Comparing models for variables given on disparate spatial scales: An epidemiological example

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund

vorgelegt von
Sibylle Sturtz

Dortmund, Juni 2007

Gutachter:

Prof. Dr. Katja Ickstadt
Prof. Dr. Claus Weihs

Tag der mündlichen Prüfung:
11. September 2007

Contents

Overview on structures and models code referred to in this thesis

Structures

+ benzene		×benzene		
low	high	low	high	
A	B	M	N	benzene only
C	D	O	P	+ latent risk covariate
E	F	Q	R	+ linear trend covariate
G	H	S	T	+ increased risk in the southern part
I	J	U	V	+ increased risk in 3 clusters
⇓	⇓	⇓	⇓	⇓
330	770	330	770	+ 330 cases

Schematic overview of the generated structures.

Models

Models with fixed locations of Gaussian kernels

Poisson–Gamma models with additive influence of benzene:
model a: no latent risk sources;
model b: 4 latent risk sources with $d_1 = 15$km;
model c: 9 latent risk sources with $d_1 = 15$km;
model d: combination of sources from b and c to 13 latent risk sources;
model e: 36 latent risk sources with $d_2 = 5$km;

Poisson–Gamma models with multiplicative influence of benzene:
model g: no latent risk sources;
model h: 4 latent risk sources with $d_1 = 15$km;
model i: 9 latent risk sources with $d_1 = 15$km;
model j: combination of sources from b and c to 13 latent risk sources;
model k: 36 latent risk sources with $d_2 = 5$km;

Poisson–Gamma model with no influence of benzene:
model w: 36 latent risk sources;
model x: 13 latent risk sources;

Other spatial models:
model y: BDCD algorithm, wards parted by river Thames are neighbours;
model z: CAR model, neighbourhood structure as used in BDCD;
model v: CAR model, wards parted by river Thames are not neighbours.

Models with random locations of Gaussian kernels

model f: Poisson–Gamma model with additive influence of benzene;
model m: Poisson–Gamma model with multiplicative influence of benzene;
model o: Poisson–Gamma model with no influence of benzene.

Chapter 1

Introduction

In spatial epidemiology interest often focuses on describing and modelling spatial variation of diseases and other spatial phenomena. The area of research can be divided into ecologic regression studies and disease mapping studies. The first group focuses on the estimation of regression coefficients in order to quantify the exposure/disease relationship, whereas the second one has the objective to estimate the spatial risk surface by highlighting areas of elevated and lowered risk. Another field of spatial models is given by cluster models which focus on determining disease etiology but provide also a popular tool in disease mapping. As the variance of the ratio between observed and expected cases, the so–called standardised mortality ratio (SMR) depends reciprocally on the number of expected cases differentiation between random variation and variation in the SMRs is difficult. Methods based on Bayesian assumptions have been used to remove sample variation. To improve prediction, measured as well as latent covariates can be included in the model. For an introduction on spatial epidemiology see for example Elliot et al. (2000).

The spatial analysis performed in this thesis was motivated by the paper by Best et al. (2001) who analyse childhood leukaemia rates in dependence on environmental benzene emissions using ecologic regression models.

Childhood leukaemia and its causes are a main research area. Compared to other diseases in economically developed parts of the world cancer in children is a rare disease. It accounts for less than 1% of new cancers each year (Wild and Kleinjans, 2003) and has an incidence rate of about 4 in 100 000 per year (Little, 1999). Nevertheless, cancer follows accidents as the second most common entry in cause of death

statistics for children. Among cancers, leukaemia is the most frequent one. Various causes for leukaemia are discussed. Dockerty et al. (2001) investigate the effects of parental age, parity (the total number of previous children live born and stillborn to the mother) and socioeconomic level on childhood cancer in a case–control study involving more than 10 000 matched pairs of children. Other risk factors are the exposure to high doses of ionising radiation, trisomy 21, certain rare diseases (Fanconis anaemia, ataxia-telangiectasia, type 1 neurofibromatosis), and certain chemotherapies (Steffen et al., 2004). Dickinson et al. (2003) analyse the proximity of railway lines to the household as an alternative risk factor for childhood leukaemia but found no significant association. UK Childhood Cancer Study Investigators (2000b) perform a case–control study involving 3 838 children with cancer and 7 629 unaffected children living in England, Scotland, and Wales in the period 1991–1998 to evaluate possible causes of childhood cancer. Among other results, they have found no association between higher radon concentrations and risk of any of the childhood cancers or the residential proximity to power lines (UK Childhood Cancer Study Investigators, 2000a). Another report on the development of childhood leukaemia in Great Britain from 1969 to 1993 is published by the Committee on Medical Aspects of Radiation in the Environment (COMARE) (2006). The authors show evidence for spatial clustering in Britain, but no evidence for clustering around nuclear installations in general, although the village next to the Sellafield power plant showed an excess of cases.

In this thesis, we model childhood leukaemia data previously analysed by Best et al. (2001). At the level of electoral wards leukaemia cases of children under 15 years old are given as well as corresponding population counts. These are related to environmental benzene exposure modelled on a 1 km $\times$ 1 km grid. As an alternative covariate we use a deprivation index which is given on ward level. The chosen index is the one of Carstairs (1995). The data set is presented in detail in Chapter 2.

When analysing the observed leukaemia cases in relation to benzene emissions, we have to cope with different spatial resolutions. The usual approach to deal with such data is to aggregate data and covariates to a common spatial scale. The frequently used Markov random field (MRF) ecologic regression model is one example for such models that have to deal with the problem of the ecological fallacy (Richardson, 1992). This term reflects the idea that group–level exposure–response relationship may not reflect individual–level relationship. As we usually

assume a Uniform distribution of risk within aggregated areas the problem of ecological fallacy tends to be larger for higher aggregations. Nevertheless, it the MRF ecologic regression model is almost exclusively used.

A random field generalisation of Poisson–Gamma hierarchical models, introduced by Wolpert and Ickstadt (1998) and generalised by Best et al. (2000) for an application in epidemiology, provides a more suitable modelling framework. Here, we are able to model data and covariates on their original spatial scales. A further advantage is the possibility to model covariates either as excess or relative risk factors (Breslow and Day, 1980) leading to different interpretations. The additive influence of excess risk factors allows alternative explanations of an event and is preferable for competing, non–interacting effects. The multiplicative influence of relative risk factors reflects different individual susceptibilities to a covariate. Latent covariates can be introduced to improve risk estimates not associated with considered covariates.

Commonly used log–linear models allow only for multiplicative modelling which may not reflect the true influence of each covariate. The class of Poisson–Gamma random field models provides a flexible tool which does not rely on multiplicative modelling of covariates and latent risk. In this thesis, we focus on investigation of the impact of a relative risk factor modelled as an excess risk factor and vice versa as well as the performance of Poisson–Gamma models in general using simulated data sets as well as the one by Best et al. (2001). This includes the effect of different settings of the latent spatial structure. Different spatial resolutions of benzene and leukaemia rates are neglected for computational reasons. Nevertheless, modelling of the latent field is on a different scale.

Spatial analyses have to deal with many sources of complexity. Therefore in modelling such data, we do not directly specify parameters for all sources of variability. Alternatively, a Bayesian hierarchical framework allows to model uncertainty in estimation of model parameters in a flexible and hierarchical way. An overview on Bayesian modelling is given in Chapter 3. This includes a brief introduction to Markov chain Monte Carlo (MCMC) methods. A description of the models applied throughout this thesis builds the main part of the third Chapter. It includes the class of Poisson–Gamma random field models, their implementation in WinBUGS, as well as alternatively considered models. These are a clustering model by Knorr-Held and Raßer (2000), the so–called Bayesian Detection of Clusters and Discontinuities (BDCD), and an ecologic regression model based on

Markov random fields (MRF) similar to the one used in Best et al. (2001).

One focus of this thesis is to implement Poisson–Gamma models in the Bayesian software programme WinBUGS (Spiegelhalter et al., 2004). This software is very popular for Bayesian applications. Selected spatial models such as the widely used MRF ecologic regression models are already implemented and ready to use, which is one reason for its popularity. By implementation of Poison–Gamma models in the software and analysing the according performance of this model class, which is the main focus of this thesis we hope to provide an alternative easy accessible modelling framework.

The main disadvantage of the software is it's non–automatised state where constant interaction between user and software is required. To improve the application of WinBUGS, the development of automatisation software is necessary. This is provided by the packages R2WinBUGS (Sturtz et al., 2005) and BRugs (Thomas et al., 2006) which allow to use WinBUGS from the statistical software R (R Development Core Team, 2006). An introduction to these packages is given in Chapter 4.

When working with MCMC methods, convergence of Markov chains needs to be assessed. The criterion of Brooks, Gelman and Rubin (Gelman and Rubin, 1992, Brooks and Gelman, 1998) is suitable to validate convergence of chains towards the posterior distribution. Once the model is fitted appropriately its fit has to be compared to alternative models. The Deviance Information Criterion (Spiegelhalter et al., 2002, DIC) is such a criterion. We employ the DIC to select the most appropriate model for the data. In Chapter 5 we discuss both, convergence diagnostics and DIC.

To analyse the performance of the proposed models and its performance within WinBUGS, we generate different spatial structures in a simulation study, the design of the study and the structures is described in Chapter 6.

For selected basic structures we use a restricted WinBUGS' implementation of Poisson–Gamma random field models. Here we limit the models' ability to estimate latent risk in order to get an idea if such an intermediate model between the discrete and the continuous version of the class of Poisson–Gamma models already gives a sufficient modelling framework. The corresponding results are presented in Chapter 7.

The simulations reveal the necessity of model refinement. We therefore extend our model and apply it on all generated structures structures described in Chapter 6. The collection of generated structures allows for a comprehensive judgement and comparison of models' performances which we present in Chapter 8.

In Chapter 9, we apply Poisson–Gamma random field models as well as the MRF models and the BDCD algorithm on leukaemia cases observed in Inner London. Used covariates include benzene emissions and the Carstairs index. We use both as either excess or relative risk factor and estimate the spatial risk surface. To improve models' performance we include a sufficient number of latent covariates.

Finally, we summarise the results in Chapter 10, where we also give an outlook to future work.

Chapter 2

Data description

In many applications, disease data as well as exposure data and covariates are measured on different geographical scales. These include individual level data and aggregated counts for administrative areas as well as regular and irregular grid structures.

In Britain, possible administrative areas are countries, districts and wards. Other geographies include health structures (health authorities, primary care trusts), electoral (parliamentary constituencies), postal (postcode sectors, unit postcodes), statistical (census output areas) and other aggregations (national parks, local education authorities).

Enumeration districts (ED) build the lowest level of census geography in Britain. Higher units, such as electoral wards and countries are merged EDs. Each of them contains approximately 200 households or equivalently around 400 persons.

The next stage of British census geography are wards. On national average, each ward contains approximately 5 500 people, although they tend to be more populous in urban areas. They represent convenient geographical units for small area epidemiological studies, as for each ward, census population and other demographic data are available. For Greater London, we have 873 wards based on the census of 1991. Inner London consists of 310 wards.

The next stage of this geography are local authority districts (LAD) comprising around 20 wards. The number of comprised wards varies between eight to 45 for the region of Greater London.

Another commonly used geography is given by the postcode of residence. It is used to survey the number of incident cases of leukaemia. Postcodes do not nest exactly within wards, but there exist postcode to ED look–up tables, see e.g. `http://census.ac.uk/cdu/Datasets/Lookup_tables/Postal/Postcode_Enumeration_District_Directory.htm`.

In this chapter we describe the different data sets used in this thesis. First of all, this is information on leukaemia incidences in the study region (Section 2.1). Additionally, we use population counts to correlate incident cases of leukaemia in different wards. These are described in Section 2.2. As covariates, we use benzene exposure data (Section 2.3) and the deprivation index as introduced by Carstairs and Morris (1991). This index is described in Section 2.4.

2.1 Leukaemia registration data

Leukaemia is the most common cancer in childhood. There are several types of leukaemia, such as the most common type in children, called acute lymphoblastic leukaemia. For medical background information on leukaemia see for example Groër and Shekleton (1979).

Incident cases of leukaemia are registered at the Office for National Statistics and the Thames Cancer Registry for the period from 1985 until 1996. Before 1985, the area that was covered comprised three separate cancer registries, namely the North West, North East and South registries (Best and Wakefield, 1999). A review of the complex cancer registration system operating in England and Wales is given in Swerdlow (1986) and Gulliford et al. (1993). They found a good documentation of cancer in general. One major reason for decreased data quality is the existence of different regional cancer registers which work at different completeness and accuracy levels. Therefore, the combination of different regional registers may lead to misinterpretations (Swerdlow, 1986). Hence, data from before the amalgamation of the three London cancer registries in 1985 is not considered in this thesis.

Cancer type, date of birth, sex, and postcode of residence (at the time of diagnosis) are available for each registration. Each registered case was checked involving matching cases with regards to postcode, sex and date of birth. In two postcode areas Health Offices of Qatar and Kuwait were located. This leads to unusually

	Greater London	Inner London
number of cases	734	295
median	1	1
arithmetic mean	0.841	0.916
variance	1.070	1.224
form of distribution	right skewed	right skewed
wards with 0 cases	418 (47.9%)	133 (42.9%)
wards with 1 cases	267 (30.6%)	101 (32.6%)
wards with 2 cases	125 (14.3%)	51 (16.5%)
wards with 3 cases	44 (5.0%)	13 (4.2%)
wards with 4 cases	11 (1.3%)	7 (2.2%)
wards with 5 cases	7 (0.8%)	5 (1.6%)
wards with 6 cases	1 (0.1%)	0 (0.0%)

Table 2.1: Some descriptive characteristics of incident cases of childhood leukaemia for Greater and Inner London.

high numbers of registered cases. Best et al. (2001) estimate the number of cases in these two areas as additional parameters in the model. In this thesis we will use the estimated numbers from that paper. Additional data checks are described in more detail in Best et al. (2001). Over the 12–year study period, we observe 734 registered cases of cancer in children under 15 years old in the area of Greater London. For Inner London, there are 295 cases. Some additional characteristics of these data are given in Table 2.1.

For the spatial distribution of the data see Figure 2.1. They look rather scattered over the whole area of Greater London. Nevertheless, there are less cases in the south–western part of the area. Additionally, high incidences of more than three cases are more likely to be observed in the north–east of Greater London.

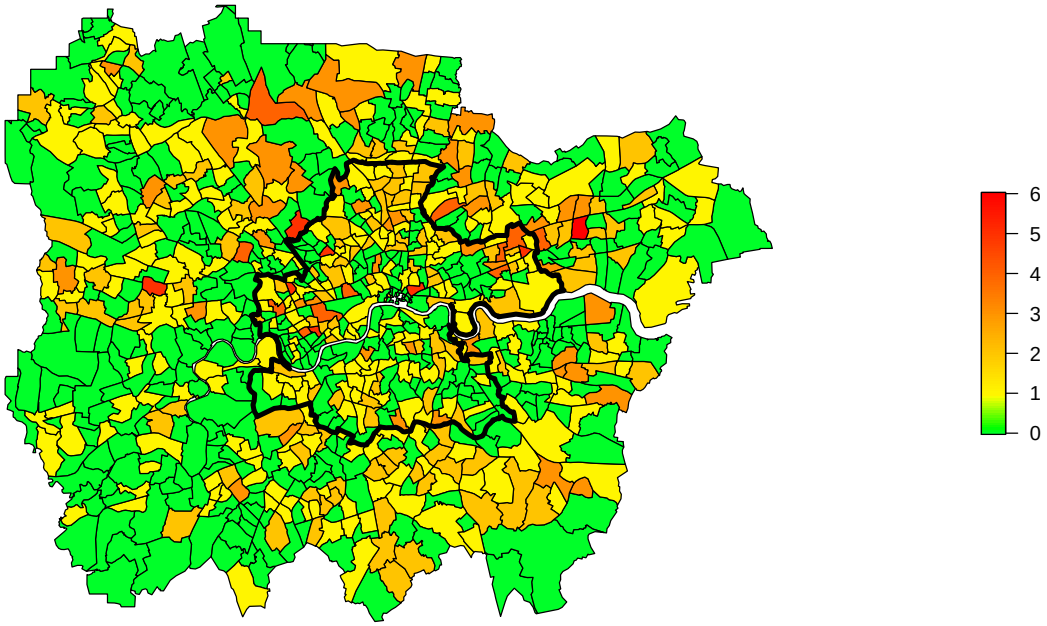


Figure 2.1: Incident cases of childhood leukaemia registered in the period of 1985–1996, bold line marks border of Inner London.

2.2 Population estimates and number of expected cases

The number of observed cases is directly influenced by the population at risk. Therefore, it is necessary to involve population figures into the analysis. Population counts stratified by sex and age are available for enumeration districts from the 1981 and 1991 censuses. For intercensual years, counts must be interpolated accounting for demographic changes (e.g. deaths and births), aging of the population, and migration. Several approaches to model population counts are discussed in Best and Wakefield (1999).

In this thesis we use a set of annual age– and sex–stratified population estimates produced for 1991 EDs following the approach by Arnold (1999) using simple linear interpolation. The interpolated strata– and ED–specific counts are rescaled to sum to the published Registrar General's mid–year population estimates which are available only for much larger geographical areas, i.e., LADs (Best et al., 2001).

Using population estimates it is possible to calculate the number of expected in-

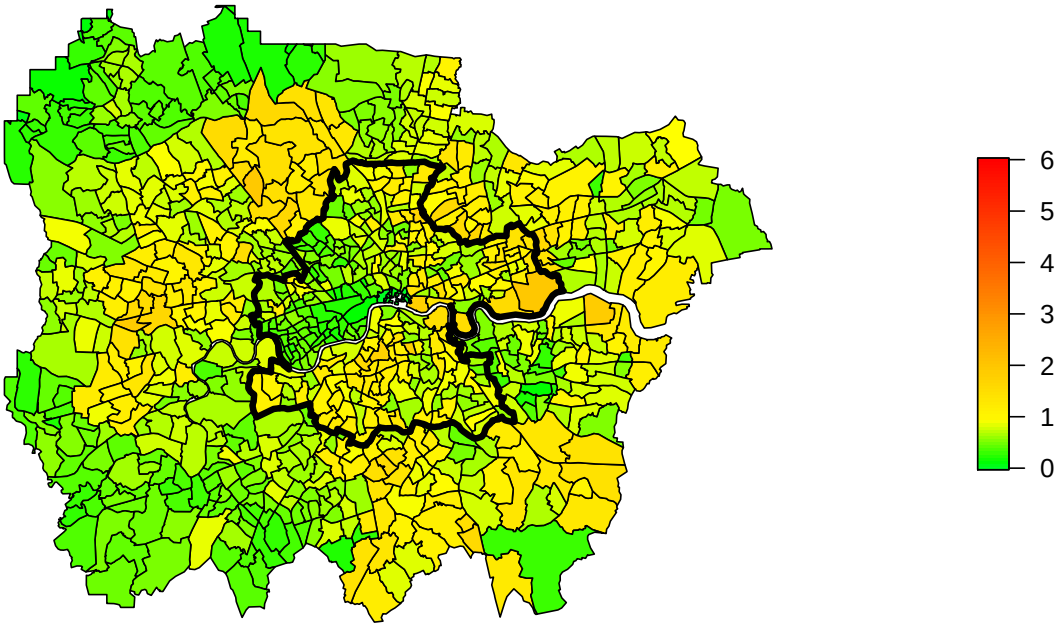


Figure 2.2: Number of expected incidences in the period of 1985–1996, bold line marks border of Inner London.

cidences E_i for each ward i, $i = 1, \ldots, n$. This is done separately for six different age–sex strata corresponding to boys and girls aged 0–4, 5–9 and 10–12 years old, for twelve time periods t corresponding to each of the years 1985–1996. Given the national leukaemia rate r_{st} for age–sex stratum s and year t, we calculate the number of expected cases

$$E_i = \sum_{st} r_{st} N_{ist}$$

using the estimated population at risk N_{ist} in ward i, age–sex stratum s and year t.

An impression of the spatial distribution of the number of expected incidences is given in Figure 2.2. Here, we can see parts with high population density, e.g., in the north of Greater London or the east of Inner London. Some characteristics of population figures are given in Table 2.2. These are very similar for both regions and show smaller values than the ones of observed cases, compare Table 2.1.

Using both, observed and expected cases, we can calculate incidences corrected for population figures. This corresponds to SMRs which are given in Figure 9.1 on page 124.

	Greater London	Inner London
number of expected cases	679.119	237.848
minimum	0.0001	0.0001
1^{st} quartile	0.543	0.533
median	0.752	0.756
arithmetic mean	0.778	0.767
3^{rd} quartile	1.005	1.025
maximum	1.924	1.924
variance	0.110	0.125

Table 2.2: Some descriptive characteristics of the expected incidences of childhood leukaemia for Greater and Inner London.

2.3 Benzene exposure data

Benzene is a colourless liquid with a sweet odour found in air, water and soil. It is produced by human activities, but comes also from natural processes like forest fires or eruption of volcanoes. It passes into the air from burning coal or oil, benzene waste and storage operations, motor vehicle exhaust or evaporation from petrol service stations (Agency for Toxic Substance and Disease Registry (ATSDR), 1997). Smoking was found to be the largest anthropogenic source of background exposure to benzene (Hattemer-Frey et al., 1990). A review of large–scale studies of personal or indoor air levels of benzene is given in Wallace (1996).

Most people are exposed to a small amount of benzene on a daily basis, mainly through breathing air that contains the substance. For small children, the daily intake of air has been estimated to be 2.3 times higher than in adults, accounting for body weight in kg (Wild and Kleinjans, 2003). Benzene is classified as a group 1 carcinogen by the International Agency for Research on Cancer. It is well known that benzene exposure increases the risk of leukaemia in adults, see for example Yardley-Jones et al. (1991), Duarte-Davidson et al. (2001), Dockerty et al. (2001), and Linet and Cartwright (1996).

Obtaining benzene exposure data via personal devices is extremely costly. An alternative is to monitor air quality at selected locations. This gives only information

	Greater London	Inner London
minimum	0.247	0.341
1^{st} quartile	0.862	1.149
median	1.123	1.319
arithmetic mean	1.101	1.321
3^{rd} quartile	1.333	1.485
maximum	2.612	2.612
variance	0.122	0.092

Table 2.3: Some descriptive characteristics of the atmospheric benzene emissions for Greater and Inner London.

about benzene level at these locations and is still expensive.

Another option is the use of an atmospheric emissions inventory. Such an inventory schedules the sources of pollutants within a particular geographic area. For each of the scheduled sources, the emission rate can be calculated by

$$\text{emission rate} = \text{activity rate} \times \text{emission factor}.$$

Activity rates are collected for all sources related to benzene emissions and applied to the activity to estimate the likely emissions in each of the observed areas. Sources of emissions include modelled traffic flows, petrol stations and commercial, residential and industrial combustion processes.

The London Research Centre (`http://www.london-research.gov.uk`) has produced such an atmospheric emissions inventory for London (Buckingham et al., 1997). Using the estimates provided by this inventory, benzene exposure data is provided for the area of Greater London on 1 km $\times$ 1 km grid squares covering the area within the M25 orbital motorway. The estimated numbers are given in tonnes per year, ranging between 0.247 and 2.612 for Greater London, see Figure 2.3. Estimates in this inventory are based on data collected in 1995. By that time traffic flows on most London roads had been at capacity for several years. Therefore it is not unrealistic to assume that these estimates are applicable from the early 1990s onwards. Table 2.3 compares some characteristics of benzene emissions for Greater and Inner London. As we see we estimate higher emissions in

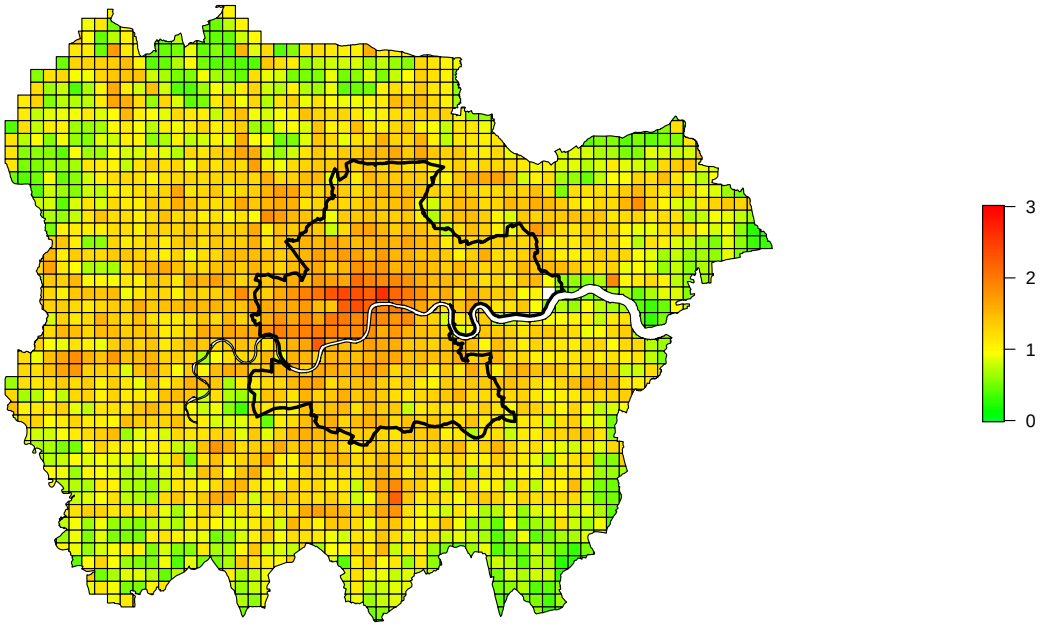


Figure 2.3: Observed benzene exposure data for Greater London on 1 km × 1 km grid cells, line marks border of Inner London.

the Inner London area leading to a smaller variance as well.

2.4 An index of deprivation

An alternative explanation of leukaemia incidence can be seen in deprivation. A prominent example of a deprivation index is the Carstairs index proposed by Carstairs and Morris (1991). It focuses on material deprivation, i.e., the access to material resources to reflect wealth and income (Carstairs, 2000). It includes the percentages of

- individuals living in overcrowded accommodation, i.e., more than one person per room;
- male unemployment;
- low social class households (head of household in social class IV of V);
- households without a car.

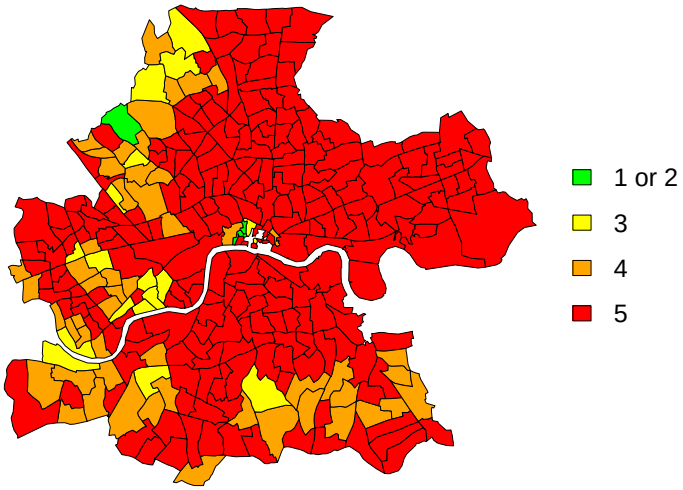


Figure 2.4: Quantiles of Carstairs index for Inner London.

These percentages are reduced to a single standardised value with mean zero and variance one without any weighting or transformation (Carstairs, 2000). There exist alternative deprivation indices differing by weighting scheme, used transformations and the variables included. The larger the score, the greater is the deprivation suffered by the according ward. For a comparison of deprivation indices see Carstairs (1995).

The Carstairs index is given on ward level for the area of Inner London. It is aggregated to quintiles referring to Greater London. For Inner London, this results in a single ward with an index of "1" representing the lowest quintile, while most wards lie in the highest quintile. For modeling purposes, we use the highest class including 233 of 310 wards as reference value. Additionally, we combine class 1 and 2 to a single class of 6 wards. The resulting spatial pattern is presented in Figure 2.4.

Chapter 3

Spatial models

There exist several alternative approaches for modelling spatially distributed data. Data sets can consist of either point–referenced data, areal data or point pattern data.

For point–referenced data, there are observations for a chosen set of locations which vary continuously over the space. Here methods like the descriptive covariogram — a plot of the covariance versus the distance — or exponential modelling are used to describe or to analyse dependence between two locations as a function of the distance, see for example Cressie (1993). Spatial prediction at points where no data are observed can be done by spatial interpolation, so–called kriging. For details on kriging see Stein (1999).

With areal data a fixed subset of the space is partitioned into a finite number of areal units with well defined boundaries. Commonly used models incorporate Markov random fields (Rue and Held, 2005, MRF) and the usage of neighbourhood information by simultaneously and conditionally autoregressive models (see Whittle, 1954, and Besag et al., 1991, respectively).

For point pattern data the location of an observation itself can be considered to be random. Therefore data indicate the occurrence of an event and possibly give additional covariate information. In the later case we refer to marked point processes. Here, tests for spatial randomness, applications of Poisson processes and Markov point processes are appropriate approaches (Diggle, 2003).

A comprehensive review on modelling all kind of spatially distributed data is given

by, e.g., Banerjee et al. (2004). Inclusion of a time dimension leads to spatial–temporal processes which are discussed for example by Finkenstädt et al. (2007).

In the context of spatial epidemiology we usually deal with areal or point pattern data. The number of observed incidences is usually assumed to follow a Poisson distribution, for nonrare diseases a binomial model is more appropriate (Knorr-Held and Besag, 1998). There exist several alternative approaches when a Poisson distribution can be assumed. Best et al. (2000) for example model the observed counts by a Poisson–Gamma model, Wakefield et al. (2000) by a Poisson–Gamma and a Poisson–log–Normal model. Different modelling approaches given by Fernández and Green (2002) and Green and Richardson (2002) introduce a mixture model for Poisson distributed data in which the weights vary according to the observations. Spatial dependencies are commonly modelled in the context of MRFs via conditional autoregressive (CAR) terms. Univariate CAR models introduced by Besag et al. (1991) are extended to multivariate CAR models for two related diseases by Held et al. (2005) and Jin et al. (2005).

Recent model approaches include the analysis of boundaries that separate areas of elevated and lowered risk by Bayesian wombling (see Banerjee et al., 2003, for point–referenced data and Lu and Carlin, 2005, for areal data).

Other approaches try to reduce ecological bias of data that are given on an aggregated level. Jackson et al. (2006) show improvement of inference for these data by combining individual–level data with aggregated data.

This chapter focusses on the models employed on observed as well as generated data in this thesis. Following an overview on Bayesian inference and Markov chain Monte Carlo (MCMC) methods in Section 3.1 and definitions of Poisson and Gamma random fields in Section 3.2, we introduce conjugate Poisson–Gamma hierarchical models (Clayton and Kaldor, 1987) in Section 3.3. This approach assumes data and covariates given for equal geographies. Wolpert and Ickstadt (1998) present a generalisation using a random field approach, allowing for data and covariates on disparate spatial scales. The random field approach is generalised by Best et al. (2000) for an application in epidemiology, allowing to model covariates either as excess or relative risk factors. This extension of Poisson–Gamma models is also described in Section 3.3.

To employ Poisson–Gamma random field models on a specific data set we have to adopt the theoretical approach presented in Section 3.3 to the given data. Sec-

tion 3.4 discusses the chosen settings and the implementation in WinBUGS.

Parameter estimation is done in a Bayesian framework using MCMC methods. Therefore, it is necessary to introduce prior distributions on the highest level of hierarchy expressing uncertainty about the parameters. The choice of prior distributions for the data analysed in this thesis is described in Section 3.4.1. Implementation in WinBUGS is discussed in Section 3.4.2 and Section 3.4.3. While the first deals with the implementation of a restricted version of Poisson–Gamma random field models, the latter discusses a more appropriate implementation of this model class.

Alternative spatial models are used to compare model performances of the different modelling approaches. We employ an alternative ecologic regression model based on a Markov random field (MRF) involving a CAR term as well as the cluster model by Knorr-Held and Raßer (2000). These are presented in Section 3.5 (MRF model) and Section 3.6 (clustering approach).

3.1 Bayesian inference

From a Bayesian point of view both the observables and the parameters of the statistical model are considered to be random. The joint probability function $P(Y, \beta)$ of observed data Y and model parameters β combines information from the data given by a likelihood function and a prior distribution $P(\beta)$ expressing uncertainty about β before taking data into account, i.e.,

$$P(Y, \beta) = P(Y|\beta)P(\beta).$$

Using Bayes theorem, we can express the distribution of the parameters β given the observed data Y by

$$P(\beta|Y) = \frac{P(\beta)P(Y|\beta)}{\int P(\beta)P(Y|\beta)d\beta}\,.$$

The posterior distribution $P(\beta|Y)$ is used for Bayesian inference (Gilks et al., 1996a) and to obtain moments, quantiles and other functions $f(\beta)$ of the parameter of interest.

In many applications, we can use the concept of conjugacy to evaluate the posterior distribution for a given likelihood and prior distribution. This will result in a posterior belonging to the same family of distribution. For the Poisson distribution for

example, the conjugate prior is a Gamma distribution, see the conjugate Poisson–Gamma model in Section 3.3. There are other examples of conjugate priors such as Binomially distributed data with a Beta prior that has a Beta posterior distribution. For details about conjugate priors see for example Gelman et al. (2003).

Hierarchical models are a more flexible approach allowing for prior distributions for the priors themselves, so–called hyperpriors. These reflect uncertainty about the true values of the parameters of the prior distribution. Additionally, this concept allows for structural dependencies of β. Information of all areas is combined via the joint influence of the hyperprior distribution.

In a hierarchical model we improve model formulation of the priors β by hyperpriors ϕ. The joint distribution is given by

$$P(\beta, \phi) = P(\beta|\phi)P(\phi),$$

leading to the posterior

$$\begin{aligned} P(\phi, \beta|Y) &\propto P(\beta, \phi)P(Y|\beta, \phi) \\ &= P(\beta, \phi)\, P(Y|\beta). \end{aligned}$$

Note that

$$P(Y|\beta, \phi) = P(Y|\beta),$$

therefore hyperpriors ϕ are independent of Y given β.

Unfortunately, integration of the numerator of the posterior distribution, namely $\int P(\beta)P(Y|\beta)\, d\beta$ can be difficult, especially for high dimensional problems.

Alternatively, we can employ Markov chains to construct samples from the posterior distribution. By Monte Carlo integration, samples are averaged to obtain the required model parameter. This approach is referred to as Markov chain Monte Carlo (MCMC) methods. The construction of a suitable Markov chain with the desired stationary distribution can be done, for example, by the Metropolis–Hastings algorithm (Metropolis et al., 1953, Hastings, 1957) or the Gibbs sampler (Geman and Geman, 1984, Gelfand and Smith, 1990, Casella and George, 1992).

Each Markov chain starts at initial values chosen by the user. A suitable algorithm constructs a chain of length ϖ. After a sufficiently long burn–in period of length ζ, the Markov chain will reach its stationary distribution, i.e., the posterior of the according parameter. When the steady state is reached, the distribution stays

stationary (Gamerman, 1997). After discarding ζ samples from the so–called burn–in period, we estimate the expected value of any function $f(\cdot)$ by the remaining $\varpi - \zeta$ iterations by averaging

$$\overline{f(\beta)} = \frac{1}{\varpi - \zeta} \sum_{t=\zeta+1}^{\varpi} f(\beta_t)$$

(Gilks et al., 1996a). Introduction of a thinning parameter reduces autocorrelation between the samples. For an overview about MCMC methods and applications see Gilks et al. (1996b). A popular criterion to determine the length of a sufficient burn–in period is the one by Brooks, Gelman and Rubin described in Section 5.1.

Gelman (1996) suggests to use multiple chains for each MCMC simulation and the usage of widely dispersed initial values. This helps to identify whether a Markov chain has reached its target distribution, especially if convergence is slow. We follow this recommendation by employing two Markov chains for each run.

For the Bayesian analysis of the leukaemia data set as well as the simulation study we use the software WinBUGS (Spiegelhalter et al., 2004) and OpenBUGS (Thomas, 2004). All other computations are done in the statistical software R (R Development Core Team, 2006). For linking WinBUGS and R, we use the R package R2WinBUGS (Sturtz et al., 2005), linking between R and OpenBUGS is done via the R package BRugs (Thomas et al., 2006). An introduction to the software packages is given in Chapter 4.

The BUGS software uses Gibbs sampling to construct transition kernels for Markov chain samplers. While compiling, it sets for each model parameter a method to draw a sample from the relevant full conditional distribution. These are chosen according to the hierarchy given in Table 3.1. A review of WinBUGS as well as an example of usage can be found in Cowles (2004).

3.2 Poisson and Gamma random fields

Before we introduce Poisson–Gamma models we need to define Poisson random fields (Definition 3.2.1) and Gamma random fields (Definition 3.2.2). For these definitions we follow the work of Ickstadt (2001).

discrete target distribution	
finite upper bound	inversion
shifted Poisson	direct sampling using standard algorithm
continuous target distribution	
conjugate	direct sampling using standard algorithms
log–concave	adaptive rejection sampling (Gilks, 1992)
restricted range	slice sampling (Neal, 1997)
unrestricted range	random walk Metropolis (Metropolis et al., 1953)

Table 3.1: Sampling methods hierarchy used by WinBUGS (Spiegelhalter et al., 2004).

Definition 3.2.1 (Poisson random field)
Let $\mathcal{Y} \subset \mathbb{R}^d$ with Borel σ–algebra $\mathcal{B}(\mathcal{Y})$ and $N(A)$ a number of points in $\mathcal{Y} \cap A$. $N(dy)$ is a Poisson random field on $\mathcal{Y}$ with non–negative σ–finite intensity measure $\lambda(dy)$ if
a) for each measurable set $A \in \mathcal{B}(\mathcal{Y})$ with $\lambda(A) < \infty$ and integer $k \geq 0$

$$P(N(A) = k) = \frac{\lambda(A)^k \exp(\lambda(A))}{k!}$$

and
b) for any disjoint measurable subsets $A_1, \ldots, A_k \subset \mathcal{B}(\mathcal{Y}), j = 1, \ldots, k$, the random variables $N(A_1), \ldots, N(A_k)$ are independent.

Therefore, for each A the number of points in A has a Poisson distribution with mean $\lambda(A)$. As the Poisson process is a point process, realisations of a Poisson random field are almost surely discrete with finitely many integer point masses or jumps.

Similar to a Poisson random field we can define a Gamma random field.

Definition 3.2.2 (Gamma random field)
Let $\mathcal{S} \subset \mathbb{R}^d$ with Borel σ–algebra $\mathcal{B}(\mathcal{S})$ and $\Gamma(A)$ a number of points in $\mathcal{S} \cap A$. $\Gamma(ds)$ is a Gamma random field on $\mathcal{S}$ with non–negative σ–finite shape measure $\alpha(ds)$ and inverse scale $\beta > 0$ if
a) for each measurable set $A \in \mathcal{B}(\mathcal{S})$ with $\alpha(ds) < \infty$ the random variable $\Gamma(A)$ has a density

$$f_{\Gamma(A)}(t) = \frac{\beta^{\alpha(A)} t^{\alpha(A)-1} \exp(-\beta t)}{\Gamma(\alpha(A))}, \qquad t > 0,$$

and
b) for any disjoint measurable subsets $A_1, \ldots, A_k \subset \mathcal{B}(\mathbb{S})$ *with* $\alpha(A_j) < \infty$, $j = 1, \ldots, k$, *the random variables* $\Gamma(A_1), \ldots, \Gamma(A_k)$ *are independent.*

It follows, that for each set A the random variable $\Gamma(A)$ has a Gamma$(\alpha(A), \beta^{-1})$ distribution with

$$\alpha(A) = \int_A \alpha(ds) = \int_A \alpha(s)\Pi(ds)$$

where $\alpha(s)$ is a density w.r.t. some reference measure $\Pi(ds)$. Realisations of a Gamma process are almost surely discrete as they consist of countably infinitely many jumps at locations $s_m \in \mathbb{S}$ with corresponding magnitudes $\gamma_m > 0$, i.e. $\Gamma(ds) = \sum_m \gamma_m \delta_{s_m}(ds)$. Wolpert and Ickstadt (1998) present the Inverse Lévy Measure (ILM) algorithm to construct a Gamma random field, for our approaches in WinBUGS and their limitations see Sections 3.4.2 and 3.4.3.

3.3 Theory of Poisson–Gamma models

In epidemiological contexts, we observe N_i cases in region i, $i = 1, \ldots, n$, due to an infection or disease or death. This number is modelled in dependence to the expected number E_i of infections or deaths in the corresponding region and a number of possible covariates. The usual approach is to assume N_i to follow a Poisson distribution with mean $\lambda_i E_i$ depending on the number of expected cases E_i and the relative risk λ_i, i.e.,

$$N_i \sim \text{Pois}(\lambda_i E_i).$$

A possible approach for modelling the relative risk λ_i is given by Poisson–Gamma random field models (Best et al., 2000).

The class of Poisson–Gamma random field models is a generalisation of **conjugate Poisson–Gamma models** first described by Clayton and Kaldor (1987). Furthermore, these models represent a generalisation of generalised linear mixed models as described by Böhning (2000) as well as McLachlan and Peel (2000). Clayton and Kaldor (1987) assume the relative risk λ_i in region i to follow a Gamma distribution with shape α and scale τ a priori, i.e.,

$$\lambda_i \sim \text{Gamma}(\alpha, \tau).$$

Thus, the posterior distribution of λ_i is

$$\lambda_i \sim \text{Gamma}(E_i + \alpha, N_i + \tau)$$

and the posterior expectation given the observed values and the prior settings of the Gamma distribution is

$$E(\lambda_i | N_i; \alpha, \tau) = \frac{N_i + \tau}{E_i + \alpha}$$

(Clayton and Kaldor, 1987). Hence, each estimate compromises the observed SMR and the prior mean α/τ. For large numbers in region i, $E(\lambda_i|N_i; \alpha, \tau)$ will be close to SMR $= N_i/E_i$, while for small numbers the expectation is close to the overall prior mean α/τ. For the estimation of α and τ, empirical or hierarchical Bayes methods can be employed.

This model is mathematically easy, but requires a sensible choice of the parameters of the Gamma distribution. The generalisation of this model by hierarchical structures is a necessary improvement.

Furthermore, the model demands the same spatial resolution of observed and expected values. When aggregating individual data to groups or combining smaller groups to larger ones in order to get a common spatial scale, the so–called ecological bias or fallacy occurs. Spatial models assume homogeneous risk within each aggregated area. This implicit Uniform distribution across the aggregation may lead to over– or underestimation of the true effect. Ecological bias is discussed in detail for example in Richardson (1992) and Greenland and Robins (1994).

A generalisation of Poisson–Gamma conjugate models by Clayton and Kaldor (1987) is presented by the papers of Ickstadt and Wolpert (1997) and Wolpert and Ickstadt (1998) as a hierarchical Poisson–Gamma model that also allows for positive association between neighbouring regions. This is modelled by introducing doubly stochastic Poisson processes whose intensities are mixtures of random fields. This approach is generalised by Ickstadt and Wolpert (1999) allowing for covariates. Contrary to other commonly used spatial models, **Poisson–Gamma random field models** use an identity link function rather than log–link, allowing to model additive and multiplicative influence of covariates, not only the latter one. This also leads to aggregation consistency. By relating all observable quantities to an underlying random field it is even possible to model data measured at disparate spatial scales, see Ickstadt and Wolpert (1999). Hence, this class of models overcomes the problem of ecological bias.

Additionally, the model by Ickstadt and Wolpert (1999) relates the intensities of the Poisson distribution to both location–specific covariates and individual–specific attributes. Spatial dependence between subregions is introduced via kernel mixtures.

Assume an observed point process in some set $\mathcal{Y}$ in Euclidean space and a number of covariates J. For any arbitrary aggregation of $\mathcal{Y}$ in region i, $i = 1, \ldots, n$, we set a Poisson regression model with identity link as follows:

$$\begin{aligned} N_i &\sim \text{Pois}(\Lambda_i \, w_i) \\ \Lambda_i &= \beta_0 + \sum_{j \in J} X_{ij} \beta_j, \end{aligned}$$

where w_i refers to a reference weight measure, e.g., the population at risk, $X_{ij}, j \in J$, is a set of covariates with corresponding coefficients β_j, β_0 corresponds to an intercept. Setting $\Lambda_i = \beta_0$ leads to the conjugate Poisson–Gamma model. We express our uncertainty about coefficients on a second stage of hierarchy by prior distributions $\beta \sim \pi(\beta)$.

This partition–based approach still leads to ecological fallacy. Therefore, we refine the partition of $\mathcal{Y}$ until ultimate refinement, leading to observations $N(dy)$ from a Poisson random field with mean

$$\begin{aligned} N(dy) &\sim \text{Pois}(\Lambda(y) \, w(dy)) \\ \Lambda(y) &= \beta_0 + \sum_{j \in J} X_j(y) \beta_j \end{aligned} \tag{3.1}$$

with reference measure $w(dy)$ on $\mathcal{Y}$ for the Poisson random field as in Definition 3.2.1.

In many epidemiologic applications we may also observe individually attributed risk, such as age and gender, which we might want to include in the model. An extension of model (3.1) is to model a point process on a space $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$ of marked points $x = (y, a)$ on location y with marks a. This leads to

$$\begin{aligned} N(dy\, da) &\sim \text{Pois}(\Lambda(y, a) \, w_Y(dy) \, w_A(da)) \\ \Lambda(y, a) &= \beta_0 + \sum_{j \in J} a_j \beta_j, \end{aligned}$$

which equals model (3.1) if $a_j = X_j(y)$ but setting $a_j = X_j(x)$ is more general. The reference weight prior $w_A(da)$ of the attributes is usually chosen to be space independent, i.e., $w_A(da|y_1) = w_A(da|y_2) = w_A(da)$ for $y_1 \neq y_2$, and set to $w_A(da) \equiv 1$,

but can also be location–specific. Choosing $w_A(da)$ to be location–independent gives $w_Y(dy)$ the role of a population reference measure (Ickstadt and Wolpert, 1999). We use the population at risk for the leukaemia data set and the simulation study.

Furthermore, we include **spatial dependencies** between regions introduced through the influence of unobserved spatially varying covariates. These are expressed by an additional component $X_*(y)\beta_*$ of unobserved but spatially correlated covariates $X_*(y)$ and regression coefficient β_* to the intensity $\Lambda(y, a)$.

We use a set $\{s_m\}_{m\in M}$ of point locations in $\mathcal{Y}$ at which the m unobserved covariates are centered. To model the decreasing influence of each latent source with decreasing distance $|y-s|$, kernel functions $k(y, s)$ are suitable, for example Gaussian–like kernels proportional to $\exp((|y - s|/\rho)^2)$ where the variance ρ determines the region of influence of the kernel. The influence of each kernel depends furthermore on the latent positive magnitudes Γ_m associated with the set $\{s_m\}_{m\in M}$, resulting in the latent term $X_*(y) = \sum_{m\in M} k(y, s_m)\Gamma_m$.

For a Bayesian analysis we need to introduce prior distributions for the additional regression coefficient β_* as well as for the magnitudes $\{\Gamma_m\}_{m\in M}$. Ickstadt and Wolpert (1999) suggest using the conjugate prior $\Gamma_m \sim \text{Gamma}(\alpha_m^\beta, \tau_m^\beta)$.

So our model is as follows:

$$\begin{aligned} N(dy\,da) &\sim \text{Pois}(\Lambda(y, a)w_Y(dy)w_A(da)) \\ \Lambda(y, a) &= \beta_0 + \sum_{j\in J} a_j\beta_j + \sum_{m\in M} k(y, s_m)\Gamma_m\,\beta_* \\ \Gamma_m &\sim \text{Gamma}(\alpha_m^\beta, \tau_m^\beta) \\ \beta &\sim \pi(\beta)d(\beta) \end{aligned} \tag{3.2}$$

including m latent risk sources on point locations $\{s_m\}_{m\in M}$ with magnitudes $\{\Gamma_m\}_{m\in M}$. We may extend model (3.2) by enlarging the number of point sources m leading to an inhomogeneous Gamma random field $\Gamma(ds)$ on space $\mathcal{S}$ as in Definition 3.2.2. Shape measure $\alpha^\beta(ds)$ and scale $\tau^\beta(ds)$ must be reduced appropriately when increasing the number of sources. This extends model (3.2) to

$$
\begin{aligned}
N(dy\,da) &\sim \text{Pois}(\Lambda(y,a)w_Y(dy)w_A(da)) \\
\Lambda(y,a) &= \beta_0 + \sum_{j\in J} a_j\beta_j + \int_{\mathcal{S}} k(y,s)\Gamma(ds)\,\beta_* \\
\Gamma(ds) &\sim \text{Gamma}(\alpha^\beta(ds), \tau^\beta(ds)) \\
\beta &\sim \pi(\beta)d(\beta).
\end{aligned} \tag{3.3}
$$

This hierarchical model representation allows for additive influence of covariates as well as for spatial dependencies. For some applications we might doubt the validity of additivity. Additive influence of covariates is more appropriate when we believe in competing, non–interacting effects of covariates, giving alternative explanations of an event. Such covariates are called **excess risk factors** (Breslow and Day, 1980) and are interpretable as the difference of stratum–specific incidences $\beta_j = \mu_{j,1} - \mu_{j,0}$ for covariate $\beta_j, j \in J_A$, for any $\mu_{j,\cdot}$. The mean of the *j-th* covariate is denoted by $\mu_{j,1}$ for diseased/dead persons, while $\mu_{j,0}$ corresponds to non–diseased/living persons. Therefore, covariate j increases the risk additively. Excess risk factors are represented in an identity link Poisson regression model by $\Lambda(x) = \sum_{j\in J_A} X_j(x)\beta_j$ for marked points $x = (y,a)$.

Best et al. (2000) extend model (3.3) for covariates to be modelled either as additive or multiplicative risk factors. Multiplicative modelling reflects different individual susceptibilities. By defining $\exp(\beta_j) = \mu_{j,1}/\mu_{j,0}$ for covariate $\beta_j, j \in J_M$, we increase a background rate of non–infected persons $\mu_{j,0}$ by $\exp(\beta_j)$. These are also called **relative risk factors** by Breslow and Day (1980). As multiplicative risk factors affect the scale of $\Lambda(x)$, we need to introduce a normalising term $c(x)$ leading to $\Lambda(x) = c(x)\exp(\sum_{j\in J_M} X_j(x)\beta_j)$ (Best et al., 2000).

Generalising model (3.3) by allowing for excess and relative risk factors leads to the formulation:

$$
\begin{aligned}
N(dy\,da) &\sim \text{Pois}(\Lambda(y,a)w_Y(dy)w_A(da)) \\
\Lambda(y,a) &= \left(\beta_0 + \sum_{j\in J_A} a_j\beta_j + \int_{\mathcal{S}} k(y,s)\Gamma(ds)\,\beta_*\right) \times c(\beta,y)\exp\left(\sum_{j\in J_M} a_j\beta_j\right) \\
\Gamma(ds) &\sim \text{Gamma}(\alpha^\beta(ds), \tau^\beta(ds)) \\
\beta &\sim \pi(\beta)d(\beta)
\end{aligned} \tag{3.4}
$$

using a normalising term $c(\beta, y)$ leading to mean relative risk factor of unity cal-

culated by

$$c(\beta, y) = \left\{ \int_{\mathcal{A}} \exp(\sum_{j \in J_M} a_j \beta_j) w_A(da|y) \right\}^{-1}$$

(Best et al., 2000). Risk factors and attributes a_j may depend on either or both the location y and attribute a in $x = (y, a)$ in Equation (3.4).

3.4 Settings and Implementation of Poisson–Gamma models

In the leukaemia example, we model the number of observed cases of childhood leukaemia N_i in ward i, $i = 1, \ldots, n$, as a realisation of the random field $N(dx)$ by

$$N_i \sim \text{Pois}(\Lambda(y)\, w_Y(dy)),$$

where we use the number of expected cases of childhood leukaemia as described in Section 2.2 as reference measure $w_Y(dy)$ for region i, $i = 1, \ldots, n$, and set $w_A(da) = 1$.

Covariate information about benzene is available and modelled either as an excess or a relative risk factor. A maximum of one relative risk factor is considered. Hence, the normalising term $c(\beta, y)$ reduces to

$$c(\beta, y) = \left\{ \int_{\mathcal{A}} \exp \left(\sum_k \lambda_{ik} \overline{(B_k - \overline{B})} \beta_{benz} \right) \right\}^{-1} = 1,$$

as $\overline{(B_k - \overline{B})} = 0$ and is therefore no longer considered. In the left–handed term B_k represents the amount of benzene in a grid cell k, $k \in K$, with overall mean $\overline{B}$. When including the Carstairs index as an alternative covariate to benzene emissions this holds accordingly. For details on the benzene covariate see below.

Furthermore, unobserved risk can be modelled by latent spatial variables. For the number and location of latent risk sources, different options are considered. Chapter 6 discusses implementation for a fixed number of kernels at fixed locations which is a restriction of the flexibility of Poisson–Gamma random field models. Chapter 8 applies unrestricted Poisson–Gamma random field models allowing for a random location of covariates.

Benzene is considered as both, an excess and a relative risk factor to compare the different model approaches. This leads to the additive model

$$\Lambda(x) = \beta_0 + X_{\text{benz}}(y)\beta_{\text{benz}} + X_*(y)\beta_* \tag{3.5}$$

and multiplicative model

$$\Lambda(x) = (\beta_0 + X_*(y)\beta_*)\exp(X_{\text{benz}}(y)\beta_{\text{benz}}) \tag{3.6}$$

including a baseline risk β_0 and unobserved risk $X_*(y) = \sum_{m\in M} k(y, s_m)\Gamma_m$ in both equations. The benzene term X_{benz} is calculated using mean polished benzene by

$$X_{\text{benz}} = \sum_k \lambda_{ik}(B_k - \overline{B}).$$

The coefficient λ_{ik} equals the amount of area grid cell k and ward i have in common. This is equivalent to an aggregation of benzene to the spatial scale of the observations. This is necessary to reduce computational time and saves up to several days for models with higher numbers of latent risk sources. Latent risk sources are used on their original spatial scales which is continuous following the Gaussian kernel. The WinBUGS code for an additive model is shown exemplarily in Appendix A.1, implementation details are discussed in Section 3.4.2 for fixed locations of latent kernels. The assumption of a random location requires a more complex implementation as discussed in Section 3.4.3. Furthermore, this includes a more flexible estimation of variance.

3.4.1 Prior settings

We need to define prior distributions for all uncertain parameters. For the Poisson–Gamma models, these are the regression coefficients β_j, $j \in J$, the latent magnitudes Γ_m, and the variance parameter ρ of the Gaussian kernel.

For the regression coefficients β_j, $j \in J = \{\beta_0, \beta_{\text{benz}}, \beta_*\}$, we assume a Gamma distribution Gamma(α, τ) with density

$$f(\beta_j) = \begin{cases} \frac{\tau^\alpha \beta_j^{\alpha-1}\exp(-\tau\beta_j)}{\Gamma(\alpha)} & \text{if } \beta_j > 0, \\ 0 & \text{else,} \end{cases}$$

and set the shape parameter equal to $\alpha = 0.575$ because this gives the ratio of the $90th/10th$ percentile of the prior distribution to be 100. This reflects a prior

probability of 80% for the number of cases associated to each factor to lie between a $1/10th$ and 10 times the prior mean. The prior scale parameter τ is chosen so that the prior mean assumes an equal amount of association for each covariate in J. Since the intensity Λ depends on the ratio of (♯ observed cases)/(♯ expected cases) the prior mean for regression coefficient β_j is

$$\beta_j^{ap} = \frac{\sum_i N_i}{|J| \sum_i E_i}. \tag{3.7}$$

For the latent magnitudes Γ_m of the kernel mixtures we use a Gamma distribution as well. Here we choose

$$\Gamma_m \quad \sim \quad \text{Gamma}(\alpha_m, \tau_m)$$

proportional to the area of the bounding box of the modelled region. To ensure aggregation consistency, the prior mean of the magnitudes is set to be

$$|\mathcal{I}|/m,$$

where $|\mathcal{I}|$ is the area of region $\mathcal{I}$ and m is the number of latent risk sources. Furthermore, the prior weight of a single latent risk source is decreased when increasing the number of modelled latent risk sources. Hence, the parameters of the corresponding Gamma distribution are given by

$$\begin{aligned} \alpha_m &= |\mathcal{I}| \times \tau_m \\ \tau_m &= \frac{1}{m}. \end{aligned}$$

For the kernel $k(y, s)$ we assume a Gaussian kernel with uncertain variance parameter ρ. Prior distribution of ρ is chosen according to a log–Normal distribution with mean 0; the precision varies between 1 and 3 where we check for consistency. This was motivated by a prior study on model adequacy assuming fixed variances of Gaussian kernels. Alternative kernels are also possible, see the discussion in Chapter 10.

3.4.2 Restricted Poisson–Gamma random field models

For a first approach to implement Poisson–Gamma random field models in WinBUGS, we restrict the model by assuming the location of the latent covariates

to be fixed. In this discrete setting of Poisson–Gamma random field models, we apply small and fixed number of Gaussian kernels with an uncertain variance. In different settings of the model, the number of included kernels is increased. By using infinitely many of such fixed kernels we can reproduce a Gamma random field where only m of these have an non–negligible influence. As in the discrete case random variables replace the random field, we can use Gamma distributions to sample the jump height Γ_m of kernel $m \in M$, i. e.,

$$\Gamma_m \sim \text{Gamma}(\alpha_m, \tau_m)$$

In the WinBUGSs' implementation described in this section, we assume a common variance for all kernels, for a generalisation of the model see Section 3.4.3.

The results of the corresponding simulation study are given in Chapter 7. Here we present the main parts of the WinBUGS code, it is given in more detail in Appendix A.1.

For the calculation of the latent risk we discretise the modelled area by dividing the area into squares of a fixed size. This is done as follows:

After the standardisation of the distance between the source itself and a chosen grid cell the cumulative density function `Fx` for source `sx` and grid cell `g` is calculated, see line 7 of the WinBUGS code below. The change in the cumulative distribution function between two points `g` and `(g+1)` (line 11) gives an estimate of the influence of the corresponding kernel `sx` in the grid cell.

```
1  for(sx in 1:nx.source)                          # loop on sources
2     {
3     for(g in 1:(nx.grid+1))                      # loop on cells
4        {
5        dx[sx,g] <- 0.001*(Sx.grid[g]-Sx.source[sx])/rho
6                                                  # "standardisation"
7        Fx[sx,g] <- phi(dx[sx,g])                 # phi=standard normal cdf
8        }
9        for(g in 1:nx.grid)
10       {
11       dFx[sx,g] <- Fx[sx,g+1] - Fx[sx,g]        # change in cdf
12       }
13    }
```

To speed up simulations, this part can be replaced by lines 68–72 in Appendix

A.1 using the Black Box function `eval.grid()` that is given in Appendix A.2.

The value of the kernel itself is calculated by matching the grid cells to the wards they lie in. This is done for longitude and latitude separately. As both directions are assumed to be independent, the value of the bivariate Gaussian kernel can be calculated by multiplication.

In this setting we typically use two Markov chains with 50 000 iterations as burn–in followed by 100 000 iterations for Monte Carlo estimation. The thinning parameter is set to be 5 in order to reduce autocorrelation.

3.4.3 Poisson–Gamma random field models

In contrast to the modelling approach described in the previous section, we now allow for random locations of the latent covariates and independent variances for each kernel in longitudinal and latitudinal direction.

For any $\epsilon > 0$ a Gamma process can be viewed as jumps of size $\Gamma \geq \epsilon$. As discussed in Section 3.2, there are countably infinitely many of such jumps, but for any $\epsilon > 0$ the number of jumps of sizes bigger than ϵ is finite with probability one. Wolpert and Ickstadt (1998) ensure to draw the largest m jumps by the Inverse Lévy Measure (ILM) algorithm. We cannot use this algorithm in WinBUGS, so we proceed as follows.

We use a fixed number m of jumps with corresponding Gaussian kernels to model the latent risk. For the jump heights, we use Gamma distributed draws

$$\Gamma_m \sim \text{Gamma}(\alpha_m, \tau_m)$$

as in the previous section. This does not ensure sampling from a Gamma random field. Nevertheless, as we allow the location of each kernel to be uncertain, we hope that chosen locations correspond to those with the highest probability of the Gamma random field.

To allow for a random location of the kernel, we need to abandon the discretisation described in Section 3.4.2 but to rely on the distance between the centroid of each ward and the location of each source. Necessary adoptions

are described in the following.

While the location of ward i is characterised by its coordinates `wardXcenter`$_i$ and `wardYcenter`$_i$, the latent risk source `sx` is situated at (`Sx.source[sx]`, `Sy.source[sx]`).

The random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is said to be p-variate normally distributed if its distribution function is given by

$$f(\boldsymbol{X}) = \left(\frac{1}{2\pi}\right)^{\rho/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{X}-\boldsymbol{\mu})\right\}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ represents the vector of means and Σ the variance-covariance matrix.

Setting $p = 2$ and covariances $\sigma_{12} = \sigma_{21} = 0$ leads to

$$\begin{aligned} f(X_1, X_2) &= \frac{1}{\sqrt{2\pi}\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left(\frac{(X_1-\mu_1)^2}{\sigma_1^2} + \frac{(X_2-\mu_2)^2}{\sigma_2^2}\right)\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{(X_1-\mu_1)^2}{\sigma_1^2} + \frac{(X_2-\mu_2)^2}{\sigma_2^2}\right)\right\}. \end{aligned} \tag{3.8}$$

After calculation of the absolute distances in lines `5` and `6` of the WinBUGS code below the kernel is calculated according to Equation (3.8).

```
1  for(i in 1:I)
2     {
3     for(sx in 1:Source)
4        {
5        distanceX[sx,i] <- abs(wardXcenter[i] - Sx.source[sx])
6        distanceY[sx,i] <- abs(wardYcenter[i] - Sy.source[sx])
7        kernel[sx,i]    <- exp(-(pow(distanceX[sx,i]/(2*rho), 2) +
8                                 pow(distanceY[sx,i]/(2*rho), 2)) )
9        }
10    }
```

This improved model representation allows variance estimation for each kernel separately which also allows for more flexibility. Furthermore, we assume different variances for longitude and latitude instead of one common ρ for both directions. For the prior distributions of the variances in x-direction

$\rho_m^{(X)}$ and in y-direction $\rho_m^{(Y)}$ of each kernel $m \in M$ we choose a Gaussian distribution of the logarithm with mean 0 and different precisions p ranging between 1 and 3 similar to Section 3.4.1, i.e.,

$$\begin{aligned} \log(\rho_m^{(X)}) &\sim \text{Gau}(0, p_X) \\ \log(\rho_m^{(Y)}) &\sim \text{Gau}(0, p_Y) \end{aligned}$$

allowing for different values p_X and p_Y for each direction. Large values for precision $p \in \{p_X, p_Y\}$ represent a concentrated influence in a small disc round the kernels location only, while smaller values indicate an influence in a larger area. Both possibilities may be present in a data set given the actual location of each kernel. Extending this approach by, e.g.,

$$p_X = \begin{cases} p_1 & \text{if } z > 0.5 \\ p_2 & \text{if } z \leq 0.5 \end{cases}$$

for an arbitrary value $z \sim \text{Unif}(0, 1)$ allows even more flexibility and improves convergence of the model.

We now implement a random location of each kernel. For each kernel, we suggest a location depending on the prior value and a Uniformly distributed random variable. The coordinates of random location of each latent covariate $(l_X^{(R)}, l_Y^{(R)})$ are given as a combination of the prior location (l_X, l_Y) and the random variable (R_X, R_Y), e.g.,

$$l_X^{(R)} = l_X + R_X$$

where

$$R_X \sim \text{Unif}(\min(C_X) - l_X, \max(C_X) - l_X)$$

and (C_X, C_Y) represent the set of coordinates of Inner London. The corresponding WinBUGS code is given in Appendix A.2 assuming multiplicative influence of benzene exemplarily.

Figure 3.1 shows a boxplot of the jump heights in a selected model of the simulation study. Here, we use all jumps in 10 000 iterations for two kernels and two chains in model Sf2. Although this is not a proof, the actual jumps heights tend to be large as to be expected for a Gamma random

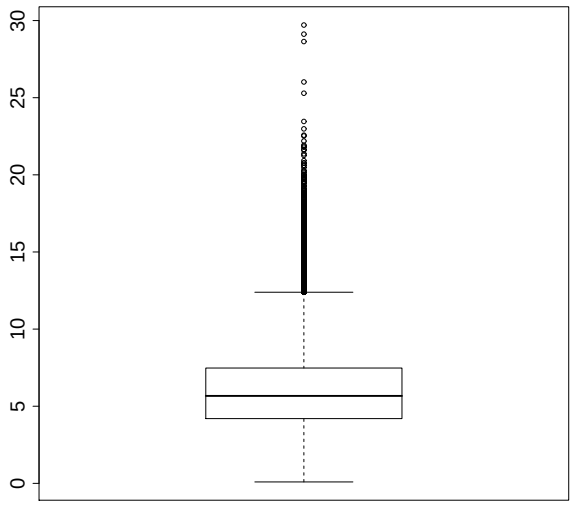


Figure 3.1: Boxplots of the jump heigths Γ_m, $m \in \{1, 2\}$, as estimated by WinBUGS in 10 000 iterations, $\nu = 2$, model Sf2.

field and therefore support the assumption that our implementation leads to an approximation of the Gamma random field. Another indication for the suitability of this approach are the actually estimated locations which are discussed in Section 8.7.

As the model is more complex compared to the restricted implementation described in Section 3.4.2, we increase the burn–in to 500 000 followed by another 500 000 iterations for Monte Carlo estimation. Again, we use two chains and set the thinning parameter to be 5 to reduce autocorrelation. To determine the number of required iterations to achieve the conjugated state we use the criterion of Brooks, Gelman and Rubin described in Section 5.1.

3.5 The Markov random field–based ecologic regression model

This alternative approach combines spatially structured and unstructured spatial effects with the influence of covariates.

We assume the observed cases N_i to follow a Poisson distribution with mean parameter μ_i. The $\log(\mu_i)$ depend on the logarithm of the expected number of deaths E_i, an overall level α and spatial random effects as well as a term depending on the mean averaged benzene observed in region i with the effect estimated by coefficient β_{benz}. For the spatial effects we differentiate between latent covariates with (V_i) and without (U_i) a spatial structure. This leads to the model

$$\begin{aligned} N_i &\sim \text{Pois}(\mu_i) \\ \log(\mu_i) &= \log(E_i) + \alpha + V_i + U_i + \beta_{\text{benz}}(B_i - \overline{B}). \end{aligned}$$

For the spatially unstructured effects V_i we assume a Gaussian distribution with mean 0 and precision parameter τ_V, i.e.,

$$V_i \sim \text{Gau}(0, \tau_V).$$

Following Clayton and Kaldor (1987), the spatially structured effects U_i are based on an intrinsic Gaussian conditional autoregressive model (CAR) which is

$$U_i | U_j, j \neq i \sim \text{Gau}(\bar{b}_i, \tau_U),$$

(Besag et al., 1991). The term $\bar{b}_i = (\sum_j w_{ij} U_j)/w_{i+}$ refers to the mean of neighbouring areas $j \neq i$ around area i, $w_{i+} = \sum_j w_{ij}$. The weights $w_{ij} = 1$ if areas i and j are neighbours, i.e., share a border, and $w_{ij} = 0$ otherwise. This definition has the intuitive interpretation for the conditional mean $E(S_i|S_{-i})$ as a weighted average of all neighbouring regions S_j. This specification of spatially structures effects leads to a Gaussian Markov Random field, see Rue and Held (2005). We will therefore refer to this model as Markov random field (MRF) model. In addition, the b_i values are constraint in summing up to zero (Besag and Kooperberg, 1995). However, this requires an improper

and unbounded uniform distribution on the real line for α, i.e.,

$$\alpha \sim \text{Unif}(-\infty, \infty).$$

For β_{benz}, we also use an uninformative Gamma distribution as prior as follows

$$\beta_{\text{benz}} \sim \text{Gau}(0, 0.0001),$$

hyperpriors for the precision parameters of the spatial effects are set to be

$$\tau_V \sim \text{Gamma}(0.5, 0.0005) \qquad \tau_U \sim \text{Gamma}(0.5, 0.0005).$$

Using the software WinBUGS we use MCMC techniques for estimation of the posterior distribution. For the MRF model two chains with a burn–in period of 200 000 iterations were chosen, followed by a sample of 400 000. To reduce autocorrelation in the Markov chains we set the thinning parameter to be 10, this leads to 40 000 iterations.

3.6 The clustering approach by Knorr–Held and Raßer (2000)

An alternative to the Poisson–Gamma model and the MRF model is a cluster or partition model, for example the so–called BDCD (Bayesian Detection of Clusters and Discontinuities in Diseases Maps) model described by Knorr-Held and Raßer (2000). Although the main goal is clustering, it is often used as a disease mapping tool (Best et al., 2005). We choose this model as it is of similar complexity as Poisson–Gamma models.

The basic idea assumes a constant mortality risk within one or more neighboured regions. These are combined to so–called clusters. Using adaptive smoothing we should be able to detect discontinuities in the modelled region.

Given the number of observed cases N_i in region i, $i = 1, \ldots, n$, and the number of expected deaths E_i, we assume a constant relative risk h_j in one or more regions, leading to

$$N_i \sim \text{Pois}(E_i\, h_j).$$

Regions are grouped to a cluster C_j with associated relative risk h_j which is a part of the set of all regions. We can write $C_j \subset \{1, \ldots, n\}$, $j = 1, \ldots, k$, where $C_{j_1} \cap C_{j_2} = \emptyset$ for $j_1 \neq j_2$ and $C_1 \cup C_2 \cup \ldots \cup C_k = \{1, \ldots, n\}$. The number of clusters k is unknown a priori.

To find a partition of the region into k clusters, we employ Reversible Jump MCMC methods introduced by Green (1995) allowing to switch between different values k.

As prior distribution for the number of cluster centers we choose a Uniform distribution on the number of regions $\{0, \ldots, n\}$ implying an equal probability for each aggregation of regions. Given the number of cluster centers k we choose k regions out of the whole area and define them as cluster centers $G_k = (g_1, \ldots, g_k)$. Remaining regions belong to the cluster center with minimal distance measured by the number of borders between region i and cluster center g_j with $j = 1, \ldots k$.

The corresponding relative risk in $H_k = (h_1, \ldots, h_k)$ in each cluster follows a log–Normal distribution a priori, i.e., $\log(h_j) \sim \text{Gau}(\mu, \sigma^2)$ and we need to specify priors for μ, which is a Uniform one on the whole real line (diffuse prior) and for σ^2 where we choose a highly dispersed Inverse Gamma distribution $\text{IG}(a, b)$ with fixed a and b. In our example we choose $a = 1$ and $b = 0.01$. Other values are tested and lead to similar results.

Additionally, we introduce possible moves for the Reversible Jump MCMC scheme. For changing the numbers of clusters, we use a *birth* and a *death* move, in which we add and delete a cluster center out of the remaining $n - k$ regions respectively. Additionally we use a *shift* move to change a cluster center g_j. In a *switch* move we exchange positions of two cluster centers in G_k possibly leading to an alternative cluster configuration. This is due to associating a region with the same distance to two cluster centers to the one with the smaller index in G_k. In a *height* move we recalculate the relative risks in H_k using

$$h_j \sim \text{Gamma}\left(y_j + \frac{\tilde{\mu}^2}{\tilde{\sigma}^2}, e_j + \frac{\tilde{\mu}}{\tilde{\sigma}^2}\right), \tag{3.9}$$

where $\tilde{\mu} = \exp(\mu + 0.5\,\sigma^2)$ and $\tilde{\sigma}^2 = \exp(\sigma^2) \times (\exp(\sigma^2) - 1) \times (\exp(2\mu))$

(Knorr-Held and Raßer, 2000). The hyperparameters μ and σ^2 of the corresponding Gamma distribution may be changed in a *hyper* move.

For details of BDCD and the chosen setting see Knorr-Held and Raßer (2000) as well as Sturtz (2002).

For application of BDCD we use a burn–in of 200 000 followed by 40 000 000 iterations using a thinning parameter of 4000. The thinning is increased compared to other models due to high autocorrelation when using Reversible Jump MCMC. This leads to 10 000 samples for Monte Carlo estimation.

Chapter 4

Computation: Linking R and WinBUGS

As already briefly described in Section 3.1, the BUGS (Bayesian inference Using Gibbs Sampling) language provides a very flexible and powerful tool for Bayesian analysis of complex models using MCMC methods. The Windows implementation of BUGS, WinBUGS, provides a graphical interface and makes the BUGS language therefore more convenient and easier accessible. The user can specify the Bayesian model including data and suitable initial values by clicking appropriate buttons. With the provided tool boxes the model can be updated, Monte Carlo estimates for specified quantities can be calculated, and lots of other functions are provided.

For the purpose of a convenient and automatised use of WinBUGS — which is especially necessary for the simulation study carried out here — Sturtz et al. (2005) develop the R package R2WinBUGS. This package uses the scripting language of WinBUGS, which is available from version 1.4 onwards.

R2WinBUGS is available under CRAN[1], the Comprehensive R Archive Network, and can be installed via the command `install.packages("R2WinBUGS")` and loaded by `library("R2WinBUGS")`. The main function of the package is `bugs()`, help is available by typing `?bugs`.

[1] http://CRAN.R-project.org

After specifying the data set necessary for modelling in WinBUGS, the R2WinBUGS package writes a file containing these data which WinBUGS can interpret. R2WinBUGS allows for a wide range of data formats in R, it can be either a *named list*, a *vector*, or a *list of the names* of the data objects. Similar things hold for the initial values of a user defined number of chains. After writing data files and files containing initial values as well as the script itself, WinBUGS is started in the batch mode and runs the script. Some outputs including trace plots and summary statistics are created automatically. Resulting values can be read in either automatically in R by the package itself or stored in ASCII files supporting the coda format. In the latter case, a wide range of inference and output diagnostics are available via the coda package (Plummer et al., 2004). A detailed description of the R2WinBUGS package including reproducible examples demonstrating the usage of the package can be found in Sturtz et al. (2005).

R2WinBUGS allows for an automatised use of WinBUGS. Nevertheless, communication between R and WinBUGS is done via exchanging text files and an interactive process of sampling/convergence diagnostics is not possible. A further development of WinBUGS called OpenBUGS contains the open source version of the BUGS language and can be embedded into R via the interface BRugs, published under CRAN. Installation is possible by `install.packages("BRugs")` similar to R2WinBUGS. The package is loaded by the command `library("BRugs")`.

The BRugs package contains the OpenBUGS software itself, refined versions of the functions that reproduce OpenBUGS functionality as well as functions for data preparation and initial values from R2WinBUGS. Using BRugs, it is possible to control OpenBUGS from R using a dynamic link library which provides a `.C()` interface to BUGS command language. Initial help is available by `?BRugs` or the online manual.

Model procedures are similar to those used in WinBUGS itself as well as the scripting language, but BRugs provides R functions that communicate with OpenBUGS components allowing for an interactive modelling process. There are various functions available, such as

- `modelCheck()` for checking a model in BUGS code stored in a specific file,
- `modelData()` for loading the data, and
- `modelInits()` for the initial values of multiple chains.

The latter two can either be stored in specified files or available as R objects only. For the initial steps of MCMC analysis a wrapper function `BRugsFit()` is available. The range of more than 60 functions reproduces OpenBUGS functionality, here is a selection of some frequently used functions:

- `samplesSet()` to set the chain for a particular variable,
- `modelUpdate()` to update the model,
- `samplesStats()` to produce summary statistics for a variable,
- `samplesHistory()` to plot the trace of a variable,
- `samplesBgr()` for the Brooks–Gelman–Rubin convergence statistics,
- `samplesDensity()` for a smoothed kernel density estimate for continuous data or a histogram for discrete data etc.

The R package is described in detail in Thomas et al. (2006). The reference also includes an example how to use the package.

Chapter 5

Convergence diagnostics and model selection

Fitting Bayesian models by MCMC methods requires to set the number of iterations needed for the model to converge and those necessary to get stable Monte Carlo estimates. While the latter one can be determined by the Monte Carlo error the choice of an appropriate burn–in period is more complicated. Methods include the visual inspection of the Markov chain, where the usage of multiple chains makes stationarity easier to determine. More objective criteria include the criteria of Geweke (1992) and Heidelberger and Welch (1983). A comparative review about possible convergence criteria can be found in Cowles and Carlin (1996). This includes the criterion of Brooks, Gelman and Rubin (BGR) which is implemented into WinBUGS. It is introduced by Gelman and Rubin (1992) and generalised by Brooks and Gelman (1998). In this thesis, BGR helps us to set the burn–in period. It is described in detail in Section 5.1.

The influence of different covariates to observed data can be expressed by alternative models. From those, the user has to identify the most appropriate one. There exist a large number of model selection criteria, for example the Bayes Information Criterion and Bayes factors described by Kass and Raftery (1995). Alternative methods include posterior predictive p–values and conditional p–values which are discussed by Bayarri and Berger (2000), Aitkin et al.

(2004), and Perez and Berger (2002) and increased their popularity during recent years. Furthermore, Vehtari and Lampinen (2004) suggest predictive explanatory power for model comparison and evaluation of model performance.

Another frequently used criterion is the Deviance Information Criterion (DIC), introduced by Spiegelhalter et al. (2002). It combines a measure of complexity and a measure of fit. This allows to compare not only the performance of the applied model but also to judge about the increased complexity by additional covariates. In this thesis, the DIC will be used as an indicator of model performance in both, simulation study and real data set application; it is described in Section 5.2.

5.1 Convergence diagnostics

A possible convergence criterion is the one by **Brooks, Gelman and Rubin (BGR)**. It is introduced by Gelman and Rubin (1992) and generalised by Brooks and Gelman (1998). This criterion is implemented in WinBUGS and available via a tool box. The idea is to monitor η iterations of a number of $\nu > 1$ chains and to compare between and within variances of those chains.

To check convergence of any scalar summary ψ three different quantities are monitored:

The within–sequence variance W is calculated by

$$W = \frac{1}{\nu(\eta - 1)} \sum_{r=1}^{\nu} \sum_{s=1}^{\eta} (\psi_{rs} - \bar{\psi}_{r\cdot})^2$$

for any scalar summary ψ. It should stabilise as the number of iterations η increases.

The mixture–of–sequences variance V is calculated as a pooled posterior average of the between–sequence variance

$$\frac{B}{\eta} = \frac{1}{\nu - 1} \sum_{r=1}^{\nu} (\overline{\psi}_{r\cdot} - \overline{\psi}_{\cdot\cdot})^2$$

and the estimated variance σ_+^2 which is calculated by a weighted average of B and W as follows:

$$\hat{\sigma}_+^2 = \frac{\eta - 1}{\eta} W + \frac{B}{\eta}.$$

This leads to

$$\begin{aligned} V &= \sigma_+^2 + \frac{B}{\nu\eta} \\ &= \frac{\eta - 1}{\eta} W + \frac{B}{\eta} + \frac{B}{\nu\eta}. \end{aligned}$$

If convergence is not achieved yet, we expect W to be less than V. Therefore, we can use the ratio of both quantities for convergence diagnostics.

The variance ratio R is given by the ratio of the within–sequence variance W and mixture–of–sequences variance V. This ratio is corrected for sampling variability depending on the degrees of freedom of estimation $d \approx 2V/\mathrm{Var}(V)$ assuming normality of the marginal distribution of each scalar quantity ψ. Therefore, the variance ratio is estimated to be

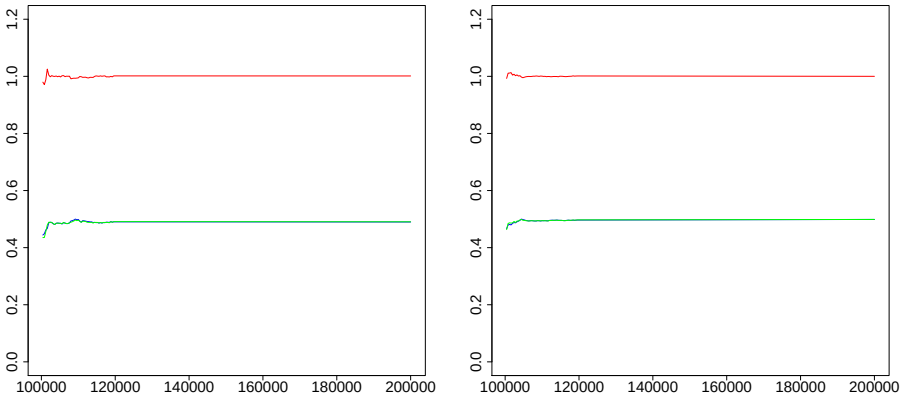
$$R = \frac{(d+3)\,V}{(d+1)\,W}$$

(Brooks and Gelman, 1998).

Dividing each chain into batches of length b, we can plot the development of V, W and R as η increases.

If assumption of normality is violated R should be modified by using interval lengths rather than variance ratios. We calculate the empirical $100(1-\alpha)\%$ interval of η simulation draws of each single chain as a substitute of within–sequence variance W. As a substitute for mixture–of–sequences variance V we use the total–sequence interval as the empirical $100(1-\alpha)\%$ interval out of all $\nu\eta$ observations. Both quantities are used to calculate

$$R_{\text{interval}} = \frac{\text{length of total–sequence interval}}{\text{mean length of the within–sequence intervals}}$$



(a) (b)

Figure 5.1: BGR plots for selected parameters in model Af1 in Chapter 8 for β_{latent} (a) and for the x–location of Gaussian kernel $l_X^{(R)}$ (b).

(Brooks and Gelman, 1998). The resulting plot is among the standard output provided by WinBUGS using batches of size 50 or larger, leading to at most 100 different values as η increases. The average length of the 80% total–sequence interval is plotted in green, the one of within–sequence variances is blue, while their ratio R_{interval} is red. For plotting purposes pooled– and within–interval widths are normalised to have an overall maximum of one (Spiegelhalter et al., 2004). Approximate convergence is attained if the green and the blue line stabilise at the same value resulting in a ratio of one plotted as a red line such as in Figure 5.1. This figure shows the BGR plot exemplarily for a Poisson–Gamma model with additive influence of benzene and one latent covariate applied on structure A (Af1), for details see Chapter 8. Here, we use a burn–in of 100 000 followed by 100 000 iterations. The thinning parameter is chosen to be 5.

5.2 The Deviance Information Criterion

The Deviance Information Criterion (DIC) can be seen as a Bayesian analogue of Akaikes Information Criterion (Akaike, 1973). Formally introduced by Spiegelhalter et al. (2002), it consists of a measure of complexity and a measure of fit. For complexity, we estimate the effective number of parameters by

$$p_D = \overline{D(\theta)} - D(\bar{\theta}), \tag{5.1}$$

where the Bayesian deviance is defined by

$$D(\theta) = -2\log\left(p(y|\theta)\right) + 2\log\left(f(y)\right), \tag{5.2}$$

see Spiegelhalter et al. (2002).

We combine the deviance as a classical estimate of fit with twice the effective number of parameters estimated by Equation (5.1) leading to

$$\begin{aligned} \text{DIC} &= D(\bar{\theta}) + 2p_D \\ &= \overline{D(\theta)} + p_D. \end{aligned} \tag{5.3}$$

The standardising term $f(y)$ of the Bayesian deviance function in Equation (5.2) is a function of the data alone. For comparison of different models used for exactly the same data it can be neglected, i.e., by setting $f(y) \equiv 1$. WinBUGS provides a tool box to calculate the DIC automatically using this setting. However, this approach is not appropriate for the simulation setting applied in Chapter 7 as we generate different data sets in each simulation leading to different values of $f(y)$.

For members of the exponential family with $E(Y) = \mu(\theta)$ we do not need to set $f(y) = 1$. Instead we can use the saturated deviance which is obtained by setting $f(y) = p(y \,|\, \mu(\theta) = y)$ in Equation (5.2). Using this definition, we expect the posterior expected deviance to be approximately the number of free parameters in θ if the model is true (Spiegelhalter et al., 2002) giving a possible check for model adequacy.

For data assumed to be Poisson distributed, i.e., $N_i \sim \text{Pois}(\theta_i\, E_i)$, the deviance $D_i(\theta_i)$ for region i, $i = 1, \ldots, n$, can be written as

$$D_i(\theta_i) = \begin{cases} 2\,\left\{ y_i \log\left(\frac{y_i}{\theta_i\, E_i}\right) - (y_i - E_i\theta_i)\right\} & \text{if } y_i > 0 \\ 2\; E_i\, \theta_i & \text{if } y_i = 0 \end{cases} \tag{5.4}$$

(McCullagh and Nelder, 1990). The total deviance $D(\theta)$ is then calculated by

$$D(\theta) = \sum_{i=1}^{n} D_i(\theta_i). \tag{5.5}$$

When implementing DIC estimation in WinBUGS negative values of p_D are possible. In the simulation study this occurs for example if we employ a hard–wired function in the software package BlackBox (Oberon microsystems, Inc., 2004) in our model. In the BUGS language, p_D is calculated by posterior means of stochastic parents. These are changed by the hard–wired function, therefore estimates are not reliable as mentioned on the WinBUGS mailing list on April, 21st, 2005, see

`http://www.jiscmail.ac.uk/cgi-bin/wa.exe?A2=ind0504&L=bugs&P=R3370&I=-1.`

Therefore we decided not to use WinBUGS' implemented calculation corrected for the saturated deviance but to use our own coding of the deviance as well as of the DIC in R.

To decide whether one model is significantly superior to other considered models the point estimate should be supported by its corresponding variance. Zhu and Carlin (2000) propose a so–called 'Brute Force' approach by rerunning the Bayesian model N times with different starting values and different seeds leading to a sequence of estimated DICs which is $\text{DIC}_1, \ldots, \text{DIC}_N$ for variance estimation. Hence, we have

$$\widehat{\text{Var}}(\text{DIC}) = \frac{1}{N-1} \sum_{l=1}^{N} (\text{DIC}_l - \overline{\text{DIC}}_l)^2. \tag{5.6}$$

This approach is very time–consuming and therefore not suitable for most applications. Alternative approaches include estimation of variances and covariances using only one setting of the model by the delta method, dealing

with effective sample sizes and accounting for autocorrelation through batching. For two independent samples of size G_1 and G_2 we separately compute $D(\bar{\theta})$ according to Equations (5.4) and (5.5) as well as $\bar{D}$ and plug those into

$$\text{Var}(\text{DIC}) = \text{Var}\left(2\bar{D} - D(\bar{\theta})\right) = 4\text{Var}(\bar{D}) + \text{Var}(D(\bar{\theta}))$$

assuming those estimates to be uncorrelated. Using batching, $\widehat{\text{Var}}(\bar{D})$ can be estimated directly from the data. A detailed explanation of the batching approach is given below.

For estimation of $\text{Var}(D(\bar{\theta}))$ we express this term as a function of $\text{Var}(\bar{\theta}_i)$ and $\text{Cov}(\bar{\theta}_i, \bar{\theta}_{i'})$ for any $i \neq i'$ using the multivariate delta method as follows

$$\text{Var}(D(\bar{\theta})) \approx \sum_i \left(\frac{\partial D_i(\bar{\theta}_i)}{\partial \bar{\theta}_i}\right)^2 \text{Var}(\bar{\theta}_i) + \sum_{i \neq i'} \frac{\partial D_i(\bar{\theta}_i)}{\partial \bar{\theta}_i} \frac{\partial D_i(\bar{\theta}_i)}{\partial \bar{\theta}_{i'}} \text{Cov}(\bar{\theta}_i, \bar{\theta}_{i'}) \tag{5.7}$$

(Zhu and Carlin, 2000). The $\bar{\theta}_i$ are posterior means of the random mean measure, their variance and covariance can be estimated from the $\{\theta^{(g_1)}\}_{g_1=1}^{G_1}$ output using batching. The first derivative of a Poisson deviance is given by

$$\frac{\partial D_i(\bar{\theta}_i)}{\partial \bar{\theta}_i} = \begin{cases} 2\sum_i(-N_i/\bar{\theta} + E_i) & \text{if } y_i > 0 \\ 2\,E_i & \text{if } y_i = 0. \end{cases}$$

When using **batching** for estimation of $\widehat{\text{Var}}(\bar{D})$, we compute the $D^{(g_2)}$ output using Equations (5.4) and (5.5) and divide this sequence of length G_2 into t successive batches of length T. For each of the batches, we calculate batch means $B_1, \ldots, B_t$ and

$$\bar{B} = \frac{1}{t}\sum_{i=1}^{t} B_i.$$

This leads to the variance estimate

$$\widehat{\text{Var}}(\bar{D}) = \widehat{\text{Var}}(\bar{B}) = \frac{1}{t(t-1)}\sum_{i=1}^{t}(B_i - \bar{B})^2$$

producing reliable estimates if T is large enough so that the correlation between batches is negligible. Furthermore, t needs to be chosen large enough to

produce reliable estimates of $\mathrm{Var}(B_i)$. Estimation of $\mathrm{Var}(\bar{\theta}_i)$ and $\mathrm{Cov}(\bar{\theta}_i, \bar{\theta}_{i'})$ follows the paper by Zhu and Carlin (2000).

Besides batching Zhu and Carlin (2000) presented two other approaches for estimating $\mathrm{Var}(\bar{\theta}_i)$ and $\mathrm{Cov}(\bar{\theta}_i, \bar{\theta}_{i'})$ in their paper, but all three lead to very poor results as presented in Zhu and Carlin (2000).

We might also build some alternative samples for the batching approach. The one we introduce in this thesis is to build the sample using **thinning**, i.e., we construct a new sample by using each $\mathfrak{N}th$ value. That would be
$1st$, $(\mathfrak{N}+1)th$, $(2\mathfrak{N}+1)th$, ..., $(\mathfrak{L}-\mathfrak{N}+1)th$ value for the first batch,
...
$\mathfrak{N}th$, $2\mathfrak{N}th$, $3\mathfrak{N}th$, ..., $\mathfrak{L}th$ value for $\mathfrak{N}th$ batch
where $\mathfrak{L} = \eta \times \nu$. Using those alternatively built batches we proceed with variance estimation as proposed by Zhu and Carlin (2000).

Additionally, we suggest the use of bootstrapping and cross–validation techniques to improve the fit of Var(DIC).

For **bootstrapping** we use the MCMC estimates of θ of length $\mathfrak{L}$ to sample a new chain with similar length. This is done with replacement. Then we estimate the DIC of this sample. The procedure is repeated $\mathcal{N}$ times to estimate

$$\mathrm{Var}^{\mathrm{Boot}}(\mathrm{DIC}) = \frac{1}{\mathcal{N}-1}\sum_{l=1}^{\mathcal{N}}(\mathrm{DIC}_l - \overline{\mathrm{DIC}}^{\mathrm{Boot}})^2,$$

where

$$\overline{\mathrm{DIC}}^{\mathrm{Boot}} = \frac{1}{\mathcal{N}}\sum_{l=1}^{\mathcal{N}}\mathrm{DIC}_l$$

and DIC_l is from the $l'th$ bootstrap sample.

An alternative approach is to adopt **cross–validation** techniques for variance estimation of DIC. We construct a new sample of θ by leaving out $\mathcal{N}$ elements successively from the whole sample of length $\mathfrak{L}$. This leads to $\mathfrak{L}/\mathcal{N}$ new samples. In this approach, the goal is not to estimate characteristics of the left–out iterations but to use the new samples to estimate the variance of the DIC. For each sample l, we calculate DIC_l, $l = 1, \ldots, \mathfrak{L}/\mathcal{N}$, which we

use to estimate

$$\mathrm{Var}^{\mathrm{Cross}}(\mathrm{DIC}) = \frac{1}{(\mathcal{L}/\mathcal{N})(\mathcal{L}/\mathcal{N}-1)} \sum_{l=1}^{\mathcal{L}/\mathcal{N}} (\mathrm{DIC}_l - \overline{\mathrm{DIC}}^{\mathrm{Cross}})^2,$$

where

$$\overline{\mathrm{DIC}}^{\mathrm{Cross}} = \frac{1}{\mathcal{L}/\mathcal{N}} \sum_{l=1}^{\mathcal{L}/\mathcal{N}} \mathrm{DIC}_l.$$

Table 5.2 presents selected results of variance estimation for three Poisson–based examples, namely

1. The conjugate Poisson–Gamma hierarchical model (Spiegelhalter et al., 2004, Examples I) employed on the numbers of failure of ten power plant pumps by George et al. (1993) with 2 chains each of length 1000 following a burn–in of 2500 (Pumps).

2. The CAR model used for disease mapping (Spiegelhalter et al., 2004, Examples in the GeoBUGS Manual): rates of lip cancer in 56 counties in Scotland as analysed by Clayton and Kaldor (1987) and Breslow and Clayton (1993) with 2 chains each of length 1000 following a burn–in of 2500 (Lip Cancer).

3. The Poisson–Gamma random field model employed on a generated data set based on the leukaemia data described in Section 2. Model and data equals combination Aa1 in the simulation study, see Section 6.2.1. We employ 2 chains with a burn–in period of 50 000 followed by 100 000 iterations when setting the thinning parameter to be 5.

Additionally, we give the corresponding DIC estimates for each of the examples as well as the number of effective parameters p_D. As the quality of estimation depends on the effective sample size Table 5.2 also contains the autocorrelation of lag one for the deviance of one chain of each of the examples.

Estimation was done using OpenBUGS (Thomas, 2004) via the R software (R Development Core Team, 2006) by the package BRugs (Thomas et al., 2006).

		Pumps	Lip Cancer	Leukaemia
DIC		17.873	103.131	315.958
pD		8.662	26.413	1.926
Autocorrelation of deviance for chain 1		0.0130	0.2130	0.0350
Brute Force	$N = 1000$	0.0352	0.1230	0.0018
Bootstrap	$\mathcal{N} = 100$	0.0034	0.0084	0.0003
	$\mathcal{N} = 1000$	0.0030	0.0084	0.0004
Batching	$T = 10$	0.1109	0.6238	0.1260
	$T = 50$	0.1143	0.7866	0.1171
	$T = 100$	0.0992	0.9296	0.1224
Batching & Thin	$T = 10$	0.0902	0.3039	0.1159
	$T = 50$	0.1064	0.2601	0.1171
	$T = 100$	0.1005	0.3608	0.1403
Cross Validation	$\mathcal{N} = 1$	0.0032	0.0101	0.0012
	$\mathcal{N} = 10$	0.0033	0.0111	0.0018
	$\mathcal{N} = 100$	0.0029	0.0124	0.0019

Table 5.1: Variances of the DIC estimated by different methods for three considered data sets.

Variance estimates are reported in Table 5.1. All values are estimated to be less than one. This holds for the Brute Force approach which can be seen as a "gold–standard" as well as for alternative approaches.

Unfortunately, faster methods than the Brute Force approach do not produce reliable estimates. Using the bootstrap approach we underestimate the true DIC. The amount of underestimation does not depend on the batch size.

The batching approach overestimates, again the results are stable for different batch sizes. For the Pumps model and the Leukaemia examples, batching and the combination of batching and thinning produce similar estimates. For the Lip Cancer examples incorporating a CAR–term a reduction of the estimate is achieved by thinning. This is probably due to autocorrelation between the samples in the original chain which can be reduced by thinning.

The cross–validation approach underestimates the variance for the Pumps and the Lip Cancer examples by factor 10, for the Leukaemia examples we achieve estimates close to those of the Brute Force approach. For all models, results are consistent for different sizes of $\mathcal{N}$.

There is just a slight relationship between the estimated variances for any of the examples. The influence of the autocorrelation in the deviance chain seems to be more important.

All together, the variance of the DIC is estimated to be less than one in all three examples which is very small compared to the DIC. It can be assumed that similar results would be achieved for models which DICs in the same order of magnitude. Therefore, an estimation of variance is not taken into account any further. Nevertheless, Zhu and Carlin (2000) report much higher DICs and Var(DIC) in their paper. A more detailed analysis of our proposed methods using different models than those applied here is necessary. For a discussion on the variance of DIC in general see Chapter 10.

Chapter 6

A simulation study: settings

For analysing characteristics and performances of Poisson–Gamma models using different settings within this model class itself and in comparison to the alternatively chosen MRF model and the BDCD cluster model, we perform a simulation study. This gives the possibility to explore the performance of the model with respect to the true underlying structure. Certain scenarios will be set up for data generation as described in Chapter 6.2.

In principle, all components such as covariates, population figures, and the number of observations can be varied within a simulation study. Nevertheless, it is sufficient to use population figures and benzene outcome like they are observed in the data set of childhood leukaemia described in Chapter 2. We used the expected number of cases as a surrogate for the real population which gives the generated numbers the interpretation of expected cases. We will only change the amount and the type of influence benzene has on the generated number of incidences in the fixed population numbers of Inner London.

Data are generated assuming a number of different scenarios as described in Section 6.2. One assumes the influence of a single covariate only, others additionally involve risk due to a latent risk factor represented by a Gaussian kernel, a linear or a plateau trend, or clusters of increased risk in combination with the covariate. The chosen covariate can be benzene, as in our real data

set, or any other variable. In a simulation study, we can set the amount of influence of the chosen covariate to result in a pre–specified number of incident cases or deaths among the population at risk.

As a first step we implement the restricted version of Poisson–Gamma models as described in Section 3.4.2. The different models employed on these data are described in Section 6.1. The results are reported in Chapter 7. Results of the more flexible implementation of Poisson–Gamma random field models as described in Section 3.4.3 are given in Chapter 8. The quality of the findings is measured using DIC and the Mean Square Error (MSE) as described in Section 6.3.

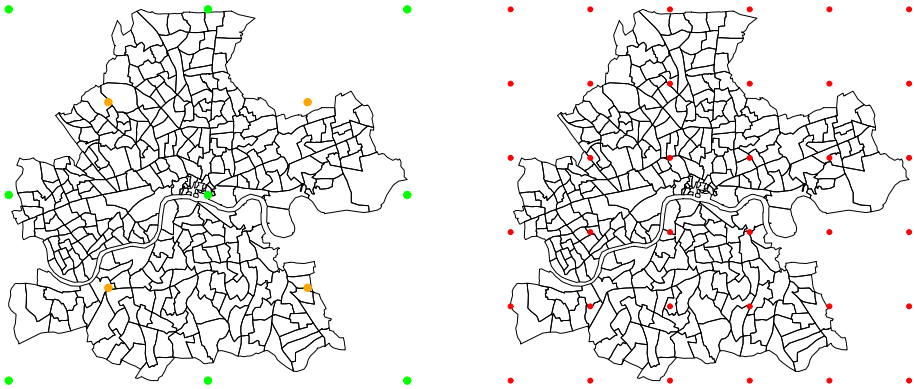
6.1 Models employed on generated data

On each of the generated data sets, several models will be employed. First, we use Poisson–Gamma models where risk is modelled depending on benzene only.

Additionally, we combine the different types of influence of benzene with latent covariates located at different distances. Here we use the restricted implementation assuming fixed locations as described in Section 3.4.2. In principal, each possible distance can be assumed. Nevertheless, given the fixed extension of the area, which is 24.6 km from east to west and 22.5 km from north to east, not all distances are suitable.

We calculate the number of points which fit into the bounding box covering the area of Inner London given a certain distance. Typically, this is not an integer. Therefore, we round this value up and recalculate the distance between two latent covariates accordingly. The deviation between the chosen distance and the recalculated one is the larger, the larger the distance between two locations is chosen.

As distances, we choose $d_1 = 15$ km (resulting in 9 latent risk sources) and $d_2 = 5$ km (leading to 36 latent risk sources) located from the edges of the grid covering the whole area of Inner London. The real distances for d_1 (d_2)



(a) (b)

Figure 6.1: Location of latent sources: (a) 4 (orange) and 9 (green) kernels with distance d_1, (b) 36 kernels with distance d_2.

are 12.316 km (4.926 km) for the latitude and 11.126 km (4.502 km) for the longitude. For computational reasons, a higher number of latent risk sources will not be considered.

There is also the possibility to locate four sources to span a square of 15 km length centered around the middle of the bounding box of Inner London. A combination of four centrally located and nine sources spread over the whole area leads to 13 latent sources as shown in Figure 6.1 (a).

Furthermore, we are interested in the performance of models where no benzene is considered, but a certain amount of latent risk sources. This gives an idea how the model performs when covariates that have an influence on the incidence outcome are not involved in the model. Here we employ two different settings: one involving 13 risk sources located as in Figure 6.1 (a) and a second setting of 36 risk sources, see Figure 6.1 (b). The location of the latent risk sources is assumed to be fixed in this part of the study and implemented in WinBUGS following Section 3.4.2. A list of the resulting Poisson–Gamma models is given below. For an implementation allowing the location of each kernel to be random see Section 3.4.3.

Summarising the models introduced above, we apply the following settings to generated data in the first part of the study:

restricted Poisson–Gamma models with additive influence of benzene:
Model a: no latent risk sources;
Model b: 4 latent risk sources with $d_1 = 15$km;
Model c: 9 latent risk sources with $d_1 = 15$km;
Model d: combination of sources from b and c to 13 latent risk sources;
Model e: 36 latent risk sources with $d_2 = 5$km;

restricted Poisson–Gamma models with multiplicative influence of benzene:
Model g: no latent risk sources;
Model h: 4 latent risk sources with $d_1 = 15$km;
Model i: 9 latent risk sources with $d_1 = 15$km;
Model j: combination of sources from b and c to 13 latent risk sources;
Model k: 36 latent risk sources with $d_2 = 5$km;

restricted Poisson–Gamma model with no influence of benzene:
Model w: 36 latent risk sources;
Model x: 13 latent risk sources.

6.2 Generation of data sets

To analyse the performance of Poisson–Gamma models, we consider five different settings.

In Section 6.2.1 we give details on the study design where only benzene determines the observed number of incidences without involving any latent risk factors in the generation procedure.

In a second part of this study, a latent risk source will be combined with the benzene covariate in data generation. This results in the structures described in Section 6.2.2.

Furthermore, we construct some other commonly used spatial structures including covariates with linear spatial trend and clusters with increased risk.

One pattern includes a covariate with linear spatial trend (Section 6.2.3), again with either a low or a high additive or multiplicative influence of benzene.

Other spatial structures increase the risk in all wards south of the Thames (Section 6.2.4) or are characterized by clusters with increased risk, see Section 6.2.5. Those scenarios again include different levels of benzene.

6.2.1 Data sets determined by benzene only

In this setting we do not account for any latent risk sources in data generation. This leads to a number of observed cases influenced only by the chosen covariate benzene. As Poisson–Gamma models give the possibility to introduce this covariate either as an excess or a relative risk factor, both interpretations will be considered in data generation leading to additive and multiplicative models which are referred to as "structures" throughout this thesis. Besides the model type, the amount of influence of benzene can be varied. Details on both are given in the following.

The model type

The number of observed cases Y_i in ward i is assumed to follow a Poisson distribution with mean depending on the expected number of cases E_i and an area–specific relative risk Λ_i, i.e.,

$$Y_i \sim \text{Pois}(\Lambda_i \, E_i).$$

For E_i we use the expectations from the example given in Best et al. (2001), see Section 2.2. For calculation of Λ_i we assume either an additive or a multiplicative influence of benzene in grid k.

Furthermore, we set the amount of influence of benzene. We do not vary the intercept parameter but fix it to have no influence at all, i.e., $\beta_0 = 0$ for the additive model and $\beta_0 = 1$ for the multiplicative one. Using a multiplicative model this gives a lower limit of around 250 observed cases; less cases can be observed setting $\beta_0 < 1$.

β_{benz}	additive model	β_{benz}	multiplicative model
0.5	55.397	0.1	246.614
1	110.795	0.3	265.522
2	221.589	0.6	297.779
3	**332.384**	**0.9**	**335.563**
4	443.178	1.0	349.583
5	553.973	2.0	544.043
6	664.768	2.5	695.583
7	**775.562**	**2.7**	**771.288**
8	886.357	3.0	905.682

Table 6.1: Number of expected cases for different parameters of benzene using additive and multiplicative models, bold numbers indicate parameter values chosen for data generation.

The amount of influence

The amount of influence of benzene is determined by the parameter β_{benz}. We calculate the number of expected cases for the area of Inner London for various settings in multiplicative and additive models as presented in Table 6.1.

Using ideas of experimental design, we select two different values for β_{benz} reflecting 'high' and 'low' influence of the covariate for each type of model according to Table 6.1. Keeping in mind that the observed number of cases in Inner London is about 290 we set $\beta_{\text{benz}} = 0.9$ leading to about 330 cases. A similar amount of expected observations is achieved by additive models with $\beta_{\text{benz}} = 3$.

For a high influence of benzene, we roughly double the expected number of observations leading to $\beta_{\text{benz}} = 2.7$ (multiplicative model) and $\beta_{\text{benz}} = 7$ (additive model) giving approximately 770 cases, see Table 6.1.

Therefore, we employ the following structures in data generation which include only benzene but no latent risk factors:

Structure A: additive model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 3$;
Structure B: additive model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 7$;
Structure M: multiplicative model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 0.9$;
Structure N: multiplicative model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 2.7$.

Note that different modelling schemes lead to a similar number of observations, but different variances. For the additive structure A we calculate a variance of 0.356, while it is 0.224 for structure M; in structure B, we get a variance of 1.940, for structure N we achieve 3.024.

According to the model formulation (see Equation (3.6) for the multiplicative model and Equation (3.5) for the additive one) the parameter of the Poisson distribution Λ_i in ward i, $i = 1, \ldots, n$, is calculated. Hence, we sample the observed number of incident cases in each situation according to the calculated Poisson mean.

6.2.2 Including a latent risk source as covariate

It is also possible to include "unobserved" risk factors in data generation additionally to benzene. This gives information on how well the selected models can detect existing latent risk sources or spatial patterns in general.

We expand data generation by just one Gaussian kernel representing the influence of an unobserved covariate. This kernel is located next to one of the four centered latent risk sources considered in models b and h, see Section 6.1. Hence, we employ models on generated data using kernels located in direct neighbourhood to this generated covariate (models b/h, d/j), in a distance of approximately 3.5 km (models c/i) or nearby as for models e and k. The chosen coordinates (527, 175) are represented by a dot in Figure 6.2. The amount of influence by the latent covariate is chosen to be similar to a low influence of benzene, i.e., causing approximately 330 cases. Two different scenarios are considered. First, both covariates are matched with a proportion of 1:1, in the second setting the number of cases caused by benzene is roughly double the amount of that caused by the latent covariate.

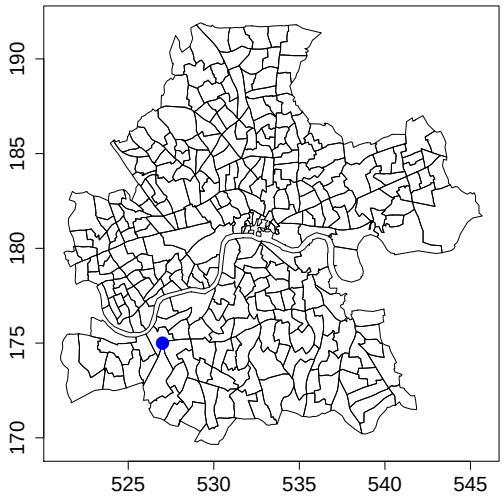


Figure 6.2: Location for the latent risk source in data generation. Coordinates are given in km, the point of origin is determined by the most western and the most southern point of the UK.

Furthermore, we consider the variance of the Gaussian kernel representing the spatial dimension of influence and the parameter β_{latent} of that kernel. The standard deviation (sd) of the Gaussian kernel is selected according to Table 6.2 to be 2.5 km which is equivalent to a variance of $2.5^2 = 6.25$ km^2. The distance of 6.25 km^2 corresponds approximately to a quarter of the circumference of Inner London. This choice of variance ensures that the Gaussian kernel has a non–neglecting influence in more than 25% of all wards which is an increase of $0.0013 \times \beta_{\text{latent}} = 0.6812$ on the parameter Λ_i. Due to a median smaller than 0.0001, the increase in at least half the wards in Inner London is less than $0.0001 \times \beta_{\text{latent}} = 0.0520$ and therefore negligible. Increases of Λ_i caused by other values of the standard deviation of the Gaussian kernel are given in Table 6.2.

The range of Inner London is about 24 km in the east–west direction and the north–south range is 22 km. Given a total influence of 0.6385 by the Gaussian kernel

$$K_{i,m} \sim N\left(\begin{pmatrix}527\\175\end{pmatrix}, \begin{pmatrix}6.25 & 0\\0 & 6.25\end{pmatrix}\right),$$

sd	minimum	1st quantile	median	3rd quantile	maximum	sum
1	< 0.0001	< 0.0001	< 0.0001	< 0.0001	0.1904	0.7231
2	< 0.0001	< 0.0001	< 0.0001	0.0005	0.0480	0.6833
2.5	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$	$\mathbf{< 0.0001}$	**0.0013**	**0.0307**	**0.6385**
3	< 0.0001	< 0.0001	0.0001	0.0020	0.0214	0.5919
5	< 0.0001	0.0002	0.0007	0.0020	0.0079	0.4445

Table 6.2: Increase of Λ_i by the latent risk source depending on standard deviation of Gaussian kernel.

we select $\beta_{\text{latent}} = 520$ resulting in 332 additionally expected cases for the additive model. As presented in Figure 6.3, influence of the generated Gaussian kernel is only observable in the south–western part of Inner London.

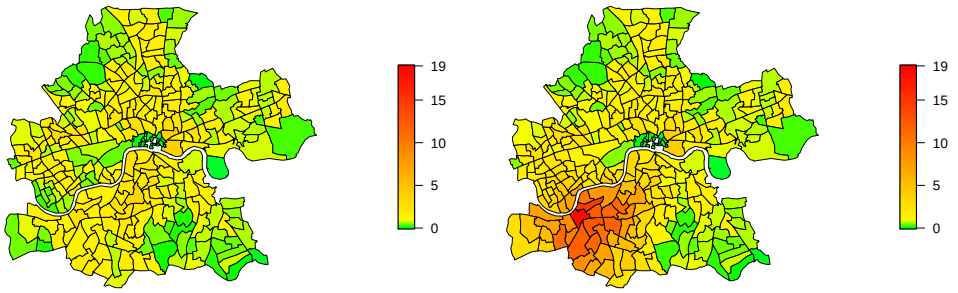
When using the multiplicative model, influences of the latent source and benzene are multiplied. This makes an adaption of β_{latent} necessary leading to a value of 350 (480) for 665 (1100) expected observations in total for a low (high) level of benzene. In addition to data generation settings when no latent variable is involved, we consider the following structures:

Structure C: additive model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 3$, $\beta_{\text{latent}} = 520$, Gaussian kernel at (527, 175) with sd=2.5;

Structure D: additive model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 7$, $\beta_{\text{latent}} = 520$, Gaussian kernel at (527, 175) with sd=2.5;

Structure O: multiplicative model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 0.9$, $\beta_{\text{latent}} = 350$, Gaussian kernel at (527, 175) with sd=2.5;

Structure P: multiplicative model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 2.7$, $\beta_{\text{latent}} = 480$, Gaussian kernel at (527, 175) with sd=2.5.



(a) (b)

Figure 6.3: Effect of involving latent variable in data generation: true underlying risk Λ_i involving only benzene ((a), structure A) and additionally a latent covariate at (527, 175) ((b), structure C).

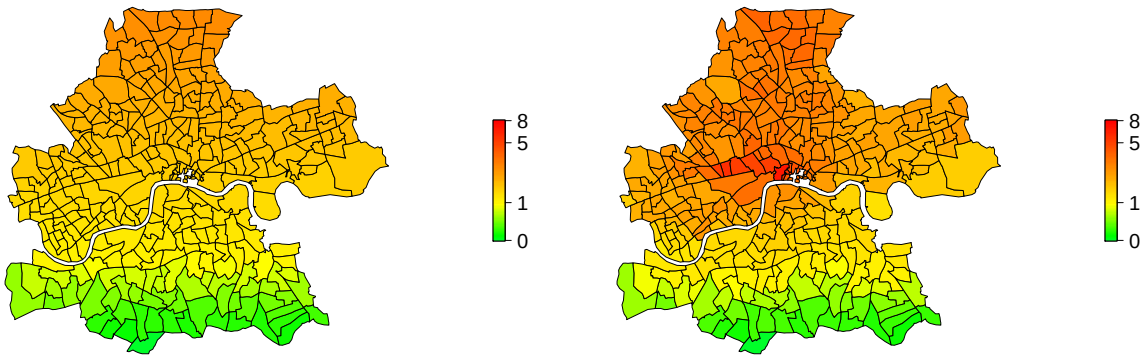
6.2.3 Including a covariate of linear spatial trend

A common scenario in spatial epidemiology is to include an additive covariate which has a linear spatial trend. In our case the trend is modelled to have no influence in the southern wards of Inner London. Influence is increased with increased distance to the minimum of all centroids of the wards, i.e.,

$$\text{trend} = \beta_{\text{trend}} \times (\text{centroid}_i - \min_i(\text{centroid}_i)),$$

where centroid_i is the centroid of region i, $i = 1, \ldots, n$, in km, $\min_i(\text{centroid}_i)$ is the smallest coordinate of the centroids of n wards, i.e., the most southern centroid, and β_{trend} is the trend coefficient. The later one is chosen to account for approximately 332 cases, which is equal to the number of cases the latent risk source in structures C/D/O/P accounts for.

The resulting spatial pattern of the trend component itself as well as in combination with a low multiplicative influence of benzene is given in Figure 6.4. All together, in the situation of including a covariate with a linear spatial trend we have the following structures for data generation:



(a) (b)

Figure 6.4: Linear spatial trend component (a) and resulting spatial pattern Λ_i (b) for structure Q.

Structure E: additive model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 3$, $\beta_{\text{trend}} = 0.137$;

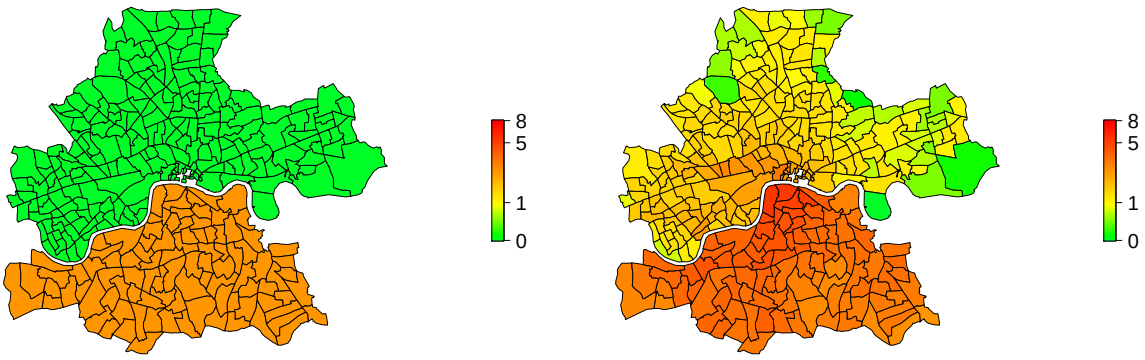
Structure F: additive model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 7$, $\beta_{\text{trend}} = 0.137$;

Structure Q: multiplicative model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 0.9$, $\beta_{\text{trend}} = 0.194$;

Structure R: multiplicative model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 2.7$, $\beta_{\text{trend}} = 0.140$.

6.2.4 Increased risk in southern areas

Another interesting spatial structure is given by an extreme of the linear spatial trend component (Section 6.2.3), built by a region of constantly increased risk. We choose the Thames as the separating factor between the high and the low-risk region. Therefore, 215 wards northwards the Thames are allocated to the low-risk region with risk determined by benzene only. The risk of the remaining 95 wards is increased such that we obtain an extra



(a) (b)

Figure 6.5: Increased risk for wards south of the river (a) and resulting spatial pattern Λ_i (b) for structure G.

330 expected cases compared to the "benzene-only" structures. The risk for each ward is calculated as follows

$$\Lambda(y) = \begin{cases} \beta_0 + X_{\text{benz}}(y)\beta_{\text{benz}} + \beta_{\text{increase}} & \text{(additive model)} \\ \beta_0 \times \exp\left(X_{\text{benz}}(y)\beta_{\text{benz}}\right) + \beta_{\text{increase}} & \text{(multiplicative model)} \end{cases}$$

where

$$\beta_{\text{increase}} = \begin{cases} 0 & \text{if northern ward,} \\ \beta^*_{\text{increase}} & \text{if southern ward.} \end{cases}$$

We choose $\beta^*_{\text{increase}}$ such that it accounts in total for 332 cases. The value of $\beta^*_{\text{increase}} = 3.7$ itself corresponds to the number of cases generated additionally in each ward. Choosing the river as the partition has another advantage: it is now possible to analyse differences in the MRF models differing in the choice of neighbouring structure across the river. The expected spatial pattern for β_{increase} as well as for the spatial pattern $\Lambda(y, a)\, w_Y(dy)$ presented in Figure 6.5 assumes a low and additive influence of benzene, denoted as structure G. All analysed structures are determined by different settings for benzene and are as follows:

Structure G: additive model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 3$, $\beta^*_{\text{increase}} = 3.7$;

Structure H: additive model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 7$, $\beta^*_{\text{increase}} = 3.7$;

Structure S: multiplicative model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 0.9$, $\beta^*_{\text{increase}} = 3.7$;

Structure T: multiplicative model, influence of benzene is high, i.e., β_{benz}=2.7, $\beta^*_{\text{increase}} = 3.7$.

6.2.5 Increased risk in cluster regions

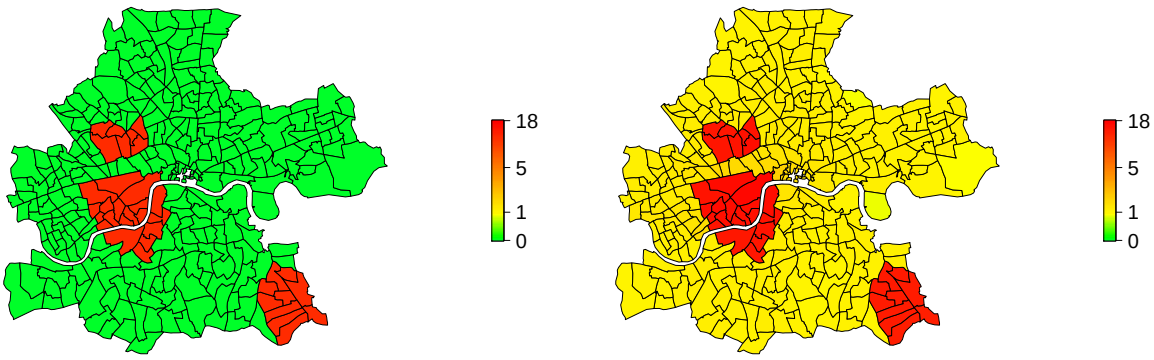
Instead of increasing risk in half of the area as in Section 6.2.4, we can assume an increased risk for some specific clusters. We design three different clusters for the area of Inner London. The location of each cluster is randomly chosen with some aspects in mind. These are:

- Clusters should be built by different numbers of wards;
- One of the clusters should be divided by the river Thames;
- There should be clusters at the border of Inner London as well as in the center; and
- Distances between the clusters should be different.

In our cluster configuration presented in Figure 6.6 the northern cluster consists of 7 wards, the southern one of 10 wards, the cluster in the center is built by 19 wards. Therefore, risk of 35 wards is increased. For each ward, the resulting risk is calculated as in Section 6.2.4 but with

$$\beta_{\text{increase}} = \begin{cases} \beta^{**}_{\text{increase}} & \text{if ward belongs to any of the clusters,} \\ 0 & \text{else.} \end{cases}$$

Again, the parameter $\beta^{**}_{\text{increase}}$ is chosen to account for approximately 332 additional cases. The resulting spatial pattern of $\beta^{**}_{\text{increase}} = 15$ is shown in



(a) (b)

Figure 6.6: Increased risk for wards on three chosen clusters (a) and resulting spatial pattern Λ_i (b) for structure U.

Figure 6.6 (a), the pattern of the expected cases when combined with a low multiplicative influence of benzene is shown in Figure 6.6 (b). The settings of the structure incorporating an increased risk in certain cluster regions are:

Structure I: additive model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 3$, $\beta^{**}_{\text{increase}}$=15;

Structure J: additive model, influence of benzene is high, i.e., $\beta_{\text{benz}} = 7$, $\beta^{*}_{\text{increase}} = 15$;

Structure U: multiplicative model, influence of benzene is low, i.e., $\beta_{\text{benz}} = 0.9$, $\beta^{**}_{\text{increase}} = 15$;

Structure V: multiplicative model, influence of benzene is high, i.e., β_{benz}=2.7, $\beta^{**}_{\text{increase}} = 15$.

6.3 Evaluation of model performance

In simulation studies, it is possible to compare the modelled parameter $\widehat{\Lambda}_i E_i$ with the generated parameter $\Lambda_i E_i$. This is done according to the criteria described in this section.

Of course, it is possible to compare the estimates $\hat{\beta}_0$ and $\hat{\beta}_{\text{benz}}$ with the data–generating parameters β_0 and β_{benz} in each structure. In models incorporating latent risk sources the parameters $\hat{\beta}_{\text{latent}}$ and $\hat{\Gamma}_m$ can also be compared with their equivalents in data generation. Nevertheless, this comparison is still insufficient as different values may lead to similar results, especially when using a different model for data generation and simulation. Therefore, we compare $\widehat{\Lambda}_i E_i$ to $\Lambda_i E_i$.

This can be done by the so–called Mean Square Error (MSE) between simulated and calculated parameter of the Poisson distribution of Y_i, which is calculated by

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\Lambda_i E_i - \widehat{\Lambda}_i E_i)^2$$

(Mood et al., 1974). If the model performs well, we will observe small values of MSE; if the given structure is not found by the applied model, larger values will result.

MSE calculated for the parameters of the Poisson distribution indicates if the true underlying structure is recovered easily but it can be calculated only if this structure is known as in a simulation study, not for real observations. An alternative measure is the DIC introduced in Section 5.2. This will be calculated in all simulations. Therefore, comparison of the performances of DIC and MSE indicates if the DIC is a sufficient method in real applications.

Chapter 7

Simulation results for restricted Poisson–Gamma models

Following the design of the simulation study described in Chapter 6 we combine the different generated structures with the chosen restricted Poisson–Gamma models as described in Chapter 6.1. In this chapter we will use structures determined by benzene only (A/B/M/N) and those characterised by the influence of benzene in combination with one latent risk source (C/D/O/P) only.

The corresponding implementation in WinBUGS is described in Section 3.4.2. We assume the location of the latent covariates to be fixed which limits the flexibility of Poisson–Gamma random field models. For the covariates we use Gaussian kernels as described in Chapter 6, their variance is assumed to be unknown. The underlying data set of each combination is generated separately. To stabilise estimation of model performance we use three runs of each combination of generating structure and model.

The resulting DIC and MSE of the simulations are summarised by their means in Tables 7.1 and 7.2 respectively. We use mean values instead of those of each single run to give a better overview. For structure A we exemplarily give the results of the three single runs on page 77.

Structures that do not involve latent risk at all such as A and B were well

model	A	B	C	D	M	N	O	P
a	336.1	328.2	780.3	665.0	347.2	388.2	718.2	3850.1
b	322.0	299.3	493.3	441.1	325.2	402.9	485.6	3602.7
c	341.8	354.6	762.9	604.4	321.3	420.0	638.3	3596.2
d	336.8	334.5	434.6	393.9	315.2	411.1	428.9	3681.0
e	329.0	336.5	444.2	427.2	337.3	427.5	446.6	1268.0
g	323.9	355.3	851.3	722.2	330.1	327.5	782.5	3276.9
h	340.6	346.7	506.2	462.8	355.4	334.8	474.8	673.4
i	345.7	348.1	657.4	543.4	339.0	323.3	616.3	1281.8
j	336.4	369.8	410.6	371.7	348.1	376.9	402.0	458.3
k	350.7	349.4	439.9	406.0	321.6	321.7	417.0	587.2
w	398.4	427.1	461.8	431.9	359.7	494.9	425.8	1257.7
x	398.5	429.3	465.8	488.4	342.0	625.5	431.4	1441.3

Table 7.1: Summary of mean DICs of structures A, B, C, D, M, N, O and P of the simulation study.

identified by all employed models. Usually, MSEs are less than one, which is really small given the total number of 310 wards where risk is estimated simultaneously.

For structures where a latent risk source is generated by a Gaussian kernel the current implementation is too restrictive. The model has difficulties in identifying the underlying structure correctly, especially if the position of the kernel is not in the vicinity of the one used for data generation. The problem is not solved by increasing the number of latent risk factors as this results in overestimation caused by additional covariates and increases computational time. Therefore it is necessary to allow for a random location of latent risk sources. This is possible within the framework of Poisson–Gamma models and shows the flexibility of the model class. It results in an extension of our model implementation as described in Section 3.4.3. The corresponding results are given in Chapter 8.

Benzene is always included as a covariate in data generation. If we do not include this covariate in the model, we are nevertheless able to achieve satisfactory results. The Gaussian kernels are very flexible already and are able

model	A	B	C	D	M	N	O	P
a	0.0011	0.0253	5.786	5.794	0.0052	0.871	22.056	98.457
b	0.0085	0.0195	1.723	1.889	0.0132	0.877	12.631	96.355
c	0.0116	0.0079	4.954	5.358	0.0176	0.886	19.850	95.883
d	0.0029	0.0284	1.040	1.153	0.0108	0.878	11.191	98.082
e	0.0183	0.0406	1.549	1.661	0.0133	0.871	12.786	63.444
g	0.0330	0.205	5.975	6.196	0.0153	0.0148	22.193	104.834
h	0.0353	0.173	1.545	2.036	0.0134	0.0078	12.103	51.161
i	0.0283	0.151	4.443	4.853	0.0124	0.0257	19.447	55.170
j	0.0350	0.167	0.977	1.106	0.0048	0.0490	10.059	55.880
k	0.0308	0.144	1.439	1.207	0.0153	0.0435	11.721	58.055
w	0.153	0.694	1.675	1.869	0.0469	1.657	13.232	62.379
x	0.205	0.908	0.993	1.601	0.0575	2.423	9.480	78.021

Table 7.2: Summary of mean MSEs of structures A, B, C, D, M, N, O and P of the simulation study.

to adopt the missing covariate. It is reasonable to assume that model performance will be improved even further when allowing for random location of the kernels.

In the following sections, we describe main results for the different structures in more detail. While Section 7.1 deals with data generated assuming additive influence of benzene but no latent risk sources (structures A and B), Section 7.2 discusses results for structures M and N. For structures where we model the latent risk at fixed locations, results are presented in Section 7.3 for additive influence (structures C and D) and in Section 7.4 for multiplicative influence of benzene, i.e., structures O and P.

As already described the restricted implementation of Poisson–Gamma models is not flexible enough for satisfying results when latent risk is involved in data generation. We therefore do not apply this implementation on other generated structures. Results of the more flexible approach for all structures is discussed in Chapter 8.

For a summary of structures' and models' names see page v.

7.1 Additive influence of benzene, no latent risk sources

Data sets analysed in this section are generated assuming an additive influence of benzene only. For structure A we assume a low influence of benzene accounting for about 330 cases, for structure B the influence of benzene produces about 770 cases.

When modelling these data by Poisson–Gamma models that do include a benzene term we calculate very small MSEs for all applied models as presented for all three runs separately in Table 7.3. Corresponding DICs are given in Table 7.4. As all have the same order of magnitude, we decide to focus on the mean of the values in the following only which are a good substitute and ease model comparison as done in Table 7.1 for the DIC and Table 7.2 for the MSE.

For the Poisson–Gamma model including benzene as an excess risk factor applied on data generated according to structure A (Aa) the amount of modelled risk explained by benzene is about 75–85%, which is a similar amount of variation as the added Poisson noise in data generation. For data generated according to structure B this amount is about 70–85%. For Poisson–Gamma models where we additionally include latent covariates in modelling the amount of risk explained by benzene reaches similar levels for both structures. By adding latent risk sources it is possible to explain some of the baseline risk. This amount is elevated by increasing the number of latent risk sources.

As the MSE's magnitude does not depend on the number of latent risk sources additionally involved in the model and all of those produce a homogeneous latent field we conclude the ability of the model to identify the data generating structure concerning the number of covariates.

In order to analyse the influence of different interpretations of covariates we compare the results of additive and multiplicative models. Compared to additive models, the MSE is slightly increased for multiplicative ones, which

	first run	second run	third run	mean
Aa	0.0012	0.0013	0.0008	0.0011
Ab	0.0138	0.0101	0.0017	0.0085
Ac	0.0027	0.0270	0.0052	0.0116
Ad	0.0006	0.0013	0.0068	0.0029
Ae	0.0095	0.0020	0.0434	0.0183
Ag	0.0260	0.0335	0.0394	0.033
Ah	0.0280	0.0289	0.0490	0.0353
Ai	0.0327	0.0248	0.0274	0.0283
Aj	0.0317	0.0287	0.0446	0.035
Ak	0.0270	0.0379	0.0276	0.0308
Aw	0.149	0.160	0.150	0.153
Ax	0.198	0.207	0.209	0.205

Table 7.3: summary of MSE in situation A

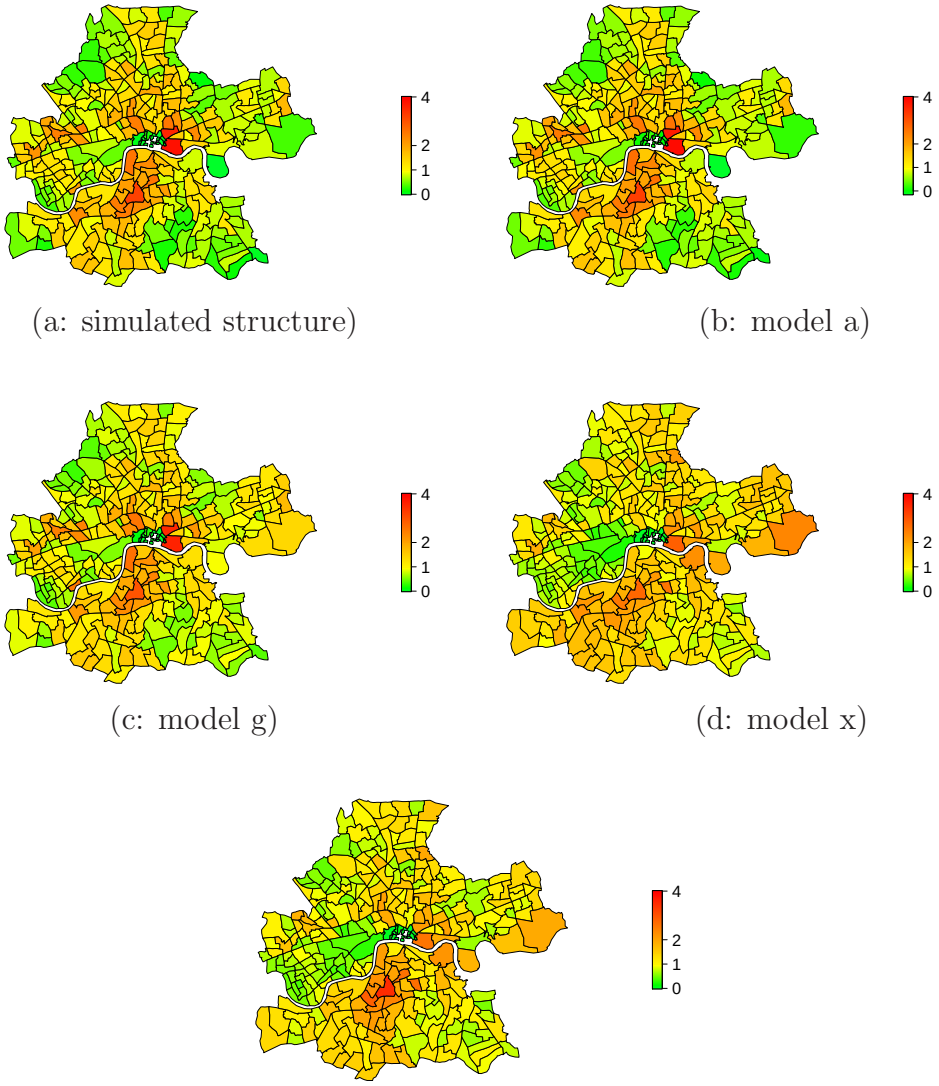
	first run	second run	third run	mean
Aa	315.958	359.796	332.665	336.140
Ab	304.397	324.184	337.455	322.012
Ac	318.813	380.457	326.073	341.781
Ad	332.325	361.980	316.182	336.829
Ae	315.155	340.755	331.089	329.000
Ag	324.607	334.977	312.215	323.933
Ah	326.854	352.175	342.707	340.579
Ai	338.147	347.535	351.436	345.706
Aj	339.557	351.712	317.891	336.387
Ak	365.749	342.209	344.246	350.735
Aw	418.239	393.501	383.341	398.360
Ax	450.744	358.570	386.145	398.486

Table 7.4: summary of DIC in situation A

is due to an overestimation of some low–risk regions at the border. This is even more apparent for data generated according to structure B assuming a high additive influence of benzene. Nevertheless, as differences are small both approaches are suitable. The spatial pattern for selected models is presented in Figure 7.1 (structure A) and Figure 7.2 (structure B) where we use the first run of the simulations. For both structures we achieve patterns close to the generated structures confirming the values of MSE which are smaller than 0.04 for A and 0.30 for B in all Poisson–Gamma models including benzene. Furthermore, both approaches lead to very similar DICs, see Table 7.1 on page 74.

If we do not include benzene in modelling but only latent covariates, estimates differ a lot more compared to previous results. In models w (36 latent covariates) and x (13 latent covariates) we explain lots of variation by latent covariates namely ≈ 95 % for structure A. For structure B we explain 98.0 % in average for Bw and 97.7 % for model Bx. Results are therefore robust to the different number of latent risk sources.

Nevertheless, compared to Poisson–Gamma models assuming additive influence of benzene, we observe higher deviations between generated and modelled risk. This is reflected by an increased MSE of around 0.150 (model Aw), 0.200 (model Ax), 0.694 (Bw), and 0.908 (Bx), compare Table 7.2. These MSEs are the highest among all applied models due to an increased variance when comparing generated and modelled values. Comparison between the results of structure A and B reveals the following relationship: the higher the influence of benzene is in data generation, the higher are the deviations between generated and modelled values and therefore the MSE. Overestimation of risk occurs especially in eastern and southern areas. Nevertheless all wards in the center of London are correctly identified as low-risk wards and the MSE is less than one, see Table 7.2. The increase of MSE is well reflected by the corresponding DICs as given in Table 7.1.

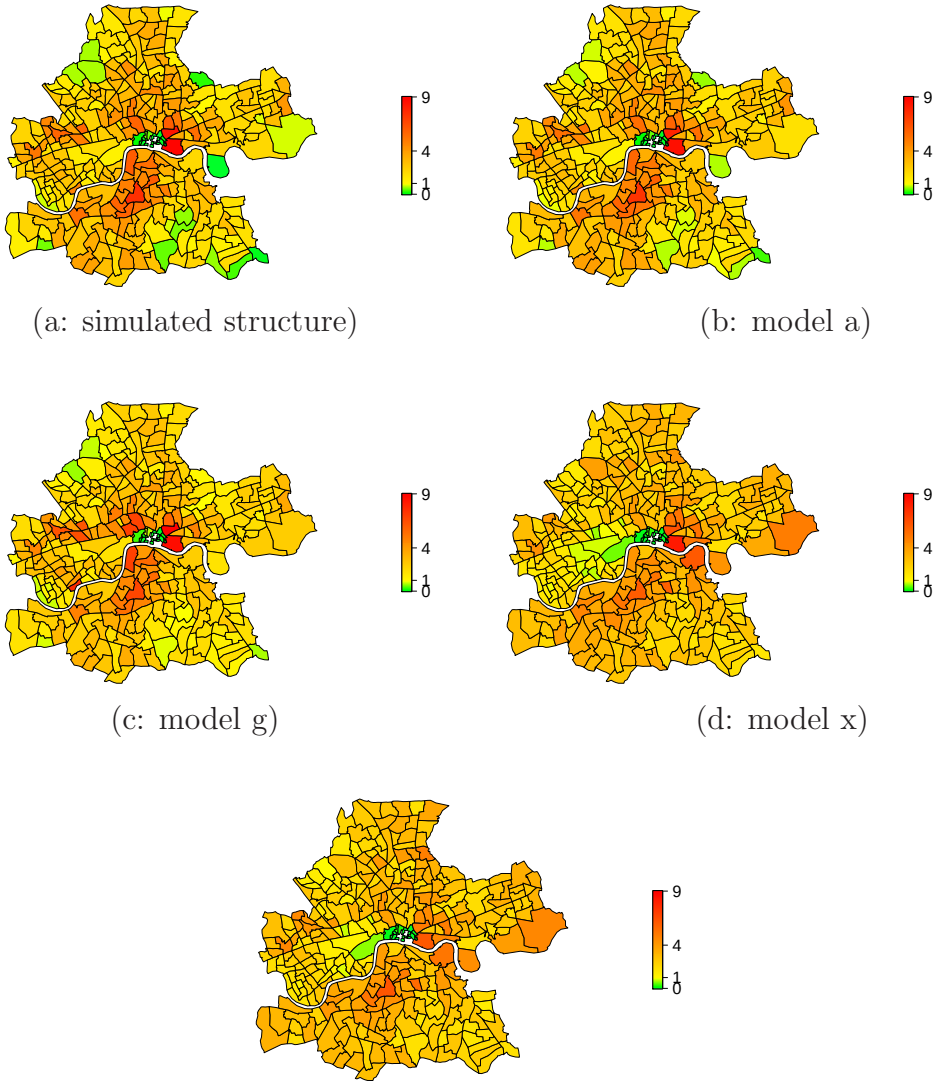


(a: simulated structure) (b: model a)

(c: model g) (d: model x)

(e: model w)

Figure 7.1: Structure A: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 13 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (e).



(a: simulated structure) (b: model a)

(c: model g) (d: model x)

(e: model w)

Figure 7.2: Structure B: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 13 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (e).

7.2 Multiplicative influence of benzene, no latent risk sources

Data analysed in this section are generated following structures M and N including multiplicative influence of benzene but no other covariates. The generated patterns are presented in Figure 7.3 (a) for structure M and in Figure 7.4 (a) for structure N.

Among other models, Poisson–Gamma models assuming benzene to be an excess risk factor are applied. For M, this class performs well and identifies the generated spatial pattern. We observe a similar behaviour for multiplicative Poisson–Gamma models with estimated MSEs of the same order of magnitude, see Table 7.2 on page 75.

For combination Mg generating structure and modelling scheme are identical. The underlying structure is well reproduced as presented in Figure 7.3 (c), the mean of the MSE is less than 0.03. If we introduce latent covariates in either additive or multiplicative models, the estimated spatial pattern remains almost constant, which is reflected by similar MSEs as given in Table 7.2. Additionally, their influence is estimated to be homogeneous over the study area. The standard deviance of each kernel is estimated to be less than 0.0001 reflecting very tight Gaussian kernels in order to minimise their influence in each ward. Similar to structures A and B it is difficult to distinguish between the results of the multiplicative and the additive Poisson–Gamma model, either by MSE or DIC, the latter is presented in Table 7.1 on page 74.

This holds only for M. When we generate a higher number of observations as for structure N, the additive model is not able to estimate more extreme values correctly. The MSE is increased by a factor of around 80, see Table 7.2. As the spatial distribution of risk due to latent covariates is still homogenous, it follows that additive models are not able to produce extreme values and skewed distributions as generated by multiplicative structures. The range of estimated risks is only a subset of the generated ones and multiplicative modelling is required. This is also reflected by the calculated DIC values,

see Table 7.1. Multiplicative inclusion of benzene in Poisson–Gamma models produces satisfying results for structure N for both MSE and DIC as well as for the estimated spatial pattern which are close to the generated structure, compare Figure 7.4 (c) and (a). Here we need to clarify whether additive modelling can be improved further if we do not rely on the restricted implementation of Poisson–Gamma models. The more flexible implementation of Poisson–Gamma random field models is discussed in Chapter 8.

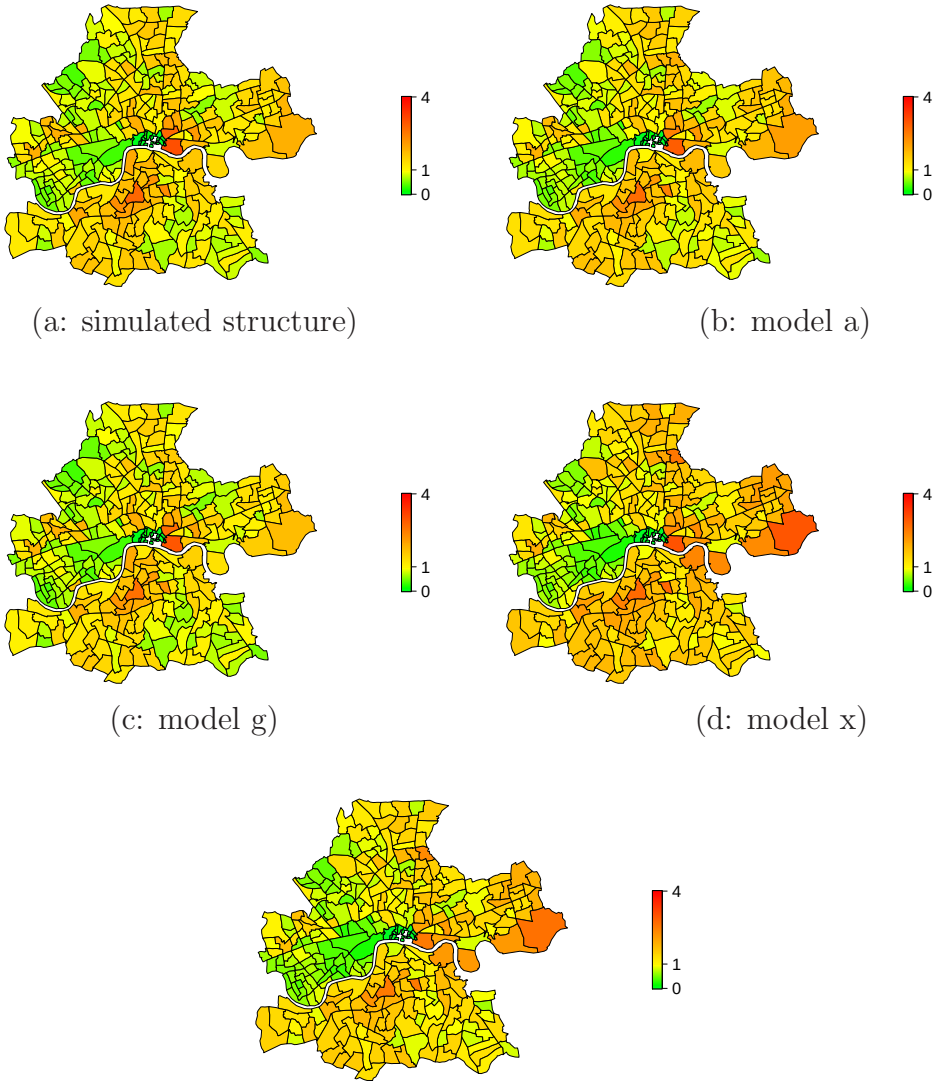
Poisson–Gamma models not taking benzene into account but a number of latent risk sources produce a reasonable fit when the influence of benzene is low as for structure M, but for increased influence of benzene as in N the model has difficulties to identify the generated structure correctly as to be seen in Figure 7.4 (d) and (e). Risk is overestimated for a large number of wards. Additionally, there are regions of rather high generated risk which are underestimated. This is reflected by high MSEs (Table 7.2) as well as highly increased DICs, see Table 7.1. This holds for models w and x involving 36 and 13 latent covariates respectively.

7.3 Additive influence of benzene, one latent risk source

In this section, we summarise the results for data generated following a Poisson–Gamma model with an additive benzene term and a latent risk source generated by a Gaussian kernel located in the south–western part of Inner London.

Here, we observe higher MSEs compared to previously analysed structures. One reason is a higher number of observed cases compared to previous structures but also worsened goodness of fit.

Poisson–Gamma models in the restricted implementation are able to model the spatial structure satisfactorily only if suitable locations for latent risk sources are provided. In Figure 7.5 some results of additive modelling of structure C are given. If too many kernels at fixed locations are provided,

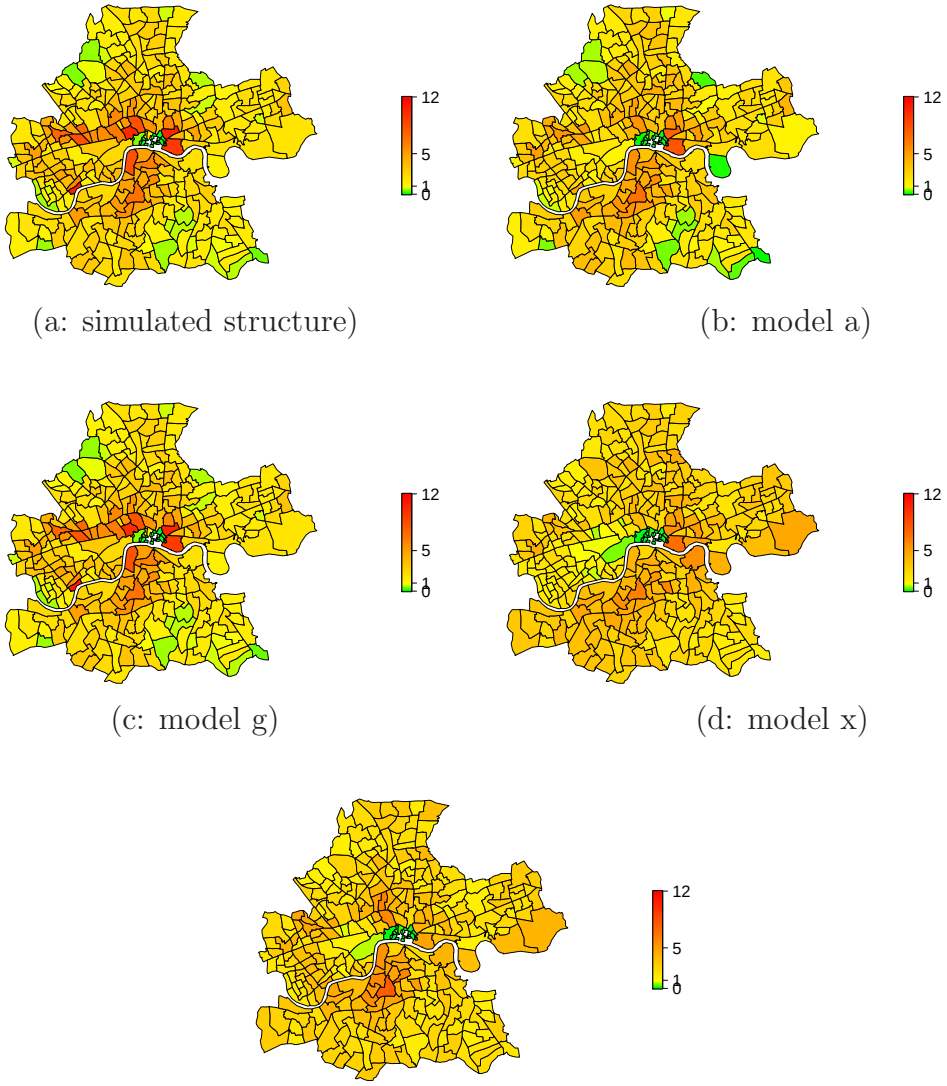


(a: simulated structure) (b: model a)

(c: model g) (d: model x)

(e: model w)

Figure 7.3: Structure M: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 13 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (e).



(a: simulated structure) (b: model a)

(c: model g) (d: model x)

(e: model w)

Figure 7.4: Structure N: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 13 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (e).

the risk of low–risk regions is likely to be overestimated (see Figure 7.5 (e) and (f)). On the other hand, if we do not consider sources located next to the location of the generated risk, the model is not able to reconstruct the spatial pattern (compare Figure 7.5 (d) and the corresponding values of MSE and DIC in Tables 7.2 and 7.1, respectively). Furthermore, if no latent risk sources are involved in modelling, the models fail to identify the pattern correctly as we expect, see Figure 7.5 and the corresponding MSE (Table 7.2) and DIC (Table 7.1). The necessity to implement a continuous version of Poisson–Gamma models concerning latent risk component follows.

The restricted implementation of Poisson–Gamma models still allows for uncertainty in the variance of the Gaussian kernel. This is well estimated if there is a kernel located at a suitable position as for example for model b, compare Table 7.5. Recall that the location of latent covariates is assumed to be fixed here. If that location is too far away like for model c, the standard deviation of the Gaussian kernel is overestimated, while it is underestimated if too many latent risk sources are provided. This can be explained by a necessary maximisation of kernels influence in order to cope with the high risk far away and the idea to minimise each kernels influence in the latter situation similar to Section 7.2. A more flexible approach for variance estimation in combination with uncertainty of kernels location can improve the model, see Chapter 8.

The results concerning the locations of Gaussian kernels can be transferred one–to–one from additive to multiplicative models and to structure D where we present a selection of estimated surfaces in Figure 7.7. Again the spatial risk is only reproduced satisfactorily when we provide Gaussian kernels in the vicinity of the generated risk source. In that case multiplicative and additive modelling of benzene leads to similar results. We also see the necessity to allow for a random location of latent covariates in our WinBUGS's implementation.

Comparison of treating benzene as excess (models a–e) or relative risk factor (model g–k) in Poisson–Gamma models reveals better results for additive modelling. We observe lower DICs and MSEs in general. As for previously

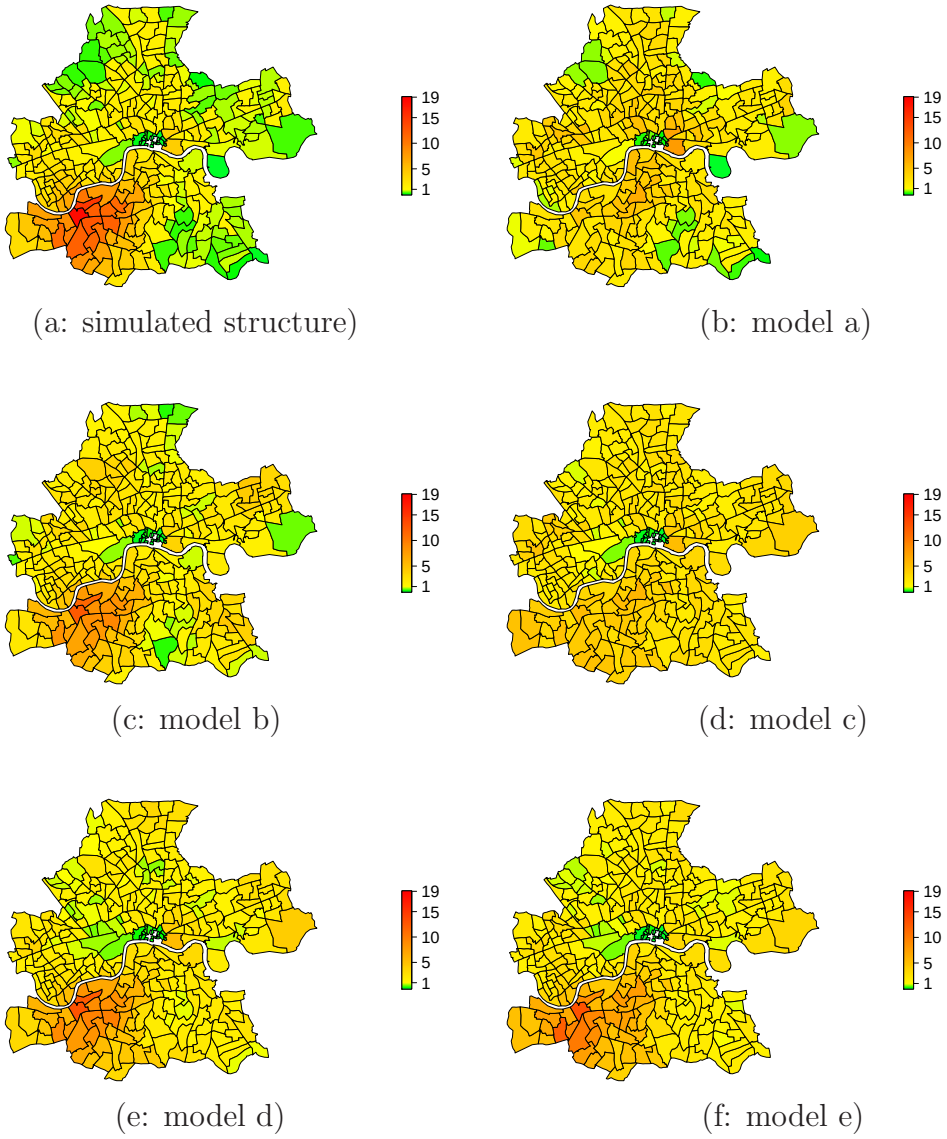


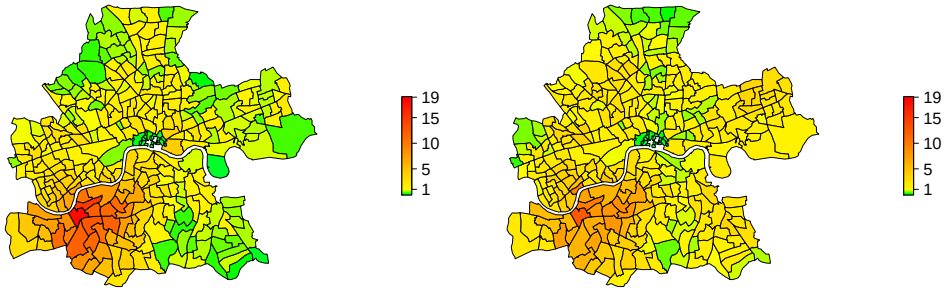
Figure 7.5: Structure C: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using different settings of Poisson–Gamma models with additive influence of benzene: no latent risk source (b), 4 latent risk sources (c), 9 latent risk sources (d), 13 latent risk sources (e) and 36 latent risk sources (f).

model	first run	second run	third run	mean
Cb	2.582	2.601	2.530	2.571
Cc	5.762	5.807	5.922	5.830
Cd	2.575	2.453	2.516	2.515
Ce	1.893	1.840	1.864	1.866
Ch	2.977	3.127	3.157	3.087
Ci	5.804	5.724	5.686	5.738
Cj	2.610	2.599	2.655	2.621
Ck	1.850	1.843	1.882	1.858
Cw	1.898	1.808	1.789	1.832
Cx	2.649	2.592	2.465	2.569

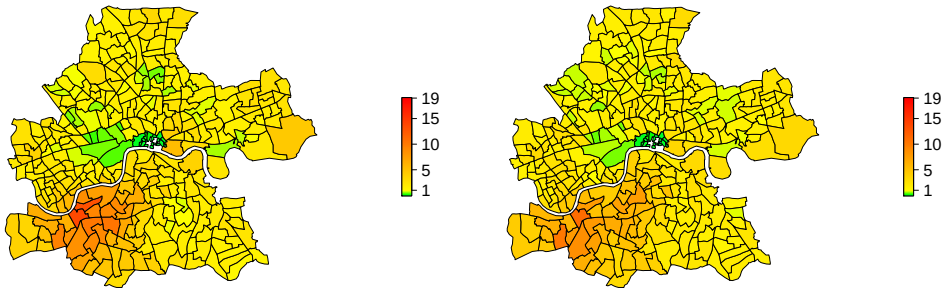
Table 7.5: Estimated variance parameter ρ (in km) for Poisson–Gamma models including a latent Gaussian kernel applied on data generated by structure C.

analysed structures it is difficult to decide whether the influence of benzene should be modelled rather additively or multiplicatively. Nevertheless it has to be noted that the lowest mean of DICs for both structures C and D are achieved for model j which assumes benzene to be a relative risk factor. The corresponding mean MSE is also the lowest for these structures. The reason for this decay can either be the suitability of the model or due to the randomly generated data sets in each run. Application of both covariate interpretations to the same data set as done in Chapter 8 allows for a fairer judgement whether this is a result of the models or of different generated data sets.

When not including a benzene term in Poisson–Gamma models as we do in models w and x, the models are able to identify the latent risk structure, compare Figure 7.6 (b) for structure C and 7.7 (b) for structure D. It has to be noted that the risk for some of the low–risk regions is overestimated by those models. This is reflected by relatively low values of MSE (Table 7.2) and DIC (Table 7.1) compared to those calculated for other models.



(a: simulated structure) (b: model h)



(c: model x) (d: model w)

Figure 7.6: Structure C: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with multiplicative influence of benzene (b), a Poisson–Gamma model without a benzene term and 13 latent risk sources (c), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (d).

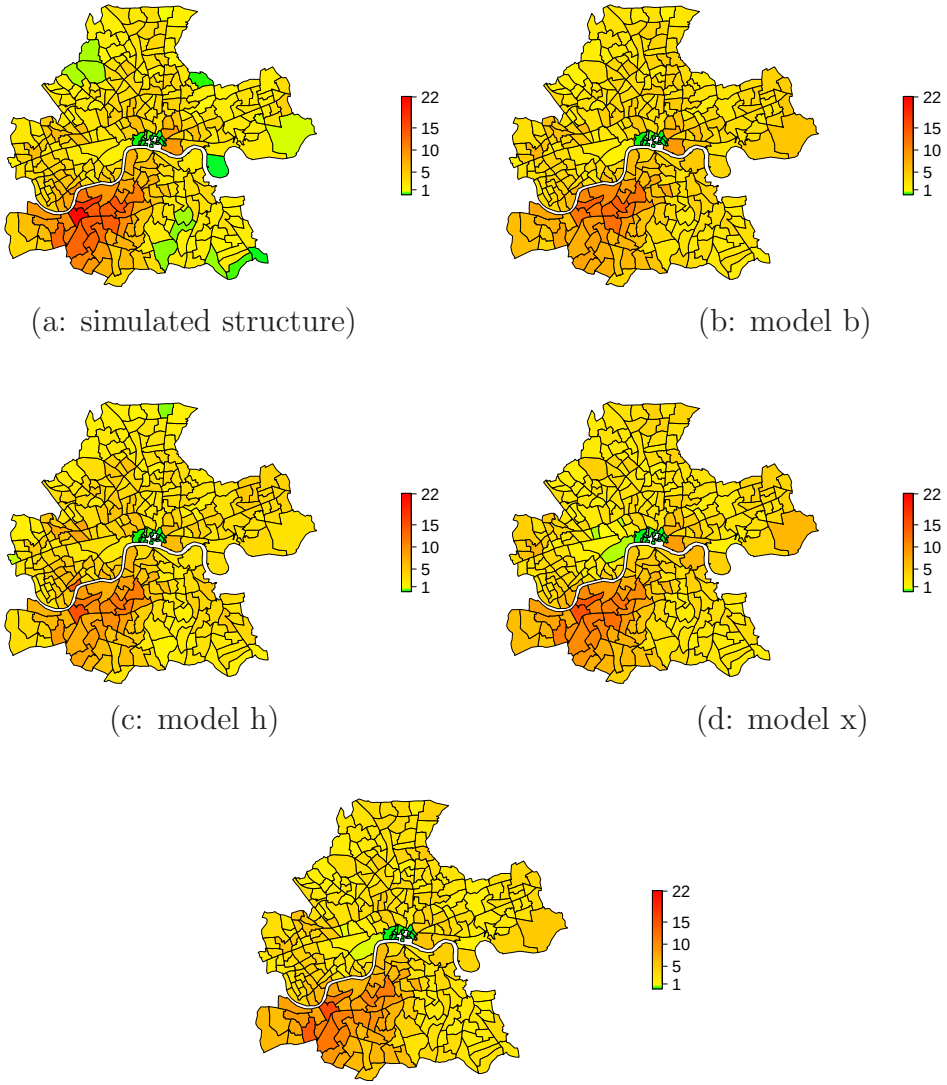


Figure 7.7: Structure D: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene and suitable locations of latent risk sources (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 13 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 36 latent risk sources (e).

7.4 Multiplicative influence of benzene, one latent risk source

Data generated by structures O and P are characterised by moderate risk in most wards, ranging between zero and five observations. Additionally, we observe very high risk in the south–western part of Inner London due to the latent risk generated by a Gaussian kernel. Here, we observe up to 20 (structure O) and 99 (structure P) cases. Especially for structure P this leads to a highly right–skewed distribution of observed risks.

We obtain similar results as described in Section 7.3 regarding the locations of the Gaussian kernels. Modelling such data without involving latent risk as for model a and model g does not produce sufficient results as we see in the DIC values (Table 7.1). Therefore we need to include latent covariates.

If we do so assuming a fixed location we note the necessity of model improvement as already concluded in Section 7.3. A small number of fixed located kernels is too unflexible to model the generated structures sufficiently, but inclusion of a high number of latent covariates is not a suitable solution.

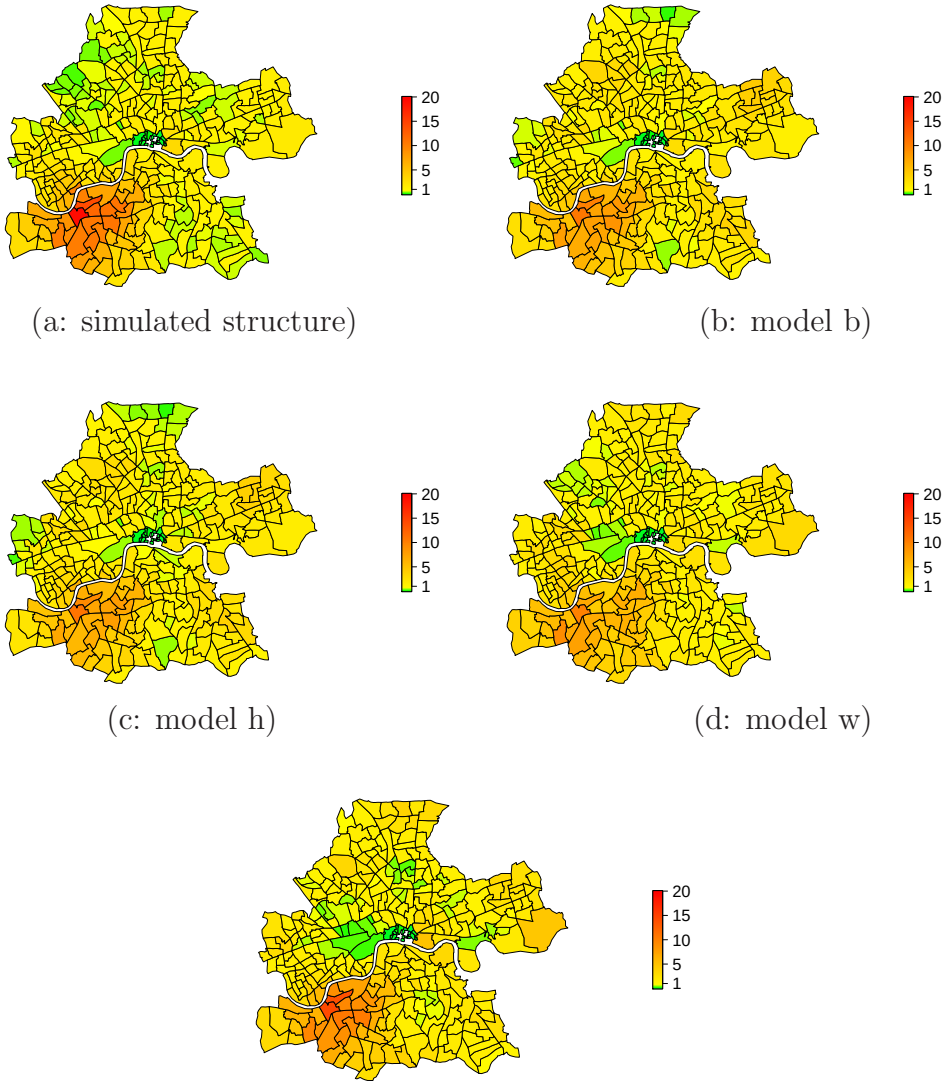
If we use unsuitably located or no latent risk sources at all, the restricted implementation of the Poisson–Gamma model is unable to reproduce the latent pattern adequately and the standard deviation of the Gaussian kernel is overestimated. Again, this causes difficulties in identifying the high–risk region. In Figure 7.9 we show the resulting spatial pattern exemplarily for multiplicative Poisson–Gamma models used on data generated according to structure P.

Additionally, multiplicative Poisson–Gamma models perform much better compared to additive models as those are able to cope with the steep descent of risk. This becomes most clear for structure P. Boxplots as presented in Figure 7.10 (d) reveal that for additive modelling of data generated by structure P we observe the third quantile of the generated values to be even lower than the median of the modelled results. This occurs even for model b which provides a Gaussian kernel close to the point used for data generation.

The corresponding multiplicative model Ph achieves good results.

The inclusion of latent risk sources in multiplicative Poisson–Gamma models at locations similar to those used in model generation leads to satisfying results, although the models are not able to reproduce risks of the same amount as generated by structure P, see Figure 7.9 (f) for example. Additionally we observe a tendency to favour models with a higher number of latent risk sources than required. In the simulations analysed in Chapter 8 we need to clarify whether this is a result of fixed locations of latent kernels, different data sets or if it is a draw back of the model implementation.

Excluding benzene from the model results in a spatial pattern with well identified risk due to the latent risk sources but also leads to overestimation in lots of low–risk regions. Inclusion of a higher number of latent covariates improves the model fit, see Table 7.1.

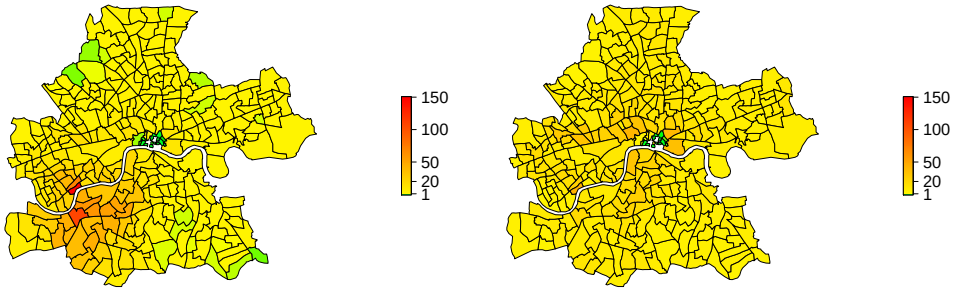


(a: simulated structure) (b: model b)

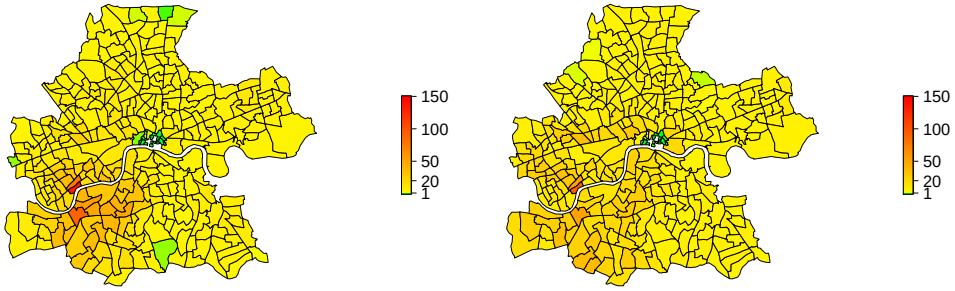
(c: model h) (d: model w)

(e: model x)

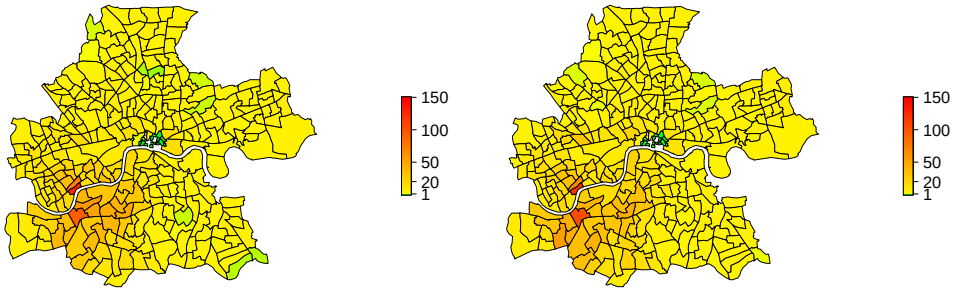
Figure 7.8: Structure O: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model with additive influence of benzene and suitable locations of latent risk sources (b), a Poisson–Gamma model with multiplicative influence of benzene (c), a Poisson–Gamma model without a benzene term and 36 latent risk sources (d), and a Poisson–Gamma model without a benzene term and 13 latent risk sources (e).



(a: simulated structure) (b: model g)

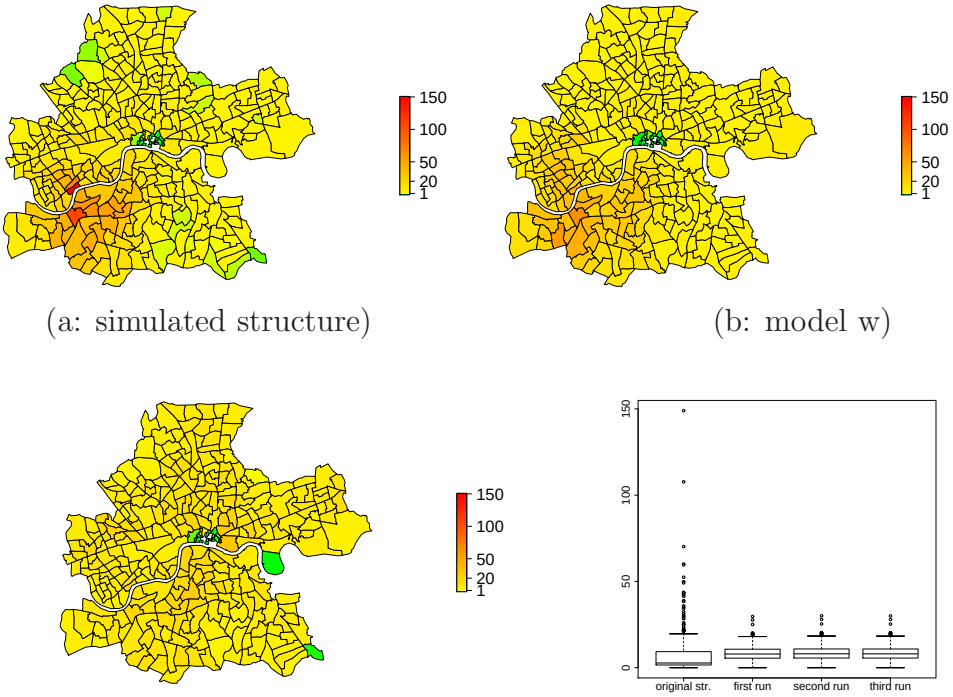


(c: model h) (d: model i)



(e: model j) (f: model k)

Figure 7.9: Structure P: Simulated cases (a) and modelled risk pattern $\Lambda_i E_i$ using different settings of Poisson–Gamma models with multiplicative influence of benzene: no latent risk source (b), 4 latent risk sources (c), 9 latent risk sources (d), 13 latent risk sources (e) and 36 latent risk sources (f).



(a: simulated structure) (b: model w)

(c: model b) (d: boxplot model b)

(e: model x)

Figure 7.10: Structure P: Simulated and modelled risk pattern $\Lambda_i E_i$ using a Poisson–Gamma model without a benzene term and 36 latent risk sources (b), a Poisson–Gamma model with additive influence of benzene and suitable locations of latent risk sources (c), the corresponding boxplot for all three runs (d), a Poisson–Gamma model without a benzene term and 13 latent risk sources (e).

Chapter 8

Simulation results for Poisson–Gamma random field models

The results described in Chapter 7 are calculated using a restricted implementation of Poisson–Gamma models applied to only two of the generated scenarios described in Chapter 6. This does not lead to satisfactory results, especially when latent risk is generated. We therefore extend our WinBUGS implementation to allow for a random location of the latent risk sources as described in Section 3.4.3. We also allow the variances of each kernel to be estimated separately.

This chapter deals with the results of this extended model formulation applied to all structures described in Chapter 6. On page v an overview of the generated structures is given.

In contrast to Chapter 7, we pick only one data set randomly for each of the selected structures. We avoid distortion of results by applying all models to exactly that data set. The changed approach appears necessary as we do not repeat each model combination three times as in Chapter 7.

To each data set, we apply the following Poisson–Gamma random field models:

model f: Poisson–Gamma model with additive influence of benzene;
model m: Poisson–Gamma model with multiplicative influence of benzene;
model o: Poisson–Gamma model with no influence of benzene.

These models are extended by a number of latent risk sources whose locations are unknown a priori. Using our WinBUGS implementation of Poisson–Gamma random fields, we successively include more latent covariates to improve the model fit until the DIC no longer improves.

In spatial epidemiology, there are also other frequently used models, for example the Markov random field (MRF) model (Besag et al., 1991) which also allows for covariates. The model is applied to our data using two different neighbourhood structures. For model v two regions are considered to be neighbours if they share a common border. As London is divided by the river Thames, wards on the North Bank are not considered to be neighbours of wards on the South Bank.

The neighbourhood structure can be extended such that contrary to the above definition, regions only parted by the river are considered to be neighbours. In model z we employ an MRF model with this extended neighbourhood definition. This definition is more adequate for a comparison to Poisson–Gamma models estimating latent risk by Gaussian kernels.

Another group of spatial models are so–called cluster models. One representative of this group of models is the BDCD algorithm (Knorr-Held and Raßer, 2000) described in Chapter 3. This model is chosen as it is of similar complexity as Poisson–Gamma models, leading to a fair comparison of models' performances. Altogether, we employ the following alternative models in addition to Poisson–Gamma models:
model y: BDCD algorithm, wards parted by river Thames are neighbours;
model v: MRF model, neighbourhood structure as used in BDCD;
model z: MRF model, wards parted by river Thames are not neighbours.

In this chapter, we present results for selected structures only, other structures are discussed in Appendix B. These include "low benzene" scenarios as these are close to our real example. In detail, we discuss:

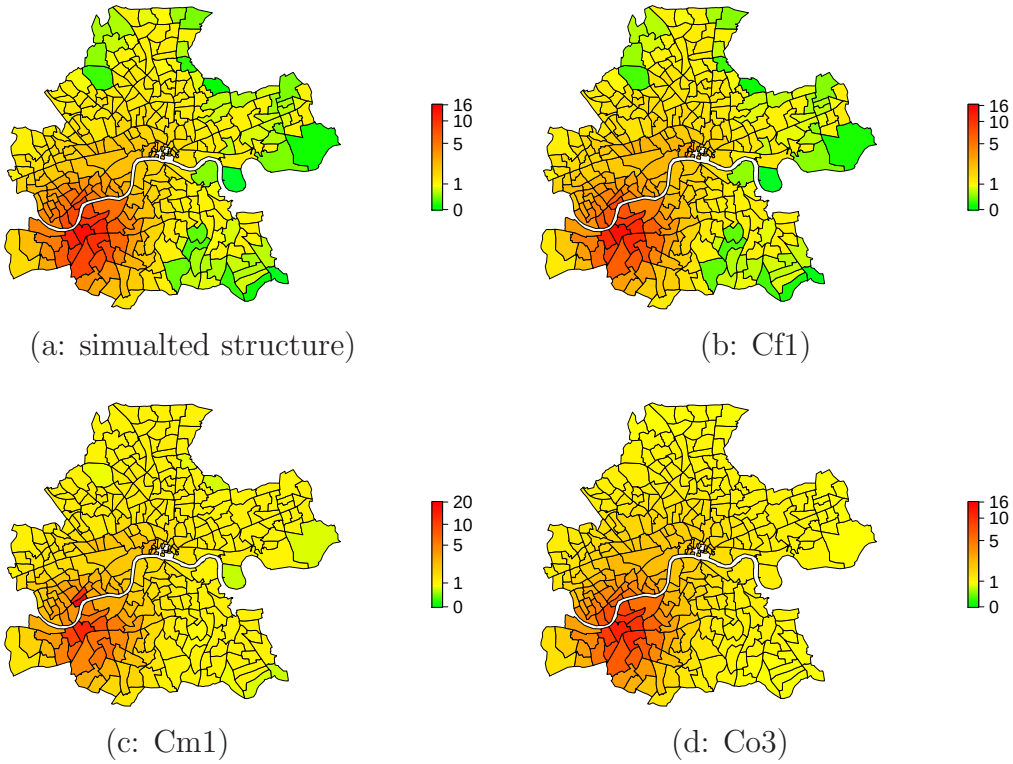
- additive influence of benzene in combination with latent risk represented by a Gaussian kernel (Structure C) in Section 8.1;
- multiplicative influence of benzene in combination with latent risk represented by a Gaussian kernel (Structure O) in Section 8.2;
- additive influence of benzene in combination with a linear spatial trend component (Structure E) in Section 8.3;
- multiplicative influence of benzene in combination with an increased risk in southern areas (Structure S) in Section 8.4;
- additive influence of benzene in combination with an increased risk in cluster regions (Structure I) in Section 8.5.

Results for structures assuming a high influence of benzene lead to similar results as described in Appendix B. Structures assuming an influence of benzene only are discussed in the appendix only as we already achieve reasonable results using the restricted implementation of Poisson–Gamma models.

For all models, we report the calculated DIC as well as the corresponding MSE. We base our model selection on the DIC values only, although we include the corresponding MSE values in our discussion. For structures assuming a low multiplicative influence of benzene and those which are discussed in this chapter we additionally give the effective number of parameters.

The notation referring to the underlying structure and the applied model is as follows: We start with a capital letter corresponding to the underlying spatial structure. A small letter indicating the model follows. When applying Poisson–Gamma models we also deal with different numbers of latent covariates. Here, the number corresponds to the number of latent covariates included in the model.

At the end of this chapter we present an overview of the results of all structures. Furthermore, in Section 8.7 we demonstrate some additional characteristics of Poisson–Gamma models which make them a flexible and powerful tool for model estimation and analysis.



(a: simualted structure) (b: Cf1)

(c: Cm1) (d: Co3)

Figure 8.1: Simualted Λ_i for structure C (a) and results: Estimated spatial pattern of $\widehat{\Lambda}_i$ for the best fitting model Cf1 (b), for model Cm1 (c), and for model Co3 (d).

8.1 Additive influence of benzene, one latent risk source

Structure C assumes an additive influence of benzene in combination with a Gaussian kernel representing latent risk. Both covariates account for a similar number of observations. A spatial plot of the generated structure is given in Figure 8.1 (a). We employ different models on the generated $\Lambda_i E_i$ according to structure C and calculate MSE and DIC for the estimated values $\widehat{\Lambda}_i E_i$ as given in Table 8.1. The MRF model and the BDCD algorithm produce high values for the DIC reflecting non-appropriate model fits. This holds for both neighbourhood structures of the MRF model. Although it includes benzene, the MRF model shows an inferior fit compared to a Poisson–Gamma model

♯ latent factors	0	1	2	3	4
model f	732.7 *2.017* (5.808)	346.2 *6.686* (0.083)	346.7 *7.528* (0.094)	352.1 *15.959* (0.092)	353.1 *18.338* (0.106)
model m	766.8 *1.713* (5.937)	374.7 *6.987* (0.276)	375.0 *7.844* (0.286)	—	—
model o	—	410.1 *7.028* (0.351)	408.0 *17.996* (0.279)	384.4 *17.348* (0.173)	385.2 *20.101* (0.173)
model y	408.0, *69.368* (0.685)				
model v	389.6, *69.246* (0.600)				
model z	387.6, *72.587* (0.586)				

Table 8.1: DIC, p_D and (MSE) values for extended models applied to structure C.

not including benzene, compare the DIC value of 387.6 for model Cz to the one of model Co3 which is 384.4. The BDCD model achieves a DIC of 408.0. Given $n = 310$ wards this fit is not appropriate. The assumption of cluster regions with a constant risk is too restrictive to model the gradual descent generated by the Gaussian covariate. In contrast, MRF models have difficulties to model sharply decreasing risk in neighbouring regions as the rather assume the mean of the surrounding regions for each ward.

In contrast, Poisson–Gamma models show a better performance in general. We conclude that those are more suitable for the underlying structure and explain our findings for those model classes in the following.

For both, the multiplicative and the additive Poisson–Gamma models we see a huge decrease in DIC and MSE from zero to one latent factor by almost 400 points. This gives immense evidence for the existence of at least one latent factor in the underlying data set reflecting the truth.

Best results are achieved if the model is equivalent to the underlying structure, namely models with an additive influence of benzene and one latent covariate. Model Cf1, the most appropriate one, has a DIC value of 346.2

(MSE of 0.083). The spatial risk surface is given in Figure 8.1 (b).

If benzene is modelled multiplicatively, performance is degraded by about 30 points. Model Cm1 shows the best results for this class. Given an MSE of 0.276, the performance of multiplicative models is still acceptable, main characteristics of the generated structure are reproduced by the model (compare Figure 8.1 (c)), although we observe a tendency to overestimate risk in low-risk regions.

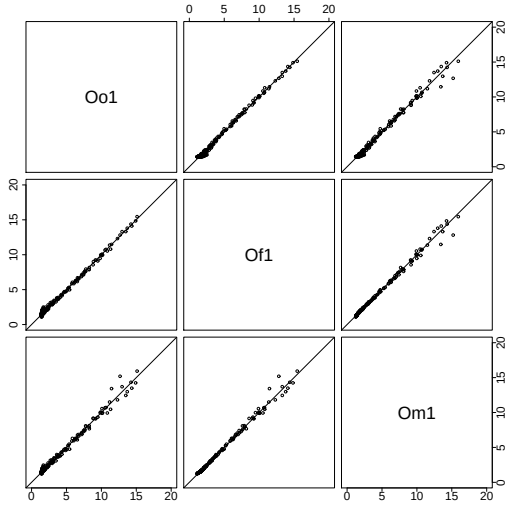
Poisson–Gamma models without benzene are estimated to have a higher DIC value, although the MSE for three latent risk factors (model Co3) is smaller than the one of model Cm1. The resulting spatial risk surface is given in Figure 8.1 (d). It especially lacks in the ability to identify low-risk regions properly, similar as discussed above for model Cm1. In conclusion, for this structure we select the model that corresponds to the generating structure.

Although both, Poisson–Gamma models Cm and log-link MRF models Cv and Cz, include a benzene term, multiplicative modelling does not give convincing results for this structure. We calculate similar DIC values as for Poisson–Gamma models without benzene. The BDCD model not including benzene leads to even higher DIC values.

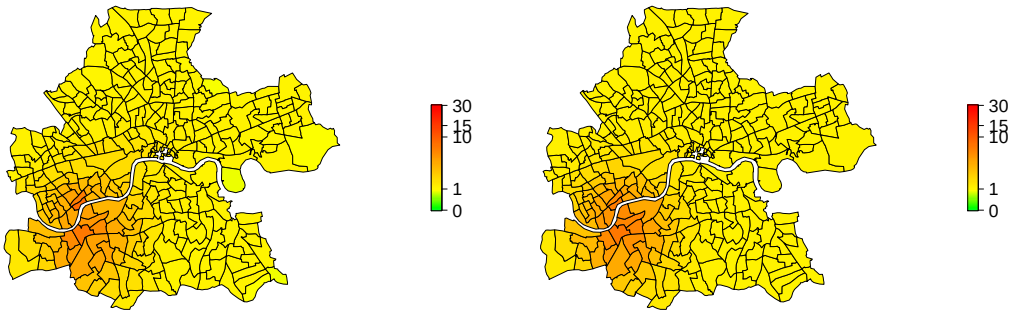
8.2 Multiplicative influence of benzene, one latent risk source

Data generated according to structure O is characterised by multiplicative influence of benzene and the presence of latent risk in the south west of Inner London. These covariates are matched in the proportion of 1:1 for the number of generated cases. A spatial plot of generated Λ_i is given in Figure 8.2 (b), resulting DICs and MSEs of applied models in Table 8.2.

Poisson–Gamma models assuming either a multiplicative or an additive influence of benzene lead to almost identical results. When we do not include any latent risk sources, DIC reaches high values, larger than 800 for both



(a)



(b: simulated structure) (c: Om1)

Figure 8.2: Results for structure O: Scatterplot matrix of the estimated parameters $\widehat{\Lambda}_i E_i$ of the Poisson–Gamma model with multiplicative influence of benzene (Om1), additive influence of benzene (Of1) and no benzene influence (Oo1), all models include one latent risk source (a); simulated structure Λ_i for O (b); rate $\widehat{\Lambda}_i$ of model Om1 (c).

♯ latent factors	0	1	2	3	4
model f	803.8 *2.0* (5.340)	373.2 *6.8* (0.176)	375.5 *9.8* (0.175)	377.3 *13.4* (0.187)	378.5 *16.3* (0.217)
model m	821.4 *1.7* (5.397)	373.5 *6.7* (0.106)	373.5 *7.0* (0.107)	377.6 *13.6* (0.121)	373.9 *7.6* (0.107)
model o	—	375.2 *6.0* (0.222)	378.2 *9.6* (0.226)	379.7 *11.3* (0.232)	380.9 *14.8* (0.273)
model y	398.0, *49.8* (0.860)				
model v	387.0, *76.9* (0.595)				
model z	386.9, *78.9* (0.559)				

Table 8.2: DIC, p_D and (MSE) values for extended models applied to structure O.

models. Both approaches show a substantial drop in the DIC when including one latent risk source in the model, leading to a DIC of 373. Addition of more latent covariates does not improve the model fit anymore. Here, the user is not able to distinguish whether additive or multiplicative modelling is more appropriate.

Similar DICs are obtained for Poisson–Gamma models without benzene (Oo1). A slight increase compared to model classes with benzene is noticeable as the DIC equals 375.2 if including only one latent covariate. Inclusion of more latent sources decreases the model fit. Therefore, the Poisson–Gamma model involving only one latent covariate is favourable.

MRF models and the BDCD algorithm lead to inferior model fits reflected by DIC values greater than 380. For the MRF model no differences due to the different neighbourhood structures are noticeable. Alternative models involve a higher number of effective parameters leading to a higher DIC. Furthermore, estimated MSE values give an indication that model fit is inferior. A reason for inferiority of BDCD is the clustering assumption which contradicts to the smoothly decreasing risk of the Gaussian kernel. This

can also not appropriately be reproduced by CAR terms involved in MRF models. Assumptions considered here do not correspond to steep slopes in decreasing/increasing risk between neighbouring regions.

In order to find the best fitting model we compare the resulting risk surfaces of the Poisson–Gamma models with the lowest DIC values. These are the

- Poisson–Gamma model with benzene as excess risk factor and one latent covariate (Of1);
- Poisson–Gamma model with benzene as relative risk factor and one latent covariate (Om1);
- Poisson–Gamma model with one latent covariate (Oo1).

Resulting values of the risk surface $\widehat{\Lambda}_i E_i$ for each model are plotted against each other using a scatterplot matrix in Figure 8.2 (a). All models result in very similar parameters of the Poisson distribution and therefore in exchangeable risks with a correlation of at least 0.998. This includes the setting used for data generation. A plot of the rates $\widehat{\Lambda}_i$ according to model Om1 is given in Figure 8.2 (c). In contrast to generated Λ_i, estimated $\widehat{\Lambda}_i$ achieve slightly lower values in high risk regions and higher values in low-risk regions. This is reflected by mean and variance of both, Λ_i and $\widehat{\Lambda}_i$. These are 2.947 (Var=33.232) and 3.068 (Var=9.174) respectively. Nevertheless, the MSE of model Om1 equals 0.106 reflecting an appropriate fit of the 310 rates.

Altogether, Poisson–Gamma models satisfy for this structure. Here, alternative covariate interpretations lead to comparable results.

8.3 Additive influence of benzene, linear decreasing trend

For structure E we assume an additive influence of benzene accounting for 330 cases. A similar amount of cases is generated following a linear spatial trend decreasing from north to south. The generated risk surface is plotted

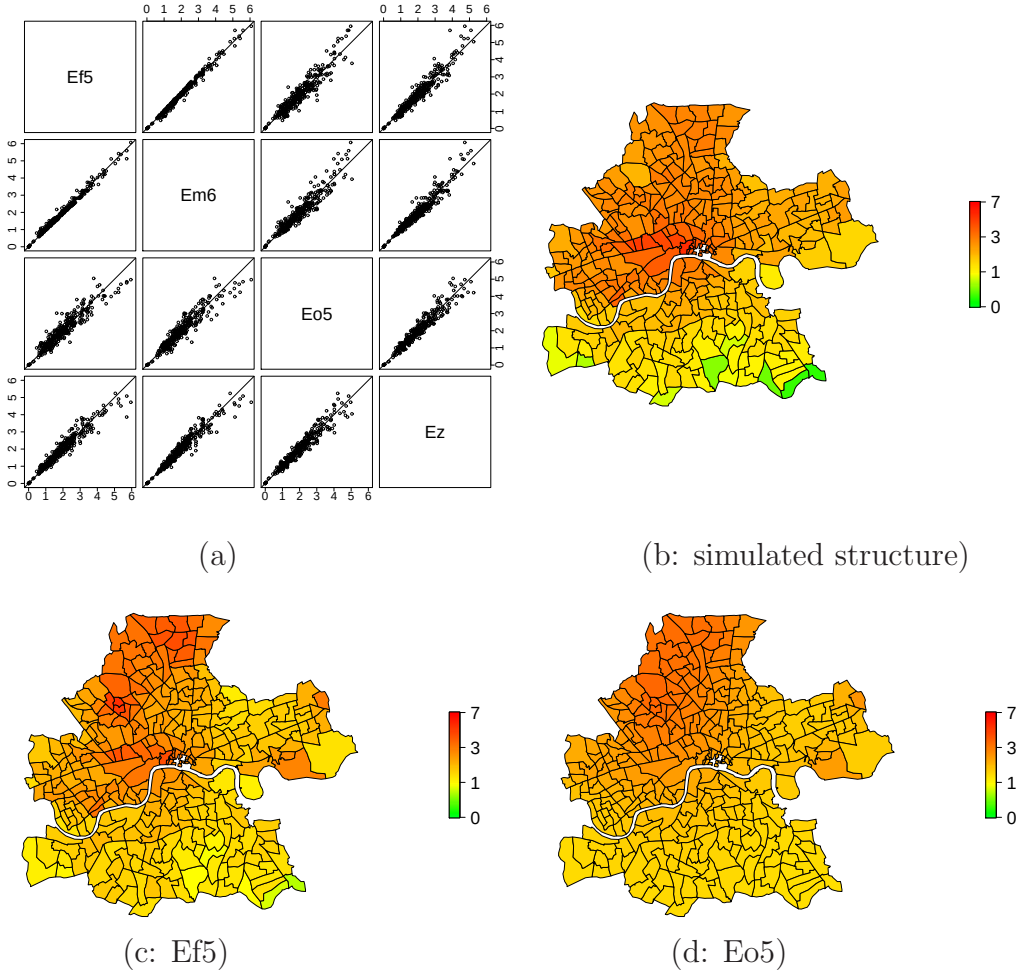
in Figure 8.3 (b). In contrast to structures characterised by latent risk built by Gaussian kernels, the generated latent pattern cannot be decomposed by such kernels. We expect a higher number of kernels to be necessary to reproduce the linear spatial trend.

We employ the models described at the beginning of this chapter and use DIC and MSE for model comparison which are given in Table 8.3. Note the high DIC values for models Ef4 (DIC = 390.5) and Em5 (DIC = 387.1). For the additive model we calculate the amount of risk explained by the different terms for model Ef3, Ef4 and Ef5, see Table 8.4. In contrast to models Ef3 and Ef5 where benzene explains more that 40%, benzene explains only 10.5% of the variation modelled by Ef4. Most of the risk is explained by the baseline. We assume similar problems for model Em5, but — as percentages cannot be calculated for multiplicative models — we cannot check for the percentages of explained risks. An improvement of the model fit is possible using different initial values. We do not pursue this possibility as incorporation of a higher number of latent factors already leads to satisfying results.

Poisson–Gamma models incorporating benzene convince with low DICs of 318.8 for model Em6 and 313.4 for model Ef5. As for previous structures, the additive influence assumed for data generation leads to the most satisfying results for the whole structure. Nevertheless, agreement between both models is high, see the scatterplot matrix in Figure 8.3 (a). Modelling the continuously decreasing trend in combination with benzene as a relative risk factor leads to best results when employing six or seven latent covariates.

The incorporation of five latent covariates achieves best results among Poisson–Gamma models not including benzene as a covariate. The DIC is increased by 25.4 points comparing model Ef5 to Eo5. This is caused by an increased variance between both parameter estimates. These differences are observable across the whole interval, in particular for the high–risk regions in the north of Inner London. Here risk is underestimated for model Eo5, see Figure 8.3 (a).

Estimated rates $\widehat{\Lambda}_i$ of model Ef5 and Eo5 are presented in Figure 8.3 (c) and (d) respectively. Both model the decreasing risk satisfactorily although incorporation of benzene is favourable.



(a) (b: simulated structure)

(c: Ef5) (d: Eo5)

Figure 8.3: Results for structure E: Scatterplot matrix of the estimated parameters $\widehat{\Lambda}_i\, E_i$ of the model with additive influence of benzene and five latent risk sources (Ef5), multiplicative influence of benzene and six latent risk sources (Em6), no benzene influence and five latent risk sources (Eo5) and the MRF model assuming wards across the Thames to be neighbours (Ez) (a); simulated spatial structure Λ_i (b); and spatial risk surfaces $\widehat{\Lambda}_i$ of model Ef5 (c) and model Eo5 (d).

♯ latent factors	0	1	2	3	4	5	6	7
model f	374.0	330.5	330.2	323.9	390.5	313.4	313.9	—
	2.020	*6.958*	*15.224*	*17.899*	*22.249*	*17.053*	*17.876*	
	(0.419)	(0.230)	(0.237)	(0.244)	(2.136)	(0.285)	(0.273)	
model m	376.2	332.5	329.2	324.6	320.6	387.1	318.8	319.0
	1.897	*6.328*	*13.005*	*16.142*	*16.779*	*22.252*	*17.991*	*18.626*
	(0.431)	(0.267)	(0.300)	(0.313)	(0.304)	(2.117)	(0.299)	(0.292)
model o	—	343.2	343.7	340.9	339.6	338.8	338.8	—
		5.967	*11.039*	*14.019*	*15.887*	*17.266*	*18.387*	
		(0.330)	(0.307)	(0.293)	(0.286)	(0.280)	(0.281)	
model y	341.7, *17.484* (0.258)							
model v	333.9, *25.509* (0.149)							
model z	330.8, *32.582* (0.164)							

Table 8.3: DIC, p_D and (MSE) values for extended models applied to structure E.

	Ef3	Ef4	Ef5
expl. risk baseline (%)	39.3	60.0	34.7
expl. risk benzene (%)	42.7	10.5	46.1
expl. risk latent (%)	18.0	29.5	19.2

Table 8.4: Amount of risk explained by baseline risk, benzene and latent term for Poisson–Gamma models including benzene additively and three to five latent factors.

Performance of the BDCD model is inferior to all classes of Poisson–Gamma models with a DIC of 373.2, see Table 8.3. As already discussed for structure C in Section 8.1, the clustering algorithm has difficulties to model a smooth decrease of the risk surface. Hence a high number of cluster centers is required which increases the number of effective parameters and therefore the DIC.

MRF models lead to worse fit compared to Poisson–Gamma models including a benzene term, but perform equally well as Poisson–Gamma models not considering benzene. The influence of the alternatively chosen neighbourhood structures is negligible. The scatterplot matrix of the estimated $\widehat{\Lambda}_i E_i$ includes model Ez. It reflects the close correspondence by the different Poisson–Gamma models to Ez which is confirmed by Pearson's correlation coefficient between Ez and Ef5 that is 0.963.

Although we use an additive setting for model generation we see no immense disadvantage when using benzene as relative risk factor as in model Em6 and Ez. The number of latent covariates involved is able to compensate differences due to different interpretations. Gaussian kernels are a suitable tool to model the latent risk. Sharp decrease is obviously not present as the increased DIC for model Ev in contrast to model Ez indicates. We point out that inclusion of benzene in any of the applied models leads to improved fits although Poisson–Gamma models perform better in general.

♯ latent factors	0	1	2	3
model f	596.3 *1.1* (2.350)	383.6 *6.8* (0.365)	373.0 *12.8* (0.243)	373.1 *15.8* (0.231)
model m	596.4 *1.1* (2.350)	383.7 *6.8* (0.361)	373.3 *13.6* (0.239)	374.0 *15.9* (0.228)
model o	—	383.5 *6.6* (0.373)	370.9 *11.9* (0.254)	373.4 *15.8* (0.238)
model y	347.8, *13.2* (0.140)			
model v	335.7, *6.5* (0.024)			
model z	376.5, *55.1* (0.257)			

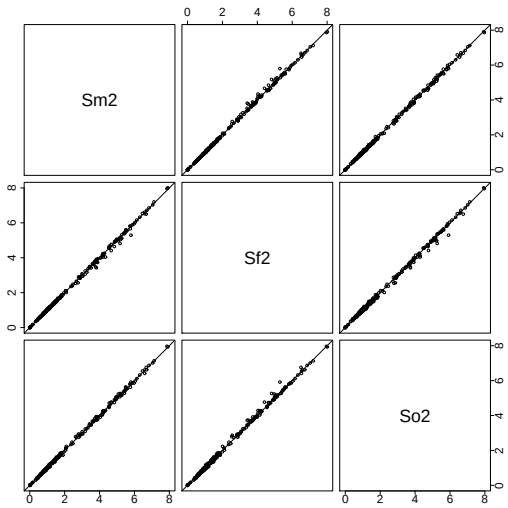
Table 8.5: DIC, p_D and (MSE) values for extended models applied to structure S.

8.4 Multiplicative influence of benzene, plateau trend

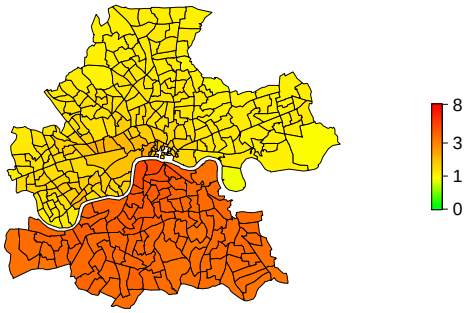
Structure S includes benzene multiplicatively as well as a plateau of higher risk in the wards south of the Thames. Both components account for 330 cases each. For a plot of generated Λ_i see Figure 8.4 (b).

For this structure, the three different classes of Poisson–Gamma models lead to very similar results. Again, inclusion of a single covariate immensely reduces the DIC. We conclude that risk due to other covariates than benzene is present.

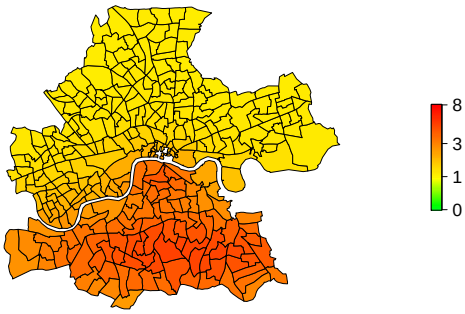
Best results are achieved when two latent covariates are considered. While model Sm2 modelling a multiplicative influence of benzene and two latent risk factors leads to a DIC of 373.3, we achieve a value of DIC = 373.0 for model Sf2 (additive influence of benzene, two latent risk factors) and of DIC = 370.9 for model So2 (two latent covariates only). Results of other numbers of covariates are given in Table 8.5. For a comparison of the estimated rates $\widehat{\Lambda}_i E_i$ of these three models see Figure 8.4 (a). We see a high agreement



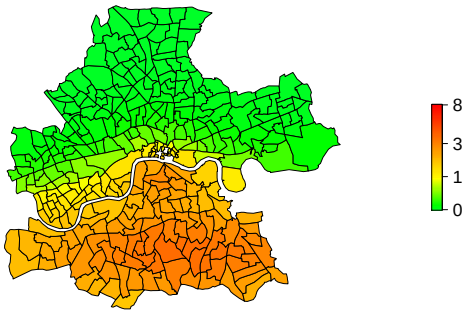
(a)



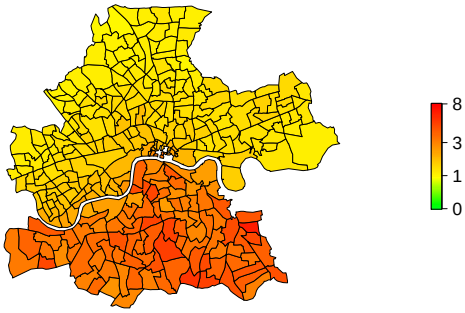
(b: simulated structure)



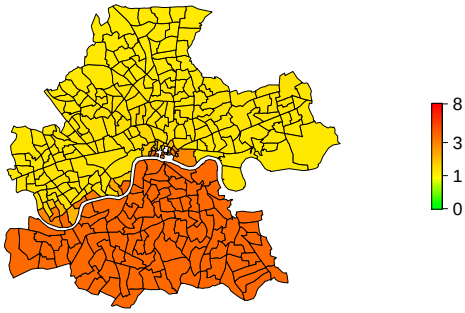
(c: Sf2)



(d: latent term of Sf2)



(e: Sz)



(f: Sy)

Figure 8.4: Results for structure S: Scatterplot matrix of the estimated parameters $\widehat{\Lambda}_i E_i$ of Poisson–Gamma models Sm2, Sf2 and So2; simulated Λ_i of structure S (b); spatial pattern of $\widehat{\Lambda}_i$ estimated by model Sf2 (c) and latent term (d); results by model Sz (e) and by model Sy (f).

between the estimated values of all three models. We pick model Sf2 to give the corresponding plot of the rates surface, see Figure 8.4 (c). The estimated risk of this model can be partitioned to evaluate the fit of the latent term alone as done in Figure 8.4 (d). As we already assume from the $\widehat{\Lambda}_i$s in Figure 8.4 (c), the change in risk at the river is not as abrupt as in the generated data, as we employ Gaussian kernels. Nevertheless we are satisfied with the results given the shape of Gaussian kernels. Poisson–Gamma models allow for other kernels such as Uniform ones or combinations of different types to improve the model fit. Furthermore, the usage of overcomplete dictionaries (Clyde and Wolpert, 2007) is also an alternative to our procedure.

Other models show a better performance. The MRF model Sv treating wards across the Thames not to be neighbours leads to a very low MSE of 0.024 for this structure. The corresponding DIC equals 335.7 which is about 40 points lower than DICs achieved by Poisson–Gamma models. We point out that the corresponding neighbourhood information strongly supports the underlying structure. This indicates that Poisson–Gamma models with more suitable kernels will reach similar values. Comparison to model Sz that has a DIC of 376.5 reveals a high dependency on the chosen neighbourhood structure, especially when comparing the number of effective parameters p_D. Lower values indicating a higher contribution of prior information are found for model Sv supporting the generating structure. DIC of Sz is comparable to that of Poisson–Gamma models. In this setting we also have corresponding situations as modelling of spatially structured and unstructured terms as in Sz corresponds to the employed Gaussian kernels.

A plot of $\widehat{\Lambda}_i$ as estimated by model Sz is given in Figure 8.4 (e). We recognise low risk regions in the north and increased risk in the south. The abrupt change in the rates $\widehat{\Lambda}_i$ at the river is not reproduced. Here, an alternative neighbourhood structure is necessary. Compared to Poisson–Gamma models, the change in model Sz is less smooth. In contrast, risk in the south is less constant but more patchy. By using the alternative neighbourhood structure as for Sv DIC drops to a value of 335.7. Here, we reproduce the sharp risk drop at the Thames.

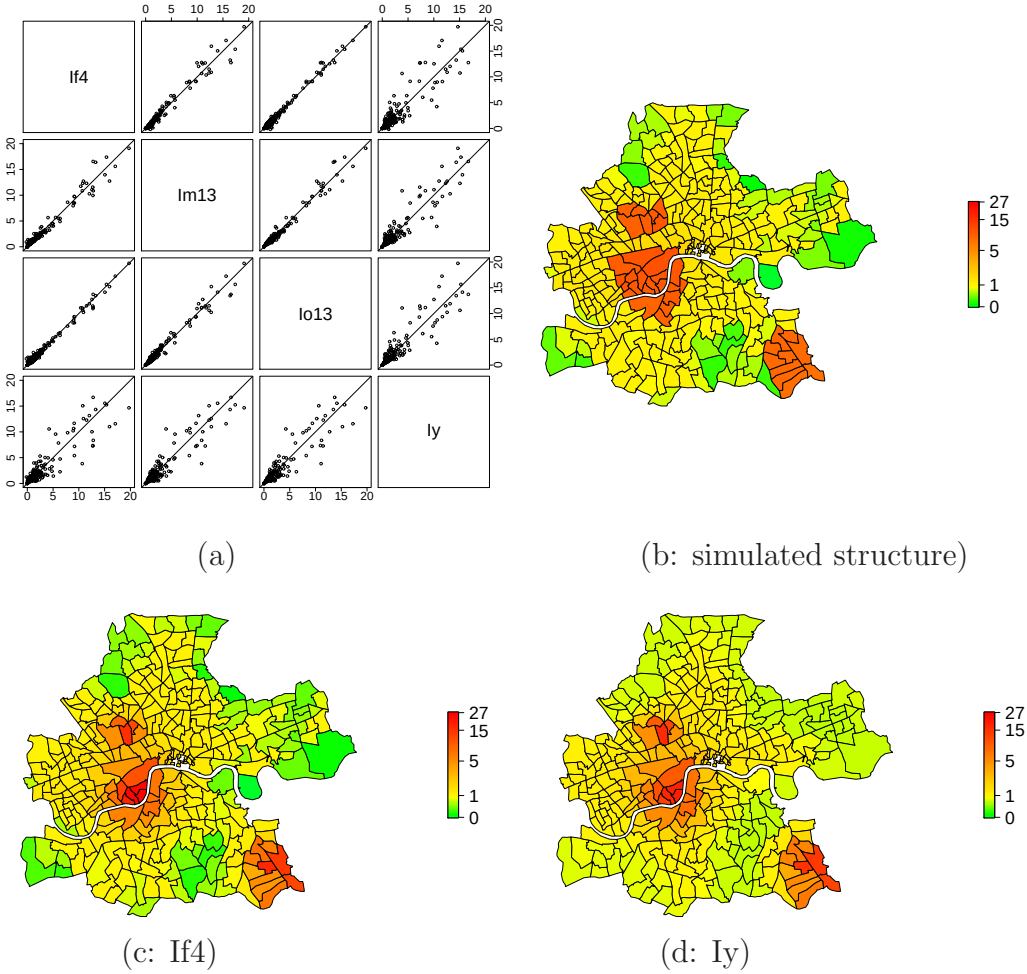
The BDCD algorithm is also able to model the risk differences appropriately. Nevertheless, the boundary between high- and low-risk regions is not estimated to equal the Thames exactly. So some low-risk regions are estimated to have higher risk than they actually have. This leads to an increase in DIC by 8 points compared to model Sv, but lower values than for Poisson–Gamma models. Note that the neighbourhood structure applied by Sy corresponds to those of Sz, i.e., wards separated by the Thames are assumed to be neighbours. Hence, the dropping line is well estimated by the partition model.

Altogether, we see advantages in MRF models when correct neighbourhood information is provided by the user. Similar DICs are gained by the BDCD model where such information is not required. Poisson–Gamma models with our choice of Gaussian kernels show similar performance as MRF models with the standard neighbourhood. Modelling of the sharp drop in the risk is successful by allowing for independent estimation of longitudinal and latitudinal variances, see Section 8.7. For this structure multiplicative modelling of benzene is not essential, the additive model leads to almost identical results.

8.5 Additive influence of benzene, increased risk in cluster regions

Data generated according to a low additive influence of benzene in combination with an increased risk in three selected cluster regions is denoted as structure I. Here, both components account for about 330 generated cases each. The resulting risk surface is given in Figure 8.5 (b).

For data generated according to this structure, we achieve the results presented in Table 8.6 when applying our selected models. Estimation of Poisson–Gamma models Io13 and Im13 stops with inclusion of any further latent kernels as this procedure does not lead to sufficient decrease of the DIC. For the resulting models, namely Im13 assuming multiplicative influence of benzene and 13 Gaussian kernels (DIC = 401.1) and Io13 modelling risk by 13 latent covariates only (DIC = 410.7), we compare the estimated $\widehat{\Lambda}_i E_i$ in a scat-



(a) (b: simulated structure)

(c: If4) (d: Iy)

Figure 8.5: Results for structure I: Scatterplot matrix of the estimated parameters $\widehat{\Lambda}_i\, E_i$ of the model with additive influence of benzene and four latent covariates (If4), multiplicative influence of benzene and 13 latent covariates (Im13), no benzene influence and 13 latent covariates (Io13), BDCD model (Iy), and the MRF model assuming wards across the Thames to be neighbours (Iz) (a); simulated spatial structure Λ_i (b); and spatial risk surfaces $\widehat{\Lambda}_i$ of model If4 (c) and model Iy (d).

♯ latent factors	0	1	2	3	4	5	6
model f	1156.6	1011.0	447.7	393.1	376.7	378.5	—
	2.0	*222.9*	*10.1*	*14.5*	*16.7*	*21.1*	
	(11.277)	(5.518)	(3.519)	(2.841)	(2.511)	(2.339)	
model m	1130.9	764.4	484.9	412.8	409.7	407.1	406.9
	2.0	*5.9*	*10.7*	*14.7*	*15.8*	*17.7*	*21.5*
	(11.199)	(8.060)	(3.743)	(2.958)	(2.653)	(2.551)	(2.361)
	7	8	9	10	11	12	13
	405.8	405.2	404.5	403.0	402.4	401.6	401.2
	23.2	*26.4*	*28.2*	*30.2*	*31.0*	*31.8*	*32.3*
	(2.263)	(2.150)	(2.074)	(1.960)	(1.940)	(1.960)	(1.878)
model o	—	1212.5	482.6	440.1	437.7	428.7	423.3
		219.0	*9.5*	*14.3*	*26.6*	*30.6*	*34.2*
		(6.090)	(3.664)	(3.064)	(2.729)	(2.540)	(2.435)
	7	8	9	10	11	12	13
	418.7	415.5	413.8	412.7	411.6	411.1	410.7
	36.0	*36.9*	*37.7*	*38.2*	*38.7*	*39.1*	*39.5*
	(2.415)	(2.391)	(2.375)	(2.371)	(2.365)	(2.362)	(2.359)
model y	381.0, *60.8* (0.633)						
model v	415.6, *115.8* (1.509)						
model z	412.1, *115.1* (1.505)						

Table 8.6: DIC, p_D (MSE) values for extended models applied to structure I.

terplot matrix as given in Figure 8.5 (a). We include the Poisson–Gamma model including benzene additively in our comparison. Here four kernels are required to achieve a minimum value of the DIC which is 376.7 in our case. Additionally we compare the results of the clustering model Iy. This model satisfies with a DIC of 381.0.

MRF models have difficulties to model the sharp decrease in risk in combination with an additive influence of benzene reflected by high DICs. These are greater than the values achieved by all settings of Poisson–Gamma models. As these models do not lead do satisfying DIC values we do not consider them any further.

The scatterplot matrix (Figure 8.5 (a)) shows a high agreement between results of all Poisson–Gamma models. The reason is probably the comparable structure of the model although different interpretations of benzene and different numbers of kernels are considered. The BDCD model shows differences across the whole interval when compared with any of the Poisson–Gamma model formulations.

We take a closer look at the resulting risk surfaces of model If4 and Iy. We choose model If4 among Poisson–Gamma models as this satisfies by the lowest DIC of this group. Figure 8.5 (c) and (d) present the surfaces of $\widehat{\Lambda}_i$. By visual comparison, identification of cluster regions leads to similar results. Estimated values do not necessarily correspond to the generated ones which lie in the interval of $\Lambda_i \in [0, 18.404]$. For model If4, we overestimate the risk as $\widehat{\Lambda}_i \in [-0.118, 26.636]$. Note that negative values are estimated by WinBUGS, although these are impossible for Poisson distributed values. We take them as estimated. For model Iy the estimated values lie in the interval of $\widehat{\Lambda}_i \in [0.520, 15.688]$, i.e., this model rather underestimates the risk in high risk regions. In contrast, low risk regions are overestimated.

The estimated surface of clusters differs from the constantly increased risk that is generated. Although we expect a clustering model to detect such structures, model Iy has similar difficulties as the Poisson–Gamma model If4 employing four smoothly decreasing Gaussian kernels for latent terms. Here we expect an improvement of the model fit when employing alternative

kernels such as Uniform ones.

Low risk regions are reasonably reproduced by the Poisson–Gamma model, whereas the clustering model tends to oversmooth. Although both models lead to similar DICs we clearly favour the Poisson–Gamma model with its potential model improvements by alternative kernels.

Due to the DIC values of Poisson–Gamma models we prefer additive modelling of the covariate as assumed for data generation. Multiplicative modelling decreases the model fit, although this model formulation is preferable over ignoring benzene's influence as in model Io13.

Altogether, for this structure we prefer the Poisson–Gamma model assuming an additive influence of benzene as in data generation. We see a huge potential for model improvement when other than Gaussian kernels, e.g. Uniform ones, are considered.

8.6 Summarised results for all structures

In this section we summarise the results of the simulation study for all structures. Table 8.7 gives the DICs achieved for the different models. For each structure we report the value for the best model, in combination with the differences between this and the DIC of every other model.

Spiegelhalter et al. (2002) suggest that differences in DIC less than two show equal support of both models and are worth considering while differences larger than seven can be interpreted as inferior support of the model by the data.

In contrast to this recommendation we observe negligible risk differences for two models while the DICs differ in less than seven to eight points, see for example the scatterplot matrices in Figures 8.3 (models Ef5 and Em6, difference 5.4) and Figure B.4 (models Ff5 and Fm7, difference 7.7). For many models, smaller DIC values almost reproduce the intersecting line. Furthermore, the goal of our study is not to find a single best fitting model but to analyse the abilities to reproduce the risk surface.

	model f	model m	model o	MRF model	BDCD
Structure A	339.9	+ 7.6	+ 23.3	+ 11.3	+ 33.3
Structure B	312.2	+ 14.6	+ 49.8	+ 15.7	+ 60.2
Structure C	346.2	+ 28.5	+ 38.2	+ 41.4	+ 61.8
Structure D	326.8	+ 31.2	+ 39.8	+ 41.6	+ 68.7
Structure E	313.4	+ 5.4	+ 25.4	+ 17.4	+ 28.3
Structure F	330.4	+ 7.7	+ 27.1	+ 4.0	+ 29.8
Structure G	+ 21.7	+ 33.8	+ 42.5	+ 25.0*	345.6
Structure H	362.0	+ 8.2	+ 40.1	10.2*	+ 27.9
Structure I	376.7	+ 24.4	+ 34.0	+ 35.4	+ 4.3
Structure J	+ 13.1	+ 68.1	+ 58.4	+ 34.3	397.6
Structure M	321.9	+ 0.9	+ 4.0	+ 0.5	+ 6.1
Structure N	+ 47.1	340.5	+ 87.1	+ 24.1	+ 80.1
Structure O	373.2	+ 0.3	+ 2.0	+ 13.7	+ 24.8
Structure P	+ 83.9	341.1	+ 137.3	+ 59.8	+ 141.1
Structure Q	+ 1.4	338.3	+ 4.4	+ 14.7	+ 7.2
Structure R	+ 60.0	327.7	+ 92.0	+ 35.3	+ 115.7
Structure S	+ 25.2	+ 25.5	+ 23.1	+ 28.7*	347.8
Structure T	+ 39.2	+ 17.8	+ 52.1	+ 54.6*	386.3
Structure U	+ 34.2	+ 34.2	+ 58.7	+ 68.1	358.9
Structure V	+ 28.9	+ 21.1	+ 87.1	+ 43.6	389.4

Table 8.7: Summary of the DICs achieved for different structures, the MRF model refers to model z where * marks structures where model v and model z lead to differences larger than seven.

Keeping in mind the recommendation of Spiegelhalter et al. (2002) we colour differences less than seven green in Table 8.7 representing models with a similar support. Models leading to much larger differences should clearly not be considered as alternatives to the "best" model. Hence we colour large values red. As a breakpoint for models that should not be considered we set twice the breakpoint between the best model and those acceptable as alternatives, i.e., differences larger than 14. Intermediate values are coloured in orange.

For the MRF model, different settings of the neighbourhood structure lead to

similar results for most structures. We therefore report only one value in Table 8.7. As reference, we choose model z as the corresponding neighbourhood structure is more adequate for a comparison to the Gaussian kernels applied for Poisson–Gamma models which also ignore the river as a boundary for modelled risk. Models where the support of both neighbourhood structures differs by more than seven points are marked by a " * ". These correspond to those structures where we generate a plateau of increased risk in southern Thames' wards, i.e., structures G/H/S/T.

Our main conclusions from the simulation study are as follows:

- Additive structures usually favour an additive model while multiplicative structures lead to a multiplicative model as best fitting model. The different interpretations of the covariate are not exchangeable in general. This holds especially for structures assuming a "high" influence of the covariate.

- When not considering benzene in the model the DIC is increased which guides us to include necessary covariates. This holds for both, Poisson–Gamma model o and BDCD model y. The increase in DIC of Poisson–Gamma models tends to be smaller compared to the BDCD model.

- The number of involved latent covariates in Poisson–Gamma models is typically small. It tends to be higher when benzene is not considered by the model.

- For MRF models, different neighbourhood structures lead to negligible differences less than seven in the DICs of most structures. Structures generating an abrupt change in risk at the Thames (G/H/S/T) favour the neighbours to be parted by the river.

- In our choice of compared models, MRF models assuming a neighbourhood structure that ignores the Thames lead to spatial dependencies most similar to multiplicative Poisson–Gamma models estimating latent risk by distance–based decrease of Gaussian kernels. Nevertheless, their results are inferior.

- Clustered structures I/J/U/V lead to lowest DICs when modelled by the BDCD model. This model also shows good results for structures assuming a plateau of increased risk south of the Thames (G/H/S/T). For other structures their deviation from the best fitting model is rather high.
- Although we do not observe differences due to different neighbourhoods for most structures we strongly recommend to apply different ones for sensitivity analysis.

From the results of the BDCD model concerning clustered structures as well as MRF models assuming neighbourhood structure v we expect an improvement of the model fit of Poisson–Gamma models when allowing for alternative kernels. For small cluster regions Uniform kernels probably provide the best alternative. Structures that assume risk to drop at a boundary such as the Thames probably require more complex kernels, half–Gaussian ones provide a possible solution. These correspond to neighbourhood v in MRF models. The advantage of Poisson–Gamma models is the estimation of the breakpoint of a half–Gaussian kernel as boundary in contrast to MRF models where the neighbourhood structure is required as model input. Working with different kernel functions presented as overcomplete dictionaries as suggested by Clyde and Wolpert (2007) provides a suitable extension of the model class. Furthermore, Poisson–Gamma models can easily be partitioned into their individual terms and are easier to interpret. They also give a guideline for possible clusters as discussed in the next section.

8.7 Identification of high-risk regions

As we have already seen in previous sections, Poisson–Gamma models provide a very flexible and powerful tool for the identification of the underlying structure. Models favour covariate information as either excess or relative risk factor corresponding to the underlying setting and therefore help to identify the correct pattern. The user can easily determine whether additional covariates should be included or do not provide any further information.

Beside mapping the risk surface alone Poisson–Gamma models are useful for identification of unknown risk factors. It is possible to map the locations of the latent random field such as in Figure 8.6. We give the locations of the Gaussian kernels for best fitting Poisson–Gamma models for structures discussed above, namely

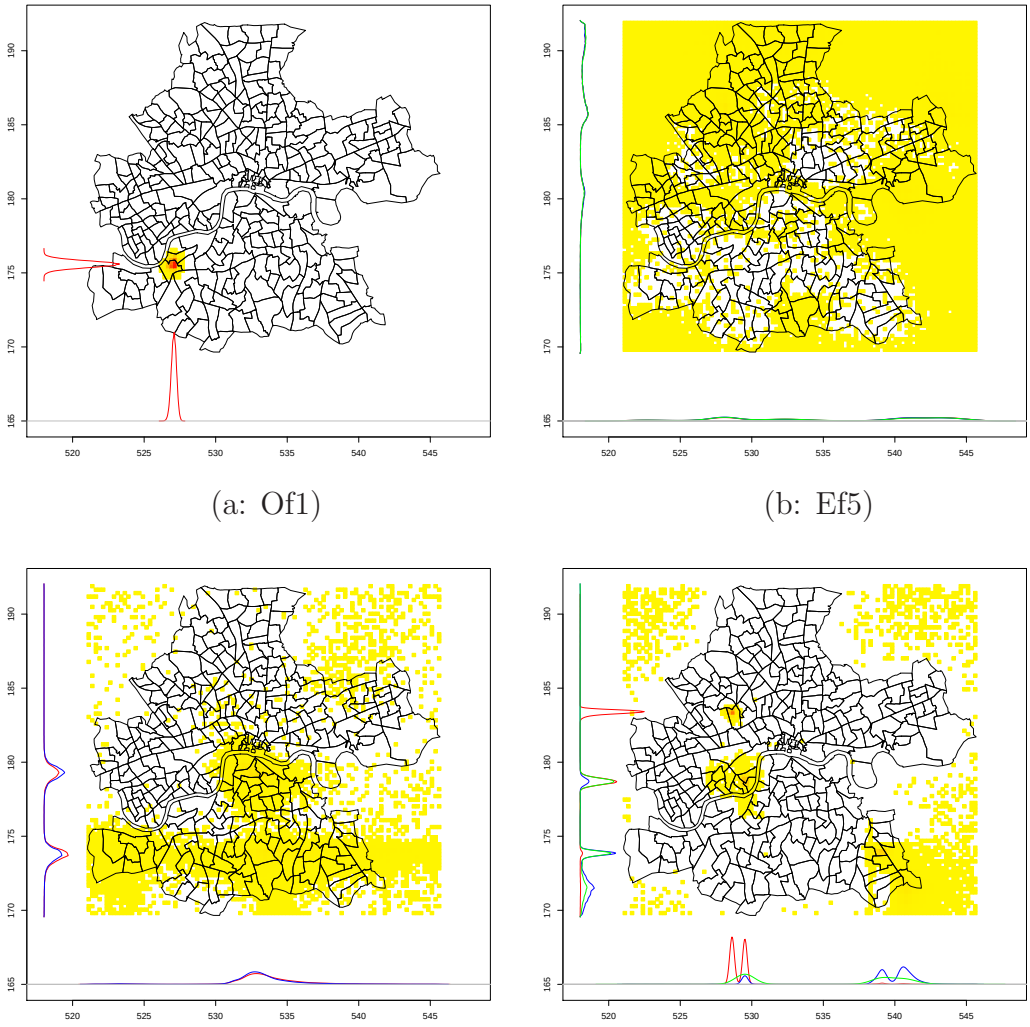
- for generation of one latent covariate: model Of1;
- for generation of smoothly decreasing risk: model Ef5;
- for generation of abrupt changing risk at the Thames: model Sf2;
- for generation of increased risk in three clusters: model If4.

The plots give estimated locations of the kernels. The frequency is represented by the colour of each cell with red values indicating frequently chosen locations. On the axes we give density estimates of the location of the Gaussian kernels for longitude and latitude. For all plots the same scale is used that ranges between 0 (yellow) and 0.2 (red).

In the first situation, the risk surface can be partitioned into Gaussian kernels. The location we have chosen for data generation in structure O is reproduced exactly by the model. In a real application we can identify the location of a point source.

The second setting assumes smoothly decreasing risk from north to south. This is also reproduced by the location of the Gaussian kernels. The location of any of the kernels is more likely in the north of the region as a higher density indicates. For longitude we see no such decreasing line which supposes no risk differences in this direction. A closer inspection of the according variances helps to figure out the smoothly decreasing risk. We observe only small peaks of the density compared to other settings.

Higher peaks in density estimates reveal sharp risk differences such as for structure S that has an increased risk in the south. Both kernels are located in the southern area, they are nearly never in any northern region. For the longitude all locations are chosen with a peak in the center. In our setting, we assume an independence of both directions making such a constitution



(a: Of1) (b: Ef5)

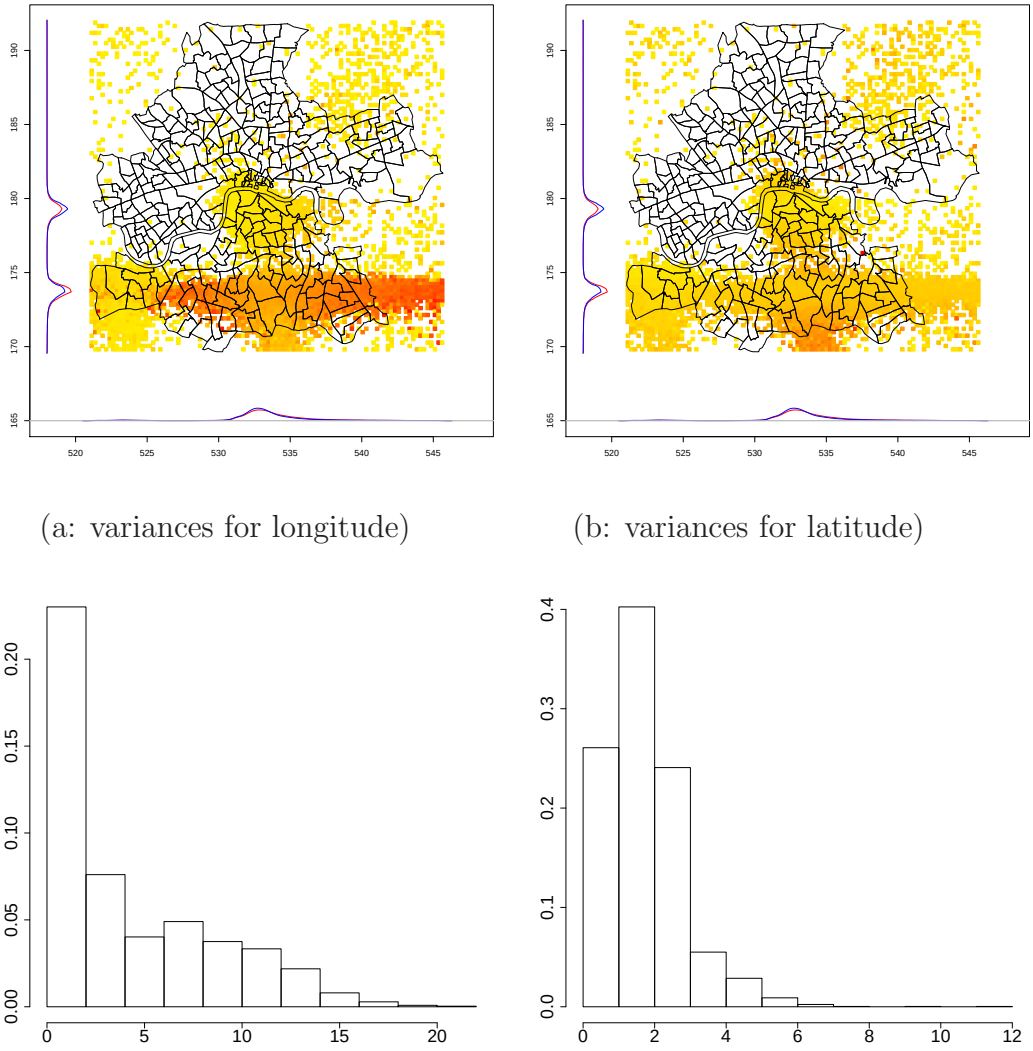
(c: Sf2) (d: If4)

Figure 8.6: Latent random field for Poisson–Gamma models Of1 (one latent Gaussian kernel in data generation), Ef5 (smoothly decreasing risk assumed), Sf2 (increased risk in southern wards assumed) and If4 (increased risk in three clusters assumed).

necessary: One kernel 'increases' the risk in all wards up to the least northern increased one while an additional kernel models latent risk in the remaining southern wards. The peak in longitudinal direction corresponds to the latter one.

The fourth situation is characterised by three clusters of increased risk. These are identifiable using the information of the cluster centers in Figure 8.6. Obviously, the size of cluster in the south-eastern area is too large to be modelled by a single Gaussian kernel given the sharp risk drop off. Here two kernels are used. Note that the model is able to separate the increased risk in the two more central clusters lying close together.

Other useful interpretation of the intensity of the influence corresponds to the variance of the kernel given its location which is demonstrated for structure S, model Sf2, in Figure 8.7. High values indicate a slowly decreasing and smooth influence of the latent covariate, while low values correspond to a small area of increased risk. This allows to model sharp decreasing risk or elevated risk in small areas. The mean variance in each of the cells varies in the interval of $[0.08, 21.19]$ for longitudinal direction and $[0.07, 11.76]$ for latitudinal direction. The scale of each plot in Figure 8.7 varies accordingly. For longitudinal direction, we estimate low variances for the centrally located kernel, while southern ones are characterised by larger variances. In contrast, variances in latitudinal direction are estimated to be smaller. Higher values are mainly observed at the transition between both kernel zones leading to a sharp decrease of the kernels. Histograms as given in Figure 8.7 (c) and (d) also show the different behaviour of longitudinal and latitudinal variances. This example also shows the necessity to allow for independent variances in both directions. We give another example for such an analysis for the leukaemia data set in Chapter 9.



(a: variances for longitude) (b: variances for latitude)

(c: histogram (longitude)) (d: histogram (latitude))

Figure 8.7: Results for model Sf2: Variances in longitudinal (x) and latitudinal (y) direction, red colour indicates high values, yellow colour low variances; the margin represents the estimated location of the Gaussian kernels for both directions. Subfigures (c) and (d) give according histograms.

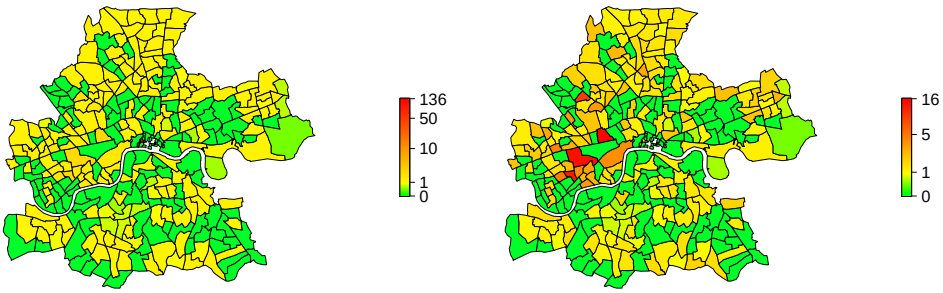
Chapter 9

Results for leukaemia data

In Chapter 8 we demonstrate the abilities of Poisson–Gamma random field models, MRF models and the BDCD algorithm to identify an underlying spatial structure. In this chapter, we apply these models to smooth the calculated SMRs for the leukaemia data set described in Chapter 2. By consideration of covariates, we analyse the dependence of observed leukaemia counts on atmospheric benzene emissions and the Carstairs deprivation index. Benzene emissions are given on 1 km $\times$ 1 km grid cells and are aggregated to match the leukaemia data set given on ward level. Deprivation data are given as quintiles referring to Greater London. A data set of quintiles than refer to Inner London is not available for this thesis. For a detailed description of the data set see Chapter 2, a plot of the SMRs is given in Figure 9.1. For reasons of computational time, we restrict our analysis to the area of Inner London.

As covariates for Poisson–Gamma models we use benzene and the deprivation index (Carstairs and Morris, 1991) leading to the following models:

model f: additive influence of benzene;
model m: multiplicative influence of benzene;
model p: additive influence of Carstairs index;
model q: multiplicative influence of Carstairs index;
model r: additive influence of benzene, multiplicative influence of Carstairs index;



(a) (b)

Figure 9.1: Calculated SMRs for the leukaemia data set (a). In (b) the increased SMR in the central region is set to be 1 for a better impression.

model s: additive influence of benzene and Carstairs index;
model t: multiplicative influence of benzene and Carstairs index;
model u: multiplicative influence of benzene, additive influence of Carstairs index;
model o: no influence of benzene and Carstairs index.

These are combined with different numbers of latent risk sources which are enlarged until model fit is not improved anymore.

Additionally, we employ MRF models and the BDCD clustering approach to the observed data. This corresponds to the following settings:

model y: BDCD algorithm, neighbours across river Thames;
model v: MRF model with neighbourhood structure as in BDCD;
model z: MRF model where Thames parts neighbourhood structure.

MRF models can also incorporate covariates just like Poisson–Gamma models. They can only be introduced as relative risk factors where we consider the following settings:

'none' is characterised by spatially structured and unstructured terms only;
'benzene' additionally includes atmospheric benzene emissions;
'deprivation' adds deprivation data to the model;
'deprivation and benzene' contains both, deprivation and benzene data.

Table 9.1 presents the calculated DICs for the applied Poisson–Gamma random field and clustering models. For MRF models DICs are reported in Table 9.2 to ensure a better overview.

For the observed data set, we slightly extend the model selection process compared to the simulation study. When the DIC is increased a single time for a model we continue to include latent factors. If the DIC still increases or remains constant we stop. We also compare the newly calculated value with those achieved in less complex models.

From the resulting DICs presented in Table 9.1 we conclude the following. Poisson–Gamma models assuming any influence of benzene and/or the Carstairs index require one latent covariate to drop the DIC substantially from values of the interval of $[408.6, 418.4]$ to $[390.8, 392.8]$. This holds for all interpretations of the covariates. The model fit of o, the model that does include only latent covariates, is of similar quality.

In simulated data we find a substantial decrease in the DIC when a required covariate is considered by the model, compare Chapter 8. Furthermore, the model fit is decreased when benzene — accounting for a certain amount of cases in data generation — is not considered in the model. For the leukaemia data we do not see such a difference when any of the covariates is not considered. Here, model o1 with a value of 390.8 is even at the lower boundary of the interval of DICs for all models with one latent variable. Following Spiegelhalter et al. (2002) models without any latent covariates lead to not inferior model fit and are therefore not considered anymore. From the DIC values it is not possible to select any of the Poisson–Gamma models at this stage.

Inclusion of more than one latent covariate does not improve the model fit for most Poisson–Gamma models. This does not hold for the following models:

♯ latent factors	0	1	2	3	4	5	6	7	8	9	10	11
model f	417.0	390.8	394.6	417.1	—	—	—	—	—	—	—	—
model m	415.5	391.0	397.3	391.8	388.9	386.9	385.2	384.1	383.4	382.8	382.5	382.5
model p	418.4	390.4	393.6	394.1	—	—	—	—	—	—	—	—
model q	417.2	391.4	393.2	392.0	388.4	385.7	383.7	382.4	381.7	381.3	381.1	381.1
model r	410.0	391.7	391.7	391.6	—	—	—	—	—	—	—	—
model s	413.8	390.4	395.3	393.5	391.6	—	—	—	—	—	—	—
model t	408.6	392.8	391.8	391.0	387.6	384.7	382.8	381.6	380.9	380.3	380.2	380.2
model u	408.8	391.1	391.8	392.7	—	—	—	—	—	—	—	—
model o	—	390.8	396.6	393.3	390.5	388.2	386.4	384.8	383.8	383.2	383.0	382.9
model y	375.6											
MRF models	see Table 9.2											

Table 9.1: DIC values of various models for observed leukaemia incidences.

covariates	neighbourhood v	z
none	391.0	384.9
benzene	390.7	384.7
deprivation	386.8	379.8
deprivation and benzene	383.8	377.7

Table 9.2: DIC values of MRF models for observed leukaemia cases.

- model m assuming benzene to be a relative risk factor;
- model q considering multiplicative influence of the deprivation index;
- model t including both benzene and the Carstairs index multiplicatively;
- model o containing only latent covariates.

These models profit from the inclusion of Gaussian kernels to model latent risk. The DIC converges until values in the interval of $[380.2, 382.9]$ are achieved which is the case for about eleven latent covariates. Highest values belong to Poisson–Gamma random field models that consider latent covariates only. Inclusion of benzene and deprivation data reduces the DIC, best fit is achieved for model t containing both benzene and deprivation as relative risk factors. Superiority of this model becomes clear for a higher number of latent covariates as the DIC converges faster to a slightly lower value of 380.2 compared to other settings. At this stage, main characteristics of the risk surface are already identified and fine adaption becomes the main goal.

Even if the DIC reduces by only ten points from the group containing only one latent covariate (best model o1 with DIC = 390.8) to those incorporating ten Gaussian kernels (model t10 with DIC = 380.2) we clearly favour the more complex model t10 over o1 corresponding to the recommendation of Spiegelhalter et al. (2002). Inclusion of additional latent covariates does not improve model fit for model t. Therefore we prefer model t10 over t11.

This does not hold for all considered models, compare the results of model class o where convergence is achieved slower making inclusion of 11 kernels necessary.

For MRF models assuming neighbourhood structure z we get similar results concerning the inclusion of covariates as for Poisson–Gamma random field models. While largest DIC values in the class are achieved by the model containing no covariates (DIC = 384.9), the DIC drops slightly when benzene is included (DIC = 384.7). Corresponding model formulations of Poisson–Gamma models achieve lower DICs for these two settings. Inclusion of the deprivation index as a single covariate in the model results in a more remarkable drop leading to a DIC of 379.8. Here the corresponding value for Poisson–Gamma random field models q11 and q10 is 381.1, i.e., greater. Consideration of both, benzene and deprivation, leads to an additional decrease of the DIC leading to a value of 377.7 for the MRF model and 380.2 for the Poisson–Gamma model t10. In terms of the effective parameters, we calculate p_D to be 41.3 for the MRF model and 24.0 for model t10, i.e., prior information contributes more information about the parameters for the MRF model compared to Poisson–Gamma models. This value is decreased for the alternative neighbourhood structure v assuming the Thames to part the neighbourhood structure where we calculate $p_D = 23.7$. In general, this alternative adjacency structure leads to increased DICs and is therefore not considered any further. We conclude that a risk change is not apparent at the Thames.

We compare the relative risks (RR) of benzene and deprivation for model t10 and the corresponding MRF model. As reference class for the deprivation index we use those 233 of the 310 Inner London wards in the highest quintile when referring to Greater London. In the fourth quintile we have 53 wards, the third quintile contains 18 wards. The last class compromises six wards in the two lowest deprivation quintiles of the Greater London data base. From Table 9.3 we see a lower RR for all covariates when applying the Poisson–Gamma random field model t10. Additionally, 95% credibility intervals (CI) are smaller when compared to the corresponding MRF model. In our example, a lower deprivation index increases the risk to suffer from

	model t10	model z
benzene	1.437 [1.002, 2.654]	2.990 [1.358, 5.796]
deprivation 2	1.894 [1.004, 4.389]	5.380 [1.708, 11.900]
deprivation 3	1.245 [1.001, 1.908]	1.759 [0.943, 2.905]
deprivation 4	1.235 [1.002, 1.712]	1.598 [1.072, 2.283]

Table 9.3: Relative risk due to benzene and deprivation index and 95% credibility intervals for model t10 and z.

leukaemia. We also observe a trend of increasing RR with decreasing deprivation. As estimation is based on a small number of wards only, especially for the lowest group, we recommend to repeat this analysis using quintiles of the deprivation index referring to Inner London.

For atmospheric benzene emissions we observe an increased RR with increased benzene emissions. For both models, the critical value 1 is not included in the 95% CI. Again, the CI of the MRF model is wider. Thus, like Best et al. (2001) in their analysis of Greater London we found a positive association of benzene to childhood leukaemia.

The third model class applied to observed leukaemia cases is the BDCD algorithm. Here, we cannot introduce any covariates. The calculated DIC value is lower than those for any other models, namely 375.6.

We compare the results of the selected models in Figure 9.2. The scatterplot matrix consists of the estimated parameters $\widehat{\Lambda}_i E_i$ of the following models:

- Poisson–Gamma model with eleven latent covariates (o11, DIC = 382.9);
- Poisson–Gamma model assuming a multiplicative influence of benzene and ten latent covariates (m10, DIC = 382.5);
- Poisson–Gamma models assuming multiplicative influence of both, benzene and deprivation, with ten latent covariates (q10, DIC = 381.1);
- Poisson–Gamma model assuming a multiplicative influence of benzene

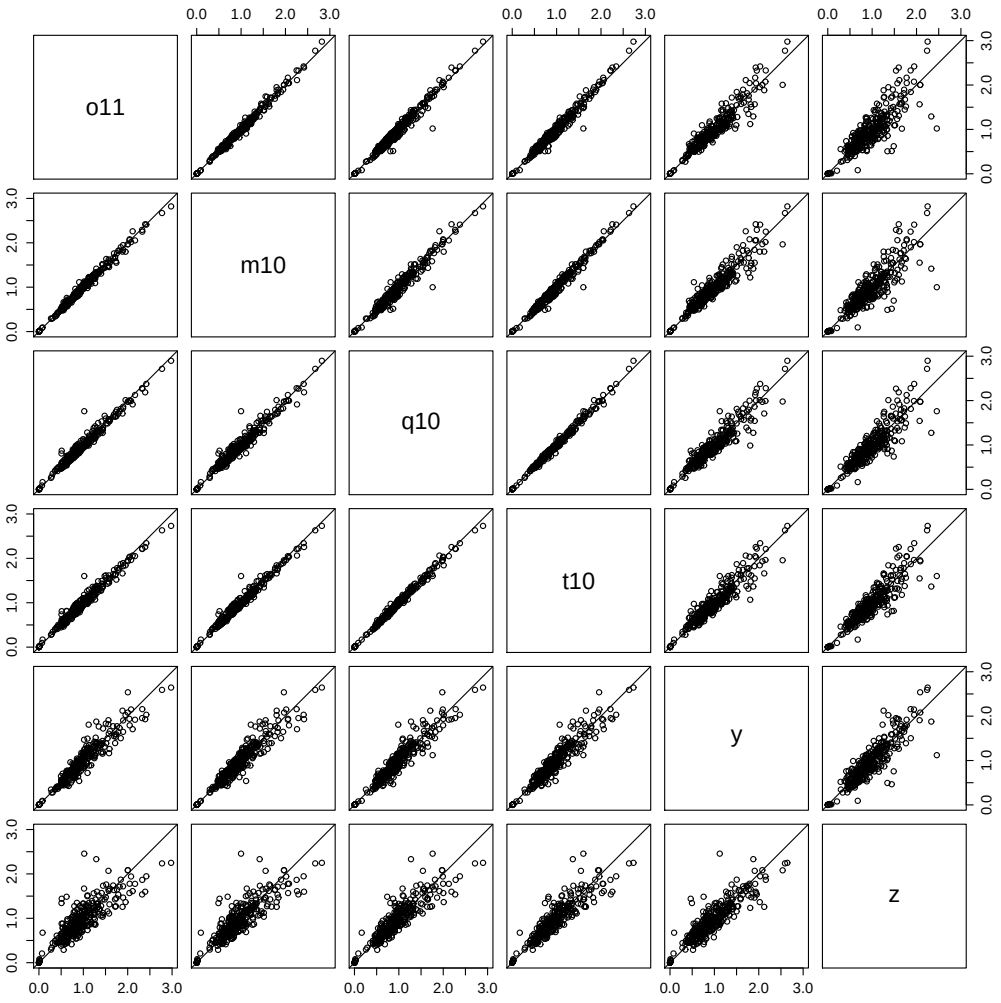


Figure 9.2: Results for the leukaemia data set: Scatterplot matrix of the estimated parameters $\widehat{\Lambda}_i\, E_i$ of selected models.

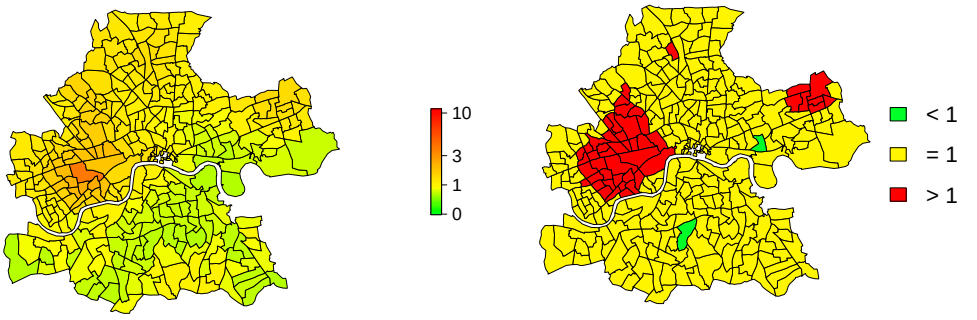
and Carstairs and ten latent covariates (t10, DIC = 380.2);

- BDCD model (y, DIC = 375.6);
- MRF model assuming wards parted by the Thames to be neighbours and multiplicative influence of benzene and deprivation (z, DIC = 377.7).

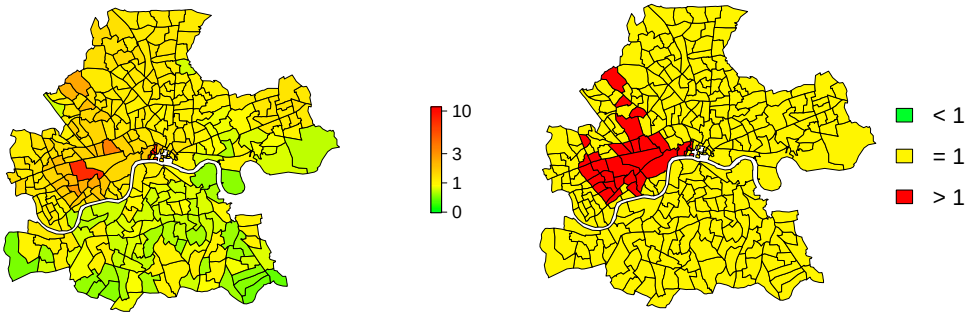
High agreement between $\widehat{\Lambda}_i E_i$ estimated by Poisson–Gamma models is obvious in the upper left corner of the scatterplot matrix. This is caused by the same structure of the models: all contain a number of latent covariates represented by Gaussian kernels. Differences result in inclusion of benzene and deprivation data, differences from a different number of latent covariates are negligible. DICs of all three models differ by only three points. We prefer model t10 as this leads to the lowest DIC as described above. Higher deviations between estimated $\widehat{\Lambda}_i E_i$ are observable between the three different model groups.

We compare risk surfaces of the Poisson–Gamma random field model t10, the corresponding MRF model with neighbourhood z and the BDCD model in Figure 9.3. The first two models contain benzene and deprivation multiplicatively. All three risk surfaces show similar patterns characterised by low risk regions compared to the UK nation-wide leukaemia rate on the southern river bank. Northern wards tend to have higher rates than the nation-wide average which is used to calculate the number of expected cases. We observe differences in the maximum value estimated by each of the models. While the maximum of the Poisson–Gamma model is 6.1, we estimate values of 4.3 for model y and 10.7 for the MRF model. Low risk regions of models t10 and y are smooth and very similar. In contrast, such regions are more patchy for the MRF model. Even if differences appear, the overall impression of the risk surface is similar for all three models. The location of high and low risk regions resembles the surface of the SMRs as given in Figure 9.1.

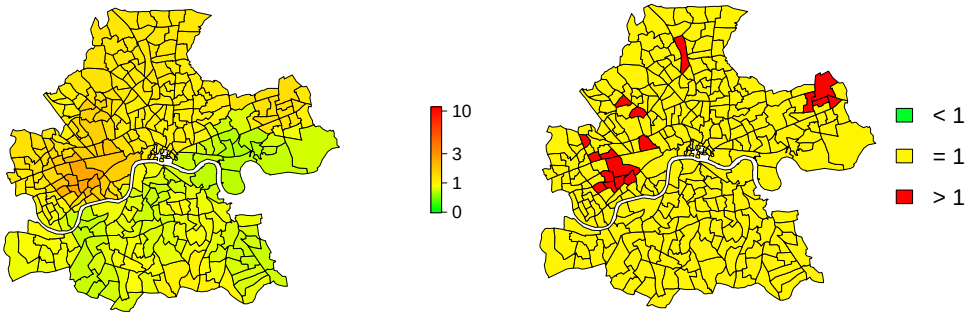
We also compare the regions where $\widehat{\Lambda}_i$ is estimated to be lie outside the 95% CI. These are given in Figure 9.3 as well. All models estimate an increased risk compared to the nation wide leukaemia rate for a small number of wards



(a: t10, DIC = 380.2) (b: 95% CI of t10)



(c: z, DIC = 377.7) (d: 95% CI of z)



(e: y, DIC = 375.6) (f: 95% CI of y)

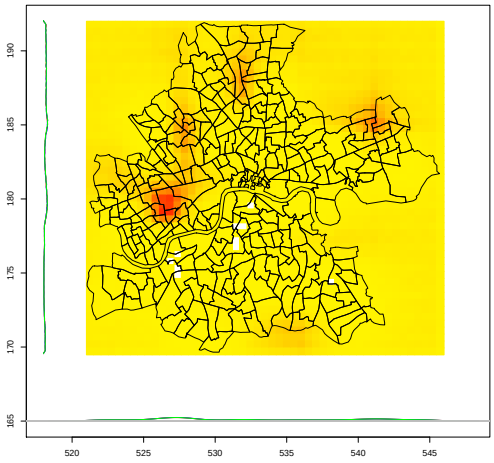
Figure 9.3: Results for the leukaemia data set: Risk surfaces $\widehat{\Lambda}_i$ of Poisson–Gamma model with ten latent covariates and multiplicative influence of benzene and deprivation (a), and the MRF model assuming the Thames not to part the neighbourhood structure involving benzene and deprivation (model z, c), and the BDCD model (model y, e) and corresponding 95% credibility intervals of $\widehat{\Lambda}_i$.

only. Model t10 is the only model that estimates the risk in two single wards to be lower than one as they lie outside the 95% CI. Concerning increased risk all models identify a region in the central western part of Inner London, although the size varies. Furthermore, model y and model t10 identify a high-risk region on the north eastern edge of the study region.

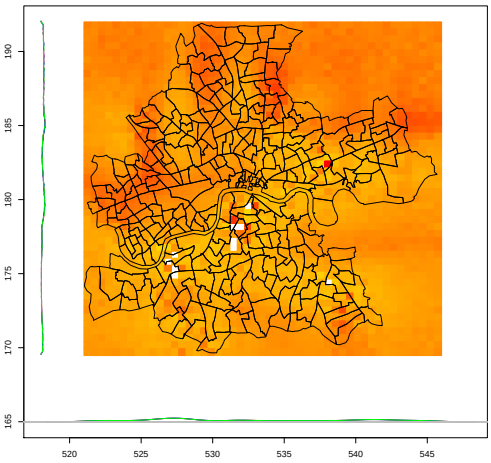
Poisson–Gamma models, especially the surface estimated by the Gamma random field, identify covariates that should be introduced in the model. We can partition the model by terms representing the covariates and latent covariates under consideration. Figure 9.4 gives an impression of the estimated location and variances of the Gaussian kernels. In contrast to previously shown figures which give the location of Gaussian kernels only such as Figure 9.4 (a), we expand our presentation to present the variances of the Gaussian kernels along both axes.

Dividing the study region into cells of 500×500 m we observe mean variances in [0.12, 3.18] for longitude and in [0.22, 2.97] for latitude among the cells. The variance is displayed in Figure 9.4 for longitudinal (Subfigure (b)) and latitudinal direction (Subfigure (c)) separately. The margin of each subfigure contains density estimates of the location of the Gaussian kernel on the same scale as in Chapter 8. In contrast, colours in the study region are adopted to the current data set.

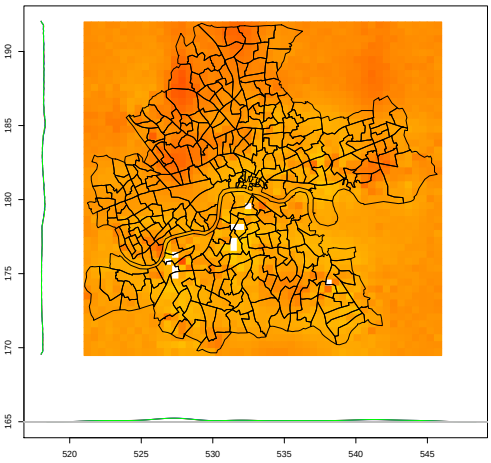
From the location of the Gaussian kernels we do not expect any particular clusters of increased risk, contrary to for example the locations of the kernels for structure I in Figure 8.6 on page 120. We rather find latent risk over the whole study region. There is a tendency to favour regions in the north-western part of the Inner London area, see Figure 9.4 (a). This corresponds to regions of significantly increased risk, compare Figure 9.3, and regions of increased SMRs, see Figure 9.1. Furthermore, frequently chosen locations are close to wards with lower deprivation, compare the plot of the Carstairs index on page 15. This holds true for the most frequently assumed position north of the Thames as well as for the two north-western ones and the slightly increased region in the south. Kernels in surrounding regions are characterised by high variances representing a smoothly decreasing influence.



(a: locations of the Gaussian kernels)



(b: variances for longitude)



(c: variances for latitude)

Figure 9.4: Results for the leukaemia data set: Location (a) and variances in longitudinal (x) and latitudinal (y) direction for the Poisson–Gamma models, red colour indicates high values, yellow colour low variances; the margin represents the estimated location of the Gaussian kernels for both directions.

As lower deprivation also increases the RR for leukaemia, these findings encourage us to use an arrangement of quintiles that refers to Inner London. Unfortunately, such data is not available for this thesis. Even so, Poisson–Gamma random field models allow for a detailed inspection of the estimated risk surface leading to a better understanding of the evaluated data set.

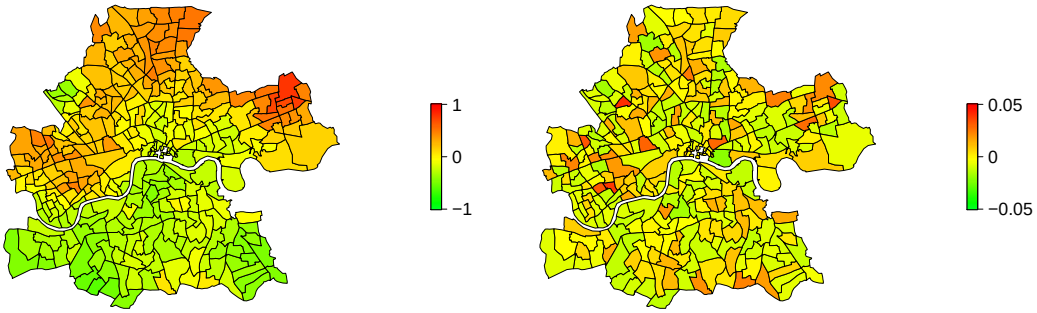
For the leukaemia data set, Poisson–Gamma random field model t10, MRF model z and partition model y lead to most satisfying results. The DICs of the three models differ by only 4.6 points. This corresponds to a strong support for the MRF model and weak support for the Poisson–Gamma model following the recommendations of Spiegelhalter et al. (2002). Benzene emissions and deprivation index are introduced as relative risk factors if possible for the model. Using the covariates as excess risk factors does not improve the model fit for all model classes. We find higher benzene levels to increase the risk to get leukaemia. Concerning the Carstairs index we find an increased risk for low deprived wards. As estimation is based on a small set of less deprived areas only we recommend to repeat this analysis with either quintiles referring to Inner London or the original deprivation data set.

For Poisson–Gamma random field models we observe a benefit in separate estimation of variances of each direction. High variances are estimated for locations inside the bounding box but also outside Inner London. Here kernels increase the risk in large parts of the study area.

For MRF models we can plot spatially structured and unstructured terms separately as done in Figure 9.5. Note that the scale of both subfigures differs.

For the spatially unstructured terms we estimate values close to zero with a very low variance. In contrast, spatially structured terms are more important for estimation of $\widehat{\Lambda}_i$. It leads to a reduced risk in wards south of the Thames and elevates risk in northern wards. Inspection of these figures helps to identify useful covariates.

In contrast, density estimates of Gaussian kernels provide a much easier interpretable information compared to structured and unstructured spatial covariates in MRF models. Poisson–Gamma models reveal the whole com-



(a: structured terms) (b: unstructured terms)

Figure 9.5: Results of the MRF model for the leukaemia data set: spatially structured terms U and unstructured terms V.

plexity of the latent risk surface. As information about the location of latent risks as well as their variance is available, investigation of potential alternative covariates is more convenient.

Additionally, other kernel functions are possible and make the class of Poisson–Gamma models more flexible. As the partition approach of the BDCD model leads to a lower DIC, such an adoption is a reasonable extension of this work.

Chapter 10

Summary and discussion

The main topic of this thesis was to analyse the performance of Poisson–Gamma random field models and its implementation into WinBUGS. We focused on disease mapping, especially on the ability of the model class to produce a precise map of the underlying risk surface by inclusion of latent risk factors and the necessity to allow for different covariate interpretations. We allowed for different spatial resolutions for latent pattern in contrast to observed data and covariates.

We set up a simulation study that allowed to evaluate models' performances with respect to possible covariate interpretations for different spatial structures. These included combinations of the covariate with other spatial characteristics such as a linear decreasing trend or clusters with increased risk.

We applied Poisson–Gamma random field models with either excess or relative risk factors in combination with varying numbers of latent covariates represented by Gaussian kernels. We also compared Poisson–Gamma random field models to other frequently used spatial models of similar complexity. These were a MRF-based ecologic regression model (Besag et al., 1991) and the BDCD model (Knorr-Held and Raßer, 2000). The first one was also implemented into WinBUGS, for the second model a software implementation is available from the authors.

Model evaluation was done using MSE and DIC. For the DIC, we also dis-

cussed different approaches for Monte Carlo error estimation as an alternative to the time-consuming Brute Force approach proposed by Zhu and Carlin (2000). In addition to the Brute Force approach the authors themselves also introduced a batching approach which we included in our comparison. Furthermore, we suggested alternatives for batch construction (batching and thin), and application of bootstrapping and cross-validation.

In contrast to Zhu and Carlin (2000) we estimated only small variances for the Brute Force approach as well as for our alternatives. One reason is the different number of free parameters. While the data set employed by Zhu and Carlin (2000) involved hospitalisations in 97 zip codes in a spatial-temporal setting over an eight-year study period, our examples included less free parameters represented by ten power plant pumps, lip cancer rates in 56 counties in Scotland, and leukaemia cases in 310 Inner London wards. As we expect the expected posterior deviance to equal the number of free parameters (Spiegelhalter et al., 2002), lower DIC values are calculated. This corresponds to smaller variances of the DIC. For future research we recommend the usage of more complex data sets. In such a comparison, e.g. the hospitalisation data set by Zhu and Carlin (2000) should be included. As the estimation of an MC error for the mean of the deviance $\bar{D}$ is straightforward, we recommend to select rather complex models where p_D tends to be large. Furthermore, we suggest to include other models not assuming Poisson-distributed data as the investigation and comparison of the estimation performance is of interest for other model classes as well.

In a first part of the simulation study we implemented a restricted version of Poisson–Gamma random field models in the WinBUGS software. Using a small number of kernels at fixed locations does not lead to a satisfying estimation of the latent field as the dependency of the model fit on an appropriate position of the kernels is very high. We therefore extended our implementation of the Gamma random field in WinBUGS.

For estimation of the random field we applied bivariate Gaussian kernels. Longitudinal and latitudinal directions are assumed to be independent, their variance and location are estimated within the MCMC framework.

Here we find the necessity to allow for alternative interpretations of covariates. In general, additive Poisson–Gamma models convince when an excess risk factor is involved in data generation while multiplicative Poisson–Gamma models tend to have higher DIC values. On the other hand, multiplicative modelling is to favour when risk due to a relative risk factor is generated. This becomes more obvious with larger influence of the covariate. When in doubt, results of the simulation study suggest to prefer the multiplicative model.

Non-consideration of a covariate that is involved in data generation leads to highest DIC values within the framework of Poisson–Gamma models.

In general, Poisson–Gamma random field models convince by their ability to reproduce the generated structure. The WinBUGS' adoption of the Gamma random field is able to reproduce latent structures. By analysing the corresponding terms we can identify covariates that should be included in the model.

We also applied MRF models and the BDCD model to generated data sets. For clustered structures the BDCD model is to favour over Poisson–Gamma models even though these models do not allow for covariates. In MRF models we included the covariate in a multiplicative setting. Here we usually find an inferiority of the model compared to corresponding Poisson–Gamma models. Exceptions are given by structures where the risk changes instantly at the Thames. Using a corresponding neighbourhood structure drops the DIC substantially and convinces by an improved estimation of the risk surface compared to all other models.

These results encourage us to enlarge the flexibility of Poisson–Gamma random field models. Distance-based Gaussian kernels lead to a comparable neighbourhood structure as the conditional auto-regressive approach where all wards are connected if they share a common border. A neighbourhood where the Thames parts this structure as assumed by one of the MRF models in our simulation study corresponds to half–Gaussian kernels that can be used alternatively for estimation of the Gamma latent field. In contrast to MRF models we do not need to fix a neighbourhood structure, the break

♯ Gaussian kernels	0	5	10	15
computational time	15 min	3 days	10 days	4 weeks

Table 10.1: Computational time of selected models

is estimated by the kernels' parameters automatically. In order to produce similar conditions for Poisson–Gamma random field models and the BDCD clustering algorithm Uniform kernels can be considered as an alternative. We expect an improvement for situations of increased risk in cluster regions. Using overcomplete dictionaries by allowing different kernel functions simultaneously such as presented by Clyde and Wolpert (2007) may also make this model class more flexible.

In our WinBUGS' implementation, we assumed each kernels' longitude and latitude to be independent. This is sufficient to reproduce generated structures. We point out that structures assuming a sharp risk increase at the west–to–east directed Thames correspond to such an independence assumption. Otherwise, random field estimation can be improved by the introduction of a covariance structure in model estimation. We emphasise that this increases computational time.

In Poisson–Gamma models the number of latent Gaussian kernels necessary to reproduce the spatial structure is typically small. For a larger number of kernels we see disadvantages of the model class as computational time is increased. For some selected runtimes see Table 10.1. We present those of typical 1GB RAM PCs. The underlying structure is simulated assuming a multiplicative influence of benzene and three clusters of increased risk (structure U). The number of Gaussian kernels was increased successively until the DIC is not reduced anymore. Hence, fitting a model that requires 15 latent covariates required several weeks. As we see in Table 10.1, this is particularly due to the complete estimation of an elevated number of Gaussian kernels. Implementation in WinBUGS/OpenBUGS cannot be optimised anymore unless the code is directly implemented by BUGS developers. For the WinBUGS implementation the user himself can slightly optimise some

code by implementation in the developers tool WBDev, for the OpenBUGS version this is not possible. As one goal of this thesis was to implement Poisson–Gamma random field models into WinBUGS, we did not code the model directly in `C++` and provided a programm such as for the BDCD model. A main disadvantage of such an a ready–to–use program is either its complexity as the model class itself or its restrictions on the flexibility of the user. In contrast, we hope that the availability of an adoptable WinBUGS implementation as presented in this thesis will provide a basis for other users that want to apply the class of Poisson–Gamma models on their data.

Speeding up calculations can also be done by assuming a fixed variance for all kernels. However, as we have seen in particular for structures with an abrupt risk change, this is a main advantage of our implementation and increases the model fit.

Another possible improvement is to allow the inclusion of the number of latent risk sources as a hyper-parameter in the model and let it being estimated in the MCMC algorithm, e.g., by reversible jump MCMC methods (Green, 1995). We expect a decrease in computational time compared to the iterative procedure proposed here. Unfortunately, this approach necessitates loops running over all possible numbers of sources and therefore depending on a random quantity. However, using WinBUGS, nodes used as bounds in `for()`-loops are not allowed to be stochastic (Spiegelhalter et al., 2004).

Recent extensions of the model class are applications for concentrations of pollutants at point sources and their dispersal over time and space, i.e., a non-stationary, spatial-temporal model. Clyde et al. (2006) use this model class to estimate the abundance of proteins in mass spectroscopy data. A good overview on such extensions is given by Clyde and Wolpert (2007). These additional applications in combinations with our findings show the power of this approach.

Keeping in mind our findings of the simulation study we applied all models to observed leukaemia counts. For covariates we considered atmospheric benzene emissions as well as the Carstairs deprivation index. The latter one was available in quintiles referring to Greater London only. Both were

included as excess and relative risk factors in different combinations. Again, we successively added latent covariates represented by Gaussian kernels. We also employed a model including only latent covariates. The DIC was used to identify the most appropriate model. As in the simulation study we applied BDCD and MRF models to the observed data. The latter model is extended to allow for deprivation as well as benzene as covariates as relative risk factors.

Best results are achieved when both, atmospheric benzene emissions and deprivation data are included in the model. This holds for both, the Poisson–Gamma random field model and the MRF-based model. Here, we prefer a neighbourhood structure that is not influenced by the course of the Thames river. MRF models lead to a slightly lower DIC value. Most appropriate fit is achieved for the BDCD partition model. The DIC value of these three models differs by 4.6 points only.

Models allowing for covariates identify benzene to increase the risk of leukaemia. Furthermore, we find an increased RR in less deprived wards. Both covariates are identified to be relative risk factors. As the Carstairs index is discretised by quintiles referring to the Greater London data base, for some deprivation quintiles the number of observations is small. Hence, we recommend to repeat this analysis using either quintiles referring to the study area only or the original data set to confirm our findings.

Best et al. (2001) perform a similar analysis in the area of Greater London using MRF-based ecologic regression models. They consider atmospheric benzene emissions as covariate and find a positive association with the risk of childhood leukaemia similar to our findings. However, their analysis on ward level found a higher increase in RR due to benzene emissions. In contrast to Best et al. (2001) we used the area of Inner London only where benzene tends to be higher while leukaemia cases and the population at risk have similar characteristics for both areas leading to increased exposure and therefore smaller variance. For a comparable analysis we recommend to apply our model on Greater London's data. As the number of wards is increased from 310 (Inner London) to 873 (Greater London), and the corresponding area from 548 km^2 (Inner London) to 2887 km^2, this increases computational

time. To be able to run such amount of data in reasonable time, WinBUGS performance needs to be improved first.

For Inner London, the model leading to the lowest DIC is the BDCD model by Knorr-Held and Raßer (2000). In our simulation study we find this model to be the most appropriate one for clustered structures. Hence, we recommend to use alternative kernels to model the latent risk in Poisson–Gamma models which are more suitable for clustered structures. Uniform kernels or half-Gaussian ones provide suitable alternatives. We expect an improvement in model performance from such an extension.

Another improvement of the model is given by treating benzene on the observed spatial resolution of a $1\,\mathrm{km} \times 1\,\mathrm{km}$ regular grid. We have aggregated this data to the same scale as the observed leukaemia data and only treated the latent field on its original scale. An extension is possible within the framework of Poisson–Gamma random fields but not considered in this thesis as computational time is very high already. Nevertheless, such an improvement is necessary for further reduction of the ecological bias.

Bibliography

Agency for Toxic Substance and Disease Registry (ATSDR) (1997): *Toxicological Profil for Benzene*. Atlanta, GA: USA: U.S. Departement on Health and Human Services, Public Health Service. URL `www.atsdr.cdc.gov/toxprofiles/tp3.pdf`.

Aitkin, M., Boys, R., and Chadwick, T. (2004): Bayesian point null hypothesis testing via the posterior likelihood ratio. *Statistics and Computing*, 15, 217–230.

Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csáki (eds.) *Proceedings of the 2nd International Symposium of Information Theory*, 267–281. Akadémia Kiadó.

Arnold, R. (1999): Small area health statistics unit procedures for estimating populations in small areas. In: R. Arnold, P. Elliot, J. Wakefield, and M. Quinn (eds.) *Population Counts in Small Areas: Implications for Studies of Environment and Health*, 10–24. London: Stationary Office.

Banerjee, S., Calin, B. P., and Gelfand, A. E. (2004): *Hierarchical Modelling and Analysis for Spatial Data*. Boca Raton: Chapman & Hall/CRC.

Banerjee, S., Gelfand, A., and Sirmans, C. (2003): Directional rates of change under spatial process models. *Journal of the American Statistical Association*, 946–954.

Bayarri, M. and Berger, J. (2000): P-values for composite null models (with

discussion). *Journal of the American Statistical Association*, 95, 1127–1142.

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

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

Best, N. G., Cockings, S., Bennett, J., Wakefield, J., and Elliott, P. (2001): Ecological regression analysis of environmental benzene exposure and childhood leukaemia: Sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society/A*, 164, 155–174.

Best, N. G., Ickstadt, K., and Wolpert, R. L. (2000): Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, 95, 1076–1088.

Best, N. G., Richardson, S., and Thomson, A. (2005): A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35–99.

Best, N. G. and Wakefield, J. (1999): Accounting for inaccuracies in population counts and case registration in cancer mapping studies. *Journal of the Royal Statistical Society/A*, 162, 363–382.

Böhning, D. (2000): *Computer-Assisted Analysis of Mixtures and Applications*. Boca Raton: Champman & Hall/CRC.

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

Breslow, N. and Day, N. (1980): *Statistical Methods in Cancer Research, Vol. 1: The Analysis of Case-Control Studies*. Lyon: International Agency for Research on Cancer.

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

Buckingham, C., Clewley, L., Hutchinson, D., Sadler, L., and Shah, S. (1997): London atmospheric emissions inventory. *Technical report*, London Research Centre, London.

Carstairs, V. (1995): Deprivation indices: their interpretation and use in relation to health. *Journal of Epidemiology and Community Health*, 49, Suppl. 2, S3–S8.

Carstairs, V. (2000): Socio-economic factors at areal level and their relationship with health. In: *Spatial Epidemiology: Methods and Applications*, 51–67. Oxford: Oxford University Press.

Carstairs, V. and Morris, R. (1991): *Deprivation and Health in Scotland*. Aberdeen: Aberdeen University Press.

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

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

Clyde, M. A., House, L. L., and Wolpert, R. L. (2006): Nonparametric models for proteomic peak identification and quantification. In: K. Do, P. Müller, and M. Vannucci (eds.) *Bayesian Inference for Gene Expression and Proteomics*, 293–308. Cambridge: Cambridge University Press.

Clyde, M. A. and Wolpert, R. L. (2007): Nonparametric function estimation using overcmplete dictonaries. In: J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West (eds.) *Bayesian Statistics*, volume 8, 1–24. Oxford: Oxford University Press.

Committee on Medical Aspects of Radiation in the Environment (COMARE) (2006): *11th Report: The Distribution of Childhood Leukaemia and other*

Childhood Cancers in Great Britain 1969-1993. Oxon: Health Protection Agency for the Committee on Medical Aspects of Radiation in the Environment. ISBN 0-85951-578-8.

Cowles, M. K. (2004): Review of WinBUGS. *The American Statistician*, 58, 330–336.

Cowles, M. K. and Carlin, B. P. (1996): Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883–904.

Cressie, N. A. C. (1993): *Statistics for Spatial Data*. New York: Wiley.

Dickinson, H., Hammal, D., Dummer, T., Parker, L., and Bithell, J. (2003): Childhood leukaemia and non-Hodgkins lymphoma in relation to proximity to railways. *British Journal of Cancer*, 88, 695–698.

Diggle, P. (2003): *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

Dockerty, J. D., Draper, G., Rowan, S. D., and Bunch, K. J. (2001): Case-control study of parental age, parity and socioeconomic level in relation to childhood cancers. *International Journal of Epidemiology*, 30, 1428–1437.

Duarte-Davidson, R., Courage, C., Rushton, L., and Levy, L. (2001): Benzene in the environment: an assessment of the potential risks to the health of the population. *Occupational and Environmental Medicine*, 58, 2–13.

Elliot, P., Wakefield, J., Best, N., and Briggs, D. (2000): *Spatial Epidemiology. Methods and Applications*. Oxford: Oxford University Press.

Fernández, C. and Green, P. J. (2002): Modelling spatial correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society/B*, 64, 805–826.

Finkenstädt, B., Held, L., and Isham, V. (eds.) (2007): *Statistical Methods for Spatial–Temporal Systems*. Boca Raton: Chapman & Hall/CRC.

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

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

Gelman, A. (1996): Inference and monitoring convergence. In: W. Gilks, S. Richardson, and D. J. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*, 131–143. London: Chapman & Hall.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003): *Bayesian Data Analysis*. Boca Raton: CRC Press, 2nd edition.

Gelman, A. and Rubin, D. (1992): Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Geman, S. and Geman, D. (1984): Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence*, 6, 721–741.

George, E., Makov, U., and Smith, A. (1993): Conjugate likelihood distributions. *Scandinavian Journal of Statistics*, 20, 147–156.

Geweke, J. (1992): Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds.) *Bayesian Statistics*, volume 4, 169–193. Oxford: Oxford University Press.

Gilks, W. (1992): Derivative–free adaptive rejection sampling for Gibbs sampling. In: J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds.) *Bayesian Statistics, Volume 4*, 641–665. Oxford: Oxford University Press.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996a): Introducing Markov chain Monte Carlo. In: W. Gilks, S. Richardson, and D. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

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

Green, P. J. (1995): Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

Green, P. J. and Richardson, S. (2002): Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97, 1055–1070.

Greenland, S. and Robins, J. (1994): Ecological studies — biases, misconceptions and counterexamples. *American Journal of Epidemiology*, 139, 747–760.

Groër, M. and Shekleton, M. (1979): *Basic Pathophysiology: A Conceptual Approach*. St. Louis: C. V. Mosby Company.

Gulliford, M., Bell, J., Bourne, H., and Petruckevitch, A. (1993): The reliability of cancer registry records. *British Journal of Cancer*, 67, 819–821.

Hastings, W. (1957): Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 97–109.

Hattemer-Frey, H. A., Travis, C. C., and Land, M. L. (1990): Benzene: Environmental partitioning and human exposure. *Environmental Research*, 53, 221–232.

Heidelberger, P. and Welch, P. (1983): Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109–1144.

Held, L., Natario, I., Fenton, S., Rue, H., and Becker, N. (2005): Towards joint disease mapping. *Statistical Methods in Medical Research*, 14, 61–82.

Ickstadt, K. (2001): *On Hierarchical Point Process Models in Spatial Statistics*. Habilitation thesis, Darmstadt University of Technology.

Ickstadt, K. and Wolpert, R. L. (1997): Multiresolution assessment of forest inhomogeneity. In: C. Gatsonis, J. Hodges, R. Kass, R. McCulloch, P. Rossi, and N. Singpurwalla (eds.) *Case Studies in Bayesian Statistics, Volume III*, 371–386. New York: Springer.

Ickstadt, K. and Wolpert, R. L. (1999): Spatial regression for marked point processes. *Bayesian Statistics*, 6, 323–341.

Jackson, C., Best, N. G., and Richardson, S. (2006): Improving ecological inference using individual-level data. *Statistics in Medicine*, 25, 2136–2159.

Jin, X., Carlin, B. P., and Banerjee, S. (2005): Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, 61, 950–961.

Kass, R. and Raftery, A. (1995): Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

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

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

Linet, M. S. and Cartwright, R. (1996): The leukaemias. In: D. Schottenfeld and J. Fraumeni (eds.) *Cancer Epidemiology and Prevention*, 843–892. New York: Oxford University Press, 2nd edition.

Little, J. (1999): *Epidemiology of Childhood Cancer*. Lyon: International Agency for Research on Cancer.

Lu, H. and Carlin, B. P. (2005): Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 37, 265–285.

McCullagh, P. and Nelder, J. (1990): *Generalized Linear Models*. Boca Raton, Florida: Chapman & Hall, 2nd edition.

McLachlan, G. and Peel, D. (2000): *Finite Mixture Models*. New York: Wiley.

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

Mood, A., Graybill, F., and Boes, D. (1974): *Introduction to the Theory of Statistics*. Singapure: McGraw Hill Book Company, 3rd edition.

Neal, R. (1997): Markov chain Monte Carlo methods based on 'Slicing' the density function. *Technical report*, Department of Statistics, University of Toronto, Toronto, Canada.

Oberon microsystems, Inc. (2004): *BlackBox Component Builder*. Switzerland. URL `http://www.oberon.ch/blackbox.html`.

Perez, J. and Berger, J. (2002): Expected posterior prior distributions for model selection. *Biometrika*, 89, 491–512.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2004): *R Package* `coda`*: Output Analysis and Diagnostics for MCMC*. URL `http://www-fis.iarc.fr/coda/`. Version 0.7-2.

R Development Core Team (2006): *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

Richardson, S. (1992): Statistical methods for geographical correlation studies. In: P. Elliot, J. Cuzick, D. English, and R. Stern (eds.) *Geographical and Environmental Epidemiology: Methods for Small-Area Health Studies*, 181–204. Oxford University Press.

Rue, H. and Held, L. (2005): *Gaussian Markov Random Fields. Theory and Applications*. Boca Raton: Chapman & Hall/CRC.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. (2002): Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society/B*, 64, 583–639.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2004): *WinBUGS Users Manual*. URL `http://www.mrc-bsu.cam.ac.uk/bugs`. Version 1.4.1.

Steffen, C., Auclerc, M., Auvrignon, A., Baruchel, A., Kebaili, K., Lambilliotte, A., Leverger, G., Sommelet, D., Vilmer, E., Hemon, D., and Clavel, J. (2004): Acute childhood leukaemia and environmental exposure to potential sources of benzene and other hydrocarbons; a case-control study. *Occupational and Environmental Medicine*, 61, 773–778.

Stein, M. L. (1999): *Interpolation of Spatial Data. Some Theory for Kriging*. New York: Springer.

Sturtz, S. (2002): *Raum-Zeitliche Untersuchung der Herz-Kreislauf-Mortalität in Nordrhein-Westfalen*. Master's thesis, Department of Statistics, University of Dortmund.

Sturtz, S., Ligges, U., and Gelman, A. (2005): R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12/3, 1–16.

Swerdlow, A. (1986): Cancer registration in England and Wales: Some aspects relevant to interpretation of the data. *Journal of the Royal Statistical Society/A*, 149, 146–160.

Thomas, A. (2004): *BRugs User Manual*. URL `http://mathstat.helsinki.fi/openbugs/`. Version 1.0.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006): Making BUGS open. *R News*, 6, 12–17.

UK Childhood Cancer Study Investigators (2000a): Childhood cancer and residential proximity to power lines. *British Journal of Cancer*, 83, 1573–1580.

UK Childhood Cancer Study Investigators (2000b): The United Kingdom Childhood Cancer Study: Objectives, materials and methods. *British Journal of Cancer*, 82, 1073–1102.

Vehtari, A. and Lampinen, J. (2004): Model selection via predictive explanatory power. *Technical Report B38*, Helsinki University of Technology, Laboratory of Computational Engineering Publications.

Wakefield, J. C., Best, N. G., and Waller, L. (2000): Bayesian approaches to disease mapping. In: P. Elliot, J. C. Wakefield, N. G. Best, and D. Briggs (eds.) *Spatial Epidemiology: Methods and Applications*, 104–127. Oxford: Oxford University Press.

Wallace, L. (1996): Environmental exposure to benzene: An update. *Environmental Health Perspectives*, 104, 1129–1136. URL `http://ehp.nies.nih.gov/docs/1996/Suppl-6/wallace.html`.

Whittle, P. (1954): On stationary processes in the plane. *Biometrika*, 86, 382–397.

Wild, C. and Kleinjans, J. (2003): Children and increased susceptibility to environmental carcinogens: Evidence or empathy? *Cancer Epidemiology, Biomarkers & Prevention*, 12, 1389–1394.

Wolpert, R. and Ickstadt, K. (1998): Poisson/gamma random field models for spatial statistics. *Biometrika*, 85, 251–267.

Yardley-Jones, A., Anderson, D., and Parke, D. (1991): The toxicity of benzene and its metabolism and molecular pathology in human risk assessment. *British Journal of Industrial Medicine*, 48, 437–444.

Zhu, L. and Carlin, B. P. (2000): Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19, 2265–2278.