Clustering Partition Models for Discrete Structures with Applications in Geographical Epidemiology

Dissertation

an der Fakultät für Mathematik, Informatik und Statistik
der Ludwig-Maximilians-Universität München

vorgelegt von

Günter Raßer

am 18. Juni 2003

1. Gutachter: Prof. Dr. L. Fahrmeir
2. Gutachter: Prof. Dr. L. Held
3. Gutachterin: Prof. Dr. K. Ickstadt

Rigorosum: 5. August 2003

Dankeschön

an alle, die mich während meiner Zeit am Institut für Statistik unterstützten und mir beim Erstellen dieser Dissertation mit Rat und Tat zur Seite standen.

Allen voran danke ich meinem Doktorvater Ludwig Fahrmeir, der mir die Freiheit gab, die Arbeit in meinem persönlichen Wohlfühltempo anzufertigen, und jederzeit für Nachfragen zur Verfügung stand. In gleichem Maße gilt mein Dank Leonhard Held, der mir die Bayesianische Seite der Statistik nahe brachte und verständlich machte und stets viel Vertrauen in meine Forschungstätigkeit setzte ("Du bist jetzt ein Selbstläufer"). Schließlich möchte ich mich bei Katja Ickstadt bedanken, die freundlicherweise und trotz eines engen "Terminplans" die Begutachtung meiner Dissertation übernahm.

Von meinen Mitdoktoranden möchte ich zwei ganz besonders hervorheben. Zum einen Leyre Osuna, die mir stets geduldig Nachilfe gab bei allen schwierigen (und auch nicht so schwierigen) Mathematik-Problemen, mich mit Kaffee versorgte (falls erwünscht) und auch anderweitig für Ablenkung sorgte. Zum anderen wären weite Teile dieser Arbeit sehr lückenhaft geblieben, wenn nicht Volker "Markov Random Man" Schmid im Zimmer neben mir gesessen hätte. Sein geduldiges Wiederkäuen aller Details über Markov Random Fields, die ich nie verstand und wohl nie verstehen werde, war mehr als hilfreich.

Die endgültige Fassung dieser Arbeit hat wesentlich von den englischen Sprachkenntnissen von Manuela Glaser und der Hilfsbereitschaft meines Zimmergenossen Thomas Kneib profitiert, denen ich für ihren Beitrag danke.

Zu erwähnen bleibt, daß die vorliegende Arbeit während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Statistik der Ludwig-Maximilians-Universität München und im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen" entstand. Die Arbeit wurde somit finanziell von der Deutschen Forschungsgemeinschaft unterstützt. Darüberhinaus entstand ein Teil der Arbeit in Zusammenarbeit mit meinen Koautoren Leonhard Held und Nikolaus Becker. Ihnen allen gebührt mein Dank.

München, im Juni 2003 *Günter Raßer*

Zusammenfassung

Diese Arbeit befaßt sich mit der Analyse von Daten, welche für endlich viele, räumlich strukturierte Einheiten vorliegen. Beispielsweise werden irreguläre Strukturen, wie politische Landkarten, oder auch reguläre Gitter betrachtet. Im Vordergrund stehen Anwendungen aus dem Bereich der geographischen Epidemiologie.

In der Arbeit wird ein Priori-Modell zur Verwendung innerhalb eines hierarchischen Bayes-Ansatzes entwickelt und theoretisch fundiert. Das vorgeschlagene Partitionsmodell faßt die Beobachtungseinheiten zu Clustern zusammen und ermöglicht die Schätzung von Parametern anhand lokaler Information. Besonderes Augenmerk liegt hierbei auf der räumlich adaptiven Glättung der Daten, wodurch mögliche Kanten in der geschätzten Oberfläche erhalten bleiben können. Die Information über das Vorhandensein von Kanten wird dabei aus den Beobachtungen gewonnen.

Eine Untersuchung verschiedener Datentypen belegt ein breites Anwendungsspektrum des Modells. Dabei erweist sich das Modell als sehr flexibel und es zeigen sich die erwünschten Glättungseigenschaften. Ein eingehender Vergleich mit in der Praxis häufig verwendeten Markov-Zufallsfeld-Modellen fällt positiv aus. In Abhängigkeit von der Qualität der Daten liefern beide Modelle entweder ähnliche Ergebnisse oder das in dieser Arbeit vorgeschlagene Modell bietet eine deutlichere Struktur in den Schätzungen und erleichtert somit die Interpretation der Ergebnisse.

Abstract

This thesis is concerned with the analysis of data for a finite set of spatially structured units. For example, irregular structures, like political maps, are considered as well as regular lattices. The main field of application is geographical epidemiology.

In this thesis a prior model for the use within a hierarchical Bayesian framework is developed, and a theoretical basis is given. The proposed partition model combines the units under investigation to clusters, and allows for the estimation of parameters on the basis of local information. Special emphasis is on spatially adaptive smoothing of the data that retains possible edges in the estimated surface. Information about the existence of such edges is extracted from the data.

The investigation of different data types supports the suitability of the model for a wide range of applications. The model seems to be very flexible and shows the desired smoothing behavior. In comparison to commonly used Markov random field models the proposed model has some advantages. With respect to the quality of the data, either both models yield similar results, or the proposed model provides more clear structure in the estimates and simplifies the interpretation of the results.

Contents

Chapter 1

Introduction

The statistical analysis of disease count data, usually summarized with respect to predefined geographical areas, has been a persistent topic in the statistical community over the past years. Though in general this is no new field in statistical research, there are two major reasons for this recent development.

First, with increasing interest in public health, nowadays in many countries data on incidence or mortality is collected routinely for severe diseases. Of particular interest are diseases, where the sources of the disease and causal connections to potential risk factors are still not fully known (e.g. cancer). Therefore, a large collection of data sets is waiting to be analyzed by epidemiologists. But usually such data suffer from low frequencies, and exhaustive statistical preprocessing is advisable. Second, although the observed number of cases is available, usually no covariates are measured, at least not on the same geographical resolution. Therefore, possible risk factors and non-observed covariates are substituted in the model by spatially structured, region-specific effects. This calls for statistically challenging spatial models, which have become feasible within a hierarchical Bayesian framework. With increasing computer power such models are also suitable for more general use.

Although the estimation of disease risks and the visual presentation of these estimates—also known as disease mapping—is probably the most prominent application, there are various other disciplines in geographical epidemiology. One example might be the modeling of the occurrence and prevalence of infectious diseases. However, we will solely focus on the disease mapping context. The great variety of models in this field can be divided into two major groups, continuous and discrete models. Sometimes the exact geographical locations of observed cases are known and proper statistical models are based on the analysis of individual cases. Thus, space is assumed to be continuous and spatial models are necessary to get estimates for the whole area of interest.

More often, data is collected (or accessible) aggregated within geographical or political districts. For aggregated disease count data in such discrete (usually irregular) space, estimates for the disease risk are available without spatial models. Still, estimates that ignore the spatial structure are of poor quality for rare diseases and sparsely populated areas. To improve such estimates, spatial models have been applied to disease count data. Often, these models are carried over from image analysis and mostly have the drawback that the data are spatially smoothed, but the amount of smoothing is the same over the whole space.

The major goal in this thesis is to develop a new methodology for disease mapping which allows for spatially adaptive smoothing. We provide a general framework that is suitable for many other applications as well. Various extensions of the basic model are proposed and investigated, e.g. the incorporation of covariate information and the modeling of space-time interactions.

In consideration of the proposed model, we start with some preliminary remarks on elementary concepts of disease mapping and statistical models in discrete space. Also, we give a brief summary of Bayesian inference and partition models.

1.1 Disease mapping

Most applications in this thesis are taken from disease mapping. A lot of research has been on this topic in recent years, especially many Bayesian approaches have been proposed. Sometimes the complexity of these models is enormous, although the nature of the data is simple. The method of a clustering partition model was motivated from an application in disease mapping. Therefore, we will briefly review the data and the basic problems in the estimation of disease risk.

Various types of disease incidence or mortality data are subject to statistical analysis. From a statistical point of view it is important to distinguish between infectious and non-infectious diseases. The latter are available on individual level or aggregated within certain areas. In this thesis we will consider aggregated count data on cancer incidence or mortality from Germany, where any death from cancer is reported and classified. Mortality data in this work is taken from the German Cancer Atlas (Becker & Wahrendorf 1997) and classification is according to ICD-9 (International Classification of Diseases, 9th revision) issued by the World Health Organization (WHO). For all types of cancer the data are classified by gender and reported for males and females separately.

1.1.1 Data and standardization

For each sex, we are given the observed number of cases y_{ij} of cancer mortality aggregated within geographical regions $i = 1, \ldots, n$, further stratified by age group $j = 1, \ldots, J$. In addition, the number of persons under risk n_{ij}, i.e. the population size in the same stratum, is reported. Without further information on individual covariates, we may postulate a binomial model

$$y_{ij} \sim \text{Bin}(n_{ij}, \pi_{ij}), \quad i = 1, \ldots, n, \; j = 1, \ldots, J, \tag{1.1}$$

where π_{ij} denotes the unknown probability or *risk* in region i and age group j. Note that the binomial assumption is still justified if individuals in region i and age group j have different probabilities (Knorr-Held & Besag 1998).

For small probabilities π_{ij}, i.e. for rare diseases, a Poisson approximation to the binomial distribution is useful. Then, y_{ij} has Poisson distribution with parameter $n_{ij}\pi_{ij}$. Now consider some reference probability p which may be calculated internally, e.g. $p = \sum_i \sum_j y_{ij} / \sum_i \sum_j n_{ij}$, or provided externally. A convenient representation of the model is in terms of the *expected* number of cases $e_{ij} = n_{ij}p$. Thus, we assume

$$y_{ij} \sim \text{Po}(e_{ij}\lambda_{ij}), \quad i = 1, \ldots, n, \; j = 1, \ldots, J,$$

where $\lambda_{ij} = \frac{\pi_{ij}}{p}$ is the *relative risk* in region i and age group j with respect to the reference probability p.

So far, both model assumptions, binomial and Poisson, are approximately the same. The advantage of the Poisson model becomes obvious, if we apply age group specific reference probabilities. Let us suppose the common proportionality assumption $\pi_{ij} = \lambda_i p_j$ for all regions $i = 1, \ldots, n$ (e.g. Wakefield, Best & Waller 2000). Then, $\lambda_i = \frac{\pi_{ij}}{p_j}$ is the relative risk in region i, independent of the age group j. The age group specific reference probabilities p_j, $j = 1, \ldots, J$, are usually derived by internal standardization, i.e. we fit a logistic regression model with age effects

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \alpha + \beta_j$$

for the binomial data (1.1). Often, the linear predictor is extended by area effects or other covariate information given on the same aggregation level.

For each region i and age group j the expected number of cases is now given by $e_{ij} = n_{ij}p_j$, and y_{ij} is assumed to have Poisson distribution with parameter $e_{ij}\lambda_i$. The observed number of cases in region i is simply $y_i = \sum_j y_{ij}$. Additivity of the Poisson distribution yields

$$y_i \sim \text{Po}(e_i\lambda_i), \quad i = 1, \ldots, n, \tag{1.2}$$

where $e_i = \sum_j e_{ij} = \sum_j n_{ij}p_j$ denotes the expected number of cases in region i. In the binomial model (1.1) there are $n \cdot J$ unknown parameters to be estimated. Under the proportionality assumption, the number of unknown parameters is reduced to n in the Poisson model (1.2).

This model is commonly used for the purpose of mapping disease risk because one needs to display only one parameter for each region. The target of statistical inference is the joint distribution of the relative risks $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. This representation allows easy interpretation of λ_i as the risk in region i relative to an overall risk. Alternatively, it offers an odds ratio interpretation by comparing the ratio λ_i/λ_j for any two regions i and j.

1.1.2 Estimating disease risk

For the Poisson model (1.2), the unknown parameters $\boldsymbol{\lambda}$ may be estimated by maximum likelihood (ML). The ML estimate for λ_i is known as the *standardized mortality ratio* (SMR) in region i

$$\text{SMR}_i = \frac{y_i}{e_i}, \quad i = 1, \ldots, n. \tag{1.3}$$

The variance of the SMRs is given by

$$\mathrm{Var}(\mathrm{SMR}_i) = \mathrm{Var}\left(\frac{y_i}{e_i}\right) = \frac{1}{e_i^2}\mathrm{Var}(y_i) = \frac{\lambda_i}{e_i}, \quad i = 1, \ldots, n,$$

and is inverse proportional to the expected counts and thus to the population size. Therefore, the variance is large for sparsely populated regions, i.e. for regions with the least reliable data. Furthermore, for the extreme case $y_i = 0$ for some region i, the ML estimate is useless. In practice, although the cases are given in aggregated form, the observed counts are often very low, especially for rare diseases.

To overcome the drawbacks of the SMRs, alternative models are proposed in the statistical literature. The main goals can be identified as: (1) smoothing of the estimated risks (i.e. the SMRs) by filtering out variation due to the Poisson model, and (2) stabilizing the estimates and improving their statistical properties. Without further information on covariates or the presence of risk factors, most commonly spatial statistical models are applied.

There are two major motivations for this. First, many severe diseases (e.g. cancer) develop in consequence of the exposure to one or more risk factors. Such risk factors might be environmental effects, but mainly are habits of people, e.g. alcohol consumption and dietary habits. Many of those potential risk factors display a spatial structure. Hence, it is reasonable to assume some sort of correlation between adjacent regions due to non-observed (or unknown) risk factors. Second, for sparse data the estimate in one region can be improved by incorporating information from adjacent regions. This is known as *borrowing strength* and commonly used in the statistical analysis of sparse data.

1.2 Markov random field models

A widely used class of models for (spatially) correlated data in discrete space are Markov random field (MRF) models. Basically, a MRF defines a joint distribution on a random vector. In a hierarchical Bayesian framework, MRFs are used to specify the correlation structure of parameters in the prior distribution. Most common are Gaussian Markov random field (GMRF) models, where the joint distribution is multivariate normal. Formulations for other distributions are possible (Besag 1974), but less common in practice.

The idea behind MRF models is to use a conditional approach to specify the joint distribution of parameters. We will give a short overview following the notation of Besag & Kooperberg (1995). Suppose we are interested in parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ corresponding to (spatial) units $i = 1, \ldots, n$. Instead of modeling the joint distribution $p(\boldsymbol{\lambda})$ explicitly, the conditional distribution $p(\lambda_i | \boldsymbol{\lambda}_{-i})$ is specified, where $\boldsymbol{\lambda}_{-i}$ is the vector of parameters without the ith element. For GMRFs, this conditional distribution is assumed to be Gaussian with conditional mean and variance

$$\mathrm{E}(\lambda_i | \boldsymbol{\lambda}_{-i}) = \sum_{j \neq i} \beta_{ij} \lambda_j \quad \text{and} \quad \mathrm{Var}(\lambda_i | \boldsymbol{\lambda}_{-i}) = \kappa_i\,, \quad i = 1, \ldots, n.$$

The coefficients $\beta_{ij},\ i \neq j$, control the conditional correlation between parameters λ_i and λ_j. For $\beta_{ij} = 0$ the corresponding elements are conditionally uncorrelated. The joint distribution of $\boldsymbol{\lambda}$ is well-defined if we demand

$$\beta_{ij}\kappa_j = \beta_{ji}\kappa_i, \quad \text{for all } i, j \in \{1, \dots, n\}.$$

Hence, the matrix $\boldsymbol{Q} = (q_{ij})$ with

$$q_{ij} = -\frac{\beta_{ij}}{\kappa_i} \quad \text{for } i \neq j \quad \text{and} \quad q_{ii} = \frac{1}{\kappa_i}, \quad \text{for } i = 1, \dots, n, \tag{1.4}$$

is symmetric. For positive definiteness of $\boldsymbol{Q}$, sufficient requirements are $\beta_{ij} \geq 0$ and $\sum_j \beta_{ij} \leq 1$ with $\sum_j \beta_{ij} < 1$ for at least one i. For given $\boldsymbol{Q}$, the joint density of parameters $\boldsymbol{\lambda}$ is

$$p(\boldsymbol{\lambda}) \propto \exp\left\{\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{Q}\boldsymbol{\lambda}\right\}, \tag{1.5}$$

a multivariate normal distribution with precision matrix $\boldsymbol{Q}$, usually called a Gaussian conditional autoregression. The inverse $\boldsymbol{Q}^{-1}$ of the precision matrix is the covariance matrix of $\boldsymbol{\lambda}$. Often the diagonal elements of the precision matrix are chosen to be $q_{ii} = -\sum_{i \neq j} q_{ij}$ for $i = 1, \dots, n$, and (1.5) can be simplified to

$$p(\boldsymbol{\lambda}) \propto \exp\left\{\frac{1}{2}\sum_{i<j} q_{ij}(\lambda_i - \lambda_j)^2\right\}. \tag{1.6}$$

This specific choice for the precision matrix implies that the elements of $\boldsymbol{Q}$ sum up to zero in each row; the precision matrix $\boldsymbol{Q}$ is only positive semi-definite and the covariance matrix does not exist. Thus, (1.6) is no proper distribution anymore and the notation as a density is slightly incorrect but intuitive. Often, this is called a *pairwise difference prior* since it is solely based on pairwise differences of the parameters whereas an overall mean is not defined. This form is frequently used as a prior distribution in Bayesian models, and sometimes called a Gaussian intrinsic autoregression. Note that the associated posterior is proper in most cases.

The advantage of the conditional approach for MRFs is the possibility to restrict conditional correlations to be non-zero for small sets of parameters. Especially for spatial applications it is convenient to assume $\beta_{ij} > 0$ only for geographically adjacent regions i and j. Such a formulation allows to perform spatial smoothing on the parameters.

An application to disease mapping data is proposed by Besag, York & Mollié (1991) and illustrated here briefly since it will repeatedly be referred to in the following chapters. They decompose the relative risk λ_i in (1.2) to

$$\lambda_i = \exp(u_i + v_i), \quad i = 1, \dots, n,$$

where $\boldsymbol{u} = (u_1, \dots, u_n)$ are spatially structured effects and $\boldsymbol{v} = (v_1, \dots, v_n)$ are region-specific uncorrelated random effects. For the structured effects a pairwise difference prior (1.6) is used with overall precision parameter κ. Note that now κ denotes the precision (the inverse variance) and not the variance. Let further $m_i,\ i = 1, \dots, n$, denote the number of regions that are adjacent to region i. The coefficients are chosen to be $\beta_{ij} = 1/m_i$ for geographically adjacent regions i and j—denoted by $i \sim j$—and $\beta_{ij} = 0$ otherwise. The conditional precisions are $\kappa_i = m_i\kappa$. The prior "density" is

$$p(\boldsymbol{u}|\kappa) \propto \exp\left\{ -\frac{\kappa}{2} \sum_{i \sim j} (u_i - u_j)^2 \right\}. \tag{1.7}$$

The random effects are assumed to be Gaussian white noise

$$p(v|\tau) \propto \exp\left\{ -\frac{\tau}{2} \sum_{i=1}^{n} v_i^2 \right\}$$

with precision τ.

The pairwise difference prior (1.6) is not limited to the Gaussian case. Modeling spatial correlations based on pairwise differences of parameters is possible with other specifications as well. Besag, Green, Higdon & Mengersen (1995) give a general formulation

$$p(\boldsymbol{\lambda}|\gamma) \propto \exp\left\{ -\sum_{i \sim j} \omega_{ij} \phi\big(\gamma(\lambda_i - \lambda_j)\big) \right\}, \tag{1.8}$$

where the summation is over all indices of pairs of adjacent regions $i \sim j$ and hence the corresponding weights ω_{ij} are assumed to be non-zero. The class of models is determined by the symmetric function ϕ, i.e. $\phi(z) = \phi(-z)$, and the scale parameter γ. For $\phi(z) = \frac{1}{2}z^2$, $\gamma = \sqrt{\kappa}$, and $\omega_{ij} = 1$ this is the Gaussian pairwise difference prior (1.7). In general, any of these models defines an improper (prior) distribution.

Although MRF models are widely used and very popular for modeling and smoothing in discrete space, there exist various other models as well. Møller & Waagepetersen (1998) introduced Markov connected component fields with applications to image analysis data. This is a general class of models with emphasis on separating clusters from background information. Later, Gangnon & Clayton (2000) applied such models to disease mapping data with the focus on finding clusters. There exist related approaches, mostly in the field of image analysis, which take the discrete nature of space into account but assume some regular structure, e.g. arrays of pixels (Johnson 1994). Wolpert & Ickstadt (1998) propose an alternative class of models for count data. In fact their approach is rather general in that it is suitable for both discrete and continuous space. All these model are rather complex and defined within a Bayesian framework. There are also non-Bayesian approaches. For example, Müller, Stadtmüller & Tabnak (1997) model aggregated count data in terms of an underlying intensity function in continuous space.

1.3 Bayesian modeling and reversible jump MCMC

The proposed methodology in this thesis is defined within a hierarchical Bayesian framework. Inference on the unknown parameters is carried out in terms of Markov chain Monte Carlo

(MCMC) techniques. Such MCMC algorithms have become very popular in statistics over the last decade. Meanwhile, there is a vast literature on theoretical properties and practical applications of MCMC samplers. In this section, we provide basic notations and definitions which are helpful for the understanding of the proposed MCMC sampler. The focus is on problems where the number of parameters in the model is unknown.

Suppose we are given data $\boldsymbol{y} = (y_1, \ldots, y_n)$ and assume some parametric observation model M with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$. The likelihood is denoted by $p(\boldsymbol{y}|\boldsymbol{\theta})$. Any Bayesian analysis requires a prior distribution for the unknown parameters with joint density $p(\boldsymbol{\theta})$. Throughout, we will assume that the elements of $\boldsymbol{\theta}$ are real valued, i.e. $\Theta \subset \mathbb{R}^k$, and that the corresponding densities exist. Statistical inference is based on the posterior density $p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$, according to Bayes' theorem.

Within a hierarchical framework, the prior distribution $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\gamma})$ depends on parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_m)$ that are subject to statistical inference themselves. Throughout, the parameters $\boldsymbol{\gamma}$ are called *hyperparameters* and the corresponding density $p(\boldsymbol{\gamma})$ is referred to as *hyperprior*. Adding an additional level to the hierarchy of the model makes the prior distribution more flexible. Moreover, without external prior knowledge the choice of the prior distribution is usually based on subjective decisions. Unknown hyperparameters that are estimated within the algorithm with respect to the data make an objective contribution to the prior.

For complex hierarchical models and high-dimensional parameter spaces the normalizing constant $p(\boldsymbol{y})$ of the posterior density prohibits any analytical calculation of posterior quantities. Therefore, sampling techniques are widely used. Such MCMC algorithms are based on Markov chains whose stationary distribution coincides with the posterior distribution of the parameters. Simulation of such a Markov chain produces a (dependent) sample from the posterior distribution. Any quantity of interest, e.g. posterior median and quantiles, can be calculated from the MCMC output by Monte Carlo techniques.

1.3.1 Reversible jump MCMC

For models with a fixed number k of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, the most general MCMC sampler is the Metropolis-Hastings (MH) algorithm (Hastings 1970). We will only give a brief overview, for a thorough discussion of statistical properties and convergence issues see the paper of Tierney (1994).

The MH algorithm provides a recipe to construct an irreducible and aperiodic Markov chain Z with state space Θ and transition kernel P such that Z has a stationary distribution with density $\pi = p(\boldsymbol{\theta}|\boldsymbol{y})$. The main issue is to define the transition kernel P in such a way that the stationary distribution is the posterior distribution of $\boldsymbol{\theta}$, and that the Markov chain Z is able to pass through the whole parameter space Θ. The principle of the simulation is to exploit the Markov property and randomly propose a new state conditional on the current one. Thus, the basic sampling scheme of the MH algorithm is to choose an arbitrary initial state $\boldsymbol{z}^{(0)} \in \Theta$ and iteratively generate states $\boldsymbol{z}^{(t+1)}$ from $\boldsymbol{z}^{(t)}$ by applying the transition kernel P, $t = 0, \ldots, T-1$.

After convergence of the chain, all or some of the states are collected. This leads to a sample $\{z^{[1]}, \ldots, z^{[S]}\}$ from the posterior distribution. The sample size $S \leq T$ can be chosen arbitrarily.

For the MH algorithm the transition kernel P is constructed as follows. Suppose the state of the Markov chain is $z^{(t)}$ at iteration t. A candidate state z^* is drawn randomly from some proposal distribution with density $q(z^{(t)}, z^*)$ depending on the current state $z^{(t)}$. This candidate state is accepted with probability $\alpha = \alpha(z^{(t)}, z^*)$ and rejected with probability $1 - \alpha$. The new state $z^{(t+1)}$ is given by

$$z^{(t+1)} = \begin{cases} z^* & \text{if } u \leq \alpha, \\ z^{(t)} & \text{otherwise,} \end{cases}$$

where u is drawn uniformly distributed on $[0, 1]$. The acceptance probability α of the MH algorithm is defined as

$$\alpha(z, z^*) = \min\left\{1, \frac{\pi(z^*)}{\pi(z)} \cdot \frac{q(z^*, z)}{q(z, z^*)}\right\} = \min\left\{1, \frac{p(\boldsymbol{y}|z^*)}{p(\boldsymbol{y}|z)} \cdot \frac{p(z^*)}{p(z)} \cdot \frac{q(z^*, z)}{q(z, z^*)}\right\}$$

for states $z, z^* \in \Theta$. Here, a transition $z \longrightarrow z^*$ and the reverse move $z^* \longrightarrow z$ are compared. If we denote by

$$\mathcal{L} = \frac{p(\boldsymbol{y}|z^*)}{p(\boldsymbol{y}|z)}, \quad \mathcal{P} = \frac{p(z^*)}{p(z)}, \quad \text{and} \quad \mathcal{Q} = \frac{q(z^*, z)}{q(z, z^*)}$$

the likelihood ratio, the prior ratio, and the proposal ratio, respectively, the acceptance probability can be abbreviated by

$$\alpha = \min\{1, \mathcal{L} \cdot \mathcal{P} \cdot \mathcal{Q}\}.$$

In many applications, e.g. for all samplers in this thesis, it is convenient to apply several proposal distributions. In general, we implement different *moves* $h = 1, \ldots, H$, where each move h performs a different type of modification of the current state. In each iteration the move type h is specified according to a fixed scheme or chosen randomly. Throughout we will work with randomly proposed moves, chosen with respect to some proposal distribution r. In this case, the proposal density is given by $r_h q_h(z, z^*)$ for a move of type h.

So far we have only considered a state space Θ of fixed dimension. The reversible jump MCMC (RJMCMC) algorithm, proposed by Green (1995), is an extension of the MH algorithm to problems where the dimension of the parameter space is variable. The main idea is to derive a more flexible algorithm that allows for data based model choice.

For a fixed parametric observation model M, it is common to omit the model M from all formulas, just like we have done above. In theory, both the prior and the likelihood are conditional on M. Still, the model M is implicit in the prior and the likelihood, and the simpler notation is justified. Often, there will be some uncertainty about the data generating process and hence a (finite or countable) set $\mathcal{M}$ of competing models may be considered.

It is convenient to classify the models in $\mathcal{M}$ in terms a model indicator $k \in \mathbb{N}$ that determines the dimension of the parameter space. Thus, the set $\mathcal{M} = \{M_1, M_2, \ldots\}$ contains models with parameter spaces of different dimensions. We assume that model M_k has parameters $\boldsymbol{\theta}_k \in \Theta_k$. Note that k is not necessarily the actual dimension d_k of $\boldsymbol{\theta}_k$, but defines it in a unique way. For simplicity, we assume that $d_1 < d_2 < \ldots$, i.e. the dimension of the parameter vector $\boldsymbol{\theta}_k$ is increasing with increasing k.

The RJMCMC algorithm allows inference on the model indicator k, i.e. on $\mathcal{M}$. For a fixed model M_k, the state space of the Markov chain is $\{k\} \times \Theta_k$. Thus, with variable model indicator k the state space of the Markov chain is $\Theta = \bigcup_{k \in \mathbb{N}}(\{k\} \times \Theta_k)$. Note that for transitions within one model M_k, i.e. for $\boldsymbol{z}, \boldsymbol{z}^* \in \Theta_k$, the RJMCMC algorithm is simply a MH algorithm.

Some care has to be taken when switching between models with different model indicators. In accordance with the definition of the model indicator such moves will be called *dimension changing* moves. In the MH algorithm, any transition and its reverse transition are enabled by the same move type. With dimension changing moves this is not possible. Whenever one move increases the number of parameters, the reverse move has to decrease it. Therefore, any RJMCMC algorithm is based on matched pairs of dimension changing moves.

Various matched pairs of moves are conceivable. Suppose we design a move that changes the model indicator from k to $k+1$. Simultaneously, the dimension of the state space is increased from d_k to d_{k+1}. Thus, this move requires the generation of at least $d_{k+1} - d_k$ parameters. Following the nomenclature of Green (1995) we will call this a *birth* move in our sampler. Accordingly, the reverse move is called a *death* move. Note that transitions from k to $\tilde{k} > k+1$ are possible but not necessarily helpful.

The fundamental idea behind the reversible jump methodology is perform a transition in fixed dimension instead of the actual transition between spaces of different dimensions. This allows to apply the concept of the MH algorithm to variable dimension problems.

Consider a birth move that changes the model indicator from k to $k^* = k+1$. Suppose the candidate state $\boldsymbol{z}^*_{k+1}$ is generated based on the current state $\boldsymbol{z}_k$ and an additional random vector $\boldsymbol{u}$ of dimension d_u with density $q(\boldsymbol{u})$. The new state is given by some deterministic function g_B of the current state and the random numbers, i.e.

$$\boldsymbol{z}^*_{k+1} = g_B(\boldsymbol{z}_k, \boldsymbol{u}).$$

Accordingly, for the reverse move $\boldsymbol{z}^*_{k+1} \rightarrow \boldsymbol{z}_k$ the state is derived by

$$\boldsymbol{z}_k = g_D(\boldsymbol{z}^*_{k+1}, \boldsymbol{v})$$

with some deterministic function g_D and a randomly generated vector $\boldsymbol{v}$ of dimension d_v with density $q(\boldsymbol{v})$. The functions g_B and g_D must be chosen to assure reversibility, i.e. in such a way that the moves birth and death match. The fundamental assumption of the reversible jump methodology is the dimension matching condition

$$d_k + d_u = d_{k+1} + d_v.$$

If this condition holds, the pair of dimension changing moves is transformed into a pair of moves in the same dimension, $(\boldsymbol{z}_k, \boldsymbol{u}) \rightarrow (\boldsymbol{z}^*_{k+1}, \boldsymbol{v})$ and reverse. The crucial point is that the parameters $\boldsymbol{u}$ and $\boldsymbol{v}$ do not appear in the state of the Markov chain.

A standard MH step can be applied to these transitions. The likelihood and the prior are straightforward, and the corresponding components of the acceptance probability do not change. What differs is the proposal ratio. Suppose a birth move is proposed with probability $r_B = r_B(k)$ which may even depend on the current model indicator k. The new state is proposed with density $q_B(z_k, z^*_{k+1}) = q(\boldsymbol{u})$. The reverse move is proposed with probability $r_D = r_D(k+1)$ and the corresponding proposal density is $q_D(z^*_{k+1}, z_k) = q(\boldsymbol{v})$. In general, the determinant of the Jacobian of the deterministic transformations g_B and g_D has to be taken into account. Thus, the proposal ratio for the birth move is given by

$$\mathcal{Q} = \frac{r_D(k+1)}{r_B(k)} \cdot \frac{q_D(z^*_{k+1}, z_k)}{q_B(z_k, z^*_{k+1})} \cdot \left| \frac{\partial(z^*_{k+1}, \boldsymbol{v})}{\partial(z_k, \boldsymbol{u})} \right| = \frac{r_D(k+1)}{r_B(k)} \cdot \frac{q(\boldsymbol{v})}{q(\boldsymbol{u})} \cdot \mathcal{J}.$$

In accordance to the notation above, the acceptance probability of the dimension changing moves is sometimes written as

$$\alpha = \min\{1,\ \mathcal{L} \cdot \mathcal{P} \cdot \mathcal{Q} \cdot \mathcal{J}\}.$$

Note that the determinant of the Jacobian $\mathcal{J}$ is not an inherent component of the reversible jump moves (Waagepetersen & Sorensen 2001), but enters solely by the use of deterministic transformations. In all applications in this thesis we have $\mathcal{J} = 1$. Moreover, even fixed dimension samplers can be constructed in such a way that a Jacobian enters in the proposal ratio. Still, this is not common and usually $\mathcal{J}$ is not written separately in the acceptance probability. For RJMCMC this is different. Often it is necessary to construct carefully designed moves based on such transformations. However, separating the Jacobian term from the proposal ratio is simply done for emphasis.

The reversible jump methodology has been used extensively in various fields of applications. The outline above is rather specific with regard to the following methodology. For a more general discussion of RJMCMC algorithms see the review by Green (2003).

1.3.2 Partition modeling

There are various applications in statistics, where data $\boldsymbol{y} = (y_1, \ldots, y_n)$ needs to be divided into groups $C_1, \ldots, C_k$. One can distinguish two cases, whether the groups are or are not known in advance, respectively. In the first case, this is a problem of classification, and the goal is to find out to which group individuals $i = 1, \ldots, n$ belong. In the second case, e.g. cluster analysis, the groups are unknown and determined on the basis of one or more characteristics of the individuals.

More general, consider a finite set of objects $V = \{1, \ldots, n\}$. We follow the definition of Hartigan (1990) and call a family of subsets $V_1, \ldots, V_k$ a partition of V if

$$V_i \cap V_j = \emptyset \quad \text{for } i \neq j \qquad \text{and} \qquad V_1 \cup \ldots \cup V_k = V.$$

A *product distribution* (or *product model*) for the possible partitions is given by

$$p(\{V_1, \ldots, V_k\}) \propto \prod_{j=1}^{k} c(V_j) \tag{1.9}$$

where $c(V_j)$ are specified non-negative cohesions for each subset.

Suppose we are given observations $\boldsymbol{y}$ and want to find a partition of the objects based on the observations. A *product partition model* for the observations is a product model for the partitions under the assumption

$$p(\boldsymbol{y}|V_1, \ldots, V_k) = \prod_{j=1}^{k} p(\boldsymbol{y}_j|V_j) \tag{1.10}$$

of conditional independence of observations $\boldsymbol{y}_j = \{y_i : i \in V_j\}$ given the partition $\{V_1, \ldots, V_k\}$. This formulation is less general than that given by Hartigan (1990) but sufficient for this thesis.

In a parametric setting, we assume an observation model M with parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ corresponding to the subsets in the partition. It is convenient to define a partition on the set of objects V by means of the parameters (Barry & Hartigan 1992). Thus, we assume that observations $\boldsymbol{y}_j$ in subset V_j arise from the same model with parameter θ_j. Now, the product model (1.9) is equivalent to

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{k} p(\theta_j).$$

Assuming further independence of the observations in V_j given the parameter θ_j, i.e.

$$p(\boldsymbol{y}_j|\theta_j) = \prod_{i \in V_j} p(y_i|\theta_j),$$

the product partition model (1.10) can be written as

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{j=1}^{k} p(\boldsymbol{y}_j|\theta_j) = \prod_{j=1}^{k} \prod_{i \in V_j} p(y_i|\theta_j).$$

Within a Bayesian framework, $p(\boldsymbol{\theta})$ is the joint prior density for $\boldsymbol{\theta}$ and $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the likelihood. In a partition model, both can be factorized with respect to the partition. Therefore, the posterior density can also be factorized accordingly.

So far, the number k of subsets of the partition was assumed to be fixed. However, in many applications, k will be unknown in advance. If we allow for a variable number of subsets, the number of parameters, i.e. the dimension of $\boldsymbol{\theta}$, will change. Still, inference on the posterior distribution is possible in terms of RJMCMC.

1.3.3 Posterior model averaging

Finally, we will briefly comment on the analysis of posterior samples. In the following we propose a clustering partition model. The terms *cluster* and *clustering* will be used repeatedly. This merely corresponds to the properties of the proposed partition on the set of units $V = \{1, \ldots, n\}$, but not to the goal of our statistical analysis.

Within a reversible jump framework, a partition into k subsets (or clusters) corresponds to a model in class M_k. Thus, the set of models $\mathcal{M}$ contains the set of all possible partitions.

Since k is variable, the samples generated by the RJMCMC algorithm belong to different models. Hence, the posterior probability for a specific model M_k is provided by the sampler and the algorithm allows for model choice, based on the posterior samples.

However, we are not interested in a cluster analysis in a classical sense. The focus of our analysis is not on finding clusters among the elements of V, but on the estimation of the corresponding parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, e.g. the relative risks in the disease mapping context.

Basically, it is possible to derive estimates for the unknown parameters conditional on the number of clusters, i.e. conditional on the model M_k. However, this neglects the uncertainty about the model indicator k. Therefore, we calculate posterior point estimates by averaging over all models, visited by the RJMCMC sampler. Throughout, all estimates for unknown parameters are derived by such model averaging. No posterior probabilities for specific partitions are calculated.

1.4 Outline of the thesis

The thesis is structured as follows. A general framework for the proposed clustering partition model is developed in Chapter 2. Some elementary properties of clustering partitions are derived, and the construction of a prior distribution for the use within a hierarchical Bayesian model is described. The first part of Chapter 3 features a published paper (Knorr-Held & Raßer 2000, Biometrics) with an application of the proposed model to disease count data. The remainder of this chapter was not included in the paper. Here, extensions of the basic model are provided; in Section 3.5 the incorporation of covariate information is described, whereas in Section 3.6 an alternative prior specification is discussed. In Chapter 4, the proposed model is further investigated. At first, the model is transferred to image analysis data. Here, we assume a Gaussian observation model. Then, the smoothing behavior of the model, some properties of the prior distribution, and computational issues are discussed. For all of these topics a comparison to Markov random field models is given. In Chapter 5, the methodology is extended to the more general case of space-time disease count data. The focus in this chapter is on the modeling of space-time interactions. Finally, the first part of Chapter 6 consists of a published paper (Knorr-Held, Raßer & Becker 2002, Biometrics) in which two model formulations for multicategorical disease data are developed and implemented in terms of Markov random field priors. The second part, not included in the paper, proposes an equivalent formulation based on a partition model.

Chapter 2

Clustering Partition Models for Discrete Structures

We are concerned with statistical analyses in discrete space and therefore have a finite set of units $\{1, \ldots, n\}$, $n \in \mathbb{N}$, under investigation. Suppose we are given corresponding observations $\boldsymbol{y} = (y_1, \ldots, y_n)$. We assume that there are no observations missing, so for each unit i there is an observation y_i. Yet, the proposed model—subject to slight changes in the construction of the prior distribution—might also be useful in the case of missing observations. A few comments on this matter will be given at the end of this chapter.

The data $\boldsymbol{y}$ are assumed to originate from some parametric observation model M with parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ for units $\{1, \ldots, n\}$. A partition model defined on a set of units assumes that observations in subsets of the units arise from the same distribution. In a Bayesian analysis such assumptions are embedded as prior information in the model. Thus, the joint prior distribution for the unknown parameters $\boldsymbol{\lambda}$ has partition model form, and partitioning is performed on the set of parameters. The assumption that observations within one subset arise from a model with the same parameters imputes some sort of similarity on the units. Hence, the partition is implicitly given if adequate prior knowledge is available.

In the absence of such knowledge, a partition model prior offers a convenient way to improve estimates within one subset and perform some sort of smoothing on the parameters. The term "smooth" itself refers to some structure within the set of units because there is no plausible way to define a smooth parameter surface in non-ordered space. Therefore it is intuitive to consider only those partitions that preserve the structure of the data. For some applications such partitions into $k \leq n$ subsets are straightforward to imagine.

Example 2.1 Consider a sequence of consecutive time points $\{t_1, \ldots, t_n\}$. A partition into k subsets $T_1, \ldots, T_k$ is defined by $T_j \subset \{t_1, \ldots, t_n\}$, $j = 1, \ldots, k$, $k \leq n$. Instead of using arbitrary subsets it seems natural to define

$$T_1 = \{t_{i_0}, \ldots, t_{i_1}\},\ T_2 = \{t_{i_1} + 1, \ldots, t_{i_2}\},\ \ldots,\ T_k = \{t_{i_{k-1}} + 1, \ldots, t_{i_k}\}$$

with ordered end points $t_1 = t_{i_0} \leq t_{i_1} < \ldots < t_{i_{k-1}} < t_{i_k} = t_n$. In Figure 2.1 a partition with

$k = 4$ subsets

$$T_1 = \{t_1, t_2, t_3\},\ T_2 = \{t_4, t_5\},\ T_3 = \{t_6\},\ T_4 = \{t_7, t_8\}$$

is given. Assuming the same parameter θ_j for each time point in subset $T_j,\ j = 1, \ldots, 4$, defines a step function on the parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_8)$

$$\begin{aligned} \theta_1 &= \lambda_1 = \lambda_2 = \lambda_3\,, \\ \theta_2 &= \lambda_4 = \lambda_5\,, \\ \theta_3 &= \lambda_6\,, \\ \theta_4 &= \lambda_7 = \lambda_8 \end{aligned}$$

and the partition of the parameters preserves the order of the time points. Alternatively, the step function can be parameterized in terms of change points. This illustrates that the partition is based on the structure of the data, i.e. the order of the time points.

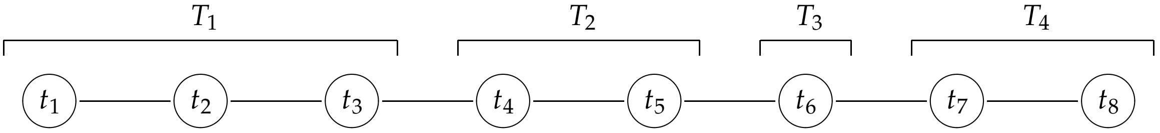


Figure 2.1: A partition of 8 time points into 4 subsets which consist of subsequent time points.

The definition of a partition is somewhat more difficult if the units under investigation are not ordered in such a regular manner. In this chapter we will propose a method to construct partitions on arbitrary finite sets of units. Given the partition, a prior model for the use within a Bayesian framework is derived. To highlight the idea of partitioning data with respect to a given underlying structure, we will call the subsets *clusters* and our prior model a *clustering partition model*.

Note that in continuous space partitions are always "clustering" in our terminology. For example, any step function on an interval $[a, b] \subset \mathbb{R}$ takes the location of single atoms (points) into account. Still, this is not true in discrete space where we may define the clusters solely in terms of covariates. However, such a definition would neglect the structure of the data, e.g. the order of the time points in Example 2.1.

2.1 Clustering partitions for discrete structures

In the following, we will propose a method to construct a partition on a finite set of units under investigation. The formulation is rather general and works for many applications, although initially the idea was used to define a partition on geographical maps. Such maps consist of regions, where the shape and size as well as the number of adjacent regions varies considerably. Due to this irregular structure, continuous settings (e.g. Euclidean space) are not suitable. Our construction is useful particularly for geographical maps, but can be applied more generally.

2.1.1 Neighborhood structure and underlying graphs

We distinguish between two major cases, whether data is observed in units that do or do not display a structure, respectively, i.e. the units

1. have a specific location to each other or

2. are mutually exchangeable.

We concentrate on the first case, where the units are structured, regardless of the observed data. We may look at the units under investigation as the vertices of an undirected graph, while the structure is given by the edges which are or are not present between any two vertices. This structure is fixed and induced by the units. We call this structure the underlying graph G. Some examples for underlying graphs arise from the applications presented in this thesis:

(a) One-dimensional sequences, e.g. for time series data:
Suppose we are given data at several (equally spaced) points in time $t = 1, \ldots, T$. For T large one might consider a continuous model; for a fairly small number of points a discrete model is more appropriate. The structure in the data is the temporal order, we want to preserve in the partition (see Example 2.1).

(b) Two-dimensional lattices, e.g. for image analysis data:
This is the two-dimensional analogue with a slightly more complex structure. Now units (or pixels) are identified by pairs $(i, j),\ i = 1, \ldots, n_1,\ j = 1, \ldots, n_2,\ n_1, n_2 \in \mathbb{N}$. The pixels are arranged as a matrix with n_1 rows and n_2 columns. An obvious generalization are lattices where some pixels are not observed or missing, e.g. an image of irregular shape where the observed pixels are arranged on a lattice.

(c) Irregular space, e.g. geographical maps:
This is a more general case and any systematic identification of units (or regions) is not possible. Yet, the underlying structure is obvious and based on the common borders of regions.

Since disease mapping is the major application in this thesis, our terminology is based on such geographical maps. Two geographical regions are called *neighbors* or *adjacent* if they share a common border. In this case an edge is present in the underlying graph. In general the regions will have different numbers of neighbors and we speak of *irregular* space. This is the most general case of an underlying graph and includes no dimension-statement.

In contrast, we will call structures *regular* if all units have the same number of neighbors (except for units on the border of the space). Such regular grids may be defined in various dimensions. Most common are lattice graphs, where units are identified as squares arranged as a matrix, see Figure 2.2. For such graphs usually two-different neighborhoods are used: (a) first-order and (b) second-order, where each unit has four or eight neighbors, respectively. Other definitions of higher order neighbors are possible, but rarely used. Identifying the units by squares allows to apply the same definition of neighborhood as for geographical regions. The assumption that two units are neighbors, if they share a common border is equivalent to the first order neighborhood as displayed in Figure 2.2 (a). This definition is often used in MRFs (e.g. Besag 1974) and we will also use only this definition. More formally, for a lattice graph where the units are arranged as a matrix with n_1 rows and n_2 columns, the neighborhood of unit (i, j), $1 < i < n_1$, $1 < j < n_2$

$$\{(i, j-1),\ (i-1, j),\ (i+1, j),\ (i, j+1)\}$$

consists of four neighbors not adjacent to each other. Regions on the border of the space have less neighbors; two for the corner units, three otherwise. This may lead to edge effects in applications and therefore sometimes artificial neighborhoods are defined, wrapping the lattice on a torus. However, this is not helpful for our applications and we will use the neighborhoods as defined above. Note that there exist various other regular grids, e.g. grids which consist of hexagons as in Figure 2.2 (c). Our model is suitable for such structures but there is no application reported in this thesis.

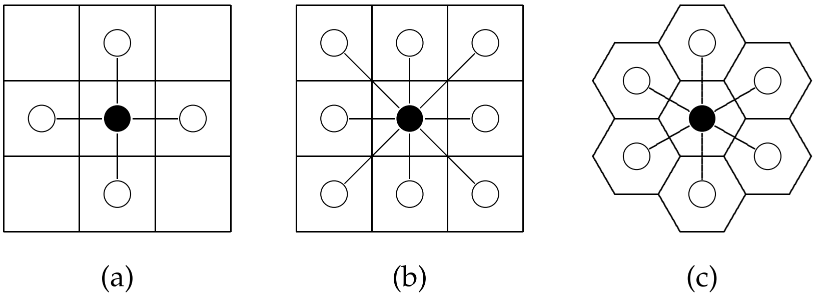


Figure 2.2: Commonly used neighborhoods for regular grids.

Besides data with an underlying graph, there exist data without any fixed neighborhood structure. The units under investigation are exchangeable. This can be seen as a special case where the underlying graph contains no edges between the vertices. For unstructured data, any partition has to be based on observations alone. This is of minor interest for the scope of this thesis and unstructured data will not be investigated.

2.1.2 Connected graphs and distance

The underlying structure for a finite set of units is best described in terms of an undirected graph. In this section some basic notations are recalled, for further details see Gould (1988). Suppose we are given a graph $G = \{V, E\}$, i.e. a finite set of vertices $V = \{1, \ldots, n\}$ and a set of edges E. Two vertices $i, j \in V$ are called *neighbors* or *adjacent*, if there is an edge $e_{ij} \in E$ between them. We will consider undirected edges only and write $i \sim j$ if vertices i and j are adjacent. The neighborhood structure of G is defined by the set of edges E only. The $n \times n$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ij} = 1$ if $i \sim j$ and 0 otherwise is called the adjacency matrix of G.

For $i \neq j$ a sequence of subsequently adjacent vertices $i, v_1, v_2, \ldots, v_p, j \in V$ is called a *walk* between i and j. The number of edges $p+1$ is the *length* of the walk. A walk in which no vertex is repeated is called a *path*. The graph G is called *connected* if there is a path between any two vertices in V. We will call a path between i and j *minimal* if it has minimal length in the set of all paths between i and j. The last statement is used to define a discrete distance metric on G.

Definition 2.1 *Let G be a connected graph with vertices V and $i, j \in V$. The distance $d(i,j)$ between i and j is the length of a minimal path between i and j.*

Although there may be two or more different minimal paths between two vertices, the length of the minimal path and therefore the distance is unique. Obviously, the distance of adjacent vertices is 1. Furthermore, Definition 2.1 ensures for all $i, j, l \in V$

1. $d(i,j) \geq 0$, and $d(i,j) = 0$ if and only if $i = j$, [Positivity]

2. $d(i,j) = d(j,i)$, [Symmetry]

3. $d(i,j) \leq d(i,l) + d(l,j)$, [Triangle inequality]

so d is a distance metric on G. Note that equality $d(i,j) = d(i,l) + d(l,j)$ is gained if and only if the intermediate vertex l is on one of the minimal paths between i and j. The $n \times n$ matrix $\boldsymbol{D} = (d_{ij})$ with $d_{ij} = d(i,j)$ will be called the distance matrix of G.

Computation of distances is initialized by

$$\begin{aligned} d(i,i) &= 0 \quad \text{for } i \in V\,, \\ d(i,j) &= 1 \quad \text{for } i \sim j\,, \end{aligned}$$

as given in the adjacency matrix. Higher distances are computed in a recursive way

$$d(i,l) = d(i,j) + d(j,l) = 2$$

for adjacent units $i \sim j$ and $j \sim l$ with $i \neq l,\ i \not\sim l$ and so on. Alternatively, computation is possible using powers of the adjacency matrix. Cell (i,j) in $\boldsymbol{A}^m$ contains the number of different walks of length m between vertices i and j (Gould 1988, Theorem 1.3.1). Since the shortest walk is a minimal path, distance $d(i,j)$ for $i \neq j$ can be computed taking powers of $\boldsymbol{A}$ until there appears a non-zero entry in cell (i,j) for the first time, i.e.

$$d(i,j) = \min\left\{ m \in \mathbb{N}\backslash\{0\} : a_{ij}^{(m)} > 0 \right\},$$

where $a_{ij}^{(m)}$ denotes the entry in cell (i,j) of $\boldsymbol{A}^m$. For any graph the distance measure is discrete with values $d(i,j) \in \{0, 1, \ldots, n-1\}$ for all vertices $i, j \in V$.

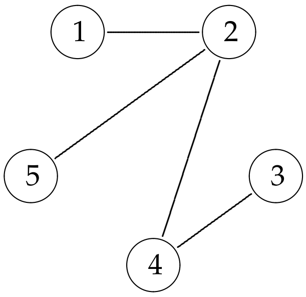


Figure 2.3: Connected graph with 5 vertices and 4 edges.

Example 2.2 In Figure 2.3 a connected graph with five vertices $V = \{1,2,3,4,5\}$ and four edges $E = \{e_{12}, e_{24}, e_{25}, e_{34}\}$ is given. The adjacency matrix $\boldsymbol{A}$ has four non-zero entries at positions corresponding to the set of edges E. The adjacency matrix and its second and third power are (written as upper triangular matrices)

$$\boldsymbol{A} = \begin{pmatrix} 0 & \boxed{1} & 0 & 0 & 0 \\ & 0 & 0 & \boxed{1} & \boxed{1} \\ & & 0 & \boxed{1} & 0 \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}, \quad \boldsymbol{A}^2 = \begin{pmatrix} 1 & 0 & 0 & \boxed{1} & \boxed{1} \\ & 3 & \boxed{1} & 0 & 0 \\ & & 1 & 0 & 0 \\ & & & 2 & \boxed{1} \\ & & & & 1 \end{pmatrix}, \quad \boldsymbol{A}^3 = \begin{pmatrix} 0 & 3 & \boxed{1} & 0 & 0 \\ & 0 & 0 & 4 & 3 \\ & & 0 & 2 & \boxed{1} \\ & & & 0 & 0 \\ & & & & 0 \end{pmatrix}.$$

The first non-zero entries are marked with boxes. With these matrices, the off-diagonal elements of the distance matrix are already defined. The distance matrix

$$\boldsymbol{D} = \begin{pmatrix} 0 & 1 & 3 & 2 & 2 \\ & 0 & 2 & 1 & 1 \\ & & 0 & 1 & 3 \\ & & & 0 & 2 \\ & & & & 0 \end{pmatrix}$$

can be computed from the information in the adjacency matrix alone.

Note that the adjacency matrix must be given explicitly for irregular graphs while for regular lattices of any dimension the distances are implicit. For a one-dimensional sequence of consecutively numbered units the distance between units i and j is simply $d(i,j) = |i - j|$. For a two-dimensional lattice with n_1 rows and n_2 columns the distance $d(i,j)$ for vertices $i = (i_1, i_2), j = (j_1, j_2)$ can be decomposed into two components, one for each dimension

$$d(i,j) = |i_1 - j_1| + |i_2 - j_2|.$$

In fact, for lattices of any dimension, the distance can be computed as the sum over the one-dimensional distances.

2.1.3 Definition of clustering partitions

A clustering partition combines units under investigation into clusters with respect to the underlying graph. More precisely, we construct a *partition* or *cluster configuration* $\mathcal{C}$ on the set of units $\{1, \ldots, n\}$ so that all units are assigned to one and only one cluster. The basic intention is to combine "similar" units and thus allow for deliberate interpretation. For this purpose we need to measure similarity which will naturally focus on the observed data. But taking into account the different data structures (cf. Section 2.1.1) we want to perform structure-preserving clustering if appropriate.

Under exchangeability of the units, i.e. for unstructured data, a partition $\mathcal{C} = \mathcal{C}(\boldsymbol{y}|M)$ depends only on the observed data $\boldsymbol{y}$, conditional on the observation model M. But if there is an underlying graph, the partition $\mathcal{C} = \mathcal{C}(\boldsymbol{y}|M, G)$ is also conditional on the graph G and information from the neighborhood structure is used for clustering. The construction of the partition $\mathcal{C}$ guarantees that for any two units in one cluster, there is a path between these two units in this cluster, hence the name clustering partition.

To perform clustering that preserves the structure of the underlying graph we will use a discrete version of *Voronoi diagrams*. For a better understanding, a brief overview on Voronoi diagrams is given. We start with a definition of Voronoi diagrams in continuous space and refer to Okabe, Boots & Sugihara (1992) for a more detailed description.

The most popular version of Voronoi diagrams (or *Voronoi tessellations*) exists for finite continuous space $S \subset \mathbb{R}^2$ in two dimensions. Here, a partition of S into k tiles is achieved by chosing k generating points $g_1, \ldots, g_k \in S$. The tiles or *Voronoi polygons* are defined to be

$$T_j = \{x \in S : d(x, g_j) \leq d(x, g_i),\ 1 \leq i \leq k,\ j \neq i\}, \quad j = 1, \ldots, k, \tag{2.1}$$

where d is some appropriate distance metric. Usually the Euclidean distance is used, i.e. $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$ for $x = (x_1, x_2)$, $y = (y_1, y_2) \in \mathbb{R}^2$. According to (2.1), any point on the border of a Voronoi polygon belongs to two or more polygons. Except for the boundaries of the polygons—a set of Lebesgue measure zero—each $x \in S$ is assigned to one and only one polygon by this definition. In other words, point $x \in S$ is assigned to polygon T_j if

$$j = \arg\min_l \{d(x, g_l)\}. \tag{2.2}$$

Extensions to the one-dimensional case or even to higher dimensional spaces are straightforward and involve only the use of different distance measures. Furthermore, this construction has the appealing feature that one can easily define a probability measure on all possible partitions. Since the partition is defined in a unique way by the choice of the generating points, any probability measure can be based on a vector of length k. The fact that boundary points belong to more than one cluster is of minor interest for statistical applications since any continuous (non-singular) probability measure on S has zero probability on the set of borders.

The easy construction of a probability measure is the main reason for adapting and extending the method of Voronoi diagrams to discrete spaces. Indeed, the formulation is very similar to the continuous case. The transfer requires the use of an appropriate distance measure as well as the definition of an updated assignment rule.

Suppose, we are given an undirected connected graph $G = \{V, E\}$, where $V = \{1, \ldots, n\}$ is a finite set of vertices representing the units under investigation and E is a nonempty set of edges. Similar to the continuous case, we construct a partition $\mathcal{C}_k$ using a set of $k \leq n$ vertices $\{g_1, \ldots, g_k\}$, $g_j \in V$, as cluster centers. The vector $\boldsymbol{g}_k = (g_1, \ldots, g_k)$ is called *generating vector* of $\mathcal{C}_k$. A unique partition of V into k clusters $C_1, \ldots, C_k$ is achieved by assigning all vertices $i \in V$ to one and only one of the cluster centers. The generated partition is called a *clustering partition*.

The assignment of vertices to cluster centers is based on the distance measure d as introduced in Definition 2.1. The construction of the cluster configuration $\mathcal{C}_k = \{C_1, \ldots, C_k\}$ is performed in two steps. First, for all vertices $i \in V$ the general rule

$$i \in C_j \quad \Leftrightarrow \quad d(i, g_j) < d(i, g_l), \quad 1 \leq l \leq k, \, l \neq j \tag{2.3}$$

is applied, i.e. each vertex is assigned to the cluster center to which it has minimal distance. Due to the discrete distance measure, this assignment is not necessarily sufficient. In a second step, for vertices with equal minimal distance to two or more cluster centers an additional rule is applied to assure uniqueness of the partition. Let $L(i)$ be the set of indexes of the cluster centers vertex i has equal minimal distance to

$$L(i) = \{l_1, \ldots, l_m\} \quad \Leftrightarrow \quad d(i, g_{l_1}) = \ldots = d(i, g_{l_m}), \, m \leq k.$$

These ties are broken by the additional rule

$$i \in C_j \quad \Leftrightarrow \quad j = \min\{l_1, \ldots, l_m\}, \, m \leq k. \tag{2.4}$$

Less formally, vertex i is assigned to cluster C_j if cluster center g_j ranks first among all candidate cluster centers in the generating vector. Hence, the cluster configuration $\mathcal{C}_k$ is unique. Furthermore, $\mathcal{C}_k$ is a partition of V (as defined in Section 1.3.2) since $\bigcup_{j=1}^{k} C_j = V$ and $C_j \cap C_l = \emptyset$ for $j \neq l$.

2.1.4 Properties of clustering partitions

Before we construct a prior distribution based on the proposed clustering partition, it is useful to investigate the properties of the clusters further. Note first, rule (2.3) assures that cluster centers are assigned to the cluster which they generate since $0 = d(g_j, g_j) < d(g_j, g_l)$ for $j \neq l$. This is a natural and expected connection between cluster centers and clusters. However, there are other desirable properties of clustering partitions.

In the following, we assume a clustering partition $\mathcal{C}_k = \{C_1, \ldots, C_k\}$ with k clusters as defined above.

Proposition 2.1 *Let vertex $i \in V \backslash \{g_j\}$ be assigned to cluster C_j, $j \leq k$, with distance $d(i, g_j) = p + 1 > 1$ to the cluster center g_j. Then, for any minimal path $g_j, v_1, \ldots, v_p, i$ between g_j and i*

$$v_l \in C_j \quad \textit{for all } 1 \leq l \leq p.$$

A formal proof of this proposition is given in Appendix A.1, although it is reasonable to assume that if a vertex is assigned to a cluster, all vertices on the minimal path to the cluster center are assigned to the same cluster. Of course, not all vertices with a smaller distance to the cluster center than vertex i have to be in the same cluster but those lying on a minimal path—in some sense "on the way"—from the cluster center to vertex i. From this proposition an important conclusion can be drawn.

Corollary 2.1 (Connectivity) *In a clustering partition for any two vertices $i_1, i_2 \in C_j$, $i_1 \neq i_2$ in the same cluster C_j, there exists a path $i_1, v_1, \dots, v_p, i_2$ between i_1 and i_2 with $v_l \in C_j$ for all $1 \leq l \leq p$. Thus, all vertices within the same cluster are connected.*

This corollary is crucial for the proposed model since it shows that the partition preserves the underlying structure. A proof is omitted, as this statement follows from Proposition 2.1 in a direct way. According to Corollary 2.1 a clustering partition can also be interpreted as a decomposition of the graph G into disjoint subgraphs $G_1, \dots, G_k$ which are connected graphs again.

Corollary 2.1 only states the existence of a path between any two vertices, but not the existence of a minimal path. Indeed, this is only possible for special cases depending on the neighborhood structure. For lattices of one or two dimensions, as in some of the applications in the following sections, we can show a discrete version of convexity.

Proposition 2.2 (Convexity for lattice graphs) *Suppose the vertices are arranged on a lattice with n_1 rows and n_2 columns. Let vertices $i_1, i_2 \in C_j$, $i_1 \neq i_2$ with distance $d(i_1, i_2) = p + 1 > 1$ be assigned to the same cluster. Then there exists a minimal path $i_1, v_1, \dots, v_p, i_2$ between i_1 and i_2 with $v_l \in C_j$ for all $1 \leq l \leq p$.*

Note that for vertices i_1, i_2 with distance $d(i_1, i_2) \leq 1$ this is true for arbitrary graphs. Furthermore, the proposition holds for the special case with $n_1 = 1$, e.g. a sequence of time points.

The statement in Proposition 2.1 concerns all minimal paths and is stronger, but only valid for paths between vertices and the corresponding cluster center. Proposition 2.2 testifies the existence of one minimal path, but not all minimal paths must be in the same cluster. This can be seen as a weak form of convexity. The proof of Proposition 2.2 is given in Appendix A.2. This property is not surprising since in Euclidean space Voronoi polygons are always convex (Okabe et al. 1992). In a discrete setting, however, this is not valid for arbitrary graphs. In fact, it is rather easy to construct counterexamples, see Appendix A.3. Therefore, the decomposition of G into subgraphs does not necessarily retain the distance measure. The distance $d_j(i_1, i_2)$ in subgraph G_j, defined as the length of the minimal path between regions $i_1, i_2 \in C_j$ with intermediate vertices in the same cluster, will be greater in general, i.e. $d_j(i_1, i_2) \geq d(i_1, i_2)$.

To summarize the properties of the clustering partition model, we have to state first of all that for arbitrary graphs clusters can be quite irregular. The distance measure defined above neglects the size and the shape of the clusters completely and is defined by the neighborhood structure alone. Yet, the terms "cluster" and "cluster center" are justified by Corollary 2.1

and Proposition 2.1, respectively. Further properties like convexity can only be derived for special cases. Still it can be pointed out that the proposed clustering procedure does not support arbitrary partitions. In general, for a given graph G, further attributes can only be investigated in simulation studies.

Beside the properties of single clusters, also the properties of the whole partition are of interest. According to rule (2.4) the order of the cluster centers is crucial for the construction of a partition $\mathcal{C}_k$ with k clusters, so the corresponding generating vector g_k is kept non-ordered. This leads to an enormous number of possible vectors g_k. For a graph with n vertices and a fixed number k of clusters, there are $n!/(n-k)!$ different vectors g_k. Although a fixed vector g_k defines a unique partition $\mathcal{C}_k$, there may be different vectors defining the same partition. This becomes obvious for $k = 1$ with n different generating vectors for the same partition. Even more extreme, for $k = n$ there are $n!$ different generating vectors, while in the resulting partition each vertex is in a separate cluster. Therefore, for most graphs, the number of different partitions will be much lower than the number of different vectors of cluster centers.

The total number of different generating vectors is $N_G(n) = \sum_{k=1}^{n} n!/(n-k)!$ and depends only on the number of vertices n of the graph G. N_G increases extremely fast (see Table 2.1) and for more than 170 vertices, the number of different vectors is no longer accessible (by standard computation).

units	1	2	3	4	5	6	7	8	9	$\cdots$	170	$\cdots$
vectors	1	4	15	64	325	1,956	13,699	109,600	986,409	$\cdots$	$1.973 \cdot 10^{307}$	$\cdots$

Table 2.1: Total number of generating vectors $N_G(n)$ for different numbers of vertices n.

The number of possible partitions depends on the neighborhood structure and cannot be derived analytically in general. For a given graph the number of partitions can be computed by checking all possible generating vectors. In practice, this is limited to graphs with very few vertices.

Example 2.3 The partition $\{T_1, T_2, T_3, T_4\}$ with $k = 4$ clusters in Example 2.1 can be generated by $g_k = (t_2, t_8, t_4, t_6)$. Still, there are various other generating vectors that produce the same partition. According to Table 2.1 for $n = 8$ time points there exist 109,600 different generating vectors.

The number of different partitions can easily be calculated using a change point formulation. For one-dimensional sequences, a clustering partition with k clusters can be equivalently parameterized by $k-1$ change points. With 7 possible positions for the change points, there are only

$$\sum_{k=0}^{7} \binom{7}{k} = 2^7 = 128$$

different partitions. Hence, the number of different partitions is much smaller than the number of different generating vectors.

Considering the applications in this thesis, the minimum number of vertices is 400. Investigation of the properties of the partition models is carried out in terms of simulation studies (see Section 4.3). In any case, the set of possible partitions is countable.

To close this section, we give some remarks on clustering partitions that are useful for the following applications.

Remark 1: The definition of the distance d only works for connected graphs with a path between any two vertices. Neither a single vertex nor a subset of vertices may be separated from the rest. Especially in geographical maps, there are often islands clearly separated from other regions. In such cases artificial neighborhoods have to be defined. Yet, the same problem exists for other discrete spatial models, e.g. for MRFs.

Remark 2: The clustering partition $\mathcal{C}_k$ is invariant to multiplication of the distance measure d with a positive constant $b > 0$. A new distance $\tilde{d} = b \cdot d$ again is a distance metric and its usage will have no effect on the assignment process according to equations (2.3) and (2.4). In fact, the partition is invariant to any strictly monotonic increasing transformation of d, although in general the new distance $\tilde{d}$ is no distance metric any more, e.g. $\tilde{d} = a + b \cdot d$ with $a \neq 0$, $b > 0$. Voronoi diagrams are well-defined with such generalized distances (Okabe et al. 1992) and so are clustering partitions.

2.2 Construction of a prior distribution

To formulate a prior model for the unknown parameters based on a clustering partition within a Bayesian framework, a suitable probability measure needs to be defined. We will construct such a probability measure in terms of clustering partitions. Still, the fact that vertices within each cluster are connected is not necessary and all formulations are valid for arbitrary partitions. All probabilities are expressed as densities as it is common within MCMC applications.

In our partition model, parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ are assumed to arise from the same (parametric) distribution in each cluster of the partition. Therefore the prior on $\boldsymbol{\lambda}$ depends on the partition. For a clustering partition $\mathcal{C}_k = \{C_1, \ldots, C_k\}$ with k clusters generated by $\mathbf{g}_k = (g_1, \ldots, g_k)$ we define the parameters $\boldsymbol{\lambda}$ on individual level to be a deterministic function of parameters $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)'$ on cluster level

$$\lambda_i = f(\theta_j) \quad \text{for } i \in C_j, j = 1, \ldots, k.$$

We replace the prior on the parameters $\boldsymbol{\lambda}$ by a prior on $\boldsymbol{\theta}_k$ together with the choice of some function f. In general, the prior $p(\boldsymbol{\theta}_k | \mathbf{g}_k, k)$ for $\boldsymbol{\theta}_k$ is conditional on the number of clusters and the generating vector. Due to the large number of possible generating vectors $\mathbf{g}_k$, we simplify this and specify the prior with respect to the number of clusters alone

$$p(\boldsymbol{\theta}_k | \mathbf{g}_k, k) = p(\boldsymbol{\theta}_k | k),$$

regardless of the cluster configuration. Nonetheless, the following results are stated in the more general notation. Any appropriate density may be chosen as a prior guess for $\boldsymbol{\theta}_k$. We will construct a reversible jump algorithm in which inference is on the number of clusters k as well as on the partition $\mathcal{C}_k$ and hence we need a prior $p(\boldsymbol{g}_k, k)$. The joint prior density is given by

$$p(\boldsymbol{\theta}_k, \boldsymbol{g}_k, k) = p(\boldsymbol{\theta}_k | \boldsymbol{g}_k, k) p(\boldsymbol{g}_k, k).$$

To summarize, a prior distribution for the unknown parameters $\boldsymbol{\lambda}$ is derived by

1. specifying an appropriate function f, relating $\boldsymbol{\lambda}$ to $\boldsymbol{\theta}_k$,
2. choosing a prior $p(\boldsymbol{\theta}_k | \boldsymbol{g}_k, k)$ for $\boldsymbol{\theta}_k$ conditional on the parameters of the partition and
3. constructing a prior $p(\boldsymbol{g}_k, k)$ for the parameters of the partition.

This defines a hierarchical prior for the unknown parameters $\boldsymbol{\lambda}$. The construction is now described in detail, starting with the lowest level in the hierarchy, the partition.

2.2.1 A prior distribution for clustering partitions

It would be desirable to assign a probability to each possible partition. Unfortunately, this is not feasible because the number of different partitions is unknown in general and cannot be computed for most applications as shown above. Yet, the construction of the clustering partition offers an alternative. Since the number of differing generating vectors can be derived by combinatorial arguments, a probability measure is straightforward to define. In accordance to the construction of the partition, this is also a hierarchical prior. First, a distribution $p(k)$ for the number of cluster centers is specified. Then a probability $p(\boldsymbol{g}_k | k)$ on all possible generating vectors $\boldsymbol{g}_k$ is defined conditional on the number of clusters.

For partition models in continuous space it is useful to restrict the number of subsets $k \leq K$ by an upper bound $K < \infty$ (Green 1995, Denison, Adams, Holmes & Hand 2002). In discrete space, this is not necessary since $K = n$ is a natural upper bound for the number of cluster centers. Therefore, any discrete probability measure for k is allowed, assuring $1 \leq k \leq n$. Throughout, we will apply one of the following three distributions: (1) uniform on $\{1, \dots, n\}$, (2) geometric with parameter $c \in [0, 1)$, or (3) Poisson with parameter $\mu > 0$. The corresponding prior probabilities $p(k)$ are

$$(1) \quad p(k) = \frac{1}{n}, \qquad (2) \quad p(k) \propto (1-c)^k, \qquad (3) \quad p(k) \propto \frac{\mu^k}{k!} \qquad \text{for } 1 \leq k \leq n.$$

Note that the geometric and Poisson distribution have to be truncated to $\{1, \dots, n\}$. With appropriate choices of the hyperparameters c and μ, both, the geometric and the Poisson distribution will favor smaller values of k.

Given the number of clusters k, the prior distribution for $p(\boldsymbol{g}_k | k)$ is based on simple combinatorics. The cluster centers are nuisance parameters in the analysis and have no further influence on the estimates. Moreover, the relation between cluster centers and partition is not fully known. Therefore, in lack of substantial prior knowledge, we choose an uninformative prior distribution. We assume that all generating vectors have equal probability

$$p(g_k|k) = \frac{(n-k)!}{n!}$$

regardless of the given graph. This prior distribution is uninformative in the sense that it does not give preference to any of the generating vectors.

2.2.2 A prior for the parameters

The prior model is completed with a prior for the unknown parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$. Of course, the choice of the function f and the prior for the parameters $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)'$ are strongly related. Throughout, we will use a piecewise constant formulation with independent parameters $\boldsymbol{\theta}_k$. In the following, some basic properties are derived.

First, the parameters are assumed to be constant within one cluster, i.e.

$$\lambda_i = \theta_j \quad \text{for } i \in C_j\,, j = 1, \ldots, k. \tag{2.5}$$

In our applications the parameters are scalars, $\lambda_i, \theta_j \in \mathbb{R}$ for $i = 1, \ldots, n,\ j = 1, \ldots, k$. Alternatively, we may rewrite (2.5) in matrix notation

$$\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k, \quad \boldsymbol{B} = (b_{ij}),\ i = 1, \ldots, n,\ j = 1, \ldots, k, \tag{2.6}$$

where $\boldsymbol{B}$ is a $n \times k$ matrix with $b_{ij} = 1$ if $i \in C_j$ and $b_{ij} = 0$ otherwise. Hence, $\boldsymbol{\lambda}$ is a linear transformation of $\boldsymbol{\theta}_k$. The matrix $\boldsymbol{B} = \boldsymbol{B}(g_k, k)$ depends on the partition and further information can be extracted. The product $\boldsymbol{B}'\boldsymbol{B}$ contains the cluster sizes

$$\boldsymbol{B}'\boldsymbol{B} = \mathrm{diag}(m_1, \ldots, m_k),$$

where m_j is the number of vertices assigned to cluster C_j. This is easily shown since cell (l, j) of $\boldsymbol{B}'\boldsymbol{B}$ is given by

$$\sum_{i=1}^{n} b'_{li} b_{ij} = \sum_{i=1}^{n} b_{il} b_{ij} = \begin{cases} \sum_{i=1}^{n} b_{ij}^2 = m_j & \text{for } l = j, \\ 0 & \text{for } l \neq j. \end{cases}$$

Second, independence of parameters $\theta_1, \ldots, \theta_k$ yields

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\lambda}) &= \mathrm{Cov}(\boldsymbol{B}\boldsymbol{\theta}_k) \\ &= \boldsymbol{B}\,\mathrm{Cov}(\boldsymbol{\theta}_k)\boldsymbol{B}' \\ &= \boldsymbol{B}\,\mathrm{diag}(\mathrm{Var}(\theta_1), \ldots, \mathrm{Var}(\theta_k))\boldsymbol{B}' \end{aligned}$$

and

$$\mathrm{Cor}(\boldsymbol{\lambda}) = \boldsymbol{B}\,\mathrm{Cor}(\boldsymbol{\theta}_k)\boldsymbol{B}' = \boldsymbol{B}\boldsymbol{B}'.$$

for the covariance matrix and correlation matrix of $\boldsymbol{\lambda}$, respectively.

Example 2.4 The parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_8)'$ in Example 2.1 are given by (2.6) with

$$\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \lambda_5 \\ \lambda_6 \\ \lambda_7 \\ \lambda_8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{pmatrix} = \boldsymbol{B}\boldsymbol{\theta}_4.$$

The cluster sizes are

$$\boldsymbol{B}'\boldsymbol{B} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix} = \mathrm{diag}(3, 2, 1, 2)$$

and the correlation matrix

$$\boldsymbol{B}\boldsymbol{B}' = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} = \mathrm{diag}(1_3, 1_2, 1_1, 1_2)$$

is block diagonal, due to the order of time points.

Note that θ_j needs not necessarily be a scalar. For example for time series data, one might assume a piecewise linear development in each cluster, $\lambda_i = \alpha_j + \beta_j t_i,\ t_i \in T_j$. In this case $\theta_j = (\alpha_j, \beta_j)$ is a vector where both components are constant within one cluster. Of course, the matrix notation (2.6) is now different. A similar approach with piecewise linear regression models is described in Holmes, Denison & Mallick (1999). Throughout, we stick to the simpler choice of scalar parameters for practical reasons. For non-ordered irregular spaces, it will be extremely difficult to define some appropriate functional relationship.

Our model assumes that parameters $\theta_1, \ldots, \theta_k$ are independent of each other. While the choice of constant parameters arises out of practicability in the first place, the assumption of independence is crucial for the adaptiveness to the data. Some comments on this matter are given in Section 4.2. The prior for the parameters $\boldsymbol{\theta}_k$ must be chosen in reference to the data or, more formally, to the specified observation model M. In general, under independence, the joint density for $\boldsymbol{\theta}_k$ is

$$p(\boldsymbol{\theta}_k | g_k, k) = \prod_{j=1}^{k} p(\theta_j | g_k, k). \qquad (2.7)$$

The hierarchical prior for $\boldsymbol{\lambda}$ consists of a clustering partition $\mathcal{C}_k$ on the set of vertices with varying number of clusters k and corresponding priors $p(\theta_j|g_k,k)$, $j = 1,\ldots,k$, for the cluster parameters. We will call this a *clustering partition model* (CPM) prior for $\boldsymbol{\lambda}$.

Let (g_k,k) define a clustering partition $\mathcal{C}_k$ with k clusters. The joint prior density for $\boldsymbol{\lambda}$ given the partition is

$$p(\boldsymbol{\lambda}|g_k,k) = \begin{cases} p(\boldsymbol{\theta}_k|g_k,k) & \text{if } \boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

The proof is given in Appendix A.4.

Now, for given data $\boldsymbol{y} = (y_1,\ldots,y_n)$ and an observation model M with parameters $\boldsymbol{\lambda}$, the likelihood $p(\boldsymbol{y}|\boldsymbol{\lambda})$ can be expressed in terms of the parameters $\boldsymbol{\theta}_k$ and the partition $\mathcal{C}_k$. If we assume that observations $\boldsymbol{y}$ are independent given the parameters $\boldsymbol{\lambda}$ the likelihood is

$$p(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} p(y_i|\lambda_i).$$

Using a CPM for the unknown parameters $\boldsymbol{\lambda}$ the likelihood conditional on (g_k,k) can always be factorized to

$$p(\boldsymbol{y}|\boldsymbol{\lambda},g_k,k) = \prod_{j=1}^{k}\prod_{i\in C_j} p(y_i|\theta_j,g_k,k) = \prod_{j=1}^{k} p(\boldsymbol{y}_j|\theta_j,g_k,k) = p(\boldsymbol{y}|\boldsymbol{\theta}_k,g_k,k), \tag{2.9}$$

where

$$p(\boldsymbol{y}_j|\theta_j,g_k,k) = \prod_{i\in C_j} p(y_i|\theta_j,g_k,k) \tag{2.10}$$

is the contribution of observations $\boldsymbol{y}_j$ in cluster C_j to the likelihood. According to (2.7) and (2.9), the posterior for the parameters $\boldsymbol{\theta}_k$ (given the partition) is

$$\begin{aligned} p(\boldsymbol{\theta}_k|\boldsymbol{y},g_k,k) &= \frac{1}{p(\boldsymbol{y}|g_k,k)} p(\boldsymbol{y}|\boldsymbol{\theta}_k,g_k,k)p(\boldsymbol{\theta}_k|g_k,k) \\ &= \frac{1}{p(\boldsymbol{y}|g_k,k)} \prod_{j=1}^{k} p(\boldsymbol{y}_j|\theta_j,g_k,k)p(\theta_j|g_k,k) \\ &= \frac{\prod_{j=1}^{k} p(\boldsymbol{y}_j|g_k,k)}{p(\boldsymbol{y}|g_k,k)} \prod_{j=1}^{k} \frac{p(\boldsymbol{y}_j|\theta_j,g_k,k)p(\theta_j|g_k,k)}{p(\boldsymbol{y}_j|g_k,k)} \\ &= \prod_{j=1}^{k} p(\theta_j|\boldsymbol{y}_j,g_k,k), \end{aligned}$$

where

$$\begin{aligned} p(\boldsymbol{y}|g_k,k) &= \int\cdots\int p(\boldsymbol{y}|\boldsymbol{\theta}_k,g_k,k)p(\boldsymbol{\theta}_k|g_k,k)d\theta_1\ldots d\theta_k \\ &= \prod_{j=1}^{k}\int p(\boldsymbol{y}_j|\theta_j,g_k,k)p(\theta_j|g_k,k)d\theta_j \\ &= \prod_{j=1}^{k} p(\boldsymbol{y}_j|g_k,k) \end{aligned}$$

is a normalizing constant. Hence, for a fixed partition $\mathcal{C}_k$ defined by (g_k, k), the joint posterior is the product of the posterior distributions in the clusters. Unfortunately, inference is not on $\boldsymbol{\theta}_k$ for a fixed partition but on $\boldsymbol{\lambda}$ with varying partitions. Using a prior on the partitions, the posterior for $\boldsymbol{\lambda}$ is

$$\begin{aligned} p(\boldsymbol{\lambda}|\boldsymbol{y}) &= \sum_{(g_k,k)} p(\boldsymbol{\lambda}|\boldsymbol{y}, g_k, k) p(g_k, k) \\ &= \sum_{(g_k,k)} p(\boldsymbol{\theta}_k|\boldsymbol{y}, g_k, k) p(g_k, k) \\ &= \sum_{(g_k,k)} p(g_k, k) \prod_{j=1}^{k} p(\theta_j|\boldsymbol{y}_j, g_k, k), \end{aligned} \tag{2.11}$$

the weighted sum over the countable set of partitions. This posterior can not be factorized anymore. Therefore, the CPM is no product partition model anymore, due to the additional prior on the partition.

2.2.3 Summary

To summarize the CPM prior some notations are given in Table 2.2. In the following applications, often the model will be specified in terms of the parameters $\boldsymbol{\theta}_k$ only. This notation is done for simplicity and rather intuitive.

individual level	vertices, regions	i	$i = 1, \ldots, n,\ n \in \mathbb{N}$
	observations	y_i	$y_i \in \mathbb{N}, \mathbb{R}, \mathbb{R}^2, \ldots$
	parameters	λ_i	$\lambda_i \in \mathbb{R}, \mathbb{R}^+$
cluster level	clusters	C_j	$j = 1, \ldots, k,\ 1 \leq k \leq n$
	cluster centers	g_j	$g_j \in \{1, \ldots, n\}$
	parameters	θ_j	$\theta_j \in \mathbb{R}, \mathbb{R}^+$

Table 2.2: Basic components of the CPM.

Note that for the most part, we will omit the partition parameters from the formulas. For example, we will denote the likelihood by $p(\boldsymbol{y}|\boldsymbol{\lambda})$ or $p(\boldsymbol{y}|\boldsymbol{\theta}_k)$ instead of $p(\boldsymbol{y}|\boldsymbol{\lambda}, g_k, k)$ or $p(\boldsymbol{y}|\boldsymbol{\theta}_k, g_k, k)$, respectively.

The posterior (2.11) cannot be derived analytically, but can be approximated by drawing samples from it. In consideration of the hierarchical prior, a reversible jump MCMC algorithm is implemented that allows for the variation of the number of clusters k, the partition $\mathcal{C}_k$, as well as the unknown parameters $\boldsymbol{\theta}_k$. Some general statements on this matter are given in the next section. Details on prior distributions are given for all applications separately, mainly following the implementation from the disease mapping example in Chapter 3.

As already mentioned at the beginning of this chapter, the CPM can be modified in case some observations are missing. Suppose that for some of the units under investigation no observations are available. Such missing data causes no problems in our Bayesian analysis as long as the likelihood (2.10) can be computed for all clusters for all partitions. This is the case, if there exists at least one observation in each cluster, i.e. for all clusters C_j there exists at least one i with $i \in C_j$ and observation y_i not missing.

Suppose, we choose the cluster centers only among those vertices where the corresponding observations are not missing. As shown at the beginning of Section 2.1.4, the cluster centers are always assigned to the cluster which they generate. Hence, there is in all clusters at least one observation. The MCMC algorithm runs without problems. Of course, the quality of the results depends on the number of missing observations. Although the posterior exists, more elaborated approaches might be useful here.

2.3 Sampling and posterior probability statements

For data $\boldsymbol{y} = (y_1, \ldots, y_n)$ inference focuses on the corresponding parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. In a Bayesian setting inference is based on the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ of the parameters given the data. Using a CPM prior for the unknown parameters $\boldsymbol{\lambda}$, the posterior (2.11) is not analytically tractable.

We construct a reversible jump MCMC algorithm to collect samples from the posterior. We produce subsequent states $\boldsymbol{z}^{(m)}$ of the Markov chain Z for iterations $m = 1, \ldots, N$, where each state is given by the model indicator k, the clusters $C_1, \ldots, C_k$ as defined by a generating vector $\boldsymbol{g}_k = (g_1, \ldots, g_k)$ and parameters $\theta_1, \ldots, \theta_k$. Thus, the state of the Markov chain will be defined on cluster level. Beside standard moves to update model parameters or hyperparameters, our algorithm implies moves to change the clustering partition. Especially dimension changing moves are implemented to increase or decrease the number of clusters throughout the algorithm. The moves are proposed by random, following some distribution r. The definition of this distribution depends on the number and type of moves, i.e. on the application.

Starting the chain with an initial state $\boldsymbol{z}^{(0)}$, we discard burn-in iterations $m = 1, \ldots, B$ and collect the states $m = B + 1, \ldots, N$ after convergence. To avoid high autocorrelations in the samples, not all iterations are stored, but equidistant steps are made with lag L. Therefore, the sample size is $S = (N - B)/L$.

Suppose that we have collected samples $s = 1, \ldots, S$ from the posterior. Sample s consists of a model indicator $k^{[s]}$, a generating vector $\boldsymbol{g}^{[s]}_{k^{[s]}}$, and the parameters $\boldsymbol{\theta}^{[s]}_{k^{[s]}}$. The samples of the parameters λ_i on individual level are sequences

$$\left\{\lambda_i^{[1]}, \ldots, \lambda_i^{[S]}\right\} = \left\{\theta^{[1]}_{j(i,1)}, \ldots, \theta^{[S]}_{j(i,S)}\right\}, \quad i = 1, \ldots, n,$$

where $j(i,s)$ is the index of the cluster, vertex i is assigned to in sample s. Point estimates for the parameters are derived by model averaging in terms of posterior means

$$\hat{\lambda}_i = \frac{1}{S} \sum_{s=1}^{S} \lambda_i^{[s]} = \frac{1}{S} \sum_{s=1}^{S} \theta^{[s]}_{j(i,s)}, \quad i = 1, \ldots, n,$$

or posterior medians

$$\hat{\lambda}_i = \text{med}\Big\{\theta^{[1]}_{j(i,1)}, \dots, \theta^{[S]}_{j(i,S)}\Big\}, \quad i = 1, \dots, n.$$

We will mainly use posterior medians as point estimates. This has the advantage that quantiles for the parameters can be calculated in a similar way. Yet, the MCMC output offers a lot more possibilities of further inference. First of all, the posterior distribution for the number of clusters k is given by

$$P(k = m) = \frac{1}{S} \left|\{s : k^{[s]} = m\}\right|, \quad m = 1, \dots, n.$$

Of special interest is the probability that two (adjacent) vertices i_1 and i_2 arise from a model with the same parameter. This can be approximated by the probability that those two vertices are in the same cluster

$$P(i_1 \text{ is in the same cluster as } i_2) = \frac{1}{S} \left|\{s : j(i_1, s) = j(i_2, s)\}\right|.$$

Some care has to be taken in interpreting probability statements as the one above, since our model does not support arbitrary partitions. Suppose, the data proposes a partition with k_1 clusters not supported by the CPM. Such a partition can well be approximated by a clustering partition with a higher number of clusters $k_2 > k_1$, where some clusters have similar parameters. Therefore, the probability that two vertices have the same parameter will be higher in general than the probability that they are in the same cluster. This approximation will be acceptable for adjacent vertices, but with increasing distance it will most likely get worse.

For the most part, we will not consider further inference on the cluster centers. Although such inference is straightforward, it is of poor explanatory power for reasons described above. Any inference can also be done conditional on the number of clusters, e.g. the number with the highest posterior probability. Inference conditional on a specific partition is also possible from a theoretical point of view. In practice, this will be impossible since the same partition will rarely be visited twice by the Markov chain, due to the extremely large number of possible partitions.

Chapter 3

Bayesian Detection of Clusters and Discontinuities in Disease Maps

This chapter addresses the statistical modeling of aggregated disease count data. As mentioned in the introduction, the CPM was originally developed for the purpose of mapping disease risk. Moreover, this is the major field of application in this thesis. Further models are implemented similar to this introductory example. Therefore, we provide a detailed description of the basic algorithm.

The first part of this chapter, i.e. Sections 3.1 to 3.4 (pp. 32–46), was originally published in the paper ”Bayesian Detection of Clusters and Discontinuities in Disease Maps” by Knorr-Held & Raßer, ©The International Biometric Society, 2000. Note that in the present version some minor modifications have been made to match the notation with other chapters in this thesis. The list of references is now included in the bibliography of the thesis. The paper is reprinted with kind permission from the International Biometric Society.

The subsequent Sections 3.5 and 3.6 are additional and not included in the original paper. Both sections discuss innovations of the fundamental disease mapping model. First, in Section 3.5 a methodology to incorporate covariate information into the basic model is provided. Such information allows to adjust the model for known risk factors, whenever the covariates are measured (or available) on the same geographical resolution. This simplifies interpretation of the results and increases the significance of the estimates. Two different model formulations for categorical and metrical covariates are described in detail. The results are compared to those gained with the standard model.

Finally, in Section 3.6 an alternative prior specification for the model without covariates is proposed. The use of a conjugate prior distribution for the cluster parameters allows to simplify the algorithm. The RJMCMC sampler is now based on the marginal likelihood, whereas the relative risk parameters are estimated separately. Again, the results are briefly discussed in comparison to the original model.

Bayesian Detection of Clusters and Discontinuities in Disease Maps

Leonhard Knorr-Held and Günter Raßer
Institute of Statistics
University of Munich
Ludwigstr. 33, 80539 Munich
Germany

Email: leo@stat.uni-muenchen.de rasser@stat.uni-muenchen.de

Abstract

An interesting epidemiological problem is the analysis of geographical variation in rates of disease incidence or mortality. One goal of such an analysis is to detect clusters of elevated (or lowered) risk in order to identify unknown risk factors regarding the disease. We propose a nonparametric Bayesian approach for the detection of such clusters based on Green's (1995) reversible jump MCMC methodology. The prior model assumes that geographical regions can be combined in clusters with constant relative risk within a cluster. The number of clusters, the location of the clusters and the risk within each cluster is unknown. This specification can be seen as a change-point problem of variable dimension in irregular, discrete space. We illustrate our method through an analysis of oral cavity cancer mortality rates in Germany and compare the results with those obtained by the commonly used Bayesian disease mapping method of Besag, York & Mollié (1991).

Key words: Cancer atlas; Clustering; Disease mapping; Oral cavity cancer; Relative risk; Reversible jump MCMC.

3.1 Introduction

Statistical methods for analyzing data on disease incidence or mortality over a set of contiguous geographical regions have gained increasing interest in the last decade. It is still very common in disease mapping to display the standard mortality ratio (SMR), the ratio of observed cases y over expected cases e, for each region either on a relative or an absolute scale. However, these maps can be seriously misleading because the SMRs tend to be far more extreme in less populated regions, especially for rare diseases. Hence, regions with the least reliable data will typically draw the main visual attention. For a thorough discussion of this issue see Clayton & Bernardinelli (1992).

As an example consider Figure 3.1, which displays the geographical variation of the standard mortality ratios for males and oral cavity cancer, 1986–1990, in all 544 districts of Germany. This data set will be analyzed later in Section 3.3. The SMRs vary between 0.15 and 2.40 with a standard deviation of the log SMRs of 0.386. However, the variation of the SMRs is reduced if we only consider highly populated regions. For example, a subsample of regions with more than 50 expected cases has a minimal SMR of 0.53 and a maximal SMR of 1.60. The standard deviation of the log SMRs is now 0.255, indicating that the SMRs tend to be more extreme in less populated regions, but this conclusion is drawn under the assumption of constant risk in the whole of Germany.

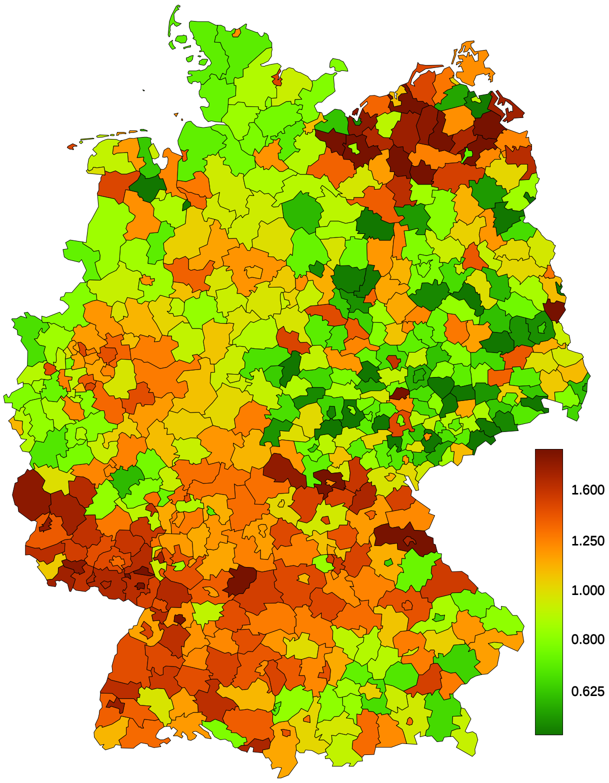


Figure 3.1: Standard mortality ratios for oral cavity cancer of males in Germany.

Indeed, an unknown part of the variation of the SMRs may be caused by geographically varying unobserved risk factors. For example, in Figure 3.1 there seem to be areas of higher risk in the north-east and in some parts of the south, especially towards the west, but a naive visual inspection can be seriously misleading and no general conclusion can be drawn from such a map. Therefore, so-called disease mapping methods have been developed to give more reliable estimates of the geographical variation of disease risk. The general goal is to identify the extra-sample variation due to unobserved heterogeneity by filtering the Poisson sample variation.

A well-known method is the empirical Bayes approach of Clayton & Kaldor (1987). Roughly speaking, this method shrinks the SMRs towards a local or a global mean where the amount of shrinkage is determined by the reliability of the data of that particular region. The two smoothing options “local” or “global” seem to be appropriate if unobserved risk factors do or do not have a spatial structure, respectively. However, one of the major goals of disease maps is to identify unobserved risk factors through the geographical variation of the disease so the spatial distribution of those unobserved factors is not known in advance. This led Besag et al. (1991) to generalize the Clayton & Kaldor method allowing for both spatially structured and unstructured heterogeneity in one model, which was later called the convolution model by Mollié (1996).

The detection of clusters in diseases is, at first sight, a separate problem. Here the goal is to identify clusters of geographically contiguous regions with elevated (or lowered) risk. Disease clusters may occur not only for infectious diseases, but also for non-infectious diseases, where risk factors do have a spatial structure. In addition one might also be interested in detecting discontinuities in the map, i.e. suspicious differences in relative risk between adjacent regions. However, results from the convolution model are often used to visually identify disease clusters, if the estimated risks exhibit a spatial pattern (e.g. Besag et al. 1991, Mollié 1996). In these cases, the Markov random field (MRF) term, which represents spatially structured heterogeneity, is dominating and the SMRs are essentially spatially smoothed. For that reason, Clayton & Bernardinelli (1992) denote the MRF term the “clustering component”.

This paper describes a new approach for the detection of clusters in disease maps. Technically, the method is based on reversible jump MCMC methodology (Green 1995) and is related to the segmentation of a spatial signal, already tackled in Green. His work has been refined by Arjas & Heikkinen (1997) and Heikkinen & Arjas (1998) who use piecewise constant step functions defined through marked point processes in continuous space. However, in our application space is discrete and irregular, which calls for several changes of the model and the methodology. Basically our prior model assumes that the area considered can be divided into several clusters, i.e. sets of contiguous regions, where each cluster has constant relative risk. The number, the size and the location of the clusters, as well as the risk within each cluster, are unknown. Risks in different clusters are assumed to be independent of each other. The model is therefore able to detect spatial discontinuities. Clusters of size one are not excluded from our model which implies that the model does necessarily smooth the SMRs. In practice it will always do so, at least to some extend, since there will always be some uncertainty whether a region forms a cluster by itself. However, the sizes of the clusters, which imply the local degree of smoothing, are variable and determined by the data, hence the smoothing is *adaptive*. This is in sharp contrast to MRF priors, where the corresponding smoothing parameter is constant and smoothing is *non-adaptive*.

The method is related to that of Schlattmann & Böhning (1993), who use mixture models within an empirical Bayes framework where each region is assigned to a component of the mixture distribution with constant relative risk. The location of the regions is, however, ignored so that members of a mixture component may be spread over the whole area. In our approach, regions are assigned to clusters with constant risk, too, but all regions in a cluster must be linked. To include location in the model we propose a construction where some regions are marked as so-called cluster centers, each of them defining a cluster. Each of the remaining regions is assigned to the cluster whose cluster center has minimal distance to the region. The distance between two regions is defined as the minimal number of boundaries that have to be crossed to move from one to the other. The construction can be seen as a modification of Voronoi tessellations (see Green 1995) in discrete, irregular space and ensures that all regions within a cluster are linked.

The output of the algorithm is very rich and can be used for Bayesian inference in several ways. First, the point estimates (mean or median) of the risk of each region incorporate all the posterior uncertainty about the number, the location and the risk level of the clusters. Since all these are variable, the posterior mean estimate will be an average over a large number of piecewise constant step functions and can be seen as essentially nonparametric (Arjas 1996, Heikkinen & Arjas 1998). A similar argument holds for all other functionals of the posterior as well, for example for the posterior median. Second, the method provides a large amount of additional probabilistic information. For example, we can calculate the probability that two or more regions belong to the same cluster. This is especially interesting for two adjacent regions where it gives an intuitive quantification for the location of discontinuities as will be illustrated in our application.

The paper is organized as follows. Section 3.2 describes our model and gives some features of the implementation by reversible jump MCMC. More details of the sampler are given in the Appendix. Section 3.3 presents results from an analysis of oral cavity cancer mortality rates of males in Germany, shown in Figure 3.1. We investigate the location of the clusters and discontinuities which have been identified by our method. We also compare our estimates with those obtained by the method of Besag et al. (1991). We close with several comments on alternative model specifications and possible extensions in Section 3.4.

3.2 The model

Suppose that data are available in the form of pairs in each of a set of n regions $i = 1, \dots, n$ giving the number of cases y_i of the disease and the number of expected cases e_i, usually calculated by internal or external standardization with respect to confounding variables.

The general idea is that the relative risk is *constant* over a set of one or more contiguous regions. This defines a cluster $C_j \subset \{1, \dots, n\}$, a set of contiguous regions with constant relative risk θ_j. The number of clusters k is treated as unknown with $k \in \{1, \dots, n\}$. Our cluster definition implies that the clusters $C_1, \dots, C_k$ cover the whole area and that they do not overlap, so $C_1 \cup \dots \cup C_k = \{1, \dots, n\}$. Note that in the limiting case $k = 1$ there is constant relative risk over the whole area whereas for $k = n$ not even two (contiguous) regions have the same relative risk.

We postulate the usual Poisson observation model (e.g. Clayton & Bernardinelli 1992), where y_i has Poisson distribution with mean $e_i\theta_j$ and θ_j is the unknown relative risk in cluster C_j with $i \in C_j$. Responses y_i, $i = 1, \dots, n$, are assumed to be conditionally independent given $\boldsymbol{\theta}_k = (\theta_1, \dots, \theta_k)$ so the likelihood function of responses $\boldsymbol{y} = (y_1, \dots, y_n)$ can be written as

$$p(\boldsymbol{y}|\boldsymbol{\theta}_k) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i\theta_j)^{y_i}}{y_i!} \exp(-e_i\theta_j). \tag{3.1}$$

3.2.1 A prior model for clustering

As a first step in the definition of the clustering model, we mark k regions $g_1, \dots, g_k$ as *cluster centers*. Each cluster center $g_j \in \{1, \dots, n\}$ defines a cluster C_j with $g_j \in C_j$. The vector of all cluster centers $\boldsymbol{g}_k = (g_1, \dots, g_k)$ defines a *cluster configuration*, i.e. an assignment of all regions to one and only one of the clusters. For that purpose, we define a measure of distance $d(i_1, i_2)$ between two regions i_1 and i_2 as the minimal number of boundaries that have to be crossed for moving from i_1 to i_2. This distance measure can be computed from the information if any two regions are adjacent or not, which is usually given in a so-called adjacency matrix. The measure of distance d is used to assign each of the remaining $n - k$ regions to one of the clusters. Region $i \notin \boldsymbol{g}_k$ will be assigned to cluster C_j if it has minimal distance to the corresponding cluster center g_j, i.e. $d(i, g_j) \leq d(i, g_l)$ for all $l \in \{1, \dots, k\}$, $l \neq j$. However, this definition is not yet unique, because some regions may have the same distance to two or more cluster centers. To ensure uniqueness we assign those regions to the cluster with the smallest index position of the corresponding cluster center in $\boldsymbol{g}_k$ among all cluster centers with minimal distance to region i. We therefore keep $\boldsymbol{g}_k$ non-ordered, otherwise clusters defined by cluster centers g_i with g_i small would tend to be larger in size than those with g_i large. For example, in our formulation a cluster configuration defined by a cluster center vector $\boldsymbol{g}_2 = (1, 2)$ will in general be different from another one defined by $\tilde{\boldsymbol{g}}_2 = (2, 1)$. Note, that the cluster centers only serve to specify a cluster configuration, they do not have any direct influence on the estimates of the relative risks.

To illustrate the flexibility of the clustering model, Figure 3.2 gives a cluster configuration of the 544 districts of Germany with $k = 20$. The cluster centers are marked with numbers 1 to 20, the corresponding index positions in $\boldsymbol{g}_k$. Note that the clusters differ considerably in size and shape. Furthermore it can be seen that, indeed, all regions within each cluster are linked. It is, however, not immediately obvious that this is true in general. Now suppose there is a cluster C_j which breaks down into two or more parts, which are not linked together. Then there must be a region $i_1 \in C_j$ with some distance m to the cluster center g_j and a neighbor i_2 of i_1 with $i_2 \in C_k, k \neq j$, and distance $m - 1$ to g_j. Otherwise all regions within C_j must be connected. Because i_1 is a neighbor of i_2 it follows, however, that i_1 must be in C_k and not in C_j which is contradictory to the assumption above and proves our claim.

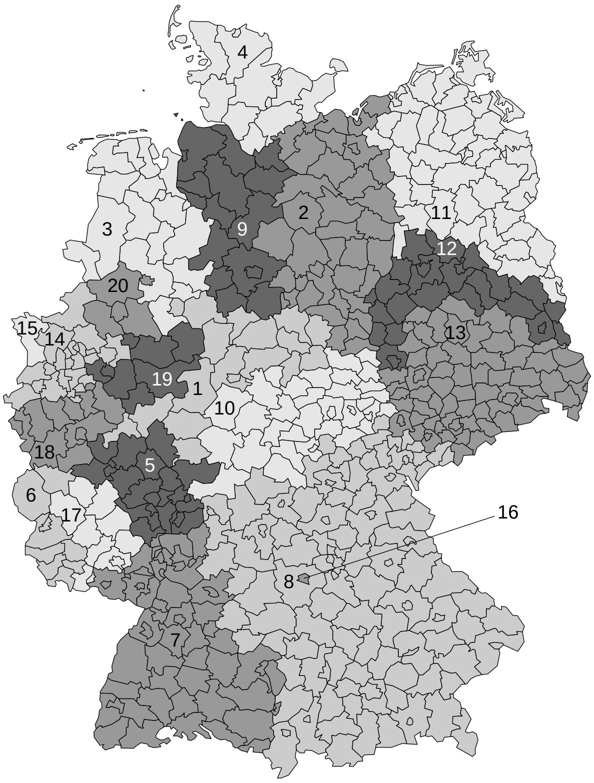


Figure 3.2: A cluster configuration for Germany with $k = 20$.

We now specify a prior distribution for the number of clusters k, the vector of cluster centers $\mathbf{g}_k$, and for the vector of relative risks $\boldsymbol{\theta}_k$. We assume that the prior for the number of clusters $p(k)$, $k = 1, \ldots, n$, is proportional to $(1 - c)^k$ with a fixed parameter $c \in [0, 1)$. The limiting case $c = 0$ gives a uniform distribution on $\{1, \ldots, n\}$, whereas $c > 0$ corresponds to a truncated geometric distribution. This choice implies that the prior ratio $p(k + 1)/p(k) = (1 - c)$, which penalizes jumps from k to $k + 1$, is constant for all k. We typically use small values for c so as to make the prior $p(k)$ close to "uninformative". Other choices might be more appropriate but, as Richardson & Green (1997) have noted, results with any prior for k can be converted to those corresponding to other priors without rerunning the algorithm.

For a given number of clusters k we assume that each vector of cluster centers $\mathbf{g}_k = (g_1, \ldots, g_k)$ has equal probability

$$p(\mathbf{g}_k|k) = \frac{(n-k)!}{n!}. \tag{3.2}$$

One could also introduce weights that take account of specific features so as to support configurations with homogeneous cluster sizes or boundary lengths, for example.

We have made extensive simulations from the prior distribution $p(\boldsymbol{g}_k|k) \cdot p(k)$ described above. For example, for each region, we have calculated the average size of the cluster the region is assigned to. Figure 4.10 (p. 81) shows the results for $c = 0.02$, grouping the regions according to the number of adjacent regions. The influence of the number of adjacent regions on the average size of the cluster appears to be minimal. Hence, the degree of smoothing is approximately the same for all regions, a priori. This is in contrast to MRF priors, where there is dependence of the smoothing parameter (the marginal variance) on the number of adjacent areas, see Bernardinelli, Clayton & Montomoli (1995a). We have also calculated the prior probability for each region to form a cluster by itself as well as the probability for being together with a neighbor. These probabilities have some variation, depending mainly on the number of neighbors. In Section 3.3, we therefore report the corresponding posterior probabilities together with the prior probabilities.

As a prior guess for the relative risks $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)$ it seems natural to assume that they are symmetrically distributed on the log scale. We therefore adopt a normal distribution for $\log(\theta_j)$, $j = 1, \ldots, k$, with unknown hyperparameters μ and σ^2. For μ, we assume a diffuse prior (uniform on the whole real line) and for σ^2 a highly dispersed but proper inverse gamma distribution $\text{IG}(a, b)$ with fixed parameters a and b. Independence of components of $\boldsymbol{\theta}_k$ yields

$$p(\boldsymbol{\theta}_k|k, \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \left(\prod_{j=1}^{k} \frac{1}{\theta_j}\right) \exp\left\{-\frac{1}{2\sigma^2} \sum_{j=1}^{k} (\log(\theta_j) - \mu)^2\right\}. \tag{3.3}$$

Conditional independence of $\boldsymbol{\theta}_k$ and $\boldsymbol{g}_k$ given k defines the prior for the unknown parameters k, $\boldsymbol{g}_k$, $\boldsymbol{\theta}_k$, μ and σ^2 as the product of the prior for k times (3.2) times (3.3) times the hyperpriors $p(\mu)$ and $p(\sigma^2)$.

3.2.2 Implementing reversible jump MCMC

This section gives an informal description of some features of our reversible jump MCMC implementation for sampling from the posterior distribution. In each iteration of the algorithm one of the following six moves is proposed:

Birth: The number of clusters is increased by introducing an additional cluster center.

Death: The number of clusters is decreased by deleting one of the cluster centers.

Shift: One of the cluster centers is moved.

Switch: The positions of two cluster centers in $\boldsymbol{g}_k$ are switched.

Height: The relative risks θ_j, $j = 1, \ldots, k$, are changed.

Hyper: The values of the hyperparameters μ and σ^2 are changed.

For a given value of k, each move is proposed with a certain probability. For some values of k certain moves are not possible, for example a death move for $k = 1$. Each move is accepted as the new state of the Markov chain with probability determined by the Metropolis-Hastings-Green ratio (Green 1995). Below we describe some features of our implementation of these six elementary moves. More details are given in the Appendix (p. 45). The main reason for choosing those moves was that they appeared to be straightforward to implement, each of them maintaining reversibility. We have included the shift and the switch move in the hope of improved mixing performance, although they seem to be not necessary. In fact, some other MCMC sampler with different proposals or different moves might be more efficient in terms of convergence, mixing or computing time but, in our experience, our algorithm gives reliable results for acceptable run lengths.

Suppose that in the current configuration k regions are marked as cluster centers. In a birth move one of the remaining $n - k$ regions is chosen randomly as a new cluster center. The new cluster center g^* is placed randomly among all possible $k + 1$ positions in the new vector of cluster centers $\boldsymbol{g}^*_{k+1}$. A value θ^* for the relative risk within the new cluster is inserted at the corresponding position in $\boldsymbol{\theta}^*_{k+1}$. In a death move from $k + 1$ to k, a randomly selected element of $\boldsymbol{g}_{k+1}$ is deleted. A sequence of a death and a birth move (or vice versa) is therefore able to restore the original configuration. In a shift move, first one of the cluster centers, whose neighbors are not all cluster centers by itself, is picked randomly. This cluster center g_j, say, is then shifted randomly to one of the neighbors that are not already cluster centers. The order in $\boldsymbol{g}_k$ is not changed. Note that the neighbors do not have to be members of the original cluster C_j which would in fact destroy the reversibility of the shift move. A switch move picks out two elements in $\boldsymbol{g}_k$ randomly and switches their position in $\boldsymbol{g}_k$ which will give a slightly different cluster configuration if there are distance ties. A height move proposes new values θ^*_j for all elements θ_j of $\boldsymbol{\theta}_k$, each of them being accepted or rejected separately. Finally, in a hyper move, values of the hyperparameters μ and σ^2 are updated by samples from the corresponding full conditional distributions.

The performance of the algorithm depends on a number of implementation issues. First, the several moves should be designed to have acceptance rates not too low. For moves that involve new values θ^*, we therefore use a proposal distribution that approximates the corresponding (fixed-dimension) "full conditional" (the prior for θ times the relevant likelihood times a normalizing constant). This device results generally in very good acceptance rates for these moves (height, birth, and—indirectly—death). Furthermore, the algorithm is now automatic, as tuning parameters are not involved. Similar proposals might be useful in many other applications of reversible jump MCMC. The shift move will have low acceptance rates, if there is very strong local information in the likelihood. Note, however, that this move is not necessary for convergence of the algorithm and could, in principle, be omitted completely.

A second problem occurs if the posterior is multimodal. This potential problem is inherent in any more complex MCMC application but seems to be of particular concern for reversible jump MCMC if only small dimension changing moves are made. If the simulated chain is trapped in one of the modes, it might be difficult for it to move to some other posterior mode, located somewhere different and clearly separated by an area of low posterior mass. This problem might be even more severe for fixed k, since the birth and death moves are known to improve mixing (Heikkinen & Arjas 1998). We routinely start several chains with different starting configurations and compare the results. Carefully designed mode jumping moves might also be useful here but require knowledge of the location of the posterior modes.

3.3 Applications

3.3.1 Simulations

To see, how well our method works, we have analyzed several artificial data sets. In particular, we have looked how well our method reconstructs a given risk surface, how sensitive our results are to the choices for $p(k)$ and $p(\sigma^2)$, and how reliable our algorithm works. The results are generally encouraging and can be found in a supplement paper (Knorr-Held & Raßer 1999). Based on these results, we recommend to use small, but positive values for c. Sensitivity with respect to $p(\sigma^2)$ was found to be small and we recommend to use $a = 1$ and $b = 0.01$ as default. Of course, sensitivity to the prior should always be studied.

3.3.2 Results for oral cavity cancer mortality in Germany

We now present results from an analysis of oral cavity cancer of males in Germany. The database records the population size and the number of deaths from oral cavity cancer, stratified by 16 age bands and 544 districts for the period 1986–1990. The total number of cases is 15,466 ranging between 1 and 501 cases with a median number of 19 cases per district. The overall mortality rate is 40.9 cases per 100,000 males. We have internally standardized the raw data with respect to all 16 age bands by maximum likelihood and have calculated the corresponding standard mortality ratios which are shown in Figure 3.1.

To examine sensitivity with respect to $p(k)$ and $p(\sigma^2)$ we have used $c = 0.0$, 0.01, and 0.02 and $(a, b) = (0.25, 0.00025)$, $(1, 0.01)$, and $(5, 0.125)$ in several combinations. For $(a, b) = (1, 0.01)$, for example, there was only slight sensitivity for k with respect to $p(k)$ with a posterior median of 45 ($c = 0.0$), 43 ($c = 0.01$), and 40 ($c = 0.02$) compared to a prior median of 272, 69, and 35, respectively. However, differences in the log relative risk estimates were found to be small. Results have been even more stable for different choices for $p(\sigma^2)$ with c fixed.

In the analysis presented here, we have set $a = 1$, $b = 0.01$, and $c = 0.02$. A plot of the prior and the posterior for k is given in Figure 3.3. The results are based on samples of 10,000 realizations, collected by saving the current state after every 10,000th basic update move after a burn-in period of 1,000,000. We have calculated autocorrelations for the corresponding relative risk samples in each region. Mixing was good with a median autocorrelation of only 0.025 for lag 1 and a maximum value of 0.51. For lag 5 the values have been below 0.1 for nearly all 544 parameters. The samples of k are shown in Figure 3.4. The acceptance rates were around 24%

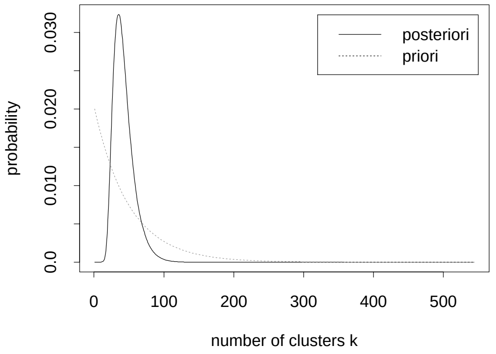


Figure 3.3: Prior and posterior distribution for the number of clusters k.

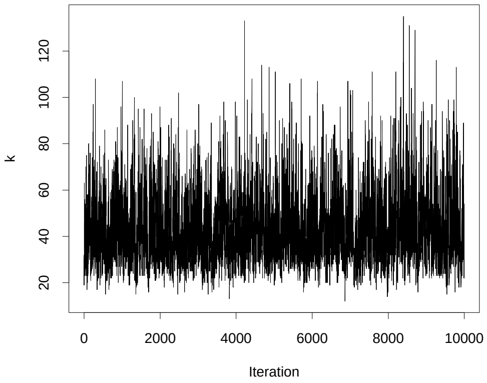


Figure 3.4: Chain for k versus iteration number.

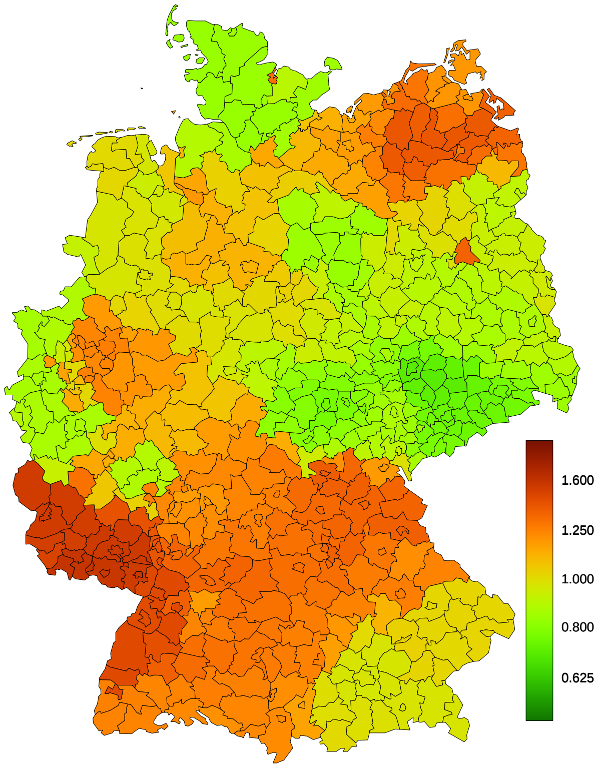


Figure 3.5: Estimated median relative risks for oral cavity cancer of males in Germany using our reversible jump MCMC algorithm.

for both the birth and the death move, 21% for a shift, 41% for a switch and 98% for a change of height.

The posterior median estimates of the relative risk vary between 0.65 and 1.42. Figure 3.5 displays those estimates on the same scale as in Figure 3.1. Most striking are three large clusters of elevated relative risk above 1.2, one in the north-east in Mecklenburg-West Pomerania, one in the south-west covering the whole Saarland and parts of Rhineland-Palatinate and Baden-Württemberg along the border to France, and the third in Franconia, the northern part of Bavaria. The latter two seem to be linked and in fact, most parts of southern Germany, excluding southern Bavaria, have an elevated relative risk above 1.0.

The most important risk factors for oropharyngeal cancers are tobacco smoking and alcohol abuse (Blot, Devesa, McLaughlin & Fraumeni 1994). The Mecklenburg-West Pomerania cluster is consistent with this, because this state has the highest per capita alcohol consumption of whole Germany (Becker & Wahrendorf 1997). Interestingly, Blot et al. (1994) note that the east-central part of France (Bas-Rhin) along the German border has the highest oral and pharyngeal incidence rate in whole Europe (1983–1987). The south-west cluster is exactly adjoining this area and might therefore continue on the other side of the border.

There are several single regions with conspicuously high risk estimates, compared to their neighbors, in particular West Berlin (estimated relative risk of 1.22) and Kiel in the very north (1.13). We have calculated the probability that each of them forms a cluster by itself. The probabilities are 0.09 for West Berlin and 0.45 for Kiel. For comparison, the median probability of all 544 regions is only 0.001. The *prior* probabilities of being alone for these regions are 0.006 and 0.03, respectively, compared to a median prior probability of 0.014. This indicates the existence of unobserved risk factors for these regions, possibly related to a higher degree of urbanization.

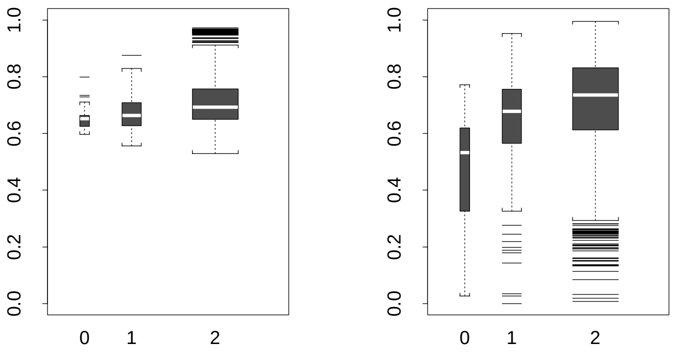


Figure 3.6: Boxplots of the prior (left) and posterior (right) probabilities that adjacent districts are within the same cluster for 0: former east-west border, 1: all other boundaries between different states, 2: boundaries within states. The width of the boxes is proportional to the number of observations.

An interesting feature of Figure 3.5 is that the map strongly retains the border between former East and West Germany, especially in the south but also for West Berlin. We have therefore calculated the probability that two regions belong to the same cluster for all 1,416 pairs of adjacent regions. Figure 3.6 compares the distribution of these probabilities for the former east-west border with all remaining ones by boxplots. To avoid a "state border" bias we have stratified the latter group in two subgroups where adjacent regions do or do not belong to the same state, respectively. Figure 3.6 gives also the corresponding plot for the *prior* probabilities. Differences between these subgroups are minimal *a priori*, however, the *posterior* probabilities are lower for the former east-west border. This indicates substantial differences between East and West Germany, either in exposure to relevant risk factors or simply in data quality. There are several hints that the latter is an important factor (Becker & Wahrendorf 1997). One reason for the apparent differences might be a lack of quality control measures in the former Democratic Republic of Germany in the process of identifying underlying diseases. For example, it might be possible that there is an underreporting of oral cavity cancer due to a relatively high rate of nonidentified cancers "of other and unspecified sites". However, it seems that noncompliance with WHO rules for the identification of underlying disease is not able to explain the differences alone. Other possible reasons are discussed in detail in Becker & Wahrendorf (1997) with relevant references.

Figure 3.7 displays the estimated median relative risks of this data set by the method of

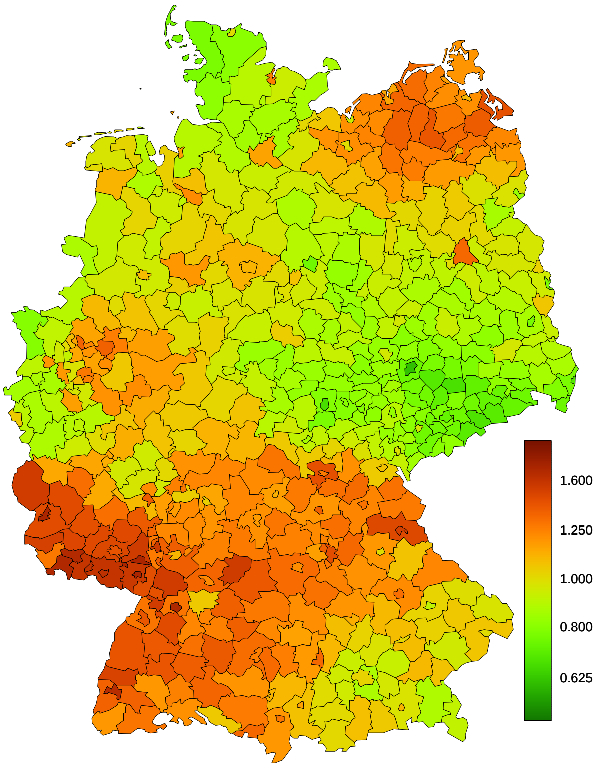


Figure 3.7: Estimated median relative risks for oral cavity cancer of males in Germany with the method of Besag et al.

Besag et al. (1991) with a Gaussian intrinsic prior for the spatial component. The estimates show slightly more variation with values between 0.56 and 1.56. The similarities between Figure 3.5 and Figure 3.7 are noticeable and relieving, although there are some apparent differences. In particular, Figure 3.7 seems to be noisier. This becomes evident from Figure 3.8, which displays the absolute difference in estimated log relative risk between adjacent regions. Overall, the median absolute difference using the Besag et al. model (0.067) is nearly four times as high as with our method (0.018). It seems that the risk variability in some parts of the map induces a considerable overall variability, because smoothing in the convolution model is non-adaptive. Our method, however, is adaptive and therefore the distribution of the absolute differences is much more skewed. An even more pronounced difference can be seen in Figure 3.8 for absolute differences between regions, where one of the regions has only one or two neighbors. Since the prior marginal variance of the MRF term is considerably larger, smoothing is much less pronounced here and the differences are very large. This can also be seen from Figure 3.7, where regions with only one or two neighbors are often in a different risk category than their neighbors. In contrast, our method, where the amount of smoothing is approximately the same for all regions a priori (see Figure 4.10), does not see much evidence for such large absolute differences, apart from the Kiel cluster. We have also tried a median-based prior instead of the Gaussian in the Besag et al. model, which gave, however, very similar results.

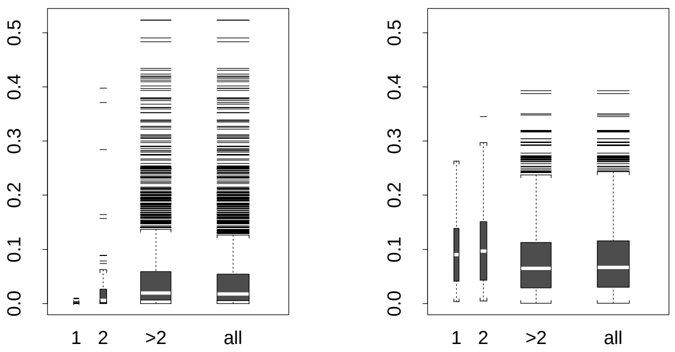


Figure 3.8: Boxplots of the absolute difference in log relative risk between adjacent districts. Left panel: Our method. Right panel: Method of Besag et al. The minimum of the number of neighbors of the two adjacent districts is used for grouping (1, 2, > 2).

3.4 Concluding remarks

We have described a novel approach to disease mapping with particular emphasis on the detection of clusters and discontinuities in disease maps. We close now with a few comments on alternative model specifications and possible extensions.

Initially, we have considered a more general cluster model, where every possible partition into k clusters has equal probability a priori, as long as all regions within each cluster are linked. However, if k is treated as unknown, we need to know the number of all possible partitions, say n_k, because this number determines the prior probability $1/n_k$ of a specific partition. These probabilities enter in the prior ratio for any birth or death move. It was and still is far from obvious to us how to calculate the n_k's in irregular space. We have therefore decided to reduce the complexity of the problem by introducing cluster centers. Of course, our model has now the slightly odd feature that, for a given partition, it is difficult to derive its prior probability. But even if this probability is zero, the partition can well be approximated by an average over a set of different configurations, that are supported by our model.

Suppose now that we define a cluster configuration by selecting a few cluster centers and assigning each of the other regions to one of the clusters based on some distance measure, just as we did. One might argue that other measures of distance as the one we propose might be more appropriate. Indeed, initially, we thought of assigning a specific point to any region, for example the centroid of the region or the location of that region's largest place. The distance between regions could then be defined as the Euclidean distance between the corresponding points. However, such a definition turned out to be not very useful, because clusters will not necessarily be connected. It is in fact rather easy to construct counterexamples, where regions, belonging to a specific cluster, are separated by regions which belong to other clusters. We therefore prefer our distance measure which ensures that clusters are connected and which does fully acknowledge the discrete nature of space.

More generally, our method might be useful for other statistical problems in discrete space.

Furthermore, it can be viewed as a module in Bayesian inference for more complex data. For example, in the current context it might be desirable to include covariate information more explicitly in the model. For categorical covariates, one could introduce an additional partition model of unknown dimension (Green 1995) for the effects of the covariate levels.

Our approach might also be useful in modeling disease risk data in time and space. Such data have been analyzed recently by Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi & Songini (1995b), Waller, Carlin, Xia & Gelfand (1997) Knorr-Held & Besag (1998). Suppose, that data (y_{it}, e_{it}) are available for n regions $1, \dots, n$ and T time points $t = 1, \dots, T$, say years. The obvious extension of our approach would be to define the neighbors of pixel (i, t) as the neighbors in space (all pixels (j, t) where region j is a neighbor of region i) and the neighbors in time (pixels $(i, t-1)$ and $(i, t+1)$ with obvious modifications for the endpoints $t = 1$ and $t = T$). Clusters of constant risk would then be defined over time and space. In particular, such a specification would be able to capture space-time interactions.

Acknowledgements

This research was supported by the German Science Foundation (DFG), SFB 386, and by the European Science Foundation Program on Highly Structured Stochastic Systems (HSSS). The authors express thanks to Nikolaus Becker for access to the data set and for the computer program for producing the map of Germany, to Iris Pigeot for help on an earlier version of this manuscript, and to the associate editor and three referees for very helpful and encouraging comments.

Appendix: Details of the sampler

Suppose a cluster configuration with k clusters is defined by a vector of cluster centers $\mathbf{g}_k = (g_1, \dots, g_k)$ and a vector of relative risks $\boldsymbol{\theta}_k = (\theta_1, \dots, \theta_k)$. In each step of the algorithm one of the six moves birth, death, shift, switch, height, and hyper is proposed with probability $r_B(k)$, $r_D(k)$, $r_{Sh}(k)$, $r_{Sw}(k)$, $r_{He}(k)$, and $r_{Hy}(k)$, respectively. These probabilities have been chosen as $r_B(k) = r_D(k) = 0.4$ and $r_{Sh}(k) = r_{Sw}(k) = r_{He}(k) = r_{Hy}(k) = 0.05$ for $k \in \{2, \dots, n-1\}$ with appropriate changes for the endpoint cases.

The six moves are now implemented as follows:

1. Birth: A uniformly distributed random variable on all $n - k$ regions, which are not cluster centers, determines the new cluster center g^*. A second uniformly distributed random variable j on $\{1, \dots, k+1\}$ determines the position of g^* in $\mathbf{g}^*_{k+1}$. A value θ^* is generated and inserted into $\boldsymbol{\theta}^*_{k+1}$ at the corresponding position. The proposal $\theta^* = \theta^*_j$ is drawn from a gamma distribution

$$\theta^*_j \sim \mathrm{G}\left(y_j + \frac{\tilde{\mu}^2}{\tilde{\sigma}^2}, e_j + \frac{\tilde{\mu}}{\tilde{\sigma}^2}\right), \qquad (3.4)$$

where $e_j = \sum_{i \in C_j^*} e_i$, $y_j = \sum_{i \in C_j^*} y_i$, $\tilde{\mu} = \exp(\mu + 0.5\sigma^2)$ and $\tilde{\sigma}^2 = \exp(\sigma^2) \cdot (\exp(\sigma^2) - 1) \cdot \exp(2\mu)$. This proposal distribution is an approximation of the (normalized) "full conditional" $\prod_{i \in C_j^*} p(y_i|\theta_j) \cdot p(\theta_j)$, where the log-normal prior $p(\theta_j)$ is replaced by a gamma distribution $G(\tilde{\mu}^2/\tilde{\sigma}^2, \tilde{\mu}/\tilde{\sigma}^2)$ with the same mean and variance. The birth step is accepted with probability $\alpha = \min\{1, \mathcal{A} \cdot \mathcal{P} \cdot \mathcal{L} \cdot \mathcal{J}\}$, where $\mathcal{A} = p(k+1)/p(k) \cdot p(\theta^*)/(n-k)$ is the prior ratio, $\mathcal{P} = r_D(k+1)/r_B(k) \cdot (n-k)/q(\theta^*)$ is the proposal ratio, $\mathcal{L}$ is the likelihood ratio and $\mathcal{J} = 1$ is the Jacobian. Here $q(\theta^*)$ denotes the density of the proposal distribution (3.4), evaluated at θ^*.

2. Death: For a death move from $k+1$ to k a uniformly distributed random variable j on $\{1, \ldots, k+1\}$ is generated which determines the cluster center g_j and the corresponding relative risk θ_j which are then removed from $\boldsymbol{g}_{k+1}$ and $\boldsymbol{\theta}_{k+1}$ respectively. The acceptance probability for the death move has the same form as for the corresponding birth move with all ratio terms inverted.

3. Shift: Among the k current cluster centers there are $n(\boldsymbol{g}_k)$ cluster centers which do not only have cluster centers as neighbors. An uniformly distributed random variable j on $\{1, \ldots, n(\boldsymbol{g}_k)\}$ determines a cluster center g_j with $m(g_j)$ "free" neighbors. A second uniformly distributed random variable on $\{1, \ldots, m(g_j)\}$ determines the new cluster center g_j^* which replaces g_j in $\boldsymbol{g}_k$. The shift step is accepted with probability $\alpha = \min\{1, \mathcal{L} \cdot n(\boldsymbol{g}_k)/n(\boldsymbol{g}_k^*) \cdot m(g_j)/m(g_j^*)\}$.

4. Switch: For a switch move two random variables i and j, uniformly distributed on $\{1, \ldots, k\}$ with $i \neq j$, are generated. The positions i and j of the corresponding cluster centers g_i and g_j in $\boldsymbol{g}_k$ are now switched. Only the likelihood ratio $\mathcal{L}$ enters in the acceptance probability for the switch move.

5. Height: For each cluster $j = \{1, \ldots, k\}$ a new value θ_j^* is proposed from (3.4) and eventually accepted or rejected separately. The acceptance probability is
$\alpha = \min\{1, \mathcal{L} \cdot p(\theta_j^*)/p(\theta_j) \cdot q(\theta_j)/q(\theta_j^*)\}$.

6. Hyper: To change the values for μ and σ^2 we use two subsequent Gibbs steps (hence $\alpha = 1$), drawing random variables from the corresponding full conditionals

$$\mu|. \sim \mathrm{N}\left(\frac{1}{k}\sum_{j=1}^{k}\log(\theta_j),\, \frac{1}{k}\sigma^2\right) \quad \text{and} \quad \sigma^2|. \sim \mathrm{IG}\left(a + \frac{k}{2},\, b + \frac{1}{2}\sum_{j=1}^{k}\{\log(\theta_j) - \mu\}^2\right).$$

Note that for moves 1–5, the likelihood ratio $\mathcal{L}$ has to be evaluated only for those regions, where the relative risk has changed in the proposal. For example, in a birth move, $\mathcal{L}$ has to be evaluated only for the regions in the new cluster and in a death move only those regions enter in the likelihood ratio that are part of the cluster, which is supposed to be removed.

3.5 Models with covariates

So far we have considered the simplest case where count data is available aggregated within specific geographical regions and no further covariates are measured. Yet, the idea to use a spatial statistical model is based on the fact that most diseases develop as a consequence of exposure to certain risk factors and that this exposure shows a spatially structured pattern. If further covariate information is available, it is desirable to extend the model accordingly in order to reduce residual variation.

The assumption that the observed count y_i in region i is a realization of a Poisson distribution with parameter $e_i \lambda_i$ offers two options to incorporate covariate information: the number of expected cases e_i and the region-specific relative risk λ_i.

Using covariate information in the preprocessing step, i.e. in the calculation of the expected cases, is the indirect way as usually done with the age effect. Age is an important risk factor for most diseases and its effect is assumed to be the same within all regions to estimate the expected number of cases. The benefit is a lower number of parameters in the model. Other covariates could be included here as well but this has no advantage because the number of unknown parameters cannot be reduced anymore. Furthermore, estimating covariate effects within the Bayesian analysis is somehow more appealing and allows further inference on the covariates as well.

Therefore, we will choose the second option and adjust the relative risks. So far, possible influences of risk factors are assumed to be absorbed within the region-specific relative risk. In other words, the relative risk λ_i is a surrogate for covariate information not measured. Any statistical inference is carried out with respect to the spatial structure. Space can be seen as the only covariate included in the model. Hence, an intuitive way to use covariate information is the decomposition of the relative risks.

Basically, we may think of different kind of covariate information, depending on the number of covariates measured and their scale, metrical or categorical. We start with the easiest model formulation for one covariate c measured on the same geographical resolution as the observed counts, i.e. for all regions the observed levels $\boldsymbol{c} = (c_1, \ldots, c_n)$ of the covariate are given. We include covariate information by decomposing the relative risk parameter λ_i in a multiplicative way

$$\lambda_i = \lambda_i^s \cdot \lambda_i^c, \quad i = 1, \ldots, n, \tag{3.5}$$

where λ_i^s is the spatial effect and λ_i^c the covariate effect for region i. In general, this model is not identifiable and restrictions have to be imposed on at least one of the effects. In a more general setting with p covariates we may decompose the relative risk parameter to

$$\lambda_i = \exp(\eta_i), \quad i = 1, \ldots, n,$$

with

$$\eta_i = \gamma_0 + \gamma_{i1} + \ldots + \gamma_{ip}. \tag{3.6}$$

Depending on the prior specifications of the covariate effects $\gamma_1, \ldots, \gamma_p$, this is a generalized linear model or generalized additive model. In this general form, the intercept γ_0 allows a rough adaptation to the data and further identifiability restrictions must be imposed on all other parameters. We might use exact as well as stochastic restrictions here; the definition of which, however, depends on the chosen prior model for the covariate. Decompositions (3.5) and (3.6) are quite general and make no further assumptions neither on the type of the covariates, metrical or categorical, nor on prior specifications. For aggregated count data, like in the disease mapping example, equation (3.6) defines a Poisson regression model, with an offset e_i and a linear predictor η_i with p covariates.

For the moment, we will concentrate on the special case with $p = 2$ parameters, space γ_i^s and covariate γ_i^c

$$\eta_i = \gamma_0 + \gamma_i^s + \gamma_i^c, \quad i = 1, \ldots, n.$$

For identifiability, we impose a stochastic restriction on the covariate effect and keep a more flexible formulation for the spatial effect. In accordance to the previous sections we rewrite the model to

$$\lambda_i^s = \exp(\gamma_0 + \gamma_i^s) \quad \text{and} \quad \lambda_i^c = \exp(\gamma_i^c), \tag{3.7}$$

which corresponds to the multiplicative decomposition (3.5). The likelihood from the Poisson model is given by

$$p(\boldsymbol{y}|\boldsymbol{\lambda}^s, \boldsymbol{\lambda}^c) = \prod_{i=1}^{n} \frac{(e_i \lambda_i^s \lambda_i^c)^{y_i}}{y_i!} \exp(-e_i \lambda_i^s \lambda_i^c). \tag{3.8}$$

This general expression on region level can be simplified within the algorithm. Sampling with covariate information requires at least one additional move for each covariate, eventually even more moves depending on the prior specification. For the spatial part $\boldsymbol{\lambda}^s$ we apply a CPM prior and consider the moves proposed in Section 3.2.2. For fixed covariate effects $\boldsymbol{\lambda}^c$ we may rewrite (3.8) to

$$p(\boldsymbol{y}|\theta_k^s) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(\tilde{e}_i \theta_j^s)^{y_i}}{y_i!} \exp(-\tilde{e}_i \theta_j^s), \tag{3.9}$$

where $\tilde{e}_i = e_i \lambda_i^c$ is the covariate-corrected expected number of cases. This is exactly the likelihood derived for the spatial model conditional on the covariate effects. Therefore, all moves for the spatial part are retained unchanged after recalculating the expected numbers of cases $\tilde{e}_i$. All formulas derived earlier hold true with this minor change. In a similar way the likelihood can be rewritten for fixed spatial effects when updating the covariate effects.

Suppose, we have additional information on one covariate in m categories. For identifiability reasons the number of categories should be clearly below the number of regions. In this work, we use covariates measured on a nominal scale and estimate the effect of each category separately. Yet, for ordered categories it might be useful to assume some kind of smooth effect over the categories, or even impose a restriction on the order of the covariate effects. Within a hierarchical Bayesian framework we could even take a further step and carry out inference on the categories of the covariate. If two or more categories have the same effect, they could be combined to one category. This defines a partition model for the categorical covariate, where the order of the categories can be preserved or not. Three different model formulations, with (1) exchangeable categories, (2) arbitrary partitions, and (3) order-preserving partitions (i.e. CPM) are described in Giudici, Knorr-Held & Raßer (2000) and will not be referred here in detail.

As a prior assumption for metrical covariates, one might assume a linear effect $\gamma_i^c = \beta c_i$ or some other functional relationship $\gamma_i^c = f(c_i)$. Moreover, a wide variety of (non-)parametric specifications are possible, e.g. either a simple random walk of first or second order, or more sophisticated approaches like Bayesian P-splines (Lang & Brezger 2003).

The notation so far is quite general. For any particular data situation, model specification and choice of prior distributions have to be considered carefully. Here, we give details for the special case of one covariate. The basic algorithm is extended accordingly and all additional moves are described in detail.

3.5.1 Model specifications

Suppose one covariate measured on a nominal scale with m categories. We decompose the relative risk according to equations (3.5) and (3.7). For the spatial part $\boldsymbol{\lambda}^s$ we use a CPM prior with risk parameters $\boldsymbol{\theta}_k^s$ as described before. Covariate effects are assumed to be exchangeable and no further restrictions are implied. For the covariate effects $\boldsymbol{\theta}^c = (\theta_1^c, \dots, \theta_m^c)$ we choose log-normal priors

$$\theta_j^c \sim \mathrm{LN}(0, \tau^2), \quad j = 1, \dots, m, \tag{3.10}$$

with

$$\lambda_i^c = \theta_j^c \quad \text{for } c_i = j,\ i = 1, \dots, n.$$

Under exchangeability, the categories are assumed to be independent and the joint prior density has the form

$$p(\boldsymbol{\theta}^c|\tau^2) = \left(\frac{1}{\sqrt{2\pi}\tau}\right)^m \left(\prod_{j=1}^m \frac{1}{\theta_j^c}\right) \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^m \left(\log\left(\theta_j^c\right)\right)^2\right\}.$$

While the location parameter in (3.10) is fixed to zero for identifiability reasons, we apply an additional hyperprior

$$\tau^2 \sim \mathrm{IG}(c, d)$$

for the dispersion parameter τ^2, again with fixed parameters c and d. Assuming further prior independence of the covariate and the spatial effects, the joint prior distribution can be factorized to

$$p(\boldsymbol{g}_k, k, \boldsymbol{\theta}_k^s, \boldsymbol{\theta}^c, \mu, \sigma^2, \tau^2) = p(k)p(\boldsymbol{g}_k|k)p(\boldsymbol{\theta}_k^s|k, \sigma^2, \mu)p(\boldsymbol{\theta}^c|\tau^2)p(\mu)p(\sigma^2)p(\tau^2),$$

where the non-covariate priors $p(k)$, $p(\boldsymbol{g}_k|k)$, $p(\boldsymbol{\theta}_k^s|k, \sigma^2, \mu)$, $p(\mu)$ and $p(\sigma^2)$ are chosen as in Section 3.2.1.

To update the covariate effects $\boldsymbol{\theta}^c$ within the algorithm, we introduce an additional move, where for each category j a proposal θ_j^* is drawn from a gamma distribution

$$\theta_j^* \sim \mathrm{G}\left(y_j + \frac{\tilde{\nu}^2}{\tilde{\tau}^2}, \tilde{e}_j + \frac{\tilde{\nu}}{\tilde{\tau}^2}\right), \quad j = 1, \ldots, m. \tag{3.11}$$

Here $y_j = \sum_{i:c_i=j} y_i$ is the total number of observed cases in all regions i with covariate effect in the jth category. Similarly, $\tilde{e}_j = \sum_{i:c_i=j} e_i \lambda_i^s$ is the corresponding number of expected cases, but now corrected for the spatial effects $\boldsymbol{\lambda}^s$. The parameters $\tilde{\nu} = \exp(0.5\tau^2)$ and $\tilde{\tau}^2 = \exp(\tau^2)(\exp(\tau^2) - 1)$ enter through the gamma-approximation of the log-normal prior as before. The acceptance probability for this move is

$$\alpha = \min\left\{1,\, \mathcal{L} \cdot \frac{p(\theta_j^*)}{p(\theta_j^c)} \cdot \frac{q(\theta_j^c)}{q(\theta_j^*)}\right\},$$

where $\mathcal{L}$ denotes the likelihood ratio and q is the density of the proposal distribution (3.11). A Gibbs sampler step for the dispersion parameter completes the algorithm. New values are drawn from the full conditional

$$\tau^2 | . \sim \mathrm{IG}\left(c + \frac{m}{2},\, d + \frac{1}{2}\sum_{j=1}^{m}\left(\log\left(\theta_j^c\right)\right)^2\right). \tag{3.12}$$

Suppose now the observed covariate values $\boldsymbol{c} = (c_1, \ldots, c_n)$ are measured on a metrical scale. We decompose the relative risk as before and apply a CPM prior for the spatial part. For the covariate we assume a linear effect on the log scale

$$\lambda_i^c = \exp(\beta c_i), \quad i = 1, \ldots, n.$$

For identifiability, we center the covariate values

$$\sum_{i=1}^{n} c_i = 0.$$

Therefore, no further restriction on the coefficient β is necessary and we use a diffuse prior

$$p(\beta) \propto \text{constant}. \tag{3.13}$$

To update the coefficient β, we apply a Metropolis-Hastings step with Gaussian random walk proposal, i.e. a new value β^* is generated by

$$\beta^* = \beta + \epsilon, \quad \text{with} \quad \epsilon \sim \mathrm{N}(0, \kappa^2) \tag{3.14}$$

with a fixed tuning parameter κ^2. This proposal is accepted with probability $\alpha = \min\{1,\, \mathcal{L}\}$, due to the uniform prior (3.13) and the symmetric proposal (3.14).

3.5.2 Comparison to previous results

To investigate the capability of the covariate-corrected models, we present two re-analyses of the oral cavity cancer data from Section 3.3.2. As mentioned before the posterior median estimates in this application visually reflect the former border between East and West Germany and so do the posterior probabilities for pairs of border regions to be in the same cluster (see Figures 3.5 and 3.6). This result provokes the assumption that these differences may be due to some non-observed state-effect and can be included in the model by introducing a state indicator variable. We therefore define a covariate c with two non-ordered categories, namely

$$c_i = \begin{cases} 1 & \text{if region } i \text{ belongs to West Germany (incl. West Berlin)} \\ 0 & \text{if region } i \text{ belongs to East Germany} \end{cases} , \quad i = 1, \ldots, n.$$

This covariate carries a strong spatial structure itself. There are 328 districts in West Germany and 216 in East Germany which are clearly separated, except West Berlin which is located in the middle of East Germany.

More interesting and meaningful is the inclusion of covariate information that covers known risk factors like tobacco consumption. Unfortunately, such information is not available in the data set. We therefore use information on lung cancer mortality as a surrogate. This is reasonable since tobacco consumption is the (only) major risk factor for lung cancer. The covariate values are defined to be the log relative risks for lung cancer, estimated by the method of Clayton & Kaldor (1987) and centered around zero (see also Natário & Knorr-Held 2003). These log relative risk estimates are displayed in Figure 3.9 showing lower rates in the south of Germany and very high rates in the west.

To assess the influence of the additional covariates on the relative risk estimates and the performance of the algorithm, we choose a parameter setting as similar as possible to the basic model. To simplify notation, we will refer to the purely spatial model from Section 3.2 as model 1, to the model with state-indicator as model 2, and to the model with smoking behavior covariate as model 3. Note that all results for model 1 presented here were gained by a rerun of the algorithm. This is due to the fact that in the original run—presented in Section 3.2—not all of the information now needed was collected. Although differences to the original run were barely noticeable, there might be some small inconsistencies to previous results.

All results were gained by runs with the same run length, burn-in, and sample size as before. We start with some details on the prior specifications for the analyses presented below. For the spatial component $\boldsymbol{\lambda}^s$, we use exactly the same choice as before: a geometric prior for the number of clusters k with constant $c = 0.02$, a diffuse prior for the mean μ, and an inverse gamma prior for the variance σ^2 with constants $(a, b) = (1, 0.01)$. There has to be remarked that with additional covariate information other choices for the prior specification of the spatial component may also be reasonable.

We now specify our prior assumption for the covariate part $\boldsymbol{\lambda}^c$ in model 2, assuring identifiability of both components. In addition to the stochastic restriction implicit in (3.10) by fixing the location parameter to zero, we choose a more informative prior for the variance τ^2. The

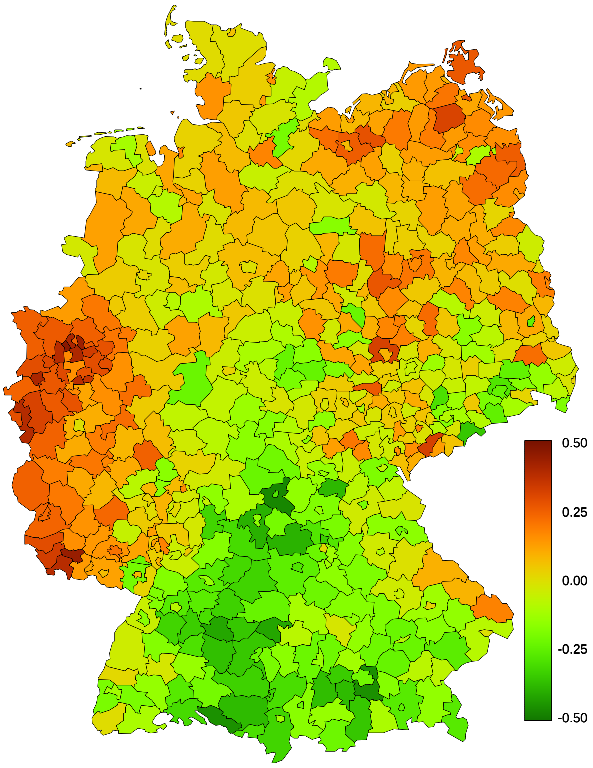


Figure 3.9: Spatial distribution of the surrogate for tobacco consumption.

main idea for the choice of parameters $(a, b) = (1, 0.01)$ for the variance σ^2 of the log spatial effects is to keep the prior uninformative in the sense that these values yield little influence on the corresponding parameters in the full conditional. By chosing the parameters $(c, d) = (5, 0.05)$ we increase the prior information in the full conditional (3.12).

In model 3, we only need to specify the tuning parameter κ^2 of the proposal distribution. This was set to $\kappa = 0.1$, determined by acceptance rates in pre-runs of the algorithm. In the run presented below, the acceptance rate was 43%.

Regarding the sampling scheme, we retain the previous choice, adding one additional move for the covariate effects. The proposal probabilities for the moves have been left unchanged and whenever a height move is proposed for the spatial effects $\boldsymbol{\theta}^s_k$, a change of the covariate effects $\boldsymbol{\theta}^c$ is performed subsequently, but accepted or rejected separately. Finally, the variance τ^2 of the covariate in model 2 is updated within the hyper move.

In model 2, for all moves regarding the spatial component acceptance rates were even higher than before, 36% for both dimension changing moves, 29% for a shift, 48% for a switch, and 98% for a change of height. Extremely high rates were gained for a change of height for the covariate component with over 99%. In model 3, all acceptance rates were about the same as in model 1.

Comparing the results, we first take a look at the model fit using the deviance information criteria (DIC) introduced by Spiegelhalter, Best, Carlin & van der Linde (2002). Opinions on

Model	$\bar{D}$	p_D	DIC
model 1	628	117	744
model 2	592	129	721
model 3	644	94	738

Table 3.1: Mean deviance, effective number of parameters, and DIC for models with and without covariate.

the DIC are quite controversial (see the discussions to the paper). From a theoretical point of view, it can be applied to problems of variable dimensions, but its suitability for such models is not tested in the statistical literature so far. Still, we will use it to get a first impression of the performance of both models; lower values indicating that a model is more appropriate. In Table 3.1 the mean posterior deviance $\bar{D}$ and the effective number of parameters p_D are reported, the sum of which is the DIC value DIC $= \bar{D} + p_D$. By comparing model 1 and model 2 we observe that the deviance is lower for model 2 which testifies a better fit to the data. Regarding the DIC, there is also a preference for model 2, although the effective number of parameters is higher than for model 1. At first sight, it is not surprising that the model with covariates has a higher number of parameters and therefore a higher model complexity. Interestingly, these additional parameters are not generated by the covariate effect but arise mostly from the spatial component as will be shown below. The partition model uses a higher number of clusters to reconstruct the spatial structure. For model 3 the results are opposite. While the model fit is slightly worse than for model 1, the number of parameters is decreased considerably. This indicates that the covariate explains a large amount of the spatial variation.

We now take a detailed look on the results of model 2 with emphasis on the comparison to model 1. Of clear interest in the discussion is the border between East and West Germany, with special focus on West Berlin.

First, the estimated covariate effects validate a much higher risk in West Germany than in East Germany, already conjecturable in the data. The posterior median estimates are 1.18 for West Germany and 0.85 for East Germany, while the corresponding means of the SMRs are 1.06 and 0.85, respectively. Looking at the ratios, the estimates suggest that the relative risk in West Germany is 1.4 times as high as in East Germany conditional on the spatial effect. According to the SMRs this ratio would only be 1.25 and thus the estimates are more extreme. This fact supports the suspicion of substantial differences in data quality between East and West Germany. Note that the sum of the log covariate effects is 0.003 which reflects the stochastic restriction on the prior distribution (3.10) with mean zero on the log-scale.

Figure 3.10 displays the posterior median estimates of the relative risks for model 2. Overall, the results show a similar spatial distribution as the corresponding results for model 1 in Figure 3.5. In accordance with the strong covariate effects, the spatial effects in Figure 3.11 show a smoother pattern than before. The clusters of elevated risk are still clearly emphasized, but the edge along the East-West border is no longer visible. All conclusions and interpretations of

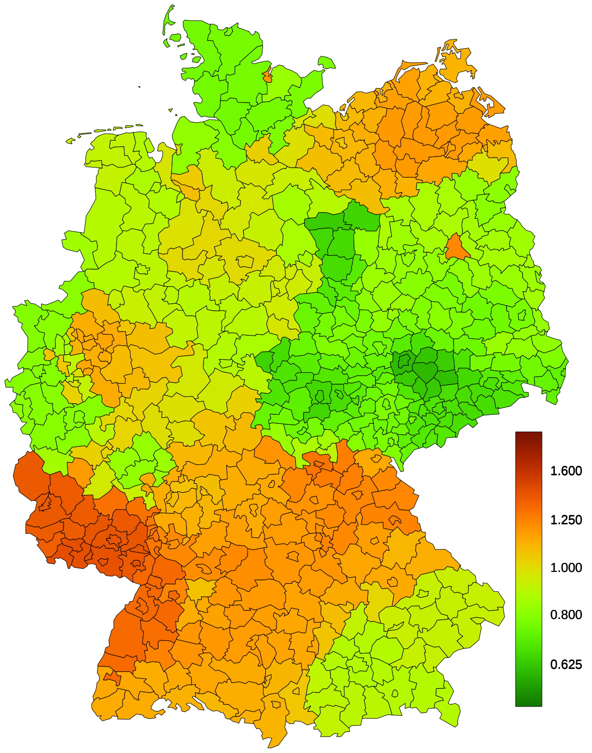


Figure 3.10: Posterior median estimates of the relative risks for model 2.

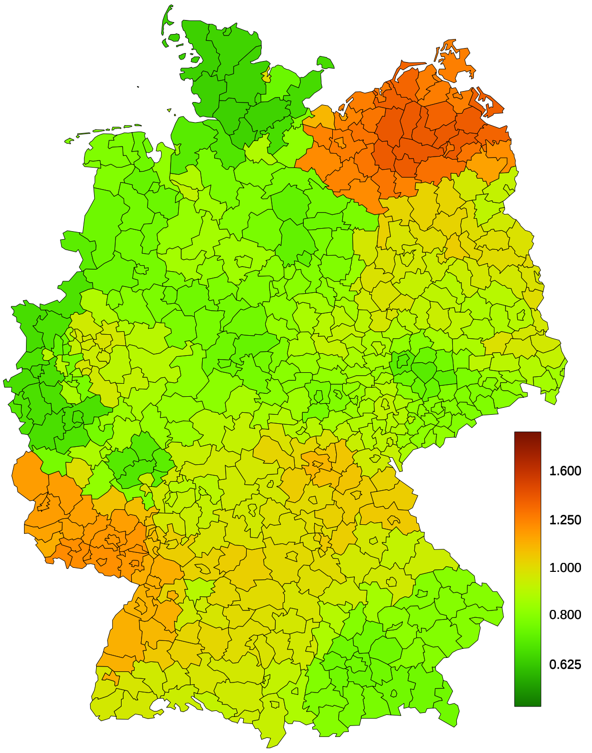


Figure 3.11: Posterior median estimates of the spatial component of model 2.

the previous analysis are still valid and we concentrate on algorithmic and statistical details in our comparison.

The smoother risk surface in the spatial component is partly due to the higher number of clusters used by the partition model. As can be seen from Figure 3.12, the posterior distribution for k has shifted to higher values with a median number of 66 clusters for model 2 compared to a median number of 41 for model 1. This allows for a more detailed reconstruction of the risk surface and yields a better adaptation to the data as indicated by the deviance.

Altogether, estimates of the risk parameters $\boldsymbol{\lambda}$ are very similar with a mean absolute difference in the log relative risks of only 0.026, but a maximum of 0.24. Taking a closer look, the differences become more obvious. Figure 3.13 shows some regions where the estimates are rather different, although for most regions differences are moderate.

Not surprisingly, the regions with the largest differences are located near the border of East Germany to West Germany or West Berlin. Furthermore, these are generally sparsely populated regions with only few expected cases. For example, there are 22 regions with an absolute difference in the log relative risk above 0.1 which corresponds to a decrease or increase of approximately 10% (or more) in the relative risk. The median number of expected cases for these regions is about 13 compared to an overall median number of nearly 20. For these regions there is little information in the likelihood and the estimates are dominated by the

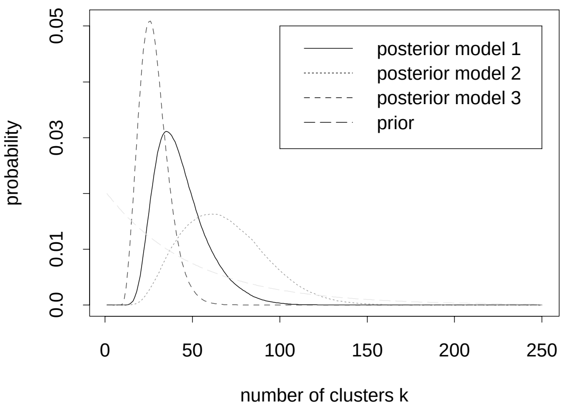


Figure 3.12: Posterior distributions of the number of clusters k for model 1, 2, and 3.

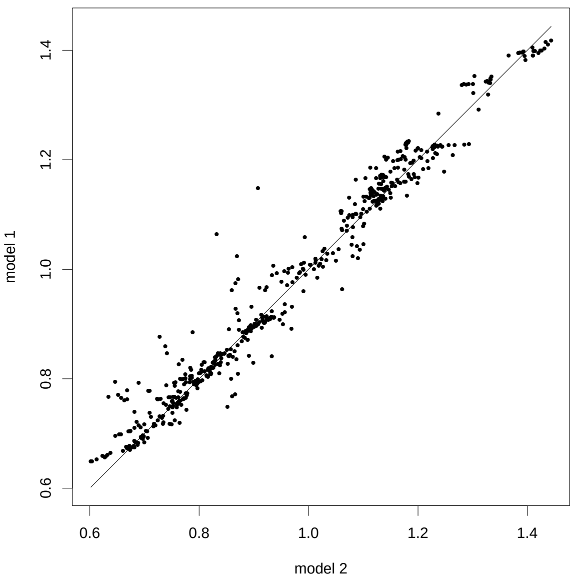


Figure 3.13: Comparison of the posterior median estimates of relative risks for model 1 and model 2.

prior information. For further investigation, consider the regions located north of West Berlin. While the results from model 1 in Figure 3.5 slightly suggest a bridge between West Berlin and the cluster of elevated risk in Mecklenburg-West Pomerania, this effect vanishes for the results of model 2, displayed in Figure 3.10. The posterior distribution for most of these regions is bimodal in model 1. Autocorrelations in the samples of the risk parameters are not high but persistent on a low level even for lag 10 or higher. As an example, consider Oranienburg, the district adjacent to West Berlin in the north. For model 1, the left column in Figure 3.14 displays corresponding plots. In model 2 (middle column), the posterior is unimodal and almost perfect mixing with extremely low autocorrelations is observed. The drawback becomes obvious in the right column. Although sampling is good for the spatial component, the uncertainty is increased and the density is much more dispersed, due to the weak stochastic restriction. Yet, this has influence on the estimates of confidence regions mainly, not on posterior median point estimates.

The discussion of the results for model 3 focuses on the influence of the covariate on the spatial component. Overall, the posterior median risk estimates from model 3 in Figure 3.15 display a similar spatial pattern as in model 1.

The absolute differences in the log relative risks are rather small, with a median of 0.036 and a maximum of 0.217. Whereas for model 2 differences appear mainly in districts located along the East-West border, for model 3 differences are spread over the whole area. Locally, the risk surface is slightly rougher with larger differences for adjacent regions than in Figure 3.5 from model 1.

The posterior for the coefficient β ranges from 0.10 to 0.87 and has a posterior median

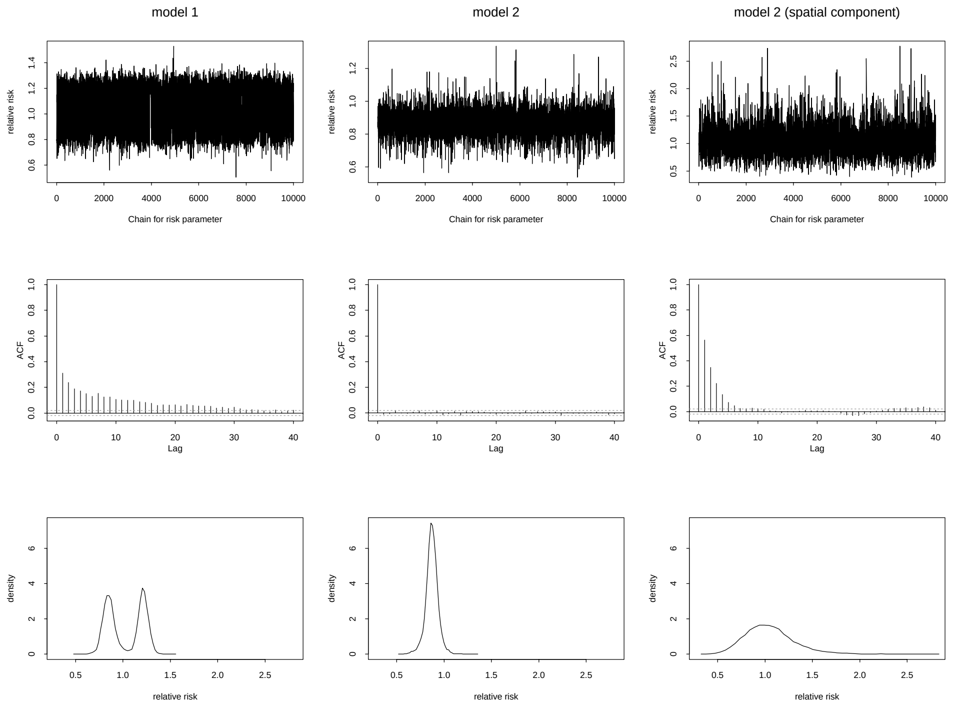
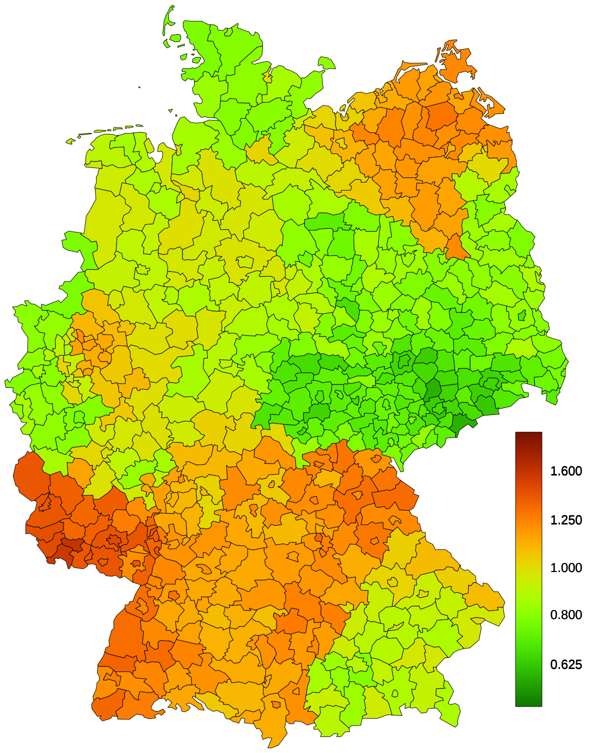


Figure 3.14: Sampling paths, autocorrelations, and kernel density estimates of the posterior for Oranienburg. Left: model 1, middle: model 2, right: spatial component of model 2.



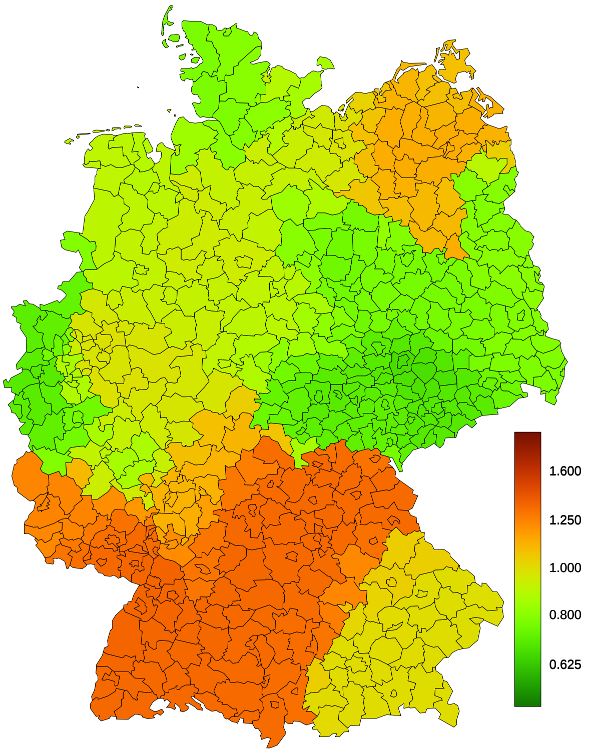


Figure 3.15: Posterior median estimates of the relative risks for model 3.

Figure 3.16: Posterior median estimates of the spatial component of model 3.

of 0.49. Therefore, a positive influence of the covariate is significant; higher tobacco consumption yields increased oral cavity cancer risk. Moreover, the covariate explains part of the spatial variation and the spatial risk surface is changed considerably by its inclusion. The posterior distribution of k in Figure 3.12 is much less dispersed with a median number of only 27 clusters. This provokes a locally smooth risk surface for the spatial component with sudden changes, see Figure 3.16. The cluster of elevated risk in Mecklenburg-West Pomerania is still visible but less accentuated. The main attention is now drawn to the south of Germany, namely Baden-Wuerttemberg, Saarland, and the north of Bavaria, which is still consistent with the results from model 1.

3.5.3 Discussion

To incorporate covariate information into the basic spatial model can be done in an intuitive way. Sampling of the covariate component does not interfere with the spatial component and the basic sampling scheme is left unchanged.

In both examples the influence of the covariate becomes apparent in the spatial component. The number of clusters is changed noticeable. Significant changes in the risk surface reduce the number of probable partitions to those preserving these edges. A smooth risk surface without dramatic changes is supported by much more partitions even likely.

In model 2, the edge in the risk surface along the former East-West border is absorbed by the covariate effect. The spatial component does not detect any dramatic differences in the residual variation anymore. The partition model concentrates on a detailed reconstruction of smaller changes. This leads not only to a higher number of clusters. One side effect are higher acceptance rates for the dimension changing moves. With an increasing number of clusters, the average cluster size decreases. Therefore, less regions are affected by the birth and death moves. Hence, the changes for the state of the Markov chain get smaller.

In model 3, smaller changes in the risk surface are explained by the covariate. Primarily, the spatial component assembles a rough pattern to compensate larger differences. For the oral cavity cancer data this pattern consists only of few larger areas, and only few clusters are needed.

Furthermore, the results from model 2 suggest that additional covariate information can affect the performance of the algorithm. Here, mixing of the risk parameters $\boldsymbol{\lambda}$ is improved while uncertainty about the specific components $\boldsymbol{\lambda}^s$ and $\boldsymbol{\lambda}^c$ is increased due to the weak stochastic restriction used for identifiability. This leads to lower autocorrelations of the risk parameter and higher autocorrelations of the spatial component. In Figure 3.17 the autocorrelations of the risk parameters for model 1 and model 2 are displayed. These autocorrelations are significantly lower for model 2. The mean autocorrelations are 0.045 (model 1) and 0.024 (model 2) for lag 1 and 0.013 (model 1) and 0.004 (model 2) for lag 5. In addition, the autocorrelations of the spatial component in model 2 are shown; these are considerably higher. As a remedy for this one could use effect coding for the categories, assuring an exact sum to zero restriction of the effects.

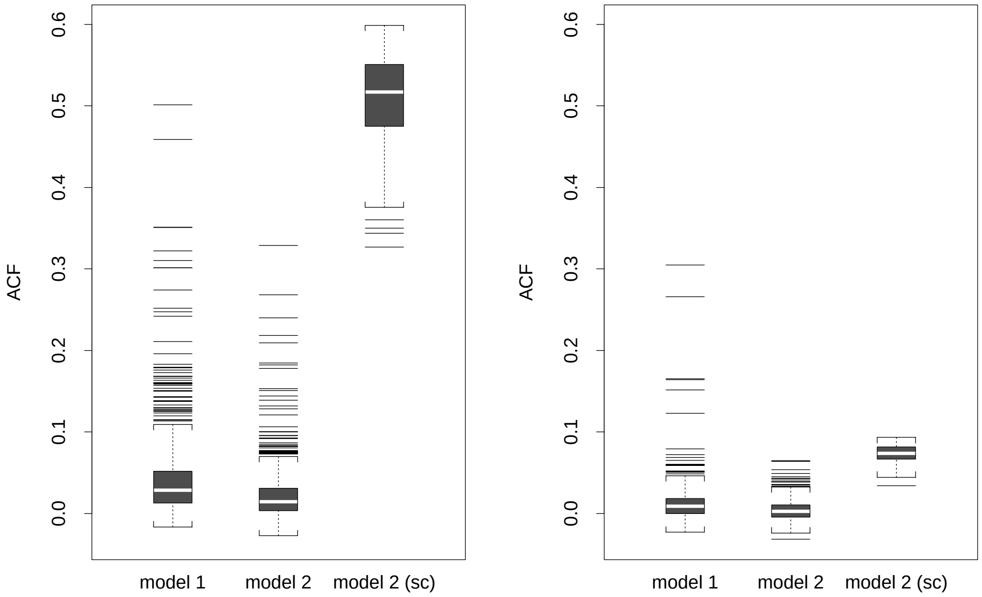


Figure 3.17: Autocorrelations of the risk parameters (model 1, model 2) and of the spatial component (model 2 (sc)) for lag 1 (left) and lag 5 (right).

3.6 Sampling with marginal likelihood

In this section, we consider again the model without covariates and propose an alternative prior specification. The decision to use a normal prior for the log relative risks was founded on the idea that the relative risks should be symmetric on the log-scale, a priori. Therefore, a relative risk θ has the same prior "probability" as the inverse relative risk $1/\theta$, i.e. $P(\theta \leq x) = P(1/\theta \leq x)$ for $x > 0$, which is an appealing and natural choice.

The proposal distribution (3.4) is chosen to be an approximation of the full conditional assuming a gamma prior for the relative risks. Alternatively, one could apply a gamma prior for the relative risks, which is the conjugate prior distribution to the Poisson observation model (see Bernardo & Smith 1994). In a continuous space setting, Denison & Holmes (2001) have proposed a gamma-Poisson model. The advantage is that a Gibbs sampler, based on marginal likelihood quantities, can be constructed. This idea also works for our discrete model.

Suppose a partition into k clusters $\mathcal{C}_k = \{C_1, \ldots, C_k\}$ with corresponding risk parameters $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)$. We replace the log-normal prior (3.3) with independent gamma priors with (fixed) parameters α and β for $\theta_j,\ j = 1, \ldots, k$. The joint prior density is the product of k gamma densities $p(\theta_j|\alpha, \beta)$

$$p(\boldsymbol{\theta}_k|k, \alpha, \beta) = \prod_{j=1}^{k} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} \exp(-\beta\theta_j).$$

Conjugacy allows to integrate over the unknown relative risk parameters. Thus, the marginal likelihood can be derived

$$\begin{aligned} p(\boldsymbol{y}|g_k, k, \alpha, \beta) &= \int p(\boldsymbol{y}|\boldsymbol{\theta}_k, g_k, k) p(\boldsymbol{\theta}_k|k, \alpha, \beta) d\boldsymbol{\theta}_k \\ &= \int \ldots \int p(\boldsymbol{y}|\boldsymbol{\theta}_k, g_k, k) \prod_{j=1}^{k} p(\theta_j|\alpha, \beta) d\theta_1 \ldots d\theta_k \\ &= \prod_{j=1}^{k} \int p(\boldsymbol{y}_j|\theta_j, g_k, k) p(\theta_j|\alpha, \beta) d\theta_j, \end{aligned} \tag{3.15}$$

where $p(\boldsymbol{y}_j|\theta_j, g_k, k)$ is the contribution of cluster C_j to the likelihood (3.1). The marginal likelihood $p(\boldsymbol{y}_j|g_k, k, \alpha, \beta)$ for cluster C_j is

$$\begin{aligned} p(\boldsymbol{y}_j|g_k, k, \alpha, \beta) &= \int p(\boldsymbol{y}_j|\theta_j, g_k, k) p(\theta_j|\alpha, \beta) d\theta_j \\ &= \int \left\{ \prod_{i \in C_j} \frac{(e_i\theta_j)^{y_i}}{y_i!} \exp(-e_i\theta_j) \right\} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha-1} \exp(-\beta\theta_j) d\theta_j \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \prod_{i \in C_j} \frac{e_i^{y_i}}{y_i!} \int \exp\left[-\left(\beta + \sum_{i \in C_j} e_i\right)\theta_j \right] \theta_j^{(\alpha + \sum_{i \in C_j} y_i)-1} d\theta_j \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \sum_{i \in C_j} y_i)}{(\beta + \sum_{i \in C_j} e_i)^{\alpha + \sum_{i \in C_j} y_i}} \prod_{i \in C_j} \frac{e_i^{y_i}}{y_i!}. \end{aligned} \tag{3.16}$$

According to (3.15) and (3.16) the marginal likelihood has product form

$$p(\boldsymbol{y}|g_k, k, \alpha, \beta) = \prod_{j=1}^{k} \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha + \sum_{i \in C_j} y_i)}{(\beta + \sum_{i \in C_j} e_i)^{\alpha + \sum_{i \in C_j} y_i}} \prod_{i \in C_j} \frac{e_i^{y_i}}{y_i!}.$$

The MCMC algorithm can now be constructed to sample the cluster configuration and the risk parameters separately. Sampling of the cluster configuration is based on marginal likelihood quantities solely, independent of $\boldsymbol{\theta}_k$. Relative risk parameters are drawn from the full conditional

$$\theta_j|. \sim \mathrm{G}\left(\alpha + \sum_{i \in C_j} y_i\,,\ \beta + \sum_{i \in C_j} e_i\right), \quad j = 1, \ldots, k,$$

given the cluster configuration $\mathcal{C}_k$. This is a Gibbs sampler step and therefore the acceptance probability is 1. However, the use of a gamma-approximation to the log-normal prior as implemented in our algorithm compensates this fact pretty good with very high acceptance rates that are close to 1. Both prior distributions, gamma and log-normal, can look quite similar. This depends on the choice of the hyperparameters. With appropriately chosen hyperpriors the posterior distributions of both models will be rather the same.

Analyzing the oral cavity cancer data with the marginal likelihood sampler yields very similar results to those reported before. The results stated below were gained with $\alpha = 21.4$ and $\beta = 22.1$. This choice approximately matches the first two moments of the gamma and the log-normal prior, i.e. the gamma prior has the same mean and variance as the log-normal prior, using the posterior median estimates of μ and σ^2 from model 1.

All acceptance rates were slightly higher than before, 31% for both the birth and the death move, 28% for a shift, and 46% for a switch. From Figure 3.18 it becomes obvious that the posterior median estimates of the relative risks are almost the same for both models. The mean absolute difference of the log relative risks is only 0.003 with a maximum of 0.043.

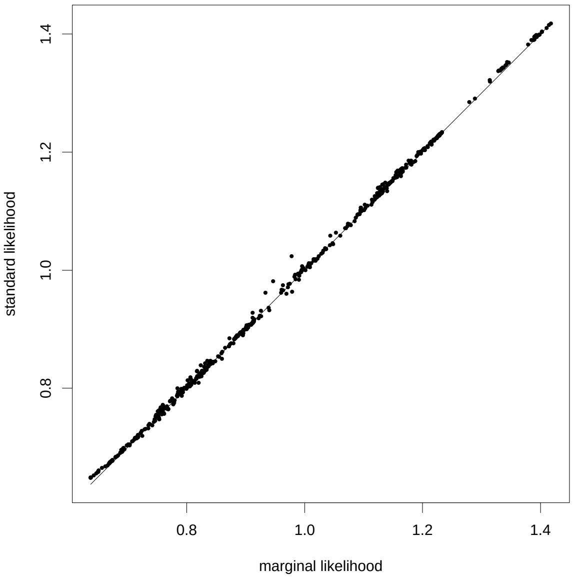


Figure 3.18: Comparison of the posterior median estimates of relative risks for the standard model (model 1) and the model based on the marginal likelihood.

The posterior distribution of the number of clusters k is very similar to model 1 with a median number of 44. Almost identical is the model fit, reported in Table 3.1 for model 1. The mean deviance of $\bar{D} = 622$ together with the effective number of parameters $p_D = 120$ gives a DIC value of 742.

Thus, with appropriately chosen hyperparameters α and β the results with both priors are rather the same. The choice of a log-normal prior is advantageous since the sampling of the hyperparameters μ and σ^2 of the relative risks is straightforward using Gibbs sampling steps. Such an easy implementation cannot be derived for the parameters α and β of the gamma prior, although it is possible to sample hyperparameters within the algorithm (Denison & Holmes 2001).

The advantage of the gamma prior approach is the Gibbs sampling of the risk parameters. Sampling the cluster configuration without risk parameters saves some computation time. Risk parameters have to be sampled only for those iterations stored for later use and not for all iterations. Note that this holds true only as long as the hyperparameters are fixed. However, sampling of the risk parameters in our original algorithm is fast since parameters are assumed independent between clusters. Thus, the computational cost is moderate.

The major drawback of the marginal likelihood approach is its limitation to one risk parameter. Consider a multiplicative decomposition of the relative risk to incorporate covariate information in the model as described in the previous section. For such a decomposition of the risk parameter, marginalization is no longer possible. Therefore, we prefer the non-conjugate log-normal prior for disease mapping applications.

Chapter 4

Further Topics in Clustering Partition Models

In Chapter 2 we have defined the CPM prior and derived some theoretical properties. In this chapter we focus primarily on practical issues. The results gained with the CPM prior are encouraging for the disease mapping data. For wider applications it is of interest if the model is useful for other data types as well. Therefore, in Section 4.1 the model is transferred to a Gaussian observation model. Here, we focus on data from image analysis. The aim is to restore some unknown true image, distorted by Gaussian white noise. Prior specifications and results are reported for simulated data sets as well as for real data from human brain mapping.

In Section 4.2, the smoothing properties of the CPM are investigated. As mentioned in the previous chapter, the CPM prior allows for adaptive smoothing with regard to the data. We take a closer look at the smoothing performance of the CPM with emphasis on the characterization of the local and global smoothing behavior. Furthermore, a comparison is given to GMRF models, for which smoothing is non-adaptive, at least in the commonly used form.

It will be shown that the smoothing behavior of the CPM is determined by the properties of the cluster configurations, mainly by the cluster sizes. Unfortunately, the corresponding prior properties cannot be derived analytically. Therefore, in Section 4.3 some simulation results from the CPM prior are reported for the graphs of all applications considered so far.

The chapter closes with some comments on computational issues of RJMCMC samplers in general and the proposed CPM sampler in particular.

4.1 Image processing

For aggregated count data, like in the disease mapping example, the sample size for each region is equal to the corresponding population size. Only for rare diseases and sparsely populated regions spatial statistical models are used to improve ML estimates of the relative risk parameters. This allows to overcome poor statistical properties of the SMRs. For sufficiently large sample sizes, e.g. for more frequent diseases and densely populated areas, such complex models are not necessary.

With decreasing sample size the need for spatial models is more urgent. There are various applications with very low sample sizes, the most prominent one probably is image analysis. Here, an array of pixels is considered with only one observation for each pixel, usually the color on some grey or color scale. Smoothing and restoration of noisy images is a vast area of statistical effort. With only one observation per pixel, estimation must be based on either rather informative statistical models or knowledge on the noise generating process. Due to the lack of information in the data, estimates are most commonly based on spatial statistical models. Such models assume some sort of similarity in nearby pixels, e.g. similar colors. Bayesian models are very popular in this area of research because such similarity assumptions are easy to incorporate into the model as prior information.

For rectangular lattice data with, say, n_1 rows and n_2 columns, pixel-labeling is usually in terms of pairs (s,t), corresponding to row s and column t of the lattice. The advantage of this notation is that one can easily see, if two pixels are adjacent or not. Like in the disease mapping context, there are (usually) no covariates available. Thus, a general model formulation can be written as

$$y_{st} = f(s,t) + \epsilon_{st}\,, \quad 1 \leq s \leq n_1\,,\, 1 \leq t \leq n_2,$$

where f denotes some unknown function varying over the lattice and ϵ is an independent spatial noise process with expectation zero (see e.g. Winkler 1995). This model is also used by Polzehl & Spokoiny (2000) who fit a piecewise constant function f in a nonparametric approach related to the Bayesian model described below.

In this section, we will apply the CPM to image analysis data. Still, there has to be remarked that essentially the model might need modifications for practical use in image analysis, especially if the number of pixels is large. We will focus on spatial applications on two-dimensional regular grids. Model formulations for other structures are essentially the same and reduce to the use of an appropriate distance measure.

4.1.1 Model formulation and prior specifications

Suppose we are given observations y_i on an array of pixels $i = 1, \ldots, n$. Instead of identifying the pixels by pairs (s,t), we use the easier labeling $y_i = f(i) + \epsilon_i$, $i = 1, \ldots, n$. The neighborhood structure of the pixels is provided separately and enters the model via the distance measure d as defined in Section 2.1.2. We assume a Gaussian observation model for y_i with spatially varying mean function $f(i) = \lambda_i$ and overall variance τ^2, i.e.

$$y_i \sim \mathrm{N}(\lambda_i, \tau^2), \quad i = 1, \ldots, n.$$

This corresponds to some unknown true quantity λ_i superposed by additive Gaussian noise ϵ_i with mean zero and variance τ^2. The pointwise maximum likelihood estimates for the unknown means are the observed values $\hat{\lambda}_i = y_i$, $i = 1, \ldots, n$, and spatial modeling is crucial to achieve separation of noise and signal.

We apply a CPM prior for the unknown mean function. For a given partition with k clusters $C_1, \dots, C_k$ and corresponding parameters $\theta_1, \dots, \theta_k$ we assume

$$\lambda_i = \theta_j\,, \quad \text{for } i \in C_j\,,\ j = 1, \dots, k.$$

Thus, the likelihood can be written as

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}_k, \tau^2) &= \prod_{j=1}^{k} \prod_{i \in C_j} \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}(y_i - \theta_j)^2\right\} \\ &= \left(\frac{1}{\sqrt{2\pi}\tau}\right)^n \exp\left\{-\frac{1}{2\tau^2}\sum_{j=1}^{k}\sum_{i \in C_j}(y_i - \theta_j)^2\right\}. \end{aligned} \tag{4.1}$$

In contrast to the Poisson model there are two unknown parameters for each pixel, λ_i and τ_i^2. With only one observation per pixel the parameters λ_i and τ_i^2 are not identifiable. Hence, we restrict the variance to be constant $\tau_i^2 = \tau^2$ over the whole image. Alternatively, one could assume either a bivariate CPM with parameters $\theta = (\lambda, \tau^2)$ or a separate CPM for τ^2.

We further assume independent conjugate priors for the unknown parameters, i.e. a Gaussian prior for the mean function

$$\theta_j \sim \mathrm{N}(\mu, \sigma^2), \quad j = 1, \dots, k,$$

and an inverse gamma prior for the variance

$$\tau^2 \sim \mathrm{IG}(\alpha, \beta).$$

For both hyperparameters additional "uninformative" priors are applied as already done for the log-normal prior in the disease mapping example, i.e. a diffuse prior for μ and a highly dispersed inverse gamma prior for σ^2 with parameters (a, b).

Similar to the Poisson-gamma setting it is possible to derive the marginal likelihood and sample the partition and the parameters separately. Note that in this case a joint conjugate prior for $\boldsymbol{\theta}_k$ and τ^2 requires a special choice for the variances, $\sigma^2 = \tau^2 v$ (Denison, Holmes, Mallick & Smith 2002) with some $v > 0$. Thus, the parameters $\theta_j|\sigma^2 \sim \mathrm{N}(\mu, \sigma^2)$ and the variance τ^2 are not independent anymore.

4.1.2 Implementation

The basic sampling scheme is retained from Section 3.2.2. We focus on the necessary changes only. The likelihood ratio $\mathcal{L}$ is now calculated according to (4.1) for all moves. The moves shift and switch are implemented as before and the formulas for the acceptance probabilities still hold true. The other moves are adapted as follows:

Height: The Gaussian likelihood and prior allow for the construction of a Gibbs sampler step to update the means $\boldsymbol{\theta}_k$. For each cluster C_j a proposal is drawn from the full conditional

$$\theta_j^*|. \sim \mathrm{N}\left(\frac{\sigma^2 m_j \bar{y}_j + \tau^2 \mu}{\sigma^2 m_j + \tau^2}, \frac{\sigma^2 \tau^2}{\sigma^2 m_j + \tau^2}\right), \tag{4.2}$$

where m_j is the size of the cluster C_j and $\bar{y}_j$ is the mean of the corresponding observations. The derivation of this full conditional is standard algebra based on combinations of quadratic forms (Box & Tiao 1992).

Birth: The generation of the new cluster C^* is performed as before. A new parameter θ^* is drawn from (4.2) with corresponding values m^* and $\bar{y}^*$ of the new cluster. The acceptance probability calculates to

$$\alpha = \min\left\{1,\, \mathcal{L} \cdot \frac{p(k+1)}{p(k)} \cdot \frac{r_D(k+1)}{r_B(k)} \cdot \frac{p(\theta^*)}{q(\theta^*)}\right\}. \tag{4.3}$$

Here, $q(\theta^*)$ denotes the density of the proposal distribution (4.2), evaluated at θ^*.

Death: As before a randomly selected cluster center and the corresponding parameter are deleted while the parameters for the remaining clusters are left unchanged. Again, the acceptance probability has the same form as for a birth move with all ratio terms inverted.

Hyper: For the variance τ^2 and both hyperparameters μ and σ^2 Gibbs sampler steps are used, drawing new values from the full conditionals

$$\tau^2|. \quad \sim \quad \mathrm{IG}\Big(\alpha + \frac{n}{2},\, \beta + \frac{1}{2}\sum_{j=1}^{k}\sum_{i \in C_j}(y_i - \theta_j)^2\Big), \tag{4.4}$$

$$\mu|. \quad \sim \quad \mathrm{N}\Big(\frac{1}{k}\sum_{j=1}^{k}\theta_j,\, \frac{\sigma^2}{k}\Big), \tag{4.5}$$

$$\sigma^2|. \quad \sim \quad \mathrm{IG}\Big(a + \frac{k}{2},\, b + \frac{1}{2}\sum_{j=1}^{k}(\theta_j - \mu)^2\Big). \tag{4.6}$$

4.1.3 Results for simulated data sets

First, we will investigate the performance of the algorithm with simulated data sets, generated from two different mean functions. We consider a lattice with 20 rows and 20 columns, i.e. a total of $n = 400$ pixels. The true mean function f_1 is piecewise constant

$$f_1(s,t) = \begin{cases} 1, & 1 \le s \le 10, \quad 1 \le t \le 10, \\ 2, & 1 \le s \le 10, \quad 11 \le t \le 20, \\ -1, & 11 \le s \le 20, \quad 1 \le t \le 10, \\ 0, & 11 \le s \le 20, \quad 11 \le t \le 20. \end{cases} \tag{4.7}$$

The surface features two edges of height one and two, dividing the left and right half, and the upper and lower half of the square, respectively. This function was introduced by Ogata (1990) and later reused by Künsch (1994). The second mean function f_2 has the same range as f_1, from -1 to 2, but is defined as an inclined plane over the lattice

$$f_2(s,t) = \frac{3}{38}(t-s) + \frac{1}{2}. \tag{4.8}$$

The functions are displayed in Figure 4.1. The true functions are distorted with independent Gaussian noise, i.e. for each pixel (s,t) we generate a Gaussian random variable with mean $f(s,t)$ and variance 1. The goal is to restore the true mean function f from the data.

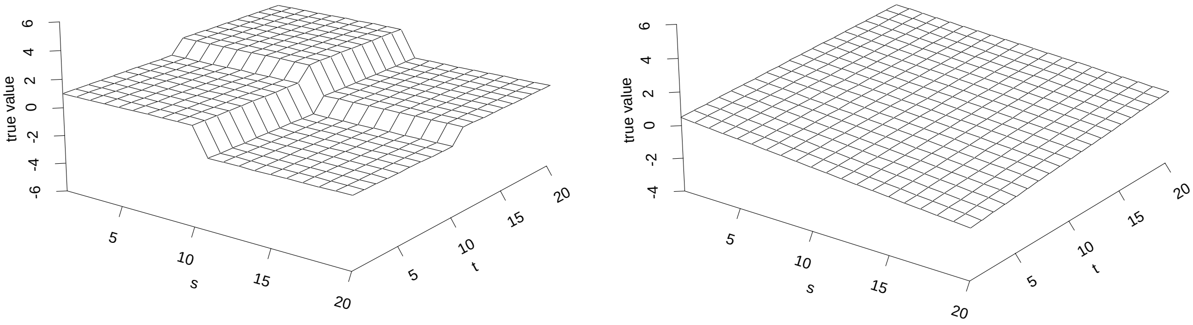


Figure 4.1: True mean functions f_1 (left) and f_2 (right).

We have simulated various replications of data sets and calculated the mean squared error (MSE) for each reconstruction

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\lambda}_i - \lambda_i)^2,$$

where $\hat{\lambda}_i$ is the posterior median estimate for λ_i. Throughout, the reconstruction of f_1 was very good. This is not surprising and one would expect a good reconstruction of a piecewise constant function by a piecewise constant partition model approach. In fact, the true function f_1 has several representations in the space of prior functions. In contrast, the smooth function f_2 can only be expressed by the prior for the extreme case of $k = n$ clusters. The results were also good but slightly reflect the prior assumption of a piecewise constant function.

Note that MRF models based on Gaussian pairwise difference priors are not suitable for data with only one observation in each pixel. As an example reconstructions for f_1 and f_2 are given in Appendix B. To achieve acceptable results, the Gaussian prior must be replaced by a more robust version, e.g. a truncated Gaussian (Künsch 1994).

For both functions we discuss only one reconstruction. Among all simulated data sets we have chosen those for which the reconstruction had the mean MSE, i.e. $\text{MSE}_1 = 0.021$ and $\text{MSE}_2 = 0.052$ for f_1 and f_2, respectively. Results are based on 21,000,000 iterations including 1,000,000 iterations burn-in. With 2,000 iterations lag between each stored iteration, this gives a total sample size of 10,000. Again, we chose a truncated geometric distribution with parameter $c = 0.02$ for the number of clusters. The results gained with a uniform prior were virtually identical.

In Figure 4.2 the simulated data and the posterior median estimates for f_1 are displayed. A visual inspection of the reconstruction reveals only minor irregularities around the intersection of the vertical and the horizontal edge. While the simulated data range from -4.7 to 4.2, the posterior median estimates are smoothed with a range of -1.26 to 2.23. Altogether, the model discovers the true setting very well. The variance of the noise process is well estimated with a posterior median of $\hat{\tau}^2 = 1.016$, and 5%- and 95%-quantiles of 0.892 and 1.175.

Moreover, the algorithm detects the simple spatial structure. The number of clusters k has

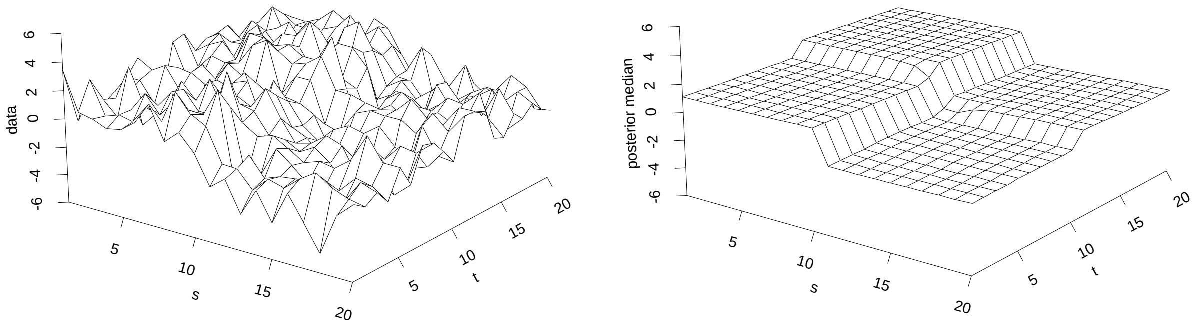


Figure 4.2: Simulated data (left) and posterior median estimates (right) for f_1.

a posterior median of 6 with a minimum number of 4 and a maximum number of 22. Finally, there is clear evidence in the posterior for the existence of the two edges. In Figure 4.3 the posterior probabilities to be in the same cluster for all 760 pairs of adjacent pixels are displayed, divided into three groups, i.e. pairs along the horizontal edge of height 2 (e2), pairs along the vertical edge of height 1 (e1), and all other pairs (other). The lowest probabilities are observed along the horizontal edge. Somewhat higher are the probabilities along the vertical edge, while all other probabilities are significantly higher. For comparison, the variation in the corresponding prior probabilities is only minimal over the three groups.

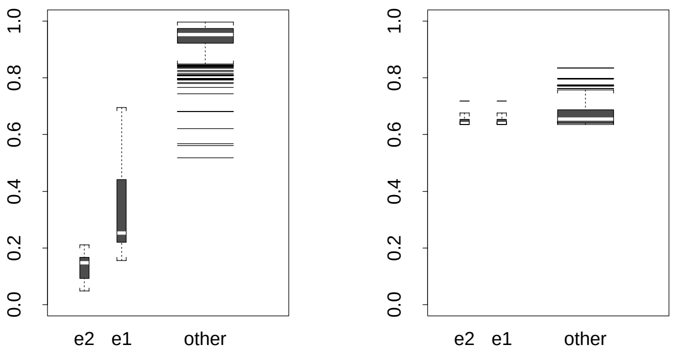


Figure 4.3: Posterior (left) and prior (right) probabilities to be in the same cluster for pairs of adjacent regions, grouped by location: along the horizontal edge (e2), along the vertical edge (e1), and all other pairs (other).

In consideration of the fact that there is only one observation for each pixel, the results for function f_1 are very convincing. For increasing number of observations, i.e. repeated measurements, the likelihood will support an even better reconstruction.

In contrast, the simulated data for the smooth function f_2 is shrunken too much to an overall mean. This is probably due to the fact that the prior gives preference to a (piecewise) constant surface. The posterior median estimates range from -0.95 to 1.44 compared to a range of the simulated data from -3.2 to 4.5. Clearly, the variation of the true function is underestimated.

This can also be seen in Figure 4.4. The posterior still reveals some slight edges, although estimates are derived by averaging over a large sample of step functions. The prior information is still visible, but again, for repeated observations at each pixel the likelihood will smooth those edges.

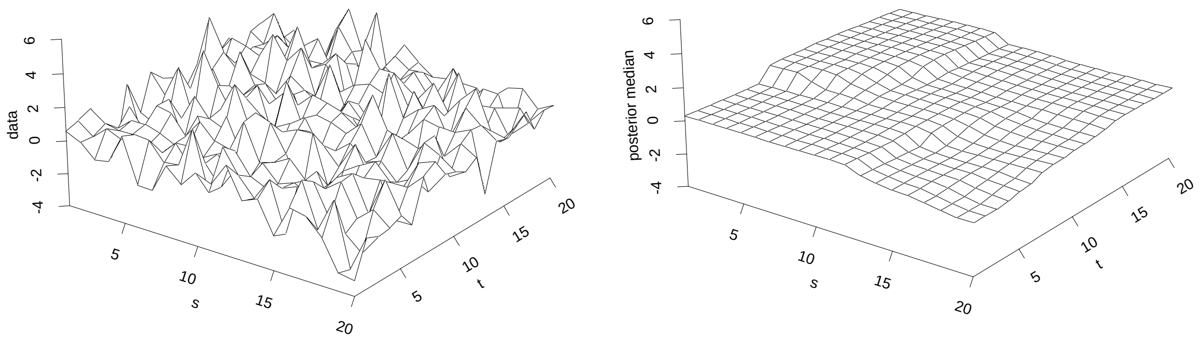


Figure 4.4: Simulated data (left) and posterior median estimates (right) for f_2.

From an objective point of view, the results are also good. The posterior distribution of the variance τ^2 is about the same as before. With a median of 1.018, a 5%-quantile of 0.901, and a 95%-quantile of 1.157 the estimates are close to the true value. The posterior of the number of clusters has a median of 8 and is slightly more dispersed than before with a minimum number of 4 and a maximum number of 40.

Although the reconstructed surface shows some edges which are not present in the true function, the model demonstrates that smooth patterns can be reconstructed well.

4.1.4 Application to fMRI data

As an application to real data, we consider measurements from functional Magnetic Resonance Imaging (fMRI). Data from fMRI experiments are used to identify activated regions in the human brain. The data presented here was collected in an experiment in which a test person was exposed to a visual stimulus for a period of 30 seconds followed by 30 seconds of rest. During the alternate sequence of 4 phases of rest and 3 phases of stimulus, data was recorded at $T = 70$ time points with 3 seconds lag in between. At each time point, brain activity is measured on a three-dimensional grid of N pixels (or voxels).

Typically, the quality of such data suffers from several sources of random error during the recording process, e.g. due to movement of the test person. In addition, there is a systematic distortion of the original ON-OFF stimulus to the signal perceived by the brain. Therefore, exhaustive preprocessing is necessary, usually carried out in form of a regression model. More precisely, observation y_{it} for voxel $i = 1, \dots, N$ at time $t = 1, \dots, T$ is gained by correcting the measurements for time trends and systematic transformations, see Gössl, Auer & Fahrmeir (2000) for a thorough discussion on this matter. After preprocessing, usually a spatial analysis is performed in order to identify activated areas in the brain. Our focus is solely on the latter part.

We consider data for one time point and one horizontal layer of pixels. Therefore, we have a two-dimensional lattice with $n = 2948$ pixels. The preprocessed data has been taken on from Lang & Brezger (2003) who also give further details on the preprocessing step. Furthermore, they report results for a spatial analysis of this data based on two-dimensional Bayesian P-splines. The aim of the analysis is to detect activated regions of the brain and separate them from non-activated regions. According to the nature of the stimulus activated pixels will mainly occur in the visual center of the brain. Although the data are discrete, a continuous model is reasonable due to the extremely large number of pixels. To speed up the analysis in our discrete model we constrain the image to 1179 pixels located in the rear part of the brain containing the visual center.

We have analyzed data for three time points: $t_1 = 18$, $t_2 = 38$, and $t_3 = 58$. These correspond to the first, second, and third period of stimulus, respectively. Detailed results are only reported for $t_3 = 58$ since this seems to be the roughest data. The large number of pixels is almost the limit of the capability of the CPM. We have increased the burn-in and lag to gain acceptable autocorrelations, especially for the number of clusters k. All results were collected in a run with 102,000,000 iterations including 2,000,000 burn-in and a lag of 20,000. Thus, posterior quantities are based on 5000 samples. We have used three different priors for the number of clusters k: uniform, geometric with parameter $c = 0.02$, and a rather informative Poisson prior with parameter $\mu = 30$. The latter choice was based on a visual inspection of the data alone. Surprisingly, differences in the posterior median estimates were found to be small. This indicates a strong spatial structure in the data which is discovered by all three priors. However, there seems to be no objective justification for an informative Poisson prior. Therefore, we will present results for the uniform prior in detail.

In Figure 4.5 the data and the posterior median estimates are displayed. Both show a strong spatial structure with rather extreme sudden changes. The estimates are plausible with large areas of almost constant values around zero. In general, this result is desired since zero corresponds to non-activated regions. Areas with estimated levels above zero mainly coincide with the known location of the visual center in the human brain.

To constitute such a clear structure, the partition model is limited to few viable partitions (compared to the enormous number of possible partitions). Therefore, one would expect low acceptance rates for the partition changing moves. However, those were passable, about 9% for the moves birth and death, nearly 10% for a switch, and over 44% for a shift. Still, the data seems to support only few partitions, while unsuitable partitions are often rejected. Further analysis of the posterior distribution of the cluster centers confirms this assumption. In Figure 4.6 the posterior probabilities of each pixel to be selected as a cluster center are displayed. There are 785 out of 1179 pixels that have a probability below 0.1, whereas there are only 31 pixels with a probability above 0.5. Note that the framed pixel in the center of the lattice indicates a pixel without observation.

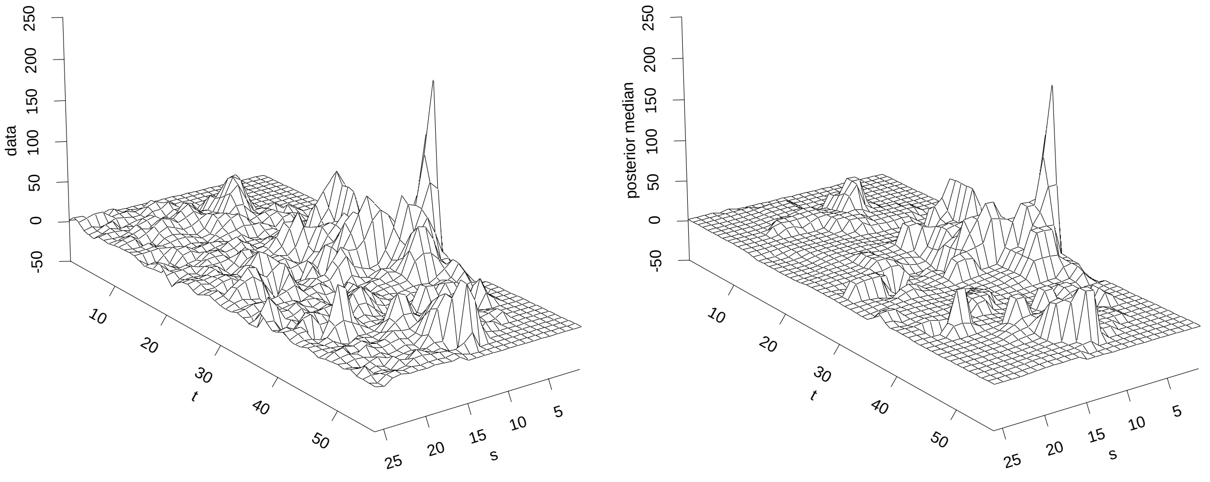


Figure 4.5: fMRI data for $t_3 = 58$ (left) and posterior median estimates (right).

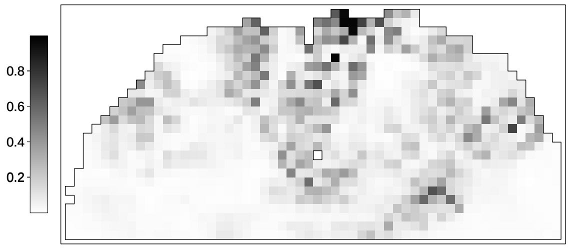


Figure 4.6: Posterior distribution of the cluster centers.

Still, the algorithm discovers a clear spatial structure. While the expected number of clusters is 590 a priori, this number is decreased considerably ranging from 80 to 185 with a median number of 128 in the posterior. From Figure 4.5, it becomes obvious that for some pixels almost no smoothing is performed. For example, the large peak of about 215 in the data is only shrunken to about 209 in the posterior.

For comparison, we will consider the results for the same data set gained by the Bayesian P-spline approach (Lang & Brezger 2003). They propose two different models. Their basic model has a global smoothing parameter, i.e. the variance is assumed to be the same over the whole space. Alternatively, they modify the model and allow spatially varying variances to account for sudden changes in the data. Note that Lang & Brezger (2003) have analyzed the whole layer, while our analysis is only based on a fraction of the pixels. Yet, a comparison will give some insight on the performance of our model.

Although the coarse structure is roughly the same with the P-spline and the CPM approach, a closer comparison of the results yields some obvious differences. Both P-spline models give very smooth estimates without sharp edges. Even for large areas of values around zero, the estimates are rather wavelike. Moreover, the data is shrunken much more than by the partition model. Especially, the large peak is estimated to 105 and 144 with global and adaptive variance, respectively.

To summarize, the CPM provides more clear structure in the estimates than the P-spline models. Extreme values are preserved while smaller changes in the surface are filtered out as noise. This indicates that the CPM prior performs spatially heterogeneous smoothing. Therefore, in the following section we will investigate the smoothing properties of the CPM prior in detail.

4.2 Some remarks on adaptive smoothing

Let us suppose we are given data $\boldsymbol{y} = (y_1, \ldots, y_n)$ with some underlying structure. We specify an observation model with parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$. Within a Bayesian framework the posterior distribution $p(\boldsymbol{\lambda}|\boldsymbol{y})$ of the unknown parameters $\boldsymbol{\lambda}$ is proportional to the likelihood times the prior, $p(\boldsymbol{\lambda}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\lambda})p(\boldsymbol{\lambda})$. Hence, the posterior is a trade-off between information in the data and prior knowledge. This is the basic concept of any Bayesian analysis: to revise prior assumptions on the unknown parameters with regard to the data.

The common conditional independence assumption that observations $\boldsymbol{y}$ are independent given the parameters $\boldsymbol{\lambda}$, neglects any underlying structure in the data. Hence, the likelihood will favor a rather rough surface for the parameters, in general. As an example, consider the disease mapping application where the SMRs, i.e. the maximum likelihood estimates, display strong variation. This complicates interpretation and may lead to incorrect conclusions. The terminology in the following considerations is based on such geographically structured data, although all findings hold true for other graphs as well.

The trade-off between likelihood and prior in a Bayesian setting offers the opportunity to smooth the estimates. For this purpose, we choose a prior that favors smooth estimates, in contrast to the likelihood. We are interested in spatial smoothing, i.e. smoothing with respect to the geographical location of the regions. To perform spatial smoothing some sort of correlation structure between parameters in (adjacent) regions has to be imposed by the prior.

We may improve the estimates by using a prior distribution, which models the correlation structure in an appropriate way. Such prior information allows us to mimic the spatial dependence of parameters if the true correlation structure is known. However, this will rarely be the case.

If the true correlation structure is unknown we are unable to tell if a fixed specified prior is appropriate or not. Thus, any choice for the prior distribution will be subjective. The only objective information we are given is the data. Therefore it would be preferable to use a spatial prior distribution which is able to adapt to the data. In the statistical context of smoothing, parametric or nonparametric, this problem is well-known and usually referred to as edge preserving smoothing, edge detecting, or change-point detection. In this context, usually the location of sharp changes in the parameter surface is unknown. Statistical models are developed, which estimate the unknown parameter in a smooth way but allow for sudden changes if there is evidence for this in the data. Most applications are taken from image analysis (e.g. Chu, Glad, Godtliebsen & Marron 1998, Polzehl & Spokoiny 2000) or time series data (e.g. Müller 1992, Barry & Hartigan 1992). Any of these models can be seen as adaptive in the sense that the smoothing effect is varying over the whole space and allows for points or regions, where the estimates are not smooth at all.

In a Bayesian context, smoothing is determined by the prior correlations of the parameters $\boldsymbol{\lambda}$. Consequently, a prior where the amount of smoothing is variable with respect to the data is a prior where the correlations are subject to statistical inference themselves. This implies the need for an additional prior and will define a hierarchical model.

4.2.1 Smoothing behavior of the CPM

First of all, there has to be remarked that for all applications considered so far the smoothing behavior depends on the hyperparameters. More precisely, the global amount of smoothing depends on the prior for the scale parameter, e.g. σ^2 in the disease mapping application. Yet, this is inherent in almost any statistical model and we concentrate on the smoothing properties apart from these parameters.

The joint prior distribution for the CPM, as defined in Section 2.2, implies a hierarchical structure itself. The joint prior $p(\boldsymbol{\theta}_k, \boldsymbol{g}_k, k)$ can be factorized in two components: the prior on the partition $p(\boldsymbol{g}_k, k) = p(k)p(\boldsymbol{g}_k|k)$ and the prior on the parameters $p(\boldsymbol{\theta}_k|\boldsymbol{g}_k, k) = p(\boldsymbol{\theta}_k|k)$. Whereas the parameters $\boldsymbol{\theta}_k$ are assumed to be independent a priori, this prior implies a specific correlation structure for the parameters $\boldsymbol{\lambda}$ conditional on the partition. For a fixed partition $\mathcal{C}_k$ with k clusters and parameters $\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k$ the correlation matrix is given by $\text{Cor}(\boldsymbol{\lambda}) = \boldsymbol{B}\boldsymbol{B}'$, cf. Section 2.2.2. Less formally, the correlation of parameters in regions i and j is

$$\text{Cor}(\lambda_i, \lambda_j|\mathcal{C}_k) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are assigned to same cluster,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

This simple correlation structure follows from the independence assumption on $\boldsymbol{\theta}_k$. According to this assumption even the parameters in two adjacent regions may be uncorrelated and the CPM prior allows for edges in the parameter surface. Still, (4.9) is a conditional statement given a fixed partition. In our model the partition is variable and the prior correlation will be non-zero for all pairs of parameters; for example, the prior $p(\boldsymbol{\theta}_k, \boldsymbol{g}_k, k)$ gives positive probability for partitions with constant parameters for all regions whenever $P(k = 1) > 0$.

By combining regions to clusters the resolution of the data is decreased. This leads to a blurring effect. In the CPM, a partition $\mathcal{C}_k$ into k clusters can be seen as a (structure-preserving) decomposition of the underlying graph. The original estimation problem is transferred to a simpler problem with less and independent parameters. Hence, given the partition, the correlation structure of the parameters $\boldsymbol{\lambda}$ is independent of the underlying graph.

In general, it is important to distinguish between local and global smoothing. The amount of smoothing on region level, i.e. the local smoothing behavior, of the CPM is determined by the size of the cluster a region is assigned to, regardless of all other clusters. In contrast, the global amount of smoothing, i.e. the global smoothing behavior, is determined by the number of non-zero elements in the correlation matrix of the parameters $\boldsymbol{\lambda}$. According to (4.9), this number strongly depends on the cluster configuration.

For a fixed partition with parameters $\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k$ the cluster sizes are given by the diagonal elements of the matrix $\boldsymbol{B}'\boldsymbol{B} = \text{diag}(m_1, \dots, m_k)$ as we have shown in Section 2.2.2. These cluster sizes also control the global smoothing behavior since $\sum_{j=1}^{k} m_j^2$ is the number of 1-entries in the correlation matrix. The n diagonal elements make no statement on the smoothing behavior of the partition and are not of interest. Hence, we may use the number of non-zero, off-diagonal entries in the correlation matrix

$$N_B = \sum_{j=1}^{k} m_j^2 - n$$

as a measure of smoothness. This offers easy interpretation since larger values of N_B indicate that smoothing is stronger globally. Furthermore, N_B allows to compare different partitions with the same number of clusters. This becomes obvious, if we investigate the properties of N_B.

Suppose $\mathcal{C}_k$ is a partition with k clusters and cluster sizes $m_1, \dots, m_k$. The average cluster size is $\bar{m} = \frac{n}{k}$. Then

$$\sum_{j=1}^{k} (m_j - \bar{m})^2 = \sum_{j=1}^{k} \left(m_j - \frac{n}{k}\right)^2 = \sum_{j=1}^{k} m_j^2 - \frac{n^2}{k} \geq 0, \tag{4.10}$$

and thus, N_B is minimized for $m_j = \bar{m},\ j = 1, \dots, k$, i.e. if all clusters have the same size. Since space is discrete, this will not be possible in general. Still, N_B is minimized if all clusters are about the same size and thus if the local amount of smoothing is approximately the same for all regions.

According to (4.10), a lower bound for N_B is given by

$$N_B = \sum_{j=1}^{k} m_j^2 - n \geq \frac{n^2}{k} - n = \frac{n}{k}(n-k), \tag{4.11}$$

which depends only on the number of clusters. Therefore, if local smoothing is homogeneous over the whole area, smoothing is less emphasized globally. On the other hand, if the global amount of smoothing increases, local smoothing becomes heterogeneous. The latter holds true since N_B is maximized if the variance of the cluster sizes is maximized, see (4.10). This is the case, if there are $k-1$ clusters of size 1 and one cluster of size $n-k+1$. More formally, an upper bound for N_B is given by

$$N_B \leq (n-k+1)^2 + (k-1) - n = (n-k+1)(n-k). \tag{4.12}$$

However, this is purely theoretical since for most graphs corresponding partitions will not exist. Still, higher values indicate that the local smoothing behavior is heterogeneous.

Both, the lower bound (4.11) and the upper bound (4.12) of N_B are monotonically decreasing with increasing k. For $k = n$, the cluster sizes are $m_j = 1$ for $j = 1, \dots, n$ and $N_B = 0$ which corresponds to no smoothing at all. In contrast, for $k = 1$, there is only one cluster of size $m_1 = n$

and with $N_B = n(n-1)$ the correlation matrix contains no zero-entires. Thus, smoothing is most extreme, i.e. the parameter surface is constant over the whole space. Therefore, the global amount of smoothing depends on the number of clusters.

To perform some kind of smoothing globally, the prior has to favor smooth estimates to oppose to the information in the likelihood. This will be the case, whenever the prior for the partition gives preference to smaller numbers of clusters. As mentioned above, the proposed prior for the partition $p(g_k, k) = p(g_k|k)p(k)$ is also hierarchical. With an appropriate prior for the number of clusters, the model will smooth the data. A rather flat prior for k, like the truncated geometric in our examples, guarantees some favor for smaller numbers of clusters. This seems to be enough to perform smoothing. Moreover, the prior probabilities for the vector of cluster centers $p(g_k|k) \propto (n-k)!$ are strictly monotonically decreasing with increasing k. Therefore, this prior also favors smaller number of clusters, whenever two partitions are approximately the same.

Altogether, in practice the CPM prior will perform smoothing globally. Moreover, this will also be true locally. Local smoothing is always performed, unless a region is alone in cluster by itself throughout the algorithm, i.e. in all samples from the posterior. Yet, a cluster of size one will increase the global amount of smoothing, unless the number of clusters is increased simultaneously. Therefore, smoothing as performed by the CPM prior is a compromise between local and global smoothing.

Now, suppose a state of the Markov chain in the RJMCMC algorithm with some partition $\mathcal{C}_k$. Our sampler allows to change this partition by the dimension changing moves birth and death, but also by the fixed dimension moves switch and shift. Changing the partition within the algorithm involves a change of the correlation structure and thus a change of the smoothing characteristic. Any proposed modification of the partition is accepted (or rejected) with respect to the information in the likelihood. Hence, inference on the partition can be seen as structural learning about the correlation matrix based on the data. In other words, smoothing is adaptive to the data.

As an example, consider the disease mapping application from Chapter 3. In Figure 4.7 a scatterplot of the posterior values of k versus N_B is displayed. It can be seen that the posterior distribution for N_B is shifted to higher values compared to the prior. Summarizing over all numbers of clusters the posterior probability to observe values above the 95%-quantiles of the prior is over 0.33, while the posterior probability for values below the 5%-quantiles of the prior is only 0.0042. Therefore smoothing is heterogeneous over the whole of Germany although the prior supports more homogeneous smoothing.

Far more extreme results are observed for the brain mapping data. The plot on the right hand side of Figure 4.7 shows that all values of N_B are greater than the 95%-quantile of the prior distribution. Clearly, there is evidence in the data to revise the prior amount of smoothing.

Finally, the smoothing behavior due to the partition is invariant to reparametrizations of the model. This becomes most obvious from the marginal approach in Section 3.6. Inference on the partition can be performed regardless of the unknown parameters $\boldsymbol{\theta}_k$. The marginal likelihood

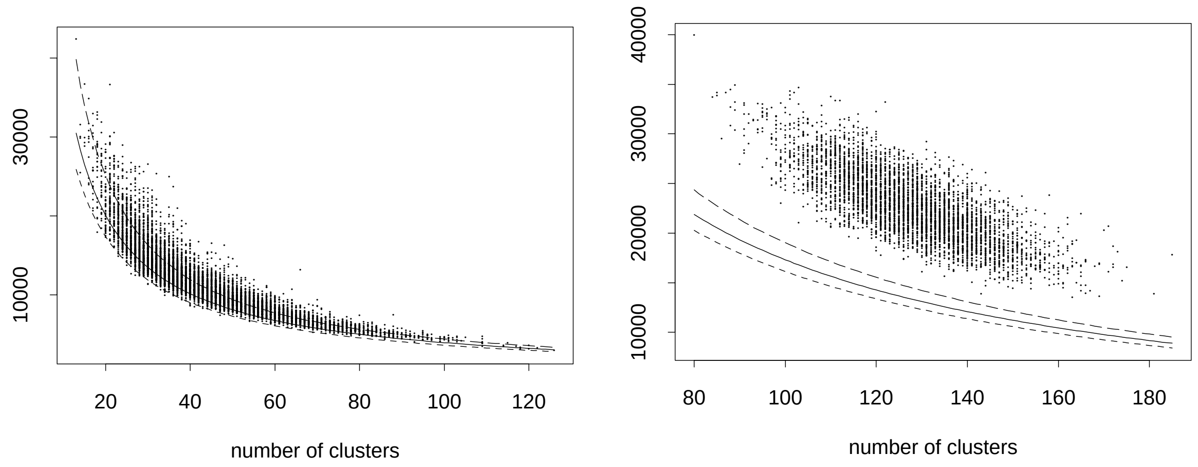


Figure 4.7: Posterior values of N_B for the oral cavity cancer data (left) and fMRI data (right). Prior median (solid line), and 5% and 95% quantiles (dashed lines) are also shown.

only uses information on the hyperparameters, e.g. α and β in the Poisson-gamma model. In principle, this is a valid approach for any combination of observation model and prior on the unknown parameters. Yet, in general, neither the marginal likelihood nor the full conditional for the risk parameters can be derived analytically.

For example, consider the log-normal prior for the relative risks θ_j, $j = 1, \ldots, k$, in the disease mapping application. We might reparameterize the model, and the two formulations

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}_k) &= \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i\theta_j)^{y_i}}{y_i!} \exp\left(-e_i\theta_j\right) && \text{with} \quad \theta_j \sim \mathrm{LN}(\mu, \sigma^2), \\ p(\boldsymbol{y}|\boldsymbol{\theta}_k) &= \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i \exp(\theta_j))^{y_i}}{y_i!} \exp\left(-e_i \exp(\theta_j)\right) && \text{with} \quad \theta_j \sim \mathrm{N}(\mu, \sigma^2) \end{aligned}$$

are equivalent as long as the hyperpriors for μ and σ^2 are identical. In general, other reparametrizations are also imaginable, but it might be impossible to adapt the hyperpriors accordingly.

4.2.2 Comparison to Markov random fields

In sharp contrast to (4.9) is the conditional approach for Markov random fields. Recall, that for MRFs two regions are called neighbors if they contribute to the full conditional of each other. Hence, the conditional correlation is always non-zero whenever two regions i and j are neighbored. This conditional correlation is determined by the precision matrix $\boldsymbol{Q}$. To perform spatial smoothing the precision matrix is chosen fixed, usually with non-zero entries for pairs of geographically adjacent regions. Thus, the definition of neighborhood for the MRF is in agreement with the definition of geographical neighborhood as given in Section 2.1.1. We will focus on this definition, although other choices are possible.

Now consider a GMRF for the parameters $\boldsymbol{\lambda}$, i.e. a pairwise difference prior with scale parameter κ

$$\begin{aligned} p(\boldsymbol{\lambda}|\kappa) &\propto \exp\left\{-\frac{\kappa}{2}\sum_{i\sim j}(\lambda_i-\lambda_j)^2\right\} && (4.13)\\ &= \exp\left\{\frac{\kappa}{2}\sum_{i<j}k_{ij}(\lambda_i-\lambda_j)^2\right\}. && (4.14)\end{aligned}$$

The precision matrix is given by $\boldsymbol{Q} = \kappa\boldsymbol{K}$, where $\boldsymbol{K} = (k_{ij})$ is a *penalty matrix* with off-diagonal elements

$$k_{ij} = \begin{cases} -1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}, \quad \text{for } i \neq j,$$

and the number of neighbors on the diagonal

$$k_i = k_{ii} = -\sum_{j\neq i} k_{ij}\,, \quad i = 1,\ldots,n.$$

Note that the off-diagonal elements in the penalty matrix are the negative entries of the adjacency matrix of the underlying graph, i.e. $k_{ij} = -a_{ij}$ for $i \neq j$ (cf. Section 2.1.2). This penalty matrix controls the conditional correlation structure and therefore Clayton (1996) also calls it the "inverse variance-covariance structure". The parameters in two adjacent regions are conditionally correlated

$$\mathrm{Cor}(\lambda_i, \lambda_j|\,.\,) = \frac{1}{\sqrt{k_i k_j}}, \quad \text{for } i \sim j.$$

Thus, the conditional correlation of two parameters solely depends on the number of neighbors of the two regions and is fixed. The local smoothing behavior of the GMRF is predefined by the specification of the precision matrix. Note that the local amount of smoothing is determined by the marginal variance of the regions, not the conditional. For pairwise difference priors, the marginal variances are not defined, but can be derived under linear constraints, see Section 4.3.4.

What varies is the global amount of smoothing according to the unknown scale parameter κ. For fixed precision κ, prior (4.13) penalizes differences in the parameters $\boldsymbol{\lambda}$ and supports a smooth parameter surface. Hence, the prior opposes the likelihood and allows for smoothing, where the global amount of smoothing depends on the scale parameter. But, smoothing is not adaptive to the observed data since the penalty matrix depends only on the underlying graph. Thus, there is no structural learning in MRFs.

Using other definitions of neighborhood will not change this, e.g. the use of second-order neighborhoods (see Figure 2.2) will lead to smoother results but not to adaptive smoothing. Other non-Gaussian approaches, e.g. based on absolute differences

$$p(\boldsymbol{\lambda}|\kappa) \propto \exp\left\{-\frac{\kappa}{2}\sum_{i\sim j}|\lambda_i-\lambda_j|\right\},$$

are more robust versions and allow for stronger edges in the parameter surface. Still, the smoothing behavior depends only on the underlying graph.

For GMRFs, adaptive smoothing requires inference on the structure of the precision matrix, i.e. inference on the elements of the penalty matrix $\boldsymbol{K}$. One approach is to interpret the k_{ij} in (4.14) as (negative) weights on the differences between the parameters. Fahrmeir, Gössl & Hennerfeind (2003) propose a model where the non-zero entries in the penalty matrix are stochastic and estimated within the algorithm. This is an appealing extension but holds the unpleasant feature that the normalizing constant of the pairwise difference prior is difficult to derive. Furthermore, smoothing is now variable and adaptive to the data but the structure of the penalty matrix is still fixed because only predefined non-zero elements of $\boldsymbol{K}$ are subject to statistical inference.

A further step would be to assume a variable neighborhood structure. For example, we may implement a move to switch off-diagonal elements of $\boldsymbol{K}$ from 0 to -1 and reverse (and simultaneously update the diagonal elements). This idea would indeed refer to structural learning based on the data. Still, some care has to be taken to assure symmetry of $\boldsymbol{K}$. In addition, the extreme case with $k_{ij} = 0$ for all pairs (i, j) has to be avoided. In this case, parameters are independent and the pairwise differences between parameters get irrelevant. Thus, the prior (4.14) does not oppose the likelihood and no spatial smoothing is performed.

4.2.3 Summary

The CPM is one possibility to perform adaptive smoothing with respect to the observed data for arbitrary graphs. Adaptiveness is achieved by inference on the correlation matrix of the parameters. The prior model, as proposed in Section 2.2.2, assumes that parameters $\boldsymbol{\lambda}$ are constant within each cluster. At first sight, this is a rather strong assumption but crucial for any spatially adaptive estimation.

In general, this assumption is not necessary. There are applications, where other formulations might be useful. Indeed, there exist related approaches in which the assumption of constant parameters is loosened. Holmes et al. (1999) propose a Bayesian partition model for applications in continuous space. In one dimension, this can be seen as regression modeling with partitions for which in every subset the unknown function is linear instead of constant. This is the continuous analogue to the piecewise linear model $\lambda_i = \alpha_j + \beta_j t_i\,,\ t_i \in C_j$, for time series data, already tackled in connection with Example 2.4. Although the parameters $\boldsymbol{\lambda}$ are not constant within each cluster anymore, the parameters defining the linear pieces still are, i.e. the intercepts $\boldsymbol{\alpha}_k = (\alpha_1, \ldots, \alpha_k)$ and the slopes $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_k)$. Thus, the more flexible model is achieved by increasing the dimension of the parameter space, i.e. $\theta_j = (\alpha_j, \beta_j)$ for cluster C_j.

More generally, any deterministic functions f_j between the unit identifier (e.g. time point i) and the parameter are conceivable, i.e. $\lambda_i = f_j(t_i)$ for $i \in C_j$. For reasons of identifiability the dimension of the parameters $\theta_j,\ j = 1, \ldots, k$, should be well below the number of observations in each cluster. However, this idea only works for certain graphs. To define such functions, the unit identifiers have to carry information on the location of the unit in the graph. Although the method might be applied for regular lattices of any dimension, it will be impractical for irregular graphs. For most irregular graphs, it will be impossible to formulate a meaningful cohesion between unit identifier and corresponding parameter due to the fact that any irregular graph (in our terminology) is non-ordered in a classical sense.

An obvious and practically tractable extension would be a stochastic version. For example, we might assume that all parameters in one cluster arise from the same distribution. Alternatively, we might apply a MRF prior for the parameters within each cluster, but still treat clusters independent. In both cases an additional level is inserted into the hierarchical model. This decreases the imputed influence of the partition model on the estimates and complicates identification of the parameters. In general, identification of parameters is simplified by a deterministic relation between unit identifier and parameter. Moreover, all results so far suggest that a CPM with deterministic (i.e. constant) cohesion is suitable for practical use.

4.3 Simulations from the prior distribution

In Section 2.1 some basic properties of the CPM prior were derived for general graphs. For specific graphs further characteristics of the prior, e.g. the local smoothing properties, are of interest. Unfortunately, it seems to be impossible to derive such properties analytically, at least, if the number of vertices is large. Therefore, we have done various simulations from the prior distribution $p(g_k, k)$ for the underlying graphs of all three applications presented so far, i.e. the map of Germany from Chapter 3 as well as the two lattices of the synthetic and the fMRI data from Section 4.1. As before, we will use the terminology of geographical data.

For a single region, the local smoothing behavior is determined by the size of the cluster the region is assigned to, as worked out in the last section. We therefore have calculated the average size of the clusters a region is assigned to as well as the probability for each region to be alone in a cluster of size one. Both terms strongly depend on the number of clusters, a priori. While the probability for being alone increases with increasing number of clusters, the average cluster sizes get smaller. In any case, the prior properties for a single region depend on its location in the graph, especially on the number of neighbors and the prior properties of the neighbors. Due to the large number of regions (between 400 and 1179) there is no convenient way to present exact results on an individual basis.

For each of the three graphs, results were gained by drawing 10^8 independent samples from the prior distribution as proposed in Section 2.2. In each sample the number of clusters k is drawn from the prior distribution $p(k)$, the elements of the generating vector g_k are chosen randomly according to $p(g_k|k)$, and finally the clustering partition is calculated.

The simulations reported below support two major conclusions. First, the probability to be alone in a cluster mainly depends on the number of neighbors of the region but not on the location of the region in the graph. This seems intuitive since for a selected region to form a cluster of its own it is necessary that the region is a cluster center itself and that there is at least one region with distance one or two also selected as cluster center. The opposite seems to be true for the average cluster size. While the number of neighbors has little effect, clusters near to the border of the graph tend to be smaller in size. Hence, smoothing is less emphasized on the border of the graph. However, this edge effect was found to be small for all graphs.

4.3.1 Results for the map of Germany

This graph is the most interesting since it is of irregular structure and results seem to be unpredictable for the $n = 544$ regions. The number of neighbors varies between 1 and 11, the distribution is displayed in Figure 4.8. Shown are the results of a simulation with a truncated geometric distribution with parameter $c = 0.02$ for k. This is the prior distribution used in the application presented in Section 3.3.2. The expected number of clusters is nearly 50.

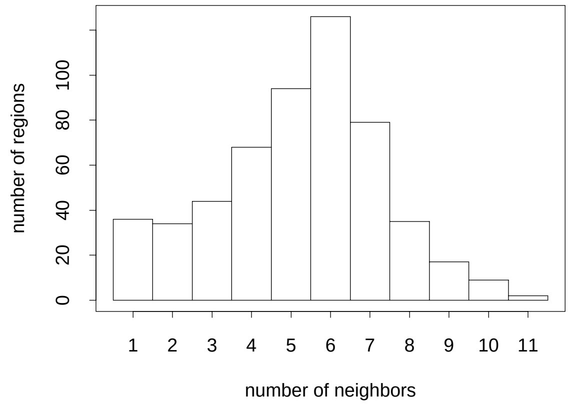


Figure 4.8: Distribution of the number of neighbors for the map of Germany.

Results are displayed in two graphs, individually as maps and summarized as boxplots grouped with respect to the number of neighbors. Note that the width of all boxplots is proportional to the square root of the number of regions in each group. In Figure 4.9 the probabilities of being alone in a cluster are quite small due to the fairly low expected number of clusters. With growing number of neighbors the probabilities are getting smaller. Consequentially, all regions with only one neighbor are clearly highlighted in the map. These are mainly medium-sized towns in Germany.

However, these probabilities have little effect on the average cluster sizes, at least if the prior for k gives preference to small numbers of clusters. From Figure 4.10 it becomes obvious that the average cluster sizes have approximately the same distribution regardless of the number of neighbors. Only those with many neighbors show slight deviations, but these regions are few.

We have done another simulation with a uniform distribution on $\{1, \ldots, 544\}$ instead of the geometric prior for k, see Appendix C.1. The results were similar in general, but the influence of the number of neighbors on the average cluster size was slightly stronger. Still, this can easily be avoided by a prior that favors small values for k.

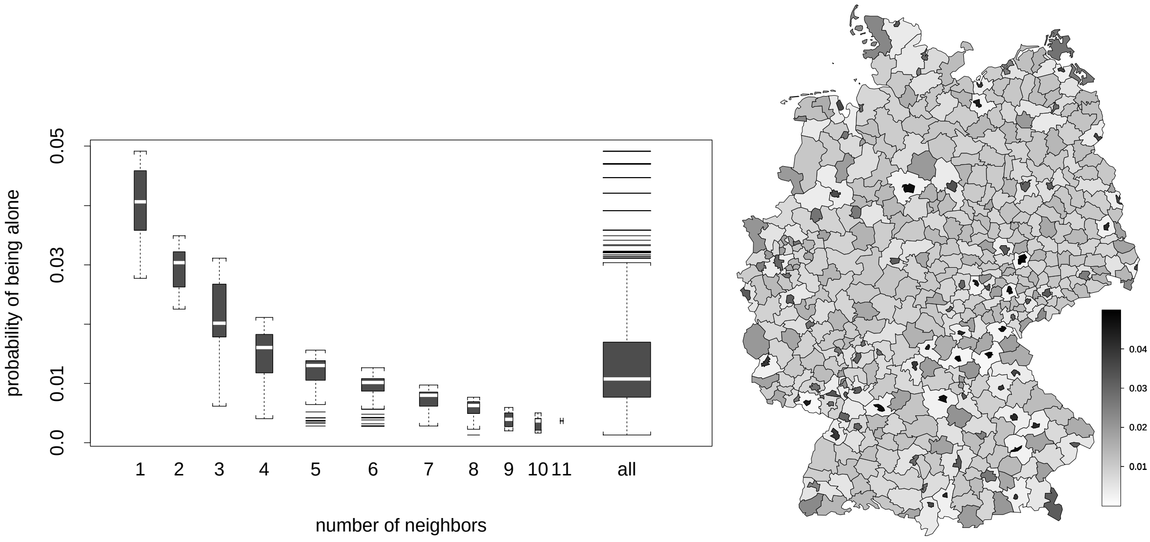


Figure 4.9: Probability of being alone in a cluster for the map of Germany.

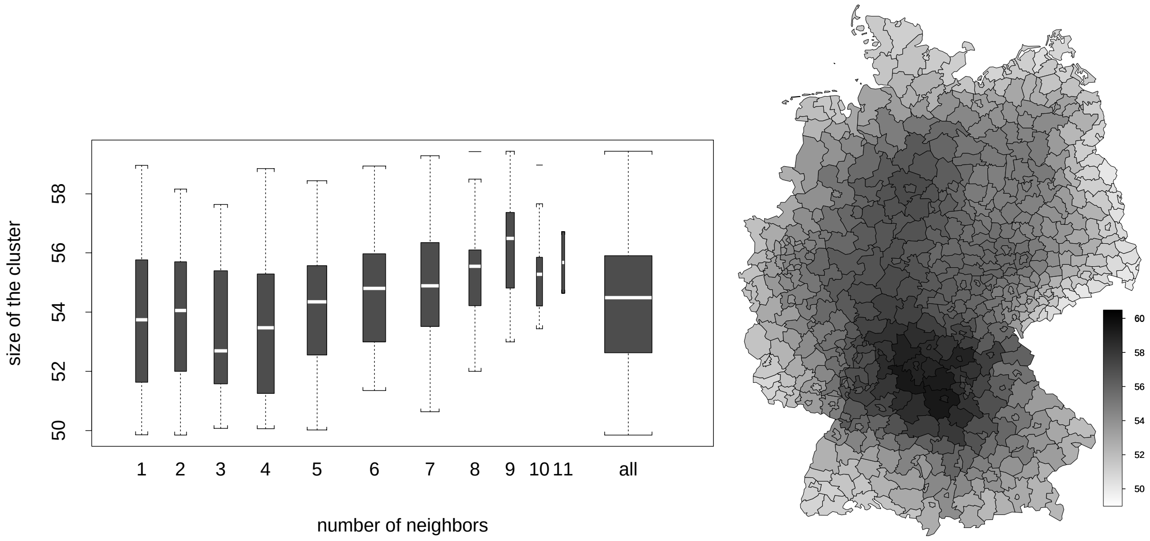


Figure 4.10: Average cluster sizes for the map of Germany.

4.3.2 Results for the 20 × 20-lattice

This is the most regular graph considered in this thesis with $n = 400$ pixels. There are 4 pixels with two neighbors, 72 pixels with three neighbors, and 324 pixels with four neighbors. Shown are the results using a geometric prior distribution for k with parameter $c = 0.02$. Again, results for a uniform distribution on $\{1, \dots, 400\}$ are given in Appendix C.2. For the geometric prior the expected number of clusters is about 50, as before.

In the left panel of Figure 4.11 the probability for each pixel to be alone in a cluster of size one is displayed. The results strongly reflect the regular structure of the graph. There are only six different values, clearly identifiable. Again, the probability of being alone decreases with increasing number of neighbors, but is constant except for pixels on the two border rows of the graph. The highest probabilities are observed at the corners, with only two neighbors. Interestingly, the lowest probabilities exist at their second-order neighbors towards the center of the graph. Obviously, the probabilities of being alone depend not only on the number of neighbors, but also on the prior properties of the neighbors. The right panel in Figure 4.11 shows the average size of the clusters each pixel is assigned to. The average size increases with increasing distance to the border of the graph. However, the range between 35 and 40 is moderate and the local smoothing behavior is rather the same for all pixels, a priori.

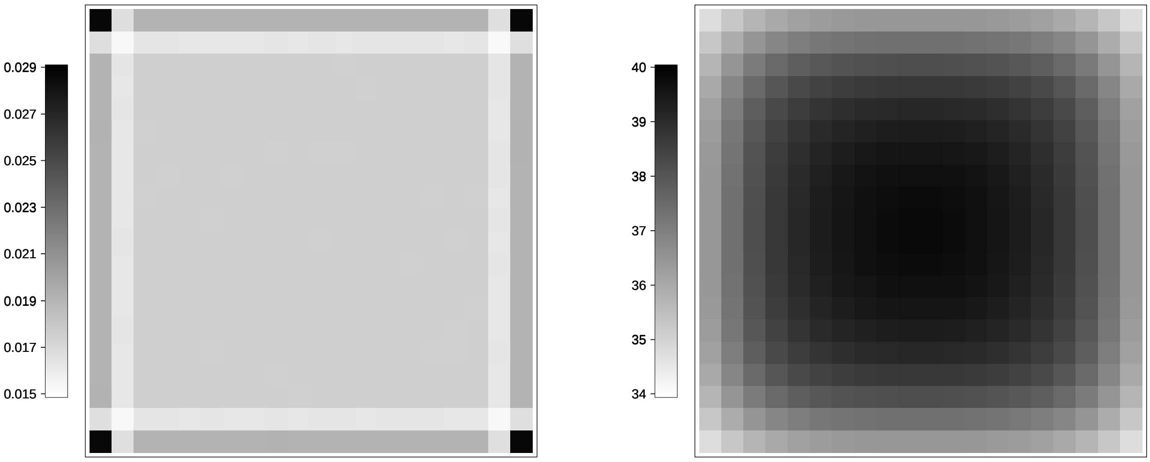


Figure 4.11: Probability of being alone in a cluster (left) and average cluster sizes (right) for the 20 × 20-lattice.

4.3.3 Results for the fMRI-lattice

This graph is basically regular but has a border of irregular shape. In addition, there is one pixel missing in the center of the graph. There are $n = 1179$ pixels and the number of neighbors varies between 1 and 4. In fact, there is only 1 pixel with one neighbor, located on the border of the graph, while there are 33, 109, and 1036 with two, three, and four neighbors, respectively. All results presented here are based on a uniform distribution on $\{1, \ldots, 1179\}$ for the number of clusters.

The probabilities of being alone in Figure 4.12 are almost constant except for pixels along the border of the graph. The pixel with one neighbor has the highest probability and with increasing number of neighbors the probabilities are decreasing, in general. Yet, as before, the lowest probabilities are observed in pixels located next to pixels with high probabilities. The right panel in Figure 4.12 shows slightly varying average cluster sizes. Clearly visible is the effect of the single missing pixel in the center of the graph leading to lower cluster sizes in nearby pixels. This effect is even more emphasized using a Poisson prior with parameter

$\mu = 30$ for the number of clusters (see Appendix C.3).

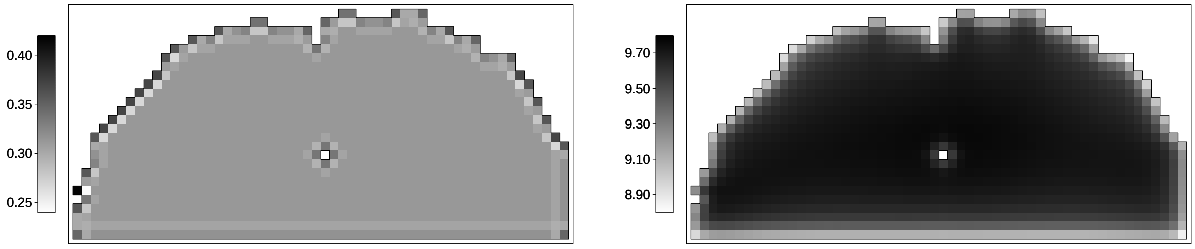


Figure 4.12: Probability of being alone in a cluster (left) and average cluster sizes (right) for the fMRI-lattice.

4.3.4 Comparison to Markov random fields

For GMRFs the local smoothing behavior is determined by the marginal variance and not the conditional variance. For the pairwise difference prior (4.13) the marginal variance is not defined since the precision matrix $\boldsymbol{Q} = \kappa \boldsymbol{K}$ is singular and thus not invertible. Yet, the covariance matrix $\boldsymbol{Q}^{-1}$ exists under linear constraints (Box & Tiao 1992). We have calculated the marginal

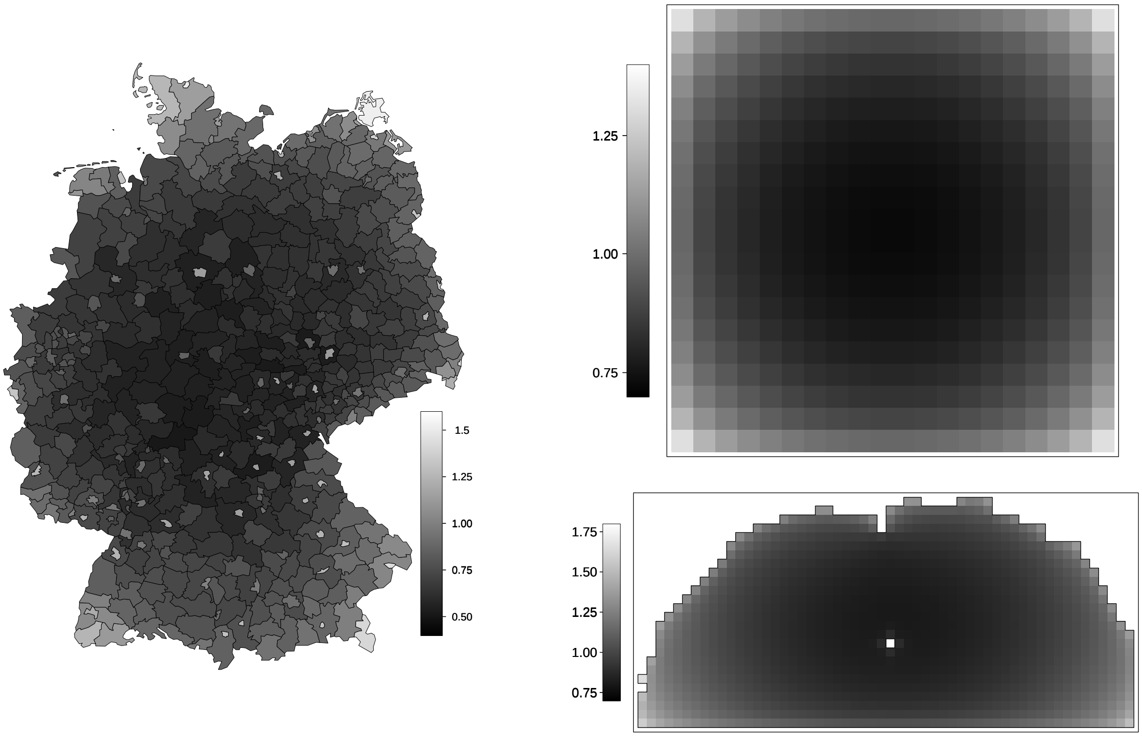


Figure 4.13: Marginal standard deviations for GMRFs with precision $\kappa = 1$.

variances under the assumption that the parameters sum up to zero. In Figure 4.13 the marginal standard deviations are displayed for all three graphs considered so far. Note that these are calculated for precision $\kappa = 1$ and thus the actual values are only proportional to those displayed.

For both regular graphs the smoothing effect is similar to the CPM prior, displayed in the right panels of Figures 4.11 and 4.12. The marginal standard deviations are higher at the border of the graphs and are decreasing with increasing distance to the border. Thus, smoothing is stronger for pixels located in the center of the graph. Note that the color scale in the figures is now inverted since large average cluster sizes in a CPM and low marginal variances in a GMRF model refer to a similar (stronger) local smoothing effect.

From the map of Germany, the difference between the GMRF prior and the CPM prior becomes obvious. While the irregular structure has little effect in the CPM (see Figure 4.10), the GMRF prior is extremely sensitive to irregularities. Here, the smoothing effect depends not only on the location of the region in the graph but also on the number of neighbors. Clearly, all regions with only one neighbor have a noticeable higher marginal variance. By the definition of a GMRF, this is true for the conditional variance, but carries over to the marginal variance, at least to some degree. This is an awkward feature of GMRFs, but one that cannot be avoided.

4.4 Computational issues

To close this chapter, some remarks on the implementation of the clustering partition model are given and computational issues are discussed. In general, for a graph with n regions, there are n associated parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ to be estimated. In addition, corresponding hyperparameters have to be estimated, e.g. μ and σ^2 in the disease mapping application. Yet, these are persistent in any hierarchical model and will be left out in this discussion.

Suppose a fixed partition $\mathcal{C}_k$ with k clusters and parameters $\boldsymbol{\theta}_k$. The CPM reduces the number of parameters to the number of clusters $k \leq n$. Thus, the number of parameters will be lower than the number of regions in general (unless we increase the dimension of the parameter space). In addition, the model requires k cluster centers, but these can be seen as nuisance parameters, in which we are not interested for further inference.

For computational speed and straightforward implementation, the assumption of (conditional) independence of the parameters $\boldsymbol{\theta}_k = (\theta_1, \ldots, \theta_k)$ is crucial. Due to this assumption, sampling from the posterior distribution can be performed by updating the k parameters one by one. This is in contrast to GMRFs based on single-site Gibbs sampler updates of the parameters. For GMRFs, sometimes high autocorrelations of the sampled values are observed (Fahrmeir & Lang 2001, Knorr-Held & Rue 2002). This leads to slow mixing behavior of the Markov chain. To remedy this drawback, block updating procedures are usually used. However, updating some or all parameters in a GMRF simultaneously involves sampling from high-dimensional normal distributions, and thus matrices of high dimensions need to be inverted. This will slow down the algorithm although fast sampling schemes are known, based on Cholesky decompositions of the precision matrix of the GMRF (Rue 2001).

Given a fixed partition in the CPM, independent sampling of the parameters is much easier and faster. The major computational cost concerns the calculation of the partition. Whereas sampling of the cluster centers is fast, the assignment of a single region to its cluster demands a pairwise comparison of distances between the region and all cluster centers, cf. assignment rule (2.3). Hence, the computational cost is enormous. For applications in two-dimensional continuous space fast algorithms for the computation of Voronoi tessellations are known (Green & Sibson 1978). Yet, this is somewhat more tricky in a discrete setting. Here, the computing time strongly depends on the number of clusters in comparison to the number of regions.

Recall, that cluster centers are always assigned to the cluster which they generate. If the number of clusters is large, the number of regions for which the assignment has to be computed is low. Hence, the computation of the partition is rather fast. Similar, if the number of clusters is low, there are only few cluster centers and therefore only few pairwise comparisons of distances are necessary. Again, the computation is fast, especially for the extreme case with only one cluster. The computation is much more expensive if the number of clusters is neither high nor low. This will often be the case, as can be seen from the applications considered so far.

Any proposed change of the partition requires the computation of a new candidate partition, regardless of acceptance or rejection of the proposal. Low acceptance rates for the partition changing moves require longer runs to achieve satisfactory mixing behavior for the model indicator k. Therefore, low acceptance rates slow down the algorithm considerably.

The performance of reversible jump MCMC samplers is discussed controversial in the statistical literature. For many applications low acceptance rates of dimension changing moves are reported. One major field of research in which reversible jump MCMC methods are used is mixture modeling. Usually, the number of components of the mixture distribution is unknown and this is an ideal application for varying dimension samplers. In a mixture model approach, Fernández & Green (2002) report acceptance rates between 4.1% and 22.1%, but mainly below 8%, for synthetic and real data sets in the context of disease mapping. Similar rates between 4% and 18% are reported by Richardson & Green (1997), also in a mixture model context. Robert, Rydén & Titterington (2000) give even lower numbers for their hidden Markov models. They speak of "virtually zero" rates for birth and death moves, while competing split and combine moves lead to acceptance rates between 0.26% and 4.4%. Slightly higher values around 10% are mentioned by Green & Richardson (2002) for a hidden Markov model approach to disease mapping. Rates above 25% are given by Denison & Holmes (2001) for their Bayesian partition model applied to individual disease incidence data in continuous space.

Compared to these reference values, the proposed CPM does fairly well. Rather high acceptance rates for the dimension changing moves were gained for the disease mapping data, around 24% and 31%, see Sections 3.3 and 3.6. For the simulated data presented in Section 4.1.3 the rates were about 8% for the step function f_1 and almost 22% for the smooth function f_2. Even for the rather extreme fMRI data those rates were still about 9%. Usually, the acceptance rates of both partition changing moves in fixed dimension—shift and switch—were even higher.

To summarize, the acceptance rates gained with the CPM prior tend to be higher than those reported for other applications of reversible jump MCMC. The rates are low if the model detects substantial (spatial) structure in the data, i.e. if there are edges in the parameter surface. This is clearly the case for the simulated step function as well as for the brain mapping data. If the surface is rather smooth, the rates increase, like in the disease mapping example and for the smooth simulated function. Similar results were observed by Robert et al. (2000). Applying their hidden Markov model to simulated iid data sets yielded rates between 22% and 33%.

Low acceptance rates for the dimension changing moves correspond to slow mixing of the hyperparameter k, which controls the number of clusters. In all our applications we have used large lags between those iterations, stored for the calculation of posterior quantities. Thus, we have used extremely long runs, and autocorrelations for the model indicator were very good in most cases. Moreover, sample sizes were also chosen to be large, between 5,000 and 10,000 samples. In accordance, the computing time was large. Still, speed is a major issue for the practical use of any statistical method.

Note that slow mixing of the model indicator k is not necessarily connected to slow mixing behavior of the parameters $\boldsymbol{\lambda}$. In fact, autocorrelations of single parameters are usually found to be good. As an example consider reruns of the synthetic data sets from Section 4.1.3. With only 210,000 samples and a burn-in of 10,000 samples this is only 1% of the run length used before. With a lag of 100 samples, the following results are based on a sample size of 2,000. Still, the posterior median estimates are rather precise. The median and mean of the absolute differences in posterior median estimates between run and rerun were only 0.013 and 0.021 for function f_1, and only 0.006 and 0.009 for function f_2.

In Figures 4.14 and 4.15 sampling paths and autocorrelations are compared for the reconstruction of functions f_1 and f_2, respectively. In the left two columns the results for the number of clusters and one parameter of the original runs are shown. Note that for the original runs, only the first 2000 samples are displayed to match with the results of the reruns. The parameter λ_{210} refers to pixel $(11, 10)$ located at the intersection of the two edges in the step function f_1. The true levels in this pixel are -1 and 0.42 for f_1 and f_2, respectively.

For both original runs the sampling paths and the autocorrelations are almost perfect. Only for function f_1, small autocorrelations are observed for the number of clusters k. Yet, these autocorrelations are decreasing fast and arise due to the strong structure discovered by the model. In contrast, the samples of k are highly autocorrelated for both reruns displayed in the two columns on the right in Figures 4.14 and 4.15. This is a consequence of the much smaller lag between stored samples. Still, autocorrelations for the single parameters are remarkably good. Especially for the smooth function in Figure 4.15, but also for the step function in Figure 4.14. The latter is even more astonishing in the face of the location of the pixel. From the sampling path it becomes obvious, that the posterior is slightly bimodal. Yet, the autocorrelations of the sample are acceptable.

In this thesis, we have used extremely long runs to achieve good mixing and low autocor-

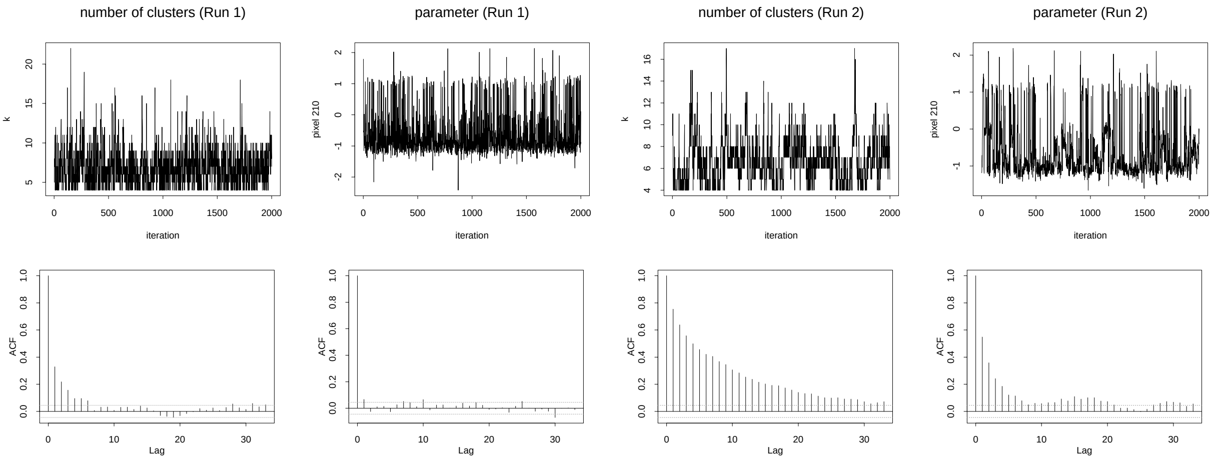


Figure 4.14: Sampling paths and autocorrelations for reconstructions of function f_1: number of clusters and one parameter for run 1 (left two columns) and run 2 (right two columns).

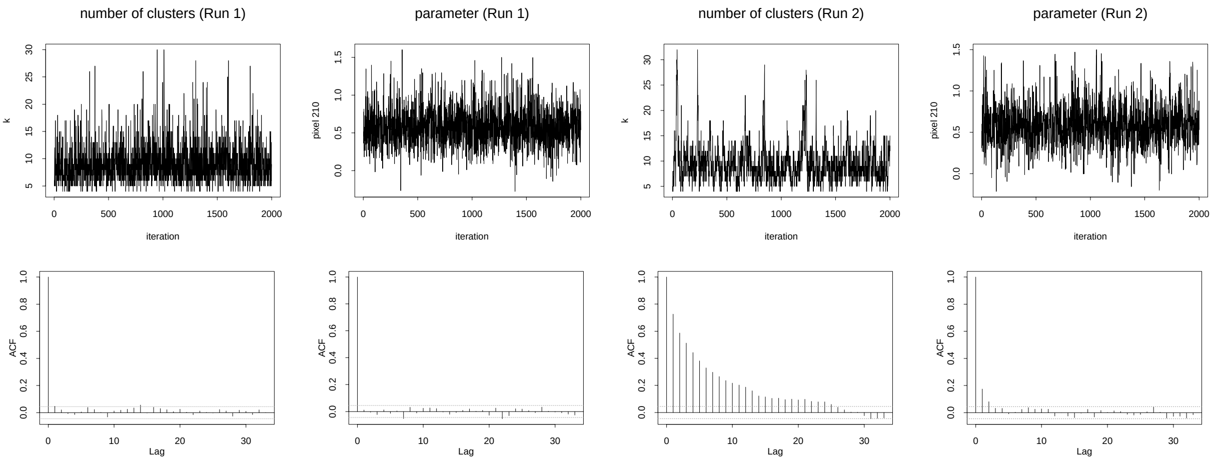


Figure 4.15: Sampling paths and autocorrelations for reconstructions of function f_2: number of clusters and one parameter for run 1 (left two columns) and run 2 (right two columns).

relations for the number of clusters k. For practical use of the CPM, both, the run length and the sample size, can be decreased to a fraction of the values used in this thesis. This will speed up the analysis considerably. For example, the reruns of the CPM sampler—coded in C++ and executed on a Sun Ultra 10 workstation—took only 49 and 80 seconds for function f_1 and f_2, respectively.

Chapter 5

Modeling Space-Time Interactions in Disease Risk

Besides the spatial variation of disease risk, the development over time is of major interest. For example, stomach cancer mortality in Germany has decreased significantly over the last two or three decades. The main reason for this time trend is a change of dietary habits of people over the years. However, care has to be taken in the interpretation of such a temporal effect. Many severe diseases, e.g. cancer, usually require years to develop and this period may even vary from person to person. Such time lag has to be considered in the interpretation of any time effect.

For many data sets, ML estimation of an overall time trend is straightforward. Based on cases aggregated over all regions, the sample size is usually sufficiently large to give reliable results. There is no need for sophisticated statistical models, unless the disease is extremely rare. In the same manner, SMRs as estimates for the spatial variation are more stable if we are given observations for a large period of time.

However, the assumption that the development over time is the same for all regions is rather strong. For example, the change of dietary habits of people is not necessarily the same for all regions. Hence, the presence of risk factors may vary over space and time. In this case, the assumption of one common time trend for all regions and one common spatial pattern for all time points is wrong. Any statistical model based on this assumption will lead to biased estimates. Therefore, some effort has been made to develop statistical models that incorporate space-time interactions.

Any model with space-time interactions has to be based on the finest resolution of the data. Thus, each region at each time point has to be considered separately. For sparse data, the quality of the SMRs will suffer. But even for reliable data, knowledge on the variation of disease risk over space and time may give useful hints on unknown risk factors.

5.1 Space-time data

In this section we will develop a model with space-time interactions for disease count data, based on a CPM prior. For this purpose, we start with a few remarks on necessary changes in the basic procedure. For example, the data have to be standardized appropriately to postulate a Poisson model. Furthermore, clustering of space-time data requires the definition of an underlying graph over space and time.

5.1.1 Standardization

Space-time data is stratified not only with respect to area and age group but also with respect to time (either time points or intervals). We are given the observed number of cases y_{ijt} in region $i = 1, \ldots, I$ and age group $j = 1, \ldots, J$, at time $t = 1, \ldots, T$. Let n_{ijt} denote the number of persons under risk in the same stratum.

As before, we postulate the usual assumption that the counts y_{ijt} have binomial distribution with unknown probabilities π_{ijt} and sample sizes n_{ijt}. Again, we approximate the binomial model with a Poisson model. Thus, we may aggregate over age group, and calculate the expected number of cases

$$e_{it} = \sum_{j=1}^{J} e_{ijt}, \quad i = 1, \ldots, I,\ t = 1, \ldots, T,$$

for region i at time t. Yet, deriving the expected number of cases is not a trivial task for data, observed at given geographical units for several time points, additionally stratified for age group (and possibly other covariates).

We will use two different (internal) standardizations of the data. In general, we fit a logit model for the raw data to adjust for age effects, see Section 1.1.1. In the simplest case we only use age effects in the linear predictor to calculate the SMR_1.

Alternatively, we may also include spatial and temporal effects. We will denote this data by SMR_2. If the number of time points is large, the data might also reflect cohort effects, related to unobserved risk factors present at the time of birth or any other fixed time point. Note that such cohort effects are not useful for our purpose. Including a cohort effect in the standardization process will not change the SMRs.

In this chapter we consider data on stomach cancer mortality of males in West Germany. There are $I = 30$ administrative regions, and data are available on a yearly basis over a period of $T = 15$ years from 1976 to 1990. Furthermore, the data are stratified by $J = 16$ age groups, defined by 5-year intervals ranging from age under 5 to age 80 and older. Due to data security reasons, data on cancer mortality in Germany is either available on a fine spatial resolution but on a low temporal resolution (cf. Section 3.3.2), or vice versa. There exist approaches to assess the disease counts on a high resolution for space and time simultaneously, see Schach (2003) for an investigation of the same data set presented here.

The total number of cases is 125,086 and varies between 42 and 870 with a median number of 243. Thus, the data are not sparse but rather informative. For this data set we have calculated both SMRs. In Figure 5.1 a scatterplot of SMR_1 and SMR_2 is shown with axes on a log-scale. The variation of the SMR_1 with a range from 0.57 to 2.26 is much stronger than the variation of the SMR_2, ranging from 0.76 to 1.25. Obviously, a large amount of the variation can be explained by separate space and time effects.

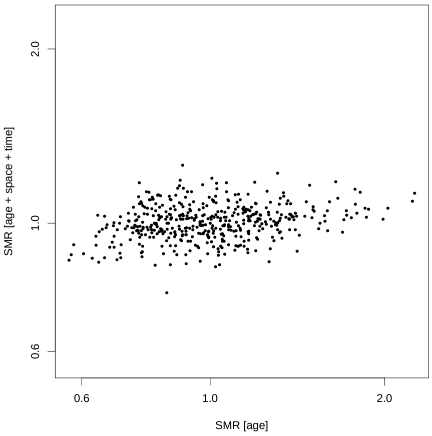


Figure 5.1: Scatterplot of SMR_1 and SMR_2. Standardization with respect to age (horizontal) and with respect to age, space, and time (vertical); both axes are on a log-scale.

5.1.2 Models for space-time data

With appropriately standardized data, we are able to formulate the usual model assumption

$$y_{it} \sim \text{Po}(e_{it}\lambda_{it}), \quad i = 1, \ldots, I, \; t = 1, \ldots, T,$$

where e_{it} is the expected number of cases in region i at time t. Again, the relative risk λ_{it} is assumed to be constant over age groups, i.e. $\lambda_{ijt} = \lambda_{it}$ for all j. Under the assumption of independence of spatial and temporal effects, the log relative risk can be decomposed additively

$$\log(\lambda_{it}) = \alpha + \beta_i + \gamma_t\,. \tag{5.1}$$

This corresponds to a generalized linear, additive, or mixed model, depending on the combination of prior choices for the spatial effect β_i and the temporal effect γ_t. We will call this the *main effects model.*

In the statistical literature various models are proposed for the introduction of space-time interactions. For the Poisson observation model, most approaches can be summarized under one general formulation

$$\log(\lambda_{it}) = \alpha + \beta_i + \gamma_t + \delta_{it}\,. \tag{5.2}$$

Here, δ_{it} is a space-time interaction term that is the focus of our statistical effort. Without further covariates this is the *saturated model.* Basically, Bernardinelli et al. (1995b) use this model assuming spatially varying linear time trends. A similar model is proposed by Assunção, Reis & Di Lorenzo Oliveira (2001) using spatially varying polynomial time trends. In both models, spatially structured coefficients of the temporal effects are assumed to follow GMRFs, a priori. Model (5.2) is also used by Lagazio, Dreassi & Biggeri (2001), but they propose to model birth cohort effects instead of temporal effects. Such a formulation may or may not be superior to "standard" space-time models, the statistical model and prior choices are the same. In fact, their model uses a GMRF directly on space-time points, a formulation which was also proposed by Knorr-Held (2000) within a binomial observation framework.

There are related approaches for space-time data that differ from the formulations above. Often, additional spatial and temporal random effects are included in the models. This is only a minor modification and such models still fit in the presented schemes. Yet, there are alternative models. For example, Böhning, Dietz & Schlattmann (2000) propose a mixture model where the counts arise according to a mixture of Poisson densities. The components of the mixture model are defined over space and time. Note that space-time interactions are also of interest for other data. Gössl, Auer & Fahrmeir (2001) propose a model for the simultaneous estimation of brain activity in voxels over time, cf. Section 4.1.4.

We model space-time interactions using a reparameterization $\nu_{it} = \alpha + \delta_{it}$, i.e.

$$\log(\lambda_{it}) = \beta_i + \gamma_t + \nu_{it} \,. \tag{5.3}$$

For ν_{it} we apply a CPM prior, the exact construction of which will be described in the next section. The CPM prior is rather flexible, and in principle the main effects β_i and γ_t can even be omitted from the linear predictor. Thus, the model can be further simplified to

$$\log(\lambda_{it}) = \alpha + \delta_{it} = \nu_{it} \,. \tag{5.4}$$

This is the space-time analogue of the purely spatial model from Section 3.2. From a theoretical point of view, the CPM should be able to incorporate the separate spatial and temporal effects included in the previous formulations. In practice, this will need a considerably larger number of clusters. It will be difficult to find a prior distribution equivalent to the saturated model. Thus, in general, the results will differ. One would assume that the estimates will be smoother (globally) due to the less flexible linear predictor.

We will investigate the performance of the saturated model (5.3) and the simplified model (5.4). For comparison, a reparameterized version of the main effects model (5.1)

$$\log(\lambda_{it}) = \tilde{\beta}_i + \gamma_t \tag{5.5}$$

with $\tilde{\beta}_i = \alpha + \beta_i$ is used as a benchmark. This allows to evaluate the significance of the interaction term. All effects are modeled using independent CPM priors. For both main effects this is not new, but for the interaction term some notes are advisable.

5.1.3 Partitioning space-time data

To apply a CPM prior for the interaction term ν_{it}, we have to define an underlying graph G that provides some convenient structure. The definition will be based on the two given graphs for space and time, i.e. the underlying graph G_s for the geographical map and the underlying graph G_t for the time points. Let the corresponding distance measures be denoted by d_s and d_t, respectively.

We define the underlying graph G for the space-time interaction with vertices, identified by pairs $(i,t),\ i = 1,\dots,I,\ t = 1,\dots,T$, for region i at time t. Thus, we have $n = I \cdot T$ vertices or data points. For the construction of a distance measure d on G we need information if any two data points are neighbors or not. Thus, we have to define neighborhood for space-time points. We will use an intuitive construction based on the neighborhood definitions of G_s and G_t. First, we define

$$(i,t) \sim (j,t) \text{ in } G, \text{for } t = 1,\dots,T, \quad \text{if } i \sim j \text{ in } G_s, \tag{5.6}$$

i.e. the definition of spatial neighborhood is carried over to space-time points for each time point. Similarly, for each region $i = 1,\dots,I$ we define the neighbors in time for point (i,t) to be

$$(i,t-1) \quad \text{and} \quad (i,t+1), \quad t = 2,\dots,T-1. \tag{5.7}$$

For the endpoints $t = 1$ and $t = T$ there is only one neighbor in time $(i,2)$ and $(i,T-1)$, respectively.

This definition of neighborhood—based on (5.6) and (5.7)—can be viewed as T spatial graphs G_s, stacked with respect to the order of the time points, or I temporal graphs G_t, arranged according to the spatial graph. The definition assures that the underlying graph G is regular for the time dimension in each region but irregular for the spatial part at each time point. Therefore, the distance for two arbitrary vertices (i_1,t_1) and (i_2,t_2) in G can easily be defined and decomposed

$$\begin{aligned} d((i_1,t_1),(i_2,t_2)) &= d_s(i_1,i_2) + d_t(t_1,t_2) \\ &= d_s(i_1,i_2) + |t_1 - t_2|. \end{aligned}$$

Obviously, this definition implies that the distance of two regions i_1 and i_2 at the same time t reduces to the spatial component, $d((i_1,t),(i_2,t)) = d_s(i_1,i_2)$. Similar, the distance of one region i at different time points t_1 and t_2 is only based on the temporal distance, $d((i,t_1),(i,t_2)) = d_t(t_1,t_2)$.

The distance measure d, as defined above, treats spatial and temporal neighborhood "identical", and the smoothing effect will be the same in both directions, a priori. Initially, the CPM prior was developed for "natural" graphs, induced by the data. Modeling space-time interactions by a partition model implies the convolution of two—substantially different—graphs. There are two different types of edges in G, either connecting two adjacent regions for the same time point, or connecting two subsequent time points for the same region. To adjust the prior model for this different types of edges, we may weight the two components. For this purpose, we introduce an additional hyperparameter ω that allows us to control our prior belief on the relation of spatial and temporal neighborhood. With $\omega > 0$, we define a *modified* distance measure

$$\begin{aligned} d_\omega((i_1,t_1),(i_2,t_2)) &= d_s(i_1,i_2) + \omega \cdot d_t(t_1,t_2) \\ &= d_s(i_1,i_2) + \omega \cdot |t_1 - t_2|. \end{aligned}$$

Partitioning, based on the distance measure d_ω, now can be tuned according to the data. More precisely, the local smoothing behavior is changed by ω, a priori. For values $\omega > 1$ spatial smoothing is preferred locally, i.e. the clusters consist preferably of regions at the same time point. In contrast for $\omega < 1$ temporal smoothing is preferred and the clusters are more likely to combine the same region over several time points.

Recall from Section 2.1.4 that the partition is invariant to any strictly monotonic increasing transformation of the distance measure, since clustering is solely based on the order of the distances, but not on the actual values. Hence, the partition is invariant to the multiplication of d_ω with any positive constant. Especially, partitioning based on

$$\begin{aligned} d^*_\omega((i_1,t_1),(i_2,t_2)) &= \frac{1}{\omega}\, d_\omega((i_1,t_1),(i_2,t_2)) \\ &= \frac{1}{\omega}\, d_s(i_1,i_2) + |t_1 - t_2|, \end{aligned}$$

defines the same CPM(ω) prior as partitioning based on the distance measure d_ω.

5.2 Prior specifications and implementation

The most general model, i.e. the saturated model, is very similar to the model with covariates as proposed in Section 3.5. Decomposition (5.3) is equivalent to a factorization of the relative risk, i.e. $\lambda_{it} = \exp(\beta_i)\exp(\gamma_t)\exp(\nu_{it})$. Hence, the likelihood for the saturated model is given by

$$p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\nu}) = \prod_{i=1}^{I}\prod_{t=1}^{T} \frac{(e_{it}\exp(\beta_i)\exp(\gamma_t)\exp(\nu_{it}))^{y_{it}}}{y_{it}!} \exp(-e_{it}\exp(\beta_i)\exp(\gamma_t)\exp(\nu_{it})),$$

with obvious changes for the simpler models (5.4) and (5.5). Note that we use a vector notation for the interaction effects $\boldsymbol{\nu}$, similar to the lattice data applications in Section 4.1.1.

For each component, we apply a CPM prior and further assume that all three components are independent of each other, a priori. As mentioned in Section 4.2.1, the CPM priors could alternatively be defined for the parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\nu}$, but in compliance to previous chapters we will work on the exponentials of the parameters. For simplicity, we use the notation $\exp(\boldsymbol{\beta})$, $\exp(\boldsymbol{\gamma})$, and $\exp(\boldsymbol{\nu})$ for the vectors of the effects.

For the interaction term $\exp(\boldsymbol{\nu})$ in (5.3) and (5.4) we assume a space-time CPM prior. As before, a partition into $k \leq n$ clusters $C_1, \dots, C_k$ is defined by a generating vector $\boldsymbol{g}_k = (g_1, \dots, g_k)$. Now, the elements of the generating vector are identified by pairs

$$g_j \in \{(i,t) : i = 1, \dots, I,\ t = 1, \dots, T\}, \quad j = 1, \dots, k.$$

The relative risks $\boldsymbol{\theta}_k$ on cluster level are assumed to have independent log-normal priors

$$\theta_j \sim \mathrm{LN}(\mu, \sigma^2), \quad j = 1, \ldots, k,$$

and the step function is defined by

$$\exp(\nu_{it}) = \theta_j, \quad \text{for } (i,t) \in C_j.$$

In general, the assignment of space-time points to clusters is based on the modified distance measure d_ω. Note that the parameter ω does not appear in any of the equations above. Still, we use a modified partition model that depends on ω.

For both main effects, $\exp(\boldsymbol{\beta})$ and $\exp(\boldsymbol{\gamma})$, we assume independent CPM priors on the set of regions $\{1, \ldots, I\}$ and the set of time points $\{1, \ldots, T\}$, respectively. The corresponding relative risk parameters have also log-normal distributions a priori. For identifiability, we use a stochastic restriction similar to the model with covariates and fix the location parameter of the log-normal priors to zero, cf. (3.10). Note that in the main effects model (5.5) the location parameter for the spatial component is not restricted.

Due to independence, the joint prior density of the parameters $\exp(\boldsymbol{\beta})$, $\exp(\boldsymbol{\gamma})$, and $\exp(\boldsymbol{\nu})$ is simply the product over three CPM priors. The hyperprior setting for each component is chosen similar to previous models. More precisely, we use a diffuse prior for the location parameter of the interaction term and of the spatial component in the main effects model. For the scale parameters we assume inverse gamma priors with appropriately chosen parameters. Similar to the covariate model, we use slightly more informative priors for the main effects than for the interaction term in the saturated model.

In general, the saturated model is not identifiable. The interaction term will be able to (partially) incorporate either spatial effects or temporal effects or even both. Still, in practice, this will lead to considerably larger numbers of clusters, and will thus be penalized by the CPM prior. For all applications presented below, we have observed no problems with identifiability. Note that exact restrictions

$$\sum_i \nu_{it} = 0 \quad \text{for } t = 1, \ldots, T \quad \text{and} \quad \sum_t \nu_{it} = 0 \quad \text{for } i = 1, \ldots, I$$

are not possible with a CPM prior for the interaction term.

The sampling scheme is chosen analogous to the previous chapters. Similar to the covariate model, updating of risk parameters is based on effect-adjusted expected numbers of cases. For example, in a height move, a new candidate parameter for cluster C_j of the interaction term is drawn from

$$\theta_j^* \sim \mathrm{G}\left(y_j + \frac{\tilde{\mu}^2}{\tilde{\sigma}^2}, \tilde{e}_j + \frac{\tilde{\mu}}{\tilde{\sigma}^2}\right), \tag{5.8}$$

where

$$\tilde{e}_j = \sum_{(i,t) \in C_j} e_{it} \exp(\beta_i) \exp(\gamma_t)$$

is the expected number of cases in cluster C_j, corrected for the spatial effects $\exp(\boldsymbol{\beta})$ and the temporal effects $\exp(\boldsymbol{\gamma})$. Besides this modification, the proposal distribution (5.8) is identical to the purely spatial model, see (3.4).

5.3 Results for stomach cancer mortality in West Germany

In this section we compare the performance of the three models proposed above. For simplicity we denote the main effects model (5.5) by model 1, the saturated model (5.3) by model 2, and the simplified interaction model (5.4) by model 3. The parameters of the prior distributions are chosen similar for all three models.

The interaction component has a geometric prior for the number of clusters with parameter $c = 0.05$ in model 2 and $c = 0.02$ in model 3. For the scale parameter we have used the usual inverse gamma prior with parameters $(a, b) = (1, 0.01)$. For both main effects, if present, the number of clusters is assumed to have a geometric distribution with parameter $c = 0.2$ for the spatial component and $c = 0.1$ for the time trend. The parameters of both inverse gamma priors were chosen identical with parameters $(a, b) = (1, 0.01)$ in model 1 and slightly more informative $(a, b) = (5, 0.5)$ in model 2.

The motivation behind these choices is simply to allow a flexible formulation for the interaction term and achieve good separation of main effects and interaction. We have tried various other combinations of prior specifications, but the influence on the results for the overall risk estimates was found to be small. Still, some differences were observable for single components.

For the beginning, we use a standard CPM prior for the interaction term with tuning parameter $\omega = 1$. The results for other choices of ω were almost identical in the saturated model. Only for model 3, a clear effect of this parameter was observable. A brief discussion of some results is given at the end of this chapter.

5.3.1 Results for SMR$_1$

As already mentioned, the mortality for stomach cancer, and thus the SMRs, are substantially decreasing over time. In Figure 5.2 the SMRs are displayed. Besides the time trend, a strong variation between regions is visible. However, it is unclear if the overall variation can be separated and represented by a time trend and a spatial effect alone. Our goal is to answer this question, and eventually estimate the space-time variation not absorbed by the main effects.

For the beginning, we discuss the results for model 1. In Figure 5.3 the posterior median estimates of the spatial component are displayed. A strong spatial structure is visible with elevated risk in the whole of Bavaria and an average risk level without dramatic changes elsewhere. The most extreme risks are estimated for Lower Bavaria and Upper Palatinate, the two administrative districts in the East of Bavaria. This effect is also observable from the SMRs in Figure 5.2, but less emphasized for the last years of the observation period.

The decreasing risk over time is clearly visible in the SMRs. The maps for the years 1988 to

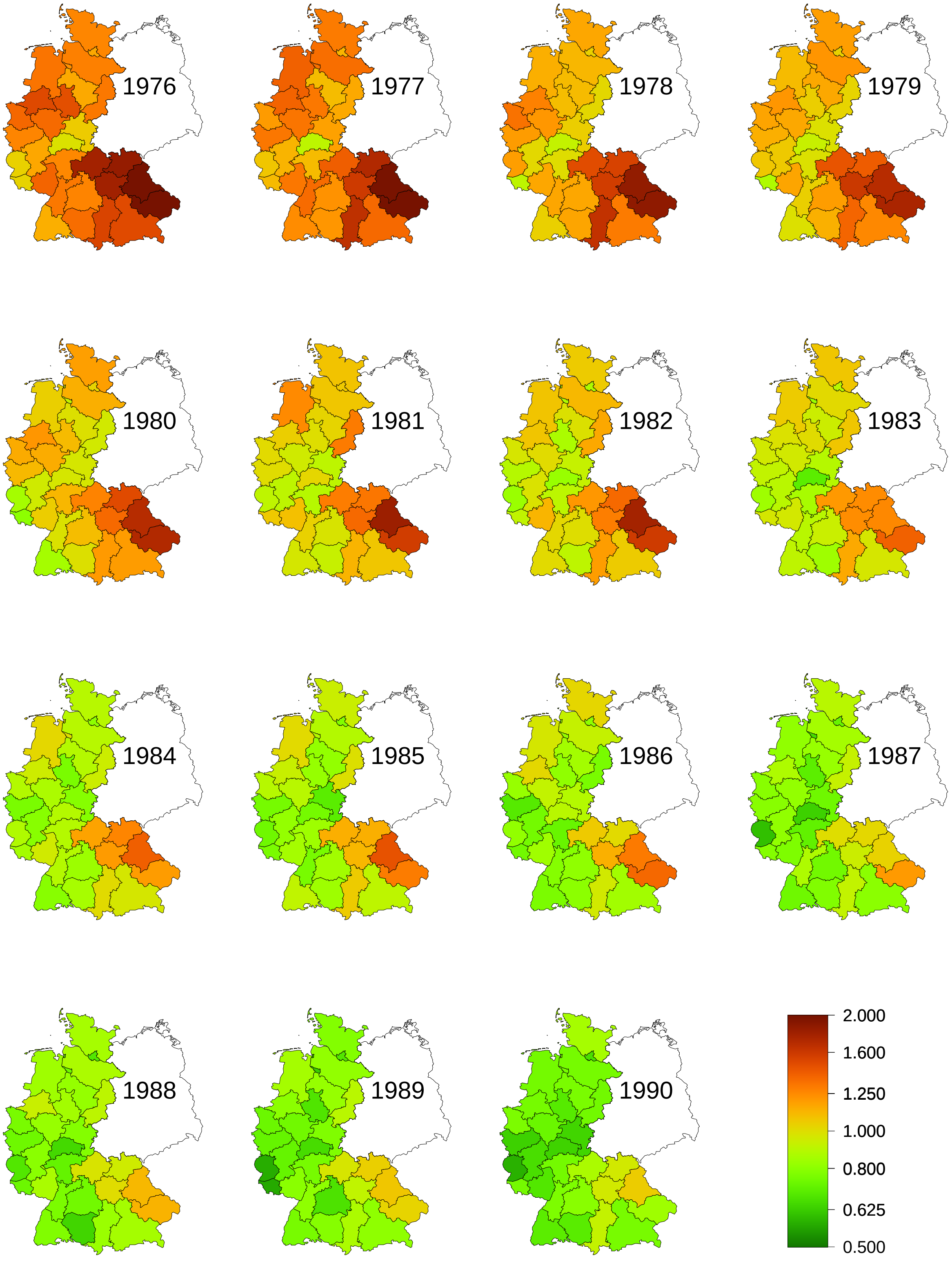


Figure 5.2: Standardized mortality ratios (SMR_1) for stomach cancer of males in West Germany.

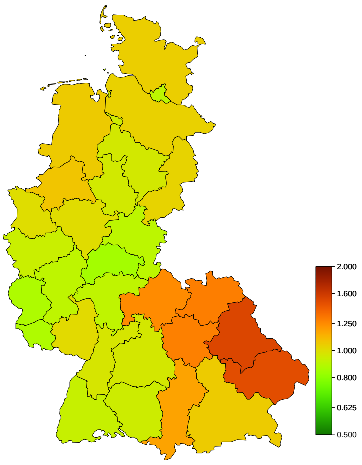


Figure 5.3: Posterior median estimates of the spatial effect for model 1 (main effects model).

1990 display rather constant risk below 1 for most regions. Extremely high risks are observed at the beginning of the observation period. This development over time is validated by model 1. The estimated time trend is displayed in the left panel of Figure 5.4. The effect—plotted on a log scale axis—is almost linear. Simultaneously, the figure shows the temporal effect for model 2 and the corresponding SMRs. The latter were calculated by

$$\mathrm{SMR}(t) = \frac{y_t}{e_t} = \frac{\sum_{i=1}^{n} y_{it}}{\sum_{i=1}^{n} e_{it}}, \quad t = 1, \ldots, T,$$

based on the observed and expected cases, cumulated over all regions. Roughly, model 1 and model 2 detect a similar time trend. Moreover, the trend for model 1 coincides with the SMRs besides a minor vertical shift. The same holds true for the spatial effects. For easier comparison we have spared out separate maps, and the spatial effects are displayed as curves in the right panel of Figure 5.4. Note that the location of the regions in the map is completely ignored. The pattern of the estimates is roughly the same for both models and the SMRs.

Altogether, model 1 offers reasonable and expected results. This is not surprising since the information in the data is very strong. Furthermore, Figure 5.4 shows that the main effects in model 2 are also rather similar. There are only some minor deviations visible which will be discussed shortly.

Model	$\bar{D}$	p_D	DIC
model 1	574	47	621
model 2	448	100	549
model 3	544	214	758

Table 5.1: Deviance summaries of all models.

First, we take a look at the model fit in terms of the posterior deviance in Table 5.1. A mean

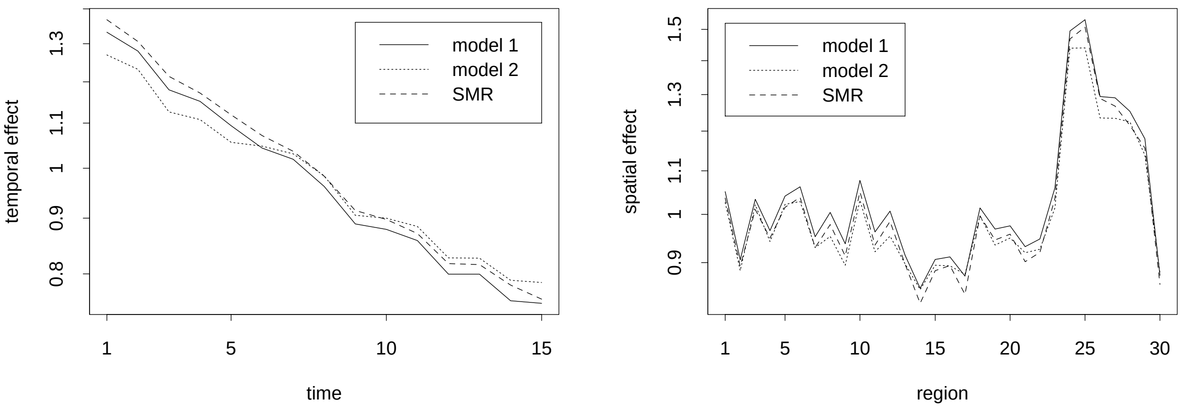


Figure 5.4: Posterior median estimates of the temporal effect (left) and the spatial effect (right) for model 1 and model 2. The corresponding SMRs are also shown. The effects are plotted on log-scale axes.

deviance of $\bar{D} = 574$ reveals an apparent lack of model fit for model 1. Together with an effective number of parameters of $p_D = 47$, this model has a moderate DIC value of 621. In contrast, model 2 offers a better performance. The deviance of $\bar{D} = 448$ indicates a remarkably better model fit. Of course, the model complexity increases, but still the advantage of an additional interaction term becomes obvious. This indicates that the variation in the data cannot be explained by main effects alone. The question arises if the residual variation is just noise or if there is some structure in it. Finally, the performance of model 3 is not convincing. Although the mean deviance is lower than for model 1, indicating a better model fit, the model complexity is enormous with an effective number of parameters of $p_D = 214$. The estimates for model 3 will be discussed later.

Before we take a look at algorithmic details, we investigate the interaction component of model 2. In Figure 5.5 the posterior median estimates are displayed. Note that these estimates are on a different scale than the SMRs in Figure 5.2. Since the main effects compensate for a large part of the variation, the interaction term covers mainly minor changes. The estimates range from 0.97 to 1.14. In the first five years the variation is rather strong with elevated risk in North Rhine-Westphalia and the south of Germany, especially Bavaria. However, this effect vanishes over time and towards the end of the observation period the effect is almost constant over the whole of Germany. This is in agreement with conclusions drawn by Becker & Wahrendorf; for Bavaria, they observe that "the differences to other parts of Germany are decreasing over time" (1997, p. 131). This can be seen from the SMRs but becomes more obvious from the interaction term. The estimated contribution of the interaction term to the relative risks is rather small, but the better model fit indicates that this effect should not be neglected. A comparison of the DIC yields the same conclusion. The DIC of 549 is superior to model 1 despite the larger number of parameters $p_D = 100$.

However, the inclusion of an additional CPM for the interaction term affects the CPMs for

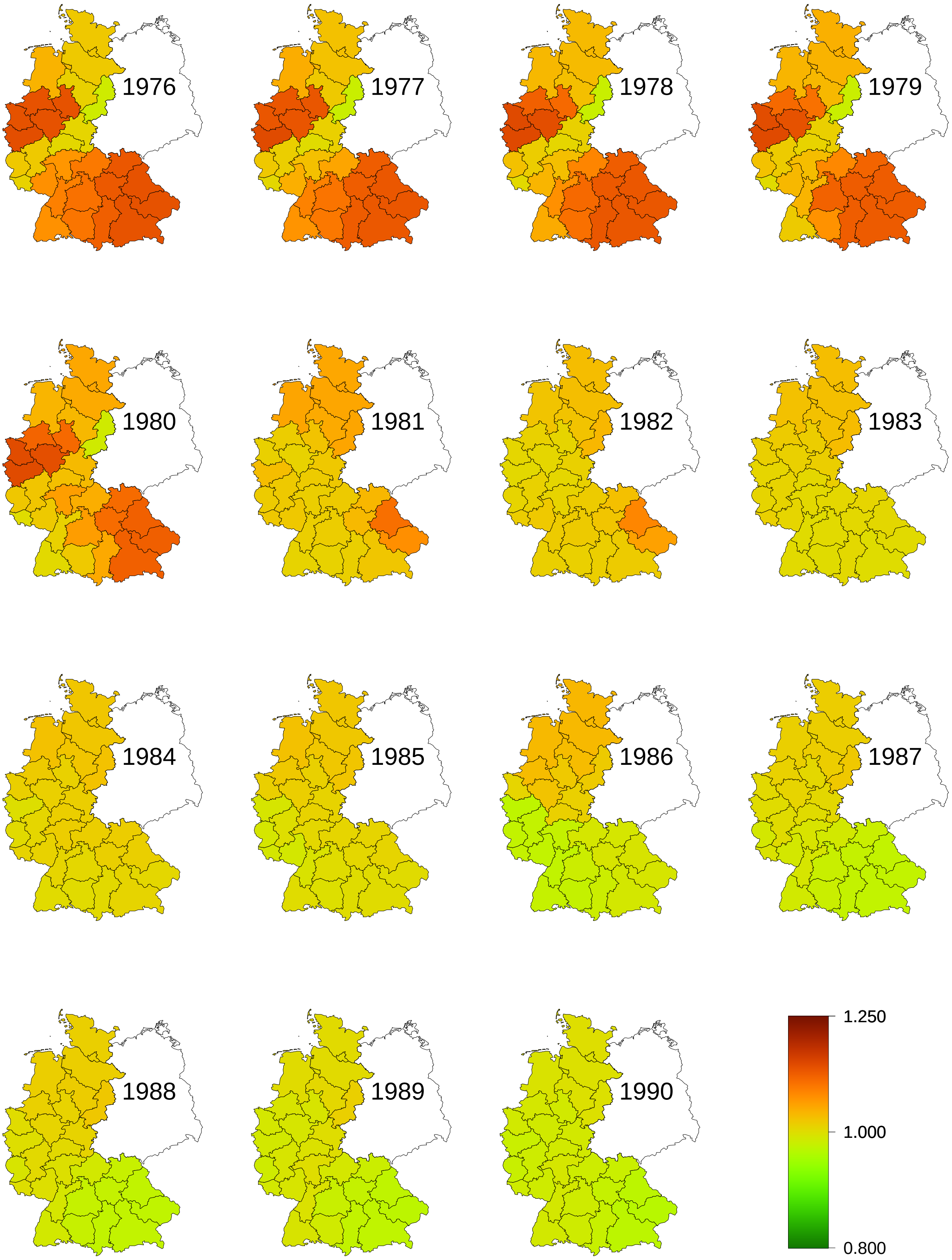


Figure 5.5: Posterior median estimates of the interaction component for model 2.

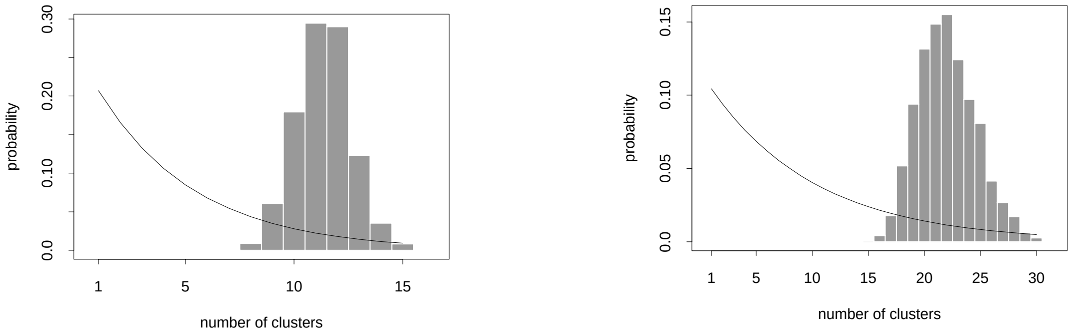


Figure 5.6: Posterior distribution of k for the temporal effect (left) and the spatial effect (right) in model 1. Prior probabilities are shown as lines.

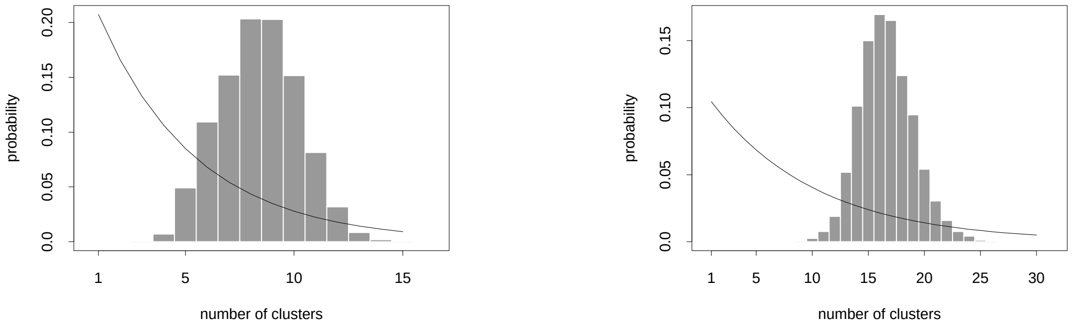


Figure 5.7: Posterior distribution of k for the temporal effect (left) and the spatial effect (right) in model 2. Prior probabilities are shown as lines.

the main effects as well. Figure 5.6 displays the posterior distribution of the number of clusters for the two main effects in model 1. For both effects the prior assumption is overruled. The algorithm requires large numbers of clusters near to the maximum numbers of T and I for the temporal and the spatial component, respectively. This corresponds to almost independent sampling of each time point or region. Hence, the estimates are similar to the SMRs. For comparison, in Figure 5.7 the posterior distribution of the number of clusters for the main effects in model 2 is shown. Clearly, the distributions are shifted to lower values. Due to the additional interaction term, the CPMs for the main effects support slightly smoother patterns. This is also visible in Figure 5.4, where the peaks in the SMRs are less emphasized in model 2 than in model 1. Still, the estimated main effects in model 2 reflect the same structure.

The overall posterior median estimates for model 2 resemble the data pretty well. In Figure 5.8 the relative risk estimates are displayed on the same scale as the data in Figure 5.2. According to the strong information in the data the maps are rather similar. Taking a closer look, the estimates are smoother, as expected. This smoothing effect is much more highlighted in model 3, see Figure 5.9. Although the basic pattern is the same, the estimates do not display smaller changes anymore. With respect to the model fit, we may draw the conclusion that model 3 is able to restore the data roughly, but smoothes to much.

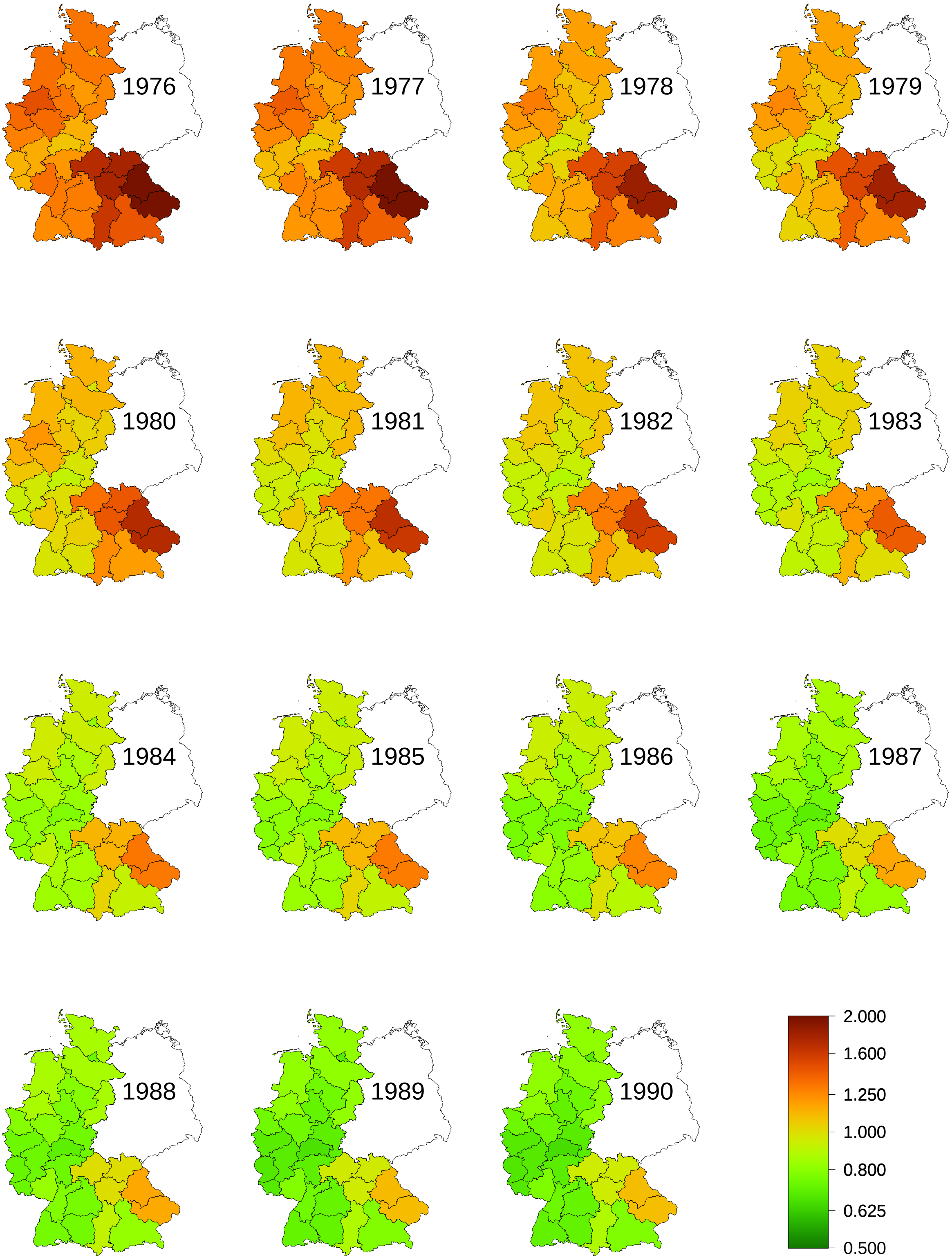


Figure 5.8: Posterior median estimates for model 2.

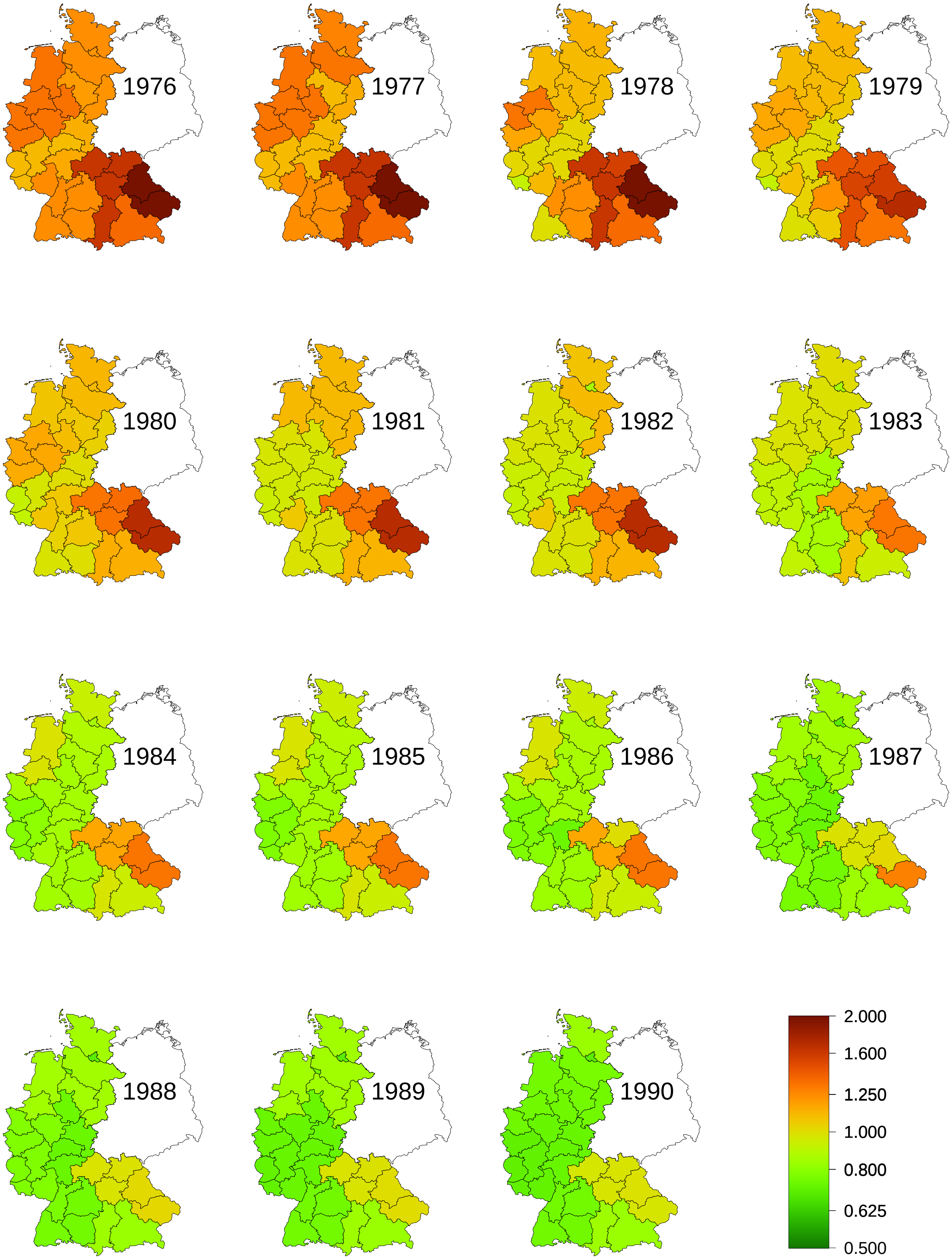


Figure 5.9: Posterior median estimates for model 3.

5.3.2 Results for SMR_2

The SMR_2 are standardized for spatial and temporal effects. This is strongly related to the results of model 1 for the SMR_1. In fact, the main effect model can be interpreted as a standardization step.

Both estimated effects in Figure 5.4 (spatial and temporal) gained with model 1 are almost identical to the exponentials of those of the logit model with which the SMR_2 were calculated. For both effects the mean absolute differences of the estimates were only about 0.03. Note that this is not necessarily the case in general. For the stomach cancer data set, there are many observed cases. Hence, the information in the likelihood assures similar results for the ML estimation (i.e. for the logit model used for internal standardization) and the Bayesian analysis (i.e. for our model 1). Still, this indicates that our prior choices in the Bayesian model were justified.

If we divide the SMR_1 by the posterior median estimates of model 1 we almost get back to the SMR_2. The mean absolute difference is only 0.007 with a maximum of only 0.03. This leads to the conclusion that one could alternatively use the SMR_2 and apply a model without main effects, i.e. model 3.

We have done an analysis of the SMR_2 with model 3. The results were similar but not identical to the estimated interaction effect for SMR_1 with model 2. For the SMR_2, the variation of the interaction term was slightly underestimated compared to model 2 for the SMR_1.

This is somewhat surprising. Still, the inclusion of an interaction term affects the main effects, as already mentioned before. In the left panel of Figure 5.4 the temporal effect for the saturated model has a lower range than the temporal effect for the main effects model. This coincides with the lower number of clusters used in model 2. Altogether, slightly more constant estimates are preferred. This leads to the effect that the time trend is slightly rotated. Higher values are decreased, while lower values are increased. Yet, the rough pattern stays the same, but less emphasized. Obviously, part of the variation of the temporal effect—and the spatial effect as well—is absorbed within the interaction component in the saturated model. This rather small shift of information from the main effect to the interaction term improves identifiability.

5.3.3 Results with modified distance

Finally, we will investigate the influence of the tuning parameter ω. As mentioned above, the interaction effect is rather small in general. For better illustration we present results for the SMR_1 gained with model 3. Although these results are not optimal in terms of model fit, the influence of the tuning parameter becomes more visible than with the saturated model.

In Figure 5.10 the posterior median estimates for Lower Bavaria are displayed. Recall that this is one of the two noticeable districts in the south-east of the map. The panel on the left shows the estimates with the default value $\omega = 1$. These estimates are identical to those in Section 5.3.1. The three panels to the right display the development of the estimates with decreasing and increasing ω, in the upper and lower row, respectively. The actual values of the tuning parameter were chosen to be symmetric on the log-scale to allow for better comparability.

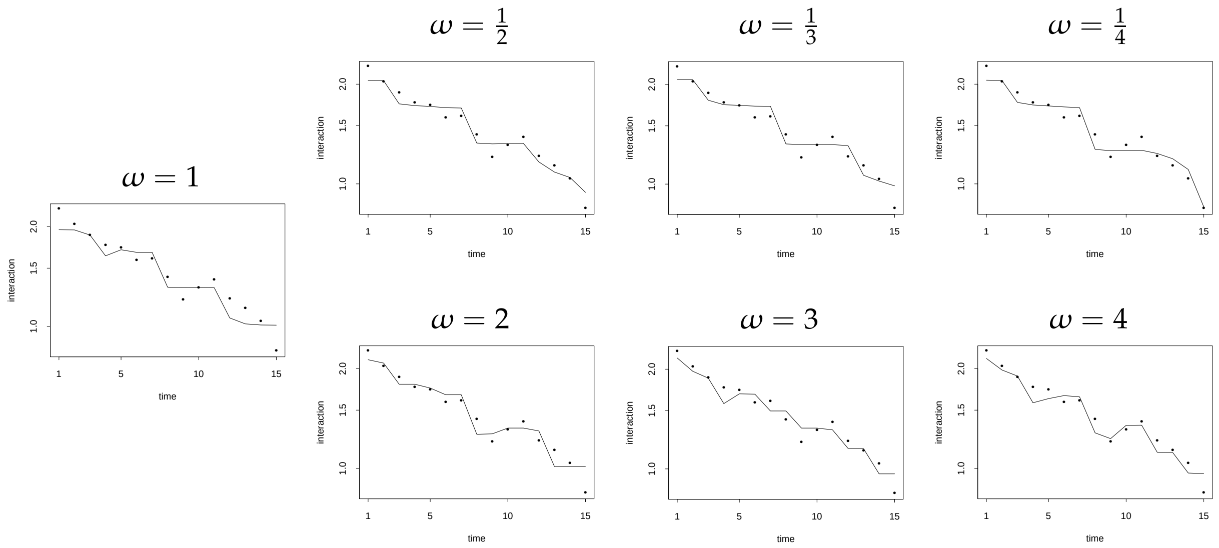


Figure 5.10: Estimated relative risks (lines) and SMRs (dots) for Lower Bavaria.

The curves show expected results. In the upper row, the time effect turns more and more into a step function. This indicates that smoothing becomes stronger for the time dimension. The values are almost constant over several time points. Simultaneously, the jumps become larger. Thus, the model concentrates all variation to few change points, but produces smooth results otherwise. The effect becomes most extreme for $\omega = \frac{1}{4}$. Here, the curve breaks at the last time point. This effect is not observable in all other curves, and most likely suppressed by information from adjacent regions.

Similar, the curve becomes rougher with increasing ω, displayed in the lower row of Figure 5.10. For the extreme case $\omega = 4$ the prior model of a step function is not visible anymore. Instead, the time points are considered almost individually.

The reverse effects are observable for the spatial variation of the risks. For $\omega > 1$ the estimates become more like a step function but get rougher for $\omega < 1$. We have spared out the corresponding maps because the differences are rather small and the color scale is not able to emphasize the effect due to the already large variation over time.

There has to be remarked that values $\omega \leq \frac{1}{3}$ and $\omega \geq 3$ are rather extreme. In fact, for only 15 time points and 30 regions, these values are too large to yield reasonable results. Still, we have included the plots for reasons of demonstration. Altogether, the modified CPM offers the possibility to adjust the prior model for different types of data. For example, if the observed counts are homogeneous over time but rather heterogeneous over the observation area, one might use values $\omega > 1$ to emphasize spatial smoothing.

The parameter ω controls the smoothing behavior of the space-time model, and can be seen as a hyperparameter of the CPM prior. Therefore, one is tempted to treat this parameter unknown as well, and estimate it from the data. In fact, the implementation of a standard MH step for ω is straightforward. For example, one might assume a log-normal prior (i.e. symmetric on the log-scale). Alternatively, one might use a discrete prior distribution since there is only a countable set (depending on the maximum distances of the spatial and the temporal graph) of values for ω that lead to potentially different partitions.

However, the CPM prior is flexible enough to adapt to the data for most (moderate) values of ω. With an increasing number of clusters k, the influence of the tuning parameter ω becomes less important. Often, the posterior distributions for ω and k will be multimodal and mixing will be poor. Therefore, we recommend to choose ω fixed with respect to the data. As a default we use $\omega = 1$ unless substantial prior knowledge suggests a different choice.

Chapter 6

Disease Mapping of Stage-specific Cancer Incidence Data

So far, we have considered count data on cancer mortality, i.e. aggregated data from binary outcomes. Throughout, we have assumed a Poisson observation model as an approximation to the binomial formulation. In this chapter we turn to cancer incidence data, where observations are available in three or more categories. Such data are rarely available, but offer various aspects of interpretation.

In this case, a Poisson approximation is no longer feasible and we will work with a multinomial model. We assume that the categories are ordered with respect to the severeness of the disease. Therefore, the proposed model is taken over from known regression models for ordinal data. Besides the spatial effect we will include age effects in the model.

The main part of this chapter, i.e. Sections 6.1 to 6.4 (pp. 108–123), contains the paper ”Disease Mapping of Stage-specific Cancer Incidence Data” by Knorr-Held, Raßer & Becker, ©The International Biometric Society, 2002. Note that some notations differ from the original version, and that the list of references is now included in the bibliography of the thesis. The paper is reprinted with kind permission from the International Biometric Society.

In contrast to the previous chapters of this thesis, the proposed model is based on MRF priors for the spatial effects and the age effects. Therefore, in Section 6.5, an equivalent model formulation is provided in terms of CPM priors. For identifiability, both effects (spatial and age) are constrained to sum up to zero. This demands for some changes in the basic sampling scheme. The results are compared to those from the MRF model. The application to stomach cancer incidence data from Germany reveals some interesting differences between the two prior models. Whereas the estimates are almost identical for informative data, the CPM prior provides more spatial structure than the MRF prior for sparse data.

Disease mapping of stage-specific cancer incidence data

Leonhard Knorr-Held

Medical Statistics Unit,
Department of Mathematics and Statistics,
Lancaster University,
Lancaster LA1 4YF,
U.K.

l.knorr-held@lancaster.ac.uk

Günter Raßer

Department of Statistics,
Ludwig-Maximilians-University Munich,
Ludwigstrasse 33,
80539 Munich,
Germany

rasser@stat.uni-muenchen.de

Nikolaus Becker

German Cancer Research Center,
Department of Biostatistics,
Im Neuenheimer Feld 280,
69120 Heidelberg,
Germany

n.becker@dkfz.de

Abstract

We propose two approaches for the spatial analysis of cancer incidence data with additional information on the stage of the disease at time of diagnosis. The two formulations are extensions of commonly used models for multicategorical response data on an ordinal scale. We include spatial and age group effects in both formulations, which we estimate in a nonparametric smooth way. More specifically, we adopt a fully Bayesian approach based on Gaussian pairwise difference priors where additional smoothing parameters are treated as unknown as well. We argue that the methods are useful in monitoring the effectiveness of mass cancer screening and illustrate this through an application to data on cervical cancer in the former German Democratic Republic. The results suggest that there are large spatial differences in the stage-proportions, which indicates spatial variability with respect to the introduction and effectiveness of Pap smear screening programs.

Key words: Cancer screening; Cervical cancer; Cumulative model; Disease mapping; Ordered categorical response; Pairwise difference prior; Sequential model; Stage-specific cancer incidence data.

6.1 Introduction

There has been much development for the spatial analysis of observational disease data within the last ten years. The work can be categorized into two groups, methodology for data where the exact location of each case is known, and methodology for aggregated data, where the total number of cases is given in predefined administrative areas, for a review see Diggle (1996). Bayesian approaches for the second type of data include the seminal work by Besag et al. (1991) who propose a Markov random field model for the spatial smoothing of disease rates. This model is nowadays widely used for "disease mapping", the study of spatial variation in disease risk, for reviews see for example Clayton & Bernardinelli (1992), Knorr-Held & Becker (2000) or Wakefield et al. (2000).

Probably the most prominent application is the statistical analysis of (age-standardized) cancer mortality rates, as such data are routinely collected throughout the world. A spatial analysis may help to identify a "spatial signal", which is particularly important for rare diseases, where the raw rates exhibit too much variation and are not particularly helpful in order to judge the variation of the underlying disease risk. The estimated spatial pattern may give hints to relevant unobserved risk factors, although some general problems of interpretation can remain due to the observational type of the data.

In this paper we extend the methodology to the analysis of cancer incidence data with additional knowledge on the stage of disease at time of diagnosis. Our aim can be described as (a) to adjust the crude observed data for effects which can be attributed to age, and (b) to assess whether there is any spatial variation left in the (adjusted) stage proportions. This is of clear public health importance for diseases for which screening programs have been implemented and spatial variation in stage proportions might indicate heterogeneity in the effectiveness of cancer screening.

We propose two formulations based on regression models for categorical data on an ordered scale (for a recent review see Fahrmeir & Tutz 2001, Ch. 3). In the first approach we model *cumulative* probabilities of disease risk, whereas in the second we model *conditional* probabilities. More specifically, in the latter approach we consider the probability that a person is diagnosed with the disease in a specific stage, given that she is diagnosed in this or in a higher stage. In each formulation, the log-odds of these (cumulative or conditional) probabilities are decomposed additively into age group and spatial effects.

We work directly on data stratified by age, which is in contrast to ordinary disease mapping methods (without stage-stratification), where the data are typically standardized by age in advance. Such a two-stage estimation procedure allows one to calculate the expected number of cases, which are subsequently used as an offset in a Poisson regression approach. However, a simultaneous estimation of age and spatial effects should in general be preferred because the uncertainty in the age estimates is then automatically incorporated. Furthermore, it is not obvious how to calculate expected cases in our multicategorical setting.

In Section 6.2 we outline the two different formulations for ordinal disease risk data, and

Section 6.3 illustrates the two approaches in an application to incidence data on cervical cancer in the former German Democratic Republic (GDR) in 1975. We compare our estimates with those obtained from a corresponding Maximum Likelihood approach with unrestricted age group and spatial effects. This corresponds to the common comparison of standardized mortality or morbidity ratios with Bayesian relative risk estimates. The results suggest that there are large spatial differences in the (age-adjusted) stage-proportions, which indicates spatial variability in the time of introduction and effectiveness of prevention programs. We close with some comments and possible extensions in Section 6.4.

6.2 Model

Let n_{ij} denote the number of person-years (or simply people) at risk in district $i = 1, \dots, I$ and age group $j = 1, \dots, J$. For each cell (i, j) let y_{ijs} denote the number of diagnosed cases of disease in stage $s = 1, \dots, S$. We assume that the stages are ordered by severity of the disease with stage S being the most severe. Finally let $y_{ij0} = n_{ij} - \sum_{s=1}^{S} y_{ijs}$ be the number of all person-years at risk, which have not being diagnosed with the disease ("stage 0"). We now assume that $\boldsymbol{y}_{ij} = (y_{ij0}, y_{ij1}, \dots, y_{ijS})$ follows a multinomial distribution with parameters n_{ij} and probability vector $\boldsymbol{\pi}_{ij} = (\pi_{ij0}, \pi_{ij1}, \dots, \pi_{ijS})$ where $\sum\limits_{s=0}^{S} \pi_{ijs} = 1$.

6.2.1 The cumulative model

In the cumulative model (McCullagh 1980) we factorize the log-odds of the *cumulative* probabilities $p_{ijs} = \pi_{ij0} + \dots + \pi_{ijs}$ into an intercept term μ_s, a spatial effect θ_{si}, and an age group effect φ_{sj}, that is

$$\operatorname{logit}(p_{ijs}) = \log\left(\frac{\sum_{t=0}^{s} \pi_{ijt}}{\sum_{t=s+1}^{S} \pi_{ijt}}\right) = \mu_s + \theta_{si} + \varphi_{sj} \qquad (s = 0, \dots, S-1). \tag{6.1}$$

Equivalently this model can be formulated in terms of *descending* cumulative probabilities $1 - p_{ijs}$; the corresponding log-odds are simply $-(\mu_s + \theta_{si} + \varphi_{sj})$. Hence the estimates from model (6.1) can easily be transformed to those corresponding to an analysis of the data with the category order reversed.

The probabilities π_{ijs} entering the multinomial likelihood can be derived from (6.1) as

$$\pi_{ijs} = \begin{cases} \operatorname{logit}^{-1}(\mu_0 + \theta_{0i} + \varphi_{0j}) & (s = 0) \\ \operatorname{logit}^{-1}(\mu_s + \theta_{si} + \varphi_{sj}) - \operatorname{logit}^{-1}(\mu_{s-1} + \theta_{s-1,i} + \varphi_{s-1,j}) & (s = 1, \dots, S-1) \\ 1 - \operatorname{logit}^{-1}(\mu_{S-1} + \theta_{S-1,i} + \varphi_{S-1,j}) & (s = S) \end{cases}, \tag{6.2}$$

where $\operatorname{logit}^{-1}(x) = 1/(1 + \exp(-x))$. To ensure that all these probabilities are positive, the unknown parameters μ_s, θ_{si} and φ_{sj} have to fulfill the constraints

$$\mu_{s-1} + \theta_{s-1,i} + \varphi_{s-1,j} < \mu_s + \theta_{si} + \varphi_{sj} \tag{6.3}$$

for all $i = 1, \dots, I$, $j = 1, \dots, J$, and $s = 1, \dots, S-1$.

6.2.2 The sequential model

The rationale for the sequential model is that a categorical response variable can take a specific value s only after the levels $0, 1, \ldots, s-1$ have been reached. This is the case in our application, where cancer diagnosis at a specific stage implies that the cancer has passed undetected through all stages below. A version of the sequential model where covariate effects do not depend on the response value is also known as the continuation ratio model (Agresti 1984).

The sequential approach to ordinal data hence models the *conditional* probability that an individual in cell (i, j) gets diagnosed of the disease in stage s, *assuming* that she gets diagnosed of the disease in stage s or higher, i.e. $q_{ijs} = \pi_{ijs}/(\pi_{ijs} + \ldots + \pi_{ijS})$. Now we decompose the log-odds of these conditional probabilities into an intercept term ν_s, a spatial effect η_{si}, and an age group effect ψ_{sj}

$$\operatorname{logit}(q_{ijs}) = \log\left(\frac{\pi_{ijs}}{\sum_{t=s+1}^{S} \pi_{ijt}}\right) = \nu_s + \eta_{si} + \psi_{sj} \qquad (s = 0, \ldots, S-1). \tag{6.4}$$

Note that, formally, the only difference to the cumulative model (6.1) is that π_{ijs} replaces the cumulative probability $\pi_{ij0} + \ldots + \pi_{ijs}$ in the numerator of the ratio within the logarithm. For $s = 0$, the cumulative and the sequential model are apparently identical (assuming compatible priors). We will comment on this further in Section 6.2.5.

The probabilities π_{ijs} can now be derived as

$$\pi_{ijs} = \begin{cases} \operatorname{logit}^{-1}(\nu_0 + \eta_{0i} + \psi_{0j}) & (s = 0) \\ \operatorname{logit}^{-1}(\nu_s + \eta_{si} + \psi_{sj}) \cdot \prod\limits_{t=0}^{s-1} \{1 - \operatorname{logit}^{-1}(\nu_t + \eta_{ti} + \psi_{tj})\} & (s = 1, \ldots, S-1) \\ \prod\limits_{t=0}^{S-1} \{1 - \operatorname{logit}^{-1}(\nu_t + \eta_{ti} + \psi_{tj})\} & (s = S) \end{cases}, \tag{6.5}$$

e.g. Fahrmeir & Tutz (2001, p. 94). Note that here the π_{ijs} are defined through products of probabilities, not through differences of probabilities as in the cumulative model. Therefore no further constraints have to be imposed on the parameters ν_s, η_{si}, and ψ_{sj}. A further difference to the cumulative model is that a sequential model applied to the data but with the category order reversed is not equivalent to model (6.4), except for the non-interesting binomial case $S = 1$. This is a consequence of the rationale underlying the sequential model where categories can be reached successively, but only in one specific direction.

6.2.3 Prior assumptions

The two alternative models proposed above are now completed by assigning prior distributions to all unknown parameters. For both the spatial and the age group parameters we will use Gaussian pairwise difference priors (Besag et al. 1995) which favor a nearly constant pattern, implied by a high prior mass on very small values of the corresponding variance parameter. However, the priors we use for these variance parameters are highly dispersed, hence the formulation will be flexible enough to capture spatial or temporal gradients or trends if there is evidence in the data for it. For the spatial effects, this corresponds to the common choice of Markov random field models while for the age-group parameters this class reduces to so-called random walk priors. Exactly the same priors have been used for disease mapping (Besag et al. 1991, Best, Arnold, Thomas, Waller & Conlon 1999), for space-time modeling of disease risk (Knorr-Held & Besag 1998) and in many other areas of application (e.g. Fahrmeir & Lang 2001). These models neither impose stationarity nor assume a specific parametric form; in fact they are closely related to non- and semiparametric smoothing methods, see Fahrmeir & Knorr-Held (2000) and Hastie & Tibshirani (2000).

In the cumulative model, we separate the spatial parameters into independent sets $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{S-1}, s = 0, \ldots, S-1$, where $\boldsymbol{\theta}_s = (\theta_{s1}, \ldots, \theta_{sI})$ and assume that, for each category s, $\boldsymbol{\theta}_s$ follows a Gaussian Markov random field (MRF) model (Besag et al. 1991)

$$p(\boldsymbol{\theta}_s|\lambda_{\theta_s}) \propto \lambda_{\theta_s}^{(I-1)/2} \cdot \exp\left\{ -\frac{\lambda_{\theta_s}}{2} \sum_{i_1 \sim i_2} (\theta_{s,i_1} - \theta_{s,i_2})^2 \right\}, \tag{6.6}$$

where the sum in the exponent goes over all pairs of adjacent areas i_1 and i_2. For some motivation for $I-1$ instead of I degrees of freedom for the precision (the inverse variance) λ_{θ_s} in (6.6) see Knorr-Held (2003).

For each unknown precision parameter $\lambda_{\theta_s}, s = 0, \ldots, S-1$, we adopt a gamma prior

$$p(\lambda_{\theta_s}) \propto \lambda_{\theta_s}^{a-1} \cdot \exp(-b\lambda_{\theta_s})$$

with suitably chosen constants a and b. The S sets of Markov random fields $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{S-1}$ are assumed to be independent. Alternatively one could specify a *multivariate* MRF model

$$p(\boldsymbol{\theta}|\boldsymbol{\Lambda}_\theta) \propto |\boldsymbol{\Lambda}_\theta|^{(I-1)/2} \cdot \exp\left\{ -\frac{1}{2} \sum_{i_1 \sim i_2} (\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})' \boldsymbol{\Lambda}_\theta (\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2}) \right\}, \tag{6.7}$$

where $\boldsymbol{\theta}_i = (\theta_{0i}, \ldots, \theta_{Si})'$. A Wishart prior would be the common choice for the precision matrix $\boldsymbol{\Lambda}_\theta$, i.e.

$$p(\boldsymbol{\Lambda}_\theta) \propto |\boldsymbol{\Lambda}|^{a-(S+1)/2} \exp\left\{-\mathrm{tr}(\boldsymbol{B} \cdot \boldsymbol{\Lambda})\right\},$$

again with suitably chosen constants a and $\boldsymbol{B}$, where a is a scalar and $\boldsymbol{B}$ is a $S \times S$-matrix. Such a multivariate MRF model might be appropriate if the MRFs $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{S-1}$ are expected to be correlated. However, note that a priori independent fields $\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_{S-1}$ can still be dependent in the posterior if they are dependent in the likelihood. So—without strong prior knowledge about the correlation—we do not expect any major differences between the two formulations and stick to the simpler form with a priori independent MRFs.

The formulation proposed in Besag et al. (1991) is more elaborate with additional parameters for *unstructured* spatial heterogeneity. It is computationally convenient to employ a reparametrized version (e.g. Carlin & Louis 1996, p. 308), where θ_{si} is independent Gaussian with mean $\tilde{\theta}_{si}$ and precision τ_s, say, and a GMRF prior is now placed on the latent vectors $\tilde{\boldsymbol{\theta}}_s$, just like in (6.6) for $\boldsymbol{\theta}_s$. In our application we have tested both models with and without the additional unstructured parameters.

For the age-group specific parameters, we assume in similar lines that, for each category s, the parameters $\boldsymbol{\varphi}_s = (\varphi_{s1}, \ldots, \varphi_{sJ})$ follow a simple Gaussian random walk in time with variance $\lambda_{\varphi_s}^{-1}$, with a flat prior for the initial value φ_{s1}. Such a formulation is the exact temporal analogue of model (6.6) as the prior can be written again in the pairwise difference form:

$$p(\boldsymbol{\varphi}_s|\lambda_{\varphi_s}) \propto \lambda_{\varphi_s}^{(J-1)/2} \cdot \exp\left\{ -\frac{\lambda_{\varphi_s}}{2} \sum_{j=2}^{J} (\varphi_{s,j} - \varphi_{s,j-1})^2 \right\}. \tag{6.8}$$

We assume prior independence for the sets of parameters $\boldsymbol{\varphi}_0, \ldots, \boldsymbol{\varphi}_{S-1}$, which again can easily be relaxed by adopting a multivariate Gaussian random walk model. Also, we use again gamma hyperpriors for the precision parameters λ_{φ_s}, $s = 0, \ldots, S-1$. Finally, for each intercept parameter $\mu_0, \ldots, \mu_{S-1}$ we adopt a flat, locally uniform prior.

Similarly, for the sequential model we use MRF priors for $\boldsymbol{\eta}_s = (\eta_{s1}, \ldots, \eta_{sI})$, random walk priors for $\boldsymbol{\psi}_s = (\psi_{s1}, \ldots, \psi_{sJ})$, and a flat prior for ν_s, $s = 0, \ldots, S-1$. The exact forms of the prior densities can easily be obtained by replacing μ_s by ν_s, $\boldsymbol{\theta}_s$ by $\boldsymbol{\eta}_s$, and $\boldsymbol{\varphi}_s$ by $\boldsymbol{\psi}_s$ in the above description of the priors in the cumulative model.

6.2.4 Model choice and parameter interpretation

At this point it might be worth noting that the posterior distribution of the *conditional* probabilities q_{ijs} can of course easily be derived from the *cumulative* model as well, as they are just simple functions of the posterior distribution of the π_{ijs}'s. Similarly, the posterior distribution of the cumulative probabilities p_{ijs} could be calculated from the sequential model. Indeed, both formulations allow the exploration of every functional of the posterior distribution of the π_{ijs}'s. The difference between the two formulations is the different parametrization of the π_{ijs}'s with different quantities being the focus for smoothing, either the cumulative or the sequential conditional probabilities. Preferences for one or the other model can either be based on interpretation issues or on more formal model choice criteria.

Regarding parameter interpretation, we are particularly interested in spatial disease risk estimates, adjusted for age. The (age-adjusted) *overall relative risk* (regardless of the stage of the disease) in district i can be obtained from the quantities $\exp(-\theta_{0i})$ and $\exp(-\eta_{0i})$. Similarly, in the cumulative model we can interpret $\exp(-\theta_{si})$, $s = 1, \ldots, S-1$, as the *cumulative adjusted relative risk* in district i. In the sequential model, $\exp(-\eta_{si})$, $s = 1, \ldots, S-1$, can be interpreted as the (age-adjusted) *odds ratio* for the conditional probability of being diagnosed in stage $s+1$ or higher, given diagnosis in stage s or higher. For the age group effects we also prefer to display $-\varphi_{sj}$ and $-\psi_{sj}$ (rather than φ_{sj} and ψ_{sj}), the age group effects on the cumulative probabilities $1 - p_{ijs}$ and on the conditional probabilities $1 - q_{ijs}$ respectively. This has the advantage that higher values in the figures displaying age effects, and darker colors in the spatial maps, can be associated with a higher (cumulative or conditional) risk of a more severe stage of the disease at diagnosis.

For assessment of the model fit, we routinely monitor the posterior distribution of the *satu-*

rated deviance (Spiegelhalter et al. 2002)

$$D = \sum_{i=1}^{I} \sum_{j=1}^{J} d_{ij}^2 \tag{6.9}$$

with the multinomial squared deviance residual

$$d_{ij}^2 = 2 \cdot \sum_{s=0}^{S} y_{ijs} \log \left(\frac{y_{ijs}}{n_{ij}\pi_{ijs}} \right) \tag{6.10}$$

(using the convention that $0 \log 0 = 0$). Each deviance residual $d_{ij} = \sqrt{d_{ij}^2}$ can be seen as a (standardized) measure of fit, comparing the observed number of cases y_{ijs} with the fitted number of cases $n_{ij}\pi_{ijs}$ for all stages $s = 0, \dots, S$. Note that this is well defined in both models, as only the multinomial cell probabilities enter. For a well fitting model, D should be asymptotically (with increasing data in each cell (i, j)) around $I \cdot J \cdot S$ (the factor S appears here due to the multinomial response with S "free" categories). The mean posterior deviance $\bar{D}$ can be used as an overall measure of model fit and can be combined with a term p_D called "the effective number of parameters" to give a *deviance information criterion* (DIC) for model choice, see Spiegelhalter et al. (2002) for further details.

6.2.5 A comparison of the two models

As an illustration, we now consider a simple example with $S = 2$ categories and no further stratification with respect to age or space (i.e. $I = J = 1$).

The difference between the two models is a different parametrization of the multinomial probabilities $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$: The cumulative model parametrizes the model with respect to cumulative probabilities $p_0 = \pi_0$ and $p_1 = \pi_0 + \pi_1$ with $p_0 < p_1$. The sequential model uses $q_0 = \pi_0$ and the conditional probability $q_1 = \pi_1/(1 - \pi_0)$. Suppose now we use independent flat $\text{Beta}(1, 1)$ priors for p_0 and p_1 in model 1, or q_0 and q_1 in model 2 respectively. No attempt is made here to choose *compatible priors* (Dawid & Lauritzen 2001); the following discussion holds for any prior choice as long as the priors are assumed to be independent (note however that the order restriction $p_0 < p_1$ already implies a dependence between p_0 and p_1 in the cumulative model).

In the sequential model it can now easily be seen that, conditional on the data, q_0 and q_1 are still independent, because the posterior is proportional to the multinomial likelihood

$$p(q_0, q_1|\boldsymbol{y}) \propto q_0^{y_0}(q_1(1 - q_0))^{y_1}\{(1 - q_0)(1 - q_1)\}^{y_2} = q_0^{y_0}(1 - q_0)^{y_1+y_2} \cdot q_1^{y_1}(1 - q_1)^{y_2} \tag{6.11}$$

which can be factorized into independent Beta terms. Therefore q_0 and q_1 are independent in the posterior with marginal distribution

$$q_0|\boldsymbol{y} \sim \text{Beta}(y_0 + 1, y_1 + y_2 + 1) \quad \text{and} \quad q_1|\boldsymbol{y} \sim \text{Beta}(y_1 + 1, y_2 + 1).$$

In the cumulative model, however, the posterior

$$p(p_0, p_1|\boldsymbol{y}) \propto p_0^{y_0}(p_1 - p_0)^{y_1}(1 - p_1)^{y_2} \text{ for } p_0 < p_1 \quad \text{and} \quad 0 \text{ elsewhere}$$

cannot be factorized, and p_0 and p_1 will be dependent. Furthermore, although $p_0 = q_0 = \pi_0$, the marginal posterior distribution of p_0

$$p_0|\boldsymbol{y} \propto p_0^{y_0} \int_{p_0}^{1} (p_1 - p_0)^{y_1} (1 - p_1)^{y_2} dp_1$$

is different from the posterior for q_0 and does not seem to be analytically tractable. We can easily sample from the posterior $p(p_0, p_1|\boldsymbol{y})$, for example by Markov chain Monte Carlo, and compare the corresponding multinomial probabilities $\boldsymbol{\pi}$ with the one obtained from the sequential model. In some empirical comparisons we have found slight differences for the posterior distribution of π_0 and stronger discrepancies for π_1 and π_2.

This simple example transfers to the general case: Although both, the cumulative and the sequential model, specify the same model for the probability π_{ij0} of not developing the disease, the posterior distributions of μ_0 and ν_0, θ_{0i} and η_{0i}, and φ_{0j} and ψ_{0j} are not necessarily exactly the same—except for the binomial case $S = 1$—because of the different parametrization of the remaining multinomial probabilities $(\pi_{ij1}, \ldots, \pi_{ijS})$.

Incidentally, the maximum likelihood estimates will be the same in both models due to the invariance property of such estimates with respect to reparametrization (e.g. Cox & Hinkley 1974). For example, in the above example the ML estimate for p_0 and q_0 is $y_0/(y_0 + y_1 + y_2)$, while p_1 is estimated by $(y_0 + y_1)/(y_0 + y_1 + y_2)$ and q_1 is estimated by $y_1/(y_1 + y_2)$.

Returning to the factorization (6.11) we note that the same independence structure holds also in the general sequential model and implies that we could—equivalently to the joint multinomial approach defined by (6.4) and (6.5)—estimate S binomial regression models

$$\begin{aligned}
y_{ij0} &\sim \text{Bin}(n_{ij}, \text{logit}^{-1}(\nu_0 + \eta_{0i} + \psi_{0j})) \\
y_{ij1} &\sim \text{Bin}(y_{ij1} + \ldots + y_{ijS}, \text{logit}^{-1}(\nu_1 + \eta_{1i} + \psi_{1j})) \\
&\vdots \\
y_{ij,S-1} &\sim \text{Bin}(y_{ij,S-1} + y_{ijS}, \text{logit}^{-1}(\nu_{S-1} + \eta_{S-1,i} + \psi_{S-1,j}))
\end{aligned}$$

completely separately. This factorization in fact reflects explicitly the conditional definition of the model. Hence, there will be no information in the likelihood about correlation between parameters for different stages. In particular, the extension to multivariate MRF and random walk priors as discussed earlier for the cumulative model does not seem to be useful here. A separate modeling approach might be advantageous if one is mainly interested in the variation of the stage-specific proportions, but not in the overall disease rate. Note that then the actual number of person-years n_{ij} is not even needed for such an analysis. This in fact opens up the possibility for *continuous* spatial modeling of the risk surface (for a similar non-Bayesian approach for spatial case-control data see Kelsall & Diggle 1998), if the exact locations of disease cases were known.

Finally, the factorization (6.11) implies that the posterior distribution of ν_0, η_{0i}, and ψ_{0j} will be the same, whether or not we further stratify by the cancer stages. This would not be exactly the case in the cumulative model.

6.2.6 Computational issues

Inference has been carried out using `C++` routines developed by the first author. We have used Markov chain Monte Carlo (MCMC) to sample from the relevant posterior distributions, applying univariate Gaussian Metropolis random walk proposals for all components of $\boldsymbol{\theta}_s$ ($\boldsymbol{\eta}_s$) and $\boldsymbol{\varphi}_s$ ($\boldsymbol{\psi}_s$), $s = 0, \dots, S-1$, while Gibbs steps have been used for the remaining precision parameter. The spread of each Metropolis proposal was tuned in an automatic fashion—prior to the collection of the posterior samples—so that the corresponding acceptance rate for each parameter was between 35 and 45%. Note that in the cumulative model one needs to check the additional restriction (6.3). If the Metropolis proposal did not fulfill the restriction it was simply rejected (formally due to a zero prior term in the numerator of the acceptance ratio).

Both formulations impose an identifiability problem on the overall risk parameter μ_s (ν_s), as those can also be absorbed by both age group and spatial effects. We have recentered both $\boldsymbol{\theta}_s$ ($\boldsymbol{\eta}_s$) and $\boldsymbol{\varphi}_s$ ($\boldsymbol{\psi}_s$) after each iteration with a corresponding adjustment to μ_s (ν_s) for $s = 0, \dots, S-1$. This is a valid approach as long as we assume a locally uniform prior for μ_s (ν_s), because it neither changes the value of the likelihood, nor of the prior (all pairwise difference priors have an implicit flat prior on the overall level), hence not of the posterior. Furthermore, it enables us to explore the posterior distribution of the age and spatial effects. Alternatively, one could impose a sum-to-zero restriction directly in the prior for each age group and spatial parameter block. However, one would need to implement a block updating algorithm, as for example suggested in Rue (2001), because single-site updating would be impossible due to degenerate full conditionals. Block updating would also be helpful for sparse data, where similar models are known to have convergence and mixing problems (Knorr-Held & Rue 2002). However, the data we considered in our application are not particularly sparse and MCMC mixing was fine for the single-site scheme we have implemented.

We finally note that Albert & Chib (1993, 2001) suggested a latent variable approach for Bayesian inference by MCMC both in the cumulative and sequential model. This can be advantageous in applications where the number of observations is small or moderate. However, in the current context the number of latent variables will be equal (in the cumulative model) or even a multiple (in the sequential model) of the number of person-years at risk. This seems to be prohibitive; for example, in our application the number of person-years, which is here simply the population number, exceeds seven millions.

6.3 Application

We now describe an application of the methodology described above to incidence data on cervical cancer in the former German Democratic Republic (GDR). The data is available on a yearly basis; here we present results for the year 1975, shortly after the introduction of Pap smear screening programs. We have used the values $a = 1.0$ and $b = 0.001$ as a default choice for the gamma hyperprior of all precision parameters, which corresponds to an extremely dispersed distribution for the (inverse gamma distributed) variances with infinite mean and variance and a prior mode at 0.0005.

The data are stratified by $I = 216$ administrative districts and $J = 15$ age groups (15–19, 20–24, . . ., 80–84 and 85+). There were no cases below age 15. The original records give information on the stage of the detected lesion in 6 categories: (I) dysplasia, (II) carcinoma in situ (both premalignant) and (III–VI) malignant cancer of increasing severity. Effective screening shifts (a) the stage of the detected lesion towards earlier stages, preferentially to a premalignant condition, and (b) the time of detection towards younger age groups. Here we focus on the effect of stage shift and combine for simplicity the premalignant categories I and II into stage $s = 1$. Similarly we aggregate the malignant categories III–VI into stage $s = S = 2$. We have deleted 35 cases (0.5%) with missing information on the stage of the disease. The total number of cases sum up to 3,466 in stage 1 and 3,540 in stage 2; the corresponding total female population in the 15 age groups is 7,262,311. The median number of cases per district (regardless of the stage) is 20.5 (range 3–759). Stage-specific medians are 9 (0–433) for stage 1 and 11 (1–326) for stage 2.

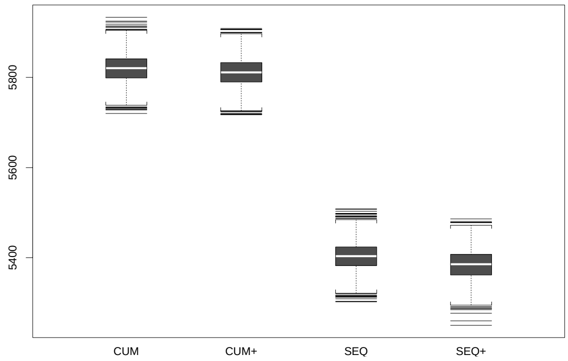


Figure 6.1: Boxplots of posterior samples from the deviance for the four different models.

In a first assessment of the model fit, Figure 6.1 compares the posterior distribution of the deviance (6.9) of the cumulative and the sequential model; both of them either without (denoted by CUM and SEQ) or with (denoted by CUM+ and SEQ+) additional unstructured random effects. Among the simpler formulations without additional unstructured random effects, the sequential model fits the data better than the cumulative model as the mean posterior deviance is is smaller (5,403 compared to 5,820) and the ranges of the posterior deviance samples of the two models are well separated. Compared to the actual number of cells times the number of stages ($I \cdot J \cdot S = 216 \cdot 15 \cdot 2 = 6,480$) this seems to be a decent fit to the data and indicates that neither interactions of age with space nor additional unstructured parameters are needed in both formulations. Indeed, the more complex formulations with additional parameters for unstructured heterogeneity give only a minor improvement in model fit, with a slightly smaller mean posterior deviance of 5,812 for the cumulative and 5,384 for the sequential model.

In the following we therefore restrict our attention to the formulations without the additional unstructured parameters. We note, however, that the DIC criterion has a slight preference for the sequential model with additional unstructured effects, see Table 6.1. It is unclear if such a small difference in DIC really matters. One would also like to ensure that this difference is not due to Monte Carlo error, in particular the assessment of the Monte Carlo error of p_D is difficult (Spiegelhalter et al. 2002). Fortunately, in our application all maps and figures are virtually indistinguishable so our conclusions are the same with or without the additional unstructured effects.

Model	$\bar{D}$	p_D	DIC
CUM	5820	130	5950
CUM+	5812	129	5941
SEQ	5403	245	5649
SEQ+	5384	259	5644

Table 6.1: Deviance summaries

First we compare the mean deviance residuals d_{ij}. Overall, 69% of the residuals from the sequential model are smaller than the corresponding ones from the cumulative model (see also Figure 6.2 for a graphical comparison), but no general pattern could be observed, that would indicate the lack of fit of the cumulative model in particular age groups or districts, say. One is tempted to study the deviance residuals further stratified by stage, but this does not proof useful, because stage-specific contributions $y_{ijs} \log \left(y_{ijs} / (n_{ij} \pi_{ijs}) \right)$ can be large in absolute size, although their sum d_{ij}^2 may still be small.

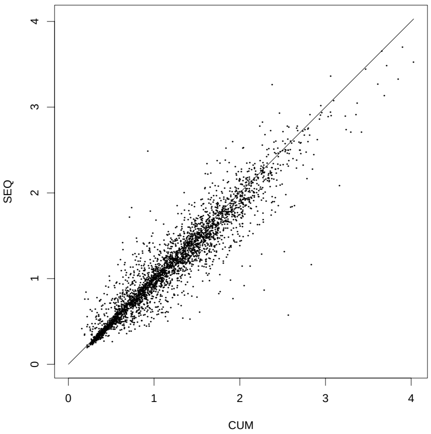


Figure 6.2: Estimated mean deviance residuals from the sequential model (y-axis) plotted against the corresponding ones from the cumulative model (x-axis).

Turning now to the estimated age effects, Figure 6.3 displays posterior median estimates within 90% pointwise credible intervals of $-\varphi_0$ and $-\varphi_1$ from the cumulative model. One can see a fairly similar inverse "bathtub" pattern of the two curves. The second curve, which describes the age pattern relevant for being diagnosed with a malignant form of the disease has a nearly constant slope for age between 30 and 70 whereas the slope of the first curve, representing the log relative risk for both the premalignant and malignant stage, is already negative in that age range. This reflects the fact that the malignant stage of cervical cancer is more likely to be diagnosed in older age groups, as the cancer needs time to progress (undetected) through the premalignant stage.

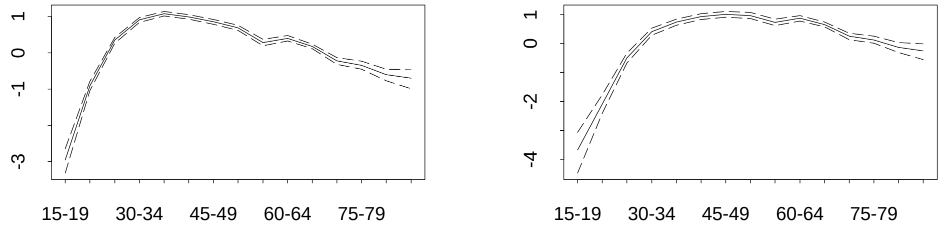


Figure 6.3: Estimated median age effects of $-\boldsymbol{\varphi}_0$ (left plot) and $-\boldsymbol{\varphi}_1$ (right plot) within 90% pointwise credible intervals from the cumulative model.

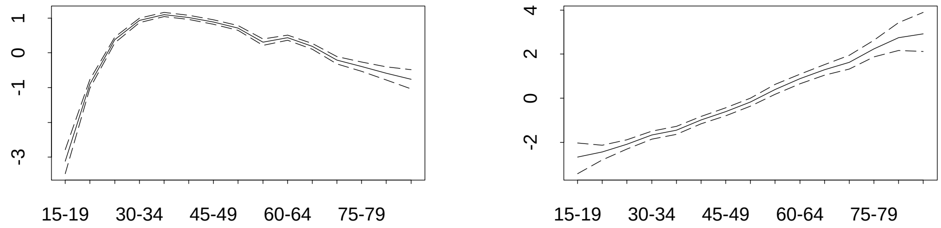


Figure 6.4: Estimated median age effects of $-\boldsymbol{\psi}_0$ (left plot) and $-\boldsymbol{\psi}_1$ (right plot) within 90% pointwise credible intervals from the sequential model.

The estimates of $-\boldsymbol{\psi}_0$ from the sequential model (Figure 6.4, left plot) are directly comparable to $-\boldsymbol{\varphi}_0$ (Figure 6.3, left plot), as both correspond to the overall log relative disease risk (keep in mind, however, that the estimates do not have to be exactly identical, as commented earlier). Here, there is virtually no difference to see. Finally, the right plot in Figure 6.4 displays the age effect on the conditional risk of the malignant disease stage 2, given a diagnosis in stage 1 or 2. As expected, an increasing conditional risk with increasing age can be seen, which is remarkably linear on the logit scale.

Figure 6.5 now displays the estimated spatial incidence pattern, regardless of the stage. The first map shows Standardized Morbidity Ratios (SMRs) calculated by internal standardization through *joint* maximum likelihood (ML) estimation; see Breslow & Day (1987, Ch. 4). More specifically, we obtained the SMRs by applying a standard logistic regression procedure to the aggregated cases in stage 1 and 2 as responses, using age group and district as factors (each of them restricted to sum up to zero). Displayed is the exponential of the estimated spatial parameters, which can hence be interpreted as (age-adjusted) relative risk estimates. The other two maps display the corresponding (posterior median) relative risk estimates $\exp(-\boldsymbol{\theta}_0)$ and $\exp(-\boldsymbol{\eta}_0)$ from the cumulative and sequential model respectively. One can see a fairly simi-

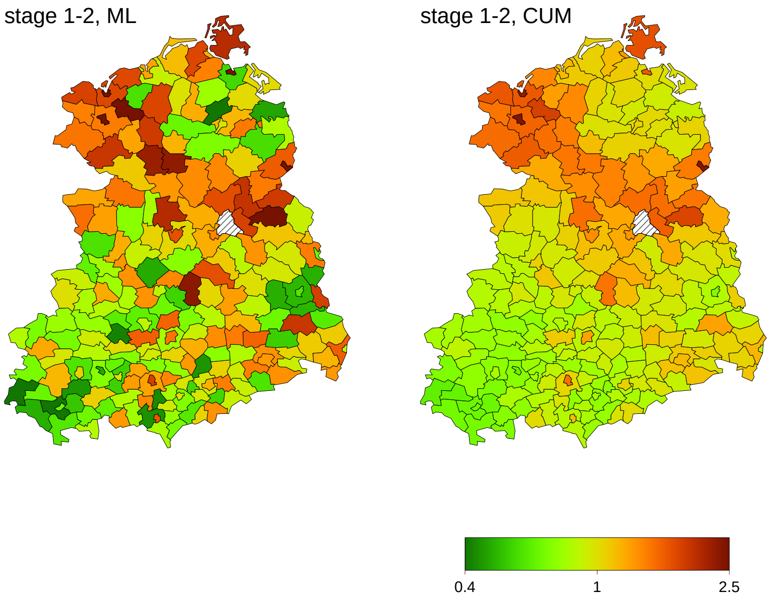


Figure 6.5: Relative risk estimates for diagnosis of the disease regardless of the stage by ML (left map), the cumulative (middle map) and the sequential model (right map).

lar pattern with the expected smoothing effect, slightly more pronounced for the cumulative model. This might be caused by the additional order restrictions (6.3). Note that we have used the same scale from 0.4 to 2.5 in all maps, which covers the estimates from the cumulative model (range 0.64–2.39), but not all of the SMRs (0.35–3.19) nor all of the estimates obtained from the sequential model (0.49–2.68). The range was chosen in order to make the spatial pattern in the smoothed maps more visible.

Figure 6.6 now displays—on the same scale as Figure 6.5—estimates of the relative risk of a tumor diagnosis in the malignant stage 2 of the disease. The left map gives ML estimates, calculated just as in Figure 6.5, but only with the cases in stage 2 as responses. The other map displays the median relative risk estimates $\exp(-\boldsymbol{\theta}_1)$ from the cumulative model. There is less spatial variation than for the overall risk $\exp(-\boldsymbol{\theta}_0)$ (Figure 6.5, middle map), with slightly higher values east of West-Berlin (the hatched region in the middle of the map).

Finally, Figure 6.7 (right map) gives the estimated odds ratio $\exp(-\boldsymbol{\eta}_1)$ from the sequential model for the probability of a diagnosis in a malignant stage of the disease, conditional on a diagnosis in stage 1 or 2. For comparison, the left map displays the corresponding ML estimates. These have considerably more variation, in fact the district-specific ML estimates did not even exist for 7 out of the 216 districts, due to no observations in stage 1. The smoothed map shows higher conditional risk of stage 2 in the south-west, and lower conditional risk in the north-east and some other parts of the country. This corresponds roughly to what is known about the local introduction of cervical cancer screening programs: Cervical cancer screening by Pap smear has been first introduced in the former GDR as a pilot project in two specific regions in 1974: East-Berlin and Mecklenburg-West Pomerania (northern coastal region). Available infor-

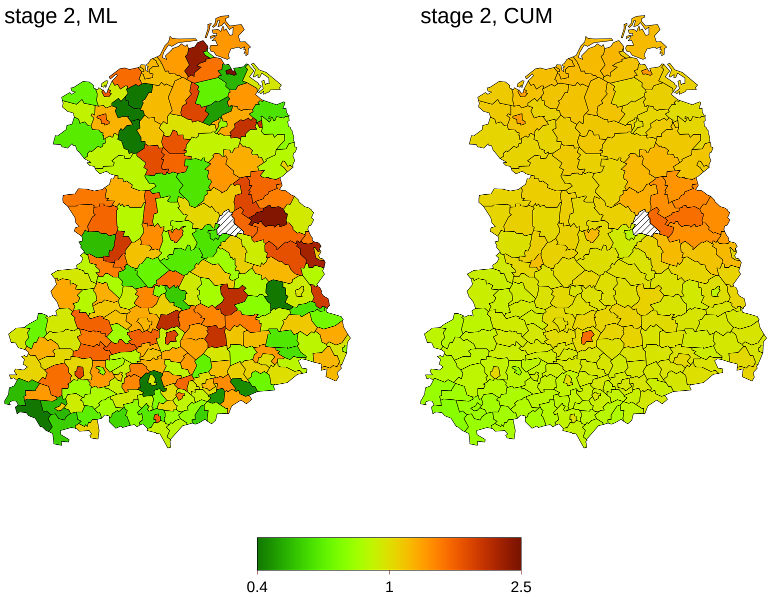


Figure 6.6: Relative risk estimates for diagnosis of the disease in stage 2 by ML (left map) and the cumulative model (right map).

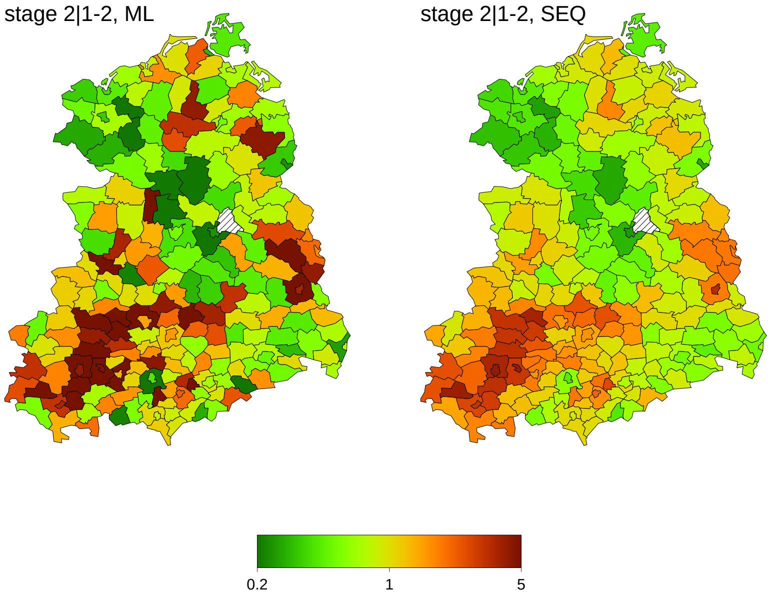


Figure 6.7: Conditional odds ratio estimates for diagnosis in the malignant stage 2, given diagnosis in stage 1 or 2 by ML (left map) and the sequential model (right map).

mation on the number of lab tests indicates that in the 1970s the highest number of tests have been carried out in these two areas, while in Saxony-Anhalt and Thuringia (the south-east of the GDR) the lowest numbers were observed (Quaas & Heinrich 1998).

The maps fit roughly into this pattern: in the north-west (initially high number of tests) they show totally a higher proportion of identified premalignant and malignant cancers (Figure 6.5), but among them low proportions of malignant cancers (Figure 6.7). In the south-west of the country (initially low numbers of tests) totally a lower proportion of identified premalignant and malignant cancers can be seen, but among them high proportions of malignant cancers. In detail, the pattern is more complicated: not the entire area of Mecklenburg-West Pomerania shows the low proportion of malignant cancers, and areas with initially low frequencies of testing show nevertheless low proportions of malignant cancers (e.g. Saxony in the south-east). These findings may be due to the fact that *several* factors influence the effectiveness of a screening programme: (a) availability of the programme, (b) quality of the programme, (c) attendance of the eligible population, and (d) quality of outcome report to the cancer registry. These factors may affect the outcome differently in the different regions of the country. The maps show only the overall effect of these factors. Thus, the method might be valuable to provide indicators to areas with unsatisfactory performance of the screening whatsoever the reasons are. Their elucidation would need more detailed epidemiological investigation.

6.4 Discussion

In this paper we have proposed methods for the spatial analysis of cancer incidence data with additional knowledge on the stage of the disease. Throughout we have used Markov random field models in order to acknowledge the spatial structure of the data. Of course, other models for spatial correlation can be used as well, for example the recently developed adaptive smoothing methods based on partition (Knorr-Held & Raßer 2000, Denison & Holmes 2001) or mixture models (Green & Richardson 2000). We are currently investigating the applicability of partition models to such data.

In terms of comparing the two proposed models it seems that most arguments are in favor of the sequential model: (a) This model is easier to implement because no order constraints are necessary; (b) We can even separate the analysis and fit S binomial regression models separately; (c) The conditional interpretation of the parameters is more useful in order to judge the effectiveness of cancer screening and shows connections to the statistical analysis of spatial case-control studies (Kelsall & Diggle 1998); (d) In our application the sequential model provided a substantially better model fit. Only if the interest lies in estimating cumulative relative risks then the cumulative model should be preferred.

An obvious extension of the two models considered is the inclusion of relevant covariates in order to reduce ("explain") the observed spatial pattern. Depending on the covariate and on the model, the effect could be assumed to be independent of the stage, or stage-specific. For example, if the number of lab tests would be available on a district-specific level, it could be included in the sequential model (6.4) for $s = 1$.

Finally we note that the incidence data from the GDR cancer registry is actually available for all years between 1961 and 1989. An interesting problem would be to construct a space-time model that captures the increasing number of cases in the premalignant stage and their temporal effect on the number of diagnosed malignant cases some time later. Here the specification of the time lag between the premalignant and malignant stage is not obvious and could possibly even be estimated from such data as well.

Acknowledgments

This research was funded by the German Science Foundation (DFG), SFB 386. The authors thank R. Stabenow from the joint cancer registry in Berlin for providing the data set on cervical cancer. The revision has benefited from helpful comments from two referees.

6.5 Model formulation with CPM prior

In this section we will propose a CPM analogue to the GMRF model. As mentioned above, the sequential model yields better results and is easier to implement. Hence, we concentrate on this approach and use separate binomial models for each category $s = 0, \dots, S-1$. For this purpose, we transfer the multinomial model (6.4) to the binomial case, omitting the stage indicator s

$$\text{logit}(q_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \nu + \eta_i + \psi_j. \tag{6.12}$$

We will apply Gaussian CPM priors for the spatial effects $\boldsymbol{\eta} = (\eta_1, \dots, \eta_I)$ and the age effects $\boldsymbol{\psi} = (\psi_1, \dots, \psi_J)$. In the multinomial GMRF model both effects have been centered for reasons of interpretation. We do the same in the CPM to facilitate a fair comparison. The restrictions

$$\sum_{i=1}^{I} \eta_i = 0 \quad \text{and} \quad \sum_{j=1}^{J} \psi_j = 0. \tag{6.13}$$

are imposed directly in the prior. Note that these restrictions are expressed in terms of parameters on region and age group level rather than on cluster level. In fact, restrictions on cluster level offer no meaningful interpretation.

Sampling under such linear constraints is slightly different than before. The algorithm used so far needs to be modified. In the following we present a sampling scheme, which is suitable for Gaussian CPMs. Yet, extensions to other distributions are possible and discussed later on.

6.5.1 Sampling under linear constraints

We will apply identical prior formulations for $\boldsymbol{\eta}$ and $\boldsymbol{\psi}$ within the logit model (6.12) under sum-to-zero restrictions (6.13). Still, the proposed Gaussian CPM sampler applies to any regression model. Therefore, we present a general formulation using our standard notation with parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$ on individual level and $\boldsymbol{\theta}_k = (\theta_1, \dots, \theta_k)$ on cluster level, cf. Table 2.2. Note that $\boldsymbol{\theta}_k$ has no relation to the spatial parameters in the cumulative GMRF model. We work in a generalized additive model framework and assume an additive predictor with intercept ν and some covariate effect $\boldsymbol{\lambda}$. With a CPM prior for $\boldsymbol{\lambda}$, the likelihood is denoted by $p(\boldsymbol{y}|\boldsymbol{\theta}_k, \nu)$.

For a partition with k clusters, a Gaussian CPM assumes that parameters $\boldsymbol{\theta}_k$ have independent normal distributions with overall mean μ and overall variance σ^2. Thus, the joint prior density is the product

$$p(\boldsymbol{\theta}_k|k,\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{k}(\theta_j-\mu)^2\right\}$$

under the linear constraint

$$\sum_{i=1}^{n}\lambda_i = \sum_{j=1}^{k} m_j\theta_j = 0, \tag{6.14}$$

where m_j denotes the size of cluster C_j. Furthermore, we assume diffuse priors for the intercept ν and the hyperparameter μ, i.e.

$$p(\nu) \propto \text{constant} \quad \text{and} \quad p(\mu) \propto \text{constant}.$$

This is a common prior assumption for location parameters, but is of special importance for the proposed sampler. As will be shown below, sampling under linear restrictions is simplified by this choice. We further use an inverse gamma prior for the variance σ^2, but this is of less importance for the following considerations.

Any proposed move that implies a change of the parameters $\boldsymbol{\lambda}$ has to account for restriction (6.14). Obviously, any proposed modification of $\boldsymbol{\theta}_k$ implies a change of the parameters $\boldsymbol{\lambda}$. Note that this is also true for any proposed new partition since this implies new cluster sizes, in general, and these enter into the restriction (6.14). For simplicity, we omit the moves shift and switch from our sampling scheme. As argued in Section 3.2.2, these moves are not necessary. Still, we have to deal with three different moves: height, birth, and death. In addition, there will be a hyper move which has no effect on the parameters $\boldsymbol{\lambda}$.

Recall that in all models so far we have used a special form of proposal distribution for the cluster parameters. More precisely, we have drawn a new candidate parameter either from the full conditional or from an approximation thereof. However, this is somewhat more difficult in the binomial model. The conjugate prior for the unknown probabilities π_{ij} is a beta distribution. Yet, we parameterize the model in terms of an additive predictor. Basically, one might use the same idea as before and draw a candidate for the probabilities (the inverse logits of the additive predictor) from the corresponding full conditional, i.e. a beta prior with matched moments. However, the transformation of this proposal to one of the parameters ($\boldsymbol{\eta}$ or $\boldsymbol{\psi}$ in our model) will be rather poor since one will have to neglect all other parameters, e.g. for updating the spatial effects $\boldsymbol{\eta}$ one would have to neglect the age effects $\boldsymbol{\psi}$ in the approximation.

Therefore, we will use a rather simple proposal scheme. First, we describe the height move in detail. The extension to the dimension changing moves will be discussed afterwards.

The construction of a naive sampler is straightforward. Suppose, for one cluster C_j an intermediate value $\tilde{\theta}_j$ is generated according to a random walk proposal $\tilde{\theta}_j|\theta_j \sim \mathrm{N}(\theta_j, \tau^2)$, where the variance τ^2 is a fixed tuning parameter of the sampler. Then, the intermediate parameters

$\tilde{\boldsymbol{\theta}}_k = (\tilde{\theta}_1, \dots, \tilde{\theta}_k)$ with $\tilde{\theta}_s = \theta_s$ for $s \neq j$ are centered. Thus, the proposal $\boldsymbol{\theta}_k^* = (\theta_1^*, \dots, \theta_k^*)$ is given by

$$\theta_s^* = \tilde{\theta}_s - \sum_{l=1}^{k} \frac{m_l}{n} \tilde{\theta}_l = \tilde{\theta}_s - \tilde{c}, \quad s = 1, \dots, k, \tag{6.15}$$

where the centering constant

$$\tilde{c} = \tilde{c}(\tilde{\theta}_j) = \frac{m_j}{n}(\tilde{\theta}_j - \theta_j)$$

can be computed solely from the intermediate proposal $\tilde{\theta}_j$. Simultaneously, the intercept ν and the mean μ are changed to

$$\nu^* = \nu + \tilde{c} \quad \text{and} \quad \mu^* = \mu - \tilde{c}. \tag{6.16}$$

The joint proposal $(\boldsymbol{\theta}_k^*, \nu^*, \mu^*)$ depends on $\tilde{\theta}_j$ alone. Therefore, given the current values $(\boldsymbol{\theta}_k, \nu, \mu)$ we have the following identity

$$p(\boldsymbol{\theta}_k^*, \nu^*, \mu^* | \tilde{\theta}_j, \boldsymbol{\theta}_k, \nu, \mu) = \begin{cases} 1 & \text{if (6.15) and (6.16) are valid,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.17}$$

On the other hand, from the current values $(\boldsymbol{\theta}_k, \nu, \mu)$ and the proposal $(\boldsymbol{\theta}_k^*, \nu^*, \mu^*)$, one can derive the intermediate state in a unique way, e.g.

$$\tilde{\theta}_j = \frac{n}{m_j}\tilde{c} + \theta_j = \frac{n}{m_j}(\nu^* - \nu) + \theta_j \tag{6.18}$$

and therefore

$$p(\tilde{\theta}_j | \boldsymbol{\theta}_k^*, \nu^*, \mu^*, \boldsymbol{\theta}_k, \nu, \mu) = \begin{cases} 1 & \text{if (6.18) is valid,} \\ 0 & \text{otherwise.} \end{cases} \tag{6.19}$$

Thus, the proposal has a joint density

$$\begin{aligned} q(\boldsymbol{\theta}_k^*, \nu^*, \mu^* | \boldsymbol{\theta}_k, \nu, \mu) &= \frac{p(\boldsymbol{\theta}_k^*, \nu^*, \mu^*, \tilde{\theta}_j | \boldsymbol{\theta}_k, \nu, \mu)}{p(\tilde{\theta}_j | \boldsymbol{\theta}_k^*, \nu^*, \mu^*, \boldsymbol{\theta}_k, \nu, \mu)} \\ &\overset{(6.19)}{=} p(\boldsymbol{\theta}_k^*, \nu^*, \mu^*, \tilde{\theta}_j | \boldsymbol{\theta}_k, \nu, \mu) \\ &\overset{(6.17)}{=} \frac{p(\boldsymbol{\theta}_k^*, \nu^*, \mu^*, \tilde{\theta}_j | \boldsymbol{\theta}_k, \nu, \mu)}{p(\boldsymbol{\theta}_k^*, \nu^*, \mu^* | \tilde{\theta}_j, \boldsymbol{\theta}_k, \nu, \mu)} \\ &= p(\tilde{\theta}_j | \boldsymbol{\theta}_k, \nu, \mu) \\ &= q(\tilde{\theta}_j | \theta_j). \end{aligned}$$

This is the density of a normal distribution according to the random walk proposal for the intermediate value. Hence, the joint proposal $(\boldsymbol{\theta}_k^*, \nu^*, \mu^*)$ is accepted with the usual Metropolis-Hastings probability $\alpha = \min\{1, \ \mathcal{L} \cdot \mathcal{P} \cdot \mathcal{Q}\}$, where

$$\begin{aligned} \mathcal{L} &= \frac{p(\boldsymbol{y} | \boldsymbol{\theta}_k^*, \nu^*)}{p(\boldsymbol{y} | \boldsymbol{\theta}_k, \nu)}, & (6.20) \\ \mathcal{P} &= \frac{p(\boldsymbol{\theta}_k^* | \mu^*, \sigma^2) p(\mu^*) p(\nu^*)}{p(\boldsymbol{\theta}_k | \mu, \sigma^2) p(\mu) p(\nu)} = \frac{p(\boldsymbol{\theta}_k^* | \mu^*, \sigma^2)}{p(\boldsymbol{\theta}_k | \mu, \sigma^2)} = \frac{p(\theta_j^* | \mu^*, \sigma^2)}{p(\theta_j | \mu, \sigma^2)}, & (6.21) \\ \mathcal{Q} &= \frac{q(\theta_j | \tilde{\theta}_j)}{q(\tilde{\theta}_j | \theta_j)} = 1. \end{aligned}$$

Note that the prior ratio $\mathcal{P}$ reduces to the ratio of two univariate normal densities for cluster C_j. The ratios for all other clusters cancel out due to the symmetry of the Gaussian prior density around μ^* and μ, respectively.

The drawback of this naive sampler is that each proposal has to be centered before the acceptance/rejection step. Thus, some computational speed is lost, unless the algorithm is tuned to produce high acceptance rates. This is possible for a random walk proposal in terms of the tuning parameter τ^2, but will lead to slow mixing behavior. Note that more complicated proposals based on simultaneous updates of several parameters are difficult to implement. In general, the proposal density of the reverse move $q(\boldsymbol{\theta}_k, \nu, \mu | \boldsymbol{\theta}_k^*, \nu^*, \mu^*)$ cannot be derived.

We now propose a modified sampler in which the intermediate proposal is accepted or rejected. Centering is performed afterwards and is therefore only necessary if the proposal is accepted. Note that in the sampler described above, a change of the intercept ν is performed. This is not necessary and enters in the acceptance probability only via the likelihood. But this change assures that the linear predictor is left unchanged in the centering step, i.e.

$$\nu + \tilde{\lambda}_i = \nu + \tilde{\theta}_j = \nu + \tilde{c} + \tilde{\theta}_j - \tilde{c} = \nu^* + \theta_j^* = \nu^* + \lambda_i^* \quad \text{for } i \in C_j.$$

Hence, the likelihood evaluated at the intermediate and the final proposal is identical

$$p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_k, \nu) = p(\boldsymbol{y}|\boldsymbol{\theta}_k^*, \nu^*).$$

The same argument holds true for the prior density. Changing the mean μ accordingly allows to rewrite the prior for $\boldsymbol{\theta}_k^*$ in terms of $\tilde{\boldsymbol{\theta}}_k$

$$p(\theta_j^*|\mu^*, \sigma^2) = p(\tilde{\theta}_j - \tilde{c}|\mu - \tilde{c}, \sigma^2) = p(\tilde{\theta}_j|\mu, \sigma^2), \quad j = 1, \ldots, k.$$

Thus, we may write the acceptance probability solely in terms of the intermediate proposal $\tilde{\boldsymbol{\theta}}_k$. Both, the likelihood ratio (6.20) and the prior ratio (6.21), are left unchanged. As already shown, the proposal depends on the intermediate proposal alone by construction. The acceptance probability can be derived without centering the proposal. Therefore, we may accept or reject the intermediate proposal. The centering step has to be performed only if the proposal is accepted.

We now turn to the dimension changing moves birth and death. We use the same idea here, and construct a sampler based on an intermediate proposal. Suppose we generate a new cluster C^* of size m^* as usual. For simplicity, the corresponding intermediate parameter $\tilde{\theta}$ is drawn from the normal prior, i.e. $\tilde{\theta} \sim \mathrm{N}(\mu, \sigma^2)$. Alternatively, one could use information in terms of the current parameters λ_i for all regions $i \in C^*$. The intermediate parameters $\tilde{\boldsymbol{\theta}}_{k+1}$ are identical to the current parameters $\boldsymbol{\theta}_k$ with the new value $\tilde{\theta}$ inserted at the correct position. Thus, the proposal $\boldsymbol{\theta}_{k+1}^*$ is given by

$$\theta_s^* = \tilde{\theta}_s - \sum_{l=1}^{k+1} \frac{m_l^*}{n} \tilde{\theta}_l = \tilde{\theta}_s - \tilde{c}, \quad s = 1, \ldots, k+1,$$

where m_l^*, $l = 1, \ldots, k+1$, are the cluster sizes of the new partition $C_1^*, \ldots, C_{k+1}^*$. The centering constant can be written as

$$\tilde{c} = \tilde{c}(\tilde{\theta}) = \frac{m^*}{n}(\tilde{\theta} - \bar{\theta}),$$

where $\bar{\theta}$ denotes the mean of the current parameters in all regions assigned to the new cluster

$$\bar{\theta} = \frac{1}{m^*} \sum_{i \in C^*} \theta_{j(i)} = \frac{1}{m^*} \sum_{i \in C^*} \lambda_i.$$

The centering constant $\tilde{c}$ depends only on the intermediate proposal $\tilde{\theta}$ and the new cluster size m^*. Simultaneous updates of the intercept and the mean according to (6.16) yield the following components of the acceptance probability

$$\begin{aligned}
\mathcal{L} &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}_{k+1}^*, \nu^*)}{p(\boldsymbol{y}|\boldsymbol{\theta}_k, \nu)} = \frac{p(\boldsymbol{y}|\tilde{\boldsymbol{\theta}}_{k+1}, \nu)}{p(\boldsymbol{y}|\boldsymbol{\theta}_k, \nu)}, \\
\mathcal{P} &= \frac{p(\boldsymbol{\theta}_{k+1}^*|\mu^*, \sigma^2)p(\mu^*)p(\nu^*)}{p(\boldsymbol{\theta}_k|\mu, \sigma^2)p(\mu)p(\nu)} = p(\theta^*|\mu^*, \sigma^2) = p(\tilde{\theta}|\mu, \sigma^2), \\
\mathcal{Q} &= \frac{1}{q(\tilde{\theta}|\mu, \sigma^2)},
\end{aligned}$$

where the usual ratios corresponding to the change of the partition and the sampling scheme have been left out for simplicity. Note that the prior ratio $\mathcal{P}$ and the proposal ratio $\mathcal{Q}$ cancel out since the proposal is drawn from the prior distribution, i.e. $q(\tilde{\theta}|\mu, \sigma^2) = p(\tilde{\theta}|\mu, \sigma^2)$. Thus, the acceptance probability for the birth move can be written

$$\alpha = \min\left\{1, \mathcal{L} \cdot \frac{p(k+1)}{p(k)} \cdot \frac{r_D(k+1)}{r_B(k)}\right\}.$$

Inverting all ratio terms yields the acceptance probability for the corresponding death move. Both dimension changing moves may also be accepted or rejected based on the intermediate proposal.

Finally, there has to be remarked that the proposed sampler can be modified. For the height move, any proposal distribution can be applied as long as we update the parameters $\theta_1, \ldots, \theta_k$ one by one. In a similar way, other proposals for the birth move are possible as long as the joint proposal density can be computed based on the intermediate (non-centered) proposal alone. Furthermore, the methodology is also suitable for prior distributions other than Gaussian. The symmetry of the prior distribution is sufficient for the construction of the sampler.

6.5.2 Implementation and prior specifications

For the analysis with CPM prior we have replaced the sequential model with two independent binomial models. For simplicity, we reintroduce stage indicators $s = 0$ and $s = 1$ in the model formulation (6.12). Thus, we have two models

$$\begin{aligned}
y_{ij0} &\sim \text{Bin}(n_{ij}, \text{logit}^{-1}(\nu_0 + \eta_{0i} + \psi_{0j})) \\
y_{ij1} &\sim \text{Bin}(y_{ij1} + y_{ij2}, \text{logit}^{-1}(\nu_1 + \eta_{1i} + \psi_{1j}))
\end{aligned}$$

corresponding to stage $1-2$ ($s = 0$) and stage $2|1-2$ ($s = 1$) in the notation of the previous sections. The prior specifications for both models were chosen identical, and we use the subscript s in the following.

For both models we use independent Gaussian CPMs with sum-to-zero restriction for the spatial effects $\boldsymbol{\eta}_s$ and the age effects $\boldsymbol{\psi}_s$. Therefore, the priors for the intercept and both means of the normal priors are diffuse. The only hyperparameters left to specify are the parameters of the inverse gamma priors for the variances. For both effects those were chosen identical $a = 1$ and $b = 0.001$. This is the same choice as for the prior of the precision parameters used in the GMRF model, see Section 6.3.

The algorithm is completed with a hyper move. Here, Gibbs sampler steps for all hyperparameters are implemented, cf. (4.5) and (4.6). Note that the intercept is only updated indirectly within the moves height, birth, and death. Still, additional Gibbs sampler steps were used for the means of the Gaussian priors to improve mixing, although this is not necessary. All results in the next section are based on 5,000 samples from the posterior. Those were collected in a run with 51,000,000 iterations, of which 1,000,000 were burn-in together with a lag of 10,000 iterations between stored iterations.

6.5.3 Comparison of the results

First, we give some comments on the performance of the algorithm. Whereas the height move can be tuned to give a satisfying acceptance rate, this is not possible with the dimension changing moves in the proposed sampler. Still, the acceptance rates for those moves were remarkably good for both effects. All acceptance rates are shown in Table 6.2. For the spatial components $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_1$ all values are very good, in particular for stage $s = 0$. For the age effects all rates are lower. This validates a strong structure of the age effects already noticeable in the GMRF model. Although the proposal distributions were chosen for reasons of practicability, the performance of the algorithm is convincing.

	stage $1-2$		stage $2\|1-2$	
	$\boldsymbol{\eta}_0$	$\boldsymbol{\psi}_0$	$\boldsymbol{\eta}_1$	$\boldsymbol{\psi}_1$
birth	33	7	18	7
death	33	7	18	7
height	54	29	65	47

Table 6.2: Acceptance rates for all moves (in percent).

We now give a brief comparison of the results gained by the sequential model with GMRF prior, discussed in Section 6.3, and the alternative formulation with CPM prior as proposed above. Basically, the results are very similar. Therefore, all conclusions drawn before are still valid.

First, we take a look at the model fit by means of the saturated deviance. In the multinomial model the deviance residual is calculated as the sum over all stages, see equation (6.10). The binomial deviance residual is a special case with only two stages. It can be shown that the sum of the deviances of separate binomial models is equal to the deviance of a multinomial model. Therefore, we may compare the two binomial models with the multinomial model by summing over the deviances. The same holds true for the effective number of parameters p_D and hence for the DIC. Note that p_D has been calculated based on the inverse logit of the posterior mean of the linear predictor, rather than from the posterior mean of the probabilities.

The values for the two binomial models, denoted by BIN_0 and BIN_1, are given in the top rows of Table 6.3. The bottom rows compare the sum of the two models, denoted by BIN, with the sequential model (SEQ) from Section 6.3. There is a slight preference for the GMRF model but the differences are small. Therefore, the decision for one of the two models (SEQ or BIN) may also be based on further inspection of the estimated effects.

Model	$\bar{D}$	p_D	DIC
BIN_0	3411	159	3570
BIN_1	2002	102	2104
BIN	5413	261	5674
SEQ	5403	245	5649

Table 6.3: Comparison of deviance summaries.

In Figure 6.8 the age effects for both models are displayed. The posterior median estimates as well as the credible intervals are very similar to the previous results, displayed in Figure 6.4. In fact, the effects $\boldsymbol{\psi}_0$ are almost identical. The effects $\boldsymbol{\psi}_1$ show some minor deviations for border age groups. For the first and the last two age groups, the CPM supports rather constant effects. In contrast, the GMRF model produces more linear trends. Clearly, this is due to the prior assumption which is stronger than the likelihood for age groups with only few observed cases.

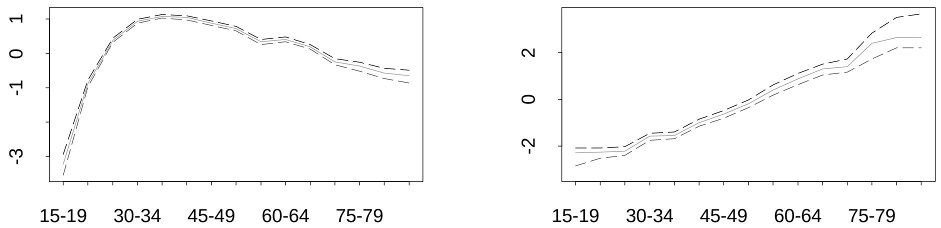


Figure 6.8: Estimated median age effects of $-\boldsymbol{\psi}_0$ (left plot) and $-\boldsymbol{\psi}_1$ (right plot) within 90% pointwise credible intervals from the two binomial models (on the same scale as in Figure 6.4).

Figure 6.9 displays the posterior median estimates of the relative risks for $s = 0$ from the CPM model (middle map). For comparison, the ML estimates and the posterior median estimates from the sequential model are also depicted; these are identical to Figure 6.5. A visual inspection reveals almost no differences. This is not surprising since the information in the data for this stage is very strong.

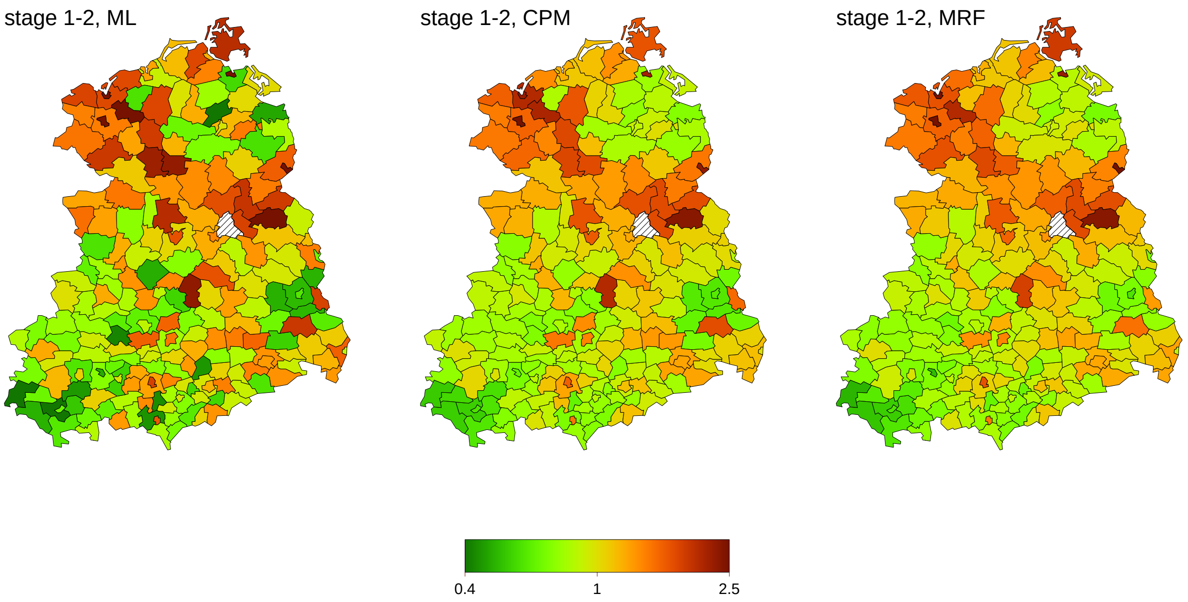


Figure 6.9: Relative risk estimates for diagnosis of the disease regardless of the stage by ML (left map), the CPM (middle map), and the GMRF model (right map).

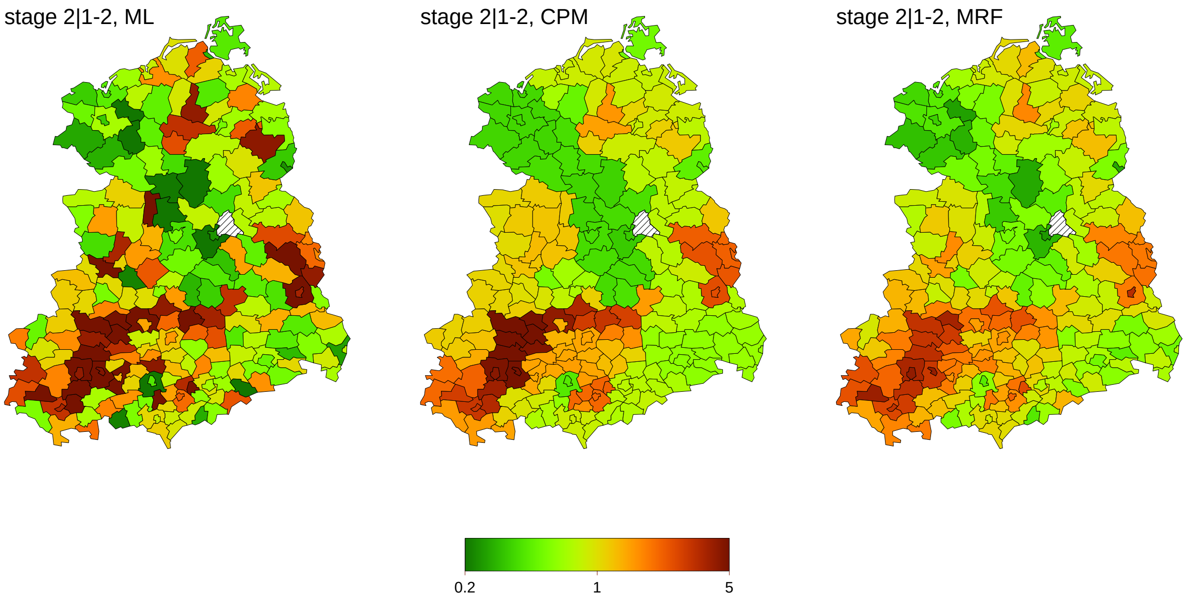


Figure 6.10: Conditional odds ratio estimates for diagnosis in the malignant stage 2, given diagnosis in stage 1 or 2 by ML (left map), the CPM (middle map), and the GMRF model (right map).

The different prior for the spatial effects becomes more obvious in the binomial model for $s = 1$. In Figure 6.10 the posterior median estimates for the conditional odds ratios are displayed, along with the corresponding previous results, see Figure 6.7. Here, the differences are clearly visible. Although the general patterns coincide, the CPM provides a more distinct structure than the GMRF prior. There is less information in the data, and correspondingly the results reflect the prior assumption. Thus, in the middle map, there are clusters of elevated or lowered risk clearly separated. In contrast, the GMRF model resembles the data to a greater extent. As a consequence, the deviance values in Table 6.3 are lower for the GMRF model. The slightly worse model fit of the CPM is probably due to this stage ($s = 1$). Yet, the clear spatial structure is appealing and offers easier interpretation than the rougher surface of the GMRF model.

Chapter 7

Conclusion

The estimation of unknown functions (or surfaces) is one of the major tasks of statistics. Often, some smoothness assumption on the unknown function is postulated in the statistical model. Either this is done in view of the nature of the data, or in order to allow for a better interpretation of the results. Of special interest for practical use are models, which are able to adapt the smoothness of the estimated function to the data. The decision on the use of such models depends on the field of application.

The main goal of this thesis was to propose a model that allows for spatially adaptive smoothing in discrete space. Originally, the CPM was developed for the purpose of estimating disease risk for a given set of geographical regions. The results for such applications were convincing and encouraged further investigation of the model. Here, the main focus was on a theoretical foundation of clustering partitions as well as on their practical applicability as a prior model within a hierarchical Bayesian framework.

The representation of a finite set of units in terms of a connected, undirected graph enables a generalization of the model to almost arbitrary discrete structures. This notation is appealing since common settings, e.g. regular arrays of pixels, are just special cases of undirected graphs with regular neighborhood structure between the vertices.

In this thesis, we have successfully applied the CPM to Poisson, Gaussian, and binomial observation models. Moreover, there are no theoretical limitations to the model concerning the transfer to other data types. Two components have influence on the smoothing properties of the CPM prior: smoothing according to the specified prior density for the parameters in the clusters, and smoothing as implied by the partition. The smoothing behavior due to the partition can be controlled by the prior distribution on the number of clusters. Given the number of clusters, we have assumed equal probabilities for all possible generating vectors. This is not necessary and it is possible to give preference to specific generating vectors, a priori. For example, we may increase or decrease the probability for a certain vertex to be selected as a cluster center. Such modifications allow for the adaptation of the CPM prior to external knowledge (if available), and may be used to deal with missing data.

One drawback of the CPM is that the prior properties cannot be derived analytically. Still, simulations from the prior are straightforward. For the graphs in this thesis, the prior shows desirable properties concerning the smoothing behavior. Furthermore, the prior properties are rather robust over the set of units, even for irregular structures.

A comparison of our CPM model with commonly used MRF models leads to the following conclusions: (1) Whenever the data provides enough information, both priors yield similar results; (2) For sparse data, the CPM prior provides more clear structure than the MRF prior. In other words, the MRF prior approximates the data under the assumption of a global smoothing parameter. This is of less importance for informative data, but leads to blurring effect for sparse data whenever sudden changes and edges are present in the surface. In contrast, the CPM prior is able to retain such edges, but for sparse data this may even lead to few unjustified edges in the surface. Basically, the CPM prior does not necessarily smooth the data, but allows for independent estimation of parameters for single vertices if there is evidence for this in the data.

For the practical use of CPM priors, the computational speed of the algorithm as well as an easy implementation are of interest. So far, there are rarely software packages available that allow for the estimation of parameters via reversible jump MCMC. All CPM samplers in this thesis were coded in `C/C++` for specific applications. Still, the implementation is simplified by the independence assumption for cluster parameters.

In general, the performance of the algorithm is good and posterior median point estimates are stable even for short runs. Still, longer runs, as used in this work, are useful for more accurate estimation, especially for complex models with several independent CPM priors for different parameters.

One limitation of the CPM prior is given by the size of the underlying graph, i.e. by the number of vertices. For large graphs, the computation time increases and the CPM prior is not suitable anymore. Still, for mid-size graphs like in most disease mapping applications the CPM algorithm is fast enough for practical use.

Appendix A

Proofs

A.1 Proof of Proposition 2.1

Let $k \leq n$ and $j \in \{1, \ldots, k\}$ be fixed. Suppose $i \in C_j$ and let a minimal path $g_j, v_1, \ldots, v_p, i$ be chosen arbitrarily. It is sufficient to show that for any v_l with $1 \leq l \leq p$ assignment rules (2.3) or (2.4) are fulfilled for cluster C_j. Since v_l is on the minimal path between g_j and i

$$d(g_j, i) = d(g_j, v_l) + d(v_l, i).$$

First, suppose $d(g_j, i) < d(g_s, i)$ for all $s \neq j$. Then, vertex i is assigned to C_j according to (2.3) and for all $s \neq j$

$$\begin{aligned} d(g_j, v_l) &= d(g_j, i) - d(v_l, i) \\ &< d(g_s, i) - d(v_l, i) \\ &\leq d(g_s, v_l) + d(v_l, i) - d(v_l, i) \qquad \text{[Triangle inequation]} \\ &= d(g_s, v_l). \end{aligned} \tag{A.1}$$

Therefore, v_l is also assigned to C_j according to (2.3).

Second, if $d(g_j, i) = d(g_s, i)$ for some $s \neq j$, then i is assigned to C_j according to (2.4) and hence $j < s$. Then, there is an equal sign in (A.1) and we only get $d(g_j, v_l) \leq d(g_s, v_l)$. But because of $j < s$, again $v_l \in C_j$ according to (2.4).

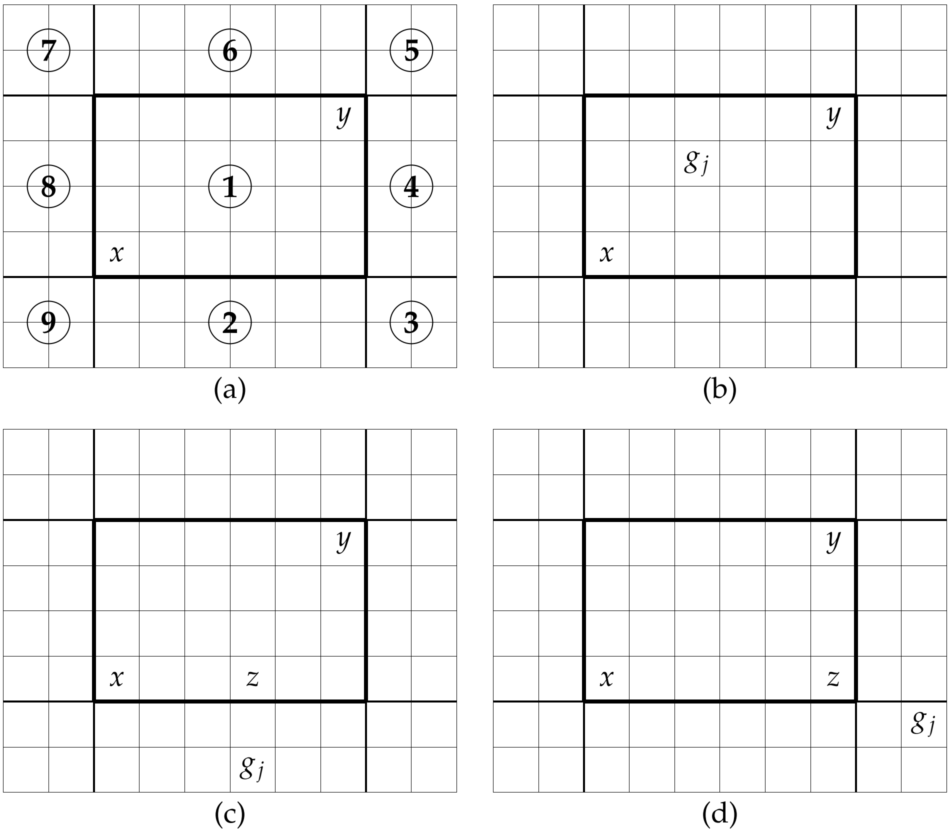
A.2 Proof of Proposition 2.2

Suppose a lattice with n_1 rows and n_2 columns and a clustering partition $\{C_1, \ldots, C_k\}$ with k clusters. Let vertices i_1, i_2 be assigned to cluster C_j. For easier notation, we use a coordinate representation of the vertices and write $i_1 = x = (\xi_1, \xi_2)$ and $i_2 = y = (\psi_1, \psi_2)$. Let vertex $g_j = (\gamma_1, \gamma_2)$ be the cluster center of C_j. We have to show that for $x, y \in C_j$ with $d(x, y) > 1$ there exists a minimal path $x, v_1, \ldots, v_p, y$ with $v_l \in C_j$ for all $1 \leq l \leq p$. If $g_j = x$ or $g_j = y$ this is true according to Proposition 2.1. Thus, let $g_j \neq x$ and $g_j \neq y$.

The vertices x and y span a rectangle R with the lower left corner $(\min(\xi_1, \psi_1), \min(\xi_2, \psi_2))$ and the upper right corner $(\max(\xi_1, \psi_1), \max(\xi_2, \psi_2))$. W.l.o.g. we assume that x is the lower left corner and y is the upper right corner of R. Other cases can be seen as a rotation of the lattice or a relabeling of the vertices, i.e. $x = i_2$ and $y = i_1$. The rectangle is given by

$$R = \Big\{(\omega_1, \omega_2) \in \{1, \ldots, n_1\} \times \{1, \ldots, n_2\} : \xi_1 \leq \omega_1 \leq \psi_1,\ \xi_2 \leq \omega_2 \leq \psi_2\Big\},$$

see Figure A.1 (a). Note that the vertices are displayed as squares, see also Figure 2.2. We distinguish three major cases determined by the location of the cluster center g_j: (1) both coordinates of g_j in R, (2) one coordinate of g_j in R, and (3) no coordinate of g_j in R.



(a) (b)

(c) (d)

Figure A.1: Rectangle R (bold lines) defined by vertices x and y and cases (medium lines). Displayed is (a) case differentiation, (b) case 1, (c) case 2 and (d) case 3.

Case 1: $g_j \in R$, see Figure A.1 (b), i.e.

$$\xi_1 \leq \gamma_1 \leq \psi_1 \qquad \text{and} \qquad \xi_2 \leq \gamma_2 \leq \psi_2$$

Then

$$\begin{aligned} d(x, g_j) &= |\xi_1 - \gamma_1| + |\xi_2 - \gamma_2| = -\xi_1 + \gamma_1 - \xi_2 + \gamma_2 \\ d(g_j, y) &= |\gamma_1 - \psi_1| + |\gamma_2 - \psi_2| = -\gamma_1 + \psi_1 - \gamma_2 + \psi_2 \end{aligned}$$

and

$$\begin{aligned} d(x,g_j)+d(g_j,y) &= -\xi_1+\gamma_1-\xi_2+\gamma_2-\gamma_1+\psi_1-\gamma_2+\psi_2 \\ &= -\xi_1+\psi_1-\xi_2+\psi_2 \\ &= |\psi_1-\xi_1|+|\psi_2-\xi_2| \\ &= d(x,y). \end{aligned}$$

Therefore, the cluster center g_j is on a minimal path between x and y. Let $d(x,g_j)=p_x+1$ and $d(g_j,y)=p_y+1$ with $p_x,p_y\in\{0,1,2,\ldots\}$. According to Proposition 2.1 there exists a minimal path $g_j,u_1,\ldots,u_{p_x},x$ with $u_l\in C_j,\ l=1,\ldots,p_x$, and a minimal path $g_j,w_1,\ldots,w_{p_y},y$ with $w_m\in C_j,\ m=1,\ldots,p_y$. Then $x,u_{p_x},\ldots,u_1,g_j,w_1,\ldots,w_{p_y},y$ is a path of length p_x+p_y+2 with all vertices in C_j. Since $d(x,y)=p_x+p_y+2$, this path is minimal.

Case 2: One coordinate of g_j in R, see Figure A.1 (c), i.e. w.l.o.g.

$$\xi_1\le\gamma_1\le\psi_1 \qquad \text{and} \qquad \gamma_2<\xi_2\le\psi_2.$$

Locations 4, 6, and 8 in Figure A.1 (a) can be seen as rotations of the lattice. We choose vertex z on the border of R with minimal distance to g_j, i.e. $z=(\gamma_1,\xi_2)$. Then

$$d(g_j,z)=|\gamma_1-\gamma_1|+|\gamma_2-\xi_2|=-\gamma_2+\xi_2,$$

and z is on a minimal path between g_j and y, since

$$d(z,y)=|\gamma_1-\psi_1|+|\xi_2-\psi_2|=-\gamma_1+\psi_1-\xi_2+\psi_2,$$

and

$$\begin{aligned} d(g_j,z)+d(z,y) &= -\gamma_2+\xi_2-\gamma_1+\psi_1-\xi_2+\psi_2 \\ &= -\gamma_2+\psi_2-\gamma_1+\psi_1 \\ &= |\gamma_2-\psi_2|+|\gamma_1-\psi_1| \\ &= d(g_j,y). \end{aligned}$$

Similar, it can be shown that z is on a minimal path between g_j and x, since

$$d(z,x)=|\gamma_1-\xi_1|+|\xi_2-\xi_2|=\gamma_1-\xi_1,$$

and

$$\begin{aligned} d(g_j,z)+d(z,x) &= -\gamma_2+\xi_2+\gamma_1-\xi_1 \\ &= |\gamma_2-\xi_2|+|\gamma_1-\xi_1| \\ &= d(g_j,x). \end{aligned}$$

Therefore, z is on a minimal path between x and y, since

$$\begin{aligned}
d(x,z)+d(z,y) &= |\xi_1-\gamma_1|+|\xi_2-\xi_2|+|\gamma_1-\psi_1|+|\xi_2-\psi_2| \\
&= -\xi_1+\gamma_1-\gamma_1+\psi_1-\xi_2+\psi_2 \\
&= -\xi_1+\psi_1-\xi_2+\psi_2 \\
&= |\xi_1-\psi_1|+|\xi_2-\psi_2| \\
&= d(x,y).
\end{aligned}$$

According to Proposition 2.1, all vertices on minimal paths between x and g_j and between y and g_j are assigned to cluster C_j, in particular, there are minimal paths between x and z and between y and z with all vertices in C_j. Since, z is on a minimal path between x and y there is a minimal path between x and y in C_j.

Case 3: No coordinate of g_j in R, see Figure A.1 (d), i.e. w.l.o.g.

$$\xi_1 \le \psi_1 < \gamma_1 \qquad \text{and} \qquad \gamma_2 < \xi_2 \le \psi_2.$$

Locations 5, 7, and 9 in Figure A.1 (a) can be seen as rotations of the lattice. We choose z as the corner of R nearest to g_j, i.e. $z=(\psi_1,\xi_2)$. Then

$$d(g_j,z)=|\gamma_1-\psi_1|+|\gamma_2-\xi_2|=\gamma_1-\psi_1-\gamma_2+\xi_2,$$

and z is on a minimal path between g_j and y, since

$$d(z,y)=|\psi_1-\psi_1|+|\xi_2-\psi_2|=-\xi_2+\psi_2,$$

and

$$\begin{aligned}
d(g_j,z)+d(z,y) &= \gamma_1-\psi_1-\gamma_2+\xi_2-\xi_2+\psi_2 \\
&= \gamma_1-\psi_1-\gamma_2+\psi_2 \\
&= |\gamma_1-\psi_1|+|\gamma_2-\psi_2| \\
&= d(g_j,y).
\end{aligned}$$

Similar, it can be shown that z is on a minimal path between g_j and x, since

$$d(z,x)=|\psi_1-\xi_1|+|\xi_2-\xi_2|=\psi_1-\xi_1,$$

and

$$\begin{aligned}
d(g_j,z)+d(z,x) &= \gamma_1-\psi_1-\gamma_2+\xi_2+\psi_1-\xi_1 \\
&= \gamma_1-\xi_1-\gamma_2+\xi_2 \\
&= |\gamma_1-\xi_1|+|\gamma_2-\xi_2| \\
&= d(g_j,x).
\end{aligned}$$

Therefore, z is on a minimal path between x and y, since

$$\begin{aligned} d(x,z)+d(z,y) &= |\xi_1-\psi_1|+|\xi_2-\xi_2|+|\psi_1-\psi_1|+|\xi_2-\psi_2| \\ &= |\xi_1-\psi_1|+|\xi_2-\psi_2| \\ &= d(x,y). \end{aligned}$$

With the same argument as before, there is a minimal path between x and y in C_j.

A.3 Counterexample for non-convexity in general graphs

Suppose a connected graph $G=\{V,E\}$ with 6 vertices $V=\{v,i,x,g_1,j,g_2\}$ and 6 edges $E=\{e_{vi},e_{vg_1},e_{ix},e_{g_1j},e_{jx},e_{xg_2}\}$ as displayed in Figure A.2. A clustering partition with generating vector $\boldsymbol{g}_2=(g_1,g_2)$ constitutes two clusters $C_1=\{g_1,i,v,j\}$ and $C_2=\{g_2,x\}$. The minimal path between i and j in C_1 is i,v,g_1,j and has length 3. But $d(i,j)=2$, and the (unique) minimal path i,x,j contains vertex $x\notin C_1$.

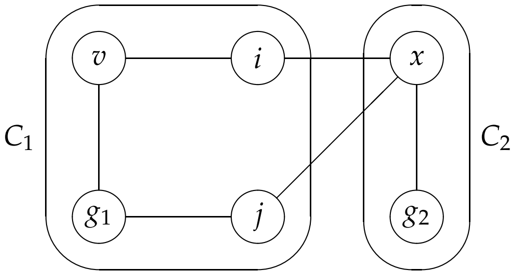


Figure A.2: A partition of G into two (non-convex) clusters.

A.4 Proof of equation (2.8)

Let $\mathcal{C}_k=\{C_1,\dots,C_k\}$ be a clustering partition with k clusters, defined by $\boldsymbol{g}_k=(g_1,\dots,g_k)$. The transformation $\boldsymbol{B}$ is uniquely defined by $(\boldsymbol{g}_k,k)$. Therefore,

$$p(\boldsymbol{\lambda}|\boldsymbol{\theta}_k,\boldsymbol{g}_k,k)=\begin{cases} p(\boldsymbol{B}\boldsymbol{\theta}_k|\boldsymbol{\theta}_k,\boldsymbol{g}_k,k)=1 & \text{if } \boldsymbol{\lambda}=\boldsymbol{B}\boldsymbol{\theta}_k, \\ 0 & \text{otherwise,}\end{cases} \tag{A.2}$$

and

$$p(\boldsymbol{\theta}_k|\boldsymbol{\lambda},\boldsymbol{g}_k,k)=p(\boldsymbol{\theta}_k|\boldsymbol{B}\boldsymbol{\theta}_k,\boldsymbol{g}_k,k)=1 \quad \text{if } \boldsymbol{\lambda}=\boldsymbol{B}\boldsymbol{\theta}_k. \tag{A.3}$$

The joint density of $\boldsymbol{\theta}_k$ and $\boldsymbol{\lambda}$ can be factorized in two ways

$$p(\boldsymbol{\lambda}|\boldsymbol{\theta}_k,\boldsymbol{g}_k,k)p(\boldsymbol{\theta}_k|\boldsymbol{g}_k,k)=p(\boldsymbol{\lambda},\boldsymbol{\theta}_k|\boldsymbol{g}_k,k)=p(\boldsymbol{\theta}_k|\boldsymbol{\lambda},\boldsymbol{g}_k,k)p(\boldsymbol{\lambda}|\boldsymbol{g}_k,k).$$

Let $\boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k$. It follows from (A.2) and (A.3)

$$\begin{aligned}
p(\boldsymbol{\lambda}|\boldsymbol{g}_k, k) &= \frac{p(\boldsymbol{\lambda}|\boldsymbol{\theta}_k, \boldsymbol{g}_k, k)p(\boldsymbol{\theta}_k|\boldsymbol{g}_k, k)}{p(\boldsymbol{\theta}_k|\boldsymbol{\lambda}, \boldsymbol{g}_k, k)} \\
&= \frac{p(\boldsymbol{B}\boldsymbol{\theta}_k|\boldsymbol{\theta}_k, \boldsymbol{g}_k, k)p(\boldsymbol{\theta}_k|\boldsymbol{g}_k, k)}{p(\boldsymbol{\theta}_k|\boldsymbol{B}\boldsymbol{\theta}_k, \boldsymbol{g}_k, k)} \\
&= p(\boldsymbol{\theta}_k|\boldsymbol{g}_k, k),
\end{aligned}$$

and thus

$$p(\boldsymbol{\lambda}|\boldsymbol{g}_k, k) = \begin{cases} p(\boldsymbol{\theta}_k|\boldsymbol{g}_k, k) & \text{if } \boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{\theta}_k, \\ 0 & \text{otherwise.} \end{cases}$$

Appendix B

GMRF Reconstructions of Synthetic Data Sets

As a comparison to the reconstruction of the two synthetic data sets, investigated in Section 4.1.3 with a CPM prior, we have analyzed the same data with a GMRF prior. More precisely, we have replaced the CPM prior with a pairwise difference prior for the parameters $\boldsymbol{\lambda}$,

$$p(\boldsymbol{\lambda}|\kappa) \propto \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (\lambda_i - \lambda_j)^2 \right).$$

Otherwise, the prior setting was chosen similar to the CPM approach. Estimation was performed using the software *BayesX* (Version 0.9, Brezger, Kneib & Lang 2002).

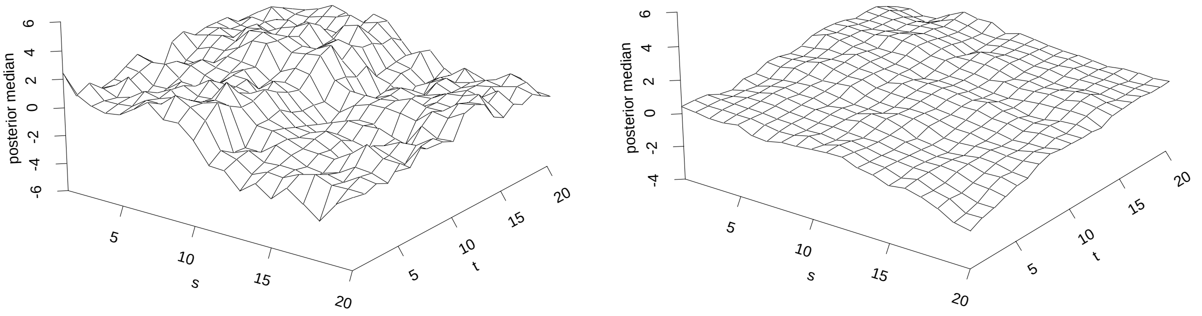


Figure B.1: Posterior median estimates for functions f_1 (left) and f_2 (right) with a Gaussian pairwise difference prior.

The reconstruction of function f_1, displayed in the left panel of Figure B.1, is rather poor. Obviously, the prior is not able to detect the two strong edges which are present in the true surface. The mean squared error of $\text{MSE}_1 = 0.318$ is clearly worse than for the CPM prior, and the error variance ($\tau^2 = 1$) is underestimated with a posterior median of $\hat{\tau}_1^2 = 0.566$. Globally, the pairwise difference prior smoothes too much.

From the right panel of Figure B.1 it becomes obvious that the results for the smooth function f_2 are better. The variance estimate $\hat{\tau}_2^2 = 0.897$ is closer to the true value, but still the true variance is slightly underestimated. Accordingly, the reconstruction shows some slight bumps. Altogether the reconstruction is about as good as with the CPM prior ($\text{MSE}_2 = 0.057$).

The pairwise difference prior assumes the same amount of smoothing over the whole lattice. Clearly, this is justified for the smooth function f_2, and the results are good. However, for the step function f_1 the assumption is wrong, and the two strong edges are blurred by the prior.

Appendix C

Further Simulations from the Prior

C.1 Map of Germany

In this simulation, the prior for the number of clusters k was uniform on $\{1, \ldots, 544\}$. Altogether the results are quite similar to those reported in Section 4.3.1. The expected number of clusters is 272.5. Thus, the probabilities of being alone in a cluster are higher and the average cluster sizes are smaller. One major difference becomes obvious in the map in Figure C.2. Here, the regions with only one neighbor are clearly visible. This effect is similar to the marginal variances of the MRF prior, although less emphasized. This phenomenon is easy to explain. With a larger number of clusters, the probability to be alone in a cluster of size one increases, especially for regions with only one neighbor, see Figure C.1. Still, this is only the case for partitions with many clusters. A small penalization of such partitions, e.g. a geometric distribution with small parameter c, is sufficient to assure that this effect vanishes, see Figure 4.10.

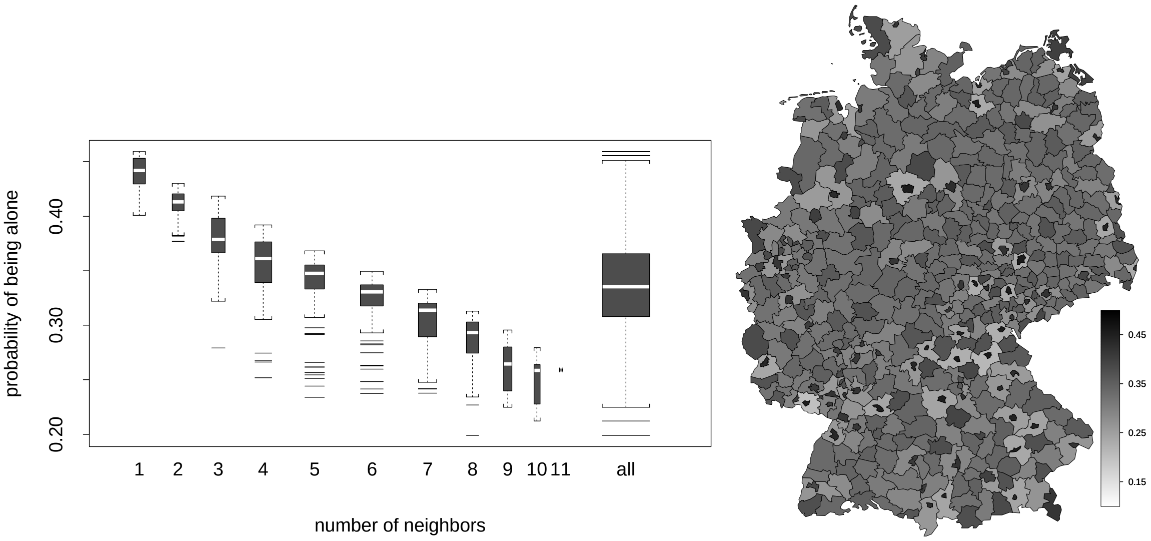


Figure C.1: Probability of being alone in a cluster for the map of Germany.

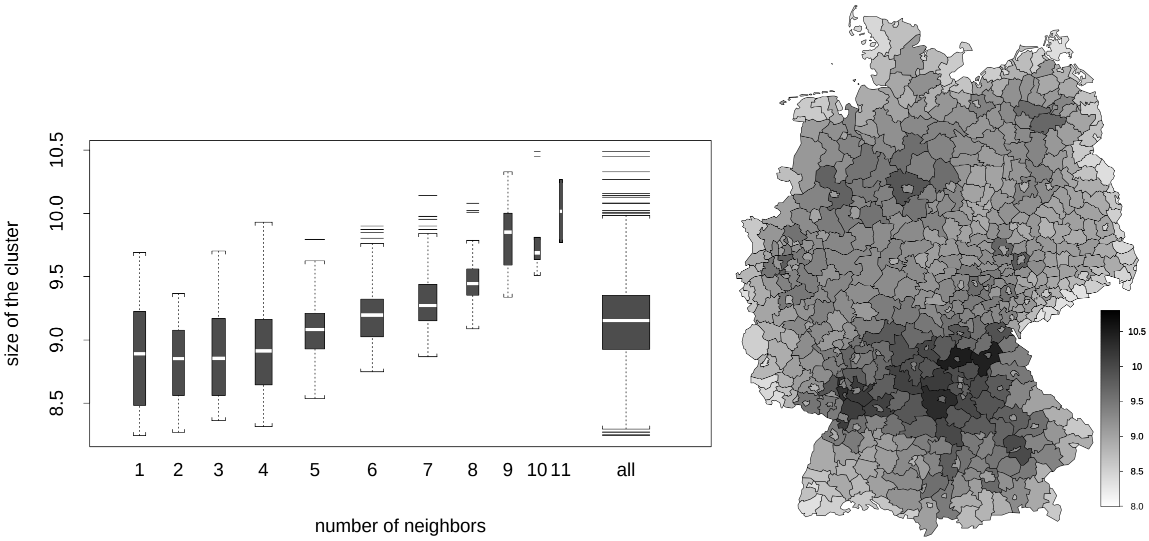


Figure C.2: Average cluster sizes for the map of Germany.

C.2 20 × 20-lattice

Here, we report a simulation from the prior using a uniform distribution on $\{1, \ldots, 400\}$ for the number of clusters k. The expected number of clusters is 200.5. Accordingly, the probabilities of being alone are higher and the average cluster sizes are smaller than with the geometric prior. Due to the regular structure of the graph there is no phenomenon for the cluster sizes as for the map of Germany. Altogether, the results appear to be similar to those with a geometric prior presented in Section 4.3.2.

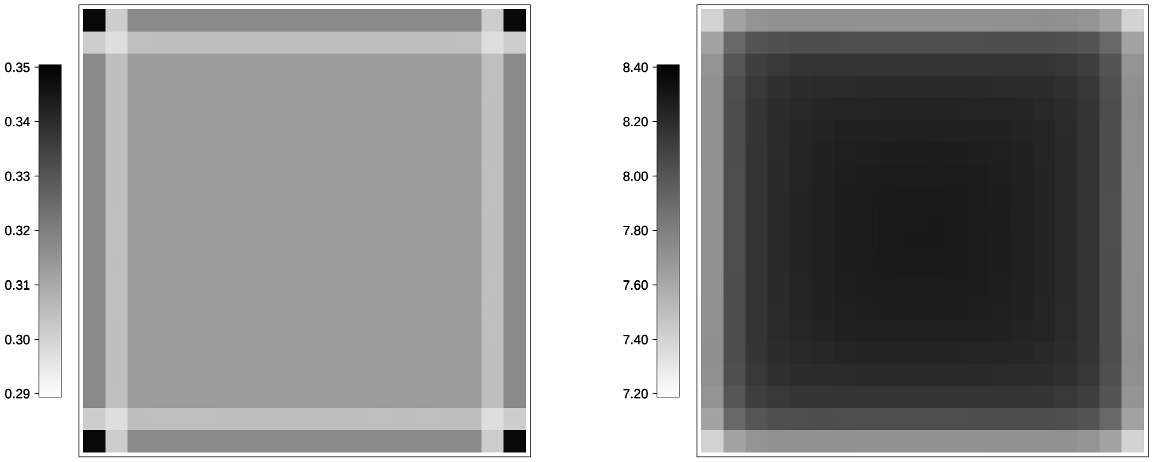


Figure C.3: Probability of being alone in a cluster (left) and average cluster sizes (right) for the 20 × 20-lattice.

C.3 fMRI-lattice

In this simulation we have used a Poisson prior with parameter $\mu = 30$ for the number of clusters k. This prior supports mainly partitions with very few clusters. Accordingly, the probabilities of being alone are extremely small (almost zero) and the average cluster sizes rather large. Striking is the influence of the one missing pixel on the average cluster sizes. Obviously, this missing pixel affects the cluster sizes of all pixels in the greater neighborhood. Still, the major conclusions drawn from the results with uniform prior in Section 4.3.3 hold true, even for this rather extreme prior.

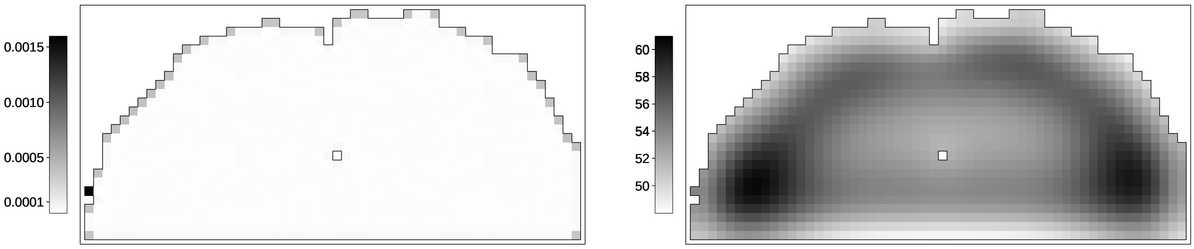


Figure C.4: Probability of being alone in a cluster (left) and average cluster sizes (right) for the fMRI-lattice.

Bibliography

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, New York: Wiley.

Albert, J. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**: 669–679.

Albert, J. & Chib, S. (2001). Sequential ordinal modeling with applications to survival data, *Biometrics* **57**: 829–836.

Arjas, E. (1996). Discussion of paper by Hartigan, *in* J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (eds), *Bayesian Statistics 5*, Oxford University Press, pp. 221–222.

Arjas, E. & Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity, *Computational Statistics* **12**: 385–402.

Assunção, R. M., Reis, I. A. & Di Lorenzo Oliveira, C. (2001). Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model, *Statistics in Medicine* **20**: 2319–2335.

Barry, D. & Hartigan, J. A. (1992). Product partition models for change point problems, *The Annals of Statistics* **20**: 260–279.

Becker, N. & Wahrendorf, J. (1997). *Atlas of Cancer Mortality in the Federal Republic of Germany 1981–1990*, Berlin: Springer.

Bernardinelli, L., Clayton, D. & Montomoli, C. (1995a). Bayesian estimates of disease maps: How important are priors?, *Statistics in Medicine* **14**: 2411–2431.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M. & Songini, M. (1995b). Bayesian analysis of space-time variation in disease risk, *Statistics in Medicine* **14**: 2433–2443.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*, Chichester: Wiley.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society Series B* **36**: 192–236.

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

Besag, J. E. & Kooperberg, C. (1995). On conditional and intrinsic autoregressions, *Biometrika* **82**: 733–746.

Besag, J. E., York, J. & Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion), *Annals of the Institute of Statistical Mathematics* **43**: 1–59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A. & Conlon, E. M. (1999). Bayesian methods for spatially correlated disease and exposure data., *in* J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith (eds), *Bayesian Statistics 6*, Oxford: Oxford University Press, pp. 131–156.

Blot, W. J., Devesa, S. S., McLaughlin, J. K. & Fraumeni, J. F. (1994). Oral and pharyngeal cancers, *in* R. Doll, J. F. Fraumeni & C. S. Muir (eds), *Cancer Surveys: Trends in Cancer Incidence and Mortality, Vol. 19/20*, New York: Cold Spring Harbor Laboratory Press, pp. 23–42.

Böhning, D., Dietz, E. & Schlattmann, P. (2000). Space-time mixture modelling of public health data, *Statistics in Medicine* **19**: 2333–2344.

Box, G. E. P. & Tiao, G. C. (1992). *Bayesian inference in statistical analysis*, New York, Chichester: John Wiley & Sons.

Breslow, N. E. & Day, N. E. (1987). *Statistical Methods in Cancer Research, vol. 2, The Design and Analysis of Cohort Studies*, Lyon: International Agency for Research on Cancer.

Brezger, A., Kneib, T. & Lang, S. (2002). *BayesX: Software for Bayesian inference based on Markov chain Monte Carlo simulation techniques, Version 0.9*, Ludwig-Maximilians-Universität München.

Carlin, B. P. & Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Chu, C. K., Glad, I. K., Godtliebsen, F. & Marron, J. S. (1998). Edge-preserving smoothers for image processing, *Journal of the American Statistical Association* **93**: 526–541.

Clayton, D. G. (1996). Generalized linear mixed models, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, London: Chapman & Hall, pp. 275–301.

Clayton, D. G. & Bernardinelli, L. (1992). Bayesian methods for mapping disease risks, *in* J. Cuzick & P. Elliot (eds), *Small Area Studies in Geographical and Environmental Epidemiology*, Oxford: Oxford University Press, pp. 205–220.

Clayton, D. G. & Kaldor, J. (1987). Empirical Bayes estimates of age–standardized relative risks for use in disease mapping, *Biometrics* **43**: 671–681.

Cox, D. R. & Hinkley, D. V. (1974). *Theoretical Statistics*, London: Chapman and Hall.

Dawid, A. P. & Lauritzen, S. L. (2001). Compatible prior distributions, *in* E. I. George (ed.), *Bayesian Methods with Applications to Science, Policy and Official Statistics*, Luxembourg: Office for Official Publications of the European Communities, pp. 109–118.

Denison, D. G. T., Adams, N. M., Holmes, C. C. & Hand, D. J. (2002). Bayesian partition modelling, *Computational Statistics & Data Analysis* **38**: 475–485.

Denison, D. G. T. & Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk, *Biometrics* **57**: 143–149.

Denison, D. G. T., Holmes, C. C., Mallick, B. K. & Smith, A. F. M. (2002). *Bayesian methods for nonlinear classification and regression*, Chichester: Wiley.

Diggle, P. J. (1996). Spatial analysis in biometry, *in* P. Armitage & H. A. David (eds), *Advances in Biometry*, New York: Wiley & Sons, pp. 363–384.

Fahrmeir, L., Gössl, C. & Hennerfeind, A. (2003). Spatial smoothing with robust priors in functional MRI, *in* M. Schwaiger & O. Opitz (eds), *Exploratory Data Analysis in Empirical Research: University of Munich, March 14-16, 2001*, Berlin, Heidelberg, NewYork: Springer, pp. 50–57.

Fahrmeir, L. & Knorr-Held, L. (2000). Dynamic and semiparametric models, *in* M. Schimek (ed.), *Smoothing and Regression: Approaches, Computation and Applications*, New York: Wiley & Sons, chapter 18, pp. 513–544.

Fahrmeir, L. & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors, *Journal of the Royal Statistical Society Series C (Applied Statistics)* **50**: 201–220.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2nd edn, New York: Springer.

Fernández, C. & Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach, *Journal of the Royal Statistical Society B* **64**: 805–826.

Gangnon, R. E. & Clayton, M. K. (2000). Bayesian detection and modeling of spatial disease clustering, *Biometrics* **56**: 922–935.

Giudici, P., Knorr-Held, L. & Raßer, G. (2000). Modelling categorical covariates in Bayesian disease mapping by partition structures, *Statistics in Medicine* **19**: 2579–2593.

Gössl, C., Auer, D. P. & Fahrmeir, L. (2000). Dynamic models in fMRI, *Magnetic Resonance in Medicine* **43**: 72–81.

Gössl, C., Auer, D. P. & Fahrmeir, L. (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging, *Biometrics* **57**: 554–562.

Gould, R. (1988). *Graph Theory*, Menlo Park, California: Benjamin/Cummings.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**: 711–732.

Green, P. J. (2003). Trans-dimensional Markov chain Monte Carlo, *in* P. J. Green, N. L. Hjort & S. Richardson (eds), *Highly Structured Stochastic Systems*, Oxford: Oxford University Press, pp. 179–198.

Green, P. J. & Richardson, S. (2000). Spatially correlated allocation models for count data, *Technical report*, University of Bristol.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping, *Journal of the American Statistical Association* **97**: 1055–1070.

Green, P. J. & Sibson, R. (1978). Computing Dirichlet tessellations in the plane, *The Computer Journal* **21**: 168–173.

Hartigan, J. A. (1990). Partition models, *Communications in Statistics: Theory and Methods* **19**: 2745–2756.

Hastie, T. & Tibshirani, R. (2000). Bayesian backfitting (with discussion), *Statistical Science* **15**: 196–223.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika* **57**: 97–109.

Heikkinen, J. & Arjas, E. (1998). Nonparametric Bayesian estimation of a spatial Poisson intensity, *Scandinavian Journal of Statistics* **25**: 435–450.

Holmes, C. C., Denison, D. G. T. & Mallick, B. K. (1999). Bayesian partitioning for classification and regression, *Technical report*, Imperial College, London.

Johnson, V. E. (1994). A model for segmentation and analysis of noisy images, *Journal of the American Statistical Association* **89**: 230–241.

Kelsall, J. E. & Diggle, P. J. (1998). Spatial variation in risk of disease: a nonparametric binary regression approach, *Journal of the Royal Statistical Society Series C (Applied Statistics)* **47**: 559–573.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in Medicine* **19**: 2555–2567.

Knorr-Held, L. (2003). Some remarks on Gaussian Markov random field models for disease mapping, *in* P. J. Green, N. L. Hjort & S. Richardson (eds), *Highly Structured Stochastic Systems*, Oxford: Oxford University Press, pp. 260–264.

Knorr-Held, L. & Becker, N. (2000). Bayesian modelling of spatial heterogeneity in disease maps with application to German cancer mortality data, *Allgemeines Statistisches Archiv (Journal of the German Statistical Society)* **84**: 121–140.

Knorr-Held, L. & Besag, J. E. (1998). Modelling risk from a disease in time and space, *Statistics in Medicine* **17**: 2045–2060.

Knorr-Held, L. & Raßer, G. (1999). Bayesian detection of clusters and discontinuities in disease maps: Simulations, *Technical Report 142*, SFB 386, University Munich. Available at `www.stat.uni-muenchen.de/sfb386/publikation.html`.

Knorr-Held, L. & Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps, *Biometrics* **56**: 13–21.

Knorr-Held, L., Raßer, G. & Becker, N. (2002). Disease mapping of stage-specific cancer incidence data, *Biometrics* **58**: 492–501.

Knorr-Held, L. & Rue, H. (2002). On block updating in Markov random field models for disease mapping, *Scandinavian Journal of Statistics* **29**: 597–614.

Künsch, H. (1994). Robust priors for smoothing and image restoration, *Annals of the Institute of Statistical Mathematics* **46**: 1–19.

Lagazio, C., Dreassi, E. & Biggeri, A. (2001). A hierarchical Bayesian model for space-time variation of disease risk, *Statistical Modelling* **1**: 17–29.

Lang, S. & Brezger, A. (2003). Bayesian P-splines, *Journal of Computational and Graphical Statistics* . to appear.

McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society Series B* **42**: 109–127.

Møller, J. & Waagepetersen, R. P. (1998). Markov connected component fields, *Advances in Applied Probability* **30**: 1–35.

Mollié, A. (1996). Bayesian mapping of disease, *in* W. R. Gilks, S. Richardson & D. J. Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*, London: Chapman & Hall, pp. 359–379.

Müller, H.-G. (1992). Change-points in nonparametric regression analysis, *The Annals of Statistics* **20**: 737–761.

Müller, H.-G., Stadtmüller, U. & Tabnak, F. (1997). Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps, *Journal of the American Statistical Association* **92**: 61–71.

Natário, I. & Knorr-Held, L. (2003). Non-parametric ecological regression and spatial variation, *Biometrical Journal* . to appear.

Ogata, Y. (1990). A Monte Carlo method for an objective Bayesian procedure, *Annals of the Institute of Statistical Mathematics* **42**: 403–433.

Okabe, A., Boots, B. & Sugihara, K. (1992). *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, Chichester: Wiley.

Polzehl, J. & Spokoiny, V. G. (2000). Adaptive weights smoothing with applications to image restoration, *Journal of the Royal Statistical Society B* **62**: 335–354.

Quaas, J. & Heinrich, J. (1998). Cervical cancer screening - a retrospective comparison between the old and new German federal states (in German), *Zentralblatt für Gynäkologie* **120**: 13–19.

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion), *Journal of the Royal Statistical Society B* **59**: 731–792.

Robert, C. P., Rydén, T. & Titterington, D. M. (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method, *Journal of the Royal Statistical Society B* **62**: 57–75.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields, *Journal of the Royal Statistical Society B* **63**: 325–338.

Schach, U. (2003). A type of Bayesian small area estimation for the analysis of cancer mortality data, *in* M. Schwaiger & O. Opitz (eds), *Exploratory Data Analysis in Empirical Research: University of Munich, March 14-16, 2001*, Berlin, Heidelberg, NewYork: Springer, pp. 366–374.

Schlattmann, P. & Böhning, D. (1993). Mixture models and disease mapping, *Statistics in Medicine* **12**: 1943–1950.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion), *Journal of the Royal Statistical Society B* **64**: 583–639.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion), *The Annals of Statistics* **22**: 1701–1762.

Waagepetersen, R. & Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping, *International Statistical Review* **69**: 49–61.

Wakefield, J. C., Best, N. G. & Waller, L. A. (2000). Bayesian approaches to disease mapping, *in* P. Elliot, J. C. Wakefield, N. G. Best & D. J. Briggs (eds), *Spatial Epidemiology: Methods and Applications*, Oxford: Oxford University Press.

Waller, L. A., Carlin, B. P., Xia, H. & Gelfand, A. E. (1997). Hierarchical spatio–temporal mapping of disease rates, *Journal of the American Statistical Association* **92**: 607–617.

Winkler, G. (1995). *Image analysis, random fields and dynamic Monte Carlo methods*, Vol. 27 of *Applications of mathematics*, Berlin, Heidelberg: Springer.

Wolpert, R. L. & Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics, *Biometrika* **85**: 251–267.

Lebenslauf

Günter Raßer
geboren am 22. April 1971 in München

Schulbildung

1977 – 1981	Grundschule an der Jahnstraße in Unterhaching
1981 – 1990	Gymnasium Unterhaching (mathematisch-naturwissenschaftlich)

Zivildienst

Okt. 1990 – Dez. 1991	Sozialstation Berg am Laim in München

Studium

Mai 1992 – Sep. 1992	Studium der Geographie an der Ludwig-Maximilians-Universität München
Okt. 1992 – Dez. 1998	Studium der Statistik an der Ludwig-Maximilians-Universität München mit den Anwendungsgebieten Psychologie und Soziologie
Dez. 1994	Diplom-Vorprüfung in Statistik
Dez. 1998	Diplom-Hauptprüfung in Statistik

Beruf

seit Dez. 1998	vollbeschäftigter wissenschaftlicher Mitarbeiter bei Prof. Dr. L. Fahrmeir am Institut für Statistik der Universität München und im Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen"