Vector-Autoregressive Time Series Models with Spatial Dependence

DISSERTATION

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik der Universität Dortmund
vorgelegt von

Ulrike Schach

aus Dortmund

im Mai 2002

Gutachter: Prof. Dr. Katja Ickstadt
Prof. Dr. Franz Hering

Tag der mündlichen Prüfung: 4. Juli 2002

Contents

Chapter 1

Introduction

When analyzing data on disease incidence or mortality, one way is to perform individual studies with individual information on exposition to risk factors. Statistical regression models can be used to identify the relationship between exposure to certain factors of influence and the health outcome of the individual. Such a statistical analysis is very detailed, with respect to the individuals under study. If the sample survey is appropriately chosen and large enough, risk factors can be identified and the interdependencies found in the sample can be generalized to a larger population. This individual-based approach to statistical analysis is commonly used in biometry and epidemiology. Nevertheless, many sources of error can be brought into the analysis through inappropriately chosen samples, sample surveys, misspecification of covariates, or leaving out of important risk factors. Some studies then include additional effects to account for unstructured heterogeneity.

An alternative to this person-related analysis is the more general observation of an area-specific appearance of disease incidence and mortality. Instead of data on individuals, with individual covariates, the analysis is based on aggregated numbers of disease counts with population related covariates. A crucial assumption in statistical modelling is the assumption of independence. When regarding aggregated data

in a certain region of study, this assumption is likely to be violated, as observations are now spatial aggregates over units of the study region with similar underlying mixtures of covariates. On the basis of specific models for spatially dependent data, an epidemiologic approach can possibly lead to the identification of risk factors on the area level. The data underlying this analysis do not only have a spatial correlation structure, but consist of repeated observations over time. Therefore, we extend the models to a temporal dimension.

In this thesis we model data on cancer mortality. Geographical differences in frequency of occurrence of the various forms of cancer are largely due to environmental risk factors and personal behavioral patterns. Person-related data, however, on variables like occupation, family status, social status, nutrition, smoking habits, or alcohol and drug consumption are usually not available on a personal basis, but sometimes on an areal level. A strong interaction and mixing of the factors within the individuals has the effect that an analysis based on only these factors must be seen with caution. Considering the fact that the time lag between exposure and development of the cancer disease usually takes years, direct or even causal effects will not be easy to identify. We therefore include area specific random effects to account for this unobservable mixture of known and unknown covariates.

The aim of this thesis is the presentation of autoregressive statistical models that account for dependence structures within the area specific data. We present modelling approaches that lead to spatial smoothing after the exclusion of covariate effects. The study area for this analysis is the Federal Republic of Germany, for which the data set is called Germany data, and the state of North Rhine Westphalia (NRW) in particular, with NRW data. For Germany we consider mortality counts of stomach cancer and lung cancer among men and women. These two cancer types show a strong spatial clustering, and are therefore suitable for a spatial modelling approach. The data on Germany have been kindly provided by the German Cancer

Research Institute in Heidelberg, and they are published in two different data sets, e.g. Becker et al. (1984), pp. 51-66 and pp. 153-168. Due to privacy protection laws, data with the highest spatial and temporal resolution are not available for Germany. Therefore, the data set is published at different levels of temporal and spatial aggregation. More precisely, one data set contains annual mortality data from 1976 to 1990 given in the 30 regions (Regierungsbezirke) of West Germany, aggregated over the 328 districts (Kreise). On the other hand, temporally aggregated data are available with a higher spatial resolution of 328 districts for Germany. Here, we have data for the three five-year blocks 1976-1980, 1981-1985, and 1986-1990.

Data on cancer mortality for east Germany (former GDR) is only available from 1981 on. When modelling the dependence structure and changes over space and time, we therefore consider only the data on West Germany, in order to have full lengths of the series. This implies for the regions of Germany, that we base the analysis on the 30 regions of West Germany, excluding West Berlin, which lies too far apart from the rest of the study area to exhibit local correlation.

Apart from the data on lung cancer and stomach cancer mortality for Germany, we analyze data for North Rhine Westphalia. Here we have chosen lung cancer among women because this data set contains, apart from the positive spatial correlation, a clear upward temporal trend. The data has been kindly provided by the Landesinstitut für den Öffentlichen Gesundheitsdienst in Bielefeld, see e.g. Nolte et al. (1997). The difference of this data set is that yearly mortality counts are available for 19 years from 1980-1998, with the highest spatial resolution based on 54 districts.

We begin with an exploratory data analysis in the following chapter, describing the temporal as well as the spatial structures within the data. In chapter 3 we give a short summary of Bayesian modelling. Emphasis is put on specific tools for Markov chain Monte Carlo simulation. As most of the inference in this thesis is drawn

from Bayesian approaches we describe sampling algorithms that are required for the stochastic simulation. Additionally we introduce diagnostic methods to check for convergence and mixing of the Markov chains. The idea of sensitivity analysis is presented to evaluate the influence of prior settings on the posterior outcome. Only recently the deviance information criterion (DIC) (Spiegelhalter et al. 2002) has been proposed as a basis for choosing between Bayesian models. We describe this measure of fit at the end of chapter 3.

In chapter 4 we present vector-autoregressive and innovation processes in preparation for a joint modelling of spatial and temporal dependence. We begin with a purely temporal analysis of lung cancer among women in NRW, and stomach cancer mortality among men for Germany. Partial autocorrelation functions are used to determine the order of the vector autoregressive process. Additionally, we introduce Markov random fields for the spatial analysis of temporal slices of the NRW data set in a Bayesian framework. We use a surrogate for smoking behavior as a covariate to model the lung cancer mortality rates among women, which we deduce from the population density of the spatial units. We compare a Gaussian and a Poisson modelling approach with respect to posterior mean estimates.

The knowledge that we have gained from separate temporal and spatial analyses will be used in chapter 5 to estimate spatial and temporal autocorrelation parameters in a combined space-time model. The idea is to present a classical and a Bayesian modelling approach, and to compare the resulting parameter estimates with supported range for the data set on stomach cancer in Germany and on lung cancer in NRW. We use the DIC to assess the model fit to the data with and without covariates. Additionally, we regard computational aspects of the frequentist and the Bayesian approach.

Based on these parameters we perform a model based type of small area estimation in chapter 6. As mentioned above, cancer mortality data with the highest

spatial and temporal resolution are not published in Germany, and we therefore aim for estimates of cancer mortality counts with that resolution. The unknown mortality counts are modelled as unknown parameters in a Bayesian setting. The problem is how to sample from the posterior distribution under the restriction of preserving the spatial and temporal marginals. To solve this problem we introduce an approach with which we can avoid to assign given data to a sum of stochastic nodes which is cumbersome in the Bayesian framework. The results of different small area allocation models will be compared and the goodness of fit is evaluated by comparing the parameter estimates with the observed data for North Rhine Westphalia.

In chapter 7 we summarize the results of the thesis and we give an outlook to further fields of application of the presented methods.

Chapter 2

Space-time Data on Cancer Mortality

2.1 Standardization

When analyzing and comparing aggregated incidence or mortality data for different years or in different regions, it is obvious that absolute case figures, i.e. raw mortality counts alone are not suitable. The underlying area specific populations usually differ in size and age structure. When aiming for the detection of regional and temporal changes, these should not be caused by superficial differences in population structure. One approach is to include the underlying population sizes and age structures into the statistical model for the mortality counts. We explain suitable Poisson models in chapter 4. Age-period-cohort (APC) models can be used when the stratification into age groups should be preserved (Knorr-Held and Rainer 2001). An alternative is to standardize the data in advance, usually with respect to age structure and gender, and model the standardized data.

If a standardization is used to make the results within different areas or years of

study comparable, the choice of a sensible standard population is crucial. Depending on the comparison one aims for, several different populations can be thought of. For instance, if the mortality rates of East and West Germany in 1976 are to be compared, the population of either one region can be chosen as the standard population. The mortality cases of the other region will be standardized with respect to that standard population. This standardization is then called internal. In our case, where we aim to draw conclusions about the same area of study in the change of years, we rather base the rates on an artificial population. This could be an aggregated population over the years for the study area. A standard population of that kind preserves the temporal trend within the data and population specific features, and because of that it is probably the most suitable standard population. However, the rates cannot be used for an international comparison, which is desirable for cancer registries. They suggest using a European standard population, or the Segi world population (Becker et al. 1984, p. 5, and Becker 1998). The cancer mortality rates of the German cancer atlas are age-standardized mortality rates with weights according to Segi's world population. We have chosen the same approach for the standardization in this thesis. Figure 2.1 gives an idea of how the artificial Segi population weights for age compare to the German age structure in 1976. As one can see clearly, there are considerably more older people in the German population, which gives these age groups more weight compared to Segi's population. The cancer mortality rates will therefore be underestimated systematically when standardizing with the world population. For the interpretation of the standardized rates this should be kept in mind.

Apart from the choice of a suitable internal or external standard population, the standardization can be undertaken directly and indirectly. From the epidemiological point of view an indirect standardization is mainly used for rare incidence or mortality (Kreienbrock and Schach 2000, p. 41). We are considering lung cancer and stomach cancer, which are relatively frequent causes of death. Therefore, a direct

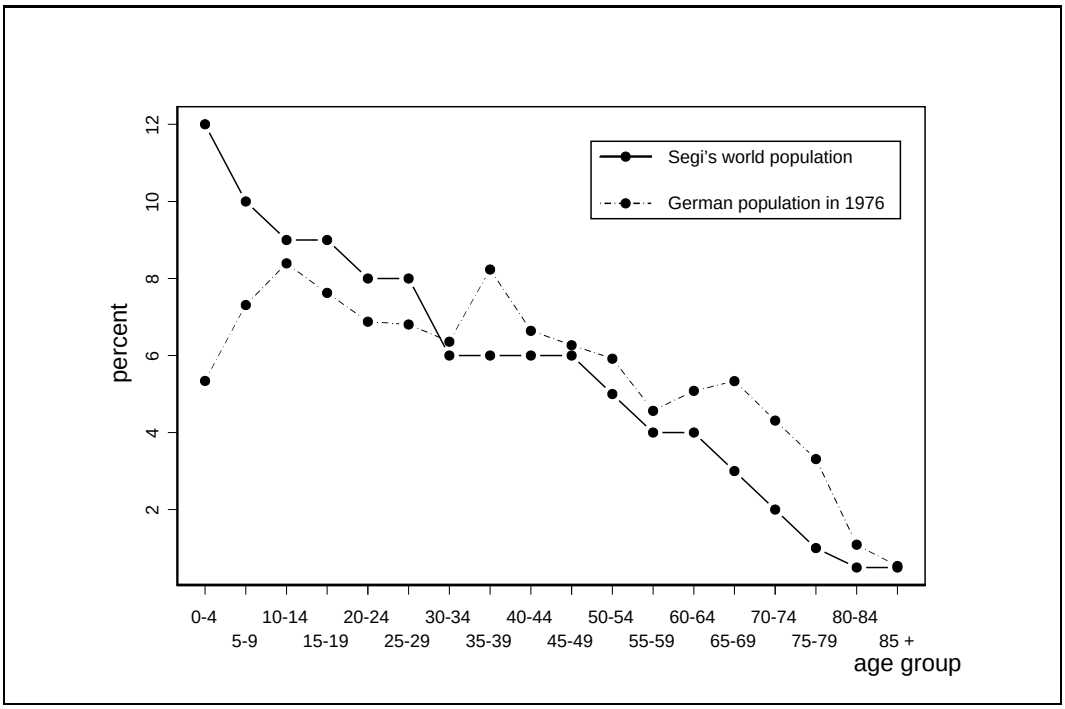

Figure 2.1: Age structure of the standard population after Segi, and of the German population in 1976.

standardization is more appropriate. In order to actually calculate the rates, we need the following notations, where k denotes the age group, with $k = 1, \ldots, K$ for the 18 individual age groups 0-4 years, 5-9 years, ... , 80-85 years, 85 years and older. The spatial index i runs from $i = 1, \ldots, I$ through the administrative units of the study region, and $t = 1, \ldots, T$ indicates the temporal index. The standardization is performed separately for each gender. Furthermore, define

D_{kti} := number of deaths of the defined cancer type in the study population

n_{kti} := number of people at risk in the study population

n^*_{kti} := number of people at risk in the standard population.

The population figures n_{kti} are based on official population figures per year. The data have been obtained from the Federal Bureau of Statistics (Statistisches Jahrbuch der Bundesländer). They are based on yearly averages of monthly compilations of

community figures (Melderegister). On the basis of these figures, the raw mortality rates mr_{ti}, with

$$\mathrm{mr}_{ti} := \frac{\sum_{k=1}^{K} D_{kti}}{\sum_{k=1}^{K} n_{kti}} \cdot 100,000 \tag{2.1}$$

can be calculated. The age-specific mortality rates are defined through $\mathrm{mr}_{kti} := (D_{kti}/n_{kti}) * 100,000$, $k = 1 \ldots, K$, and based on them, the mortality rates can be rewritten as

$$\mathrm{mr}_{ti} = \frac{\sum_{k=1}^{K} n_k \, \mathrm{mr}_{kti}}{\sum_{k=1}^{K} n_k} = \sum_{k=1}^{K} W_k \, \mathrm{mr}_{kti}. \tag{2.2}$$

The weight $W_k := n_{k..} / \sum_{k=1}^{K} n_{k..}$ adjusts for the proportion of people in the k-th age group of the study population. If instead of the age structure of the study population the age structure of the standard population is used, the resulting mortality rate is a standardized mortality rate, as stated in Kreienbrock and Schach (2000), p. 38. Assuming a constant structure of the standard population over space and time, the standardized mortality rate can be written as

$$r_{ti} := \frac{\sum_{k=1}^{K} n_k^* \, \mathrm{mr}_{kti}}{\sum_{k=1}^{K} n_k^*} = \sum_{k=1}^{K} W_k^* \, \mathrm{mr}_{kti}, \tag{2.3}$$

where W_k^*, $k = 1, \ldots, K$, is the weight function for the age structure of the standard population. An asterisk indicates that the population figures are based on the standard population. The use of this type of standardization, with a constant standard population after Segi, preserves the temporal trend within the data.

The data on both lung cancer and stomach cancer have been standardized with the described method, each cancer type stratified by the variables gender, individual age group, region, and year.

2.2 Cancer types under examination

There are several reasons, why lung cancer and stomach cancer have been chosen for this analysis. First of all, both cancer types show relatively high average standardized mortality rates of more than 20 per 100,000 for stomach cancer in the study period and even more than 50 per 100,000 for lung cancer among men. The corresponding figures for women lie around 12 per 100,000 for stomach cancer and between 5 and 10 per 100,000 for lung cancer. Therefore, the analysis of the spatial and temporal tendencies and changes is of outstanding importance. As mentioned before, the aim of this analysis is only to a minor extent to look for yet unknown causes for these cancer types. We rather aim for the detection of regional differences and temporal changes in order to make spatio-temporal predictions and extrapolations. Both lung cancer and stomach cancer show distinct regional patterns, and strong spatial clustering. For stomach cancer, a strongly decreasing trend is visible, and for lung cancer, the temporal structure seems to depend on the region.

2.2.1 Lung cancer

There is much information available about the causes of lung cancer. The fact that lung cancer mortality is increasing, is partly due to a refusal to apply this knowledge about established risk factors. Smoking is by far the most important risk factor, but also there have been surveys to show that asbestos dust, especially with exposition at the work site, plays an important role. Air pollution, however, does not seem to have a considerable influence, as described by Becker et al. (1984), p. 155. However, Wichmann et al. (1991) claim, that for those regions in Germany, in which the air pollution is high, the mortality rates of lung cancer are higher than expected when only the risk factors smoking and exposition at work are taken into account.

The time lag between being exposed to one or more of the major risk factors

and developing or even dying of lung cancer ranges between 10 to 30 years. As described by Becker et al. (1984), p. 153, there is an age specific increase in mortality, especially when the persons at risk are 60 years and older. This tendency has been found for both genders. Regarding the regional distribution of death rates, above average death rates have been found in western parts of Germany, especially in Rhineland-Palatinate, Saarland, North Rhine Westphalia, and the city states Bremen and Hamburg. The regional differences are very similar for males and females. However, female lung cancer death cases are only about 1:10 of those of men. For men the general tendency is an increase of the observed rates. For women, except for the regions Schleswig-Holstein, Lower Saxony, Hesse, Baden Wurthemberg, Bavaria, and Saarland, the same holds. When regarding the level of urbanization of the observed region, there seems to be an indication for lower death rates in more rural areas and vice versa. According to Becker et al. (1984), p. 155, this is due to the fact, that there are more smokers in urban than in rural areas. Additionally, not only the percentage of smokers (for both male and female) grows with the number of inhabitants of a city, also the number of cigarettes smoked among those people who smoke increases, as stated by Wichmann et al. (1991). On the other hand, Munich and Stuttgart have strikingly low mortality rates compared to other conurbations, in all of which mortality is elevated, as stated by Becker et al. (1984), p. 155. This phenomenon remains unexplained by knowledge about lung cancer risk factors. Figure 2.2 shows the spatial and temporal relative change of age-standardized lung cancer mortality rates for women in North Rhine Westphalia. A clear temporal increase in lung cancer mortality among women over the 15 years from 1981 to 1995 can be seen in figure 2.2. The spatial clustering also seems to increase over that period of time. Höpker (1998), however, points out that maps of un-smoothed standardized rates should be interpreted with caution. The districts with the highest rates tend to cluster around the densely populated area called Ruhr area (Ruhrgebiet).

When analyzing area specific count data, the problem usually arises that areas

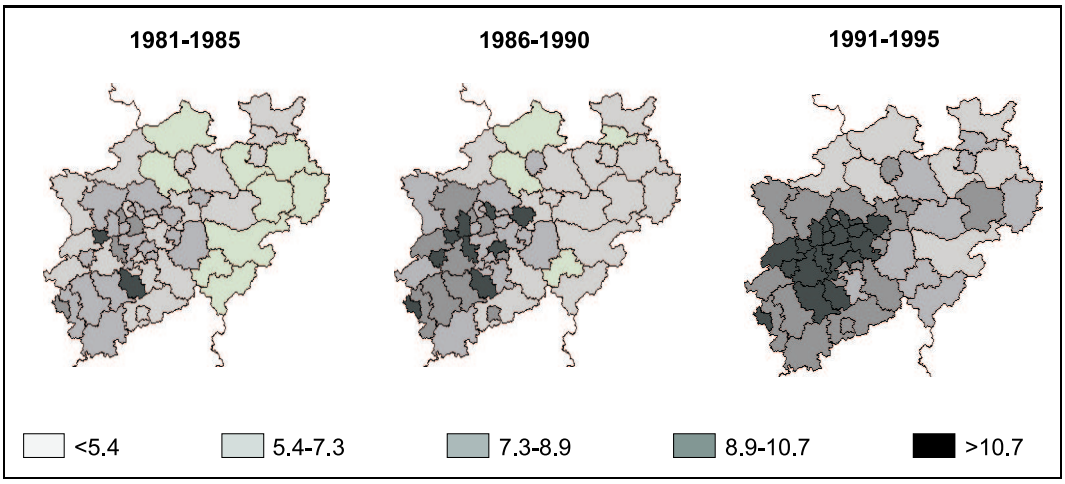

Figure 2.2: Lung cancer among women in NRW. Age-standardized mortality rates over 15 years.

with small population sizes show more variation than densely populated areas. This problem is called over-dispersion. It can be accounted for through a suitable transformation. Cressie (1993), p. 395, proposes to use a square-root transformation, a log transformation, or a transformation after Freeman and Tukey (1950). Figure 2.3 displays the dependence of the standard deviation on the mean of the 54 districts in North Rhine Westphalia over the study period from 1981-1995. As the dependence of the standard deviation on the mean is approximately linear, a log transformation has been chosen as it is "stronger" than a square root transformation in order to remove this dependence. Figure 2.4 shows how the dependence can be removed with the transformation. For Gaussian modelling we will use log transformed age standardized mortality rates throughout the analysis, also in the case of stomach cancer.

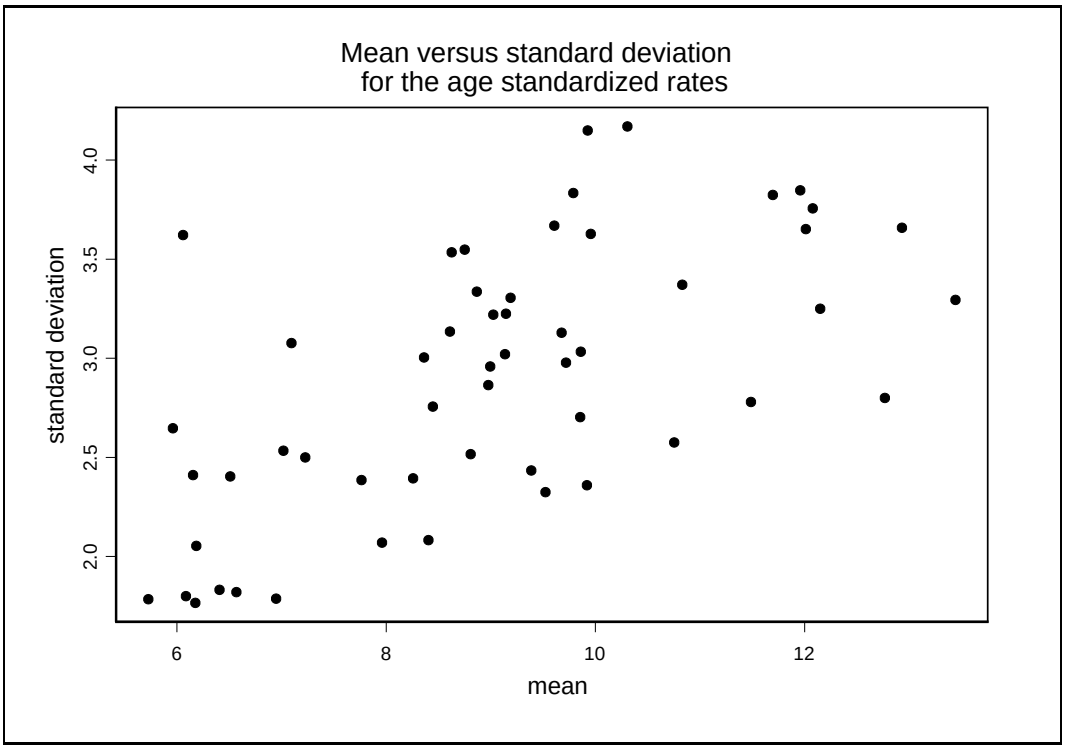

Figure 2.3: Lung cancer among women in NRW. Age standardized rates: mean versus standard deviation.

2.2.2 Stomach cancer

Non only in Germany the frequency of stomach cancer is clearly decreasing. Especially in the Western industrialized countries and with a time lag also in Eastern Europe, incidence as well as mortality have decreased over the past 20 years. The comparison between male and female mortality shows, that the mortality among males with mortality rates of 15-20 per 100,000 is more than double the mortality among women with rates among 5-15 per 100,000. The spatial distribution of stomach cancer mortality reveals large regional differences with relatively low rates in the north of Germany. In the south of Germany, especially in the area of Bavaria, the mortality rates for men as well as for women are rather high. Although the temporal trend shows a steady decrease, according to Becker et al. (1984), p. 51,

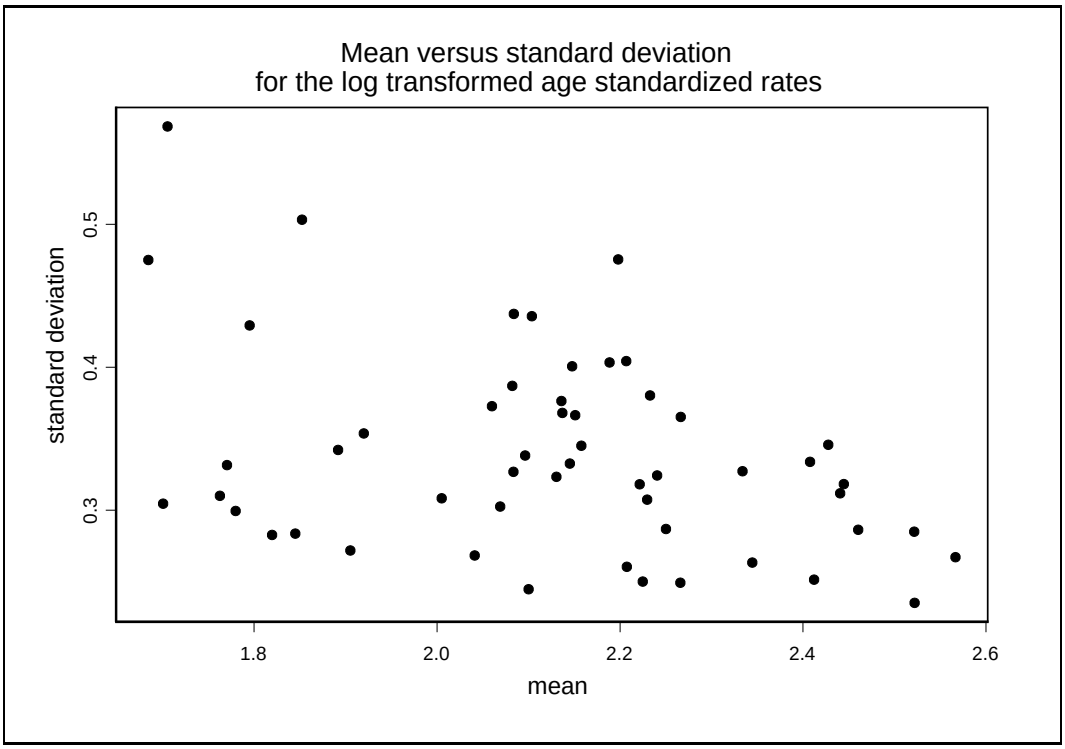

Figure 2.4: Lung cancer among women in NRW. Log-transformed age-standardized rates: mean versus standard deviation.

the regional differences remain constant. Figures 2.5 and 2.6 display the temporal changes in age standardized mortality rates for the districts within West Germany. In the box plots the length of the whiskers is 1.5 times the inter quartile range, if the minimum or maximum has not been reached before. Mortality rates that are not lying within the interval of 1.5 times the inter quartile range are displayed as individual lines, above or below the whiskers. The two outlying regions in both figures for males and females are Lower Bavaria and Upper Palatinate, which show extraordinary high rates throughout the study period.

The etiology of stomach cancer is discussed for example in Risch et al. (1985), or in Willett and McMahon (1984). Various endogenous and exogenous factors have been identified either as protective or as risk factors. An increased consumption of

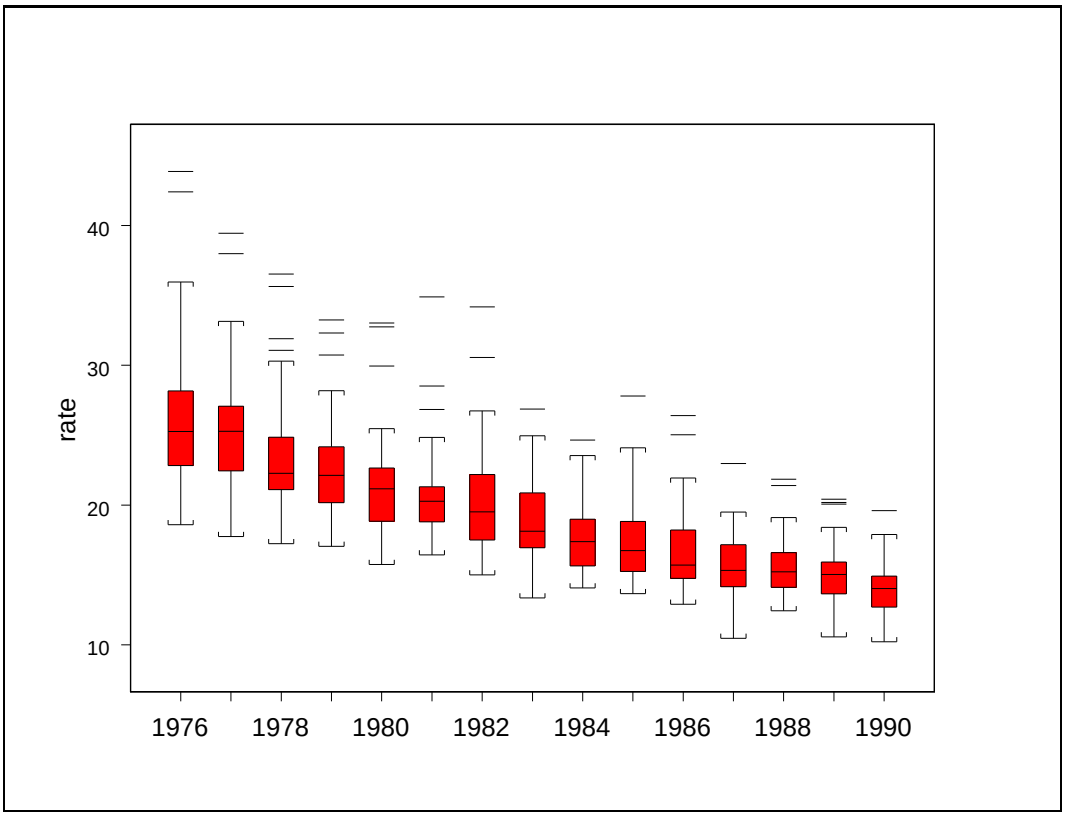

Figure 2.5: Stomach cancer mortality rates among men in Germany. Box plots of the temporal trend.

starchy cereal products, strongly salted food, preserved vegetables, and smoked food seems to increase the risk of stomach cancer. Becker et al. (1984), p. 51, however, name personal and familial disposition, and typical occupational groups as special risk factors.

2.3 Exploratory data analysis

The descriptive analysis in this section will be performed for the Germany data set, i.e. a data set with a yearly temporal resolution, which is spatially aggregated over 30 administrative regions (Regierungsbezirk, RB) for the period from 1976 to 1990.

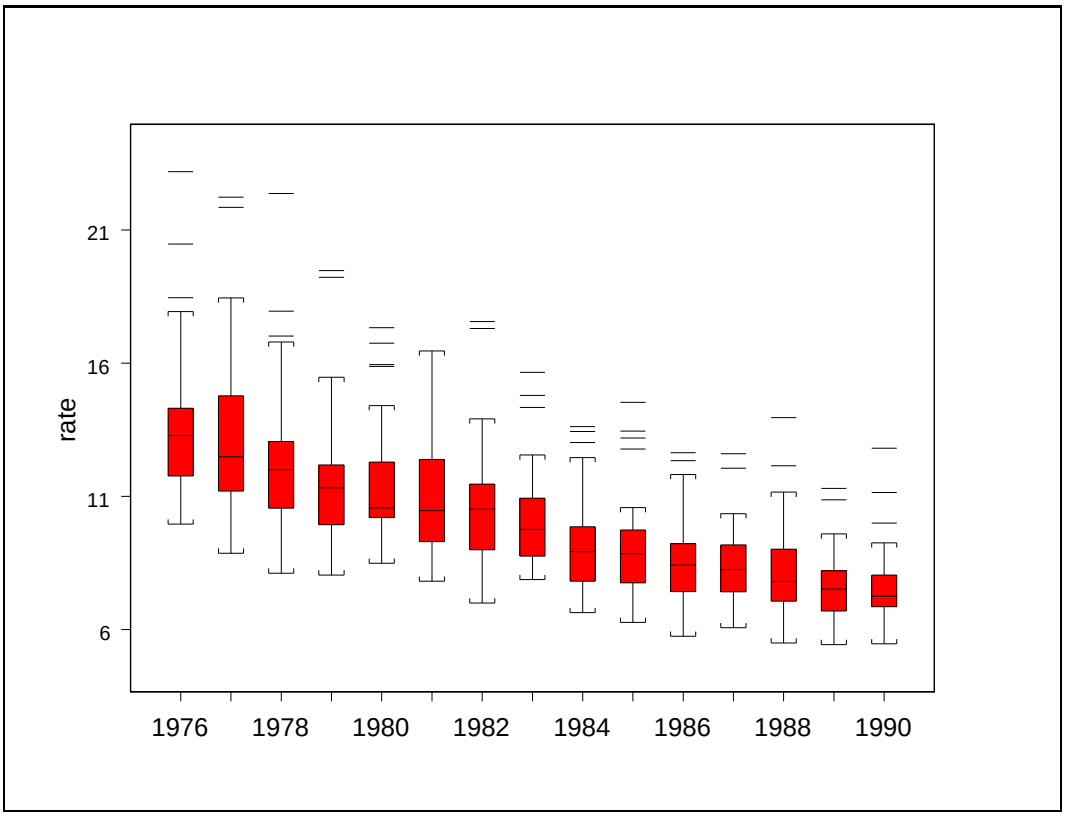

Figure 2.6: Stomach cancer mortality rates among women in Germany. Box plots of the temporal trend.

We begin with the analysis of lung cancer among men. The highest mortality rate for men of 75.38 per 100,000 persons at risk has occurred in the region of Trier in the year of 1981. The mortality rate of the Saarland with 75.34 in 1986 is only a little smaller. The lowest rates have been found in Tübingen. The rate here was only 30.42 in the year of 1989. When considering the data set by region, the maximum mean has been found in Düsseldorf, with an average rate of 66.14 over the 15 years from 1976 to 1990. The minimum rate occurred in Stuttgart. Here the mean over the observed period of time is 33.47. It remains to mention that mean and median of the observed time series for the rates differ only to a very small extent, as displayed in figure 2.7. This indicates that the data can be considered

symmetric. The standard deviation of the regions varies between a minimum of 1.13

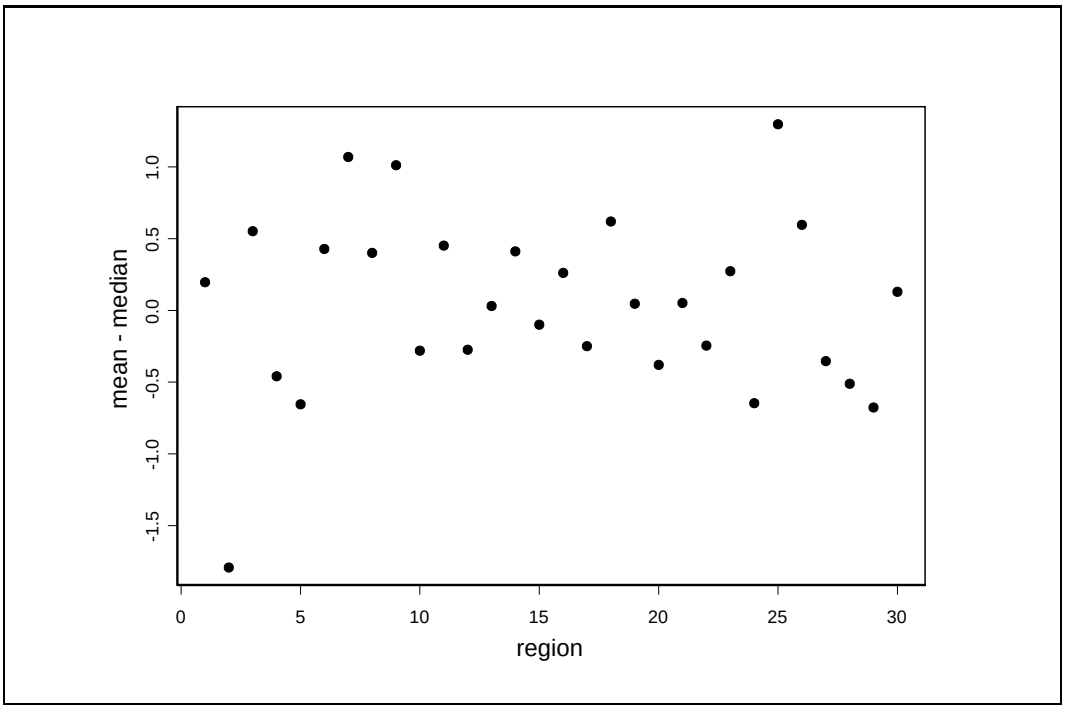

Figure 2.7: Lung cancer among men in Germany. Difference of mean and median.

in Stuttgart and a maximum of 5.51 in the region of Trier. Considering the data set now over time instead of region, the year with the lowest average mortality rate is 1976, the beginning of the observation period. In 1978, the highest average rates have been found. The mean lies at 47.66 per 100,000.

A descriptive analysis of female lung cancer mortality rates reveals that the lowest cancer rate of 2.8 for women in the observed period of time has been found in Upper Franconia in the year of 1979. The maximum rate has been observed in Bremen in 1990. Here the mortality rate was 13.45. Again, the regional differences over time shall be examined. The resulting average rate is 6.0, with a minimum of 4.07 for Tübingen and a maximum of 10.07 for Hamburg. The standard deviation varies between 0.61 and 2.64 among the regions. The analysis of female lung

cancer for the 15-year observation period reveals a strong upward temporal trend, as visible in figure 2.8. The increase of female lung cancer mortality over the years

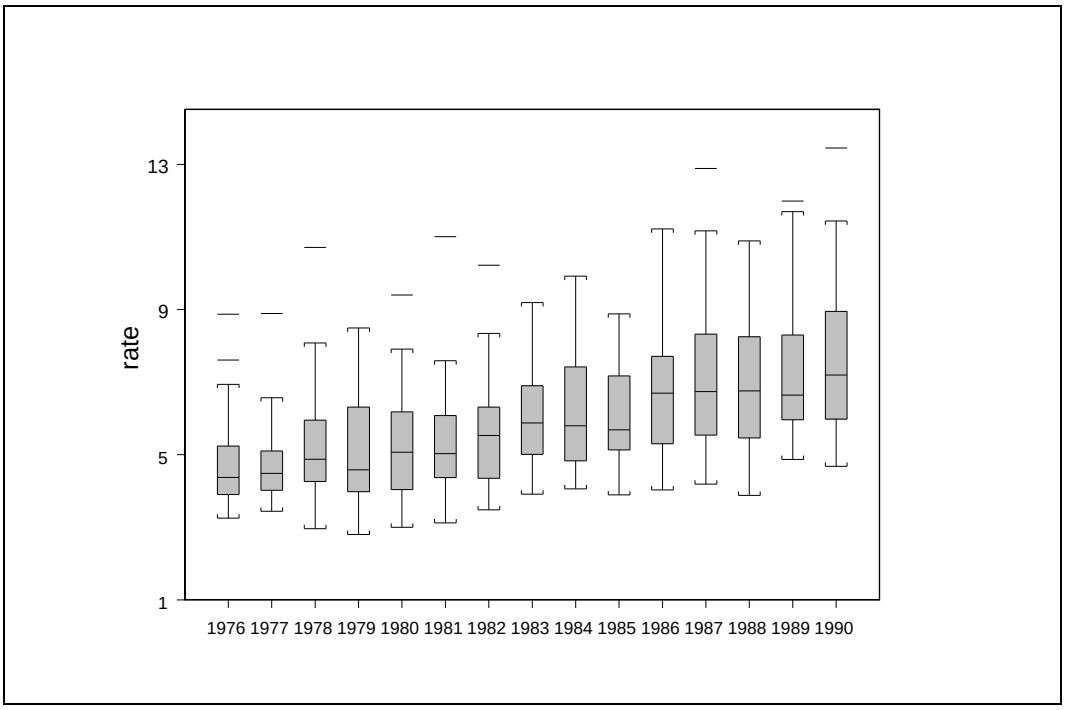

Figure 2.8: Lung cancer mortality among women in Germany. Box plots of the temporal trend.

is indicated by the increasing medians over the observation period. As mentioned in section 2.2.1, the city states show higher mortality rates for women, than the more rural areas. Figure 2.9 shows the age standardized rates for the 54 districts (5 RB's) in 1995 in North Rhine Westphalia. The two different symbol sizes indicate whether a district is urban or rural. So called "Stadtkreise" are considered as urban, whereas "Landkreise" are labeled rural. Figure 2.9 shows a tendency of an increased mortality rate in urban districts. A question is, whether there are not only higher rates in urban districts, but also whether a stronger increase over time can be observed. Figure 2.10 displays the average mortality rates of the urban districts compared to the average mortality rates of the rural districts for NRW over time.

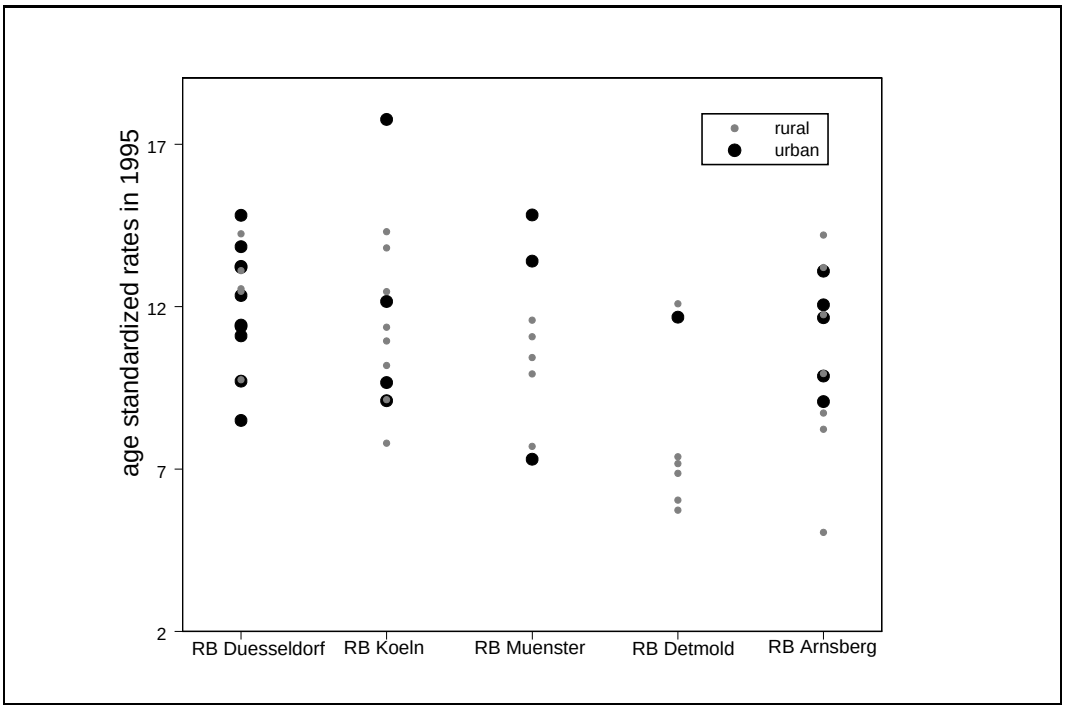

Figure 2.9: Lung cancer mortality among women in NRW. Rural and urban districts in 1995.

We can see that throughout the observation period the average rates of the urban districts lie above those of the rural districts, which clearly shows an effect of the risk factor urbanization. Least squares fits of the data demonstrate that the slope for the urban time series is significantly higher than that for the rural population. Because of that, we must consider the degree of urbanization as an effect modifier.

For an exploratory analysis of stomach cancer we again consider male and female mortality separately. Looking at the stomach cancer mortality rates of men, the lowest rate of 10.22 can be found in Trier in the year of 1990. The highest rate of 43.87 has been observed in the region of Upper Palatinate in 1976. This indicates that there has been a decrease in mortality, as will be described in more detail. When averaging over time of the 30 regions under examination, the lowest mean has been found in Giessen, the highest in Upper Palatinate. The standard deviation

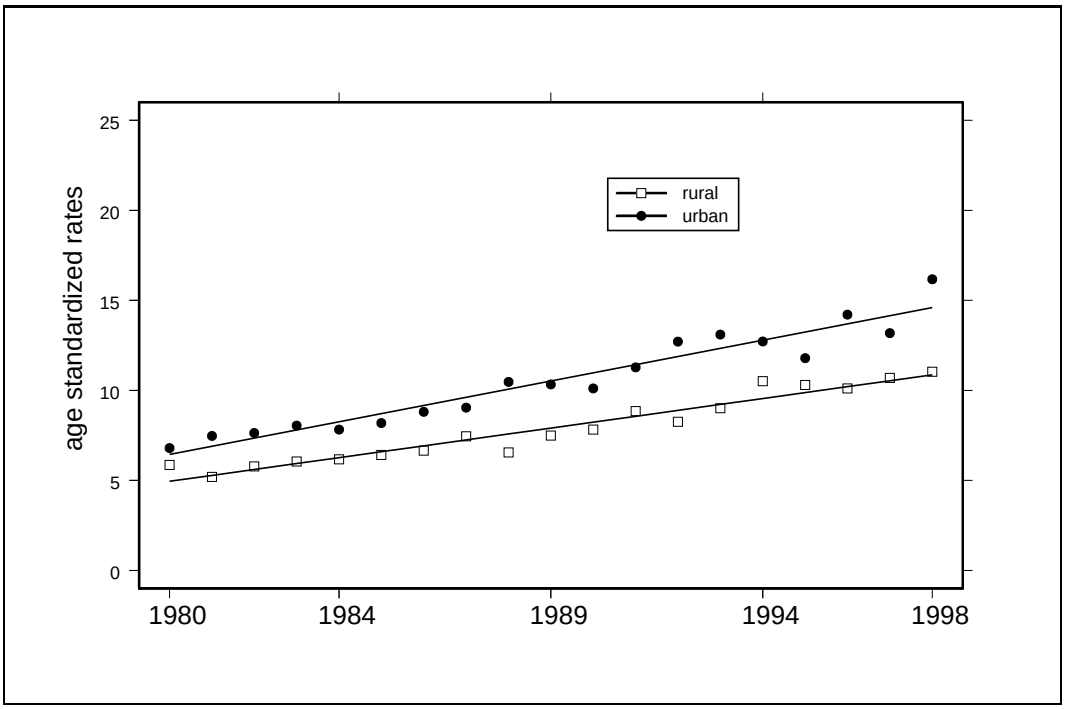

Figure 2.10: Lung cancer mortality among women in NRW. Least squares smoothed rates over time.

lies between a minimum of 2.83 and a maximum of 7.58, with a mean value of 4.20. The temporal development of the mortality rates is clearly decreasing, as described in the previous section. The highest average rate over the regions of 26.64 has been observed in 1976, the beginning of the study period. The lowest average of 14.21 occurred in 1990, the end of the study period. Therefore, even though regional differences can be found within male stomach cancer, the general temporal trend turns out to be strongly monotonously decreasing over time.

The female stomach cancer mortality rates seem to develop rather similar to those of men. The minimum observation of 5.44 has been found in Trier in 1989, in the same region with the lowest male stomach cancer rate. The highest rate of 23.19 occurred in Lower Bavaria in 1976. The year of this observation is the same, as the year of the highest male rate, and the region of Lower Bavaria is a neighboring region

of Upper Palatinate. Looking at the regional averages over time, the minimum rate of 8.10 can be found in Giessen, as for men, and the highest rate of 15.81 has been observed in Upper Palatinate, again as for men. The standard deviation varies between 1.35 and 3.94, with an average of 2.24. When averaging over the regions and analyzing the data temporally, the highest average rates have been found in the year of 1976, the lowest in the year of 1989. The temporal development of male and female stomach cancer mortality therefore runs almost parallel, as visible in figures 2.5 and 2.6.

2.4 Related data analyses

In the foregoing sections we have described data on cancer mortality that are suitable for a combined spatio-temporal analysis. For this thesis we have concentrated on lung cancer and stomach cancer mortality. Obviously other cancer types, other causes of death, or disease incidence data can be analyzed with the methods presented in the following chapters. Well-known data sets suitable for an analysis with the methods of this thesis are the data set on Ohio lung cancer (Xia et al. 1997), and the data set on Scottish lip cancer (Cressie 1993, p. 537). Not only epidemiologic data, but data on socio-demographic variables like unemployment rates, average income, or deprivation indices can be used for a spatio-temporal analysis, as well. Furthermore, when considering regular lattices, the methods are useful for image analysis, as described by Besag, York, and Mollié (1991). Recently published work by Gössl, Auer, and Fahrmeir (2001) deals with spatio-temporal imaging of functional magnetic resonance, or applications in ophthalmology (Krahnke 2001).

Chapter 3

Bayesian Modelling and Inference

This chapter deals with Bayesian aspects of modelling and inference in preparation for the analysis of our space-time data. We begin with a short passage about preliminaries of Bayesian theory, and we continue with the specification of Markov chain Monte Carlo (MCMC) tools for Bayesian inference. We assume that the observations supporting the statistical analysis, $z_1, \ldots, z_n$, are sampled from a parameterized probability distribution. Thus, $f(z_i \mid \theta)$ in $\mathbb{R}^p$ is the density from which z_i is drawn and the parameter or parameter vector θ is unknown. In order to analyze parameter vectors of this type, we introduce the definition of a parametric statistical model.

Definition 3.1 *Let θ denote a subset of a finite dimensional vector space, let $f(z \mid \theta)$, $\theta \in \Theta$ be a family of probability distributions, and let Z be a random variable with probability density $f(z \mid \theta)$. Then a sample $z_1, \ldots, z_n$ of Z is a parametric statistical model.*

3.1 Introduction to Bayesian theory

At the end of the eighteenth century, Bayesian statistics was often called inverse probability, as statistical analysis is fundamentally based on an inversion. Statistical inference aims at retrieving the causes from the effects, which are the observations (Robert 1994, p. 7).

Theorem 3.1 *[Bayes Theorem] If A and E are events such that $P(E) \neq 0$, $P(A \mid E)$ and $P(E \mid A)$ are related by*

$$P(A \mid E) = \frac{P(E \mid A)\, P(A)}{P(E \mid A)\, P(A) + P(E \mid A^c)\, P(A^c)} = \frac{P(E \mid A)\, P(A)}{P(E)}. \tag{3.1}$$

According to Robert (1994), p. 8, equation 3.1 appears as a major conceptual step in the history of Statistics, being the first inversion of probabilities. Especially a transformation of equation 3.1 shows that for two equally probable causes A and B, given a specific effect E, the ratio of their probabilities is the same as the ratio of the probabilities of the effect E, given the two causes A and B. A continuous version of this result has been proved by Bayes (1763), where the two random variables Z and θ are specified, with the conditional density $f(z \mid \theta)$ and marginal density $g(\theta)$. Then the conditional density of θ given z is

$$\begin{aligned} g(\theta \mid z) &= \frac{f(z \mid \theta)\, g(\theta)}{\int f(z \mid \theta)\, g(\theta)\, d\theta} \\ &= \frac{f(z \mid \theta)\, g(\theta)}{f(z)}. \end{aligned} \tag{3.2}$$

In this context the term density refers to both the Lebesgue measure and the counting measure. Hence, the discrete case is included. If one models the uncertainty on a parameter vector θ through a probability distribution for discrete random variables, or a probability density π on Θ for continuous variables, $\pi(\theta)$ is the resulting distribution or density called prior distribution. 3.2 can then be rewritten

as

$$\pi(\theta \mid z) = \frac{f(z \mid \theta)\, \pi(\theta)}{f(z)}, \tag{3.3}$$

which incorporates the main idea of Bayesian modelling that parameters are assigned a probability distribution.

Definition 3.2 *A Bayesian statistical model is composed of a parametric statistical model, $f(z \mid \theta)$, and a prior distribution on the parameters, $\pi(\theta)$.*

$f(z \mid \theta)$ is often referred to as sampling distribution or observation model. As the denominator of 3.3 is independent of θ, an equivalent form of 3.3 is

$$\pi(\theta \mid z) \propto f(z \mid \theta)\, \pi(\theta). \tag{3.4}$$

The conditional density is obtained by normalizing the right hand side of the equation yielding the posterior density. The observed data affect the posterior distribution of the unknown parameters $\pi(\theta \mid z)$ only through the sampling distribution $f(z \mid \theta)$. Thus, for fixed z, the likelihood function $f(z \mid \theta)$ can be considered as a function of θ. In this notation θ is the vector which includes all unobservable quantities. If covariates are included in the model, the distribution of the unknown parameters is conditional on these as well.

We can summarize the process of Bayesian modelling through the three following steps. Initially, a full probability model needs to be set up, with a prior distribution for the unknown parameter vector $\pi(\theta)$, and the distribution $f(\theta, z)$ for the observable quantities, based on the observation model $f(z \mid \theta)$. Secondly, the posterior distribution of the unknown parameters is obtained by conditioning on observed data. The major aim in Bayesian modelling is the evaluation of the posterior distribution, which leads to the estimation of posterior quantities. The third step includes assessment of the model fit to the data. This issue will be discussed in more detail in section 3.3.2.

The specification of a suitable prior distribution is of great importance in Bayesian modelling. In the case that prior information is available, e.g. from earlier or related studies, the additional knowledge about certain parameters should be built into the model via a sensibly chosen prior distribution. However, it may be difficult in some cases to elicit a priori information or even a well-defined prior distribution. A lack if communication between the statistician and the decision maker, time delay, or costs are just a few reasons that may lead to little prior knowledge about the parameters. The following two subsections about conjugate and noninformative priors reveal possible approaches of choosing appropriate prior distributions.

3.1.1 Conjugate prior distributions

A common parametric approach to find a suitable prior distribution is to use a so-called conjugate prior distribution. According to Robert (1994), p. 97, this useful approach involves a subjective input as limited as possible.

Definition 3.3 *A family $\mathcal{F}$ of probability densities $f(z \mid \theta)$ on Θ is said to be conjugate if, for every $\pi \in \mathcal{F}$, the posterior density $\pi(\theta \mid x)$ also belongs to $\mathcal{F}$.*

A justification of the conjugate prior approach has been proposed by Raiffa and Schlaifer (1961). When the observation of Z is distributed according to $f(z \mid \theta)$, then the change of $\pi(\theta)$ to $\pi(\theta \mid z)$ is contributed through the data z only. This additional information about the prior distribution should then not lead to a complete structural change, but only to a change in the parameters of the distribution. Additionally, conjugate priors are attractive to use, as the posterior distributions are always computable (Robert 1994, p. 98). According to Gelman et al. (1995), p. 37, conjugate priors can also be interpreted as additional data. However, using conjugate priors only for simplicity and technical reasons can be dangerous, especially if proper prior information is available.

It has been shown by Brown (1986) and Robert (1994), p. 99ff., that conjugate prior distributions are associated with the class of sampling distributions that are exponential families. As there exists an automated way to deduce a conjugate distribution from the model $f(z \mid \theta)$, conjugate distributions for exponential families are usually referred to as natural conjugate distributions.

One conjugate family that shall be described in more detail is the family of Gaussian distributions. The family of Gaussian distributions is conjugate to itself. This implies that for a Gaussian observation model with

$$Z \mid \theta \sim \text{Gau}(\theta, \sigma^2)$$

the corresponding natural conjugate prior is again Gaussian, parameterized as $\theta \sim \text{Gau}(\mu, \tau^2)$. The resulting posterior distribution can be calculated analytically as

$$\theta \mid z \sim \text{Gau}(\rho(\sigma^2\mu + \tau^2 z), \rho\sigma^2\tau^2)$$

with $\rho^{-1} = \sigma^2 + \tau^2$. Let us consider the example of a Gaussian observation model with unknown mean θ, but known variance σ^2. Then the posterior mean of θ is a weighted average of the mean of the prior distribution μ and the sample mean θ. DeGroot (1986), p. 326, gives additional properties of the relative weight given to θ. Moreover, it can be seen that the posterior variance of θ depends on the number of observations, but does not depend on the observed magnitudes. Therefore, the posterior variance can be calculated in this case before any observations have been taken. A further discussion of this example with a generalization to the case of a multivariate Gaussian distribution will be given in section 3.1.3.

Advanced algorithms to sample from the posterior distribution make the use of conjugate families less important. However, in the context of this work, we will use conjugate prior distributions in the following situations. We assign a Gamma prior distribution to the inverse of the normal variance in the chapters 4 and 5. We also use a Wishart prior distribution for the multivariate normal covariance matrix in

chapter 6. Throughout this thesis we use the fact that the Gaussian distribution is a conjugate family to itself.

3.1.2 Noninformative prior distributions

As we have described above, conjugate priors can be a useful approximation to a prior distribution. However, when no information about the parameters is available, the use of conjugate priors is solely justified for analytical reasons. Alternatively, so called flat, diffuse, or noninformative priors can be used. Due to the lack of prior information they are derived from the sampling distribution alone. That means the model determines the family of the prior distribution to achieve conjugacy which can lead to closed form expressions of the posterior quantities. The idea is to construct a prior distribution, that contains as little a priori information about the parameter as possible. As it is often impossible to assign uniform priors over the entire support of θ, $\pi(\theta)$ is usually assigned a Gaussian distribution with large variance, or a uniform prior restricted to a certain area of the support. Depending on the parameter, of course, other distributions, such as the Gamma distribution for dispersion parameters on the positive support of θ may be used. The advantage of noninformative priors, at least for large samples, is that they are guaranteed to play a minimal role in the posterior distribution and allow the data to "speak for themselves", unaffectedly of additional information.

Noninformative priors have the disadvantage that transformations of the parameter of interest may lead to an informative prior distribution. This in turn leads to the invariance re-parameterization problem. For example the problem occurs, when assigning a uniform prior on $[0, 1]$ to the dispersion parameter σ^2 of the normal

distribution. Then

$$F_{\sigma^2}(u) = P(\sigma^2 < u) \; = \; \begin{cases} 0 & u < 0 \\ u & 0 \leq u \leq 1 \\ 1 & u > 1 \end{cases}$$

and for $0 \leq u \leq 1$ we obtain the density $\frac{dF_{\sigma^2}(u)}{du} = 1$. In the same interval, we obtain for σ itself

$$F_{\sigma}(u) = P(\sigma < u) = P(\sigma^2 < u^2) = u^2$$

and hence $\frac{dF_{\sigma}(u)}{du} = 2u$ which is clearly informative.

A second problem with noninformative priors is that they easily lead to improper priors, which implies that their densities do not integrate to 1. It can be desirable, e.g., to assign a uniform prior on $(-\infty,\infty)$ to some parameter. In that case, an improper prior can be used, but it should not be interpreted as probability distribution. There exist prior proposals like Jeffreys noninformative prior (Jeffreys 1961) that is based on the Fisher information matrix. Jeffreys prior usually results in improper priors, however it fulfills the invariance re-parameterization property for specific transformations. Especially when considering subvectors of the parameters of interest, problems with Jeffreys prior may occur. Bernardo (1979) has developed a modification of Jeffreys prior, namely the reference prior approach, that distinguishes between parameters of interest and nuisance parameters.

In the context of this work noninformative and improper priors play a role in three further situations.

1. In many situations the ideal form of a noninformative prior is a uniform distribution over an interval of infinite length. Suitable software has been developed which generates samples from such flat distributions. The WinBUGS software (Spiegelhalter, Thomas, and Best 2002) uses the specification dflat(), and in BayesX (Lang and Brezger 2002) such diffuse priors are the default choice.

2. Under extreme circumstances the posterior distribution for a well defined problem with proper prior might not exist, since the integral $\int f(z \mid \theta)\pi(\theta)d\theta = \infty$. In this case the posterior distribution cannot be computed.

3. In situations where a noninformative improper prior is desired, one frequently uses a proper prior that is essentially constant over an interval in which the parameter can be expected to lie. For example it is common use to substitute a flat prior by a $\text{Gau}(0, \sigma^2)$ distribution with large σ^2 ($> 10^4$), or a uniform prior over a large interval.

3.1.3 Example for multivariate Gaussian data

In this section we consider the case, that our observations are drawn from a multivariate normal distribution, i.e. the model is of the following form

$$Z \mid \mu, \Sigma \sim \text{Gau}_p(\mu, \Sigma),$$

where μ is a column vector of length p and Σ is a $p \times p$ covariance matrix, symmetric and positive definite. The likelihood for a sample of iid. observations $z_1, \ldots, z_n$ is

$$f(z_1, \ldots, z_n \mid \mu, \Sigma) \propto |\Sigma|^{-n/2} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(z_i - \mu)' \Sigma^{-1}(z_i - \mu)\}. \qquad (3.5)$$

Consider the case of an unknown mean vector μ but known covariance matrix Σ. Then the conjugate prior distribution for μ is again the multivariate normal distribution which is parameterized as $\mu \sim \text{Gau}_p(\mu_0, \Lambda_0)$. After rearranging terms, it is easily seen that the posterior distribution of μ is

$$\pi(\mu \mid z, \Sigma) \propto \exp\{-\frac{1}{2}(z - \mu)'\Sigma^{-1}(z - \mu)\} \exp\{-\frac{1}{2}(\mu - \mu_0)'\Lambda_0^{-1}(\mu - \mu_0)\}. \qquad (3.6)$$

Expanding and completing the quadratic form for μ leads to the posterior distribution

$$\mu \mid z, \Sigma \sim \text{Gau}_p(\mu_n, (\Lambda_0^{-1} + n\,\Sigma^{-1})^{-1})$$

with

$$\mu_n = (\Lambda_0^{-1} + n\,\Sigma^{-1})^{-1}\,(\Lambda_0^{-1}\mu_0 + n\,\Sigma^{-1}\bar{z}).$$

From this result we can draw some conclusions about the posterior mean and variance. In the multivariate Gaussian context the posterior mean is a weighted average of the prior mean and the sample mean. The weight of the sample mean increases as the number of observations increases. In the limit the posterior mean is equivalent to the sample mean. The posterior covariance is smaller than the one prior to data collection. With an increasing number of observations, the posterior covariance approaches the covariance of the mean of the sample. In this example the posterior distribution can be calculated explicitly due to the conjugacy of the Gaussian distribution which in general proves to be difficult.

To conclude it can be said that if no information on the parameter vector is available, the choice of noninformative priors is a good alternative that leads to a posterior distribution which solely depends on the observation model and the data.

3.2 Likelihood inference

In the beginning of this section we have described the inversion principle of statistical analysis. According to Robert (1994), p. 7, this aspect results in the notion of the likelihood function, which is just the sampling distribution rewritten as a function of the unknown parameter vector θ

$$l(\theta \mid z) = f(z \mid \theta),$$

depending on the given and fixed observed data. The inversion principle, and especially equation 3.1, can be seen as an actualization principle. In the context of the likelihood function, it can be transferred to consider the likelihood of event A, and update $P(A)$ to $P(A \mid E)$ given the observed event E.

According to Gelman et al. (1995), p. 107, we can often interpret classical point estimates as exact or approximate posterior summaries based on some implicit full probability model. Asymptotic theory can be used to construct a theoretical Bayesian justification for classical maximum likelihood inference under certain regularity conditions. DeGroot (1986), p. 339, mentions a possible difficulty of Bayes estimation, if the parameter θ is actually an unknown (high dimensional) parameter vector. Then multivariate prior distributions must be specified for all components of the parameter vector θ, even if there are only one or two components of interest. In such a problem it is especially difficult to specify meaningful priors on the multidimensional parameter space Θ. However, we have discussed the influence of prior specifications in Bayesian modelling, and therefore we aim to perform both likelihood and Bayesian inference for the case of the space-time model in chapter 5. We will show the modelling strategies in the frequentist and the Bayesian case, we will compare the resulting parameter estimates and posterior quantities, and we will discuss computational issues of frequentist and Bayesian inference for that example.

Maximum likelihood estimators

In order to obtain point estimates of the desired parameters in the frequentist case, we use maximum likelihood estimation. This subsection gives a short summary on the theory of maximum likelihood point estimation. Suppose that the random variables $Z_1, \ldots, Z_n$ form a random sample from a discrete or continuous distribution with probability distribution function or density $f(z \mid \theta)$, where the unknown parameter or parameter vector θ belongs to the parameter space Θ. For any observed vector $z = (z_1, \ldots, z_n)$ in the sample, the value of the joint probability distribution function as a function of θ for a given vector z is called the likelihood function $l(\theta, z)$. The idea of maximum likelihood estimation is now to obtain a parameter estimate $\hat{\theta}(z) \in \Theta$ for the observation z such that for this parameter estimate the proba-

bility density of the observation is maximized. Thus $\hat{\theta} = \hat{\theta}(z)$ is called maximum likelihood estimate for θ if

$$l(\hat{\theta}, z) \geq l(\theta, z) \qquad \text{for all } \theta \in \Theta.$$

For $n \to \infty$, and under regularity conditions the maximum likelihood estimator is a sufficient statistic, and according to Gelman et al. (1995), p. 107, so is the posterior mode, mean and median under the Bayesian point of view. Sufficiency implies that the estimator contains essentially all the information available about θ from the data. Therefore, asymptotic irrelevance of the prior distribution can be shown, which allows for the use of noninformative priors for a large number of observations.

An important feature of this thesis is that the analyzed data are not independent. Even in the case of dependent observations, maximum likelihood theory can be applied under very general conditions using Martingale limit theory. In this context, asymptotic behavior of maximum likelihood estimators can be derived, i.e. for the much more general case of departure from independence. These topics will be discussed in more detail in chapter 5.

3.3 Bayesian inference with Markov chain Monte Carlo

In recent years Markov chain Monte Carlo techniques have been rapidly developed, providing new methods firstly in the field of Physics, and then later also for statisticians. Monte Carlo (MC) importance sampling as introduced by Hammersley and Handscomb (1964) has been used to simulate directly from the distribution of interest, as the integrals that are required for Bayesian computation need an analytic or numerical approximation. In a Bayesian setting the aim is usually the evaluation of

the posterior distribution. However, Monte Carlo simulation is restricted to simple problems, with a small number of dimensions. The idea of the more sophisticated MCMC simulation is to construct a Markov chain that has the posterior distribution as its stationary distribution. This technique is applicable for more complex and high dimensional models, and its applications within various sampling algorithms will be explained in more detail in the following sections. The idea underlying the generation of an ergodic Markov chain with the posterior distribution as its stationary distribution is made possible through the following theorem (Kulkarni 1995, p. 105). Let $p_{ij}^{(t)}$ be the probability that the Markov chain is in state j at time t when it starts from state i at time 0.

Theorem 3.2 *[Limit theorem for ergodic Markov chains]*
Assume that $\{Z_t, t \geq 0\}$ *is an irreducible, aperiodic, and positive recurrent Markov chain. Then we have*
(a) $\lim_{t\to\infty} p_{ij}^{(t)} = \lim_{t\to\infty} p_{jj}^{(t)} =: \pi_j$,
(b) $\{\pi_j, j \in S\}$ *are given by the unique solution to*

$$\pi_j = \sum_{i\in S} \pi_i \, p_{ij}, \qquad \sum_{j\in S} \pi_j = 1.$$

The proof can be found in standard literature for stochastic processes, such as Kulkarni (1995), p. 105f., or Fahrmeir, Kaufmann, and Ost (1981), p. 59. Irreducible, aperiodic Markov chains are also called ergodic. Transferring the content of theorem 3.2 to our application in MCMC simulation, it follows that for any irreducible and aperiodic chain there exists a unique, positive solution, i.e. the limiting distribution $\{\pi_j\}$ of Z_t as $t \to \infty$. The limiting distribution is also called the steady-state distribution. It can be shown that

$$P\{Z_t = j\} = \pi_j \qquad \text{for all } n \geq 1.$$

The distribution of Z_t is independent of t if the Markov chain starts with the initial distribution $\{\pi_j, j \in S\}$. For this reason $\{\pi_j, j \in S\}$ is called a stationary distribu-

tion. The consequences of this theorem for Markov chain Monte Carlo methods will be discussed after the introduction of specific sampling techniques.

In the context of MCMC methods it is clear that the state space of the Markov chain is finite since there is only a finite collection of real numbers available on the computer. Hence, even in the case of continuous distributions the theory for discrete Markov chains can be applied, provided that the chain is irreducible and aperiodic. The latter assumption is usually given for granted.

3.3.1 Sampling techniques

Suppose it is the aim to generate a sample from a distribution π on $\mathcal{X} \subseteq \mathbb{R}^n$, but it is impossible to do it directly. However, it is possible to construct a Markov chain with state space $\mathcal{X}$, which is straightforward to simulate from and whose equilibrum distribution is π. π is typically the posterior distribution, and observations drawn from the Markov chain are dependent. However, if the chain is run long enough simulated values of the chain can be used as a basis for summarizing features of π, such as mean, median, or measures of dispersion. The crucial part is to find algorithms for constructing chains with specific equilibrum distributions. Consider $x^{(1)}, x^{(2)}, \ldots$ as a realization from an appropriate chain, then according to Smith and Roberts (1993) typically available results in the sense of theorem 3.2(a) include

$$\lim_{t\to\infty} X^{(t)} \stackrel{d}{=} X \sim \pi \tag{3.7}$$

and

$$\frac{1}{M}\sum_{m=1}^{M} f(x^{(m)}) \to E_\pi\{f\} \qquad \text{almost surely.} \tag{3.8}$$

Equation 3.8 implies that ergodic averaging of a function of interest over realizations from a single run of the chain provides a consistent estimator of its expectation.

Typical sampling algorithms that shall be described in more detail are the Gibbs sampler, and the Metropolis-Hastings sampler.

As mentioned, successive $X^{(t)}$ will be correlated, thus suitable spacings or long runs of the chain are required to form the sample. The autocorrelation can be reduced by subsampling the chain by keeping only every kth iteration. This method is called 'thinning'. Another possibility is to run parallel independent chains. The question which of the methods is preferable will be discussed at the end of this chapter. Keeping in mind that the chains need a certain time to converge towards the stationary distribution, a so called 'burn-in' phase is necessary for every chain, and must be discarded for the analysis. Moreover, sophisticated models require sampling algorithms that usually need even longer phases to adapt. These cases will be discussed in the context of Metropolis sampling.

Gibbs sampler

One of the best known sampling algorithms is the Gibbs sampler. It has been introduced by Geman and Geman (1984) and transferred to the use for Bayesian inference by Gelfand and Smith (1990). Applications can be found, e.g., in Casella and George (1992), Brooks (1998), and Brooks and Gelman (1998). To apply the Gibbs sampling algorithm, firstly, the Bayesian model needs to be specified as a joint distribution of all parameters and observable quantities. Given the observed data, it is the aim to sample values of the unknown parameters from their conditional posterior distributions. The sampling is done by conditioning the distribution of each single node successively on all remaining nodes of the model. More formally, let $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ be a p-dimensional vector of parameters, and let $\pi(\theta \mid z)$ be its posterior distribution given the data. Then the basic scheme of the Gibbs

sampler, according to Chen, Shao, and Ibrahim (2000), p. 20ff., is:

Step 0 Choose an arbitrary starting point $\theta_0 = (\theta_{1,0}, \theta_{2,0}, \ldots, \theta_{p,0})'$ and set the iteration index $j = 0$

Step 1 Generate $\theta_{j+1} = (\theta_{1,j+1}, \theta_{2,j+1}, \ldots, \theta_{p,j+1})'$ as follows:

Generate $\theta_{1,j+1} \sim \pi(\theta_1 \mid \theta_{2,j}, \ldots, \theta_{p,j}, z)$

Generate $\theta_{2,j+1} \sim \pi(\theta_2 \mid \theta_{1,j+1}, \theta_{3,j} \ldots, \theta_{p,j}, z)$

$\vdots$

Generate $\theta_{p,j+1} \sim \pi(\theta_p \mid \theta_{1,j+1}, \theta_{2,j+1} \ldots, \theta_{p-1,j+1}, z)$

Step 2 Set $j = j + 1$ and continue with step 1.

The conditional distribution of the individual components of θ given all the other components including the data used in step 1 of the algorithm are called full conditional distributions.

The basic justification of the Gibbs algorithm is as follows. Consider a random element $\theta \in \mathbb{R}^p$ with density π with respect to a measure ν. If a new element θ^* is constructed by keeping the elements $\theta_2, \ldots, \theta_p$ fixed and drawing a new θ_1^* from the conditional distribution $\pi(\theta_1 \mid \theta_2, \ldots, \theta_p)$ then the new vector $\theta^* = (\theta_1^*, \theta_2, \ldots, \theta_p)$ has the same distribution as θ, since

$$\begin{aligned} P(\theta^* \in B) &= \int_B \pi(\theta_2, \ldots, \theta_p)\, \pi(\theta_1^* \mid \theta_2, \ldots, \theta_p) d\nu(\theta) \\ &= \int_B \pi(\theta) d\nu(\theta) = P(\theta \in B). \end{aligned} \tag{3.9}$$

This process is repeated for the components of $\theta_2, \ldots, \theta_p$. We denote the resulting vector by $\tilde{\theta}$. The transition of θ to $\tilde{\theta}$ can be repeated independently and the sequence $\theta, \tilde{\theta}, \tilde{\tilde{\theta}}, \ldots$ obviously constitutes a Markov chain. From equation 3.9 it follows that this Markov chain is stationary and if we take π to be the posterior density $\pi(\theta \mid x) = \frac{\pi(\theta) f(x|\theta)}{f(x)}$ with fixed x, then the stationary distribution coincides with the posterior distribution.

Thus, in practice we do not construct the posterior distribution, but rather start at an arbitrary starting value for the vector θ and rely on the fact that under very general conditions the Markov chain converges to its stationary distribution. Details have been worked out by Geman and Geman (1984), Gelfand and Smith (1990), and Besag, Green, Higdon, and Mengersen (1995). There exists a sufficient condition that guarantees geometric convergence, as proved by Schervish and Carlin (1992).

In the context of the spatial and spatio-temporal models considered in later chapters, a natural ordering of the site specific random effects is impossible. Therefore, the approach using full conditionals is not directly applicable, and this requires a modification of the Gibbs sampler, the so called sliced sampler (Neal 1997 and Neal 2002) must be applied. The simulation software WinBUGS (Spiegelhalter et al. 2002) uses this modification for inference problems that cannot be expressed in the context of directed acyclic graphs (dag). The idea is to simulate from the joint distribution, to avoid an ordering, in the sense of chain graphs.

For the more complex allocation models introduced for the small area estimation in chapter 6, we need an algorithm which is not restricted to sampling from the full conditional distributions. We employ the Metropolis-within-Gibbs sampler, a special case of the Metropolis-Hastings algorithm.

Metropolis-Hastings sampler

The very general Metropolis-Hastings (MH) algorithm has been firstly introduced by Metropolis et al. (1953), and it has then been generalized by Hastings in 1970. As mentioned above the Gibbs sampler generates random draws from the full conditional distributions, which are often unknown or only known up to a normalizing constant. Application of an algorithm of the much more general class of Hastings algorithms has the advantage that it suffices to know the ratio $\frac{\pi(\theta^*|z)}{\pi(\theta|z)}$. The idea of Metropolis-Hastings or Hastings steps is that a proposed new value for a given

component is accepted only with a certain probability less or equal to one. The steps of the algorithm can be described as follows, according to Chen et al. (2000), p. 23. Chib and Greenberg (1995) introduce the term candidate generating density for the so-called proposal density $q(\theta, \theta^*)$.

Step 0 Choose an arbitrary starting point θ_0 and set $j = 0$

Step 1 Generate a candidate point θ^* from $q(\theta_j, \cdot)$ and u from the uniform distribution $U(0, 1)$

Step 2 Set $\theta_{j+1} = \theta^*$ if $u \leq a(\theta_j, \theta^*)$ and $\theta_{j+1} = \theta_j$ otherwise, where the acceptance probability is given by $a(\theta, \theta^*) = \min\{\frac{\pi(\theta^*|z)q(\theta^*,\theta)}{\pi(\theta|z)q(\theta,\theta^*)}, 1\}$

Step 3 Set $j = j + 1$, and go to step 1.

Many well-known algorithms such as the Gibbs sampler and the Metropolis sampler are special cases of the MH sampler, dependent on suitably chosen proposal densities. The Gibbs sampler has an acceptance probability of 1. The Metropolis sampler has a symmetric proposal density with $q(\theta, \theta^*) = q(\theta^*, \theta)$, which leads to a simplification of the acceptance probability. The MH sampler depends strongly on the choice of the proposal, especially on the spread of the proposal density. According to Chen et al. (2000), p. 24, it affects the behavior of the chain in the following two directions: the acceptance rate and the region of covered sample space. For the covered sample space, it can be said that if the spread is extremely large, some of the generated candidate values will have low probability of being accepted. On the other hand, if the spread is chosen too small the chain will take longer to traverse the support of the density. Both of these situations are likely to be reflected in high autocorrelations across sample values. The acceptance rate is the percentage of times a move to a new value is made. Spiegelhalter et al. (2002) aim for acceptance rates between 20 and 40 percent, and Lang and Brezger (2002) propose

acceptance rates between 30 and 80 percent. Chen, Shao, and Ibrahim (2000), p. 24, let the acceptance rate depend on the number of dimensions. For one-dimensional problems the acceptance rate is tuned to 45 percent, for increasing dimensionality it is reduced to, e.g., 25 percent in a six-dimensional example. In the WinBUGS software (Spiegelhalter et al. 2002) the following "Metropolis-within-Gibbs sampler" has been implemented.

Metropolis-within-Gibbs sampler

A special case of the MH algorithm is the Metropolis-within-Gibbs sampler. The algorithm works as follows: an intractable full conditional density is sampled with the general form of the MH algorithm, as described above. The other values are sampled directly from their full conditionals using the Gibbs algorithm. In the WinBUGS context, the standard deviation of the proposal distribution is tuned to obtain acceptance rates that lie between 20 and 40 percent. Common proposal distributions include the Gaussian distribution, the t-distribution, and the triangular distribution. The tuning is performed within the first 4,000 iterations, which are discarded from further summary statistics. These iterations will simultaneously be considered as the burn-in phase for the models presented here.

In general there is great flexibility in constructing Markov chains that converge to a given distribution, and in theory all such chains will converge eventually. In practice however, fast convergence is important. Approaches like block sampling improve the mixing of the Markov chains by aiming for a more rapid moving of the chain through the state space $\mathcal{X}$. Especially in models where the components of the parameter vector θ are already correlated in the prior, single and successive updating leads to slow mixing of the chains. This usually results in strong autocorrelation and slow convergence. The software BayesX (Lang and Brezger 2002) uses the possibility of a simultaneous updating of blocks of parameters.

3.3.2 Convergence diagnostics

Despite the existence of literature concerning the convergence of MCMC methods (e.g. Tierney 1994, Brooks and Gelman 1998, Kass, Carlin, Gelman, and Neal 1998), results do not easily translate into clear guidelines for the practitioner (Smith and Roberts 1993). Theoretical assessment does not provide useful bounds on rates of convergence and any kind of pragmatic output analysis to estimate the length of the transient phase. A sensible output analysis, however, is crucial. Without it one is endangered to overlook important aspects of the multidimensional behavior of the chain. Ideally, the samples from the posterior distribution will be independent, and the Monte Carlo (MC) standard error of the mean will be minimal. The MC error is calculated as $\sigma/\sqrt{N}$, where σ is estimated according to a method introduced by Roberts (1996), and N is the number of iterations of the chain. Simple diagnostic methods for convergence and mixing behavior of the chain are trace plots of parameter samples, autocorrelation functions based on parameter samples, and sometimes bivariate scatter plots of pairs of parameters.

The first approach to output analysis is to monitor visually the ergodic averages of selected scalar quantities for stationarity. The length of the burn-in phase, i.e. the number of observations until the chain has reached its stationary distribution needs to be determined. Geyer (1992) recommends discarding the first one or two percent of a run long enough to give sufficient precision.

The number of samples to be drawn from the Markov chains depends on the strength of the correlation between the samples. The accuracy of the estimated quantities of interest is more dependent on the correlation between the realizations than on the number of realizations. If the autocorrelation of successive states of the generated chain is high, a large number of iterations may be required for reliable estimates. Thinning, i.e. keeping only every kth iteration for the estimation, can be used to reduce the dependence between the samples, and the necessary storage size

for the chains decreases. An empirical autocorrelation can be calculated to determine the amount of thinning needed. According to Knorr-Held (1997), p. 25, one or two thousand nearly independent samples are usually sufficient for the estimation of standard posterior characteristics, such as the mean, median, or other quantiles of the marginals. Variances need to be considered with more care, and the mixing of the chains should be monitored with caution. Gelman and Rubin (1992) suggest monitoring multiple runs from dispersed starting points. An output analysis based on several runs is more likely to detect multimodality of the posteriors, as described by Besag and Green (1993). For complex models, however, a long adaptive phase and a suitable burn-in for each chain may be time consuming and therefore ineffective. Specific diagnostic tools for convergence are implemented in the WinBUGS software, which are based on the CODA language and which use a convergence statistic according to Brooks and Gelman (1998). In general, Smith and Roberts (1993) propose to apply an exploratory data analysis to the chains, similar to one for multivariate observations.

3.3.3 Estimating posterior quantities

The fundamental goal of Bayesian computation is to assess posterior quantities of interest. After convergence of the Markov chain, we obtain typically dependent samples from the posterior distribution. Posterior mean and standard deviation, median, and other quantiles can easily be estimated through the empirical evaluation of the posterior distribution. Furthermore, credible intervals, or supported range, and other characteristics like functionals of the parameters result directly from the posterior marginal densities of the parameters. More formally, posterior quantities are usually of the following form (Chen et al. 2000, pp. 67-68)

$$E(h(\theta) \mid z) = \int_{\mathbb{R}^p} h(\theta)\, \pi(\theta \mid z)\, d\theta, \tag{3.10}$$

where $h(\cdot)$ is a real valued function of $\theta = (\theta_1, \ldots, \theta_p)'$.

3.3.4 Sensitivity analysis and model selection

Two topics left to mention in this section are sensitivity analysis and Bayesian ways of model selection. Sensitivity analysis is the assessment of the influence of the prior distribution on the posterior outcome. Especially in the context of Bayesian hierarchical modelling the hyperprior settings of the precision parameters can have a large influence on posterior results. The prior distributions used in this thesis are often non-informative and sometimes even improper. In that case, according to Robert (1994), p. 120, the arbitrary part of the prior distribution should not get predominant. If prior knowledge is not available, different choices of prior settings should be compared in order to detect discrepancies, and dependencies in the results on the prior. Further discussion about sensitivity analysis can be found in Robert (1994), p. 120, and Gelman, Carlin, Stern, and Rubin (1995), p. 161. Bernardinelli, Clayton, and Montomoli (1995) discuss the importance of prior specifications in the context of disease mapping. Mollié (2000) compares different choices of the prior mean for inverse variance parameters in a hierarchical Bayesian framework. In chapter 5 we evaluate the influence of prior specifications on the posterior densities of the temporal and spatial autocorrelation parameters, on covariate effects, and on the dispersion parameter. The prior settings are modified with respect to the type of distribution, and within one distribution type we have additionally changed the parameters.

Model selection in Bayesian (hierarchical) modelling is yet challenging and a topic of current research. Han and Carlin (2001) discuss the use of Bayes factors for complex models to calculate posterior probabilities for a collection of competing models. However, they suggest that less formal Bayesian model choice methods may offer a more realistic alternative in many cases. For this thesis we use a Bayesian

measure of comparing models according to Spiegelhalter, Best, Carlin, and van der Linde (2002). It is based on the effective number of parameters in the model (denoted by p_D) and the posterior mean deviance ($\bar{D}$). Then the deviance information criterion (DIC) is defined as

$$\text{DIC} := p_D + \bar{D}.$$

The complexity of the model, p_D, is measured through the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest. The effective number of parameters will be close to the nominal number of parameters in simple models. In more complex, or hierarchical models, p_D is typically much smaller than the nominal number of model parameters, due to prior dependence between parameters. The posterior mean deviance is suggested by Spiegelhalter et al. (2002) as a measure of fit. The DIC can be seen as a Bayesian version of the Akaike's information criterion, and model specifications with a smaller DIC are preferable. For this thesis the DIC has been used to compare purely spatial, and spatio-temporal models with and without covariates.

Chapter 4

Temporal and Spatial Markov Dependence Structures

his chapter deals with modelling approaches for temporally and spatially dependent data using stochastic process theory. Stochastic processes are families of random variables Z_t, depending on a parameter t. We begin with the statistical theory for vector-autoregressive processes. The area specific data over the observation period form the multivariate time series. The time series definition of a Markov process will be transferred to spatially dependent data in section 4.2. Here we regard yearly slices of observations in the spatial units of the study region. Markov random fields are introduced to specify spatial departures from independence for each year. The separate temporal and spatial analyses are used as preparation for the combined analysis in chapter 5.

A K-dimensional vector stochastic process or multivariate stochastic process is a function $Z : \mathcal{T} \times \Omega \rightarrow I\!R^K$, where for each fixed $t \in \mathcal{T}$, $Z_t(\omega)$, $\omega \in \Omega$, is a K-dimensional vector, a measurable function on an underlying probability space $(\Omega, \mathfrak{A}, P)$. The underlying stochastic process is said to have generated the multivariate time series Z_t.

4.1 Vector-autoregressive processes

We begin with modelling the multivariate time series of yearly observations of cancer mortality over the study region. The spatial units within the region form the vectors of temporally dependent data. A special case of multivariate stochastic processes are autoregressive processes, on which we will concentrate for our applications. Finite order autoregressive processes are based on the idea of a structural analysis with linear functions dependent on past observations. We investigate first order vector-autoregressive (VAR) processes that satisfy relationships according to the following definition.

Definition 4.1 *Let C be a fixed $K \times K$ coefficient matrix, $\nu = (\nu_1, \ldots, \nu_K)'$ is a fixed $(K \times 1)$ vector of intercept terms. Furthermore, let $\epsilon_t = (\epsilon_{1t}, \ldots, \epsilon_{Kt})'$ be a K-dimensional white noise vector, or innovation process, with the properties that $E(\epsilon_t) = 0$, $E(\epsilon_t \epsilon_t') = \Sigma_\epsilon$ and $E(\epsilon_t \epsilon_s') = 0$ for $s \neq t$. Σ_ϵ is assumed to be non-singular. Then a vector process $Z_t = (Z_{1t}, \ldots, Z_{Kt})'$ satisfying*

$$Z_t = \nu + C\, Z_{t-1} + \epsilon_t \qquad t = 1, \ldots, T \tag{4.1}$$

is called a vector-autoregressive process.

Due to the structure of the error term, processes as specified in definition 4.1 are usually seen to be determined by their innovation process ϵ_t. An important case to consider is the assumption that ϵ_t is Gaussian white noise, i.e. $\epsilon_t \sim \text{Gau}_K(0, \Sigma_\epsilon)$ for all t, and ϵ_t and ϵ_s are independent for $s \neq t$. In that case Z_t is a Gaussian process which implies that all subcollections $Z_{t_1}, \ldots, Z_{t_n}$ have multivariate normal distributions for all $t_1, \ldots, t_n$. The data under examination in this thesis are cancer mortality data. We assume that the multivariate time series has started in the infinite past. This assumption can easily be justified, as this process will have reached equilibrum, i.e. it will have converged to its stationary distribution.

In the following we analyze the relationship between the innovation process ϵ_t and the Z process. If we assume that the generating mechanism starts at time $t = 1$ we can rewrite 4.1 as follows

$$\begin{aligned}
Z_1 &= \nu + C\, Z_0 + \epsilon_1 \\
Z_2 &= \nu + C\, Z_1 + \epsilon_2 = \nu + C(\nu + C\, Z_0 + \epsilon_1) + \epsilon_2 \\
&= (U_K + C)\nu + C^2\, Z_0 + C\, \epsilon_1 + \epsilon_2 \\
&\vdots \\
Z_t &= (U_K + C + \ldots + C^{t-1})\, \nu + C^t\, Z_0 + \textstyle\sum_{i=0}^{t-1} C^i\, \epsilon_{t-i},
\end{aligned}$$

with K-dimensional identity matrix U_K. This shows that $Z_0, \epsilon_1, \ldots, \epsilon_t$ uniquely determine the vectors $Z_1, \ldots, Z_t$, and the joint distribution of $Z_0, \epsilon_1, \ldots, \epsilon_t$ determines the joint distribution of $Z_1, \ldots, Z_t$. The above equation can also be rewritten as

$$\begin{aligned}
Z_t &= \nu + C\, Z_{t-1} + \epsilon_t \\
&= (U_K + C + \ldots + C^j)\, \nu + C^{j+1}\, Z_{t-j-1} + \textstyle\sum_{i=0}^{j} C^i\, \epsilon_{t-i}.
\end{aligned}$$

Under certain conditions concerning the matrix C, the sequence $\{C^i, i = 0, 1, \ldots\}$ converges to the zero matrix of dimension $K \times K$ and $\sum_{i=0}^{j} C^i$ is absolutely summable, which results in the existence of the infinite sum

$$\sum_{i=1}^{\infty} C^i\, \epsilon_{t-i} \tag{4.2}$$

in mean square.

A sufficient condition for the convergence of 4.2 is that a suitable matrix norm of C is less than 1 (Fahrmeir and Hamerle 1984). Lütkepohl (1991), p. 10, proposes to use the following spectral matrix norm based on an eigenvalue condition on C

$$\|C\| := \max\{\sqrt{\lambda} : \lambda \text{ is eigenvalue of } C'C\}. \tag{4.3}$$

According to Horn and Johnson (1985), p. 295f., definition 4.3 satisfies the usual properties of a norm, including the triangle inequality and submultiplicativity.

We assume that C satisfies the norm condition

$$\|C\| < 1.$$

Since then C^j converges to a matrix of zero as $j \to \infty$, we can ignore the term $C^{j+1} Z_{t-j-1}$ in the limit because $\{Z_{t-j-1}\}$ is stationary. Hence, Z_t is the well-defined stochastic process

$$Z_t = \mu + \sum_{i=0}^{\infty} C^i \, \epsilon_{t-i}, \qquad t = 0, \pm 1, \pm 2, \ldots, \tag{4.4}$$

where $\mu = (U_K - C)^{-1}\nu$ is the limit of the geometric series of constant terms. Of special interest in this context are the first and second moments of the process Z_t. These are

$$E(Z_t) = \mu \qquad \text{for all } t$$

and

$$\begin{aligned} E(Z_t - \mu)(Z_{t-h} - \mu)' &= \lim_{n\to\infty} \sum_{i=0}^{n} \sum_{j=0}^{n} C^i \, E(\epsilon_{t-i} \, \epsilon'_{t-h-j} \, (C^j)' \\ &= \lim_{n\to\infty} \sum_{i=0}^{n} C^{h+i} \Sigma_\epsilon (C^i)' \\ &= \sum_{i=0}^{\infty} C^{h+i} \Sigma_\epsilon (C^i)' \\ &= C^h \sum_{i=0}^{\infty} C^i \Sigma_\epsilon (C^i)' =: \Delta_Z(h), \end{aligned} \tag{4.5}$$

as $E(\epsilon_t \epsilon'_s) = 0$ for $s \neq t$ and $E(\epsilon_t \epsilon'_t) = \Sigma_\epsilon$ for all t. Thus, neither the first nor the second moment of the process are dependent on t, which implies second order stationarity or weak stationarity. In case of a Gaussian vector-autoregressive process, the joint density of $\{Z_t\}_{t\in\mathcal{T}}$ is Gaussian as well, and hence weak stationarity implies strong stationarity, i.e. all joint distributions of the process are time invariant. For $h = 0$ we obtain the covariance matrix of the stationary process as

$$\text{cov}(Z) = \sum_{i=0}^{\infty} C^i \, \Sigma_\epsilon \, (C^i)'. \tag{4.6}$$

The innovation process can also be specified through the Z process. To see this let us consider a second order stationary Gaussian vector-autoregressive process with zero mean

$$Z_t \sim \text{Gau}_K(0, \Sigma) \qquad t = 1, \ldots, T.$$

Then Z_t can be represented on the basis of its innovation process ϵ_t as follows

$$\begin{aligned} & Z_t &=& C\, Z_{t-1} + \epsilon_t \\ \Leftrightarrow \qquad & \epsilon_t &=& Z_t - C\, Z_{t-1} \qquad t = 0, \pm 1, \pm 2, \ldots . \end{aligned} \tag{4.7}$$

If C is chosen as

$$C = \text{cov}(Z_t, Z_{t-1})\, \Sigma^{-1},$$

then ϵ_t is independent of Z_{t-1}. Since ϵ_t and Z_t are Gaussian, it suffices to show that

$$E(\epsilon_t\, Z'_{t-1}) = 0.$$

This follows from

$$\begin{aligned} E(\epsilon_t\, Z'_{t-1}) &= E[(Z_t - C\, Z_{t-1})\, Z'_{t-1}] \\ &= E(Z_t\, Z'_{t-1}) - C\, E(Z_{t-1}\, Z'_{t-1}) \\ &= \text{cov}(Z_t, Z_{t-1}) - C\, \Sigma \\ &= \text{cov}(Z_t, Z_{t-1}) - \text{cov}(Z_t, Z_{t-1})\, \Sigma^{-1}\, \Sigma \\ &= 0. \end{aligned}$$

From now on we consider "mean polished" processes Z_t, centered around zero. This implies a simple transformation of the data for the analysis.

The results obtained for first order vector-autoregressive processes can be easily extended to higher order vector-autoregressive processes of order p, as illustrated by Lütkepohl (1991), p. 11. For this thesis we will restrict to first order processes and rather deduce ways of incorporating the multivariate spatial dependence structure

within the vector time series. Thus, more emphasis is put on a sensible choice of the C matrix, which will be extended to joint temporal and spatial modelling in chapter 5.

Application to the data

We begin with a time series analysis of the districts in North Rhine Westphalia, with respect to lung cancer among women (dataset NRW). In this context, we consider the multivariate time series of $I = 54$ spatial units, which we assume for the moment to be independent. The length of the series is 19 years, from 1980 to 1998. To justify the assumption of a Gaussian distribution of the process, we have chosen to display histograms of the log-transformed the age standardized rates of two years within the study period in figure 4.1.

These figures show that we can consider the multivariate time series of lung cancer mortality rates to be Gaussian. If we shift the data by mean polishing, the assumption of a $\mathrm{Gau}_I(0, \Sigma)$ Z-process can be justified. As a first order vector-autoregressive process is to be fitted, we plot the partial autocorrelations to visualize the time series correlation. The partial autocorrelation in this context measures the correlation of the time series with lagged series of itself. The average partial correlations with lags from 1 to 10 over the 54 time series of log-transformed age standardized mortality rates have been calculated and plotted in figure 4.2. The dotted line indicates the upper 95 % confidence limit of the partial correlations. The figure shows a first order autoregressive structure of the multivariate process through an exponential decrease of the partial autocorrelations.

Besides lung cancer mortality among women in North Rhine Westphalia we have analyzed the temporal structure of stomach cancer in West Germany (dataset Germany). Figure 4.3 shows the corresponding partial autocorrelations for men. The correlation between subsequent observations over the years is apparently even

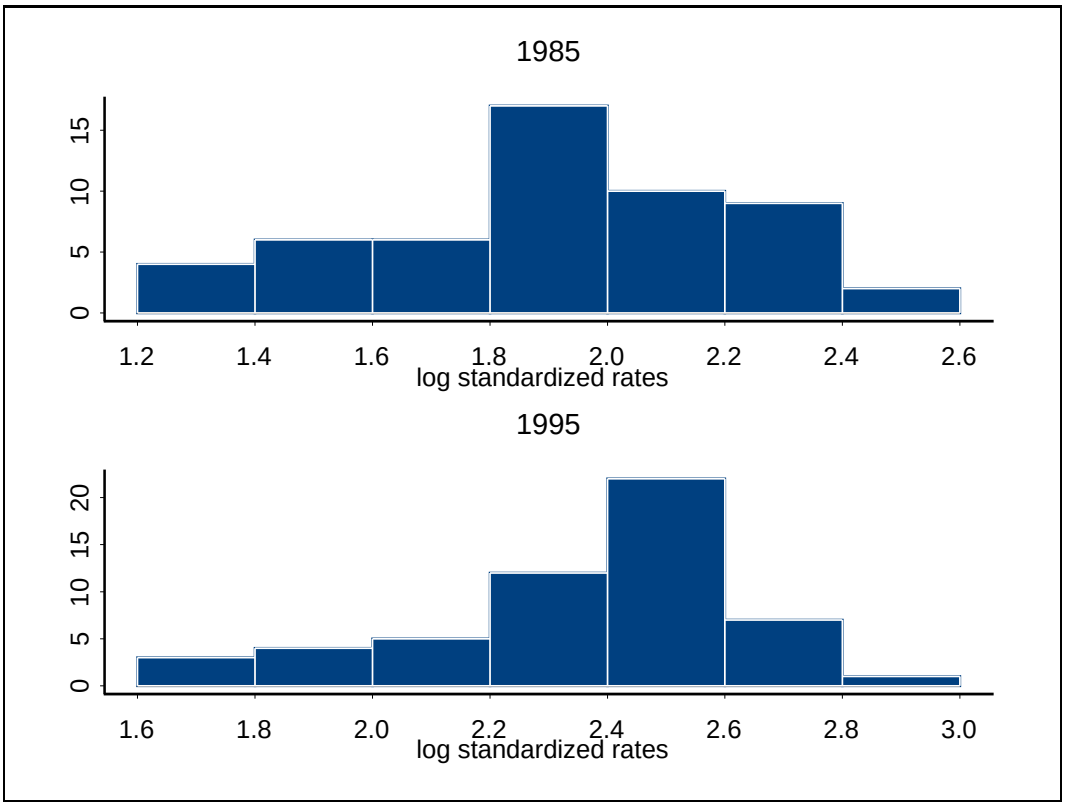

Figure 4.1: Lung cancer mortality among women in NRW. Histograms of the log-transformed rates for 1985 and 1995.

stronger than for lung cancer. Again the assumption of a first order autoregressive structure can be justified.

4.2 Markov random fields

In the foregoing section, we have introduced vector-autoregressive processes to model the temporal dependencies between annual cancer mortality data. We have discussed the multivariate time series, but we have not considered any specific correlation structure between the single vectors of the time series. Now we take into account that we have observations over the spatial units of the study region for each year.

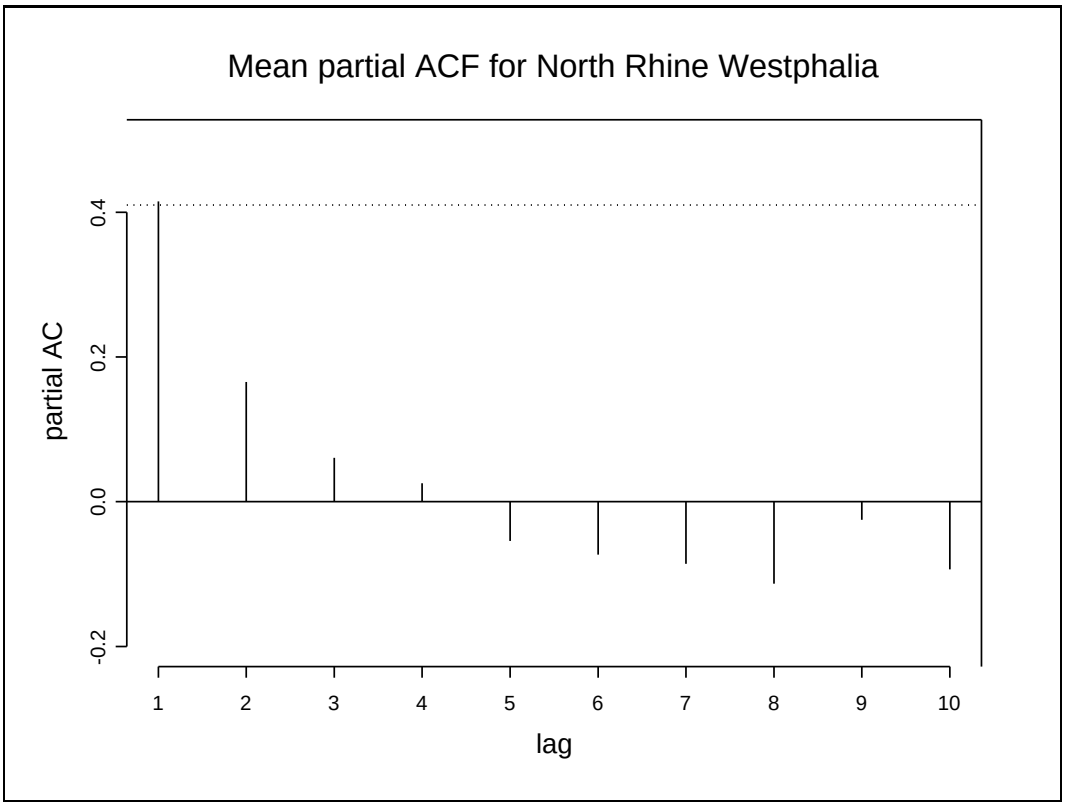

Figure 4.2: Lung cancer among women in NRW. Mean partial autocorrelation.

The data on cancer mortality collected in the $i = 1, \ldots, I$ sites should be modelled as (spatially) dependent data, since there is much spatial variation and large regional differences as described in chapter 2. For most cancer types, the pattern of variation does not seem to be spatially random. More specifically the lung cancer and stomach cancer maps of the age-standardized rates show strong clustering mainly caused by similar underlying distributions of risk factors or behavioral patterns. From page 66 we apply the presented methods to the data, and Moran's I statistic is used to justify a spatial modelling approach. Unlike for infectious diseases we do not need to incorporate the dynamic and fast moving spread over space (and time), where transmission of the disease between individuals is likely (Cargnoni, Müller, and West 1997). However, for our non-infectious cancer data our models should rather aim

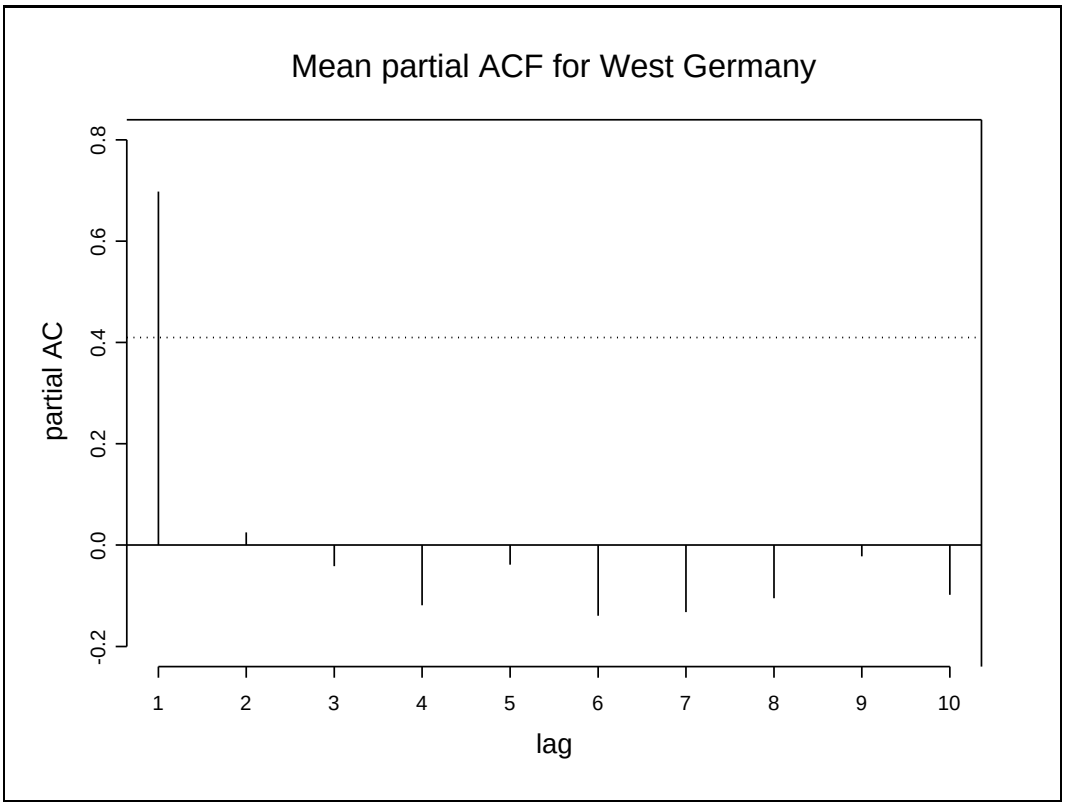

Figure 4.3: Stomach cancer among men in West Germany. Mean partial autocorrelation.

at a basis for prediction incorporating the spatial dependence. As mentioned, the detection of causality is difficult, not only because of long time lags between exposure to risk factors and dying of the specific cancer type.

When modelling spatially dependent data, the common idea of stochastic independence between observations is usually not applicable. Hence, it can be assumed that spatial units lying closer to each other are more alike than those which are further apart. Besag (1974), Cressie and Chan (1989), and Cressie (1993), p. 410, explain the idea of a spatial probability distribution at each of the given sites for fixed time t.

There are two different modelling approaches towards the specification of such

a probability measure. One is based on geometric model assumptions on a regular or irregular grid, considering two sites as neighbors if they share a common border, or if their centers lie within a certain distance of each other. As every model, this is only an attempt to depict the real underlying (correlation) structure of the data. However, depending on the number of sites and on the shape of the study region, this approach can be useful.

The alternative, as introduced by Besag (1974), is to begin with the examination of the probability distribution of Z_i, given all other sites on the lattice, ignoring their location. This approach can be seen as a probabilistic approach and the full conditional distributions for every site can be calculated. Often, it can be assumed that only a few sites have an influence on a particular site i. Hence, the (full) conditional distribution for site i depends only on those. However, symmetry of the correlation matrix is often needed for computational reasons, which can cause problems when using the probabilistic approach. Although these two models arise from two completely different views, they often lead to very similar dependence structures. The approach by Besag (1974) can also be seen as an example of the wider class of mixed, or random effects models (Best, Spiegelhalter, Thomas, and Brayne 1996). Generalized linear mixed models (GLMM) or generalized linear models with random effects are a very general class of models and they can be used in a wide range of statistical inference problems. Due to the introduction of random effects, besides the fixed effects, the approach allows for correlation or heterogeneous variability among the data. Thus, mixed models extend the generalized linear models by allowing for a more flexible specification of the covariance matrix of the error terms. The model can be formulated as follows

$$Z = X\,\gamma + D\,b + \epsilon. \tag{4.8}$$

Here, $X\gamma$ consists of the fixed effects, whereas Db incorporates the random effects. In our case, these are spatially structured. D is the design matrix and b is the parameter

vector of the random effects. When implying a Gaussian modelling approach it is easy to show that

$$Z \sim \text{Gau}_I(X\,\gamma, D\,\text{var}(b)\,D' + \text{var}(\epsilon)). \tag{4.9}$$

From the distribution of Z in 4.9 it is clear that the fixed effects enter the model through the mean only, whereas the random effects enter only through the covariance matrix. If $D = 0$ and $\text{var}(\epsilon) = \sigma^2\,I$, the generalized linear mixed model reduces to a generalized linear model.

The estimation of the parameters for fixed and random effects is more difficult in the mixed model environment, than in generalized linear models. Least squares estimation is no longer the best method. With frequentist approaches, strong restrictions are needed to reduce the number of parameters. These are then obtained with estimated generalized least squares. Maximum likelihood or restricted maximum likelihood methods are preferred to an expectation maximization (EM) approach. A short excursion of a frequentist analysis of spatial linear models is given in section 4.2.3. We rather use MCMC techniques for the estimation of the random effects, as will be explained in the sections 4.2.1 and 4.2.2.

To model the spatial dependence structure we need the following definition of a set of neighbors of site i.

Definition 4.2 *[Besag 1974] Site j ($\neq i$) is said to be a neighbor of site i, denoted by $i \sim j$, $i,j = 1, \ldots, I$, if and only if the functional form of $P(Z_i \mid Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_I)$ is dependent upon the variable Z_j.*

We introduce the notation of the set of neighbors of site i as $\mathcal{N}_i := \{k : k \text{ is a neighbor of } i\}$. As a simple example for illustration, Markov chains of time series theory can be used. Direct transformation yields that every site i, $2 \leq i \leq I-1$ has the two neighbors $i-1$ and $i+1$. The two edge sites 1 and I have only one

neighbor, i.e. 2 and $I-1$ respectively. In the spatial context, an example on a regular (rectangular) lattice is illustrated in figure 4.4.

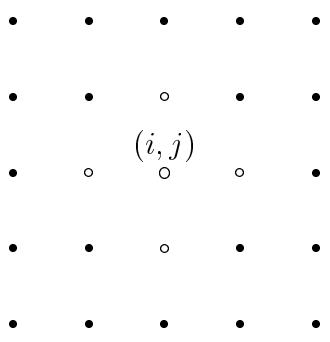

Figure 4.4: First order spatial neighbors on a regular lattice.

For this example only, the sites within the lattice are labelled by integer *pairs* (i, j). Then, if $P(Z_{i,j} \mid \text{all other sites})$ depends only upon $Z_{i-1,j}, Z_{i+1,j}, Z_{i,j-1}$ and $Z_{i,j+1}$ for each internal site on the lattice, then we have a so called "nearest-neighbor" lattice scheme, as described by Besag (1974). This model for spatial dependence results in a geometric approach. Figure 4.4 displays spatial neighbors of order one, second order neighborhood structures can be thought of. Besag and Kooperberg (1995) give an idea of second-order properties of lattice processes. However, difficulties arise when increasing the order of the dependence structure. There will be proportionately more sites lying on an edge leading to more complex likelihood terms. Cressie (1993), p. 422, gives an idea of how to incorporate the problem of edge sites into the likelihood function.

Coming back to first order nearest neighbor models, two definitions of Markov chains exist in spatial stochastic process theory. One is based on a joint, or simul-

taneous specification for the process $\{Z_i, i = 0, 1, \dots\}$ through

$$P(Z_1, \dots, Z_j \mid Z_0) = \prod_{i=1}^{j} Q_i(Z_i, Z_{i-1}) \qquad \text{for all } j \geq 1, \tag{4.10}$$

where Q_j is some function of z_j and z_{j-1}, which are generic realizations of Z_j and Z_{j-1}. However, there is a more familiar definition of a Markov chain based on conditional probabilities

$$P(Z_j \mid Z_0, \dots, Z_{j-1}) = P(Z_j \mid Z_{j-1}) \qquad \text{for all } j \geq 1. \tag{4.11}$$

In time series problems the two specifications for Markov chains via 4.10 and 4.11 are equivalent, due to the unidirectional flow of time. In the spatial context, however, these two modelling approaches can lead to considerable differences (Cressie 1993, p. 404).

4.2.1 Auto-Gaussian models

In the Gaussian context of modelling spatially dependent data, we now describe the simultaneous and the conditional approach in detail.

Simultaneous autoregressive approach (SAR)

First order autoregressive processes are characterized by the equation

$$Z_t = \mu + C\,(Z_{t-1} - \mu) + \epsilon_t$$

or equivalently

$$(U - C)(Z_t - \mu) = \epsilon_t \tag{4.12}$$

where U is the unity matrix of dimension $I \times I$, and ϵ_t is independent of $Z_0, Z_1, \dots, Z_{t-1}$.

Let $\{Z_i, i = 1, \ldots, I\}$ be a Gaussian random vector. If we generalize equation 4.12 to the spatial context we obtain

$$(U - C)(Z - \mu) = \epsilon,$$

where ϵ has expectation vector 0 and covariance matrix Σ_ϵ. It is then easy to show that Z must have the structure (Cressie 1993, p. 406)

$$Z \sim \text{Gau}_I(\mu, (U - C)^{-1}\, \Sigma_\epsilon\, (U - C')^{-1}. \tag{4.13}$$

Conditional autoregressive approach (CAR)

If we assume a spatial regression approach of the form

$$E(Z_i \mid Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_I) = \mu_i + \sum_{j=1, j\neq i}^{I} b_{ij}(Z_j - \mu_j), \qquad i = 1, \ldots, I,$$

satisfying the conditions

$$b_{ij}\, \sigma_j^2 = b_{ji}\, \sigma_i^2,$$

and $b_{ii} = 0$, then it can be shown (Cressie 1993, p. 407) that Z must have the structure

$$Z \sim \text{Gau}_I(\mu, (U - B)^{-1}\, M), \tag{4.14}$$

where $M = \text{diag}(\sigma_1^2, \ldots, \sigma_I^2)$, $\mu = (\mu_1, \ldots, \mu_I)'$ and $B = (b_{ij})$ is an $I \times I$ matrix whose (ij)th element is b_{ij}. Probability measures of equations 4.13 and 4.14 are called Markov random fields, which can be seen as a generalization of temporal Markov processes.

Definition 4.3 *A probability measure with conditional distributions that define a neighborhood structure $\{\mathcal{N}_i, i = 1, \ldots, I\}$, where $\mathcal{N}_i$ are sets of neighbors, is called Markov random field.*

According to Kaluzny, Vega, Cardoso, and Shelly (1998), p. 129, the conditional approach leads to consistent estimates. Cressie (1993), p. 408, mentions that least-squares estimates of the simultaneous autoregressive approach with a Gaussian model yield correlated residuals. In terms of building spatial models for Markov-type departures from independence, it is, as stated by Cressie (1993), p. 404, the conditional approach that turns out to be more natural. We will therefore assume conditional dependence for the following spatial, as well as spatio-temporal models, based on neighboring sites in space and time. Neighboring sites in time are subsequent observations of the time series, and two sites are considered as neighbors in space if they share a common border. An extension of Markov random fields for a spatio-temporal analysis can be found in Lavine and Lozier (1999), where the authors include not only longitude and latitude in an ecological setting, but also the temporal dimension.

From now on we consider the $b_1, \ldots, b_I$ to be spatially structured random effects, constructed on the basis of a conditional autoregressive approach. In a Bayesian setting, this leads to the well-known Markov random field prior (Besag et al. 1991, Bernardinelli, Clayton, and Montomoli 1995, Bernardinelli, Clayton, Pascutto, Montomoli, Ghislandi, and Songini 1995)

$$\pi(b_1, \ldots, b_I \mid \sigma^2) \propto \exp\{-\frac{1}{2\sigma^2} \sum_{i \sim j} (b_i - b_j)^2\}$$

which is equivalent to the assumption that the observations of neighboring sites are identical if the number of neighbors is the same. This prior is also called spatial smoothing prior. The dispersion parameter σ^2 is used to determine the strength of the spatial correlation or smoothing of the quantities of interest. In hierarchical Bayesian modelling σ^2 is used as a hyperparameter. The conditional distribution of one parameter b_i given all the other parameters $b_{\{-i\}}$ has an intuitive form as it

depends only on those parameters b_j of adjacent areas. It can be shown that

$$b_i \mid b_{\{-i\}} \sim \text{Gau}(1/n_i \sum_{i \sim j} b_j, \frac{\sigma^2}{n_i}),$$

where n_i denotes the number of neighbors of site i. Hence the conditional mean is modelled as the average of the neighboring parameters.

Application to the data

We apply the spatial random effects model to lung cancer among women in North Rhine Westphalia. To justify the need for spatial modelling, we measure the strength of spatial autocorrelation for this data set. Moran's correlation coefficient as an exploratory measure of clustering (Cliff and Ord 1973, p. 17, and with application to cancer mortality data in Zöllner 1991) is applied to the nearest neighbor conditional model of the age-standardized mortality rates. Under the assumption of constant mean and variance of the process $\{Z_i, i = 1, \ldots, I\}$ the spatial correlation coefficient after Moran is defined as

$$\rho_I = \frac{I}{2A} \frac{\sum_{i \neq j} b_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{\sum_{i=1}^{I}(Z_i - \bar{Z})^2},$$

where b_{ij} are the elements of the neighborhood matrix $B = (b_{ij})_{i,j=1,\ldots,I} \in I\!R^{I \times I}$, and $A = 1/2 \sum_{i \neq j} b_{ij}$. In this context, $b_{ij} = 1$ if $i \sim j$, i.e. share a common border, and 0 otherwise.

Figure 4.5 shows the correlations with Moran's I over the study period from 1980 to 1998. One can see clearly that the spatial clustering for female lung cancer in NRW tends to increase over the years from 1980 to 1998. The values of ρ_I lie between 0.1 in the beginning of the study period and 0.4 at the end. A permutation test based on Moran's I can be constructed to test for spatial autocorrelation. The correlations for the NRW data set over 19 years prove to be significantly different from 0 on a 95 % level. The maps of North Rhine Westphalia lung cancer in figure 2.2

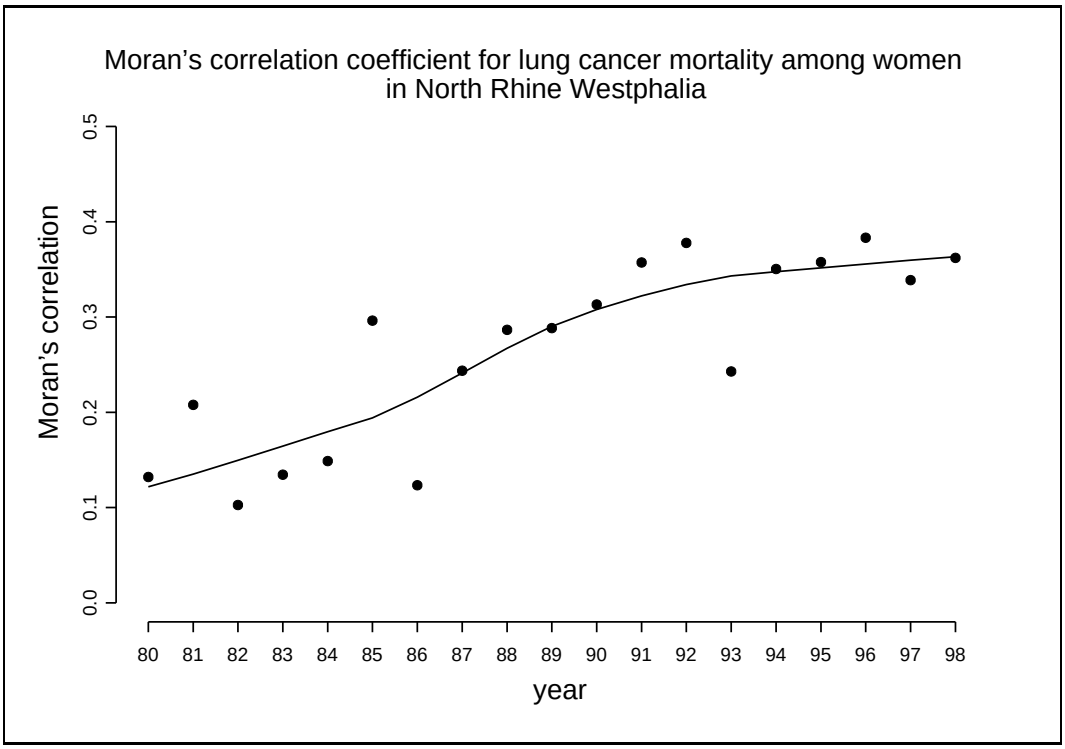

Figure 4.5: Lung cancer among women in NRW. Spatial autocorrelation with CAR model and smoothed with lowess curve.

have already given an indication for this tendency. More sophisticated approaches in the direction of the detection of clusters can be found in Knorr-Held and Raßer (2000). For this thesis, however, Moran's correlations simply serve as an indication for the need of spatial modelling.

We begin with an application of the auto-Gaussian model. The following modelling assumption will be made for the log-transformed age standardized lung cancer mortality rates (denoted by r_i) among women in North Rhine Westphalia

$$\log(r_i) \sim \text{Gau}(\mu_i, \sigma_i^2) \qquad i = 1, \ldots, I. \tag{4.15}$$

As mentioned, the dispersion parameter is modelled as $\sigma_i^2 = \sigma^2/n_i$, which implies that with an increasing number of neighbors (n_i) the variance decreases. The cor-

responding functional relation for μ_i can be denoted as an overall level γ_0, plus the individual spatial random effects b_i

$$\mu_i = \gamma_0 + b_i \qquad i = 1, \ldots, I.$$

Prior distributions have been assigned to b_i, γ_0 and σ^2 as follows

$$\begin{aligned} b_i &\sim \text{Gau}(\bar{b}_i, \sigma_i^2), \qquad i = 1, \ldots, I \\ \gamma_0 &\sim \text{U}(-\infty, \infty) \\ \sigma^2 &\sim \text{InvGamma}(r^*, d^*). \end{aligned}$$

The mean of the Gaussian distribution for the spatial random effects, $\bar{b}_i$, is constructed based on the assumption of a conditional autoregressive distribution of b_i, given its neighboring sites $b_{\{-i\}}$. As the overall mean of the b_i is not defined, the model specification is improper. We need a suitable sum-to-zero constraint on the spatial random effects to ensure identifiability. In this model, the parameters of the inverse Gamma distribution r^* and d^* have been chosen to be equal to one, according to Spiegelhalter et al. (2002). This implies a mean and variance of one. Other specifications are possible, such as choosing both parameters to be 0.01. In that case the mean is equal to 1, but the variance results to be 100. Spiegelhalter et al. (2002) mention however, that with a mean equal to one, the prior mass is placed away from zero, where we expect it to lie. This can lead to an artefactual spatial structure. Kelsall and Wakefield (1999), p. 151, discuss the prior belief that the standard deviation of the random effects is centered around 0.05, with one percent prior probability of being smaller than 0.01 or larger than 2.5. According to Breslow and Clayton (1993), the CAR structure of the random effects leads to an improper prior for the intercept term, as implemented for the parameter γ_0.

We have chosen a burn in of 10,000 iterations. The WinBUGS software uses a modification of the Gibbs sampler to generate the chains. The resulting mean estimates are based on additional 5,000 draws from the posterior distribution, with a

	1981-85		1986-90		1991-95	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
γ_0	1.88	0.03	2.08	0.04	2.34	0.03
σ	0.26	0.02	0.26	0.03	0.25	0.02

Table 4.1: Lung cancer among women in NRW. Parameter estimates for the Gaussian spatial model.

thinning of 100. The model has been applied to three 5-year blocks within the study period, i.e. 1981-85, 1986-90, and 1991-95. The analysis provides mean estimates for the vector of spatial random effects, the overall level γ_0 and the dispersion parameter σ. Table 4.1 displays the resulting estimates for the parameters γ_0 and σ over the three 5-year blocks. The level parameter γ_0 increases over time, whereas its standard deviation remains constant. The dispersion parameter σ has been estimated rather constantly around 0.26 for the three 5-year periods. Figure 4.6 shows the smoothed rates over time, which have been back-transformed to their original scale. Each map displays the 20 % quantiles within the corresponding period.

Smoking is by far the most important risk factor for lung cancer. The smoking behavior of women especially depends on the area of living. As described in chapter 2, fewer women in rural areas are smokers, whereas women in urban living areas have a greater prevalence of smoking. However, data on smoking behavior with the high spatial resolution of districts, cannot be obtained. Kafadar and Tukey (1993) propose to use the logged population density of the largest city within every spatial unit to adjust for urbanization. However, in North Rhine Westphalia there are many rural districts with no large cities at all. To avoid a binary splitting into rural and urban areas, as in figure 2.9, we use a continuous covariate as a surrogate for smoking behavior. We assume the log-transformed population density of the district in the fixed year of 1995 (denoted by x_i) to represent the "urbanicity" of a district. We

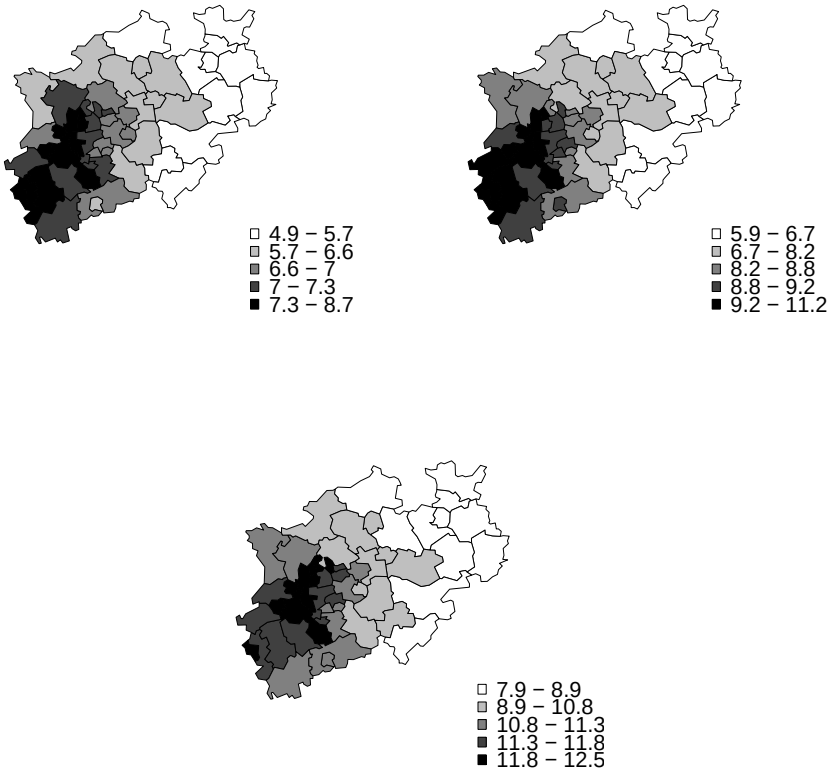

Figure 4.6: Lung cancer among women in NRW. Smoothed rates in 1981-85, 1986-90, 1991-95.

use this covariate to improve the model for lung cancer mortality rates. The model assumption for μ_i can be specified as follows

$$\mu_i = \gamma_0 + \gamma_1 x_i + b_i \qquad i = 1, \ldots, I.$$

A priori we have chosen $\gamma_1 \sim \text{Gau}(0, 10000)$, which results in a proper but flat prior with large variance. The prior specifications for γ_0, b_i, and σ^2 are as described on page 68. The resulting mean parameter estimates with estimated standard deviations are given in table 4.2.

	1981-85		1986-90		1991-95	
	mean	st. dev.	mean	st. dev.	mean	st. dev.
γ_0	0.92	0.30	0.99	0.30	1.35	0.28
γ_1	0.15	0.05	0.16	0.05	0.15	0.04
σ	0.24	0.02	0.24	0.02	0.23	0.02

Table 4.2: Lung cancer among women in NRW. Parameter estimates for the Gaussian spatial model with covariate population density.

The estimated level γ_0 is clearly increasing over the three periods of observation. The influence of the surrogate for smoking behavior remains relatively constant over time. Figure 4.7 shows the mean smoothed mortality rates in 20 % quantiles over time for North Rhine Westphalia. Compared to the smoothed maps without the covariate, we can see that the extreme clusters in the south west of NRW have been broken up.

Finally, we take a look at the residuals of the Gaussian models with and without covariate. Figure 4.8 shows the mapped residuals shaded in 20 % quantiles for the model without the population density (left), and with the covariate (right). One can see clearly that the range of the residuals in the plot without covariate lies between -2.1 and 4.1, whereas the residuals of the map including the covariate only range between -1.5 and 3. Furthermore, the spatial structure of the map without the covariate still reveals clusters in the north and in the south of NRW, whereas the map of the residuals including the covariate are spatially at random.

4.2.2 Auto-Poisson models

An alternative modelling approach is to use a Poisson model to describe the underlying variation in disease risk. Especially when the number of death cases is small,

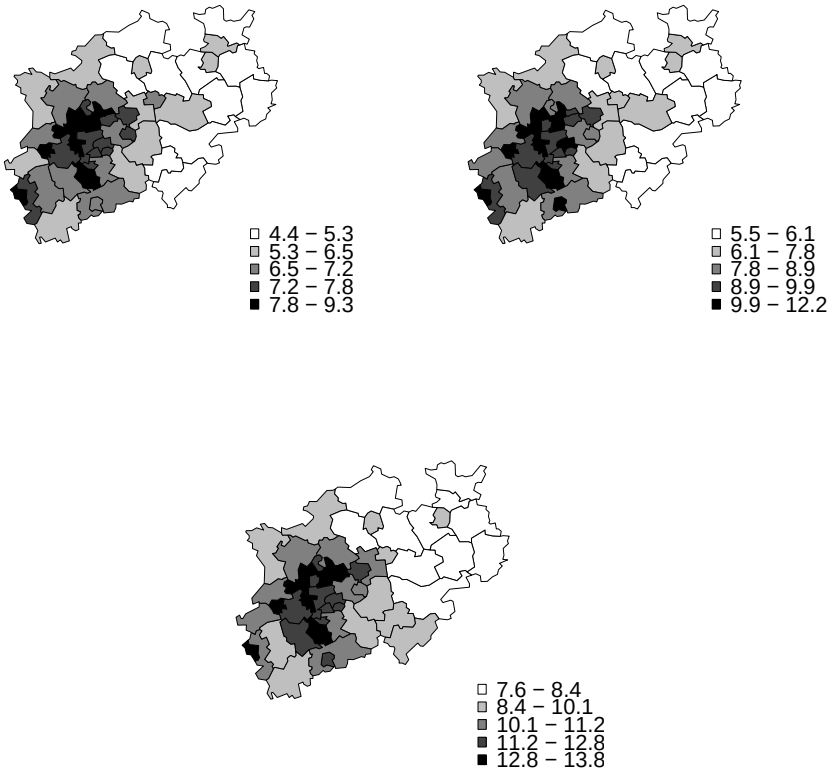

Figure 4.7: Lung cancer among women in NRW. Smoothed rates with covariate urbanization in 1981-85, 1986-90, 1991-95.

Poisson models have the advantage to allow for over-dispersion, without transformation of the data. In addition to the observed counts of cancer mortality O_i for every site we need to calculate the expected numbers of death cases, E_i, depending on gender and age structure of the underlying population. For simplicity we did not calculate the expected counts based on an artificial population with an additional weighting, but on the basis of the population (N_i) in each site of the study area directly. Again we consider aggregated 5-year blocks of mortality and population, like in the auto-Gaussian modelling approach. We obtain the expected counts according

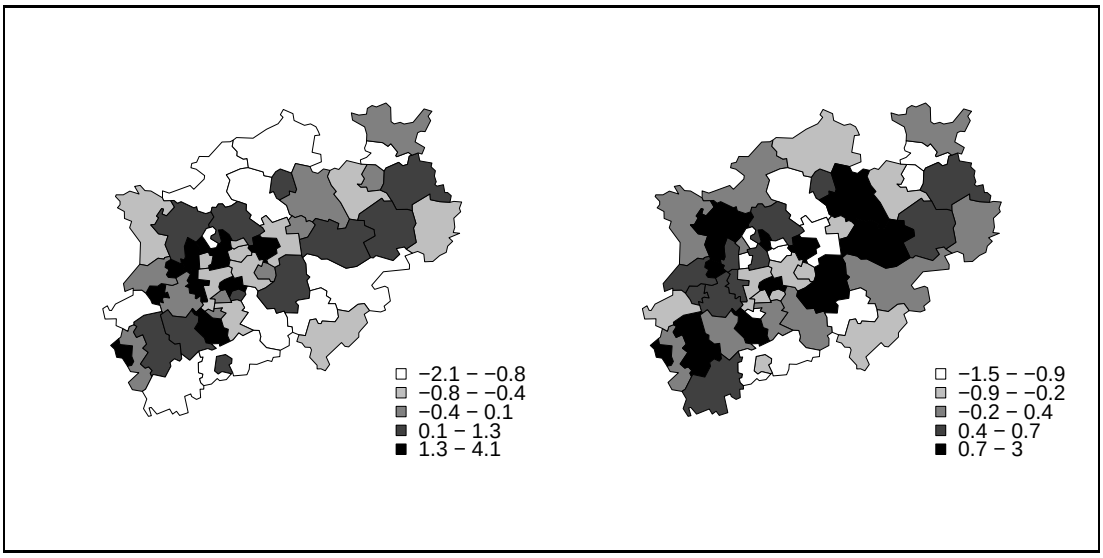

Figure 4.8: Lung cancer among women in NRW, 1986 - 1990. Comparison of the residuals without (left) and with (right) the surrogate for smoking behavior.

to

$$E_i = N_i \cdot \frac{\sum_{i=1}^{I} O_i}{\sum_{i=1}^{I} N_i} \cdot 100,000 \tag{4.16}$$

Following Breslow and Clayton (1993) we consider a random effects Poisson model for the observed mortality counts. The model can be specified as

$$O_i \sim \text{Poi}(\lambda_i), \tag{4.17}$$

with the natural link function $\log(\lambda_i) = \log E_i + \gamma_0 + \gamma_1 \, x_i + b_i$. As for the Gaussian model, the surrogate for smoking behavior (x_i) serves as a covariate proxy. The vector of spatial random effects b_i is assigned a Gau$(\bar{b}_i, \sigma_i^2)$ distribution, again implying a conditional autoregressive structure. $\bar{b}_i$ is the arithmetic mean of the neighboring sites of b_i, i.e. of the sites sharing a common border with site i. As in the Gaussian model, the dispersion parameter σ_i^2 serves as a hyperparameter for the spatial random effects. $\sigma_i^2 = \sigma^2/n_i$, where n_i is the number of neighbors of district i, and σ^2

	1986-90	
	mean	st. dev.
γ_0	-1.59	0.26
γ_1	0.23	0.04
σ	0.36	0.04

Table 4.3: Lung cancer among women in NRW. Parameter estimates for the Poisson spatial model with covariate population density.

is assigned an inverse Gamma prior, i.e. $\sigma^2 \sim \text{InvGamma}(r^*, d^*)$. γ_0 is given a flat prior, whereas γ_1 is assigned a $\text{Gau}(0, 10000)$ non-informative prior.

Application to the data

We apply the Poisson model to lung cancer mortality among women in North Rhine Westphalia for the 5-year period from 1986-90. After a burn-in of 10,000 iterations we have recorded the posterior quantities based on 5,000 draws from the posterior distribution with a thinning of 10. The resulting parameter estimates are displayed in table 4.3.

We have calculated the DIC for the spatial Poisson model without the covariate population density to be 460.519. The DIC for the model including the estimation of the parameter γ_1 as in table 4.3 is 454.843. The reduction of the DIC shows the slightly improved fit of the model to the data using the surrogate for smoking behavior.

We want to compare the spatial smoothing of the Gaussian and the Poisson model, both including the covariate proxy for smoking behavior. Therefore, we plot a so called probability map of the estimated smoothed rates in the Gaussian case, which we have transformed back to its original scale. For the Poisson data

we calculate the quotient of observed and expected case counts, a rate ratio called standardized mortality ratio (SMR)

$$\mathrm{SMR}_i = \frac{O_i}{E_i}, \qquad i = 1, \ldots, I.$$

We have chosen the 5-year period from 1986 to 1990 for the comparison between the auto-Gaussian and the auto-Poisson model. Figure 4.9 shows the 90 % quantiles of the resulting mean rates and rate ratios for the Gaussian and the Poisson model. The areas with the most extreme 10 % of the rates and ratios have been colored black. The two figures are similar in the sense that they indicate high lung cancer mortality

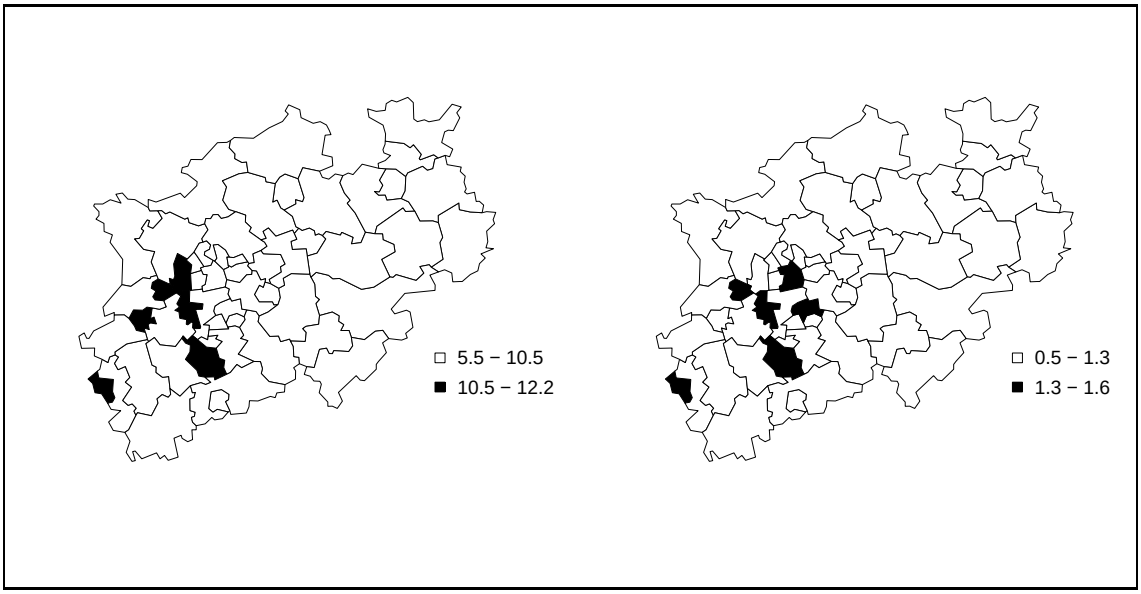

Figure 4.9: Lung cancer among women in NRW, 1986 - 1990. Comparison of the districts with the highest 10 % of mortality with Gaussian model (left) and Poisson model (right).

in the densely populated areas of NRW. However, the districts they identify of having the highest 10 % of lung cancer mortality are not exactly the same. For the districts Aachen, Düsseldorf, Cologne, and Krefeld the results are the same in both models. Furthermore, with Gaussian smoothing Duisburg and Mönchengladbach

are identified, whereas with the Poisson model Essen and Wuppertal lie within the highest 10 % of the SMR's.

However, the use of either a Gaussian or a Poission model, or age-standardized rates or the SMR's respectively, must also be seen from the epidemiologist point of view. The age-standardized rates are easier to calculate. For the calculation of the SMR's we need the expected number of death counts, and additional assumptions concerning the population structure. Furthermore, it is common to use the SMR's only for rare events, which lung cancer death cases are clearly not. The most popular example for an application of the Poisson model based on SMR's is the data set on lip cancer in Scotland with observed cases varying between 0 and 39 for the 56 regions of Scotland. Using the SMR's for rare events makes the results more robust with respect to small changes in the data which can have a great influence. In our data sets on lung cancer and stomach cancer, however, this problem does not occur.

4.2.3 A spatial linear model approach

In this section an additional spatial modelling approach and its application to the data set will be discussed in brief. It is a simplification of the most general correlation structure resulting in a conditional autoregressive approach, the spatial linear model (SLM). It is a model of the class of generalized linear models, that incorporates the correlation structure of the data through a specified covariance matrix. A general form of these models is

$$Z_i = \mu_i + \epsilon, \qquad i = 1, \ldots, I,$$

separating the levels of variation in area specific large scale variation μ_i and iid. random errors, $\epsilon \sim \text{Gau}_I(0, \Sigma_\epsilon)$ accounting for the small scale variation. The small scale variation can be modelled in different approaches, e.g. a conditional, a simultaneous, or a moving average approach. Again, we imply a conditional autoregressive

structure, which enters in the model through the specific type of covariance matrix. μ_i may be constant or a linear model with covariates. We have set up the dispersion matrix following Kaluzny, Vega, Cardoso, and Shelly (1998), p. 128 as

$$\Sigma = \sigma^2 \, (U - \rho \, B)^{-1}. \tag{4.18}$$

ρ and σ are scalar parameters, estimated iteratively along with regression coefficients of the GLM. B is a weighted neighbor matrix, that consists of the assumed neighborhood structure. The covariance of 4.18 is a special case of the class of more general spatial covariance functions described by 4.14. With this simplification it is possible to obtain the parameter estimates through a frequentist estimating approach using generalized least squares. The parameters of small scale and large scale variation interact, which is why the model is fitted iteratively.

Application to the data

The log-transformed age standardized rates for the three five year blocks have been examined with respect to lung cancer among women in North Rhine Westphalia. μ_i is modelled as a linear function of the intercept term and the covariate as follows.

$$\mu_i = \gamma_0 + \gamma_1 \, x_i \qquad i = 1, \ldots, I$$

The log-transformed population density of 1995 has again been used as a covariate. The number of necessary iterations lies between 8 and 11 for these data. Table 4.4 gives the results for the fitted spatial linear model. The scalar ρ is a measure of spatial correlation, as described in equation 4.18. The results of the spatial linear model compared to those obtained with Bayesian inference in table 4.2 are very similar. The estimated level γ_0 is increasing over time, whereas the parameter γ_1 is almost constant.

To conclude, the spatial linear model requests a simplification of the more general covariance matrix of the Markov random field approach described in 4.14. However,

	1981-85	1986-90	1991-95
γ_0	0.80	0.83	1.35
γ_1	0.16	0.19	0.15
σ	0.17	0.17	0.14
ρ	0.14	0.15	0.15

Table 4.4: Lung cancer among women in NRW. Parameter estimates for the Gaussian spatial linear model.

the assumption of equal variances of the site specific observations seems to be reasonable in this application, since cancer rates as well as underlying population sizes do not differ very much between the sites. Furthermore, it is also plausible to assume that only neighboring sites influence each other, which is the justification for using a particular neighborhood matrix B. ρ then acts as a regression parameter for the degree of spatial influence.

Using Bayesian methods, it is possible to estimate ρ and σ. Additional to fixed covariate effects, site specific random effects can be estimated.

Finally, it remains to mention that the spatial linear model as presented in the GLM environment needs the assumption of a Gaussian observation model. Hence, possibly more appropriate models in other applications, such as Poisson models are not feasible.

Chapter 5

Autoregressive Spatio-Temporal Modelling

When modelling data dependent on space and time, it becomes clear that separate approaches for a temporal and a spatial analysis are not satisfactory. This chapter deals with the combined analysis of temporal and spatial dependence structures, which leads to an increased complexity of the model. The results obtained in the foregoing chapter will be used, where we have learned from the purely temporal and purely spatial analyses. We aim to apply the model to stomach cancer among men and women in Germany over the 15 year observation period from 1976 to 1990 (Germany data). Figure 5.1 gives an idea of the relative changes of the quintiles of stomach cancer among men over time. In the south of Germany the rates are constantly high. In the north, however, there is a clear relative increase, especially in the region of Schleswig-Holstein. We use a frequentist approach based on maximum likelihood and a Bayesian approach to estimate autocorrelation parameters of the spatio-temporal model.

Furthermore, we use the data set on lung cancer among women in NRW (NRW data) for a spatio-temporal analysis. As the observation period is 19 years, and

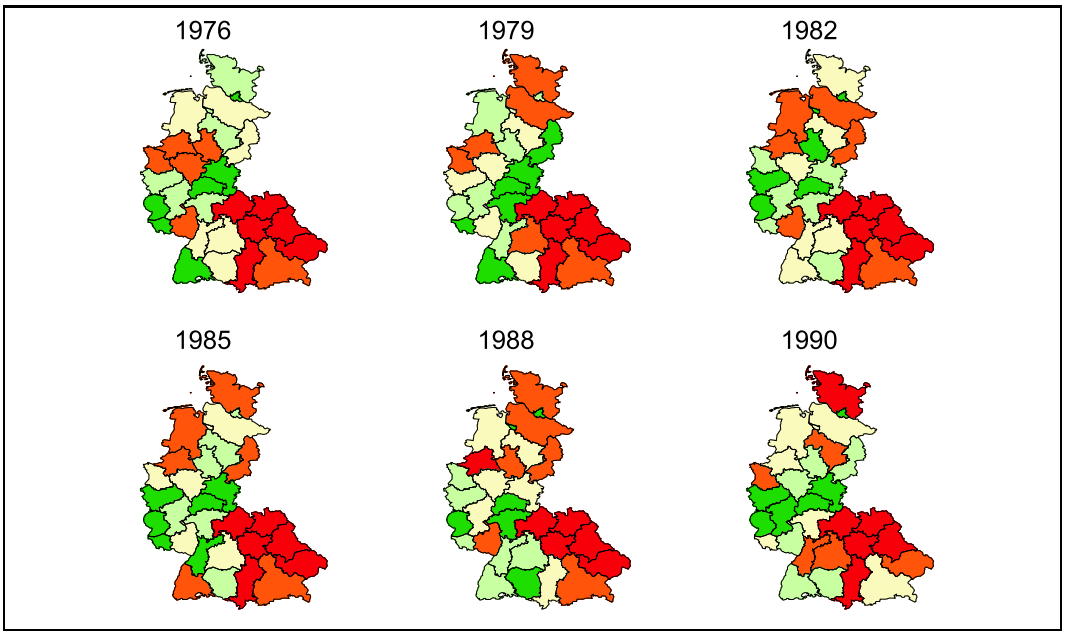

Figure 5.1: Stomach cancer among men in the regions of Germany. Relative changes of the rates over time.

we have 54 spatial units, computing times in the likelihood approach are largely increased, and hence the frequentist approach is infeasible. Therefore, we use solely Bayesian inference to estimate the unknown parameters and perform model selection according to the DIC.

At the end of the chapter we compare the results of the likelihood and the Bayesian approach for stomach cancer. Emphasis will be put on the comparison of the parameter estimates with supported range, and on computational aspects.

5.1 Space-time vector-autoregressive processes

In the first section of chapter 4 we have described the autoregressive structure of the temporal component of the stomach cancer and the lung cancer data set. Therefore, we again assume a first order VAR structure. In order to model the spatial depen-

dencies we employ the conditional autoregressive approach, with a neighborhood structure based on common borders through the adjacency matrix B. In combination, we have chosen a spatio-temporal modelling approach that uses a Markov type dependence in space and time. Figure 5.2 displays the assumed autoregressive correlation structure in time and space.

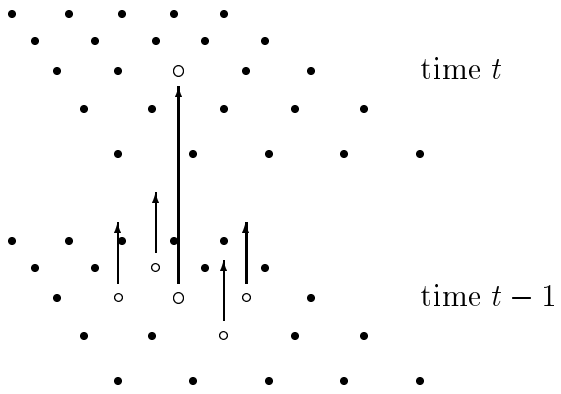

Figure 5.2: First order neighborhood scheme on a regular lattice in space and time.

In words, we consider the outcome of site i at time t to be influenced by the neighbors of i at time $t-1$, the set of units $\{-i\}$, and the outcome of site i at time $t-1$. We have chosen this approach to model cancer mortality data, as we find it reasonable to assume that subsequent sites in time are likely to be highly correlated. Neighboring sites are assumed to have an influence on the outcome at time t as well, but not directly. There is a time lag of one year, in this application, after which the the influence expresses. More formally, we assume $\{Z_t, t = 1, \ldots, T\}$ to be a multivariate stochastic process of spatially dependent time series. In section 4.1 we have introduced multivariate VAR time series to model the area specific cancer mortality rates over time. However, we have considered the time series to be independent. If we generalize the assumptions of section 4.1 to the case of

dependent time series, we may assume $\{Z_t\}$ to be a stochastic process with the following characteristics (Schach 2002a):

i) $Z_1, \ldots, Z_t \sim \text{Gau}_I(0, \Sigma)$,

ii) $\{Z_t\}$ has the Markov property,

iii) $\{Z_t\}$ is second order stationary.

Let ϵ_t be an innovation process as in equation 4.7. The aim is to build the spatial dependence structure into the process via a suitable matrix C.

We have chosen to model the influence of the neighbors of i with a temporal lag of 1, as we assume a stronger temporal than spatial relationship. In a simultaneous approach of spatial and temporal correlation, identifiability problems can occur for the autocorrelation parameters. We assume additivity of spatial and temporal effects, and hence arrive at the model according to Schach (2002a)

$$Z_t = \alpha\, Z_{t-1} + \beta\, B\, Z_{t-1} + \epsilon_t, \qquad t = 1, \ldots, T. \tag{5.1}$$

Again, $\epsilon_1, \epsilon_2, \ldots$ are independently and identically distributed random errors, distributed according to $\text{Gau}_I(0, \Sigma_\epsilon)$. Especially, ϵ_t are independent of $Z_0, \ldots, Z_{t-1}$. The parameter α measures the temporal autocorrelation, and β is a spatial regression parameter. B is the neighborhood, or adjacency matrix. The (i, j)th element of B is zero, if the regions i and j share no common border, or if $i = j$. Otherwise, $b_{ij} = 1/n_i$, where n_i is the number of neighbors of region i. The question is, whether the process as specified in 5.1 is admissible, in the sense that it converges to a stationary distribution. Therefore it is the aim to investigate, whether we can construct a matrix, that incorporates the type of spatial and temporal dependence structure as specified in figure 5.2. If we define

$$C := \alpha\, U + \beta\, B \tag{5.2}$$

then equation 5.1 is a special case of 4.1 for a mean polished process with $\nu = 0$. The process $\{Z_t, t = 1, \ldots, T\}$ can be expressed recursively on the basis of the constant matrix C, based on the initial distribution at time $t = 1$, and the independent innovation terms $\epsilon_{t-j}, j = 0, \ldots, \infty$ through the relation

$$Z_t = \sum_{j=0}^{\infty} C^j \, \epsilon_{t-j}. \tag{5.3}$$

Parameters of the model can be estimated iteratively, including the estimation of the covariance matrix of the process (Σ) based on the innovation covariance (Σ_ϵ), as will be described in the following section. The conditions under which the matrix C converges are determined indirectly through the autocorrelation parameters α and β. We have chosen the spectral matrix norm, which satisfies the triangle inequality and hence

$$\begin{aligned} \|C\| &= \|\alpha\, U + \beta\, B\| \\ &\leq \|\alpha\, U\| + \|\beta\, B\| \\ &\leq \mid \alpha \mid \, \|U\| + \mid \beta \mid \, \|B\|. \end{aligned}$$

The spectral norm of the identity matrix is 1, the spectral norm of the neighborhood matrix B depends on the study area, of course. If we consider the regions of West Germany, we have 30 spatial units, and the corresponding spectral norm is 1.089. For North Rhine Westphalia, with its 54 districts, the result is 1.171. Obviously, this resulting matrix norm is strongly dependent upon the choice of the specific neighborhood structure with corresponding weights. It follows that a sufficient condition for convergence of the infinite series and thus for the admissibility of the model is

$$\mid \alpha \mid + 1.089 \mid \beta \mid < 1$$

for the spatial structure of West Germany, and

$$\mid \alpha \mid + 1.171 \mid \beta \mid < 1$$

for the structure of North Rhine Westphalia.

5.2 Parameter estimation with likelihood approach

Due to the Markov property of the stochastic process Z_t, the likelihood function can be written as a product of the conditional transition probabilities, although the observations are not independent. Yet, the covariance matrix of the process Σ is unknown, but we know that $\Sigma = C\,\Sigma\,C' + \Sigma_\epsilon$. Σ can be derived after transforming the Z-process as follows. The recursive specification of 5.3 leads to the following equation for the covariance matrix, similar to equation 4.6

$$\Sigma \;=\; \sum_{j=0}^{\infty} C^j\,\Sigma_\epsilon\,(C^j)'.$$

If we assume homogeneous variances and independent innovation terms, the innovation covariance matrix reduces to $\Sigma_\epsilon = \sigma_0^2\,U$. Hence, Σ can be rewritten as

$$\Sigma \;=\; \sigma_0^2\,\sum_{j=0}^{\infty} C^j\,(C^j)'. \tag{5.4}$$

We consider ϵ_t as noise, which acts on the components of the process. Hence, it is reasonable to assume that all components have the same variance and are independent of each other and the assumption $\Sigma_\epsilon = \sigma_0^2\,U$ is plausible. We use a sensible stopping rule for the approximation of the infinite geometric series in 5.4 dependent on the spectral matrix norm.

The likelihood function for the three unknown parameters α, β, and σ_0^2 for the above model can be formulated as follows

$$\begin{aligned} l(\alpha,\beta,\sigma_0^2 \mid z_1,\ldots,z_T) \;=\; & \frac{1}{(2\pi)^{\frac{I}{2}}\,\sqrt{\det\Sigma}}\,\exp\{-\frac{1}{2}\,z_1'\,\Sigma^{-1}\,z_1\}\cdot \\ & \prod_{t=2}^{T}\frac{1}{(2\pi)^{\frac{I}{2}}\,(\sigma_0^2)^{\frac{I}{2}}}\,\exp\{-\frac{1}{2}\,\frac{1}{\sigma_0^2}\,\|z_t - C\,z_{t-1}\|^2\}. \end{aligned} \tag{5.5}$$

When including a matrix of covariates and a vector $\gamma = (\gamma_0, \ldots, \gamma_p)'$ of regression coefficients in the model, two different cases can be formulated. In the first case the covariates are constant over time, which results in the following model formulation

$$Z_t = \alpha\, Z_{t-1} + \beta\, B\, Z_{t-1} + \gamma\, X + \epsilon_t, \qquad t = 1, \ldots, T.$$

Thus the likelihood of 5.5 can be generalized to

$$\begin{aligned} l(\alpha, \beta, \gamma, \sigma_0^2 \mid z_1, \ldots, z_T) &= \frac{1}{(2\pi)^{\frac{I}{2}} \sqrt{\det\Sigma}} \exp\{-\frac{1}{2}\,(z_1 - X\,\gamma)'\,\Sigma^{-1}\,(z_1 - X\,\gamma)\} \\ &\quad \prod_{t=2}^{T} \frac{1}{(2\pi)^{\frac{I}{2}}\,(\sigma_0^2)^{\frac{I}{2}}} \exp\{-\frac{1}{2}\,\frac{1}{\sigma_0^2}\,\|z_t - C\,z_{t-1} - X\,\gamma\|^2\}. \end{aligned}$$

Here, $\gamma_0, \ldots, \gamma_p$ are the unknown regression coefficients and X is a constant regressor matrix.

In the second case, the covariates have a temporal structure, i.e. X is replaced by $X^{(t)}$. Hence

$$Z_t = \alpha\, Z_{t-1} + \beta\, B\, Z_{t-1} + \gamma\, X^{(t)} + \epsilon_t, \qquad t = 1, \ldots, T.$$

The regressor matrix can be considered to form a separate stochastic process over the observed time period. In that case, the resulting likelihood is written as

$$\begin{aligned} l(\alpha, \beta, \gamma, \sigma_0^2 \mid z_1, \ldots, z_T) &= \frac{1}{(2\pi)^{\frac{I}{2}} \sqrt{\det\Sigma}} \exp\{-\frac{1}{2}(z_1 - X^{(1)}\,\gamma)'\,\Sigma^{-1}\,(z_1 - X^{(1)}\,\gamma)\} \\ &\quad \prod_{t=2}^{T} \frac{1}{(2\pi)^{\frac{I}{2}}\,(\sigma_0^2)^{\frac{I}{2}}} \exp\{-\frac{1}{2}\,\frac{1}{\sigma_0^2}\,\|z_t - C\,z_{t-1} - X^{(t)}\,\gamma\|^2\}. \end{aligned} \tag{5.6}$$

This specification of the likelihood function can be used to model a linear temporal trend within the data. More specifically, with $X^{(t)} = t$ the model formulation including fixed covariate effects results in

$$Z_t = \alpha\, Z_{t-1} + \beta\, B\, Z_{t-1} + \gamma\, X + \delta\, t + \epsilon_t, \qquad t = 1, \ldots, T. \tag{5.7}$$

In this special case the time points themselves will be viewed as a covariate dependent on time.

As mentioned in section 3.2, the parameters of the space-time model will be estimated using a frequentist approach based on the maximum likelihood estimator. The ML estimator has been chosen for this problem, since it has good asymptotic behavior for ergodic Markov chains under relatively weak regularity conditions, as examined in Schach (2000). According to Hall and Heyde (1980), p. 156, it can be shown that the ML estimator is asymptotically normal, consistent, and sufficient, even in the case of dependent observations. Hence, statistical tests and confidence intervals based on the normal distribution can be built. However, it is not always possible to maximize the log-likelihood analytically, especially in a high dimensional multivariate problem. We consider a model without covariates but with temporal trend δ as specified in 5.7, therefore $\theta = (\alpha, \beta, \sigma_0^2, \delta)'$. In this case the estimates can only be obtained using a numerical or iterative procedure. One possible algorithm is based on a modified quasi-Newton method, which is used by S-Plus in order to avoid calculating the Hessian matrix of second derivatives for four unknown parameters. As the quasi-Newton method is endangered to provide local instead of global extrema, different starting values will be given to obtain reliable results. Further discussion on m-dimensional time series and the calculation of the exact likelihood function for spatially and temporally correlated data can be found in Abraham (1983) and Aroian (1980). Cressie and Huang (1999) discuss classes of spatio-temporal stationary covariance functions in the context of point source data, or geostatistics.

Application to the data

We begin with the application of the space-time model to the stomach cancer mortality data in West Germany. The observed period of time begins in 1976 and runs

for 15 years until 1990. We consider the rather coarse spatial structure of 30 regions. We begin with a model without covariates, and the likelihood of equation 5.6 is used for the analysis. We include the linear temporal trend parameter δ, as the figures 2.6 and 2.5 show clearly a nonconstant decreasing mean for stomach cancer. The maximization of the likelihood function from dispersed starting values has lead to the following parameter estimates for men and women.

	parameter	men	women
temporal AC	α	0.691	0.699
spatial AC	β	0.197	0.115
dispersion	σ_0	0.119	0.194
temporal trend	δ	-0.005	-0.003

Table 5.1: Stomach cancer in Germany. Parameter estimates of the space-time model with likelihood approach.

We can see that the temporal correlation is stronger than the spatial dependence. It turns out that the results are very similar if we compare men and women. The largest difference lies in the parameter estimation of the spatial autocorrelation. As we have expected, the temporal trend is negative.

We have performed the purely temporal and purely spatial analyses for lung cancer among women in North Rhine Westphalia, but we would like to apply the spatio-temporal model to this data, as well. However, the state consists of 54 districts and the observation period is 19 years, which leads to an infeasible increase in computing time with the likelihood approach compared to the 30 regions over 15 years. For that reason the parameters of the space-time model for lung cancer cannot be estimated using the ML method.

5.3 Parameter estimation with Bayesian approach

We have indicated that apart from the classical maximum likelihood estimation, we aim to use Bayesian inference for the space-time model. Similar approaches to a combined space-time modelling can be found in Knorr-Held and Besag (1998), where the number of cancer deaths is assumed to follow a binomial distribution dependent on time, space, age-group, and site-specific covariate effects. The two models presented there aim for the adjustment of unmeasured spatial covariates via random effects, and with a surrogate for smoking behavior. Waller et al. (1997) introduce a Poisson modelling approach, and propose to use a first order temporal autoregressive structure, and an auto-Gaussian model (Besag 1974) for the spatial component.

According to Schach (2002a), we begin with a three-stage Gaussian hierarchical model for the age standardized and log-transformed mortality rates r_{ti}. The parameters α, β, and δ are assigned non-informative Gaussian prior distributions, centered around zero. As the time series is modelled assuming an autoregressive structure, we need to consider time $t = 1$ separately from the following points in time $t = 2, \ldots, T$, similar to the likelihood approach. For $t = 1$ we introduce a generalized linear mixed model for the spatially structured random effects b_{1i} with a conditional autoregressive structure. In the Bayesian approach we deal with two dispersion parameters σ_i^2 and σ_0^2. $\sigma_i^2 = \sigma^2/n_i$ is the dispersion parameter for the spatial random effects at time $t = 1$, and it is assigned an inverse Gamma distribution with shape parameter 0.0001, and inverse scale parameter 0.01. Furthermore, σ_0^2 is the dispersion of the rates, which we assign an inverse Gamma distribution with the same parameters. Based on the model assumptions, the Bayesian space-time model is set up as follows.

$t = 1$: **Modelling of the spatial random effects** b_{1i}

$$b_{1i} \sim \text{Gau}(\bar{b}_{1i}, \sigma_i^2)$$

As in the purely spatial random effects model, $\bar{b}_{1i}$ denotes the mean of the adjacent regions of i at time $t = 1$.

$t = 2, \ldots, T$: **Autoregressive modelling of the rates** r_{ti}

$$r_{ti} \sim \text{Gau}(\mu_{ti}, \sigma_0^2)$$

The expectation of r_{ti} is modelled through the functional relation

$$\mu_{ti} = \alpha\, r_{t-1,i} + \beta \bar{r}_{t-1,\{-i\}} + \delta\, t \quad t = 2, \ldots, T,\ i = 1, \ldots, I.$$

Here, $\bar{r}_{t-1,\{-i\}}$ indicates the average rate among the neighbors of $r_{t-1,i}$. In contrast to the purely spatial model with spatial random effects, as presented in section 4.2.1, the mixed model approach is only entering at time $t = 1$. From $t = 2$ we imply a temporally autoregressive structure, and we do not aim for the estimation of spatial random effects for $t > 1$. Again, the stochastic simulations of the Markov chains have been carried out with the WinBUGS software.

Application to the data

We have chosen a burn-in phase of 5,000 iterations. The parameter estimates are based on 1,000 observations recording every 10th iteration. Table 5.2 gives the resulting mean estimates again for stomach cancer among men and women for West Germany. The resulting parameter estimates are very similar to those obtained with the likelihood approach. Additionally, the dispersion of the spatial random effects (σ) has been estimated with 0.154 (standard deviation 0.005) for men, and 0.179 (standard deviation 0.006) for women.

We have carried out a sensitivity analysis to evaluate the dependence of the posterior on the prior distribution. Instead of assigning non-informative Gaussian

	parameter	men	women
temporal AC	α	0.691	0.698
spatial AC	β	0.197	0.114
dispersion	σ_0	0.170	0.152
temporal trend	δ	-0.005	-0.004

Table 5.2: Stomach cancer in Germany. Bayesian parameter estimates of the space-time model.

priors to the autocorrelation parameters α and β and the trend parameter δ, we have replaced them by improper flat priors, by Gau(0,25) priors, and by U(-2,2) prior distributions. The corresponding posterior distributions for the parameter α are displayed in figure 5.3. Figure 5.3 shows identical posteriors of the parameter α for the considerably different prior distributions. The evaluation of β and δ leads to similar results. We have also changed the hyper-parameters of the inverse Gamma distribution for σ_0^2 and σ^2 to InvGamma$(0.001, 0.01)$ (according to Spiegelhalter, Thomas, and Best 2002), and to InvGamma$(0.5, 0.0005)$ as proposed by Kelsall and Wakefield (1999). The resulting parameter estimates for the dispersion parameters, as well as the other parameters were not affected.

Apart from modelling stomach cancer mortality in Germany, we have applied the spatio-temporal model to lung cancer mortality among women in North Rhine Westphalia. After a burn-in phase of 5,000 iterations, and again based on a sample of 1,000 from the posterior distribution with thinning 10, we have obtained the parameter estimates as displayed in table 5.3. It is interesting to see that the spatial autocorrelation seems to be much stronger for lung cancer than in the stomach cancer example. The trend is estimated to be positive, as the lung cancer rates among women increase. Again, the dispersion parameter σ for the spatial random effects at time $t = 1$ has been estimated as 0.298 with an estimated standard deviation of

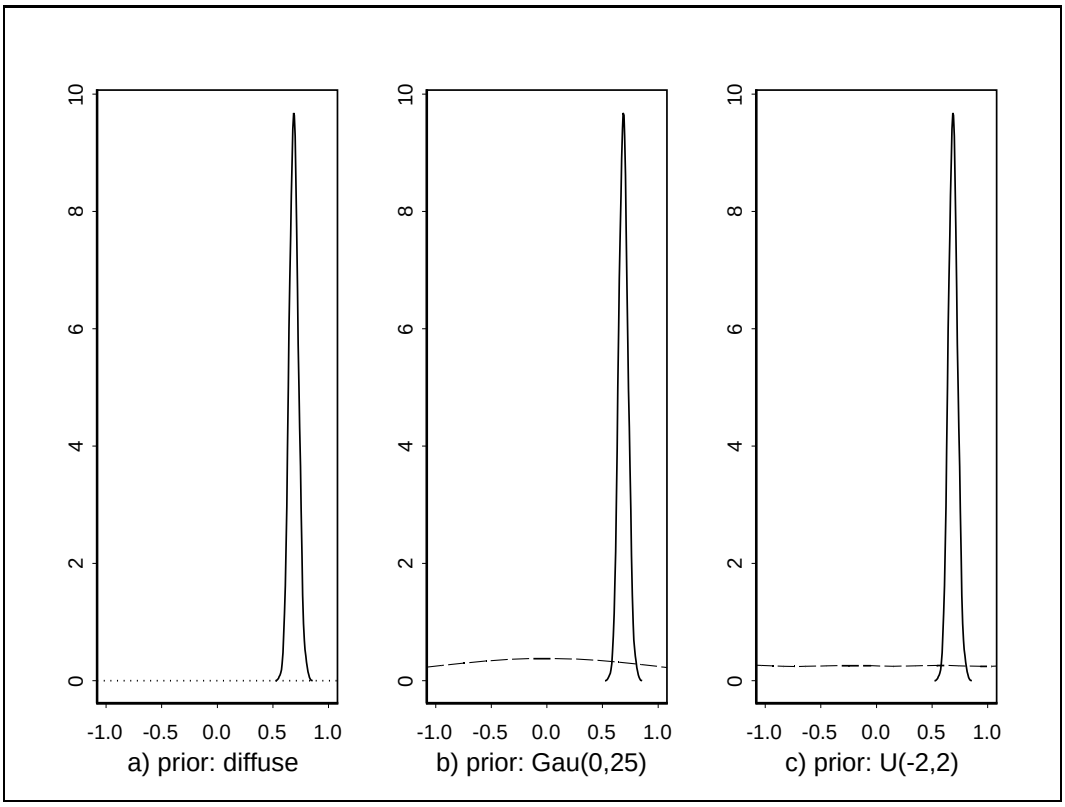

Figure 5.3: Evaluation of the prior influence on the estimated posterior density for the autocorrelation parameter α; improper prior (indicated by a dotted line at 0) for a), proper priors for b) and c).

0.007.

We have calculated the DIC, which results to be 400.667 for the model without covariate. We want to quantify the influence of the covariate smoking behavior. Similar to the purely spatial model in chapter 4 we use the log-transformed population density in 1995 as a constant regressor over time, denoted by x. This changes the functional relation of μ_{ti} as follows

$$\mu_{ti} = \alpha\, r_{t-1,i} + \beta \bar{r}_{t-1,\{-i\}} + \gamma_1\, x_i + \delta\, t \quad t = 2, \ldots, T,\ i = 1, \ldots, I.$$

We have assigned a non-informative Gaussian prior to the corresponding parameter

	parameter	post. mean	st. dev.
temporal AC	α	0.373	0.029
spatial AC	β	0.622	0.032
dispersion	σ_0	0.177	0.028
temporal trend	δ	0.008	0.003

Table 5.3: Lung cancer among women in NRW. Bayesian parameter estimates of the space-time model.

γ_1. The resulting parameter estimates are displayed in table 5.4. The dispersion

	parameter	post. mean	st. dev.
temporal AC	α	0.212	0.031
spatial AC	β	0.276	0.039
dispersion	σ_0	0.172	0.029
temporal trend	δ	0.031	0.003
smoking	γ_1	0.130	0.011

Table 5.4: Lung cancer among women in NRW. Bayesian parameter estimates of the space-time model with covariate.

parameter at time $t = 1$ has been estimated as 0.278, with an estimated standard deviation of 0.006. Compared to the model without covariate effects, much of the spatial correlation seems to be explained by the surrogate for smoking behavior. The reduction in spatial correlation is clearly visible, and smoking has been estimated to have a significantly positive effect in the sense that 0 does not lie within the 95 % confidence interval. The DIC is reduced to 267.826 for this model, and therefore indicates the superiority over the model without covariate.

In the following section the parameter estimates of the frequentist and the

Bayesian approach with respect to stomach cancer among men in West Germany will be compared and discussed.

5.4 Comparison

The tables 5.1 and 5.2 show, that the parameter estimates obtained with classical ML estimation and with Bayesian inference are approximately the same, especially for the temporal autocorrelation parameter α, and the temporal trend δ in the case of stomach cancer in Germany.

We have described the computational infeasibility of the frequentist approach if the number of spatial units, or the length of the times series increases. The parameter estimates obtained with ML estimation are based on an average of six iterations until convergence. Computing times are relatively long, whereas in the Bayesian approach hundreds of samples of the posterior distribution can be drawn within seconds of time.

Furthermore, the dependence of the parameter estimates on the starting values is problematic in the likelihood approach, because for many combinations of starting values, the maximization procedure of the likelihood function breaks off and no estimates can be obtained. The independence of the parameter estimates of the starting values in the Bayesian approach is visible in figure 5.4, with the dispersion parameter of the rates σ_0 for female stomach cancer as an example. Starting values have been set to 1, 0.5, and 0.1 for the parallel chains. After the first five iterations, all three chains have passed the transient phase independently of the starting values, whereas starting values for σ_0 larger than 0.2 have turned out to be infeasible in the likelihood approach. The same behavior can be observed for the other parameters.

The estimation of the confidence intervals with the likelihood approach is rather difficult, as it is necessary to calculate the variance-covariance matrix of the unknown

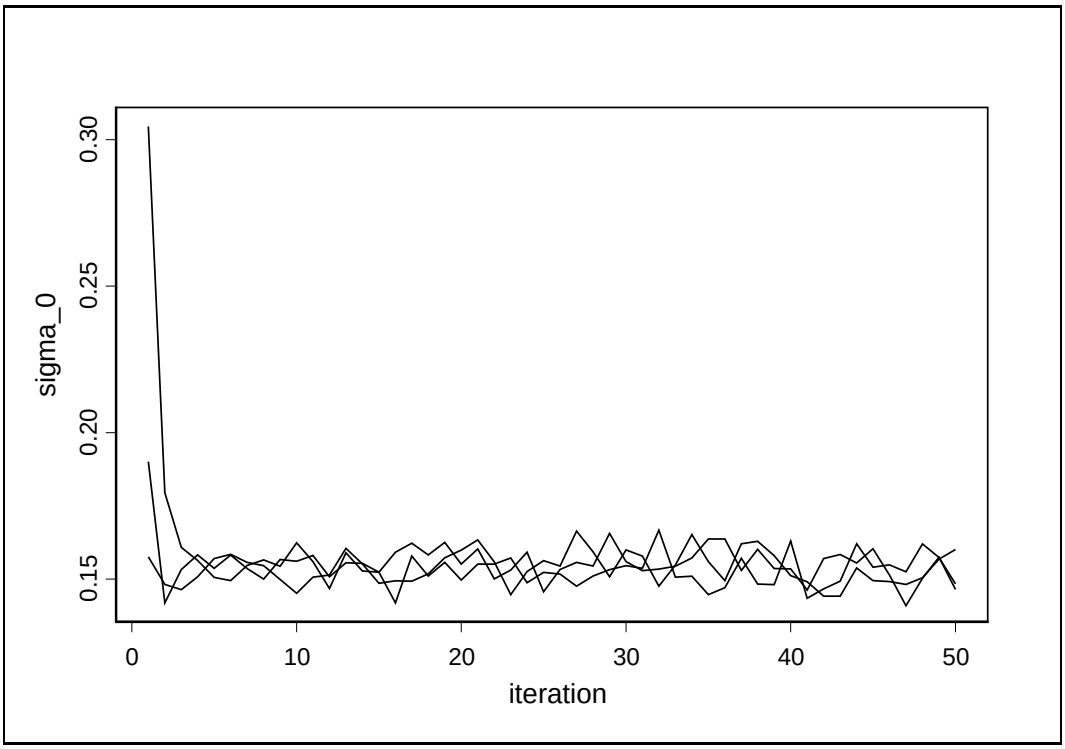

Figure 5.4: First 50 iterations of 3 chains for the standard deviation of the rates in the space-time model.

parameters. We have used S-Plus functions for the calculation, where computing times for the 4 unknown parameters have been relatively long. The corresponding credible regions can easily be calculated for the Bayesian posterior quantities. Table 5.5 gives the results for stomach cancer mortality data among men. The results obtained with the two approaches match to a large extent. The confidence intervals of the likelihood approach show a slight tendency to be larger than those obtained with Bayesian inference.

We have mentioned the difficulty to obtain parameter estimates with likelihood inference for this model. Computing times are extremely prolonged when calculating the variance-covariance matrix to test for significance of the parameters. It does not seem possible to obtain the parameter estimates and the variance-covariance matrix

	parameter	Bayesian	frequentist
temporal AC	α	(0.691 ± 0.066)	(0.691 ± 0.080)
spatial AC	β	(0.197 ± 0.077)	(0.197 ± 0.090)
dispersion	σ_0	(0.170 ± 0.061)	(0.119 ± 0.012)
temporal trend	δ	(-0.005 ± 0.001)	(-0.005 ± 0.001)

Table 5.5: Stomach cancer among men in Germany. 95 % credible regions and confidence intervals for the Bayesian and the frequentist approach.

analytically. On the other hand, the generation of credible regions with Bayesian inference is fast and easy. To conclude, we have seen clear advantages of the Bayesian approach especially with respect to computational feasibility, computational speed, and independence of starting values.

Chapter 6

Application to Small Area Estimation

The frequency and distribution of cancer mortality in Germany is published in two different data sets, with a different spatial and temporal resolution. One set has a coarse, aggregated spatial structure of the 30 regions ("Regierungsbezirke") of West Germany, but contains yearly data. The other data set contains cancer mortality data for the 328 districts ("Landkreise") within the same study region, but the data are aggregated over 5-year periods in time. The data with the highest spatial and temporal resolution is not available for further analysis due to privacy protection and tabulation procedures in Germany. When regarding further subdivision by age-group, or rare cancer types, the time-by-location cell frequencies become too small to be published. We have examined the spatial and temporal dependence structure of the marginal data in the foregoing chapter. Based on the knowledge gained from the data sets, we now aim for a spatio-temporal small area estimation. We present here a model based small area estimation to obtain estimates with the highest temporal resolution. We begin with a brief description of the idea of a typical synthetic (section 6.1) and model based small area estimation (section 6.2).

6.1 Conventional small area estimation

Many large socio-economic or health related samples were originally designed to give national and regional estimates for population specific parameters. For the high level of aggregation, they are constructed to provide a certain accuracy. In Germany for example, the so called "Mikrozensus" survey is a yearly survey based on one percent of German households. The survey is designed to obtain estimates only for the whole of Germany, or for large areas within, e.g. states, and it is based on a stratified cluster sample. However, estimates are also needed for regions, districts, cities, or even smaller administrative units, which are not well represented by the large survey data. One way to meet this demand is to redesign these surveys, or design new surveys, but this is usually too expensive and time consuming. Small area estimation is an alternative, as it can be seen as an intra-polation of information collected on a larger spatial scale to local areas within the study region, see Schach (2002b). It is concerned with the estimation of parameters corresponding to small geographical areas or subpopulations. The underlying idea is to pool data from other areas to estimate the parameters for a particular area. One approach would be to use direct or design based estimators. These, however, are not available, or would have unacceptably large standard errors. Schaible, Brock, Casady, and Schnack (1979) give an overview of conventional and synthetic estimators. These estimators are, however, poststratified based on area specific covariates like age structure, gender, race, and so on. Synthetic estimators additionally use survey results collected within the small area. Such additional information on cancer mortality is not available in the context of this thesis. Nandram (2000) describes that some countries like Canada attempt to design large-scale surveys to include small areas directly.

6.2 Model based small area estimation

Hierarchical Bayes and empirical Bayes estimators are well applicable for the systematic connection of small areas. Clayton and Kaldor (1987) introduce empirical Bayes estimates, Ghosh and Rao (1994) and Nandram (2000) give an overview of current Bayesian and non-Bayesian approaches. An attractive approach in the context of disease mapping with empirical Bayes are shrinkage estimators towards the grand mean. The shrinkage is done adaptively, estimates with large sample size are shrunk less than those based on smaller samples, as proposed by Marshall (1991). The crude rates are shrunk towards a local neighborhood rate, with a given specified neighborhood structure.

Areas of application of small area estimation need not be data on geographical areas, but Hulting and Harville (1991), e.g., present an example of small are estimation in which the areas correspond to batches of raw material in an industrial application. Nandram and Petruccelli (1997) give another example of transferring small area estimation to the problem of prediction of time series, by pooling similar series. The hierarchical Poisson regression model introduced by Waller et al. (1997) includes area specific auxiliary data and models regional disease rates over space and time with space-time interactions.

We have introduced the possible use of census data to strengthen the small area estimation. For cancer mortality, however, we face the additional problem that a direct linkage between person related data and cancer mortality is impossible. Thus, the conventional approaches to improve the estimation are not directly applicable. In our case, we employ spatial and temporal marginals, the knowledge about spatial and temporal dependence structures, and further covariates to strengthen the estimation. We introduce the following notation:

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

Table 6.1: Given marginal data for the small area estimation.

D_{ti} : number of cancer cases at time t in district i

n_{ti} : population size at time t in district i (known)

r_{ti} : mortality rate at time t in district i,

where $i = 1, \dots, I$ denotes the districts within the study region, and $t = 1, \dots, T$ represents the time points (years) for the analysis. The aim of the small area estimation is to obtain cancer death estimates $\hat{D}_{ti}$ for the unknown random variable D_{ti} using the given marginals $D_{.i}$ and $D_{t.}$ where a dot indicates summation over the dotted index. The underlying population size n_{ti} with the highest spatial and temporal resolution is known and available for the analysis. Table 6.1 shows the given marginals and it is the aim to fill up the table and obtain data with the highest spatial and temporal resolution.

6.2.1 Proportional partitioning

An intuitive way of partitioning the $D_{t.}$ into $\{\hat{D}_{ti}, i = 1, \dots, I\}$ consists of splitting the given sum $D_{t.}$ according to the corresponding population sizes $\{n_{ti}, i = 1, \dots, I\}$. The number of cancer mortality cases is weighted by the underlying population size of the district. In that sense we can speak of a proportional partition for every year $t = 1, \dots, T$. The overall estimated rate

$$\hat{p}_t = D_{t.}/n_{t.} \qquad t = 1, \dots, T$$

is used for the area specific estimation of cancer death cases. It is natural to define

$$\hat{D}_{ti} = n_{ti}\,\hat{p}_t \qquad t = 1, \ldots, T,\ i = 1, \ldots, I.$$

The binomial model can be used to calculate confidence intervals for the estimated number of cancer deaths by standard methods. We assume that the unknown number of cancer death cases D_{ti} is distributed according to $\mathrm{Bin}(n_{ti}, p_t)$. In that case

$$\begin{aligned} \mathrm{Var}(\hat{D}_{ti}) &= \mathrm{Var}(n_{ti}\,\hat{p}_t) \\ &= n_{ti}^2\,\mathrm{Var}(\hat{p}_t) \\ &= n_{ti}^2\,\frac{\hat{p}_t(1-\hat{p}_t)}{n_{t.}}, \end{aligned}$$

which results in a confidence interval for D_{ti}

$$[\hat{D}_{ti} \pm u_{1-\alpha/2}\,n_{ti}\sqrt{\frac{\hat{p}_t(1-\hat{p}_t)}{n_{t.}}}]. \tag{6.1}$$

The proportional partitioning, however, based on a binomial model has some disadvantages. The data in different regions are considered to be independent, which is obviously not true in our example. In the chapters 4 and 5 we have discussed the complex spatial and temporal dependencies within the data. Thus, the resulting confidence intervals usually are anti-conservative and therefore misleading. Furthermore, the proportional partitioning can only be performed individually for each year $t = 1, \ldots, T$. Even if we could implement the spatial dependence into the model, the problem of temporal autocorrelation would remain unsolved. For the case that covariates are to be included in order to strengthen the small area estimation, the weight of the single covariates must be specified prior to the estimation. Without doing so the estimation is not straightforward in the proportional case. We aim to avoid problems like these by a Bayesian model formulation, which is capable of incorporating the spatial as well as the temporal correlations, and which allows for posterior quantification of the covariate effects.

6.2.2 Allocation with space-time dependence

In this section we describe model based approaches for small area estimation, that include the space-time dependence structure within the data. Again, we consider table 6.1, with its given spatial and temporal marginals. We begin with a Poisson modelling approach of the D_{ti} as described by Schach (2002c).

We assume space-time specific mortality rates r_{ti} as in chapter 5, and include the underlying population n_{ti} with the highest spatial and temporal resolution. The Poisson parameters λ_{ti} can be specified as follows

$$\lambda_{ti} = n_{ti}\, r_{ti} \qquad t = 1, \ldots, T,\ i = 1, \ldots, I, \tag{6.2}$$

where r_{ti} represents the rate at time t in district i. We have learned in the previous chapter that the multivariate time series can be modelled via a multivariate vector-autoregressive process including a spatial dependence structure with time lag 1. If we aim to incorporate this structure into the small area model, we again need to differentiate between the time $t = 1$ and the remaining time points $t > 1$. We set up the functional relation for the rates r_{ti} at site i at time t as follows

$$\begin{aligned} t = 1\text{:} \quad & r_{1i} = \mu_1 + b_{1i} \qquad i = 1, \ldots, I \\ t > 1\text{:} \quad & r_{ti} = \mu_t + \alpha\,(r_{t-1,i} - \mu_{t-1}) + \beta\,(\bar{r}_{t-1,\{-i\}} - \mu_{t-1}) \qquad i = 1, \ldots, I. \end{aligned}$$

At time $t = 1$ we begin with modelling the rate r_{1i} dependent on the overall level μ_1 plus the spatial random effects vector b_{1i}. This vector is used to model the spatial heterogeneity of the mortality rates at time $t = 1$. For $t > 1$ we imply the vector autoregressive structure of the space-time model of section 5. Thus, the rates r_{ti} for $t > 1$ are modelled based on three additive components. One is again the overall level μ_t, the others are the components of temporal and spatial autoregression. For the temporal autoregression we compare the rate in district i at time $t-1$ ($r_{t-1,i}$) with the overall level at time $t-1$ (μ_{t-1}). For the spatial autoregression, the neighboring sites of i at time $t-1$ ($\bar{r}_{t-1,\{-i\}}$) are compared to the overall level at time $t-1$.

An alternative approach that rather aims for the estimation of a spatial and temporal trend surface than using autoregressive structures can be specified as an overall mortality rate for every year, estimated through $\hat{p}_t = D_{t.}/n_{t.}$, assuming that this overall rate $\hat{p}_t$ needs to be adjusted for time specific and area specific random effects, and even interactions. This leads to the following relationship

$$\hat{p}_{ti} = \hat{p}_t \exp\{\alpha_t + \beta_i + \delta_{ti}\} \qquad t = 1, \ldots, T,\ i = 1, \ldots, I.$$

If we assume a Poisson model, the Poisson parameters again are $\lambda_{ti} = p_{ti}\, n_{ti}$. In general $p_{ti} = p_{ti}(\theta)$, with $\theta = (\alpha_t, \beta_i, \delta_{ti})$ for this case.

Coming back to our autoregressive allocation model, the number of known data points (which are the row and column sums) is rather small compared to the number of missing data (entries of the table). With an increasing number of rows and columns (i.e. years and districts) the unknown entries of the table increase proportionately. In the Bayesian framework, it is easy to treat the missing data D_{ti} as unknown parameters. These parameters can be estimated given the spatial and temporal marginals. As the model is highly parameterized frequentist solutions are difficult or even impossible to apply. We have introduced proportional partitioning and a binomial model for the construction of confidence intervals. However, the complex dependence structure within the data cannot be accounted for. Other frequentist approaches can be thought of, but restrictive assumptions need to be made. Therefore, we concentrate on Bayesian approaches throughout the rest of this chapter.

Combinatorial approach

Apart from the choice of a suitable model for the D_{ti} along with a link function, we need to find a way of MCMC sampling given the marginals. In the updating process of the Markov chains it is cumbersome to assign given data, i.e. the marginals, to

a sum of estimated parameters. A combinatorial approach to solving this problem would be to update the cell counts D_{ti} by choosing a 2×2 subtable and then adding

$$\begin{array}{cc} +c & -c \\ -c & +c \end{array}$$

to the counts in the subtable. Moving the subtable across the rows and columns of the contingency table necessarily preserves the marginals. Adding or subtracting c at the intersection of two rows and two columns, where rows represent years, implies a type of temporal Markov dependence.

However, a sensible choice of a proposal distribution for c goes beyond standard Bayesian estimation. Additionally, we would possibly face the problem of non-identifiability when estimating c. Apart from computational problems the interpretation of c in this context is more complex than in conventional contingency table, or 4-fold table applications. We therefore concentrate on the stochastic approach of MCMC sampling with given marginals. For further discussion of MCMC sampling preserving the marginals see McDonald, Smith, and Forster (1999).

Stochastic approach

In trying to obtain parameter estimates for the unknown cell frequencies, we have though of the following computational trick (Schach 2002b). We use an indirect adjustment of the sum of the estimated parameters towards the observed data. As the small area estimation is two-dimensional, in the direction of time and space, the adjustment is also two-dimensional. We assign the following highly informative Gaussian prior distributions to the observed marginal numbers of cancer deaths

$$\begin{aligned} D_{t.} &\sim \text{Gau}(\hat{D}_{t.}, 0.001) \\ D_{.i} &\sim \text{Gau}(\hat{D}_{.i}, 0.001). \end{aligned}$$

In every updating step of the chain, the sum of the estimated number of cancer deaths is indirectly pulled towards the observed marginals. As the variances of the

prior distributions are small, the adjustment has strong influence on the parameter estimates, and thus the marginals are preserved. The estimation is invariant under the change of the order of the two adjustments.

6.3 Parameter estimation

We have set up the Poisson model for the spatio-temporal small area estimation in section 6.2.2. In order to perform the parameter estimation of the missing cell frequencies, along with the estimation of spatial and temporal autocorrelation parameters, trend, and dispersion, we need suitable prior distributions. We begin with the vector of spatial random effects, b_0, at time $t = 1$ which we assign a non-informative multivariate Gaussian prior

$$b_0 \sim \text{Gau}_I(\gamma, V),$$

where γ and V are hyperparameters. Instead of assigning a conditional autoregressive structure to the b_0 directly, we expect the spatial dependence between neighboring sites at time $t = 1$ to arise through the model assumptions and the data. γ is assigned a non-informative Gaussian prior, $\gamma \sim \text{Gau}_I(0, U)$. U represents the unity matrix of dimension $I \times I$. The covariance matrix V is assigned a Wishart(U, I) distribution. The model specification of 6.2 for the rates includes the restricted vector of spatial random effects b_{1i}, as proposed by Besag and Kooperberg (1995). Again, the sum to zero constraint on b_0 assures identifiability at time $t = 1$.

For the vector autoregressive development of the rates we use the following prior for the overall levels μ_t, dependent on t

$$\mu_t \sim \text{Gau}(\bar{\mu}_t, 0.001), \qquad \bar{\mu}_t = D_{t.}/n_{t.}.$$

The Gaussian prior for μ_t has the overall rate for each year t as expectation, with a small variance. The temporal and spatial autoregression parameters, however, are

assigned non-informative Gaussian priors centered around zero:

$$\alpha \sim \text{Gau}(0, 1000)$$
$$\beta \sim \text{Gau}(0, 1000).$$

Figure 6.1 displays the graphical representation of our allocation model. Full arrows indicate stochastic links to which a probability distribution is attached. Double arrows denote deterministic relationships, i.e. logical links. Boxes contain given data, and circles represent stochastic nodes. The matrix V and the vector γ are hy-

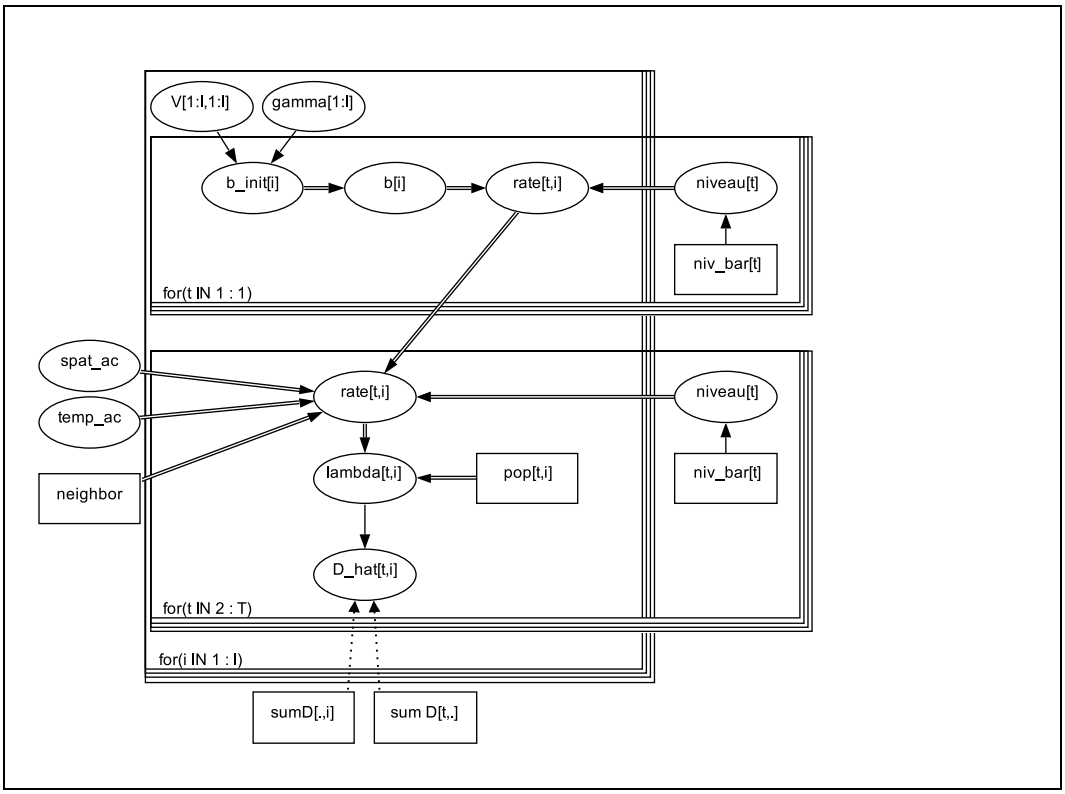

Figure 6.1: Graph of the small area model.

perparamters for the vector of spatially structured random effects (b.init$_i$) at time $t = 1$. After a sum-to-zero constraint on b.init$_i$ we obtain b_i. The overall level (niveau) plus the spatial random effects determine the rate at time $t = 1$. For all subsequent time points the rate depends on the spatial and temporal autocorrelation

parameters (temp_ac and spat_ac), and the overall level. The estimated number of cancer deaths in location i at time t follows a Poisson distribution with parameter λ_{ti}, which is determined by the rate r_{ti} and the underlying population. The dashed arrows at the bottom of the graph indicate the indirect adjustment of the estimated number of cancer deaths through the spatial and temporal marginals based on the stochastic approach.

6.4 Application of allocation model

We begin with the application of the small area Poisson model to stomach cancer among men in the region of Braunschweig, which is located in Lower Saxony. The region consists of 11 districts, thus $i = 1, \ldots, 11$. We consider the aggregated 5-year ($t = 1, \ldots, 5$) period from 1986 to 1990. Due to the complexity of the model, the full conditionals for the parameters cannot be set up and the Gibbs sampler is not applicable. We use a Metropolis-within-Gibbs algorithm for the updating of the Markov chains. We have chosen the WinBUGS default burn-in of 4,000 iterations. The parameter estimation is based on 400 recorded updates, keeping every 20th iteration.

Figure 6.2 shows the acceptance rates of the sampler based on the proposal distribution chosen within the WinBUGS software. Displayed are the successive minimum, maximum and average acceptance rates over 100 iterations each. The horizontal axis shows the number of random quantities generated. Apparently, the rates lie well within the recommended range. To check for low autocorrelation of the parameters, we have chosen the level in 1990 as an example and figure 6.3 shows that the chain is mixing well.

Apart from the application of the space-time allocation model to stomach cancer among men in the region of Braunschweig, we apply the model to the estimation

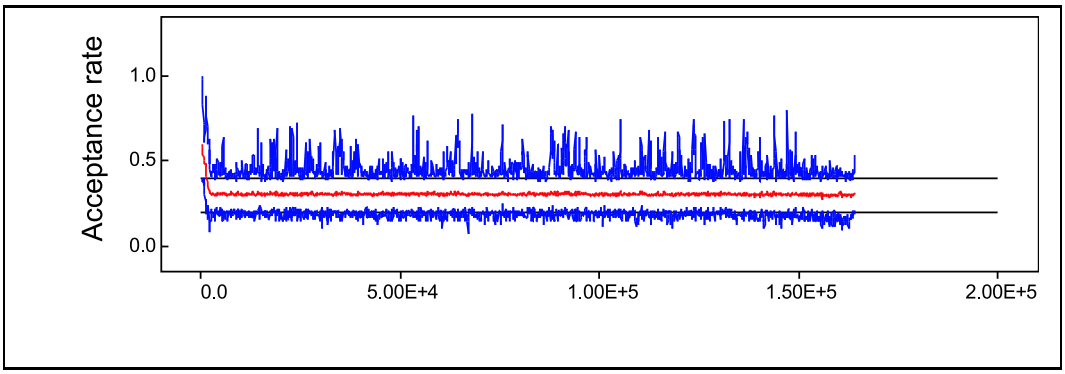

Figure 6.2: Metropolis-within-Gibbs acceptance rates for stomach cancer among men in RB Braunschweig.

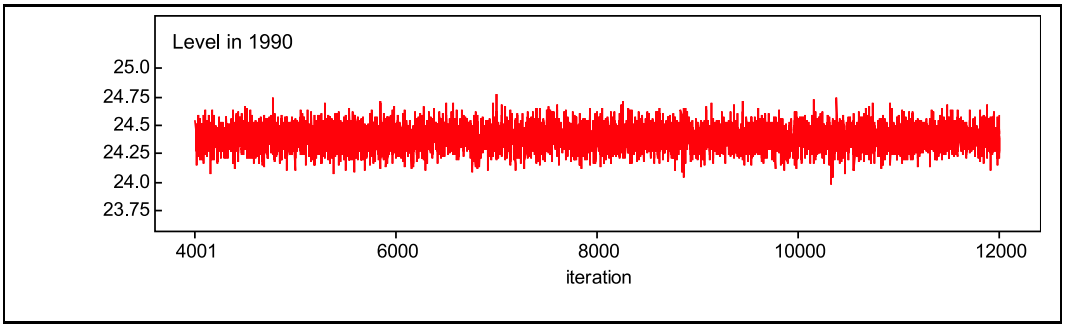

Figure 6.3: Trace plot for the level μ_5 for stomach cancer among men in RB Braunschweig in 1990.

of lung cancer mortality among women in North Rhine Westphalia. Data with the highest spatial and temporal resolution are available for this state, and therefore an evaluation of the fit is easier. We have chosen the region of Arnsberg, which includes 12 districts. Three 5-year periods from 1981 to 1995 have been used to consider the whole length of the time series over 15 years. The Metropolis-within-Gibbs acceptance rates for the allocation model in Arnsberg are displayed in figure 6.4.

Trace plots of the unknown parameters have been used to check for convergence

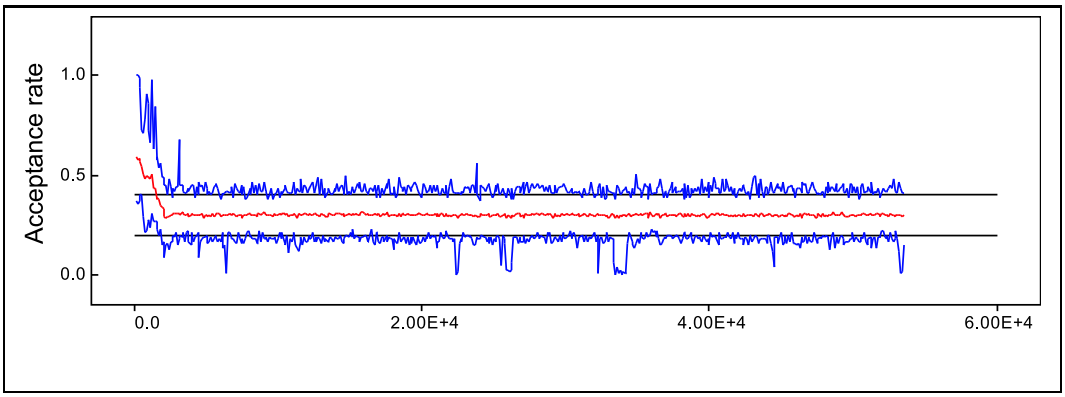

Figure 6.4: Metropolis-within-Gibbs acceptance rates for lung cancer among women in RB Anrnsberg.

and for mixing behavior of the chains, and the results have been satisfactory. We have modified the hyperpriors of the dispersion parameter for the spatially structured random effects. Whereas in a typical Markov random field prior specification the amount of spatial smoothing is usually strongly dependent on the choice of the dispersion hyperprior, in our space-time allocation model the multivariate distribution of the spatial random effects enters only at time $t = 1$. Therefore changes in the dispersion hyperprior in this case have only a minor influence on the estimated numbers of cancer deaths $\hat{D}_{ti}$.

6.4.1 Results

Firstly, we display the estimated stomach cancer mortality rates among men in the region of Braunschweig. Figure 6.5 shows the estimated parameters for the years 1986 to 1990.

In general, the explorative data analysis of male stomach cancer has shown a clearly decreasing trend. The parameter estimates for the overall levels $\mu_1, \ldots, \mu_5$, as well as the estimated rates from 1986 to 1990 indicate this.

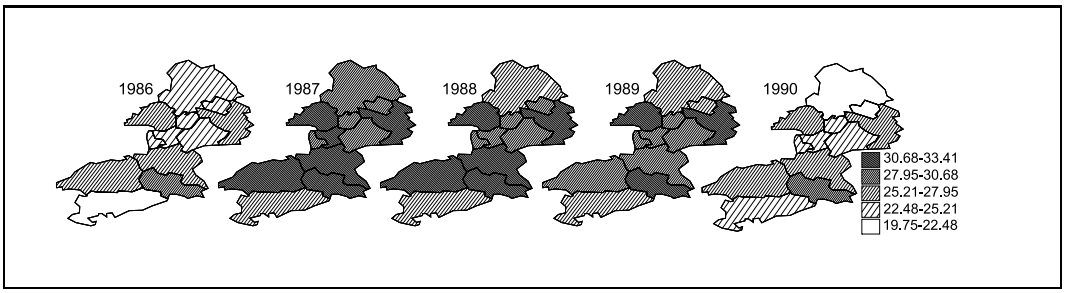

Figure 6.5: Stomach cancer among men. Spatial and temporal resolution of the rates for RB Braunschweig for the years 1986-1990.

On the other hand, we have applied the small area estimation to lung cancer among women in North Rhine Westphalia. The resulting parameter estimates for the region of Arnsberg are displayed in figure 6.6. The credible regions are based on a posterior linkage of the three 5-year periods. The dots represent the estimated raw cancer mortality rates per 100,000 inhabitants at risk, with 99 percent credible intervals for all 12 districts within Arnsberg. The results show the clear increasing temporal trends.

In this section we have presented a model based small area estimation, which guarantees a preservation of the given spatial and temporal marginals. However, as figure 6.6 shows, there is a tendency of the presented method to produce "laced" credible regions after each 5-year period. This is due to the fact, that the estimation is performed separately for each 5-year period. A subsequent small area estimation of all 3 periods is not possible, as the number of unknown parameters with the present version of the software is too large when considering 12 districts. In the following we demonstrate the estimation for two 5-year periods, but based on a reduced number of spatial units. We have divided the region of Arnsberg into 4 spatial units, with 3 districts per unit as follows:

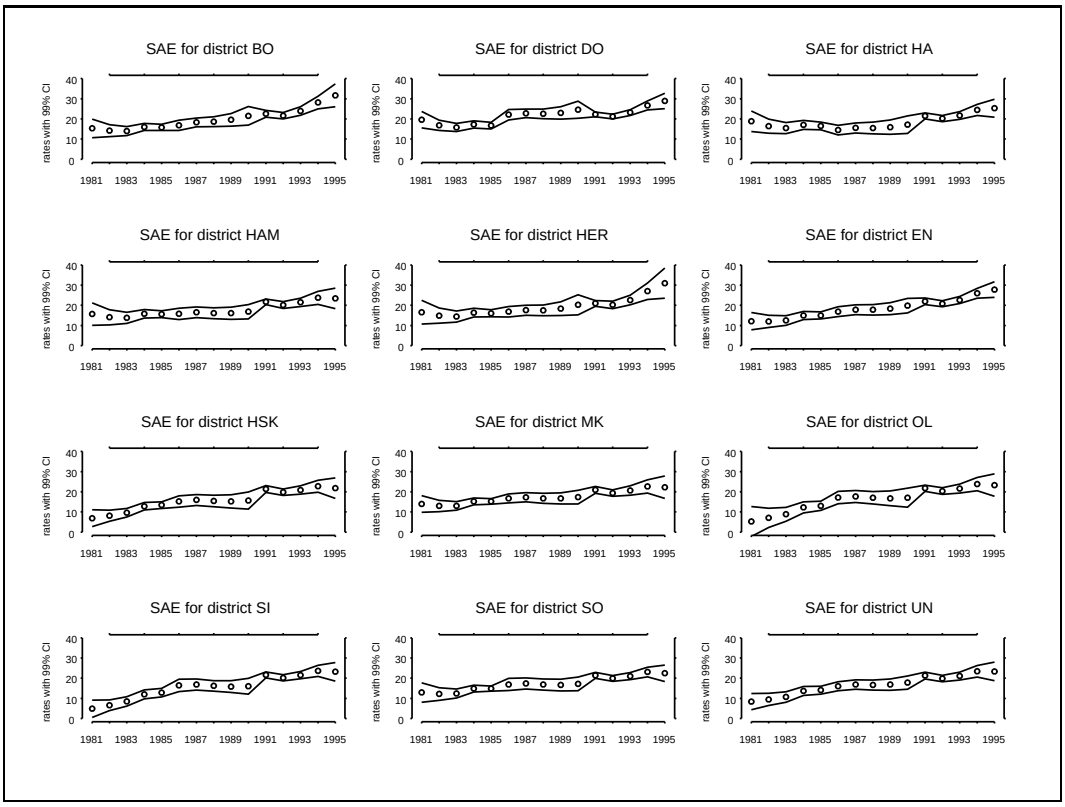

Figure 6.6: Lung cancer among women in NRW. Spatial and temporal resolution of the smoothed rates (SAE) for RB Arnsberg 1981-1995.

area 1 : Bochum, Dortmund, Herne,

area 2 : Hamm, Soest, Unna,

area 3 : Hagen, Ennepe, Märkischer Kreis,

area 4 : Hochsauerlandkreis, Olpe, Siegen.

The four new units have been constructed to form coherent regions, with similar levels of urbanization. We have applied the small area allocation model, and we have obtained the parameter estimates with a burn-in phase of length 4,000, and additional 400 recorded updates with a thinning of 20. Figure 6.7 shows the estimated posterior means with 95 % credible regions of the levels for each year, fitted with a lowess curve. The autocorrelation parameters with lower and upper 95 % credible bound have been estimated as $\hat{\alpha} = 1.058$ (1.00,1.11), and $\hat{\beta} = 0.066$ (-0.36,0.38).

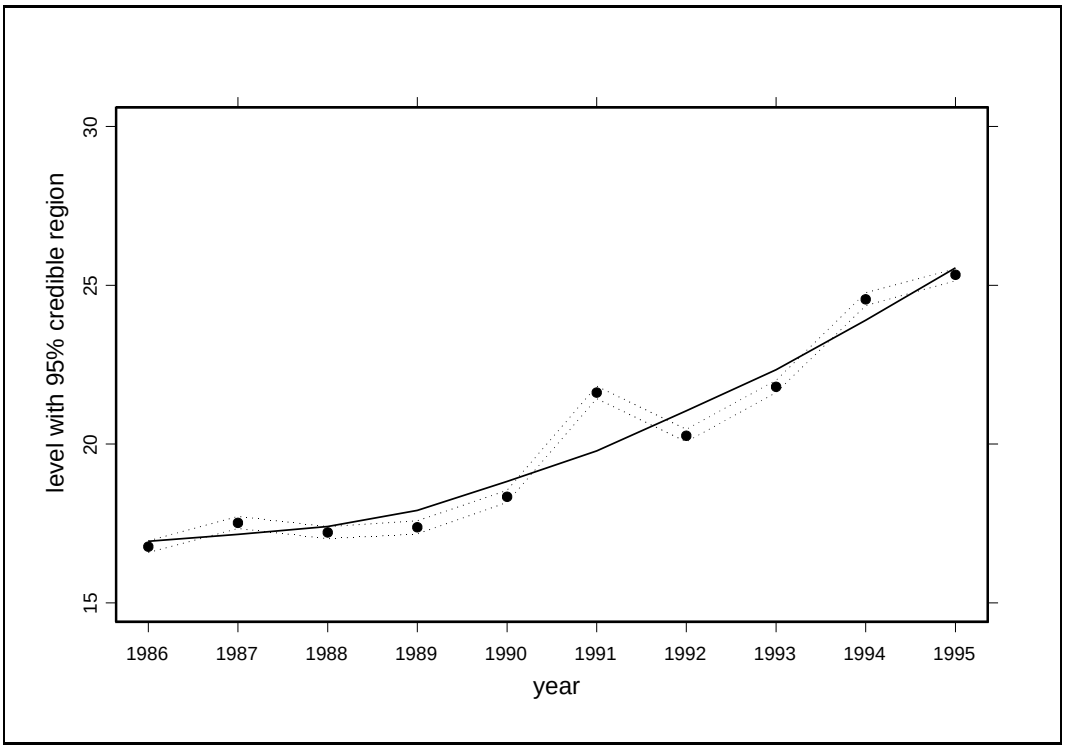

Figure 6.7: Lung cancer among women in NRW. Estimated levels with 95 % credible regions for RB Arnsberg 1986-1995.

Finally, it remains to mention that we have expected the estimated rates to be strongly correlated over space and time which can result in poor mixing of the chains as discussed in section 3.3.2. Therefore, we find it necessary to demonstrate that suitable thinning and long enough runs of the chains have lead to satisfactory mixing. Figure 6.8 shows the autocorrelations in area 1 for the years 1986, 1989, 1993, and 1995.

Furthermore, we have described the possibility of including covariates when using the Bayesian allocation model. Hence, we again take the area specific data on population density as a surrogate for smoking behavior among women to improve

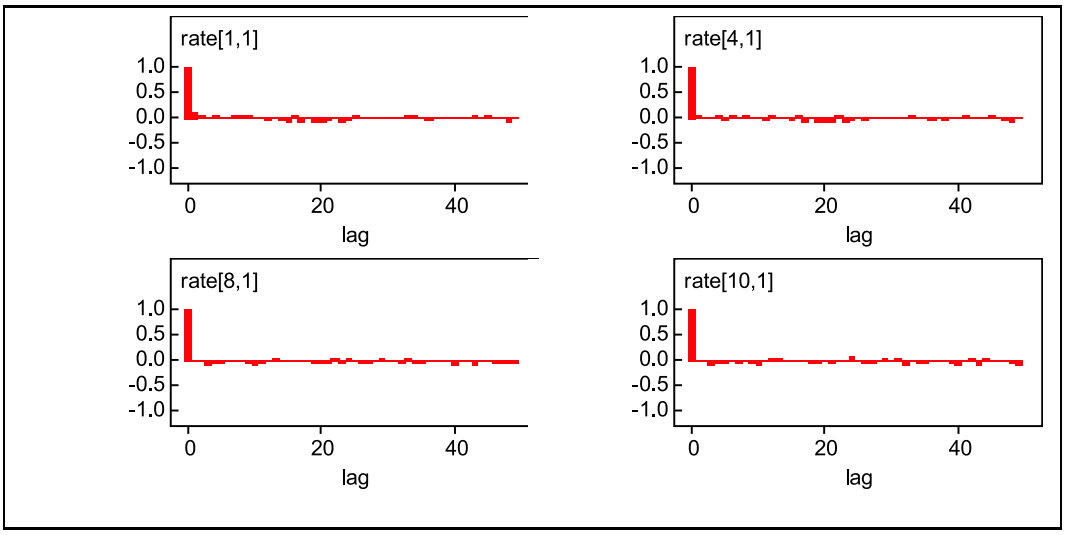

Figure 6.8: Lung cancer among women in NRW. Autocorrelations for the rate in area 1 for the years 1986, 1989, 1993, and 1995

the fit. We generalize the functional relation for the rates of the allocation model to

$$t = 1: \quad r_{1i} = \mu_1 + b_{1i} + \gamma\, X_i \qquad i = 1, \ldots, I$$

$$t > 1: \quad r_{ti} = \mu_t + \alpha\,(r_{t-1,i} - \mu_{t-1}) + \beta\,(\bar{r}_{t-1,\{-i\}} - \mu_{t-1}) + \gamma\, X_i \qquad i = 1, \ldots, I.$$

We have assigned a non-informative Gaussian prior distribution to the parameter γ. The parameters of spatial and temporal autocorrelation, and the covariate effect with 95 % credible regions have been estimated as follows: $\hat{\alpha} = 1.083$ (1.04,1.13), $\hat{\beta} = -0.170$ (-0.67,0.27), and $\hat{\gamma} = -0.028$ (-0.09,0.002).

6.4.2 Evaluation of fit

We have introduced a proportional partitioning given the spatial and temporal marginals, and with confidence intervals based on the binomial model. Figure 6.9 shows the parameter estimates resulting from the spatio-temporal allocation model and the proportional partitioning.

As the raw mortality counts are displayed, the exceptionally large numbers for

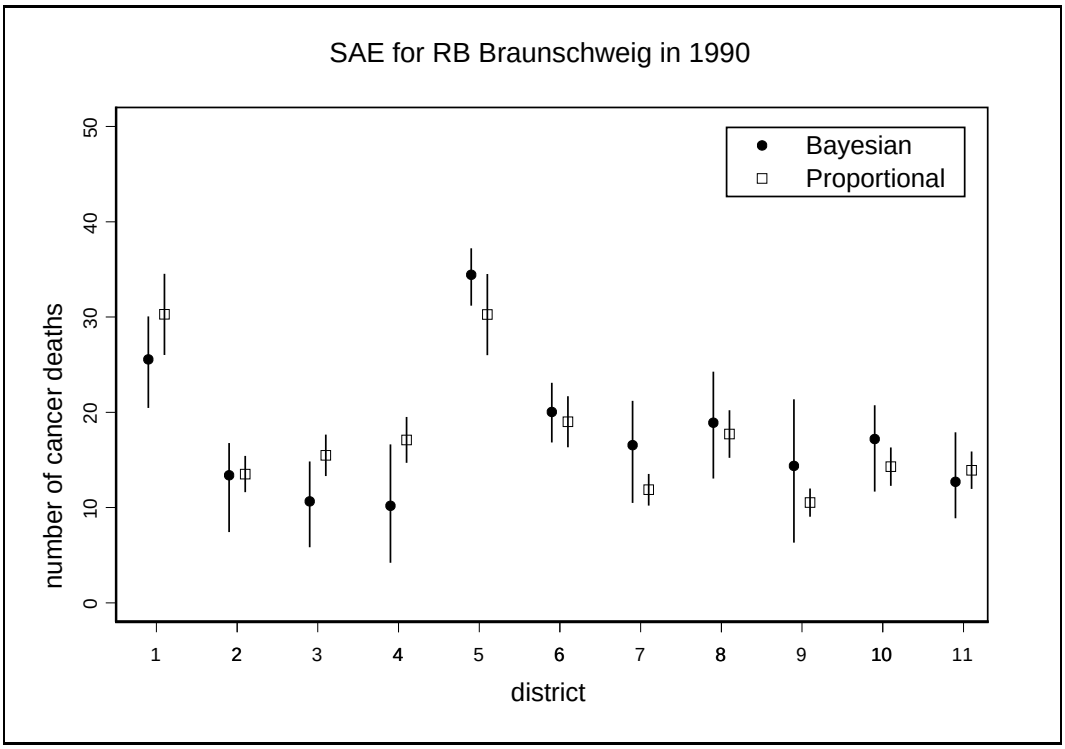

Figure 6.9: Stomach cancer among men in the region of Braunschweig. Small area estimates and 95 % supported range of the Bayesian approach and the binomial model in 1990.

the first and the fifth district are not surprising, as these are the two most densely populated urban districts Braunschweig and Göttingen. The comparison between the allocation model and the proportional partitioning reveals very similar mean estimates, but the standard deviations differ considerably. The binomial model leads to much smaller confidence intervals. However, the larger credible regions of the allocation model are more realistic. They account for the spatial and temporal dependence structures within the data, whereas the binomial small area estimation model is based on the assumption of independence.

When considering the data sets within North Rhine Westphalia the advantage is that data with the highest spatial and temporal resolution are available. Therefore

an evaluation of the fit is easy. Figure 6.10 shows the true observed rates as solid dots, and the small area estimates as circles. The 99 % credible regions are again based on separate analyses of the three 5-year periods.

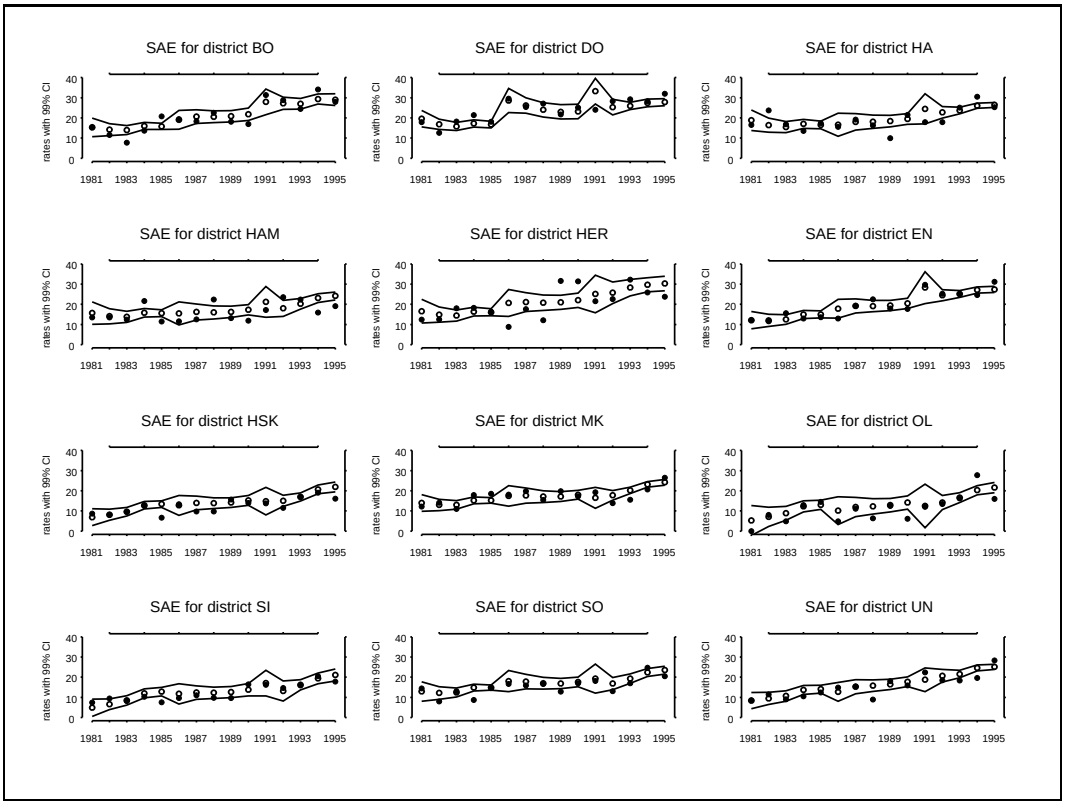

Figure 6.10: Lung cancer among women in NRW. Estimated and observed rates with 99 % credible regions for RB Arnsberg 1981-1995.

It turns out that the small area estimation is not of the same quality in all 12 districts. In districts with lower population density the fit is better than in the others. The observed rates for the districts Ennepe (EN), Märkischer Kreis (MK), Siegen (SI), and Unna (UN) lie almost completely within the 99 % credible intervals. For the other districts, however, there are more data points lying outside the supported range.

An explanation for this can be that the smoking behavior of women in highly

urbanized districts differs from that in more rural areas. Throughout this thesis we have found that the model fit could be improved by introducing a covariate for smoking behavior when analyzing lung cancer data.

For the reduced number of spatial units in the region of Arnsberg, we have been able to include the covariate proxy for smoking behavior, and in the Bayesian framework it was straightforward to estimate its effect. Figure 6.11 shows the estimated mortality rates per 100,000 people at risk, displayed as circles, over 10 years from 1986 to 1995 with estimated 95 % credible regions. The true observed rates are included as solid dots. The observed rates lie relatively well within the credible

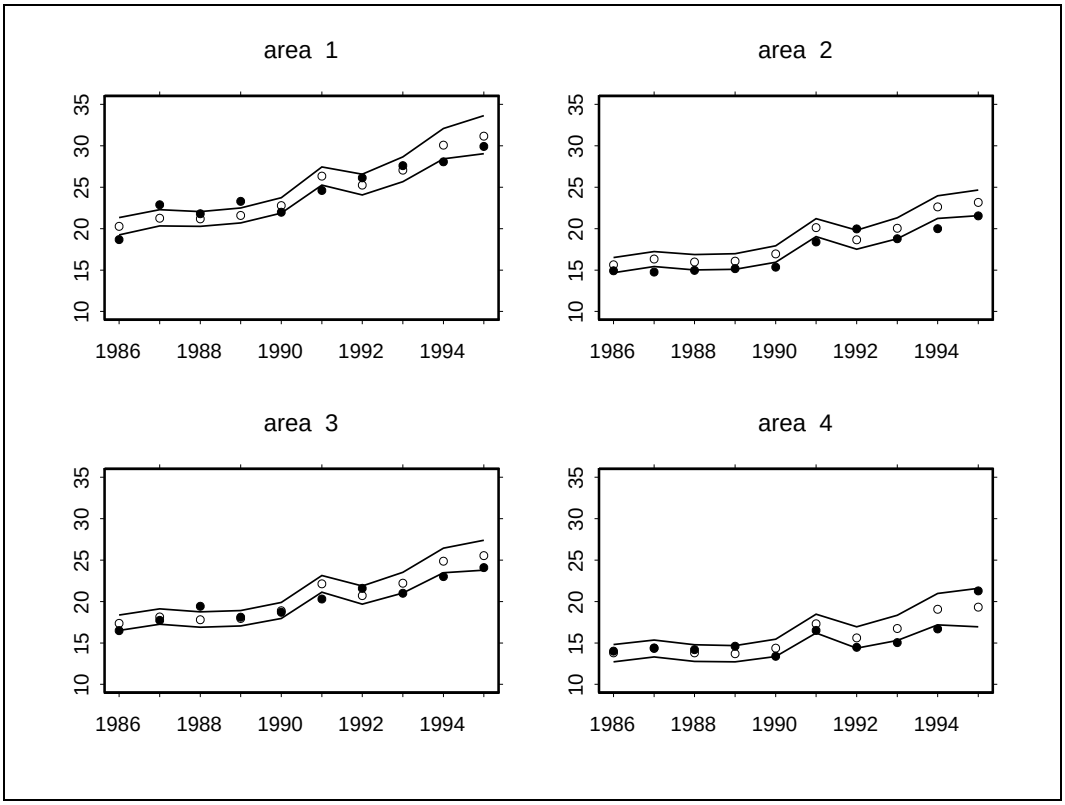

Figure 6.11: Lung cancer among women in NRW. Estimated and observed rates with 95 % credible regions for RB Arnsberg 1986-1995.

regions obtained with the model including the population density as covariate.

We have discussed the question, whether the "lacing" of the credible regions of the estimated rates can be reduced by a simultaneous estimation over more than one 5-year period. In figure 6.11 one can see clearly the "smooth" supported range at the "critical" transition from year 1990 to 1991.

6.5 Generalization of allocation model

We have introduced a model based small area estimation for the missing cell frequencies given the spatial and temporal marginals. We have included the covariate proxy urbanization for smoking behavior. However, additional to the implementation of data on smoking behavior among women, the inclusion of other covariables is straightforward. The influence of those covariates can be estimated automatically and simultaneously within the model, without the necessity of prior specification of the impact of the covariate. Compared to a classical model based partitioning, where the extension of the (proportional) partitioning in the direction of including covariates implies a prior knowledge of the covariate weights, the Bayesian modelling approach is more general. It allows the data to express the weight of one or more covariates through the MCMC simulation, which makes its superiority compared to the frequentist approach clear.

When analyzing mortality data, it is common among epidemiologists to aggregate the data over several years, or to calculate running means in order to avoid extreme variation among the data, possibly due to errors in data collection. Our allocation model for small area estimation can be applied in this kind of problem, as well. When we split up aggregated data into data with the highest resolution in space and time, the results will be "smoothed" automatically, which can lead to the fact that for some areas or years the observed rates tend to lie outside of the estimated credible regions.

Chapter 7

Discussion and Outlook

The subject of this thesis has been the analysis of dependent data. The field of application was data on lung cancer and stomach cancer mortality in different regions of Germany over time. The dependence structure within the data arises through the fact that both cancer types show distinct spatial patterns, but also are dependent over time. We have applied separate time series analyses for the areal data, and we have analyzed temporal slices of the data with Markov random field modelling approaches to learn about the spatial dependence structure. The knowledge gained from separate analyses has been used to build a spatio-temporal model on the basis of vector-autoregressive process theory of spatially dependent time series. The parameters of the space-time model have been estimated with Bayesian and with likelihood methods. Different cancer types, and different regions of study have been analyzed. Models with and without covariates have been fitted and compared on the basis of the deviance information criterion. We have discussed the computational difficulties of the likelihood approach, and the elegance of the evaluation of the posterior quantities in the Bayesian approach with MCMC methods.

Finally, we have transferred the spatial and temporal autoregressive dependence structures to a small area estimation problem. Due to the fact that cancer mor-

tality data with the highest spatial and the highest temporal resolution are not published, we have applied a model based small area estimation. Instead of solely using covariates to base the partition on, we have additionally implied knowledge of spatial and temporal dependence within the data. We have modelled the missing data as parameters, and we have estimated data with the highest spatial and temporal resolution given the spatial and temporal marginals. We have compared the resulting estimates with a proportional partitioning according to a binomial modelling approach. The comparison reveals larger and more reliable credible regions of the allocation with space-time dependence. Finally, we have used data on lung cancer with the highest spatial and temporal resolution for the state of North Rhine Westphalia for a comparison with the estimated data according to the allocation model. In the case of lung cancer among women it turns out that for the rural areas the estimation is better than for the urban districts. We believe this to be due to different behavioral patterns among women in rural compared to urban areas.

In chapter 4 and chapter 5 we have worked with first order vector-autoregressive processes. Further research could be undertaken in the direction of higher order autoregressive structures. For the estimation of the temporal trend in chapter 5, we have assumed linearity. Alternatively, non-parametric approaches like first or second order random walks can be applied to improve the fit of the underlying temporal trend. The small area estimation in this application has been used to break down spatial aggregates of regions into districts, and temporally aggregated data to yearly figures. Without considering computational feasibility problems, the allocation with our model can be used to break down spatial data to a finer resolution than districts. In Germany, for example, this could be postal area codes. A new law on notifiable infectious diseases has come into force in North Rhine Westphalia in 2001. From then on incidence cases must be notified on 3-digit area code level. Further research could be undertaken in the direction of spatial dynamic modelling or disease surveillance of these infectious diseases. In the direction of the temporal component, the small

area estimation model could be extended to the case of breaking down yearly data into monthly data, with additional seasonal effects. For a more dynamic modelling, it might be worth thinking of a simultaneous spatio-temporal structure.

Appendix A

Probability Distributions

A.1 Discrete distributions

A.1.1 Binomial distribution

The binomial distribution, $\text{Bin}(n, p)$ is used to represent the number of "successes" Z in a sequence of n iid. Bernoulli trials, with probability of success p in each trial. For $0 \leq p \leq 1$ the probabilities for Z are given by

$$f(z \mid p) = \binom{n}{z} p^z \, (1-p)^{n-z} \, I\!I_{\{0,\ldots,n\}}(z).$$

The expectation of Z equals np and the variance is $np\,(1-p)$. The usage in the WinBUGS language is `z ~ dbin(p,n)`.

A.1.2 Poisson distribution

The Poisson distribution, $\text{Poi}(\lambda)$ is used to represent count data. According to Gelman et al. (1995) simulation for the Poisson distribution can be somewhat cumbersome, as it is a problem to invert the cumulative distribution function. For

$\lambda > 0$ the Poisson distribution can be written as

$$f(z \mid \lambda) = e^{-\lambda} \frac{\lambda^z}{z!} I\!\!I_{I\!N}(z).$$

For this discrete distribution the expectation and the variance are λ. The WinBUGS code is `z` $\sim$ `dpois(`λ`)`.

A.2 Continuous distributions

A.2.1 Multivariate Gaussian distribution

Consider $\mu \in I\!R^p$ and Σ a $(p \times p)$ symmetric positive-definite matrix. Then the parameterization of the normal distribution, $\text{Gau}_p(\mu, \Sigma)$ results in a density

$$f(z \mid \mu, \Sigma) = (2\pi)^{-p/2}(\mid \Sigma \mid)^{-1/2}\exp\{-(z-\mu)' \Sigma^{-1} (z-\mu)/2\}.$$

The vector of expectations is $\text{E}(Z) = \mu$ and the covariance matrix $\text{E}[(Z-\mu)(Z-\mu)'] = \Sigma$. For $p = 1$ the log-normal distribution is defined as the distribution of e^Z with $Z \sim \text{Gau}(\mu, \sigma^2)$, see Robert (1994), p. 381. The WinBUGS language uses the function `z[]` $\sim$ `dmnorm(`μ`[],`Σ^{-1}`[])`.

A.2.2 Gamma distribution

The gamma distribution, $\text{Gamma}(\alpha, \beta)$ is the conjugate prior distribution for the inverse of the normal variance and for the mean parameter of the Poisson distribution. A non-informative distribution is obtained, as $\alpha \to 0$ and $\beta \to 0$. With positive parameters α and β the gamma density can be expressed as follows

$$f(z \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} \exp\{-\beta\, z\}\, I\!\!I_{[0,\infty)}(z).$$

Then the expected value of the gamma distribution is $\text{E}(Z) = \alpha/\beta$ and the variance $\text{var}(Z) = \alpha/\beta^2$. Special cases of the gamma distribution are the Erlang distribution,

the exponential distribution, and the χ^2-distribution. To call the gamma distribution in WinBUGS, use `z` $\sim$ `dgamma(`α, β`)`.

A.2.3 Inverse Gamma distribution

The inverse gamma distribution InvGamma(α, β) is the distribution of Z^{-1} when $Z \sim$ Gamma(α, β), and it is the conjugate prior distribution for the normal variance. A non-informative distribution is obtained in the limit as α and $\beta \to 0$. Again with positive values for α and β the inverse gamma density can be formulated as

$$f(z \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{e^{-\beta/z}}{z^{\alpha+1}} I\!\!I_{[0,\infty)}(z).$$

$\mathrm{E}(Z) = \alpha/\beta$ and $\mathrm{var}(Z) = \alpha/\beta^2$ as for the gamma distribution.

A.2.4 Wishart distribution

The Wishart distribution, Wishart$_p(\Sigma, N)$ is the conjugate prior distribution for the inverse covariance matrix in a multivariate normal distribution. It is a multivariate generalization of the χ^2_1-distribution, for the case that $p = 1$, or the gamma distribution. A random random variable Z of dimension p follows a Wishart distribution, if it is proportional to

$$f(z \mid \Sigma, N) \propto \mid z \mid^{(N-p-1)/2} \exp(-\frac{1}{2}\mathrm{tr}(\Sigma\, z))\, I\!\!I_{\mid z \mid > 0}(z).$$

The usage in the WinBUGS language is `z[]` $\sim$ `dwish(`Σ, N`)`.

A.2.5 Inverse Wishart distribution

If $Z^{-1} \sim$ Wishart$_p(\Sigma, N)$ then Z has the inverse-Wishart distribution. The inverse Wishart distribution is the conjugate prior distribution for the covariance matrix in a multivariate normal distribution.

List of Figures

Bibliography

Abraham, B. (1983). The exact likelihood function for a space time model. *Metrika 30*, 239–243.

Aroian, L. A. (1980). Time series in m-dimensions: Definition, problems and prospects. *Communications in Statistics B 9*, 453–465.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of changes. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Becker, N. (1998). Wie mißt man das Krebsrisiko? *Deutsches Ärzteblatt 95*, B1667–B1670.

Becker, N., R. Frentzel-Beyme, and G. Wagner (1984). *Atlas of cancer mortality in the Federal Republic of Germany*. Berlin, New York: Springer.

Bernardinelli, L., D. Clayton, and C. Montomoli (1995). Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine 14*, 2411–2431.

Bernardinelli, L., D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine 14*, 2433–2443.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference (disc: p128-147). *Journal of the Royal Statistical Society B 41*, 113–128.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (disc: p226-236). *Journal of the Royal Statistical Society B 36*, 192–225.

Besag, J. and P. Green (1993). Spatial statistics and Bayesian computation (disc: p53-102). *Journal of the Royal Statistical Society B 55*, 25–37.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems (disc: p41-66). *Statistical Science 10*, 3–41.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic autoregressions. *Biometrika 82*, 733–746.

Besag, J., J. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics (disc: p21-59). *Annals of the Institute of Statistical Mathematics 43*, 1–20.

Best, N. G., D. J. Spiegelhalter, A. Thomas, and C. E. G. Brayne (1996). Bayesian analysis of realistically complex models. *Journal of the Royal Statistical Society A 159*, 323–342.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Brooks, S. P. (1998). Markov chain Monte Carlo method and its application. *The Statistician 47*, 69–100.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*, 434–455.

Brown, L. D. (1986). *Foundations of exponential families*. Hayward, California: IMS Lecture Notes - Monograph Series 6.

Cargnoni, C., P. Müller, and M. West (1997). Bayesian forecasting of multinomial time series through conditionally Gaussian dynamic models. *Journal of the American Statistical Association 92*, 640–647.

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

Chen, M. H., Q. M. Shao, and J. G. Ibrahim (2000). *Monte Carlo methods in Bayesian computation*. Berlin, New York: Springer.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–335.

Clayton, D. and J. Kaldor (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics 43*, 671–681.

Cliff, A. D. and J. K. Ord (1973). *Spatial autocorrelation*. London: Pion.

Cressie, N. and N. H. Chan (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association 84*, 393–401.

Cressie, N. and H.-C. Huang (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association 94*, 1330–1340.

Cressie, N. A. C. (1993). *Statistics for spatial data (Revised edition)*. New York: Wiley.

DeGroot, M. (1986). *Probability and statistics (2nd edition)*. Reading: Addison-Wesley.

Fahrmeir, L. and A. Hamerle (Eds.) (1984). *Multivariate statistical procedures (German)*. New York: DeGruyter.

Fahrmeir, L., H. Kaufmann, and F. Ost (1981). *Stochastische Prozesse*. Munich: Hanser.

Freeman, M. and J. Tukey (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics 21*, 607–611.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398–409.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian data analysis*. London: Chapman and Hall.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences (disc: p483-501, 503-511). *Statistical Science 7*, 457–472.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEEPAMI 6*, 721–741.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (disc: p483-503). *Statistical Science 7*, 473–483.

Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal (disc: p76-93). *Statistical Science 9*, 55–76.

Gössl, C., D. P. Auer, and L. Fahrmeir (2001). Bayesian spatiotemporal inference in functional magnetic resonance imaging. *Biometrics 57*, 554–562.

Hall, P. G. and C. C. Heyde (1980). *Martingale limit theory and its applications*. New York: Academic Press.

Hammersley, J. M. and D. C. Handscomb (1964). *Monte Carlo methods*. London: Methuen.

Han, C. and B. P. Carlin (2001). Markov chain Monte Carlo methods for computing Bayes factors: a comparative review. *Journal of the American Statistical Association 96*, 1122–1132.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Horn, R. A. and C. R. Johnson (1985). *Matrix analysis*. Cambridge: University Press.

Höpker, C. (1998). Flickenteppich mit falschen Farben. *Deutsches Ärzteblatt 95*, B1650–B1652.

Hulting, F. L. and D. A. Harville (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: Computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association 86*, 557–568.

Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: University Press.

Kafadar, K. and J. W. Tukey (1993). U.S. cancer death rates: A simple adjustment for urbanization. *International Statistical Review 61*, 257–281.

Kaluzny, S. P., S. C. Vega, T. P. Cardoso, and A. A. Shelly (1998). *S+SpatialStats: User's manual for Windows and UNIX*. Berlin, New York: Springer.

Kass, R. E., B. P. Carlin, A. Gelman, and R. M. Neal (1998). Markov chain Monte Carlo in practice: A roundtable discussion. *The American Statistician 52*, 93–100.

Kelsall, J. and J. Wakefield (1999). Discussion of 'Bayesian models for spatially correlated disease and exposure data' by Best et al. In J. Bernardo, J. Berger, A. Dawid, and A. Smith (Eds.), *Bayesian Statistics*, Volume 6, Oxford. University Press.

Knorr-Held, L. (1997). *Hierarchical modelling of discrete longitudinal data*. Ph. D. thesis, University of Munich.

Knorr-Held, L. and J. Besag (1998). Modelling risk from a disease in time and space. *Statistics in Medicine 17*, 2045–2060.

Knorr-Held, L. and G. Raßer (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics 56*, 13–21.

Knorr-Held, L. and E. Rainer (2001). Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics 2*, 109–129.

Krahnke, T. (2001). *On the estimation of smooth parameter fields with applications in ophthalmology*. Ph. D. thesis, University of Dortmund.

Kreienbrock, L. and S. Schach (2000). *Epidemiologische Methoden* (3rd ed.). Stuttgart: Gustav-Fischer Verlag.

Kulkarni, V. G. (1995). *Modeling and analysis of stochastic systems*. London: Chapman and Hall.

Lang, S. and A. Brezger (2002). *BayesX version 0.6: User manual*. Department of Statistics, University of Munich.

Lavine, M. and S. Lozier (1999). A Markov random field spatio-temporal analysis of ocean temperature. *Environmental and Ecological Statistics 6*, 249–273.

Lütkepohl, H. (1991). *Introduction to multiple time series analysis*. Berlin, New York: Springer.

Marshall, R. J. (1991). Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics 40*, 283–294.

McDonald, J. W., P. W. F. Smith, and J. J. Forster (1999). Exact tests of goodness of fit of log-linear models for rates. *Biometrics 55*, 620–624.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations for the state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Mollié, A. (2000). Bayesian mapping of Hodgkin's disease in France. In P. Elliott, J. Wakefield, N. Best, and D. Briggs (Eds.), *Spatial Epidemiology*, Oxford, pp. 267–285. University Press.

Nandram, B. (2000). Bayesian generalized linear models for inference about small areas. In D. Dey, S. Ghosh, and B. Mallick (Eds.), *Generalized linear models: a Bayesian perspective*, New York, pp. 89–109. Marcel Dekker.

Nandram, B. and J. D. Petruccelli (1997). A Bayesian analysis of autoregressive time series panel data. *Journal of Business and Economic Statistics 15*, 328–334.

Neal, R. (1997). Markov chain Monte Carlo methods based on 'slicing' the density function. Technical Report 9722, Department of Statistics, University of Toronto.

Neal, R. (2002). Slice sampling. *Annals of Statistics*, to appear.

Nolte, E., U. Laaser, D. Bardehle, and R. Annuß (1997). Die gesundheitliche Lage der Bevölkerung in Nordrhein-Westfalen fünf Jahre vor der Zielstellung der WHO: Gesundheit für alle im Jahr 2000. *Bundesgesundheitsblatt 40*(9), 322–332.

Raiffa, H. and R. Schlaifer (1961). *Applied statistical decision theory.* Harvard: University Press.

Risch, H. A., M. Jain, and N. W. Choi (1985). Dietary factors and the incidence of cancer of the stomach. *American Journal of Epidemiology 122*, 947–959.

Robert, C. P. (1994). *The Bayesian choice: A decision-theoretic motivation*. Berlin, New York: Springer.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice*, London. Chapman and Hall.

Schach, U. (2000). Spatio-temporal models on the basis of innovation processes and application to cancer mortality data. Technical Report 16/2000, University of Dortmund, SFB 475.

Schach, U. (2002a). Spatio-temporal modelling of cancer mortality rates. In W. Gaul and G. Ritter (Eds.), *Classification, Automation, and New Media*, Berlin, New York, pp. 491–498. Springer.

Schach, U. (2002b). Spatio-temporal small area estimation for aggregated mortality data and application to lung cancer among women in North Rhine Westphalia. In I. Zöllner and D. Böhning (Eds.), *Disease Mapping in Health Reporting and Spatial Models for Surveillance*, Munich, pp. 63–70. SPSS Print.

Schach, U. (2002c). A type of Bayesian small area estimation for the analysis of cancer mortality data. In O. Opitz and M. Schwaiger (Eds.), *Explanatory Data Analysis in Empirical Research*, Berlin, New York, pp. 362–370. Springer.

Schaible, W., D. Brock, R. Casady, and G. Schnack (1979). Small area estimation, an empirical comparison of conventional and synthetic estimators for states. *Vital and Health Statistics II 82*, 1–19.

Schervish, M. J. and B. P. Carlin (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics 1*, 111–127.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (disc: p53-102). *Journal of the Royal Statistical Society B 55*, 3–23.

Spiegelhalter, D., N. Best, B. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, to appear.

Spiegelhalter, D., A. Thomas, and N. Best (2002). *WinBUGS version 1.3.1: User manual*. Cambridge, London: MRC Biostatistics Unit, and Department of Epidemiology and Public Health.

Tierney, L. (1994). Markov chains for exploring posterior distributions (disc: p1728-1762). *Annals of Statistics 22*, 1701–1728.

Waller, L. A., B. P. Carlin, H. Xia, and A. E. Gelfand (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association 92*, 607–617.

Wichmann, H., K. H. Jöckel, and B. Molik (1991). Luftverunreinigungen und Lungenkrebsrisiko. Umweltforschungsplan des Bundesministeriums für Umwelt, Naturschutz und Reaktorsicherheit.

Willett, W. C. and B. McMahon (1984). Diet and cancer - an overview. *New England Journal of Medicine 310*, 633–638.

Xia, H., B. P. Carlin, and L. A. Waller (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics 8*, 107–120.

Zöllner, I. (1991). *Statistische Methoden zur Analyse räumlicher Konzentrationen - Anwendung auf die Verteilung von Krebsfällen in Deutschland*. Ph. D. thesis, University of Dortmund.