5335	0.7415	0.8003	0.8302	1.2206	0.0077	0.0103
$\sigma_4^{2'}$	1.1097	0.5112	0.6683	0.7586	0.7893	1.1798	0.4414	0.3978
$\sigma_5^{2'}$	0.9860	0.4731	0.7384	0.7231	0.7385	1.0765	0.2477	0.2512
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}^{\mathbf{2}'}$	2.2399	0	1.5306	1.6213	1.6838	0	0.7894	0.3524*
$\tau^{2'}$	0.6400	0.3233	0.4807	0.5226	0.5496	0.8484	0.1593	0.2489
ψ'	0.9615	0.6524	0.9667	0.9003	0.8739	0.9924	−0.0052	0.0054
Data Set 3								
α_1'	−1.0000	−0.9418	−0.8082	−0.8095	−0.8108	−0.6827	−0.1918	0.1918
α_2'	0.0500	0.0274	0.0341	0.0346	0.0348	0.0432	0.0159	0.3173
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012	0	0.8089	0.8102	0.8115	0	0.1924	0.1922*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	4.6563	0	4.0244	4.0878	4.1292	0	1.8784	0.4034*
$\sigma_2^{2'}$	1.5008	0.5220	0.7748	0.7943	0.8176	1.1984	0.7260	0.4837
$\sigma_3^{2'}$	0.7492	0.2933	0.4030	0.4352	0.4529	0.6666	0.3462	0.4621
$\sigma_4^{2'}$	1.1097	0.4854	0.6971	0.7233	0.7557	1.1439	0.4126	0.3718
$\sigma_5^{2'}$	0.9860	0.5081	0.7337	0.7705	0.7986	1.1933	0.2523	0.2559
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}^{\mathbf{2}'}$	2.2399	0	1.3368	1.3918	1.4431	0	0.9385	0.4190*
$\tau^{2'}$	0.6400	0.2434	0.3637	0.4042	0.4180	0.6550	0.2763	0.4317
ψ'	0.9615	0.6352	0.9752	0.8984	0.8710	0.9926	−0.0137	0.0142
Data Set 4								
α_1'	−1.0000	−1.0670	−0.9145	−0.9241	−0.9255	−0.7878	−0.0855	0.0855
α_2'	0.0500	0.0435	0.0527	0.0528	0.0528	0.0625	−0.0027	0.0535
$\|\|\cdot\|\|$ for $\boldsymbol{\alpha}'$	1.0012	0	0.9160	0.9256	0.9270	0	0.0855	0.0854*
$\|\|\cdot\|\|$ for $\mathbf{b}'$	6.5191	0	5.9122	5.9885	6.0411	0	2.3161	0.3553*
$\sigma_2^{2'}$	1.5008	0.9740	1.4046	1.5061	1.5645	2.3310	0.0961	0.0641
$\sigma_3^{2'}$	0.7492	0.3386	0.4914	0.5073	0.5178	0.7205	0.2578	0.3441
$\sigma_4^{2'}$	1.1097	0.5981	0.8296	0.8658	0.8897	1.2562	0.2801	0.2524
$\sigma_5^{2'}$	0.9860	0.8343	1.2150	1.2627	1.3016	1.9401	−0.2290	0.2323
$\|\|\cdot\|\|$ for $\boldsymbol{\sigma}^{\mathbf{2}'}$	2.2399	0	2.0926	2.2068	2.2806	0	0.4546	0.2029*
$\tau^{2'}$	0.6400	0.2955	0.4263	0.4556	0.4763	0.7303	0.2137	0.3339
ψ'	0.9615	0.8727	0.9944	0.9686	0.9560	0.9978	−0.0329	0.0342

* value indicates $\frac{\|\text{True} - \text{Estimator}\|}{\|\text{True}\|}$, where $\|\|\cdot\|\|$ is the Euclidian norm.

Table 5.3: Estimated Posterior Mode, Median, Mean and Quantiles for Simulation Study 3 Based on Model (4.7)

5.4 Summary of Simulation Results

We summarize now briefly the main results of our simulation study. First of all, the trace plots of the realized MCMC Chains from Models (4.9) (see Figure 5.10) and (4.7) (not presented here) indicate identifiability. These models and the corresponding estimating MCMC algorithms can therefore successfully be used for data from the unidentifiable primary Logit (4.1) and Probit (4.2) Models with individual cluster effects.

The next point is to consider the random variability of the estimates. For each model we simulated 4 data sets, which provided us with 4 posterior estimates for the corresponding parameter vector. For Model (3.2) all 4 posterior estimates for the fixed $\boldsymbol{\alpha}$, spatial $\mathbf{b}$ and cluster $\mathbf{c}$ parameters lie quite closely around the corresponding true values. The true values are also well inside 90% credible intervals (see Figures 5.4, 5.5, 5.6 and Table 5.1). However the spatial hyperparameter ψ, if close to 1, is often overestimated when a prior $[\psi] \sim \frac{1}{(1-|\psi|)^{1-a}}$ with $a = 0.5$ (see Figures 5.4) and underestimated when $a = 1$ is chosen. For the choice $a = -1$, which corresponds to an improper prior, the resulting posterior for ψ is obviously also improper for the modified Pettitt's CAR model. For Model (4.9) the hyperparameter ψ is not overestimated by a prior choice of $[\psi] \sim \frac{1}{(1-|\psi|)^{1-a}}$ with $a = 0.5$, while the estimates for the other parameters remain close to the corresponding true values (see Figures 5.8 and 5.9 and Table 5.2). Finally for the Probit Model (4.7) the estimates for the parameters $\boldsymbol{\alpha}', \mathbf{b}', \tau^{2'}$ and ψ' are usually still sufficient, but not as precise, as for former models. This is especially true for the cluster parameter $\boldsymbol{\sigma}^{\mathbf{2}'}$ (see Figure 5.12, 5.13 and Table 5.3).

The MCMC estimating algorithms are fast enough so, that we were able to simulate relatively long Markov Chains (50,000 iterations for Models (3.2), (4.9) and 25,000 iterations for Model (4.7)), with every 50th or 25th iteration respectively recorded. The running time was between 10 and 15 hours. 10 pilot runs with 300 iterations per each pilot run were used to determine optimal proposal standard error values for the MH-step (see Section 3.2). This allows us to achieve acceptance rates between $30\% - 60\%$ for the MH-steps. All this provided a good mixing for each model, which can also be seen from low autocorrelations (not shown). This holds especially for Model (4.7), where the estimating MCMC algorithm allows block updating and reduces the number of parameters, whose updates require a MH-step to only one. Finally we note, that the length of the "burn in" phase, which consists of 10 pilot runs is enough to achieve the stationary phase of the corresponding Markov Chains (see for example Figure 5.10 for Model (4.9)).

Chapter 6

Application: Mobility Data

6.1 Data Description

In this chapter we analyze the mobility behavior of private households in Munich. In particular, the choice between individual and public transport options is recorded for each trip. One central question is to identify areas of low/high utilization of public transport after adjusting for explanatory factors such as trip, individual and household related attributes. The goal is to find flexible statistical models which incorporate covariates together with spatial and cluster information.

The data was collected within the study "Mobility 97" (see Zängler 2000). It was sponsored by the BMW AG, Munich, Germany, and conducted by the Institute "Sozialökonomie des Haushalts" at the Munich University of Technology in collaboration with the Infratest Burke Wirtschaftsforschung GmbH & Co, Munich. The participants of the survey are German-speaking persons not younger then 10 years, which live in a private household in the state of Bavaria.

In order to take into consideration seasonal fluctuations in mobility behavior of the participants, the survey was carried out in three waves in March, June and October of 1997. Each participant reported all his or her trips conducted by public or individual transport during a period of two or three days.

We consider part of the data which includes 1375 trips taken by 296 persons in 167 households in the city of Munich, Germany.

For each trip the binary variable of interest Y has value 1, if individual transport (car) was used and value 0, if public transport was used. In addition to the response Y person, household and trip related covariates were recorded. Using standard model selection techniques for GLM's we selected the following set of covariates as starting point for our models. The person related covariates are age (metric), sex, personal income, car usage (main, secondary or not user) and whether the person possesses or not a public

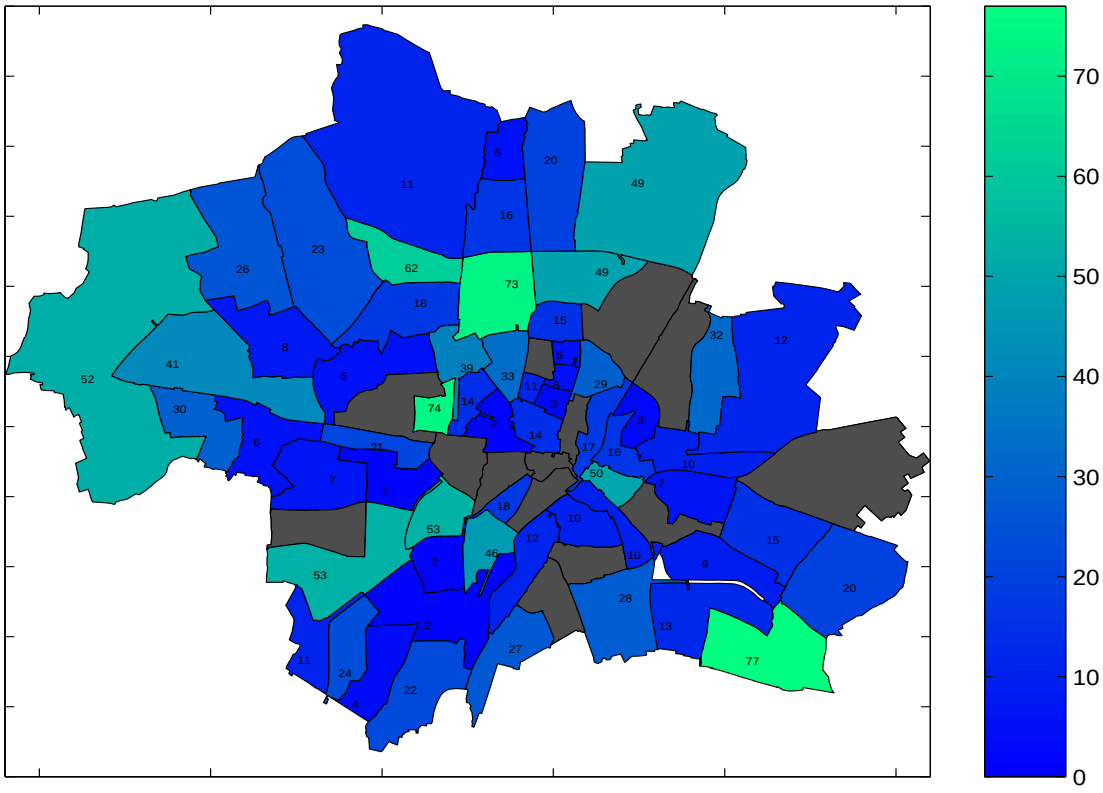

Figure 6.1: Number of Available Trips over Postal Codes of Munich, Germany

transport net card. We retain only one household related covariate, namely household type (single, single parent or not single). The trip related covariates are day type (work day or weekend), day time (day or night), distance and whether the person took the trip alone or not alone.

Table 6.1 shows the covariates, which will be utilized in our models. There we also give the distribution of trips classified by the variable levels. For the covariate **USAGE**, note that both main and secondary users must be not younger than 18 years and must have a driver license and a car available in the household. In addition to the main effects we selected the following 10 interactions:

WAY ALONE:NET CARD	USAGE:SEX
WAY ALONE:USAGE	DISTANCE:USAGE
DAY TYPE:NET CARD	USAGE:DAY TIME
SEX:DAY TIME	PERSONAL INCOME:NET CARD
DISTANCE:AGE	DAY TYPE:AGE

Since the MCMC algorithms presented in Sections 3 and 4 exclude an intercept, we have to estimate 36 regression parameters. The corresponding parameter estimates ignoring

Variable (coding)	**Levels**	**Number of Trips Using**		
		Individual Transport	**Public Transport**	**Total**
DAY TYPE (d.ty.)	*WORK DAY*	595	297	892
	WEEKEND	425	58	483
HOUSEHOLD TYPE (hh.)	*SINGLE*	156	125	281
	SINGLE PARENT	84	10	94
	NOT SINGLE	780	220	1000
PERSONAL INCOME (p.i.)	*NO INCOME* ($<$ 200 *DM)*	24	31	55
	MIDDLE (200 $-$ 3000 *DM)*	475	193	668
	HIGH ($>$ 3000 *DM)*	521	131	652
DISTANCE (d.)	*SHORT* ($\leq$ 3.5 *km)*	294	71	365
	MIDDLE (3.6 $-$ 21.5 *km)*	571	257	828
	FAR ($>$ 21.5 *km)*	155	27	182
WAY ALONE (w.a.)	*ALONE*	507	267	774
	NOT ALONE	513	88	601
USAGE (u.)	*MAIN USER*	731	100	831
	SECONDARY USER	213	99	312
	NOT USER	76	156	232
NET CARD (n.c.)	*YES*	235	247	482
	NO	785	108	893
SEX (s.)	*MALE*	549	172	721
	FEMALE	471	183	654
DAY TIME (d.t.)	*DAY (6 a.m. - 9 p.m.)*	905	336	1241
	NIGHT (9 p.m. - 6 a.m.)	115	19	134
AGE (POLY.AGE.1 and POLY.AGE.2)	metric (quadratic, normalized with Splus function *poly(age,2)*)			
T O T A L		1020	355	1375

Table 6.1: Description of Covariates

Number in Figure 6.2	Postal Code	Number in Figure 6.2	Postal Code	Number in Figure 6.2	Postal Code
1	80331	26	80805	51	81477
2	80333	27	80807	52	81479
3	80335	28	80809	53	81539
4	80336	29	80933	54	81541
5	80337	30	80935	55	81543
6	80339	31	80937	56	81545
7	80469	32	80939	57	81547
8	80538	33	80992	58	81549
9	80539	34	80993	59	81667
10	80634	35	80995	60	81669
11	80636	36	80997	61	81671
12	80637	37	80999	62	81673
13	80638	38	81241	63	81675
14	80639	39	81243	64	81677
15	80686	40	81245	65	81679
16	80687	41	81247	66	81735
17	80689	42	81249	67	81737
18	80796	43	81369	68	81739
19	80797	44	81371	69	81825
20	80798	45	81373	70	81827
21	80799	46	81375	71	81829
22	80801	47	81377	72	81925
23	80802	48	81379	73	81927
24	80803	49	81475	74	81929
25	80804	50	81476		

Table 6.2: Connection of Postal Codes with Its Numeration, Used in the Models

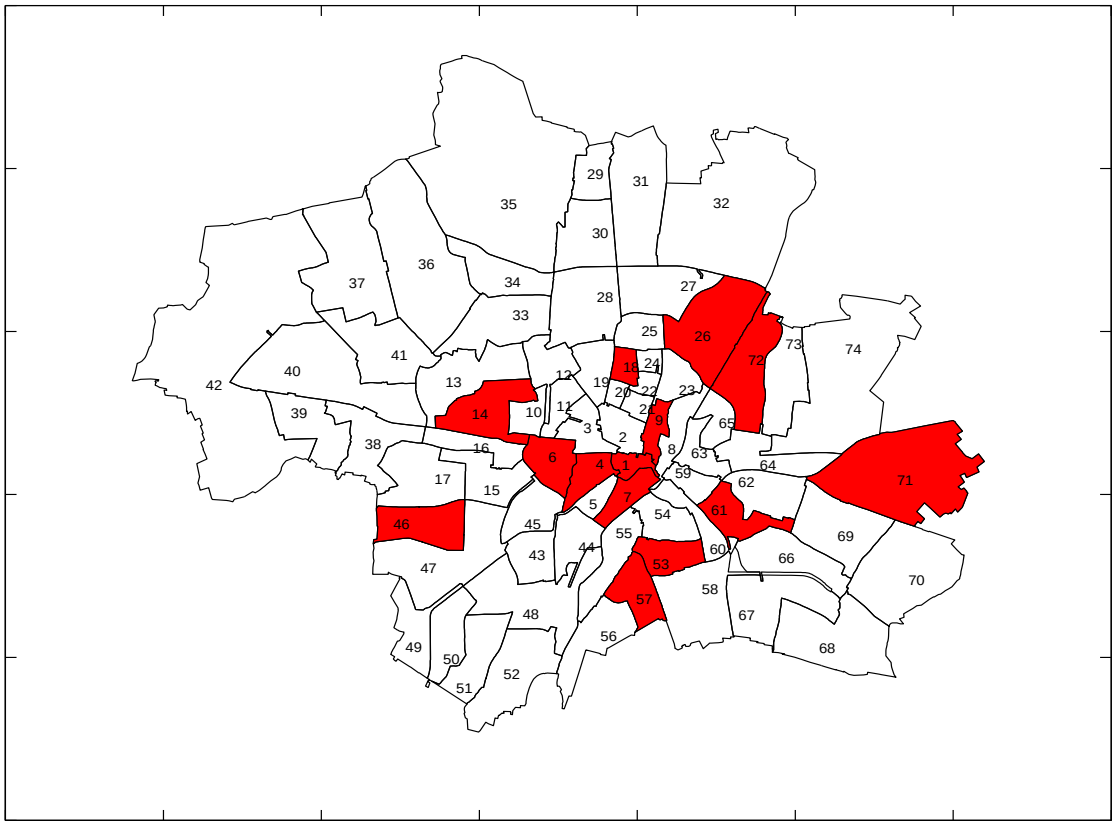

Figure 6.2: Enumeration of Postal Codes of Munich, Germany, Used in the Models; red: Postal Codes Have no Data

spatial and cluster effects using a logit model, provided by the function *glm()* in Splus, are given in Table 6.3.

	Value	Std. Error	t value
(Intercept)	-3.54	0.81	-4.33
day.type.weekend	2.15	0.38	5.65
household.type.single.parent	3.60	1.05	3.41
household.type.not.single	1.04	0.24	4.28
pers.income.middle	2.38	0.68	3.49
pers.income.high	2.70	0.72	3.72
distance.middle	-0.57	0.31	-1.80
distance.far	1.05	0.61	1.72
way.alone.not.alone	1.90	0.40	4.71
poly(age, 2)1	27.84	7.87	3.53
poly(age, 2)2	1.81	7.89	0.23
usage.second.user	0.70	0.61	1.14
usage.not.user	-4.66	0.69	-6.73
net.card.no	4.75	0.98	4.81
sex.female	0.45	0.29	1.54
day.time.night	-0.73	0.43	-1.69
way.alone.not.alone:net.card.no	-2.42	0.43	-5.52
usage.second.user:sex.female	-2.27	0.51	-4.41
usage.not.user:sex.female	0.43	0.56	0.77
way.alone.not.alone:usage.second.user	1.39	0.49	2.80
way.alone.not.alone:usage.not.user	2.69	0.55	4.88
distance.middle:usage.second.user	-1.44	0.59	-2.43
distance.far:usage.second.user	-1.97	0.91	-2.15
distance.middle:usage.not.user	1.06	0.62	1.70
distance.far:usage.not.user	-1.57	1.07	-1.46
day.type.weekend:net.card.no	-1.42	0.47	-3.01
usage.second.user:day.time.night	7.87	6.41	1.22
usage.not.user:day.time.night	0.27	0.95	0.29
sex.female:day.time.night	2.89	1.15	2.50
pers.income.middle:net.card.no	-1.21	0.96	-1.25
pers.income.high:net.card.no	-2.51	0.98	-2.55
distance.middle:poly(age, 2)1	-30.34	8.69	-3.49
distance.far:poly(age, 2)1	-0.35	13.73	-0.02
distance.middle:poly(age, 2)2	-10.46	9.09	-1.15
distance.far:poly(age, 2)2	-5.81	14.35	-0.40
day.type.weekend:poly(age, 2)1	2.06	7.52	0.27
day.type.weekend:poly(age, 2)2	-32.54	9.07	-3.58

Table 6.3: Parameter Estimates Ignoring Spatial and Cluster Effects Using a Logit Model Applying the Function *glm()* in Splus

6.2 Results

We present briefly the results for 11 models we chose. Model 1, based on Model (3.2) includes only fixed and spatial effects. Models 2 — 5, also based on Model (3.2) include additionally group cluster effects with different choices of clusters. For each of these 5 models 25000 MCMC iterations of the algorithm presented in Section 3.2 were run and every 25^{th} iteration was recorded, giving acceptable low autocorrelations (not shown). In Models 6, 8, 10, based on Logit Model (4.9) individual cluster effects are modeled. The algorithm for parameter estimation is given in Section 4.2.1. Finally, Models 7, 9, 11 have the same cluster choice as Models 6, 8, 10, respectively, but are based on Probit Model (4.7) with individual cluster effects. For probit models we run 20000 iterations (and recorded every 20^{th} iteration) of the MCMC algorithm based on latent variable representation (4.8), presented in Section 4.2.3. We note, that 10 pilot runs (5 pilot runs for probit models) with 300 iterations per each pilot run, which we used to determine optimal proposal standard error values for the MH-step (see Section 3.2), provided in each model good acceptance rates for the MH steps between $30\% - 60\%$ and, besides this, the sufficient "burn in" phase.

As a starting point for the choice of fixed effects for each of the 11 models we used the covariates identified in the standard logit model presented in Table 6.3. So we need to estimate 36 regression parameters $\alpha_1, \ldots, \alpha_{36}$. The intercept effect is modeled within the spatial and cluster part. As prior distributions for $\alpha_1, \ldots, \alpha_{36}$ we chose independent normal distributions with zero mean and standard error equal to 5. We consider an interaction as insignificant when the corresponding estimated 90% credible interval contains the zero value for all interaction terms. If an interaction is found to be insignificant, then the corresponding terms will be removed and the model parameters will be estimated again using the appropriate MCMC algorithm. Continuing with this procedure we arrive at a model where all interactions are significant. The estimated posterior modes for the regression parameters are shown for all 11 models in Table 6.8 (main effects) and in Table 6.9 (interactions).

6.2.1 Model with only Fixed and Spatial Effects (Model 1)

As noted before, this model is based on Model (3.2). The independent binary responses $Y_i \sim \text{Be}(p_i),\ i = 1, \ldots, 1375$ are modeled here as follows:

Model 1:

$$\theta_i := \text{logit}(p_i) = \mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)}$$

$$b_j | b_{j'}, j \neq j' \sim N\left(\frac{\phi}{1+|\phi|N_j}\sum_{j\sim j'} b_{j'}\,,\, \frac{(1+|\phi|)\tau^2}{1+|\phi|N_j}\right),\ j = 1, \ldots, 74\,.$$

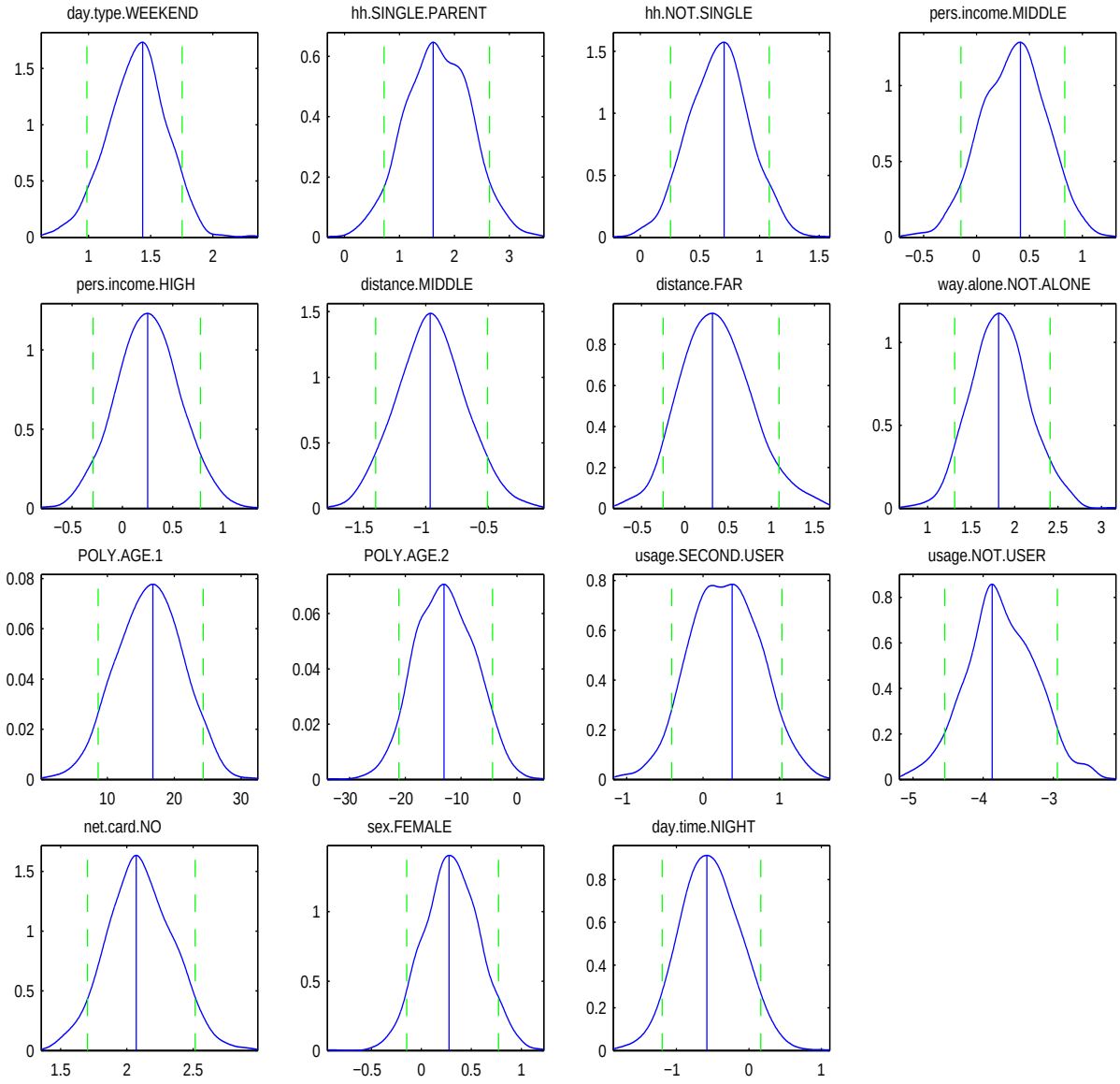

Figure 6.3: Estimated Posterior Densities for Main Effects in Model 1 (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

We chose as prior distribution for the spatial hyperparameter $\psi = \frac{\phi}{1+|\phi|}$ an uniform distribution on $(-1, 1)$, while for the other spatial hyperparameter τ^2 we chose an non-informative prior, i.e. $[\tau^2] \propto 1$. The corresponding estimated posterior densities for regression parameters are presented in Figure 6.3 (main effects) and Figure 6.4 (interactions), respectively. Vertical solid lines indicate estimated posterior modes, while vertical dashed lines give 90% credible intervals. For the spatial effects estimates we show on Figure 6.6 three maps with estimated posterior modes, medians and means, respectively. On the left bottom map we show which spatial effects significantly differ from zero. Finally, we note

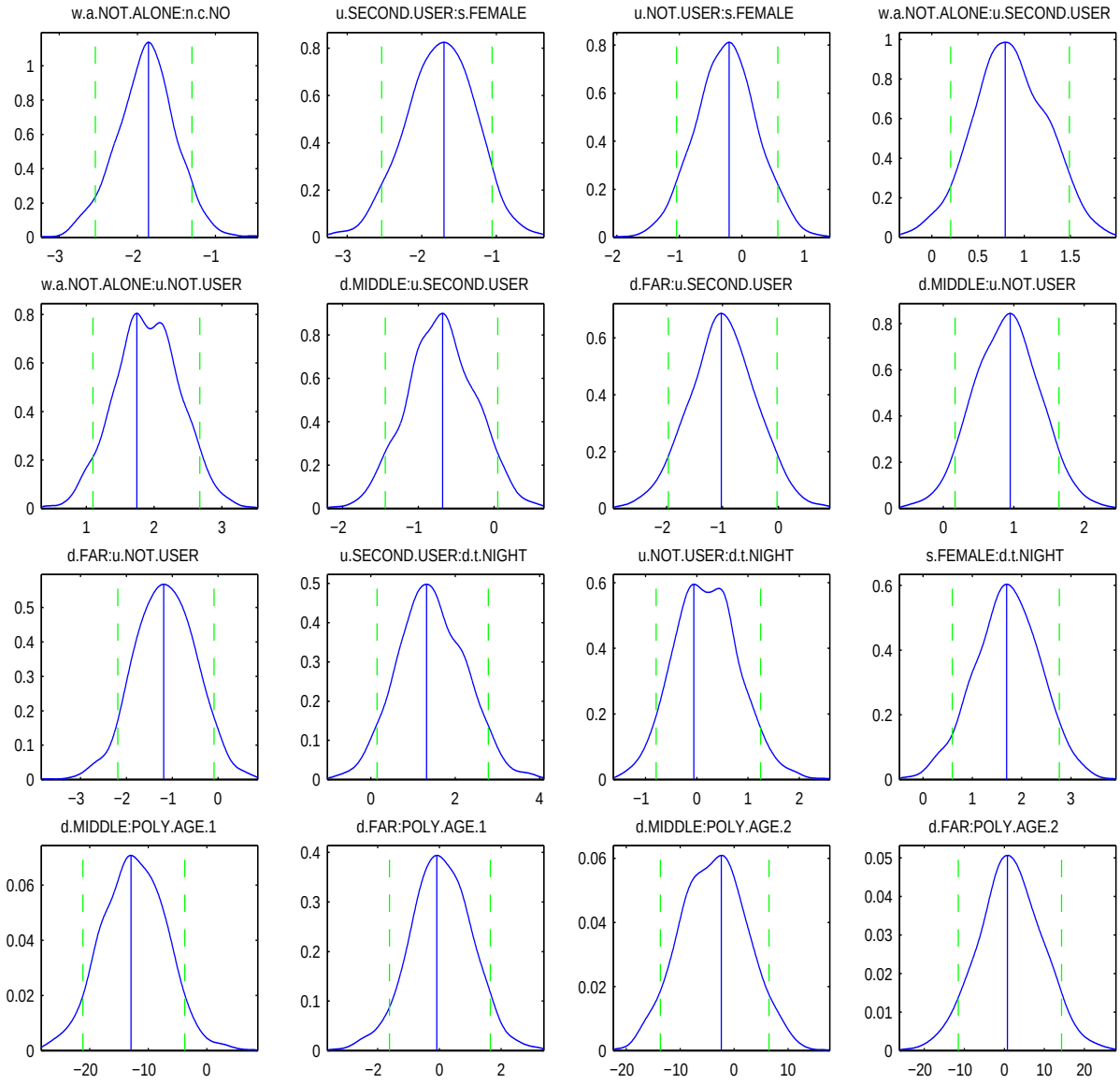

Figure 6.4: Estimated Posterior Densities for Interaction Effects in Model 1 (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

that spatial effects of the postal codes with no observations are insignificant.

In Figure 6.5 we present estimated posterior densities for the hyperparameters τ^2 and ψ, respectively. Note that the spatial dependence parameter ψ is negative, which indicates that large positive spatial effects in an area can be surrounded by negative spatial effects and vice versa. This behavior is seen on the left bottom map of Figure 6.6.

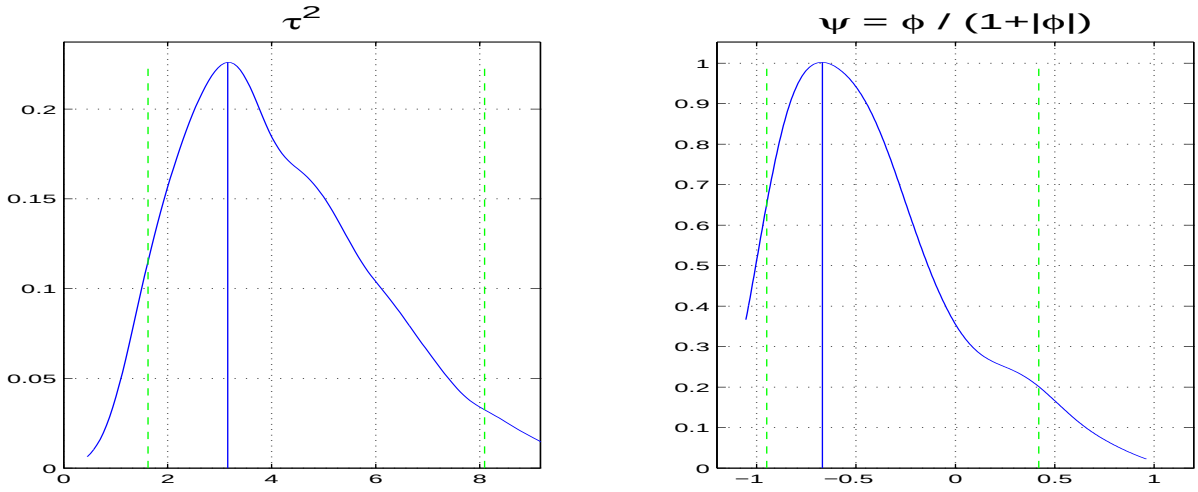

Figure 6.5: Estimated Posterior Densities of Spatial Hyperparameters in Model 1 (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

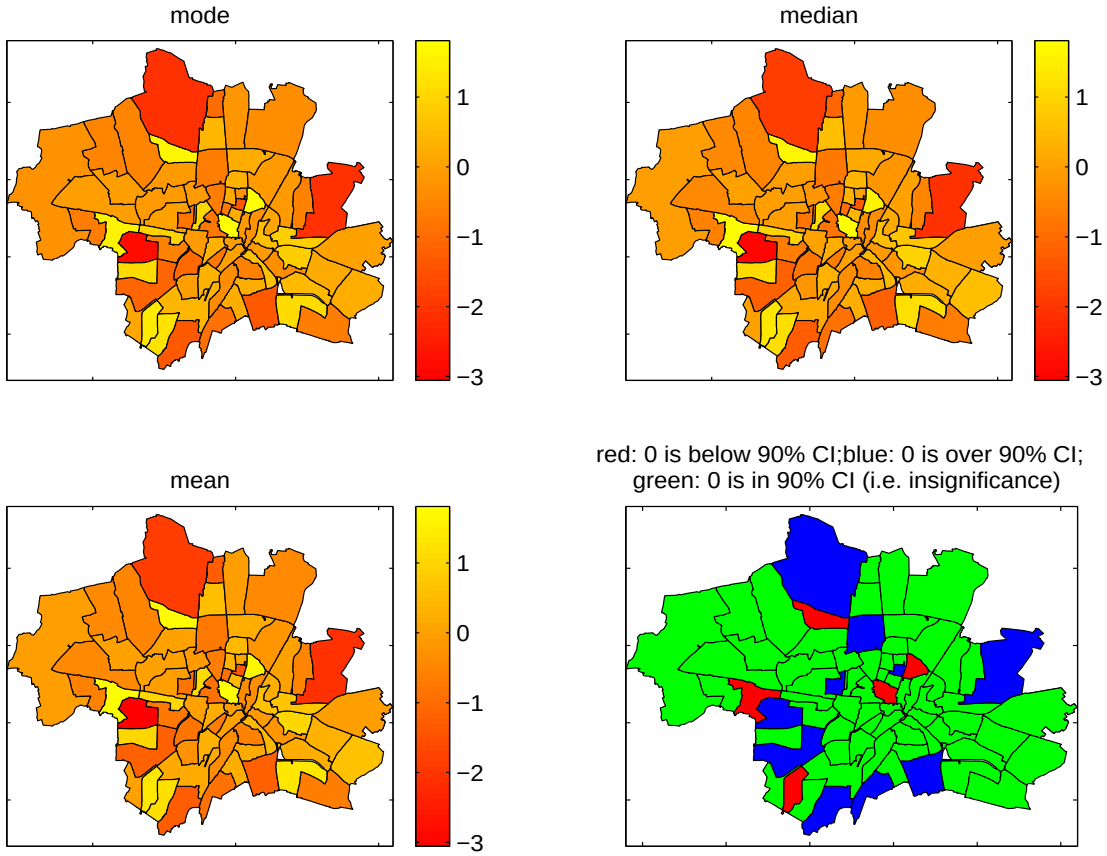

Figure 6.6: Estimated Spatial Effects $\hat{b}_j$, $j = 1, \ldots, 74$ in Model 1.

6.2.2 Models with Fixed, Spatial and Group Cluster Effects (Models 2 - 5)

These logit models are based on the Model (3.2) discussed in Section 3. In particular we consider the following setup:

Models 2 - 5:

$$\theta_i := \text{logit}(p_i) = \mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)} + c_{m(i)}$$
$$b_j|b_{j'}, j \neq j' \sim N\left(\frac{\phi}{1+|\phi|N_j}\sum_{j\sim j'} b_{j'}\,,\,\frac{(1+|\phi|)\tau^2}{1+|\phi|N_j}\right),\ j = 1,\ldots,74$$
$$c_m \sim N(0,\sigma_c^2),\ m = 1,\ldots,M\,.$$

For the cluster hyperparameter σ_c^2 we choose an inverse gamma prior distribution given by $\sigma_c^2 \sim IG(3, 0.5)$, while prior choices for fixed and spatial parameters remain the same as in Model 1. Only in Model 2, in order to avoid numerical problems (clustering around border values -1 and 1) we chose $[\psi] \propto (1-|\psi|)^{0.5}$ instead of $[\psi] \propto 1$ on the interval $(-1, 1)$. The posterior centrality estimates of the spatial τ^2, ψ and cluster σ_c^2 hyperparameters and their 90% credible intervals are given in Table 6.6.

In **Model 2** we chose as cluster groups the 74 postal codes. Therefore both structured ($b_j,\ j = 1,\ldots,74$) and unstructured ($c_j,\ j = 1,\ldots,74$) spatial effects are included in Model 2. In the Figure 6.7 we present spatial maps with estimated posterior modes and means for the structured spatial effects b_j (top row) and unstructured spatial effects c_j (middle row). In the bottom row we present modes and medians from the estimated posterior distribution of the sum $b_j + c_j$ of structured and unstructured spatial effects. From the right column of maps it is remarkable that both structured and unstructured effects are insignificant, while their sum is, and form a similar spatial pattern as in Model 1. Therefore it is not surprising that the posterior density of ψ, in particular the posterior mode estimate in Models 1 and 2 are also similar (see Figure 6.8).

In **Model 3** five clusters are formed by classifying numbers of trips per household into 5 groups, i.e. $m = 1,\ldots,5$. The lowest class has the highest numbers of trips, the highest class has the fewest numbers of trips. The distribution of trips into clusters are given in Table 6.4. In Figure 6.9 we present corresponding estimated posterior densities of the group cluster effects $c_m,\ m = 1,\ldots,5$. The cluster effect is significant (marked with *), if its 90% credible interval does not include zero. Note that cluster effects for households with large numbers of trips are positive and cluster effects for households with few numbers of trips are negative.

Finally Figure 6.17 presents map with the estimated spatial effects for this model.

In **Model 4** we use 12 clusters instead of 5 clusters formed by the numbers of trips a household has taken. In Figure 6.10 we show estimated posterior densities for the corresponding cluster effects $c_m,\ m = 1,\ldots,12$. We emphasize the similar behavior of the

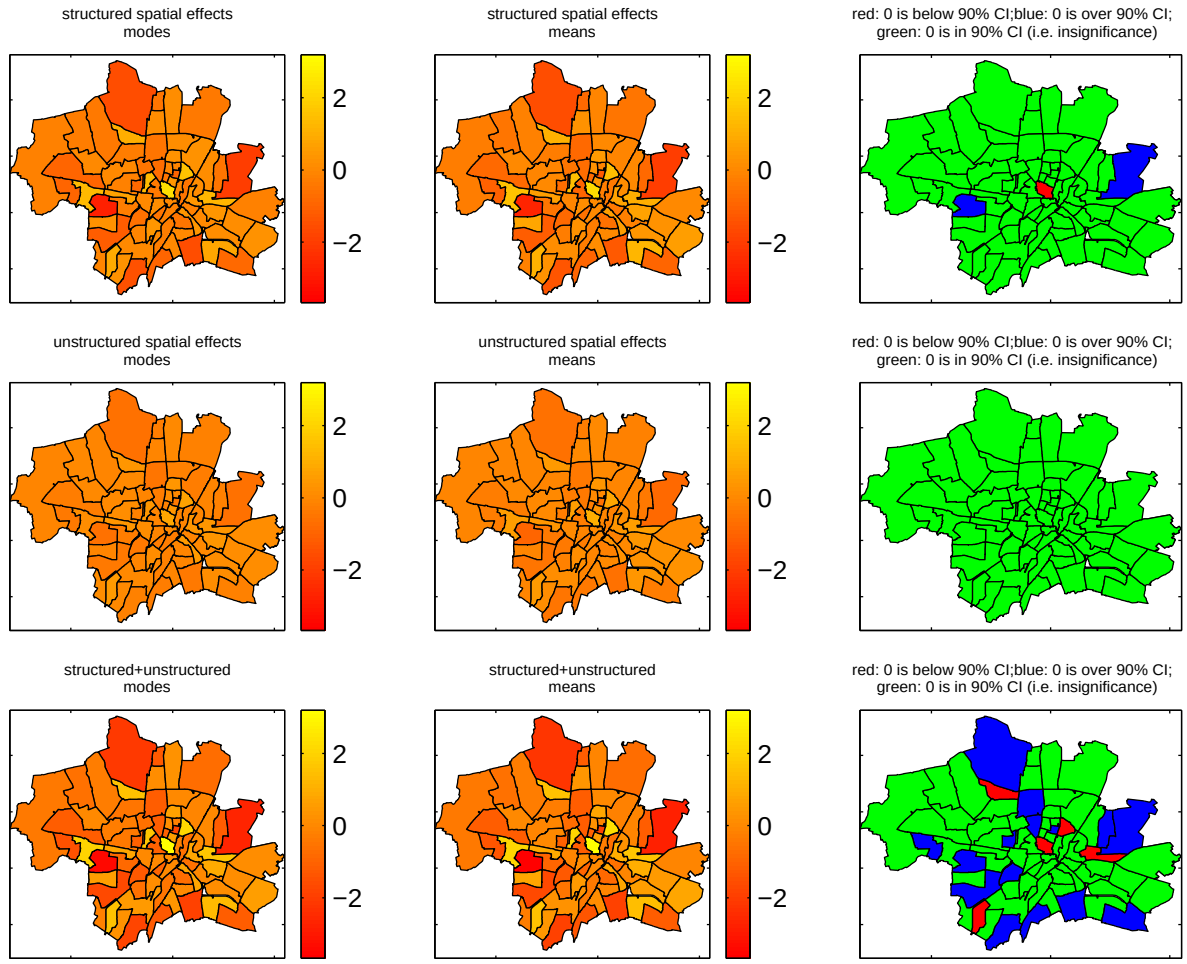

Figure 6.7: Estimated Spatial Effects in Model 2: Structured $\hat{b}_j, j = 1 : 74$ (top), Unstructured $\hat{c}_j, j = 1 : 74$ (middle) and their Sum $\widehat{b_j + c_j}, j = 1 : 74$ (bottom)

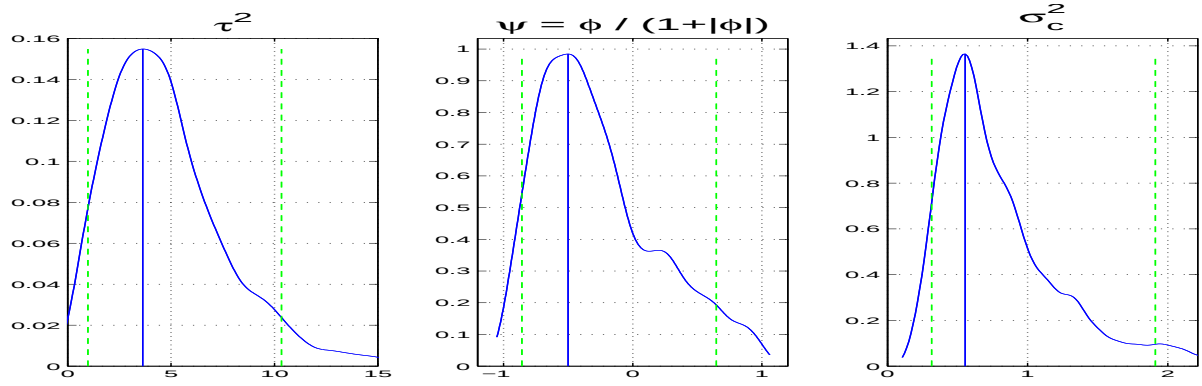

Figure 6.8: Estimated Posterior Densities of Hyperparameters in Model 2. (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

cluster effects, as mentioned before for Model 3.

The last model with group cluster effects, **Model 5**, has 5 clusters, formed by classifying numbers of trips per person (instead of household, as before) into 5 groups. The distribution of trips into clusters is given in Table 6.5. We note that only the 4^{th} and the 5^{th} cluster effect (with fewest numbers of trips) are significant in this model. Both have

Cluster	**Description**		**Total**
1^{st}	Households which conducted	≥ 23 trips	275 trips
2^{nd}	Households which conducted	$16-22$ trips	296 trips
3^{rd}	Households which conducted	$12-15$ trips	250 trips
4^{th}	Households which conducted	$8-11$ trips	275 trips
5^{th}	Households which conducted	≤ 7 trips	279 trips

Table 6.4: Distribution of Trips into Clusters for Model 3

negative values, namely around -1, i.e. the probability to use public transport for the corresponding trips is higher. We omit the corresponding density plots to save space. For the last 2 models we also omit figures presenting maps with the estimated spatial effects

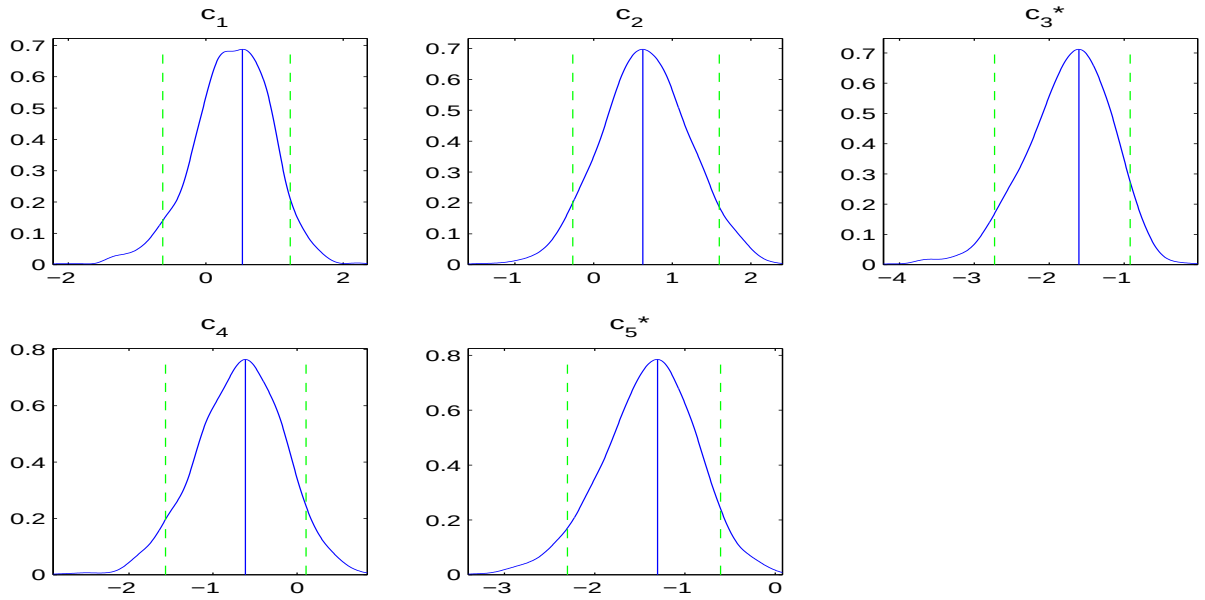

Figure 6.9: Estimated Posterior Densities of Group Cluster Effects c_m, $m = 1, \ldots, 5$ in Model 3. (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

since their spatial patterns are similar to the one of Models 1,3 or Model 2 when the joint effect of structured and unstructured spatial components is considered. Therefore

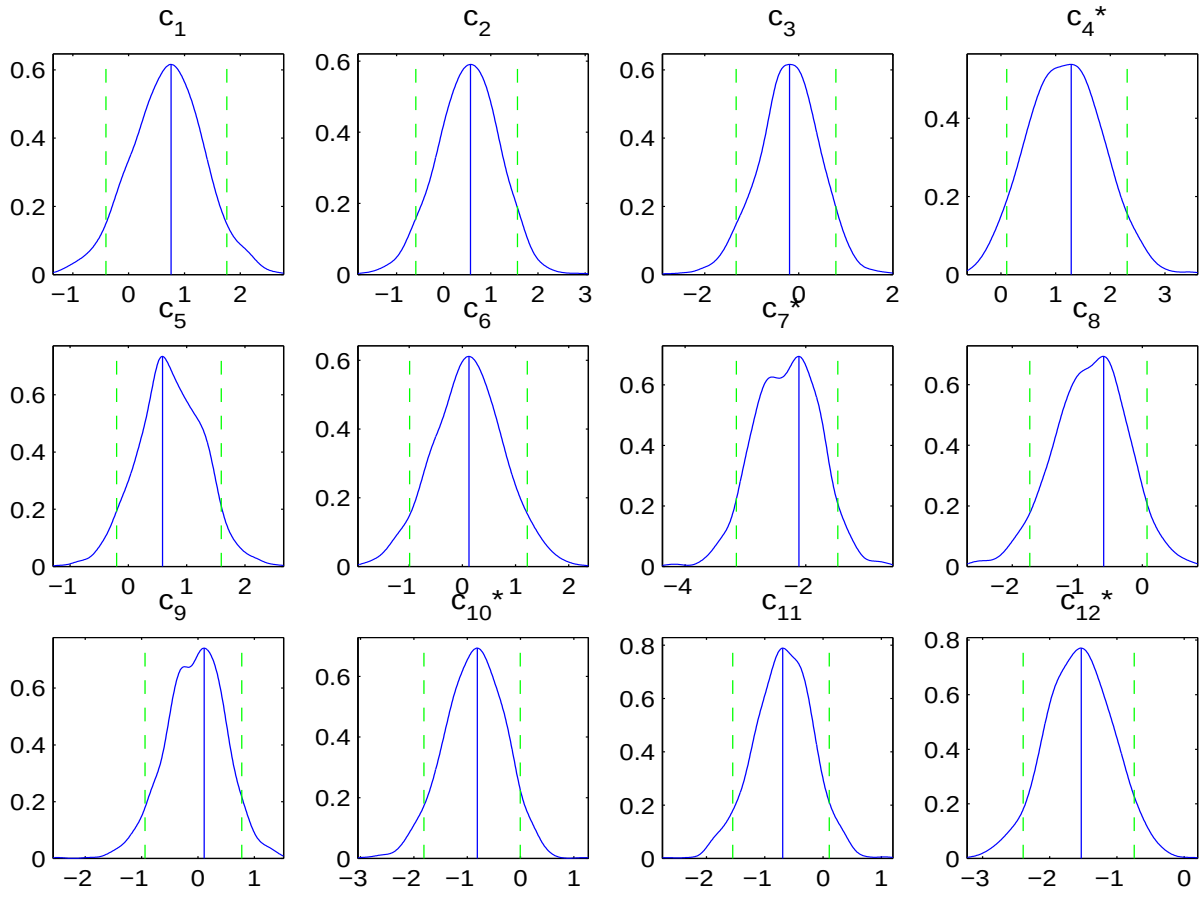

Figure 6.10: Estimated Posterior Densities of Group Cluster Effects c_m, $m = 1, \ldots, 12$ in Model 4. (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

the posterior density of the spatial dependence parameter ψ, in particular the estimate (the posterior mode) also remains similar (not shown for Models 3,4 and 5). This can be seen in Table 6.6 where posterior centrality estimates and 90% credible intervals for the hyperparameters are given.

6.2.3 Models with Fixed, Spatial and Individual Cluster Effects (Models 6 - 11)

We assume here that natural parameters θ_i of the independent binary responses $Y_i \sim \text{Be}(p_i)$, $i = 1, \ldots, 1375$ are generated by the following model (class introduced

Cluster	**Description**		**Total**
1^{st}	Persons which conducted	≥ 12 trips	309 trips
2^{nd}	Persons which conducted	$9-11$ trips	301 trips
3^{rd}	Persons which conducted	$7-8$ trips	240 trips
4^{th}	Persons which conducted	$5-6$ trips	285 trips
5^{th}	Persons which conducted	≤ 4 trips	240 trips

Table 6.5: Distribution of Trips into Clusters in Model 5

Model	**Parameter**	**Mode**	**Mean**	**Median**	**90%**	**CI**
2	ψ	-0.500	-0.271	-0.372	-0.857	0.646
	τ^2	3.628	4.777	4.313	0.981	10.335
	σ_c^2	0.554	0.836	0.678	0.315	1.912
3	ψ	-0.541	-0.422	-0.446	-0.930	0.149
	τ^2	6.262	9.124	8.233	3.358	18.417
	σ_c^2	0.802	1.270	1.076	0.486	2.797
4	ψ	-0.507	-0.516	-0.538	-0.954	0.031
	τ^2	6.293	8.299	7.452	3.194	16.067
	σ_c^2	0.880	1.272	1.122	0.589	2.398
5	ψ	-0.874	-0.543	-0.594	-0.956	0.058
	τ^2	4.025	5.298	4.777	2.020	9.685
	σ_c^2	0.526	0.753	0.646	0.324	1.585

Table 6.6: Point and Interval Estimates for the Hyperparameters in Models 2 - 5 (with Group Cluster Effects)

in Section 4.1):

Models 6 - 11:

$$\theta_i = \mathbf{x_i^t}\boldsymbol{\alpha} + b_{j(i)} + c_{m(i),k(i)}, \text{ where } \theta_i = \log\left(\frac{p_i}{1-p_i}\right) \text{ or } \theta_i = \Phi^{-1}(p_i)$$
$$b_j|b_{j'}, j \neq j' \sim N\left(\frac{\phi}{1+|\phi|N_j}\sum_{j\sim j'} b_{j'}\,, \frac{(1+|\phi|)\tau^2}{1+|\phi|N_j}\right),\ j = 1,\ldots,74$$
$$\forall m = 1,\ldots,M: \quad c_{m,k} \sim N(0,\sigma_m^2),\ k = 1,\ldots,K_m\,.$$

Recall that for both logit and probit link functions this model is unidentifiable. Therefore we reparametrize it (according to (4.6)) to achieve identifiability. For the probit link this leads to Model (4.7) given by

$$Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with}$$
$$p_i = \mathbf{P}\{Y_i = 1|\mathbf{x_i}, \boldsymbol{\alpha}', b'_{j(i)}, \sigma'^2_{m(i)}\} = \begin{cases} \Phi\left(\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}\right) & \text{if } m(i) = 1 \\ \Phi\left(\frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}}\right) & \text{if } m(i) = 2,\ldots,M \end{cases}$$

and for the logit link to Model (4.9):

$$Y_i|p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with}$$
$$\log\left(\frac{p_i}{1-p_i}\right) = \begin{cases} \mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)} & \text{if } m(i) = 1 \\ \frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}} & \text{if } m(i) = 2,\ldots,M\,. \end{cases}$$

Models (4.9) and (4.7) contain one parameter less than the original Models (4.1) and (4.2), respectively, and are identifiable and therefore suitable for the parameter estimation.

As before, for the spatial variance parameter we chose a flat prior $[\tau^{2'}] \propto 1$, while for the spatial dependence parameter in order to avoid numerical problems (clustering around border values -1 and 1) we take the prior distribution $[\psi'] \propto (1-|\psi'|)^{0.5}$ instead of $\psi \sim Uni(-1,1)$, which is common choice for models with group cluster effects. In Probit Models 7, 9, 11 the cluster parameters $\sigma_2^{2'},\ldots,\sigma_M^{2'}$ have an inverse gamma prior distribution given by $IG(3,0.5)$ (with expectation and variance equal to 1). This choice allows direct Gibbs sampling for these parameters. In Logit Models 6, 8, 10 for cluster components we simulate Markov Chain from posterior distribution of $\sigma'_2,\ldots,\sigma'_M$ with the prior choice taken as normal $N(1,1)$ truncated to $(0.2,+\infty)$.

In **Models 6 and 7** (with logit and probit link function respectively) we use 3 clusters formed by household type. In both models we omitted this covariate as a fixed effect. In Figure 6.11 we present estimated posterior densities for the cluster parameters $\sigma_2^{2'}$ and $\sigma_3^{2'}$. Note that $\sigma_m^{2'}$ is significantly different to $\sigma_1^{2'} = 1$ if $1 \notin 90\%\,CI$ for $\sigma_m^{2'}$, which is the case for all $m \geq 2$ in Model 7.

Logit **Model 8** and Probit **Model 9** have the same cluster choice, as Model 3 (see Table 6.4). Note that the fewest heterogeneity is within the group with the largest numbers

of trips per household (see estimates of the cluster parameters $\boldsymbol{\sigma}'$ for Model 8 or $\boldsymbol{\sigma}^{\mathbf{2}'}$ for Model 9 in Table 6.7). Five clusters in Logit **Model 10** and Probit **Model 11** coincide with the cluster choice in Model 5 (see Table 6.5). Heterogeneity within the group with the fewest numbers of trips per person (the 5th cluster) is the largest, so we observe the similar behavior as in Models 8 and 9, where clusters are formed by numbers of trips per household. We omit figures presenting maps with the estimated spatial effects for Models 6 - 11 with individual cluster effects because of similarity with the former

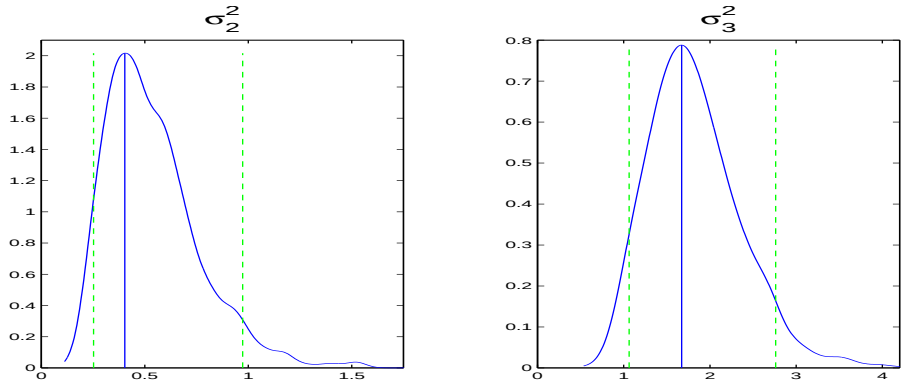

Figure 6.11: Estimated Posterior Densities of Individual Cluster Effects $\sigma_2^{2'}, \sigma_3^{2'}$ in Model 7. (Solid Line = Estimated Posterior Mode, Dashed Line = 90% CI)

model classes. To save space we also do not present plots with the estimated posterior parameter densities. The posterior centrality estimates of the spatial hyperparameters $\tau^{2'}, \psi'$ and cluster parameters $\sigma_2', \ldots, \sigma_M'$ and their 90% credible intervals for Models 6 - 11 are given in Table 6.7, while estimates for the fixed effects $\boldsymbol{\alpha}'$ are given in Table 6.8 (main effects) and in Table 6.9 (interactions). Posterior mode estimates are marked with *, when the corresponding parameter is insignificant, i.e. the 90% credible interval contains zero. If all terms of some interaction effects were insignificant, the model was reduced on this interaction and the model parameters were estimated again using the appropriate MCMC algorithm. Those interactions are marked with "*n.r.*", which stands for "not represented" in the model.

Model	Parameter	Mode	Mean	Median	90%	CI
6	ψ	−0.468	−0.396	−0.418	−0.870	0.181
	$\tau^{2'}$	4.861	6.854	5.931	2.553	14.196
	σ_2'	0.277	0.484	0.430	0.226	0.921
	σ_3'	1.439	1.461	1.443	1.068	1.943
7	ψ	−0.525	−0.408	−0.444	−0.874	0.111
	$\tau^{2'}$	1.685	2.283	2.095	0.903	4.418
	$\sigma_2^{2'}$	0.404	0.540	0.497	0.252	0.973
	$\sigma_3^{2'}$	1.672	1.831	1.763	1.062	2.763
8	ψ	−0.410	−0.413	−0.422	−0.865	0.075
	$\tau^{2'}$	10.769	17.101	14.799	6.002	36.512
	σ_2'	0.922	1.010	0.973	0.648	1.464
	σ_3'	2.842	2.951	2.913	2.240	3.734
	σ_4'	1.430	1.486	1.459	1.078	2.019
	σ_5'	1.822	1.797	1.789	1.313	2.343
9	ψ	−0.512	−0.382	−0.412	−0.854	0.195
	$\tau^{2'}$	3.588	5.380	4.678	1.817	11.032
	$\sigma_2^{2'}$	0.797	0.906	0.845	0.477	1.534
	$\sigma_3^{2'}$	5.844	6.949	6.477	3.707	11.627
	$\sigma_4^{2'}$	1.543	1.958	1.837	0.996	3.370
	$\sigma_5^{2'}$	2.743	3.174	2.953	1.600	5.474
10	ψ	−0.476	−0.403	−0.439	−0.876	0.199
	$\tau^{2'}$	7.538	9.468	8.232	3.167	19.895
	σ_2'	1.027	1.058	1.041	0.752	1.430
	σ_3'	1.168	1.180	1.166	0.797	1.610
	σ_4'	1.271	1.300	1.287	0.897	1.768
	σ_5'	1.553	1.681	1.642	1.196	2.255
11	ψ	−0.350	−0.369	−0.383	−0.867	0.2220
	$\tau^{2'}$	1.773	2.517	2.182	0.849	5.2690
	$\sigma_2^{2'}$	0.782	0.887	0.846	0.510	1.4190
	$\sigma_3^{2'}$	0.976	1.067	1.009	0.557	1.7630
	$\sigma_4^{2'}$	1.021	1.340	1.250	0.731	2.1950
	$\sigma_5^{2'}$	1.972	2.209	2.088	1.169	3.7700

Table 6.7: Estimated Spatial Hyperparameters and Cluster Parameters for Models 6 - 11 (with Individual Cluster Effects)

Main Effect	Model										
	1	2	3	4	5	6	7	8	9	10	11
DAY TYPE											
WEEKEND	1.44	2.21	2.46	2.52	2.11	2.25	1.19	3.32	1.70	2.78	1.29
HOUSEHOLD											
SINGLE.PARENT	1.61	3.15	3.42	2.92	3.31	n. r.	n. r.	4.24	2.62	3.65	2.17
NOT.SINGLE	0.70	0.68	0.25*	0.27*	0.90	n. r.	n. r.	0.85*	0.40*	0.96	0.44
PERSONAL INCOME											
MIDDLE	0.41*	0.48*	1.63	1.41	0.71*	1.06	0.58	1.62	0.80	0.64*	0.37*
HIGH	0.25*	0.42*	1.27	1.14	0.12*	0.76	0.35*	1.46	0.56	0.24*	0.05*
DISTANCE											
MIDDLE	−0.96	−1.15	−1.06	−1.17	−1.05	−1.29	−0.61	−1.90	−1.04	−1.16	−0.59
FAR	0.32*	0.81*	0.98*	0.83*	0.97*	1.21*	0.64*	0.78*	0.65*	0.85*	0.33*
WAY ALONE											
NOT.ALONE	1.82	2.09	2.07	2.30	1.93	2.17	1.36	3.21	1.76	2.30	1.20
AGE											
POLY.AGE.1	16.80	8.73	11.64	11.53	9.95	6.11	5.57	7.97	8.82	9.81	7.46
POLY.AGE.2	−13.07	−8.96	−9.03	−8.64	−9.67	−8.93	−7.85	−9.69	−8.65	−7.63	−8.03
USAGE											
SECOND.USER	0.38*	1.09*	1.27	1.16*	0.88*	1.11	0.73	1.23*	0.51*	1.51	0.80
NOT.USER	−3.87	−6.41	−6.52	−6.52	−5.90	−6.38	−3.63	−9.99	−5.39	−7.44	−3.95
NET CARD											
NO	2.07	2.67	3.03	3.32	2.72	2.78	1.70	3.90	2.28	3.11	1.64
SEX											
FEMALE	0.28*	0.16*	−0.19*	−0.47*	0.10*	0.30*	0.11*	−0.48*	−0.20*	0.01*	−0.05*
DAY TIME											
NIGHT	−0.58*	−1.02	−1.12	−1.29	−1.13	−1.19	−0.63	−1.99	−1.08	−1.30	−0.64

Table 6.8: Posterior Mode Estimates for Main Effect Parameters

Interaction	Model										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
WAY ALONE:NET CARD											
NOT.ALONE:NO	−1.86	−2.39	−2.37	−2.76	−2.37	−2.54	−1.54	−3.10	−1.69	−2.53	−1.48
USAGE:SEX											
SECOND.USER:FEMALE	−1.70	−2.13	−2.07	−1.81	−2.01	−2.30	−1.45	−2.80	−1.73	−2.50	−1.22
NOT.USER:FEMALE	−0.20*	0.66*	0.58*	0.79*	0.40*	0.26*	0.22*	1.39*	0.62*	0.80*	0.62*
WAY ALONE:USAGE											
NOT.ALONE:SECOND.USER	0.79	1.21	0.80	0.76*	1.22	1.09	0.62	1.20*	0.66*	1.32	0.77
NOT.ALONE:NOT.USER	1.75	3.65	4.19	3.76	3.41	4.35	2.30	5.08	3.06	4.22	2.03
DISTANCE:USAGE											
MIDDLE:SECOND.USER	−0.68*	−1.03	−1.39	−0.97	−1.19	−1.31	−0.74	−1.44*	−0.52*	−1.54	−0.87
FAR:SECOND.USER	−1.02	−2.25	−2.12	−1.72	−2.22	−2.73	−1.52	−2.41	−1.52	−3.61	−1.62
MIDDLE:NOT.USER	0.95	1.68	1.52	1.64	1.27	1.20	0.63	2.47	1.66	1.53	0.89
FAR:NOT.USER	−1.19	−1.19*	−1.55*	−2.01	−1.51*	−2.31	−1.27	−2.68	−1.93	−1.94	−1.15
DAY TYPE:NET CARD											
WEEKEND:NO	n. r.	−0.91	−1.23	−1.23	−1.07	−0.82*	−0.50*	−1.51	−0.83	−1.25	−0.45
USAGE:DAY TIME											
SECOND.USER:NIGHT	1.32	5.01	5.22	6.63	5.71	5.07	3.53	6.17	4.53	5.67	4.96
NOT.USER:NIGHT	−0.06*	0.31*	0.45*	0.38*	0.26*	0.32*	0.56*	0.72*	0.69*	0.68*	0.59*
SEX:DAY TIME											
FEMALE:NIGHT	1.70	2.88	3.36	3.55	3.49	3.02	1.11	2.94	1.65	3.30	1.22
PERSONAL INCOME: NET CARD											
MIDDLE:NO	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
HIGH: NO	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
DISTANCE:AGE											
MIDDLE:POLY.AGE.1	−12.93	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
FAR:POLY.AGE.1	−0.09*	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
MIDDLE:POLY.AGE.2	−2.41*	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
FAR:POLY.AGE.2	0.76*	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
DAY TYPE:AGE											
WEEKEND:POLY.AGE.1	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.
WEEKEND:POLY.AGE.2	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.	n. r.

Table 6.9: Posterior Mode Estimates for Interaction Parameters

6.3 Model Comparison

In this section we would like to compare the goodness of fit for the proposed 11 models with regard to their spatial effects. For this goal we choose as measure D the sum of weighted squared residuals over all postal codes of Munich defined by

$$D(\mathbf{Y}) := \sum_{j=1}^{74} n_j (p_j^{empir} - p_j^{estim})^2 \,, \tag{6.1}$$

where $n_j :=$ number of trips in the j^{th} postal code. Empirical probabilities p_j^{empir} are equal to $p_j^{\text{empir}} := \frac{1}{n_j} \Sigma_{i:j(i)=j} Y_i$, and the posterior probability estimates p_j^{estim} are based on the MCMC run, and equal the mean value over R iterations:

$$p_j^{\text{estim}} := \frac{1}{n_j * R} \sum_{i:j(i)=j} \sum_{r=1}^{R} h^{-1}(\eta_{ir}), \tag{6.2}$$

where $h(\cdot)$ is logit (for Models 1 - 5, 6, 8, 10) or probit (for Models 7, 9, 11) link function, and

$$\eta_{ir} = \begin{cases} \mathbf{x_i^t}\boldsymbol{\alpha}_r + b_{j(i),r} & \text{for Model 1} \\ \mathbf{x_i^t}\boldsymbol{\alpha}_r + b_{j(i),r} + c_{m(i),r} & \text{for Models 2-5} \\ \frac{\mathbf{x_i^t}\boldsymbol{\alpha}'_r + b'_{j(i),r}}{\sigma'_{m(i),r}} & \text{for Models 6-11 .} \end{cases}$$

In Table 6.10 we present value D for all 11 models. According to Table 6.10 the best

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
D	2.35	1.23	0.95	1.02	1.44	1.9	1.7	3.25	2.88	1.84	1.72

Table 6.10: Model Fit Comparison with Regard to Spatial Probabilities

fit with regard to spatial probabilities has Model 3 (with group cluster effects). For this model we present a color map with estimated spatial probabilities over postal codes of Munich (see Figure 6.12, bottom map), which coincides quite well with the map showing the empirical spatial probabilities (see Figure 6.12, top map). This indicates graphically that Model 3 has a reasonably good fit of the data with respect to the spatial resolution. Recall that we presented in Figure 6.1 the map with the number of available trips over postal codes of Munich.

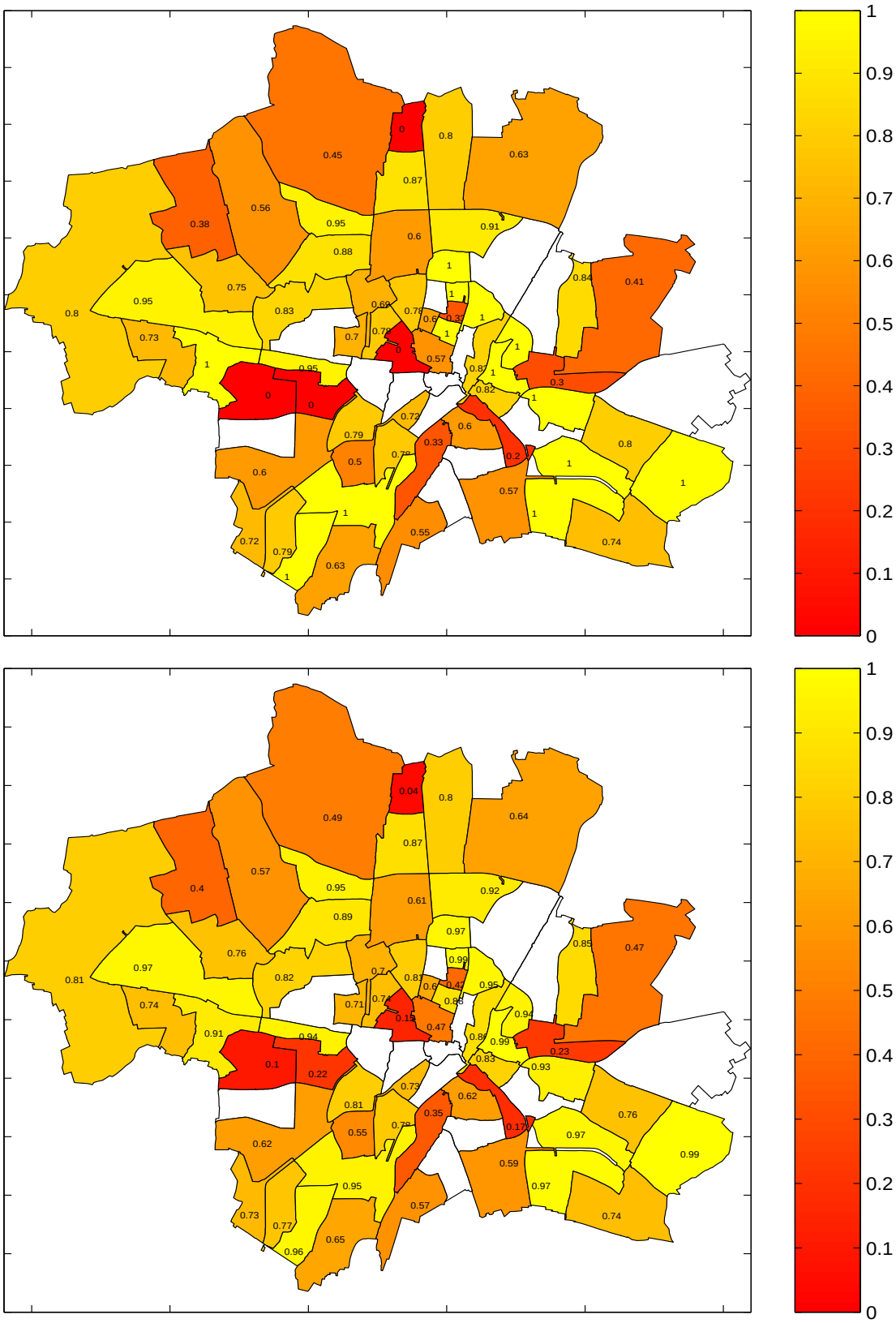

Figure 6.12: Top Map: Observed Probabilities of Individual Transport Use by Postal Codes in Munich, Germany; Bottom Map: Posterior Mean Probability Estimates of Individual Transport Use by Postal Codes in Munich, Germany for Model 3

6.4 Model Interpretation

After model fitting and model selection one is interested in what can be learned about the transport behavior. For this we now investigate the implications of Model 3, which exhibits the best goodness of fit (see Table 6.10) with regard to individual transport probabilities. First we will estimate these probabilities when individual or combinations of two covariates change. The remaining covariates in the model are set to their "most usual values", corresponding to the modus for categorical covariates and median values for quantitative covariates. These "most usual values" are presented in Table 6.11 and are derived from Table 6.1. The only quantitative covariate **AGE** has a median age of 42 years. Since the effect of age is modeled quadratically, we use the numerical more stable orthogonal parameterization **POLY.AGE.1** and **POLY.AGE.2** for the linear and quadratic effect respectively. Since Model 3 includes spatial effects we have to specify a postal code for which we estimate these probabilities. We have chosen postal code area 81377 (see Table 6.2 and Figure 6.2 for exact location), since this postal code area has a large observed number of trips (see Figure 6.1) and the smallest 90% credible interval for its spatial effect. Finally Model 3 contains group cluster effects with regard to the number of trips a household has taken. Since each cluster group contains the similar number of individual trips, for our investigations we chose the last, i.e. the 5th cluster group corresponding to households with ≤ 7 trips (see Table 6.4), which has the smallest 90% credible interval for its cluster effect c_5. Posterior mean estimates for individual transport probabilities for a fixed covariate vector $\mathbf{x}$ for Postal code 81377 (corresponding to b_{47} and the 5th group cluster can be calculated as

$$p_{mean}(\mathbf{x}) := \frac{1}{R-B}\sum_{r=B+1}^{R} p_r(\mathbf{x}) = \frac{1}{R-B}\sum_{r=B+1}^{R}\left(\frac{\exp(\mathbf{x^t}\hat{\boldsymbol{\alpha}}_r + \hat{b}_{47,r} + \hat{c}_{5,r})}{1+\exp(\mathbf{x^t}\hat{\boldsymbol{\alpha}}_r + \hat{b}_{47,r} + \hat{c}_{5,r})}\right), \quad (6.3)$$

where $\hat{\boldsymbol{\alpha}}_r$, $\hat{b}_{47,r}$ and $\hat{c}_{5,r}$ are the r^{th} MCMC estimate of $\boldsymbol{\alpha}$, b_{47} and c_5 respectively. Here R is the total number of MCMC iterations and B is the burn-in. We can also determine 90% credible bounds which are defined as 5% and 95% quantiles of the sample $\{p_r(\mathbf{x})\}_{r=B+1}^{R}$. For "the most usual" trip, which is associated with postal code 81377 and 5th cluster group, the estimated posterior mean probability for taking individual transport is equal to 0.7.

Figure 6.13 gives the estimated posterior mean probability together with 90% credible bounds for choosing individual transport as age changes in Postal code area 81377 and trips associated with the 5th group cluster when the remaining covariates are set to their "most usual value" given in Table 6.11. It is not very surprising that the probability of using a car increases rapidly to an age of about 35 years, remains reasonably stable between 35 years and 65 years and decreases slowly after the 65 years. Young people have

Variable	**Category**	**Value**
DAY TYPE	*WORK DAY* *WEEKEND*	1 0
HOUSEHOLD TYPE	*SINGLE* *SINGLE PARENT* *NOT SINGLE*	0 0 1
PERSONAL INCOME (p.i.)	*NO INCOME* ($<$ 200 *DM*) *MIDDLE* (200 $-$ 3000 *DM*) *HIGH* ($>$ 3000 *DM*)	0 1 0
DISTANCE (d.)	*SHORT* ($\leq$ 3.5 *km*) *MIDDLE* (3.6 $-$ 21.5 *km*) *FAR* ($>$ 21.5 *km*)	0 1 0
WAY ALONE (w.a.)	*ALONE* *NOT ALONE*	1 0
USAGE (u.)	*MAIN USER* *SECONDARY USER* *NOT USER*	1 0 0
NET CARD (n.c.)	*YES* *NO*	0 1
SEX (s.)	*MALE* *FEMALE*	1 0
DAY TIME (d.t.)	*DAY (6 a.m. - 9 p.m.)* *NIGHT (9 p.m. - 6 a.m.)*	1 0
AGE	**POLY.AGE.1** **POLY.AGE.2**	-0.0028648 (42 years) -0.0220355

Table 6.11: Design Vector of the Trip Chosen for the Interpretation of Fixed Effects

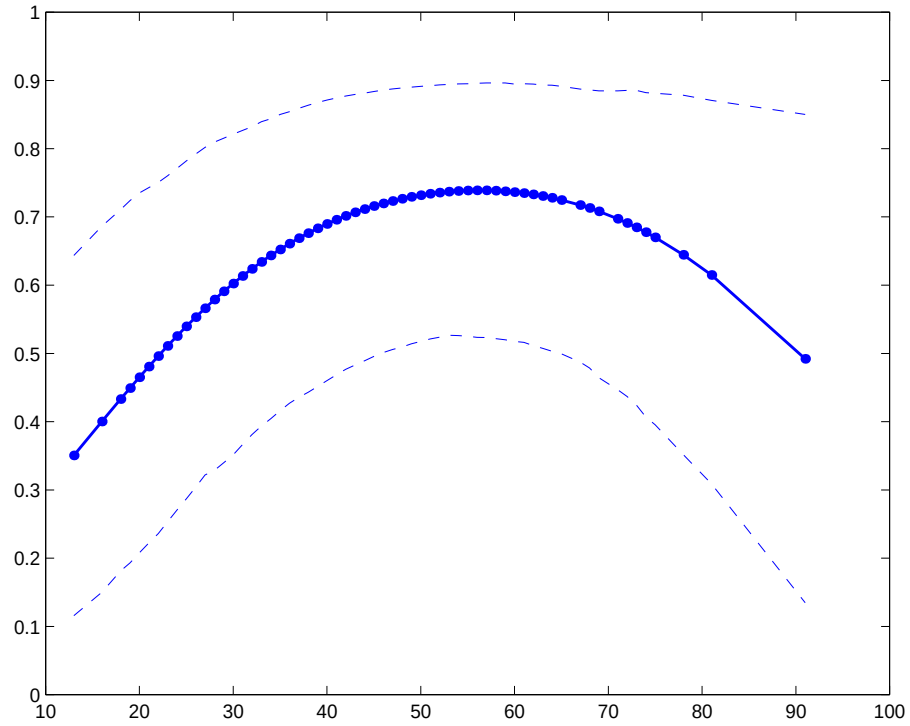

Figure 6.13: Estimated posterior mean probabilities for using an individual transport in Postal code area 81377 and 5th cluster group for different **AGE**, while other covariates are set as in Table 6.11 (dotted lines correspond to 90% credible bounds)

a lower probability to possess a car, while the older people might prefer public transport options.

Note that we can interpret the effect of age directly, since no interaction terms include age. For almost all other covariate effects we have to consider covariate combinations corresponding to interaction terms. In particular note that Model 3 includes 7 interaction terms. In order to interpret effects of the categorical covariates we plot for each of the 7 interactions the estimated posterior mean probabilities (6.3) for using individual transport. We consider now briefly each of the 7 plots in Figures 6.14, 6.15 and 6.16. We expect that people *With Net Card* will use public transport more often. The top plot of Figure 6.14 confirms our expectation. However, the estimated probability to take a car for the persons *With Net Card* increases, if this person travels *Not Alone*. This effect might arise from comfort reasons, if a passenger offers his car, or from costs reasons, if not all passengers have a net card.

In contrast to the covariate **WAY ALONE** the covariate **DAY TYPE** seems to have less interaction with the covariate **NET CARD**, when the top panel of Figure 6.14 is

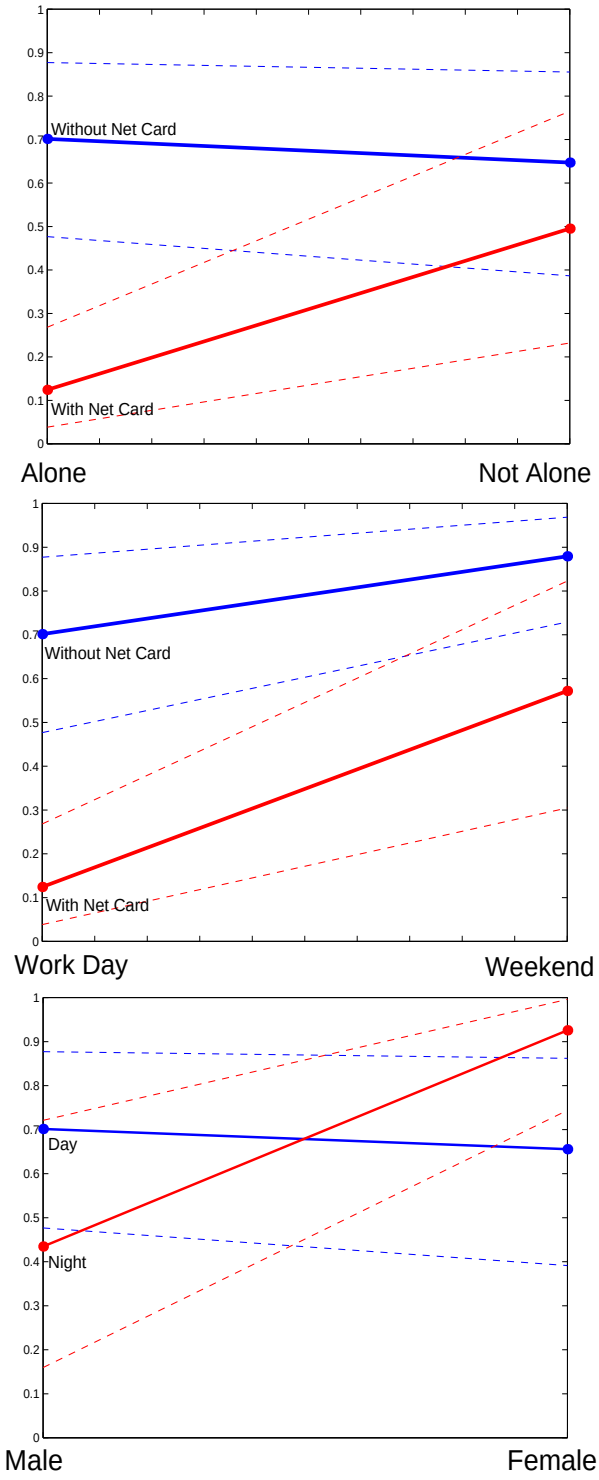

Figure 6.14: Estimated posterior mean probabilities for using an individual transport in Postal code area 81377 and 5th cluster group for different combinations of the covariates which form the interaction, while other covariates are set as in Table 6.11;
Top: WAY ALONE:NET CARD; Middle: DAY TYPE:NET CARD; Bottom: SEX:DAY TIME (dotted lines correspond to 90% credible bounds)

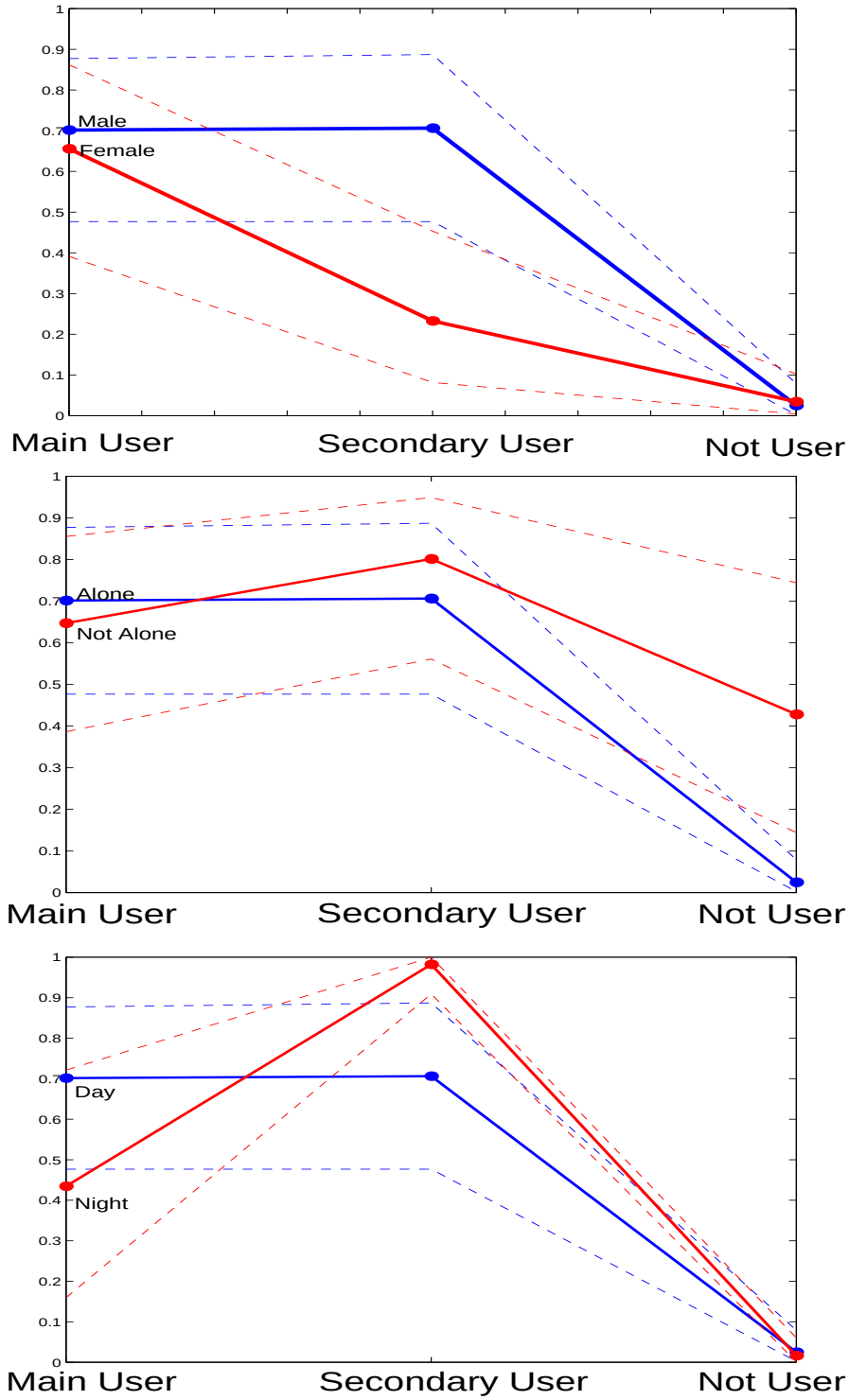

Figure 6.15: Estimated posterior mean probabilities for using an individual transport in Postal code area 81377 and 5th cluster group for different combinations of the covariates which form the interaction, while other covariates are set as in Table 6.11;
Top: USAGE:SEX; Middle: USAGE:WAY ALONE; Bottom: USAGE:DAY.TIME (dotted lines correspond to 90% credible bounds)

compared to the middle panel of Figure 6.14. Further the estimated probability to take a trip using individual transport for a person *With Net Card* increases from *Work Day* to *Weekend* similarly as from travelers *Alone* to *Not Alone*.

The bottom plot of Figure 6.14 shows strong interaction between the covariates **SEX** and **DAY TIME**. During *Day (6 a.m. - 9 p.m.)* both *Males* and *Females* use public transport with similar probability. At *Night (9 p.m. - 6 a.m.)* this probability decreases for *Males*. This might reflect the avoidance of car usage after alcohol consumption. However it rapidly increases for *Females* almost up to 1. An explanation might be that women are afraid to use public transport at night because of low usage and deserted stops.

The covariate **USAGE** is involved with the next 4 interactions effects considered. The corresponding plots are presented in Figures 6.15 and 6.16. The "dominant" profile curve of the transport choice, which corresponds to "the most usual" trip for 3 different categories of **USAGE** are *Male* on the top panel of Figure 6.15, *Alone* on the middle panel of Figure 6.15, *Day* on the bottom panel of Figure 6.15 and *Middle Distance* on Figure 6.16.

The top panel of Figure 6.15 shows the estimated posterior mean probabilities for different **USAGE** and **SEX** combinations. As expected, the probability to take individual transport decreases for *Not Users* almost down to 0. The profile curve for *Female* differs from the profile curve for *Male* only for *Secondary Users*, where this probability for *Females* is only about half that of *Males*. This shows that *Males* are much more likely to use the car when they are secondary users compared to *Females*. Further, both *Male Main Users* and *Secondary Users* use the car in about 70% of their trips.

We consider now the 6 subgroups formed by the levels of **USAGE** and **WAY ALONE** (see middle panel of Figure 6.15). The estimated probability to use public transport for *Not Alone* increases slightly from *Main User* to *Secondary User*, but remains close to the one for *Alone*. The profile curve for *Not Alone* differs significantly from the profile curve for *Alone* only for *Not Users*. This is not surprising since a *Not User* traveling alone has zero probability using a car.

If we consider the different **USAGE** and **DAY TIME** combinations (the bottom panel of Figure 6.15) we see that during *Day* time there is no difference between *Main Users* and *Secondary Users*, while during *Night* time *Secondary Users* nearly always use individual transport. Note that we consider *Male* users here, since there are more *Male* than *Female* users in this data set. This is consistent with bottom panel of Figure 6.14, when averaged over usage.

Figure 6.16 shows the estimated probabilities for different **USAGE** and **DISTANCE** combinations. It is remarkable that for *Main Users* and *Secondary Users* both profile curves for *Short Distance* ($\leq 3.5km$) and *Far Distance* ($> 21.5km$) lie significantly higher

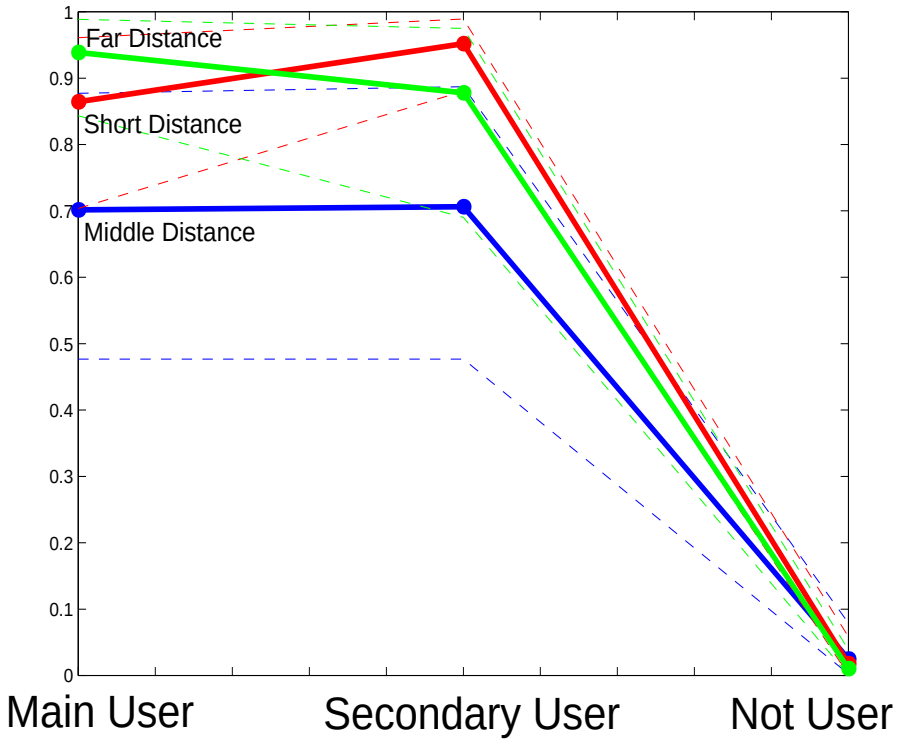

Figure 6.16: Estimated posterior mean probabilities for using an individual transport in Postal code area 81377 and 5th cluster group for different **USAGE** and **DISTANCE** combinations, while other covariates are set as in Table 6.11 (dotted lines correspond to 90% credible bounds)

than the profile curve for *Middle Distance* and are close to 1. For the high individual transport probability for *Far Distance* trips, the sparse suburban railway net, compared to the road net, might be responsible, as well as the high ticket costs and shorter travel duration. In contrast, the most *Short Distance* trips are within the city. Therefore the interpretation of the above-average high probability to use a car in this case is not so simple.

We continue now this section with the interpretation of spatial effects. There are 24 postal codes whose 90% credible intervals do not include zero and therefore are significant (see bottom right panel of Figure 6.17). 17 of them have a significant negative spatial effect and are indicated by a blue color. This shows that the probability of taking individual transport regardless of the covariate and group effects is reduced. In contrast 7 postal code areas have significant positive spatial effects, showing that the probability of individual transport increases. These postal code areas are indicated by a red color. We expect that the interpretation of the spatial effects is connected with the subway (U-Bahn) net and

suburban railway (S-Bahn) net. Table 6.12 confirms our assumption in general. The left column shows the numbers of postal code areas, which have U- or S-stops inside. The

	with U- or S-stops	without U- or S-stops
90% CI over 0	2	5
90% CI below 0	11	6

Table 6.12: Interpretation of Spatial Effects in Context of Presence/Absence of the U-or S- Stops inside of postal codes

right column contains the numbers of postal code areas without stops. There are only 24 postal codes with the significant spatial effects under consideration. The estimated odds ratio is $\frac{2 \cdot 6}{11 \cdot 5} \approx 0.22$, which is clearly below 1 (the 90% confidence interval is $[0.044, 1.091]$) which confirms that presence of U- and S-stops are connected with negative spatial effects and therefore reduces probability to use a car and vice versa. While there is a general dependency between significant spatial effect and the presence of the U+S-net in this postal areas, some areas do not follow this pattern (see Table 6.13). For their exact

with U- or S-stops but 90% CI over 0	without U- or S-stops but 90% CI below 0
80333	80999
81476	80634
	80797
	81243
	80689
	81373

Table 6.13: Untypical Postal Code Areas

locations see Table 6.2 and Figure 6.2. These areas should therefore be of special interest to the city planners, which seek to improve the public transport net, since these areas indicate areas of low/high public transport usage even after adjustment of trip, person and household specific effects.

We also note that the estimate of the spatial dependency parameter $\hat{\psi} \approx -0.5$ is negative. This can be explained by the specific form of S- and U-Bahn net, whose lines run from the center to suburbs like a star. Since the sign of the spatial effects correlates with the presence/absence of the U-or S- stops, it is not surprising, that especially far from the center the neighboring postal codes have often spatial effects with opposite signs.

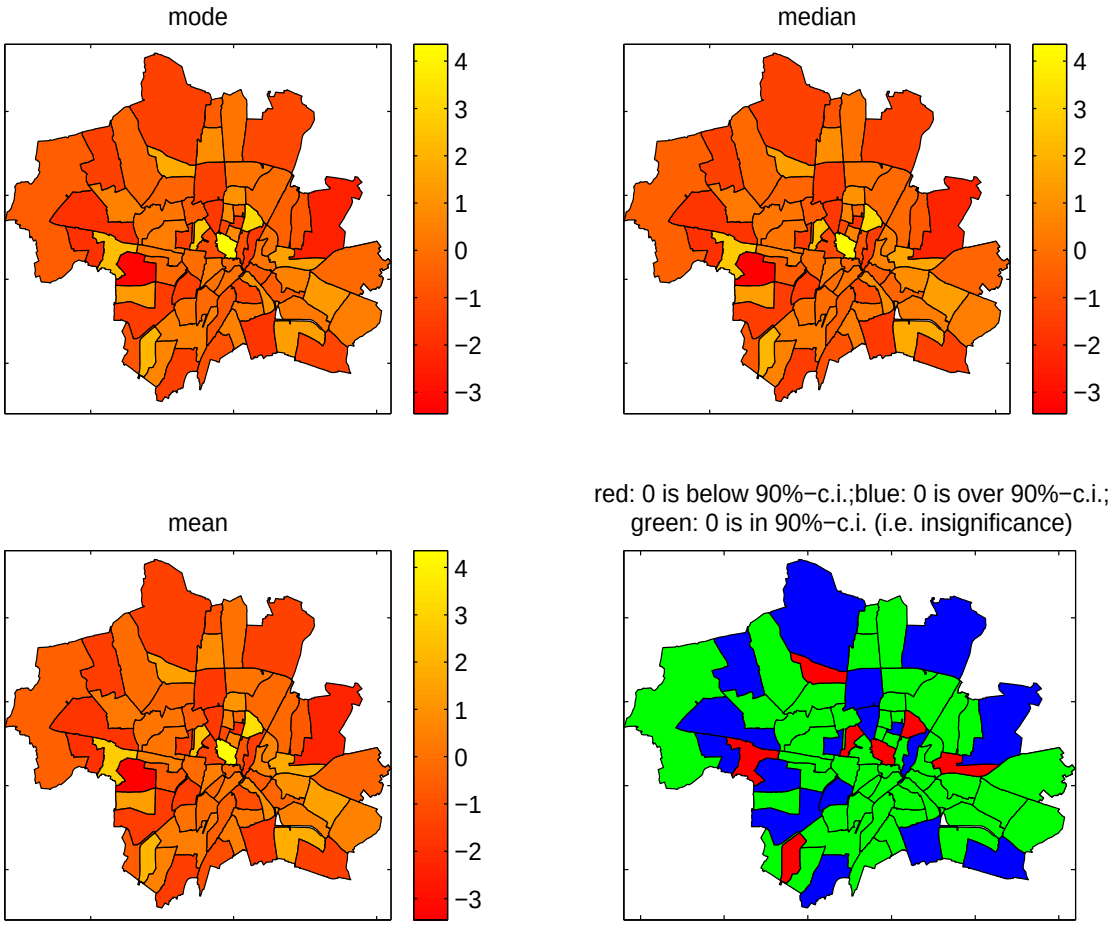

Figure 6.17: Estimated Spatial Effects $\hat{b}_j,\ j = 1, \ldots, 74$ in Model 3.

Finally we mention that cluster effects for households with large numbers of trips are positive and cluster effects for households with few numbers of trips are negative (see Table 6.4 and Figure 6.9). This corresponds that households with high mobility needs to use a car more often than households with low mobility needs.

Chapter 7

Discussion: Summary and Outlook

7.1 Summary

In this thesis we developed hierarchical binary spatial regression models with group (3.2) and individual (4.9 and 4.7) cluster effects. Both models allow for joint adjustment of covariate, spatial and cluster effects. This provides a method for identifying areas of low/high utilization of public transport after adjusting for cluster and explanatory factors such as trip, individual and household attributes.

The models we developed are based on the Bayesian approach. For parameter estimation MCMC algorithms were used. We modeled spatial effects using Gaussian CAR models which we introduced and discussed in Section 2. Such spatial priors allow us to take into consideration possible spatial dependence such as spatial smoothing. In contrast to the widely used geostatistical Gaussian kriging approach (see Diggle, Tawn, and Moyeed 1998), Gaussian CAR models allow for fast individual updating of spatial effects in a Gibbs sampler scheme. Moreover with CAR models we avoid also the difficult updating of the covariance hyperparameters in general kriging especially if the number of spatial locations is large.

For our models we modified the Pettitt CAR process, which was introduced in Pettitt, Weir, and Hart (2002). Our modification (2.9) still has all nice properties of the Pettitt CAR models: proper joint distributions, a similar interpretation of parameters, the same conditional correlations and, the most important, it allows to apply a similar method for fast updating of the spatial hyperparameters (see Section 2). But in contrast to the Pettitt CAR model our modification includes in the limit a specific intrinsic CAR model (2.4), which is often used in the modeling smoothing effect. We investigated the asymptotic behavior of the joint distribution of the modified Pettitt CAR model (see 2.10) and compared it with the joint distribution of intrinsic CAR model, which we interpreted in Section 2 (see 2.7).

The usage of CAR distributions for the modeling of spatial effects in our application requires to determine the neighboorhood structure of the 74 postal code areas of Munich. For this goal we first created a data set, which contains the polygon coordinates for each postal code area. Next we wrote a fast program which creates the neighborhood matrix using this data set. Two postal code areas are evaluated as neighbors if its polygons have at least one joint line. Figure 7.1 shows the program output for 74 postal code areas of Munich (for their exact locations see Table 6.2 and Figure 6.2). The right panel presents

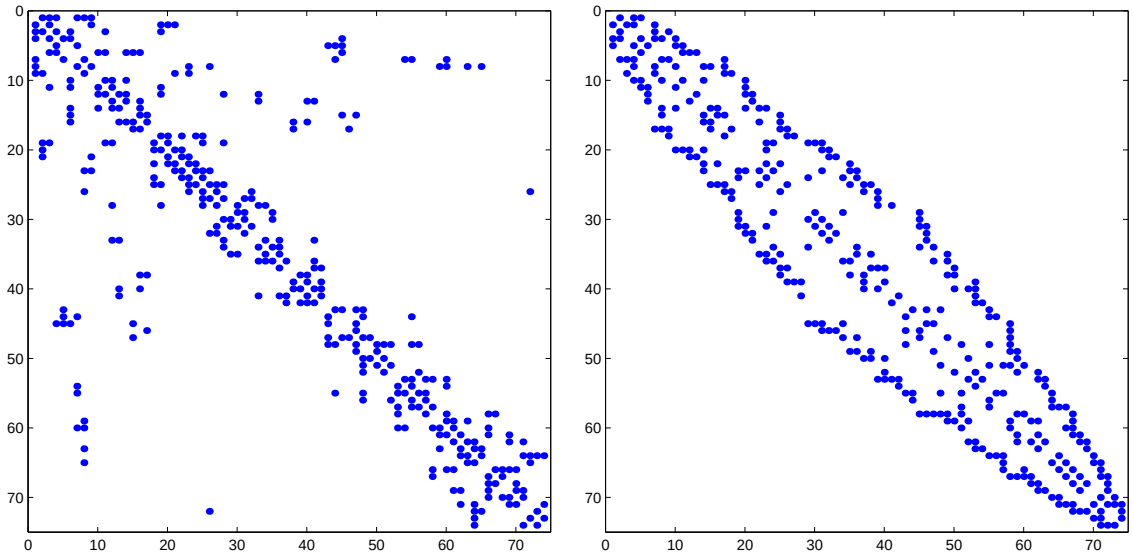

Figure 7.1: Left: Neighborhood Matrix of 74 Postal Code Areas of Munich; Right: Neighborhood Matrix after Reordering of Postal Code Areas which Minimizes Its Bandwidth

the neighborhood structure of postal code areas after their optimal reordering which minimizes the bandwidth of the neighborhood matrix. This is important for Bayesian inference for Probit Model (4.7) with individual cluster effects (see Section 4.2.3). This reordering can be accomplished in MATLAB by the function *symrcm*.

In addition to spatial effects we extended the binary regression model by cluster random effects. The first approach (Section 3) models heterogeneity between clusters. This is realized by the Logit Model (3.2) defined as

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i)}}_{\textit{random cluster effect}},$$
$$c_m \sim N(0,\sigma_c^2)\ i.i.d.,\ m = 1.\ldots,M.$$

It assumes as usual that all members of a cluster have the same random effect. In Section 4 we developed also an approach, which models heterogeneity within a cluster. This is

realized in Model (4.1) given by:

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i),k(i)}}_{\textit{random cluster effect}},$$
$$c_{mk} \sim N(0, \sigma_m^2)\ i.i.d.,\ m = 1.\ldots, M.$$

The cluster effects within a cluster are modeled here as independent normally distributed random variables with zero mean and a cluster specific variance. We call this approach individual cluster effect modeling. The non identifiability problem, which arises here, was solved for Probit Model (4.2) using reparametrization (4.6), which results in the identifiable equivalent Model (4.7). We applied the same idea for the Logit Model (4.1), which led us to Model (4.9). This model has one parameter less, compared to Model (4.1):

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \begin{cases} \mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)} & \text{if } m(i) = 1 \\ \frac{\mathbf{x_i^t}\boldsymbol{\alpha}' + b'_{j(i)}}{\sigma'_{m(i)}} & \text{if } m(i) = 2, \ldots, M\ . \end{cases}$$

The simulation study presented in Section 5 indicated its identifiability. We mention here again that large values of the cluster parameter $\sigma_m^{2'}$ indicate a large heterogeneity within cluster m. This interpretation of the cluster parameters follows from reparametrization (4.6). If the numbers of observations in clusters differ strongly, we recommend to denote the cluster with the largest number of observations as the first cluster when the individual cluster approach is to be used.

For the hierarchical spatial binary regression Logit (4.9) and Probit (4.7) Models with individual cluster effects we developed two MCMC algorithms for parameter estimation (Section 4.2.). The first one is direct and the second one is based on latent variables. The first algorithm is useful if the likelihood of the data, given covariates and unknown parameters, can be easily computed as in binary logistic models. Markov Chains are then generated using Metropolis-Hastings steps. But in contrast to the logit case the Metropolis-Hastings update step is computationally much more expensive for probit models. It is slow and less precise because of the numerical complexity of the computations for the values of $\Phi(\cdot)$. This arises from the computation of $\Phi(\cdot)$ for large values, since this distribution function converges to 1 much faster, than the more heavy-tailed logit distribution. We proposed a method how to overcome this problem in Section 4.2.2. The implementation of this method for the direct MCMC algorithm for Probit Model (4.7) is still under preparation and will be investigated in the future. The second approach, which is particularly useful for probit models, is based on latent variables, where the observed binary responses are generated by a threshold mechanism. For latent Gaussian variables this leads to binary probit models as discussed for example in Albert and Chib (1993). In particular, for our Probit Model (4.7) this leads to the latent variables representation

(4.8). For MCMC inference, Gaussian latent variables are considered as unknown additional "parameters" and are generated with the other parameters in a Gibbs sampling scheme. We note again that block updating for regression parameter $\boldsymbol{\alpha}$ and spatial parameter $\mathbf{b}$ is available in this estimating algorithm. Therefore we achieved considerably better mixing than in the direct algorithm, where parameters are updated individually. This allowed us to reduce the number of iterations in the corresponding Markov chains. Furthermore, this method reduces the number of parameters, which require a numerically more complicated Metropolis-Hastings step, to only one. The corresponding MCMC algorithm is developed in Section 4.2.3. We finally note, that Holmes and Knorr-Held (2003) give new ideas on how to use latent variables in a logit model. These ideas might yield a new MCMC algorithm for Models (4.9) and (3.2).

The final remark with regard to the applicability of the MCMC algorithms in this thesis is as follows. Since the original Model (4.1) is not identifiable we expect that a MCMC algorithm based on (4.1), i.e. Markov chains $\{\boldsymbol{\alpha}^r\}, \{\mathbf{b}^r\}, \{\boldsymbol{\sigma}^{\mathbf{2}r}\}, \{\tau^{2r}\},\ r = 1, 2, \ldots$ may not converge. However, with the parameter transformation given in (4.6) we achieve the identifiable Model (4.9) and therefore we expect the transformed chains

$$\left\{\frac{\boldsymbol{\alpha}^r}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\mathbf{b}^r}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\boldsymbol{\sigma}^{\mathbf{2}r}}{\sqrt{1+\sigma_1^{2r}}}\right\}, \left\{\frac{\tau^{2r}}{1+\sigma_1^{2r}}\right\}, \{\phi^r\}, \quad r = 1, 2, \ldots$$

to converge. The simulation study for Model (4.1) supported this conjecture. These chains can be also used for the parameter estimation of $\boldsymbol{\alpha}', \mathbf{b}', \boldsymbol{\sigma}^{\mathbf{2}\prime}, \tau^{2'}, \phi'$ from the identifiable Model (4.9) as an alternative to constructing a MCMC algorithm directly based on Model (4.9). Generally, such alternatives compared to the common approach to simulate Markov Chains only from the identifiably reparametrized model might be preferable if parameter updates in the unidentifiable model are simpler than in its identifiable analog.

All 3 Models (3.2), (4.9) and (4.7) were successfully applied to data from the Munich mobility study. The starting set of covariates for the fixed effects in our models we selected using standard model selection techniques for GLM's such as partial deviance tests. The results are presented in Section 6. A goodness of fit assessment for each model with regard to their spatial effects was carried out in Section 6.3.

7.2 Outlook

We now want to consider the following extensions to our models:

- Modeling of spatial/cluster interactions;
- Modeling of simultaneous heterogeneity within and between clusters.

Some first ideas on how to approach these questions will be discussed in the following 2 subsections.

7.2.1 Modeling of Spatial/Cluster Interactions

The first important question we consider is how to allow for a possible interaction between spatial and cluster effects. We propose two possibilities to do this. The first one uses a general linear kriging approach (Diggle, Tawn, and Moyeed 1998). As already noted in the introduction, this approach is appropriate for data collected at specified point locations. But we can apply this idea also for data regions (whose number J is constant and does not depend on the number of observations n). Here we model the vector of spatial effects $\mathbf{b} = (b_1, \ldots, b_J)^t$ as a realization of some Gaussian stationary process $\mathbf{B}$ with zero mean instead as a realization from the CAR process.

To take into consideration an interaction between spatial and cluster effects we propose to define a Gaussian stationary process $\mathbf{G}$ with zero mean at the locations g_{jm}, $j = 1, \ldots, J$, $m = 1, \ldots, M$. Here the location g_{jm} is a vector with spatial and cluster components: $g_{jm} = (s_j, r_m)^t$. For example the spatial component can be defined by the coordinates of the region center. Cluster components can be metric (for example, scored age levels if clusters represent age groups), or categorical. Then the predictor can be modeled as follows (compare with (3.2)):

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{G(g_{j(i),m(i)})}_{\textit{spatial and cluster effect}} . \tag{7.1}$$

A similar idea about defining locations is used in Kuhnert, Mengersen, and Smith (2002), who collect in clusters similar design vectors of observations. But the authors seem to assume that all covariates are metric. For a discussion on how to choose an appropriate distance metric they refer to Kaufman and Rousseeuw (1990). We can define the distance between two locations $h(g_{jm}, g_{j'm'})$ for example as follows:

$$h^2(g_{jm}, g_{j'm'}) = h_s^2(s_j, s_{j'}) + q \cdot h_r^2(r_m, r_{m'}) , \tag{7.2}$$

where h_s, h_r stands for distances in the spatial and in a finite cluster space, respectively, and q is a known weight coefficient. To define a stationary Gaussian process $\mathbf{G}$ we need to

choose a covariance function $\sigma_g^2 \cdot \rho(h;\boldsymbol{\delta})$, where $\rho(\cdot;\cdot)$ denotes the correlation function of the process $\mathbf{G}$, and $\boldsymbol{\delta}$ stands for the correlation parameter vector. In the introduction we mentioned the estimation problem of correlation parameter $\boldsymbol{\delta}$ in kriging models without interaction when the number of regions is large. Including interaction as in (7.1) increases the number of region/cluster combinations, thus the method will not be feasible for even a medium number of regions and a large number of clusters. A second problem with this approach is how to interpret the absence of interaction between spatial and cluster effects. One possible solution is to require a splitting of process $\mathbf{G}(\mathbf{g_{jm}})$ in a sum of two independent zero mean stationary Gaussian processes $\mathbf{S}(\mathbf{s_j})$ and $\mathbf{R}(\mathbf{r_m})$ for some value of parameter $\boldsymbol{\delta}^*$ when no interaction is present. For this assume, that processes $\mathbf{S}$ and $\mathbf{R}$ have covariation functions, denoted by $\sigma_s^2 \cdot \rho_s(h_s)$ and $\sigma_r^2 \cdot \rho_r(h_r)$, respectively. For $\boldsymbol{\delta}^*$ must hold:

$$\begin{aligned}\sigma_g^2 \cdot \rho(h(g_{jm}, g_{j'm'});\boldsymbol{\delta}^*) &= \operatorname{cov}(G(g_{jm}), G(g_{g'm'})) = \operatorname{cov}(S(s_j) + R(r_m), S(s_{j'}) + R(r_{m'})) \\ &= \sigma_s^2 \cdot \rho_s(h_s(s_j, s_{j'})) + \sigma_r^2 \cdot \rho_r(h_r(r_m, r_{m'})).\end{aligned}$$

If $s_j = s_{j'}, r_m = r_{m'}$ we obtain $\sigma_g^2 = \sigma_s^2 + \sigma_r^2$. So a splitting of process $\mathbf{G}(\mathbf{g_{jm}})$ in a sum of two independent stationary Gaussian processes with zero mean is possible only if the choice of correlation function $\rho(h;\boldsymbol{\delta})$ allows for some $\boldsymbol{\delta}^*$ the following representation:

$$\rho(h(h_s, h_r);\boldsymbol{\delta}^*) = \beta\rho_s(h_s) + (1-\beta)\rho_r(h_r)\,, \quad \text{for some } \beta \in (0,1)\,. \tag{7.3}$$

At the moment it is unclear if there exist such correlation functions which satisfy (7.3).

As an alternative approach we suggest to use multivariate CAR models mentioned for example by Pettitt, Weir, and Hart (2002). The multivariate CAR model is a model for $\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_J})^t$, where the components $\mathbf{b_j} = (b_{j1}, \ldots, b_{jM})^t$, $j = 1, \ldots, J$ are $M-$dimensional vectors instead of scalars, as before. The joint distribution of the vector $\mathbf{b}$ is defined as follows:

$$\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_J})^t \sim N_{J\times M}\left(\mathbf{0}, \tau^2(Q^{-1} \otimes V)\right), \quad V = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \in \mathbb{R}^{M\times M}, \tag{7.4}$$

where $A \otimes B$ stands for *Kronecker product* of matrix $A = \begin{pmatrix} a_{11} & \ldots & a_{1J} \\ \vdots & \ddots & \vdots \\ a_{J1} & \ldots & a_{JJ} \end{pmatrix}$ and matrix B given by

$$A \otimes B = \begin{pmatrix} a_{11}\cdot B & \ldots & a_{1J}\cdot B \\ \vdots & \ddots & \vdots \\ a_{J1}\cdot B & \ldots & a_{JJ}\cdot B \end{pmatrix}.$$

In particular for the multivariate modified Pettitt CAR, the conditional distribution is given then as follows (compare with (2.9)):

$$\mathbf{b_j}|\mathbf{b_{j'}}, j \neq j' \sim N_M \left(\frac{\phi}{1+|\phi|N_j} \sum_{j \sim j'} \mathbf{b_{j'}} \, , \frac{(1+|\phi|)\tau^2}{1+|\phi|N_j} \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \right) .$$

The parameter ρ measures the strength of the cluster dependence. If $\rho = 0$ then all M components of vector $\mathbf{b_j}$ are iid. As before, the parameter ϕ measures the strength of the spatial dependence. If $\phi = 0$ then the vectors $\mathbf{b_j}, j = 1, \ldots, J$ are independent and normally distributed with mean zero and covariance matrix $\tau^2 V$. Further interesting properties of the multivariate CAR model can be found in Pettitt, Weir, and Hart (2002). The authors use multivariate Gaussian CAR models for a data augmentation approach. In their application the binary data concerns the presence or absence of two tree varieties represented at 469 sites. Since the presence of these two kinds of trees can depend on each other, the authors model this data using a multivariate CAR model with $J = 469$ and $M = 2$ by a threshold mechanism. Further, Carlin and Banerjee (2002) use multivariate Gaussian CAR models in generalized linear mixed models, namely for spatial survival data analysis, where the Gaussian variable enters in the linear predictor. Here M stands also for the number of types of observations, namely types of cancer, whose presence can depend on each other. We now propose to apply the multivariate Gaussian CARs in a new way, namely for modeling spatial-cluster interactions. More precisely, we propose to model spatial and cluster effects jointly as some multivariate CAR. As usually, J denotes the number of regions, while M stands for the number of clusters. Then logits are modeled as follows (compare with (3.2)):

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i),m(i)}}_{\textit{spatial and cluster effect}} \quad ,$$

where $\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_J})^t$, $\mathbf{b_j} = (b_{j1}, \ldots, b_{jM})^t$, $j = 1, \ldots, J$ is modeled as a realization of the multivariate CAR (7.4). In this model we have to estimate one additional parameter ρ, which measures strength of a space-cluster interaction. The absence of interaction is indicated by $\rho = 0$. In this case the M vectors $(b_{1m}, \ldots, b_{Jm})^t$, $m = 1, \ldots, M$ are independent identically distributed Gaussian CAR models.

Note that this principle can be used to interpret the absence of the spatial-cluster interaction also for the general kriging approach (7.1). This means that no interaction is present if there is no correlation between $G(g_{jm})$ and $G(g_{j'm'})$ for $m \neq m'$ while $(G(g_{1m}), \ldots, G(g_{Jm})^t, m = 1, \ldots, M$ are independent and identically Gaussian distributed

with distance metric

$$h^2(g_{jm}, g_{j'm'}) = h_s^2(s_j, s_{j'}),$$

which is independent of m. This can be achieved from (7.2) when q is large and $\rho(h, \boldsymbol{\delta}) = 0$ if $h > R$ (for some fixed R). This means we have to allow to estimate q together with correlation parameters $\boldsymbol{\delta}$. It is an open problem how to do this.

An important advantage of the multivariate CAR approach compared to general kriging approach is fast and easy updating for all variance-covariance parameters, namely τ^2, ϕ and ρ even in large dimensions.

Finally we show how to interpret the modeled interaction present in the multivariate CAR model (7.4) as a product of spatial and cluster effects. By this we mean that the distribution of the multivariate Gaussian CAR vector $\mathbf{b}$ with the variance-covariance matrix $\tau^2(Q^{-1} \otimes V)$ has the same mean and covariance matrix as the random vector $B := (B_{11}, \ldots, B_{JM})^t$ with components

$$B_{jm} = B_j \cdot A_m,$$

where $B_j, j = 1, \ldots, J$ and $A_m, m = 1, \ldots, M$ are independent random vectors. Here

$$\begin{array}{l} (B_1, \ldots, B_J)^t \sim N_J(\mathbf{0}, \tau^2 Q^{-1}) \quad \text{is a Gaussian CAR and} \\ \mathbf{A} := (A_1, \ldots, A_M)^t \quad \text{has zero mean and covariance } V. \end{array}$$

If in addition for the vector $\mathbf{A}$ the following distribution is chosen:

$$\begin{array}{l} P(A_m = 1) = P(A_m = -1) = \frac{1}{2}, \\ P(A_{m'} = 1 | A_m = 1) = P(A_{m'} = -1 | A_m = -1) = \frac{1+\rho}{2}, \\ P(A_{m'} = 1 | A_m = -1) = P(A_{m'} = -1 | A_m = 1) = \frac{1-\rho}{2}, \end{array}$$

then the components $B_{jm},\ j = 1, \ldots, J,\ m = 1, \ldots, M$ have the same distribution as the corresponding spatial-cluster effects b_{jm}, i.e. they are also normal with the same mean and variance. However their joint distributions are different.

7.2.2 Modeling of Simultaneous Heterogeneity within and between Clusters

In this thesis we first considered the model with group cluster effects, Model (3.2), which implements heterogeneity between clusters:

$$\begin{array}{l} Y_i | p_i \sim \text{Bernoulli}(p_i) \text{ conditionally independent with} \\ \theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i)}}_{\textit{random cluster effect}}, \quad c_m \sim N(0, \sigma_c^2)\,. \end{array}$$

Next we introduced the model with individual cluster effects, Model (4.1) (and further we developed its identifiable representation, Model (4.9)), which implements heterogeneity within clusters:

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i),k(i)}}_{\textit{random cluster effect}}, \quad c_{mk} \sim N(0, \sigma_m^2)\,.$$

A natural proposal would be to unite these two approaches, in order to be able to model both heterogeneity between and within clusters:

$$\theta_i := \log\left(\frac{p_i}{1-p_i}\right) = \underbrace{\mathbf{x_i^t}\boldsymbol{\alpha}}_{\textit{fixed effect}} + \underbrace{b_{j(i)}}_{\textit{random spatial effect}} + \underbrace{c_{m(i),k(i)}}_{\textit{random cluster effect}}, \quad c_{mk} \sim N(c_m, \sigma_m^2)\,.$$

We only note here, that this proposed model is also unidentifiable and we need to find a similar identifiable representation, as for Model (4.1).

APPENDIX

A Proofs of Some Results in Chapter 2

Proof of (2.7) To show (2.7), first note that from (2.6) it follows immediately that $Q_{12} = -Q_{11}1_{J-1}$ and $Q_{22} = -Q_{21}^t \mathbf{1}_{J-1}$, which implies

$$Q_{11}^{-1}Q_{12} = -\mathbf{1}_{J-1} \quad \text{and} \tag{A.1}$$

$$Q_{22} = Q_{21}(Q_{11}^{-1}Q_{12}) \; . \tag{A.2}$$

Further, since the neighborhood structure has no isolated regions or groups of regions, the matrix Q_{11} is clearly positive definite as a symmetric diagonally dominant matrix (see Theorem 12.2.16 in Graybill (1983)). Therefore the rank of the positive semi-definite matrix Q_0 is equal to $J-1$. We can write now the density of $\mathbf{b}$ as

$$\begin{aligned}
[\mathbf{b}] \; &\propto \exp(-\tfrac{1}{2\tau^2}\mathbf{b}^t Q_0 \mathbf{b}) \\
&= \exp\{-\tfrac{1}{2\tau^2}[\mathbf{b_{-J}}^t Q_{11}\, \mathbf{b_{-J}} + b_J Q_{21}\mathbf{b_{-J}} + \mathbf{b_{-J}}^t Q_{12} b_J + b_J Q_{22} b_J]\} \\
&\overset{(A.2)}{=} \exp\{-\tfrac{1}{2\tau^2}[\mathbf{b_{-J}} + (Q_{11})^{-1}Q_{12}\, b_J]^t\, Q_{11}\, [\mathbf{b_{-J}} + (Q_{11})^{-1}Q_{12}\, b_J]\} \\
&\overset{(A.1)}{=} \exp\{-\tfrac{1}{2\tau^2}[\mathbf{b_{-J}} - \mathbf{1}_{J-1} b_J]^t\, Q_{11}\, [\mathbf{b_{-J}} - \mathbf{1}_{J-1} b_J]\} \cdot 1 \\
&\propto [\mathbf{b_{-J}}|b_J] \cdot [b_J] \; .
\end{aligned}$$

This implies (2.7).

Proof of (2.10)

To show (2.10), we first note that simple direct computations from (2.2) and comparison with (2.6) lead us for $\phi \geq 0$ to the following equality for the precision matrix of the modified Pettitt's CAR, denoted by $\frac{1}{\tau^2}\, Q^{m.P}$:

$$Q^{m.P} = \psi Q_0 + (1-\psi) I_J, \text{ where } \psi = \frac{\phi}{1+\phi} \; .$$

If we write the variance matrix of the modified Pettitt's Gaussian CAR as

$$\tau^2 \Sigma^{m.P} = \tau^2 (Q^{m.P})^{-1} = \tau^2 \begin{pmatrix} \Sigma_{11}^{m.P} & \Sigma_{12}^{m.P} \\ \Sigma_{21}^{m.P} & \Sigma_{22}^{m.P} \end{pmatrix} \quad \text{and } Q^{m.P} = \begin{pmatrix} Q_{11}^{m.P} & Q_{12}^{m.P} \\ Q_{21}^{m.P} & Q_{22}^{m.P} \end{pmatrix} ,$$

where $\tau^2\Sigma_{11}^{m.P} \in \mathbb{R}^{(J-1)\times(J-1)}$ denotes the covariance matrix of the vector $\mathbf{b_{-J}}$, then the variance of b_J, $\sigma_J^2(\psi) = \tau^2\Sigma_{22}^{m.P}$ can be written (see Rao 1973, p.33) as follows:

$$\begin{aligned}\sigma_J^2(\psi) &= \tau^2(Q_{22}^{m.P} - Q_{21}^{m.P}\,(Q_{11}^{m.P})^{-1}\,Q_{12}^{m.P})^{-1} \\ &= \tau^2\,(\psi Q_{22} + (1-\psi) - \psi Q_{21}(\psi Q_{11} + (1-\psi)I_{J-1})^{-1}\psi Q_{12})^{-1}\ .\end{aligned}$$

From (A.2) it follows that $\Sigma_{22}^{m.P} \to +\infty, \quad \text{if } \psi \to 1$. To proof the first equation of the (2.10) we must show that

$$\phi^{-1}\Sigma_{22}^{m.P} \to \frac{1}{J}, \quad \text{if } \phi \to +\infty\ .$$

So,

$$\begin{aligned}&\phi^{-1}\Sigma_{22}^{m.P} = \phi^{-1}\,(\psi Q_{22} + (1-\psi) - \psi Q_{21}(\psi Q_{11} + (1-\psi)I_{J-1})^{-1}\psi Q_{12})^{-1} \overset{\psi=\frac{\phi}{1+\phi}}{=} \\ &= \frac{\frac{1}{\phi}}{Q_{22}+\frac{1}{\phi}-Q_{21}(Q_{11}+\frac{1}{\phi}I_{J-1})^{-1}Q_{12}}\frac{1+\phi}{\phi} \overset{x:=\frac{1}{\phi}}{=} \frac{x(1+x)}{Q_{22}+x-Q_{21}(Q_{11}+xI_{J-1})^{-1}Q_{12}}\ . \\ &\lim_{\phi\to\infty}\frac{\Sigma_{22}^{m.P}}{\phi} = \lim_{x\to\infty}\frac{x(1+x)}{Q_{22}+x-Q_{21}(Q_{11}+xI_{J-1})^{-1}Q_{12}} \overset{\text{Th. of L'Hospital}}{=} \\ &= \lim_{x\to\infty}\frac{1+2x}{1+Q_{21}(Q_{11}+xI_{J-1})^{-2}Q_{12}} = \frac{1}{1+(Q_{21}Q_{11}^{-1})(Q_{11}^{-1}Q_{12})} \overset{A.1}{=} \frac{1}{1+(J-1)} = \frac{1}{J}\ .\end{aligned}$$

The conditional mean of $\mathbf{b_{-J}}$ given b_J can be written (see Anderson 1958) as:

$$\mu_J(\psi) = \Sigma_{12}^{m.P}(\Sigma_{22}^{m.P})^{-1}b_J,$$

what again can be written in terms of $Q^{m.P}$ (Rao 1973, p.33) as

$$\mu_J(\psi) = -(Q_{11}^{m.P})^{-1}Q_{12}^{m.P}b_J$$

Thus the second equation of the (2.10) follows immediately from (A.1) and the fact, that $Q^{m.P} \to Q_0$, if $\phi \to +\infty$, in particular $\lim_{\phi\to+\infty}\Sigma_{11}^{m.P}(\phi) = Q_{11}^{-1}$, since the matrices involved are positive definite.

B Generalized Linear Models (*GLM's*)

There is considerable literature available discussing GLM's. A standard reference for all types of GLM's is the book by McCullagh and Nelder (1989). Generalized linear models are, as the name says, a generalization of linear models based on the normal distribution. This generalization consists of two parts:
First, other distributions than the normal distribution for the response variable can be used, as long as they belong to the class of exponential family densities. Second, the mean is not directly modeled, but a transformation of it. This is facilitated by using special link functions. So, the observations $Y_i|\mathbf{x}_i,\ i = 1, \ldots, n$ are assumed to be conditionally independent with a density $p(y_i|\mathbf{x}_i)$ from the exponential family. Here the design vector $\mathbf{x}_i \in \mathbb{R}^p$ describes covariates which are used to explain the response Y_i. Recall that the density $p(y_i)$ of a random variable Y_i belongs to the exponential family, if it has a form

$$p(y_i|\theta_i, \phi) = p(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right), \tag{B.1}$$

where

- $\theta_i = \theta(\mu_i)$ is called the natural or canonical parameter. It is a monotonic function of the expected value μ_i of Y_i, i.e. $\mu_i := E(Y_i)$;

- ϕ is a dispersion parameter, which might be known or unknown;

- $b(\cdot)$ and $c(\cdot)$ are known functions which determine the given distribution. Given a density of the form (B.1) it can be shown that

$$\begin{aligned} b'(\theta_i) &= E(Y_i) = \mu_i \\ \phi \cdot b''(\theta_i) &= \mathrm{var}(Y_i) \end{aligned} \tag{B.2}$$

To the exponential family belong almost all standard distributions such as normal, binomial, Poisson and Gamma. To construct a density for the $Y_i|\mathbf{x}_i$ the linear predictor $\eta_i := \mathbf{x_i}'\boldsymbol{\beta}$ is linked with the expected value $\mu_i = E(Y_i|\mathbf{x}_i)$ via a monotonic response function $h(\eta_i)$:

$$\mu_i = h(\eta_i) = h(\mathbf{x}_i^t\boldsymbol{\beta}) \quad \forall i = 1, \ldots, n,$$

where $\boldsymbol{\beta} := (\beta_0, \ldots, \beta_p)^t$ is a unknown regression parameter vector. The aim in model fitting is to replace the data $\mathbf{Y} := (Y_1, \ldots, Y_n)^t$ with a set of fitted values $\hat{\mathbf{Y}}$ derived from a model. The number of included parameters $p + 1$ is very important. Three different types of models can be considered. First of all we have the "zero model". In that model no parameters are included, except the intercept term β_0. Its disadvantage is that it is

too simple and therefore often does not model the relationship between $\mathbf{Y}$ and $X := (\mathbf{x_1}, \ldots, \mathbf{x_n})^t$ realistically. On the other side, there is the *"saturated"* or *"full model"*. In this case as many parameters are included in the model as observations are available, i.e $p + 1 = n$. If the number of parameters increases, the mean vector $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_n)^t$ is represented more and more realistically, up to the saturated model, where $\boldsymbol{\mu}$ is fitted perfectly. But on the other hand, the saturated model is unsuitable for forecasting since it provides no data structure explanation. Therefore we need to take into consideration a third class of models, which have more parameters than the zero model, but less than the saturated model.

If $\eta_i = \theta_i$, i.e. $h(\eta_i) = \theta^{-1}(\eta_i)$, then $h(\cdot)$ is called the *canonical response function*. The inverse function $g = h^{-1}$ is called *link function*. The link function corresponding to $\eta_i = \theta_i$ is called the *canonical link function*. For the binomial distribution the canonical link function is the *logit* function defined by

$$\text{logit}(p_i) := \log\left(\frac{p_i}{1 - p_i}\right)$$

The classical parameter estimation procedure is based on the *Maximum Likelihood* approach. For this we consider maximization of the *loglikelihood function* $L(\boldsymbol{\beta})$ which is given as

$$\begin{aligned} &L(\boldsymbol{\beta}) := \log p(\mathbf{Y}; \boldsymbol{\beta}) = \log \textstyle\prod_{i=1}^n p(Y_i; \boldsymbol{\beta}) = \sum_{i=1}^n \log p(Y_i; \boldsymbol{\beta}) = \sum_{i=1}^n L_i(\boldsymbol{\beta}), \text{ where} \\ &L_i(\boldsymbol{\beta}) \overset{(B.1)}{\propto} \tfrac{Y_i\theta_i - b(\theta_i)}{\phi} + \log(c(Y_i, \phi)). \end{aligned} \tag{B.3}$$

Note that we can disregard the term $\log(c(Y_i, \phi))$, since it does not depend on $\boldsymbol{\beta}$. Next we define the *score function* $\mathbf{s}(\boldsymbol{\beta})$ as the vector of derivatives of the loglikelihood:

$$\begin{aligned} &\mathbf{s}(\boldsymbol{\beta}) := \tfrac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \textstyle\sum_{i=1}^n \frac{\partial L_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{s}_i(\boldsymbol{\beta}), \text{ where} \\ &\mathbf{s}_i(\boldsymbol{\beta}) := \tfrac{\partial L_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \tfrac{\partial L_i(\boldsymbol{\beta})}{\partial \theta_i} \cdot \tfrac{\partial \theta_i}{\partial \mu_i} \cdot \tfrac{\partial \mu_i}{\partial \eta_i} \cdot \tfrac{\partial \eta_i}{\partial \beta} \overset{(B.3),(B.2)}{=} \tfrac{1}{\phi}(Y_i - h(\mathbf{x_i}^t\boldsymbol{\beta})) \cdot \tfrac{1}{b''(\theta_i)} \cdot h'(\eta_i) \cdot \mathbf{x_i}. \end{aligned} \tag{B.4}$$

The solution of the system $\mathbf{s}(\boldsymbol{\beta}) = \mathbf{0}$ is called the ML solution. When this solution is the global maximum it is the *maximum likelihood estimator*, $\hat{\boldsymbol{\beta}}_{ML}$.

The (expected) Fisher information matrix $F(\boldsymbol{\beta}) := \sum_{i=1}^n F_i(\boldsymbol{\beta})$, which is needed in the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{ML}$, is defined as the covariance matrix of the score vector $\mathbf{s}(\boldsymbol{\beta})$. Under regularity conditions $F(\boldsymbol{\beta})$ can be calculated alternatively as the expected value of the observed Fisher information matrix $F_{obs}(\boldsymbol{\beta})$, where

$$F_{obs}(\boldsymbol{\beta}) := -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t}.$$

To calculate the individual Fisher information $F_i(\boldsymbol{\beta})$ of the i^{th} observation note that

$$F(\boldsymbol{\beta}) = \operatorname{cov} \mathbf{s}(\boldsymbol{\beta}) \overset{E\mathbf{s}(\boldsymbol{\beta})=\mathbf{0}}{=} \sum_{i=1}^{n} E(\mathbf{s}_i(\boldsymbol{\beta}) \cdot \mathbf{s}_i^t(\boldsymbol{\beta})) \overset{(B.4),(B.2)}{\Rightarrow} F_i(\boldsymbol{\beta}) = \frac{1}{\phi} \frac{[h'(\eta_i)]^2}{b''(\theta_i)} \cdot \mathbf{x_i} \cdot \mathbf{x_i}^t.$$

For the canonical response function (i.e. $\eta_i = \theta_i$) due to (B.2) we obtain simple expressions for the score vector and the Fisher information matrix:

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) &= \tfrac{1}{\phi} \textstyle\sum_{i=1}^{n} (Y_i - h(\mathbf{x_i}^t \boldsymbol{\beta})) \cdot \mathbf{x_i} \\ F(\boldsymbol{\beta}) &= \tfrac{1}{\phi} \textstyle\sum_{i=1}^{n} h'(\mathbf{x_i}^t \boldsymbol{\beta}) \cdot \mathbf{x_i} \cdot \mathbf{x_i}^t. \end{aligned}$$

Note that for the canonical response function the differentiation of the $\mathbf{s}(\boldsymbol{\beta})$ eliminates the random terms involving Y_i so that $F := E(F_{obs}) = F_{obs}$ in this case.

The important result for the maximum likelihood estimation in GLM's is that under regularity conditions asymptotically holds (McCullagh and Nelder 1989, appendix S.4 d.):

$$\hat{\boldsymbol{\beta}}_{ML} \sim N_{p+1}(\boldsymbol{\beta}, F^{-1}(\hat{\boldsymbol{\beta}}_{ML})). \tag{B.5}$$

In particular, the diagonal elements of the matrix $F^{-1}(\hat{\boldsymbol{\beta}}_{ML})$ can be used therefore for constructing of confidence intervals for $\beta_j, j = 0, \ldots, p$. These are used for testing the significance of β_j. If the $(1-\alpha)100\%$ confidence interval for β_j includes zero, β_j is assumed to be non-significant at level α. This test is called the *Wald test* for GLM's. Another possibility to test parameter significance is to use the so-called *likelihood quotient* or *partial deviance* test. Define the *deviance* statistic D as

$$D := -2 \left(L(\hat{\boldsymbol{\beta}}_{ML}) - L(\tilde{\boldsymbol{\beta}}_{ML}) \right),$$

where $L(\tilde{\boldsymbol{\beta}}_{ML})$ denotes the maximum log likelihood in the full model. Let d be the deviance of a reduced model with parameter vector $\boldsymbol{\beta}_r := (\beta_1, \ldots, \beta_{p_r})^t$, where $p_r < p$. Under the hypothesis, that parameters $(\beta_{p_r+1}, \ldots, \beta_p)^t = \mathbf{0}$, the value

$$d - D$$

has asymptotically χ^2 distribution with $p - p_r$ degrees of freedom (McCullagh and Nelder 1989, Section 5.5 a.). The deviance D can also be used for testing of goodness of the model. Under the hypothesis, that chosen model is true, we have asymptotically

$$D \sim \chi^2$$

with $g - p$ degrees of freedom , where g is number of the observation groups with equal design vectors. Note that we can apply this residual deviance test only if the binary data can be grouped to form binomial response data (Collett 2002, Section 3.8.2). With this result we close our short introduction in the generalized linear models.

C Bayesian Inference and Markov Chain Monte Carlo (MCMC) Methods

Good advanced introduction to Bayesian Inference and MCMC methods is given for example in Chib (2001) and in Gamerman (1997).

C.1 Bayesian Inference

In contrast to the classical or frequentist approach the Bayesian approach considers parameters are considered as random quantities, which will be updated in the presence of the observed data. In particular, let $\boldsymbol{\beta} := (\beta_0, \ldots, \beta_p)^t$ the parameter vector and let $p(\boldsymbol{\beta})$ its density or probability function, which express our uncertainty about the parameter $\boldsymbol{\beta}$ **before** sampling the data. This distribution is called the *prior distribution*. We now observe a random sample $\mathbf{Y} = (Y_1, \ldots, Y_n)^t$ with joint density or probability function $p(\mathbf{Y}; \boldsymbol{\beta})$. Further we assume that $Y_i, i = 1, \ldots, n$ are conditionally independent given the parameter $\boldsymbol{\beta}$. If we consider $p(\mathbf{Y}; \boldsymbol{\beta})$ as a function of the parameter $\boldsymbol{\beta}$ for given $\mathbf{Y}$, we speak of the likelihood. In particular we denote the *likelihood* by

$$l(\boldsymbol{\beta}) := p(\mathbf{Y}; \boldsymbol{\beta}).$$

Since the observations $\mathbf{Y}$ contain information about $\boldsymbol{\beta}$, we update our knowledge about $\boldsymbol{\beta}$ by considering the conditional distribution of $\boldsymbol{\beta}$ given $\mathbf{Y}$. This distribution is called the *posterior distribution* and expresses our uncertainty about $\boldsymbol{\beta}$ **after** taking into account the data. It can be calculated by *Bayes' theorem*, i.e.

$$p(\boldsymbol{\beta}|\mathbf{Y}) = \frac{p(\mathbf{Y}; \boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{Y})} = \frac{p(\mathbf{Y}; \boldsymbol{\beta})p(\boldsymbol{\beta})}{\int p(\mathbf{Y}; \boldsymbol{\beta})p(\boldsymbol{\beta})d\boldsymbol{\beta}}. \tag{C.1}$$

Since $p(\mathbf{Y})$ does not depend on $\boldsymbol{\beta}$, we can write

$$p(\boldsymbol{\beta}|\mathbf{Y}) \propto l(\boldsymbol{\beta}) \times p(\boldsymbol{\beta}). \tag{C.2}$$

A prior distribution with $p(\boldsymbol{\beta}) \propto 1$ is called *non-informative*. It is not a proper prior if the parameter space for $\boldsymbol{\beta}$ is not bounded. In this case we have $\int p(\boldsymbol{\beta})d\boldsymbol{\beta} = \infty$. The usage of such non-informative priors has to be done with care to insure that the resulting posterior is proper. For example for binary models we found that by using an improper prior there exist a few extremal states of data $\mathbf{Y}$, which cause an improper posterior. Once the posterior distribution is determined, the main location measure such as *posterior mode* $\hat{\boldsymbol{\beta}}_{Bayes}$ can be used for parameter estimation of $\boldsymbol{\beta}$. Note that when a non-informative prior $p(\boldsymbol{\beta}) \propto 1$ is used the posterior mode estimate formally coincides (by (C.2)) with the maximum likelihood estimate $\hat{\boldsymbol{\beta}}_{ML}$. The classical approach uses confidence intervals as

interval estimates for $\boldsymbol{\beta}$, while the Bayesian inference uses *credible intervals*. A $100(1-\alpha)\%$ credible interval for a scalar β is given by

$$C \quad \text{with} \quad \int_C p(\beta|\mathbf{Y})d\beta = 1-\alpha.$$

This can be interpreted that the posterior probability β falling into the interval C is $1-\alpha$. Such an easy interpretation cannot be made for confidence intervals.

If a future observation Y^* needs to be predicted using the data $\mathbf{Y}$, the Bayesian approach uses the *predictive distribution*, which is the conditional distribution of Y^* given $\mathbf{Y}$. If Y^* and $\mathbf{Y}$ are conditionally independent given $\boldsymbol{\beta}$, then the predictive distribution is given by

$$p(Y^*|\mathbf{Y}) = \int p(Y^*|\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathbf{Y})d\boldsymbol{\beta}.$$

Posterior calculations in (C.1) are tractable, in particular if one considers *conjugate prior distributions*. Assume that the data $\mathbf{Y}$ arises from a class of parametric distributions which we call M. A class of prior distributions P is conjugate to an observational model M if for every prior $p \in P$ and for any observational distribution $l \in M$, the posterior distribution $p(\cdot|\mathbf{Y})$ is also an element of P. Table C.1 presents some examples of conjugate Bayesian models. However in more complex situations it will often not be possible to get an

Prior	Likelihood	Posterior
Normal (with known variance)	Normal	Normal
Beta	Binomial	Beta
Dirichlet	Multinomial	Dirichlet
Gamma	Poisson	Gamma

Table C.1: Examples of Conjugate Bayesian Models

analytically closed expression for the posterior distribution, since the normalizing constant $p(\mathbf{Y})$ defined in (C.1) of the posterior distribution requires a possibly high dimensional integration. In particular, sometimes it is useful to consider an additional probabilistic structure for the parameter $\boldsymbol{\beta}$. In this case we speak about *hierarchical models*. Model (3.2) is a good example for a hierarchical model. Figure 3.1 presents its hierarchical structure. In this example the prior distribution is specified in two stages and the parameters τ^2, ϕ and σ_c^2 are called *hyperparameters*.

To overcome the analytic intractability MCMC methods are widely used. The idea behind these methods is simple and extremely general. In order to sample from a given probability distribution that is referred to as the target distribution, a suitable Markov chain is constructed with the property that its (limiting) distribution is the target distribution. Further, MCMC methods allow the parameter space to include for example

missing data or even a model class choice. Finally, sometimes the target distribution is the posterior distribution of the parameters augmented by latent data, in which case the MCMC scheme operates on a space that is also considerably larger than the parameter space. This strategy is called *data augmentation* and is applied for Model 4.7 (see Section 4.2.3).

Depending on the specific problem, Markov chains can be constructed by using the Metropolis-Hastings algorithm, the Gibbs sampling method, or hybrid mixtures of these two algorithms.

Before we will present these two algorithms, we will recall some important definitions and results from the theory of Markov chains.

C.2 Markov Chains

An advanced introduction to Markov chains with general state space can be found in Meyn and Tweedie (1993), Nummelin (1984) and Tierney (1996).

Let F be some σ-algebra of a state space S. A *Markov Chain* X over (S, F) with index space $T = \mathbb{N} \cup \{0\}$ is a process $X_n, n = 0, 1, \ldots$, where given the present state, past and future states are independent, i.e.

$$P(X_{n+1} \in A | X_n = x, X_{n-1} \in A_{n-1}, \ldots, X_0 \in A_0) = P(X_{n+1} \in A | X_n = x) \tag{C.3}$$

for all sets $A_0, \ldots, A_{n-1} \in F$. If the probabilities in (C.3) do not depend on n, we say that the Markov chain is *homogeneous*. In this case a *transition function* or *kernel* $P(h, x, A) := P(X_{n+h} \in A | X_n = x)$ must satisfy the following conditions:

(i) $\forall x \in S: \quad P(x, \cdot)$ is a probability distribution over (S, F);

(ii) $\forall A \in F$: the function $P(\cdot, A)$ is measurable;

(iii) $\forall s, t \in \mathbb{N},\ x \in S,\ A \in F: \quad P(s+t, x, A) = \int_S P(t, y, A) P(s, x, dy)$
(Kolmogorov-Chapman condition).

Here $P(x, A)$ is defined by $P(x, A) := P(1, x, A) = P(X_{n+1} \in A | X_n = x)$. The Kolmogorov-Chapman condition can be written in operator notation as

$$P^{s+t} = P^s \cdot P^t. \tag{C.4}$$

If S is a discrete space with d states $s_1, \ldots, s_d$, than the transition function is given as matrix $P \in \mathbb{R}^{d \times d}$, where $P_{ij} = P(X_{n+1} = s_i | X_n = s_j)$. In this case condition (C.4) corresponds to matrix multiplication.

A probability distribution π over (S, F) is called *P-invariant* or *P-stationary*, if

$$\forall x \in S, A \in F : \quad \int_S P(x, A)\pi(dx) = \pi(A), \tag{C.5}$$

or, in the operator notation:

$$\pi P = \pi.$$

The meaning of the invariance is clear: if we choose π as the initial distribution, i.e. $P(X_0 \in A) = \pi(A)$, then all X_n have the distribution π; the Markov Chain X is stationary.

A probability distribution π over (S, F) is called *P-reversible*, if

$$\forall x, y \in S : \quad \pi(dx)P(x, dy) = \pi(dy)P(y, dx). \tag{C.6}$$

The reversibility means that with initial distribution π the sequences (X_0, X_1) and (X_1, X_0) have the same distribution. Using induction it follows that this statement holds even for sequences $(X_0, X_1, \ldots, X_n)$ and $(X_n, X_{n-1}, \ldots, X_0)$ $\forall n \in \mathbb{N}$, i.e. the time direction does not play any role. Integration of (C.6) over $A \times S$ shows that the reversible distribution is invariant. The inverse statement does not hold in general.

A transition kernel P is called *irreducible*, if there exists some probability distribution ψ over (S, F) such that

$$\forall x \in S, \ \forall A \in F \text{ with } \psi(A) > 0 : \quad \sum_{k=1}^{\infty} P^k(x, A) > 0. \tag{C.7}$$

The irreducibility induces that the probability for the Markov chain X to reach $A \in F$ such, that $\psi(A) > 0$ is positive from any state $x \in S$. Irreducibility is required for showing that the invariant distribution is unique and that the following law of large numbers for Markov chains holds:

Proposition C.1 *Let P be a irreducible transition kernel with the invariant distribution π. Than π is the unique invariant distribution and*

$$P\left(\frac{1}{n+1}\sum_{t=0}^{n} f(X_t) \to \int f(x)\pi(dx), \ n \to \infty | X_0 = x\right) = 1$$

for π-almost all $x \in S$ and all f with $\int |f(x)|\pi(dx) < \infty$.

For the proof see for example p.49 of Guttorp (1995).

MCMC methods allows us to compute $\int f(x)\pi(dx)$ using recursive simulation of $X_{n+1}|X_n = x \quad \sim \quad P(x, \cdot)$ such that $\frac{1}{R-B}\sum_{t=B+1}^{R} f(X_t)$ approximates the integral one is interested in. For this we need 3 conditions:

(*i*) P is irreducible;

(*ii*) π is P-invariant;

(*iii*) Simulation from $P(x, \cdot)$ is simple for each $x \in S$.

Often instead of checking condition (ii), condition (ii'): π is reversible for P is easier to verify. Recall that (ii') implies (ii).

C.3 Metropolis-Hastings (MH) Algorithm

The MH algorithm allows us to construct a transition kernel P, which is irreducible and reversible with respect to the probability distribution π. It is easy to show directly the following proposition:

Proposition C.2 *Let π be absolutely continuous with respect to some measure μ and let $q(x, y) > 0$ be a transition kernel (measurable with $\int q(x, y)\mu(dy) = 1 \; \forall x \in S$). Then the following transition kernel P is π-irreducible and π-reversible:*

$$\begin{array}{l} P(x, A) = \int_A q(x, y)a(x, y)\mu(dy) + \mathbf{1}_A(x) \left(\int_S q(x, y)(1 - a(x, y))\mu(dy) \right), \\ \textit{with} \quad a(x, y) = \min \left(1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)} \right). \end{array} \tag{C.8}$$

According to condition (*iii*) in the last algorithm choose the transition kernel q such that simulation from $q(x, \cdot)$ is simple for each $x \in S$. Note that

$$P(x, x) = \int_S q(x, y)(1 - a(x, y))\mu(dy) > 0$$

implies that $P(x, S) = 1$. The MH algorithm to generate a Markov chain $\boldsymbol{\beta}^{(j)}, j = 0, 1, \ldots$ with stationary distribution $\pi(\boldsymbol{\beta}) := p(\boldsymbol{\beta}|Y)$ can now be described as follows:

(1) Set the iteration counter to $j = 1$ and specify an initial value $\boldsymbol{\beta}^{(0)}$;

(2) Generate a candidate value $\boldsymbol{\beta}$ from density $q(\boldsymbol{\beta}^{(j-1)}, \cdot)$;

(3) Accept $\boldsymbol{\beta}$ with the probability $a(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\beta})$ given in (C.8). If $\boldsymbol{\beta}$ is accepted, move the Markov chain to $\boldsymbol{\beta}^{(j)} := \boldsymbol{\beta}$, else $\boldsymbol{\beta}^{(j)} := \boldsymbol{\beta}^{(j-1)}$ and the Markov chain does not move;

(4) Change counter j to $j + 1$ and return to Step (2) until convergence is reached.

We call the transition kernel $q(\cdot, \cdot)$ the *proposal kernel* and the function $a(\cdot, \cdot)$ the *acceptance probability*. We speak about convergence in step (4) since the initial value $\boldsymbol{\beta}^{(0)}$ does not come from the P-invariant distribution π. Until convergence to stationary distribution

π is reached, the chain is in the so-called *burn in* phase. Its length $B \in \mathbb{N}$ depends on the choice of the initial value and on the mixing rate of the Markov chain.

One should observe that the target density π appears as a ratio in the acceptance probability $a(\cdot, \cdot)$ and therefore the algorithm can be implemented without knowledge of the normalizing constant of $\pi(\cdot)$. Different proposal densities give rise to specific versions of the MH algorithm, each with the correct invariant distribution π. One family of candidate-generating densities, which is very popular in applications and which we used in our MCMC algorithms, is given by *symmetric* kernels $q(x, y) = q(|x - y|)$. The candidate y is thus drawn according to the process $y = x + \xi$, where ξ follows the distribution $q(\cdot)$. One has to be careful, however, in setting the variance of ξ; if it is too large it is possible that the chain may remain stuck at a particular value for many iterations while if it is too small the chain will tend to make small moves and move inefficiently through the support of the target distribution. Both circumstances will tend to generate draws that are highly serially correlated, which is undesirable. Finally note that when q is symmetric, the acceptance probability $a(x, y)$ of a move is only determined by the ratio $\pi(y)/\pi(x)$.

C.4 Gibbs Sampler

An elementary introduction to another MCMC method, Gibbs sampling, is provided by Casella and George (1992). In this algorithm all parameters are grouped into g blocks $(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_g)$ and each block is sampled according to the *full conditional distribution* of block $\boldsymbol{\beta}_k$, $k = 1, \ldots, g$, defined as the conditional distribution under π of $\boldsymbol{\beta}_k$ given all other blocks $\boldsymbol{\beta}_{-k} := (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_g)$ and denoted as $\pi(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k})$. Derivation of the full conditional distributions is usually quite simple since, by Bayes theorem, $\pi(\boldsymbol{\beta}_k|\boldsymbol{\beta}_{-k}) \propto \pi(\boldsymbol{\beta})$. In addition, the powerful device of data augmentation, due to Tanner and Wong (1987), in which latent or auxiliary variables are artificially introduced into the sampling, is often used to simplify the derivation and sampling from the full conditional distributions. As already mentioned, we apply the data augmentation approach in MCMC algorithm for Model (4.7) (see Section 4.2.3). The Gibbs sampling algorithm can now be described as follows:

(1) Set the iteration counter to $j = 1$ and specify an initial value $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1^{(0)}, \ldots, \boldsymbol{\beta}_g^{(0)})^t$;

(2) Obtain a new value $\boldsymbol{\beta}^{(j)} = (\boldsymbol{\beta}_1^{(j)}, \ldots, \boldsymbol{\beta}_g^{(j)})^t$ through successive generation of values

$$
\begin{array}{lrcl}
(2.1) & \boldsymbol{\beta}_1^{(j)} & \sim & \pi(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2^{(j-1)}, \ldots, \boldsymbol{\beta}_g^{(j-1)}) \\
(2.2) & \boldsymbol{\beta}_2^{(j)} & \sim & \pi(\boldsymbol{\beta}_2|\boldsymbol{\beta}_1^{(j)}, \boldsymbol{\beta}_3^{(j-1)}, \ldots, \boldsymbol{\beta}_g^{(j-1)}) \\
 & \vdots & & \\
(2.g) & \boldsymbol{\beta}_g^{(j)} & \sim & \pi(\boldsymbol{\beta}_g|\boldsymbol{\beta}_1^{(j)}, \ldots, \boldsymbol{\beta}_{g-1}^{(j)});
\end{array}
$$

(3) Change counter j to $j+1$ and return to step (2) until convergence is reached.

It is clear that the Gibbs sampler defines a homogeneous Markov chain. Its transition kernel is given by

$$P(\boldsymbol{\beta}^{(j-1)}, \boldsymbol{\beta}^{(j)}) = \prod_{k=1}^{g} \pi(\boldsymbol{\beta}_k^{(j)} | \boldsymbol{\beta}_1^{(j)}, \ldots, \boldsymbol{\beta}_{k-1}^{(j)}, \boldsymbol{\beta}_{k+1}^{(j-1)}, \ldots, \boldsymbol{\beta}_g^{(j-1)}). \tag{C.9}$$

The Gibbs sampler scheme is a special case of the MH method (see for example Chib (2001)). The connection to the MH algorithm can be seen by setting in Step $(2.k)$ $(k = 1, \ldots, g)$ of the Gibbs sampling algorithm the proposal transition probability $q_k(\boldsymbol{\beta}, \boldsymbol{\beta}')$ as:

$$q_k(\boldsymbol{\beta}, \boldsymbol{\beta}') = \pi(\boldsymbol{\beta}'_k | \boldsymbol{\beta}_{-k}) \cdot \delta_{\beta_{-k}}(\boldsymbol{\beta}'_{-k}), \tag{C.10}$$

where $\delta_x(\cdot)$ denotes the Dirac function, defined by:

$$\int f(u)\delta_x(u)du = f(x).$$

The right side of (C.10) is further equal to

$$\frac{\pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}'_k, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_g)}{\int_{\beta'_k} \pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}'_k, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_g) d\boldsymbol{\beta}'_k} \cdot \delta_{\beta_{-k}}(\boldsymbol{\beta}'_{-k}).$$

Then the acceptance probability is given by

$$a(\boldsymbol{\beta}, \boldsymbol{\beta}') = \min \left(1, \frac{\pi(\boldsymbol{\beta}')}{\pi(\boldsymbol{\beta})} \frac{\pi(\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_{k-1}, \boldsymbol{\beta}_k, \boldsymbol{\beta}'_{k+1}, \ldots, \boldsymbol{\beta}'_g)}{\pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}'_k, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_g)} \cdot \right.$$
$$\left. \cdot \frac{\int_{\beta_k} \pi(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}, \boldsymbol{\beta}'_k, \boldsymbol{\beta}_{k+1}, \ldots, \boldsymbol{\beta}_g) d\boldsymbol{\beta}_k}{\int_{\beta'_k} \pi(\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_{k-1}, \boldsymbol{\beta}_k, \boldsymbol{\beta}'_{k+1}, \ldots, \boldsymbol{\beta}'_g) d\boldsymbol{\beta}'_k} \cdot \frac{\delta_{\beta'_{-k}}(\boldsymbol{\beta}_{-k})}{\delta_{\beta_{-k}}(\boldsymbol{\beta}'_{-k})} \right)_{\boldsymbol{\beta}'_{-k} = \boldsymbol{\beta}_{-k}} 1.$$

Thus, the Gibbs sampler algorithm is a special case of the MH algorithm, which shows that the corresponding Markov chain is stationary in the limit with distribution π. Note that often one cannot generate observations from all full conditionals, even when simple blocks are used. In this case we can generate the k^{th} block $\boldsymbol{\beta}_k$ using a MH step within the Gibbs sampler scheme. This was done in each of our MCMC algorithms, for example, by ψ-updating step. Such an MCMC algorithm is called a hybrid MCMC algorithm. Following from the MH-representation of the Gibbs sampler the hybrid MCMC procedure can be presented then as a complex MH algorithm and convergence to the stationary distribution π still holds.

So far nothing has been said on how to form the components of the vector $\boldsymbol{\beta}$. A first choice might be to use scalar components. However high correlation among components leads to slow convergence of the Gibbs sampler and inefficient moving through the support

of the target distribution. So it might be preferable to update highly correlated components jointly. In general, one can advise to use as large blocks as possible in the Gibbs sampler. This means to block components in such a way that it is easy to sample from the full conditional for this block. Note that if it is possible to block all components into a single block, then this means that it is easy to sample from the posterior distribution directly and a Gibbs sampler is no longer needed.

List of Figures

List of Tables

Bibliography

Albert, J. and S. Chib (1993). Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass. 88*, 669–679.

Anderson, T. (1958). *An introduction to multivariate statistical analysis*. New York: Wiley.

Bennett, J., A. Racine-Poon, and J. Wakefield (1996). MCMC for nonlinear hierarchical models. In *Markov Chain Monte Carlo in Practice*, pp. 339–57. Chapman & Hall.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc. B,36*, 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Besag, J. and P. Green (1993). Spatial statistics and Bayesian computation (with discussion). *J. R. Statist. Soc. B,55*, 25–37.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science 10*, 3–66.

Breslow, N. and D. Clayton (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass. 88*, 9–25.

Carlin, B. and S. Banerjee (2002). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. *Bayesian Statistics 7, to appear*, 7–21.

Casella, G. and E. George (1992). Explaining the Gibbs sampler. *Amer. Statistician 46*, 167–174.

Chib, S. (2001). Markov Chain Monte Carlo methods: Computation and inference. In *Handbook of Econometrics, Volume 5*, pp. 381–394. North Holland.

Collett, D. (2002). *Modelling Binary Data, second Edition*. London: Chapman & Hall.

Diggle, P., J. Tawn, and R. Moyeed (1998). Model-based geostatistics (with discussion). *Applied Statistics 47, Part 3*, 299–350.

Fahrmeir, L. and S. Lang (2000). Bayesian semiparametric regression analysis of multicategorical time-space data. *http://www.stat.uni-muenchen.de/sfb386/, Discussion Paper 202*.

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. London: Chapman & Hall.

Gelfand, A., N. Ravishanker, and M. Ecker (2000). Modeling and inference for point-referenced binary spatial data. In *Generalized Linear Models: A Bayesian Perspective*, pp. 381–394. Marcel Dekker.

George, A. and J. Liu (1981). *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. New York: Chapman & Hall.

Graybill, F. (1983). *Matrices with Applications in Statistics, 2nd edition*. California: Wadsworth.

Guttorp, P. (1995). *Stochastic Modelling of Scientific Data*. London: Chapman & Hall.

Heagerty, J. and S. Lele (1998). A composite likelihood approach to binary spatial data. *J. Am. Statist. Ass. 93*, 1099–1111.

Holmes, C. and L. Knorr-Held (2003). Efficient simulation of bayesian logistic regression models. *http://www.stat.uni-muenchen.de/sfb386/, Discussion Paper 306*.

Huffer, F. and H. Wu (1998). Markov Chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics 54*, 509–524.

Kaufman, L. and P. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kelsall, J. and P. Diggle (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Appl.Statist. 47,Part 4*, 559–573.

Knorr-Held, L. and H. Rue (2000). On block updating in Markov random field models for disease mapping. (revised, may 2001). *http://www.stat.uni-muenchen.de/sfb386/, Discuttion Paper 210*.

Kuhnert, P., K. Mengersen, and G. Smith (2002). Spatial regression with smoothing over the covariates. *Technical Report*.

McCullagh, P. and J. Nelder (1989). *Generalized Linear Models, 2nd Edition*. London: Chapman & Hall.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Volatility*. Springer Verlag.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-negative Operators*. Cambridge University Press.

Pettitt, A., I. Weir, and A. Hart (2002). Conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statistics and Computing 12*, 353–367.

Rao, C. (1973). *Linear Statistical Inference and Its Applications, 2nd Ed.* New York: Wiley.

Robert, C. (1995). Simulation of truncated normal variables. *Statistics and Computing 5*, 121–125.

Tanner, M. and W. Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association 82*, 528–549.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In *Markov Chain Monte Carlo in Practice*, pp. 59–74. Chapman & Hall.

Zängler, T. (2000). *Mikroanalyse des Mobilitätsverhaltens in Alltag und Freizeit*. Berlin: Springer Verlag.