



Stockholms
universitet

Spatio-temporal Modeling of Hantavirus in Germany

Andreas Hicketier

Masteruppsats 2015:8
Matematisk statistik
September 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Spatio-temporal Modeling of Hantavirus in Germany

Andreas Hicketier*

September 2015

Abstract

In this master thesis several spatio-temporal models are being fitted to the spatio-temporal occurrence of hantavirus in Germany. Hantavirus is an infectious disease transmitted by bank voles. The relationship between several covariates related to the number of bank voles and the disease incidence is explored. These covariates include forest area, fructification of trees and proximity between urban and forest areas. The inference is carried out in a Bayesian framework and the data is modeled as a generalized additive model to include spatial effects. The INLA method as presented by Rue et al. (2009) is explained in depth and a general summary of the necessary background material is given. Model fitting is carried out with the R-package R-INLA. We find evidence for our hypothesis that a high fructification increases the hantavirus case numbers in the following year. A high fructification of trees means an abundance of food for bank voles and as such larger case numbers can be expected. Data and models are visualized and the models are checked for adequacy.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: andreas.hicketier@posteo.de. Supervisor: Michael Höhle.

ACKNOWLEDGEMENT

I would like to express my gratitude to my supervisor, Michael Höhle, both for the idea of this thesis as well as his excellent support during the writing process.

I would also like to thank my girlfriend, Isabelle Hughes, who always supported me and gave me many helpful comments.

Lastly, I would like to thank the Robert Koch-Institut and the Johann Heinrich von Thünen-Institut for making the data available to me and thus making this thesis possible.

Contents

1	Introduction	4
1.1	The Data	5
1.1.1	Hantavirus	5
1.1.2	CORINE Land Cover	10
1.1.3	Forest Area	10
1.1.4	Fructification	11
1.1.5	Urban Proximity	17
2	Prerequisites for the Statistical Modeling	18
2.1	Bayesian Generalized Additive Models	18
2.2	Gaussian Markov Random Fields	19
2.2.1	Definition through Joint Density	19
2.2.2	Definition through Full Conditionals	20
2.3	Model Selection	21
2.3.1	Deviance Information Criterion	21
2.3.2	Probability Integral Transforms	21
3	Integrated Nested Laplace Approximations	22
3.1	Outline	22
3.2	Gaussian Approximations	23
3.3	Laplace Approximations	24
3.4	Integrated Nested Laplace Approximations	25
3.4.1	Equivalence of Laplace Approximations	26
3.5	Finding Evaluation Points for $\tilde{\pi}(\boldsymbol{\theta} \mathbf{y})$	27
3.6	Approximations of $\pi(x_i \mathbf{y}, \boldsymbol{\theta})$	28
3.6.1	Gaussian Approximation	28
3.6.2	Full Laplace Approximation	28
3.6.3	Simplified Laplace Approximation	29
3.6.4	Computing $\pi(x_i \mathbf{y})$	29
3.7	Implementation in R-INLA	29

4	Analysis of the Hantavirus Data	31
4.1	Fitting a Generalized Linear Model	31
4.2	Fitting Generalized Additive Models	33
4.3	Model Selection	34
4.4	Results of the GLM	36
4.5	Results of the GAM	37
5	Conclusion	41

Chapter 1

Introduction

The aim of this master thesis is to fit a spatio-temporal model to the occurrence of human hantavirus cases in Germany. Hantavirus is an infectious disease transmitted through inhalation of aerosolized dried excretions of bank voles to humans. It can cause severe fever with renal syndrome. As part of the Act on the Prevention and Control of Infectious Diseases (IfSG), the occurrence of the disease in humans in Germany has to be reported to the Robert Koch-Institut (RKI), the federal institute for disease control and prevention, which is part of the Federal Ministry of Health. The aim of the modeling is to help us understand which factors are important influences on the number of cases. Our main hypothesis is that there is a strong connection between the fructification of trees in a given year and the number of cases in the following year. This can be explained by the fact that the fructification of trees can be used as a proxy variable for the number of bank voles that carry hantavirus. The use of proxy variables is necessary, because we have no knowledge of the number of bank voles in an area. Since the fructification serves as a measure for the food available to bank voles, an abundance of food for bank voles in one year implies an increased number of surviving offspring and thus a greater total number of bank voles in the following year. The motivation for our hypothesis can be seen in Figure 1.1, where the first plot shows the number of yearly reported human hantavirus cases. The second and third plot show the weighted average fructification of beech and oak of all counties per year with the weights given by the counties' area. It is clearly visible that most years with high average fructification are followed by years with high reported case numbers. This gives a very rough picture as fructification and cases are not equally distributed across Germany but it serves as a good motivation on which to base a further, more detailed analysis. A connection between fructification and number of cases is also considered in Boone et al. (2012).

We will also consider the forest and field areas available as habitats to bank voles and the distance between these and urban areas. The habitats serve, again, as a proxy for the number of animals and the interaction is thought to be important as no other attack vectors for the disease (e.g. mosquitoes) seem to exist, so that proximity between humans and bank voles is necessary for the disease to be transmitted. These hypotheses are to be investigated in more detail. We will use these variables which provide us with

information on a county level to perform an ecological regression.

The analysis is carried out in a Bayesian framework using structured additive regression models. The posterior distributions are estimated using Integrated Nested Laplace Approximations (INLA). INLA was introduced by Rue et al. (2009) and is a deterministic approximation alternative to the widely used MCMC approximations offering accurate results in combination with reduced computational time. R and the R-INLA package (<http://www.r-inla.org/>) are used to fit the model. As a consequence, a secondary aim of this thesis is to present the INLA method in depth. We will explain several details in more depth than in the original paper and will offer proofs that were omitted in the paper. As INLA works on latent Gaussian models, a subclass of structured additive regression models, an overview of these model types will be provided together with a presentation of Gaussian Markov random fields, which are essential for INLA and spatial modeling in general.

1.1 The Data

The RKI provides the data of reported hantavirus infections in Germany during 2001-2012 down to the county level (German: Landkreise).

Data on the area of tree species in counties as well as the fructification of these trees, i.e. the amount of fruit/nuts carried in a year, is provided by the Johann Heinrich von Thünen-Institut, the Federal Research Institute for Rural Areas, Forestry and Fisheries. These fructification data serve as a proxy variable for the number of bank voles of which we have no knowledge.

Information on the size and location of both forests and urban areas will be obtained from CORINE land cover raster data (resolution: $100\text{m} \times 100\text{m}$) from the German Aerospace Center. This will allow the computation of covariates such as the average distance between settlements and forests in a county. And lastly, a vector layer with the country and county borders from the Federal Agency for Cartography and Geodesy will be used.

The following subsections present the data that was used for the analysis. Initial data management and preprocessing was done according to an internal report at the Robert Koch-Institut by Faber and Höhle (2013), and then extended upon as part of this master thesis. Extensions include redoing and extending the preprocessing to accommodate newer packages versions for the R packages used in the preprocessing, the inclusion of additional cases, due to a broader case definition, and assignment of cases to different counties when the counties were changed due to re-organization of the political regions (in German: *Kreisgebietsreformen*).

1.1.1 Hantavirus

Hantavirus is a single-stranded, enveloped, negative sense RNA virus which can be found worldwide. Small rodents are the main host carriers, and transmission occurs through

contact with hantavirus-infected rodents, their urine or their droppings, see Faber et al. (2010). Inhalation of aerosolized virus particles from excreta of the infected rodents is also a mean of transmission of the hantavirus.

In Europe, three types of hantaviruses can be found: Puumala, Dobrava and Saaremaa, see Vaheri et al. (2013). The first one causes a mild form of nephropathia epidemica, leaving infected patients with fever and influenza-like symptoms, including headaches, muscle pain and renal impairment. The second type of hantavirus, Dobrava, has similar symptoms but often presents hemorrhagic complications. Few cases of Saaremaa virus have been reported and documented, but the symptoms appear to lie between the ones of Puumala and Dobrava viruses. Incubation time ranges from 5 to 60 days, and the severity of the symptoms varies greatly, with some patients showing little clinical signs, which results in an under-reporting of cases.

In Germany, Puumala virus is the predominant human pathogenic hantavirus species, see Boone et al. (2012). Its host carrier is the bank vole (*Myodes glareolus*). The rural regions endemic to the virus are lower Franconia, the Bavarian forest, Swabian Alb, Münster and Osnabrück regions (Faber et al., 2010). While the bank vole is found anywhere in Germany, not all areas have a population that carries the virus.

In this analysis we consider all reported cases in the RKI database that were classified as 'Hantavirus', 'Puumalavirus' and 'not classified'. This is done to account for regional differences in analyzing and interpreting the reported cases, e.g. most 'Hantavirus' could be further analyzed to 'Puumalavirus'. The data of 2001 are somewhat unreliable, as this was the first year where the disease had to be reported according to the specifications of the German Act on the Prevention and Control of Infectious Diseases - hence it will be left out of the analysis.

As a handful of the counties were merged over the years, we assigned the newer cases randomly to the old counties that made up the new county. Some other counties were split into new ones, in which case the newer cases are assigned to the old county. In total, we have 412 counties in this analysis. A threshold of 0.25 for the average yearly incidence was included and counties with an incidence below the threshold were not considered endemic regions. The average yearly incidence for county i is given by

$$\frac{n_i \cdot 10^5}{\text{pop}_i \cdot T},$$

where n_i is the total number of cases 2002-2012 in county i , pop_i the population and T the number of years, i.e. 11. This was done to ignore very small case numbers in some counties that could be explained by e.g. people moving instead of the local bank vole population carrying hantavirus. After applying the threshold we are left with 150 endemic counties. Data of reported hantavirus incidence per county and per year are shown in Figure 1.3 and the total incidence is shown in Figure 1.2. Table 1.1 shows the total number of reported hantavirus cases in Germany per year.

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
209	127	223	403	61	1618	207	160	1920	267	2643

Table 1.1: Total number of reported hantavirus cases in Germany per year.

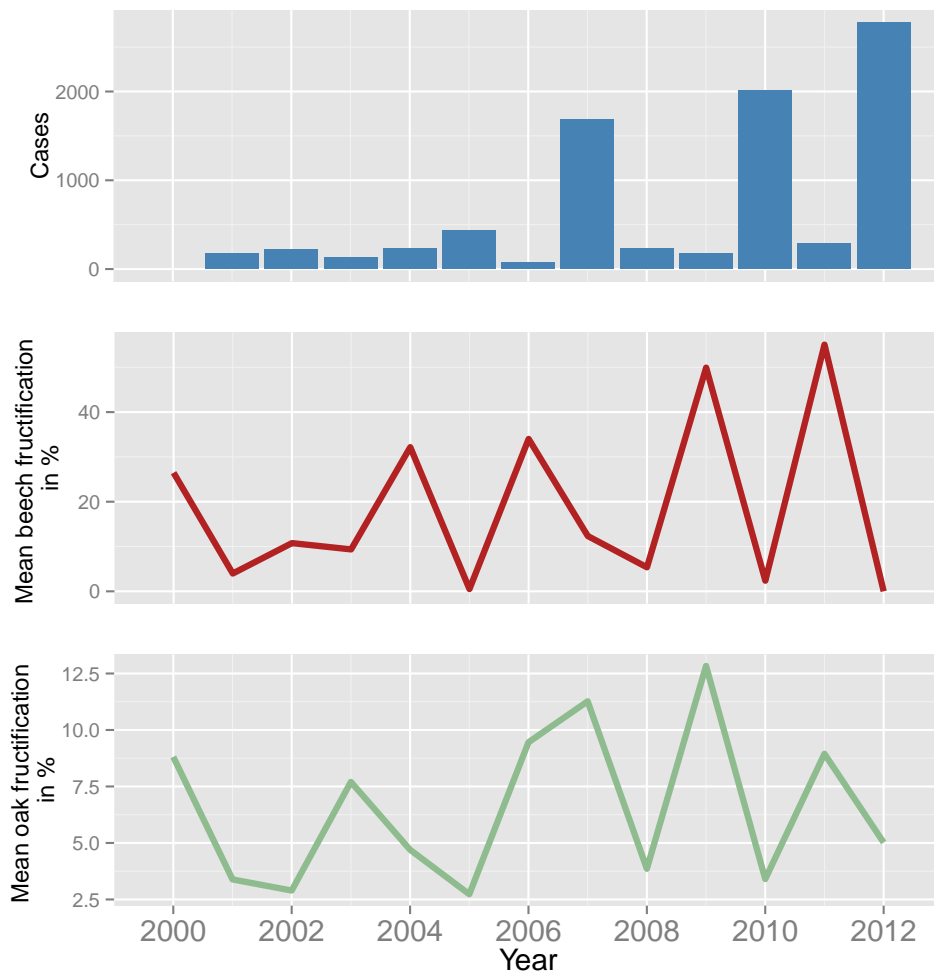


Figure 1.1: The top plot shows the number of reported hantavirus cases per year. The middle and lower plots show the yearly weighted mean fructification of all counties for beech and oak, where weights are given by counties' area.

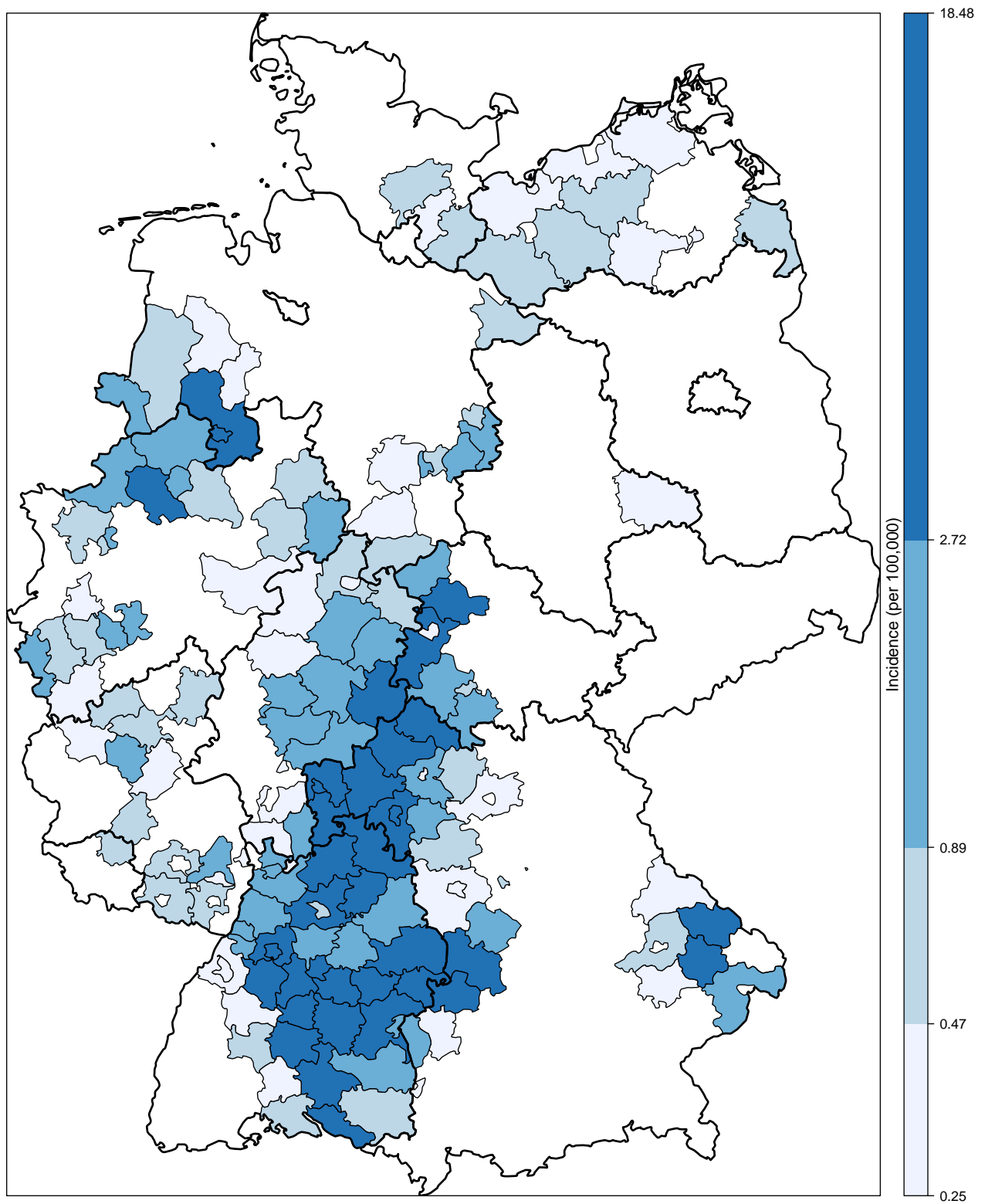


Figure 1.2: Overall incidence 2002-2012 (per 100,000 inhabitants) for each county

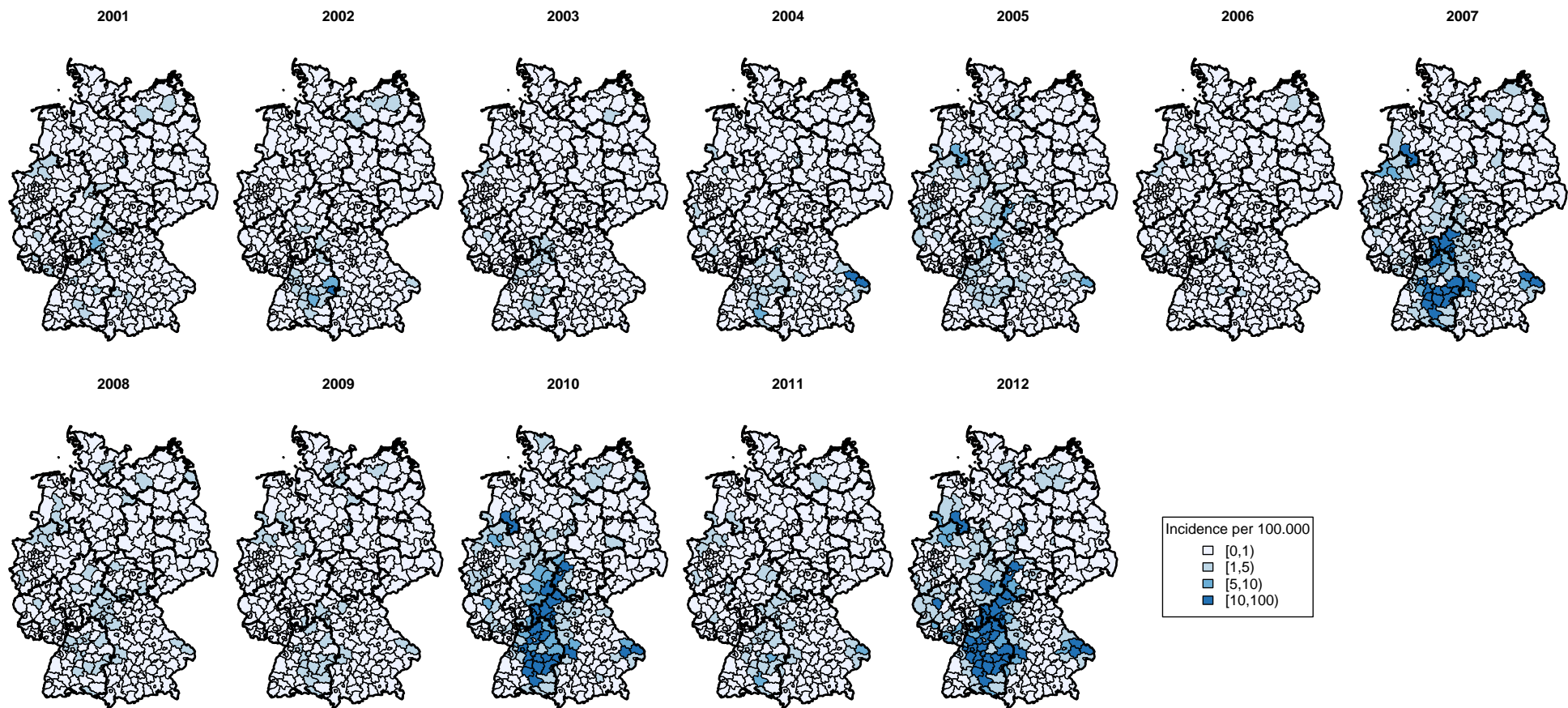


Figure 1.3: Yearly incidence (per 100,000 inhabitants) of hantavirus in the 412 districts of Germany.

1.1.2 CORINE Land Cover

The Coordination of Information on the Environment (CORINE) Land Cover is a project initiated by the European Commission and executed by the European Environment Agency that aims to provide land cover information for countries in the European Union. This is done through the use of satellite images. The land is divided into 44 different land cover classes with a resolution of 1:100,000. We will use three land cover classes as covariates: fields (German: Wiesenweiden), grassland (Grasland) and bushes (Waldstrauch). The land cover data are freely available at <http://www.eea.europa.eu>. We get the proportion of county area that is covered by three different land cover classes in Figure 1.4.

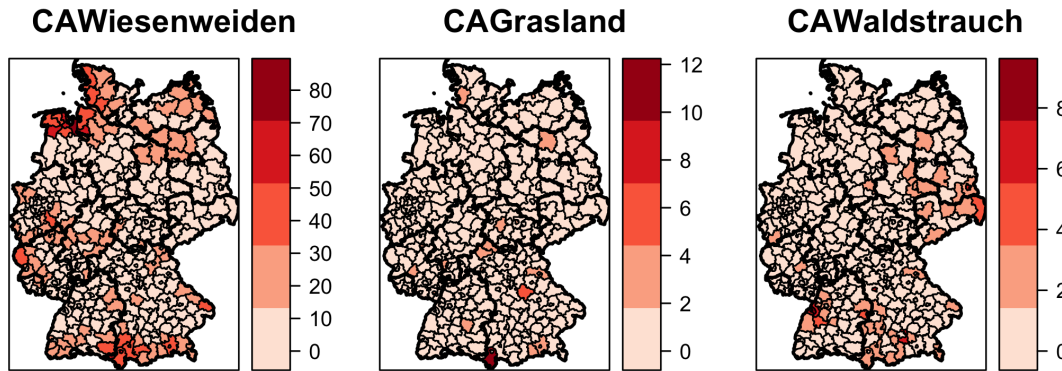


Figure 1.4: Proportion of county area (measured in %) occupied by one of the three CORINE land cover classes: fields (Wiesenweiden), grassland (Grasland) and bushes (Waldstrauch).

1.1.3 Forest Area

The Federal Forest Inventory is performed every ten years by the Thünen-Institut in accordance with Federal Forest Act and provides information as to the number and volume of trees in Germany. The most recent version was done in 2011/2012 and is freely available at <http://www.bundeswaldinventur.de>. For the inventory all of Germany is overlaid by a $4\text{km} \times 4\text{km}$ grid, with some federal states having double or quadruple that intensity. At each grid point there are four sampling points, the corners of a $150\text{m} \times 150\text{m}$ square around the grid point. In total, at around 60,00 sample points an approximate 420,000 trees are measured. This is extrapolated to an estimated total of ninety billion trees in Germany. Fig. 1.5 shows the area covered by oak, beech and fir for each county. In the analysis we also consider **ARest**, which is the remaining forest area of the total forest area that is not covered by beech, oak or fir.

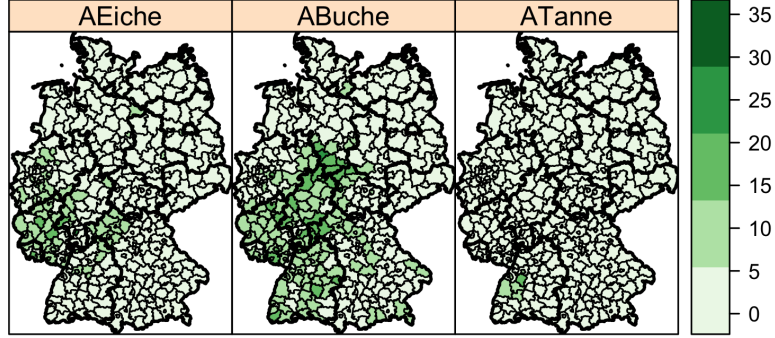


Figure 1.5: Percentage of county area covered by oak (Eiche), beech (Buche) and fir (Tanne), respectively.

1.1.4 Fructification

Fructification is measured yearly in several locations in Germany. In total we have 2157 measurements of beech fructification and 1376 measurements of oak fructification, both from 2000 to 2012. Following the approach of Faber and Höhle (2013), these measurements are then used to compute a smoothed 2D-spline for all of Germany and finally the values of the spline at the centroid of each county are taken. This 2D-spline is necessary as we otherwise could not have fructification values for each county, because the sampling points are spread irregularly across Germany, see Figures 1.6 and 1.7. This is done for both beech and oak, as these were determined to be the most important food sources for bank voles.

As this is one of the most important variables, we first describe the extrapolation and then extend upon it. The fructification status of each tree type is determined yearly at each sample location. Four different statuses are possible: non, weak, medium, strong fructification.

We now consider this for beeches. Let $n(\mathbf{s}_i, t)$ be the number of beeches sampled at location $i = 1, \dots, n_t$, where $\mathbf{s}_i = (s_{ix}, s_{iy})'$, and at time $t \in \{2000, \dots, 2012\}$. Let then $n_{\text{heavy}}(\mathbf{s}_i, t)$ denote the number of beeches with fructification statuses medium or strong. We then get the proportion of trees with medium and high fructification by

$$p(\mathbf{s}_i, t) = \frac{n_{\text{heavy}}(\mathbf{s}_i, t)}{n(\mathbf{s}_i, t)}.$$

These proportions are shown in Figures 1.6 and 1.7.

We now construct a smoothing spline for each year. The distribution of $n_{\text{heavy}}(\mathbf{s}_i, t)$ is modeled through a binomial distribution in a generalized additive model (GAM), see also

section 2.1, as follows

$$n_{\text{heavy}}(\mathbf{s}_i, t) \sim \text{Bin}(n(\mathbf{s}_i, t), p(\mathbf{s}_i, t)), \text{ with} \\ \text{logit}(p(\mathbf{s}_i, t)) = f_t(\mathbf{s}_i),$$

where $f_t(\mathbf{s})$ is the two-dimensional smoothing thin plate spline for year t . For more details on the fitting of such spline models, consult Wood (2006).

Lastly, we take the value of the smoothing spline at the centroid of each county through

$$\text{hfruc}_{jt} = \hat{p}(\mathbf{s}_j, t) = \text{logit}^{-1}(\hat{f}_t(\mathbf{s}_j)),$$

where \mathbf{s}_j , $j = 1, \dots, 412$ denotes the location of the centroid of region j . Figure 1.8 shows the estimated proportion of middle and high statuses for each county. The same process is done for oaks and the equivalent plot is given in Figure 1.9. In the analysis, a county's fructification value is given in percent.

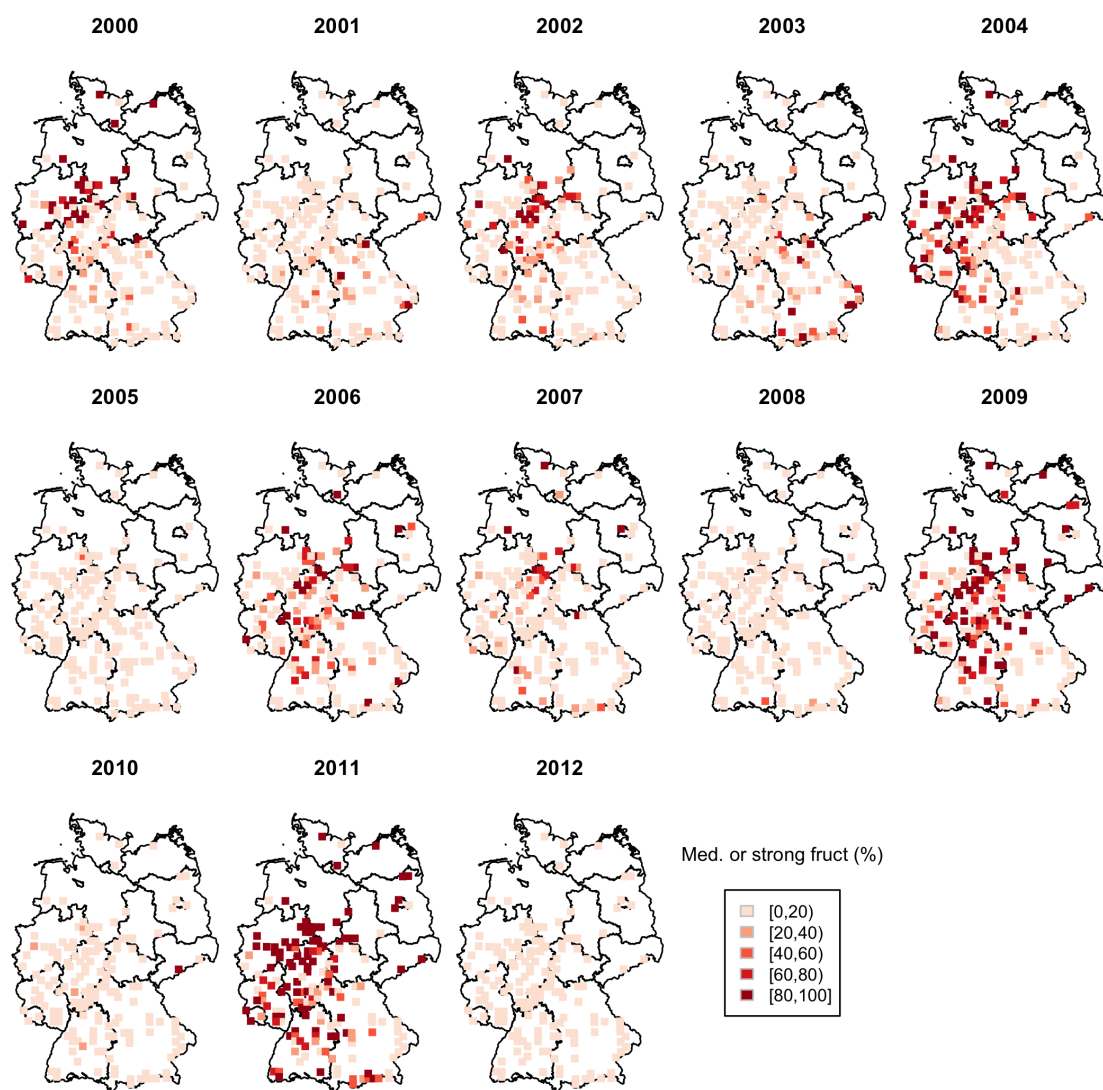


Figure 1.6: Fructification index of the common beech in Germany 2000–2012 at the observed sample points.

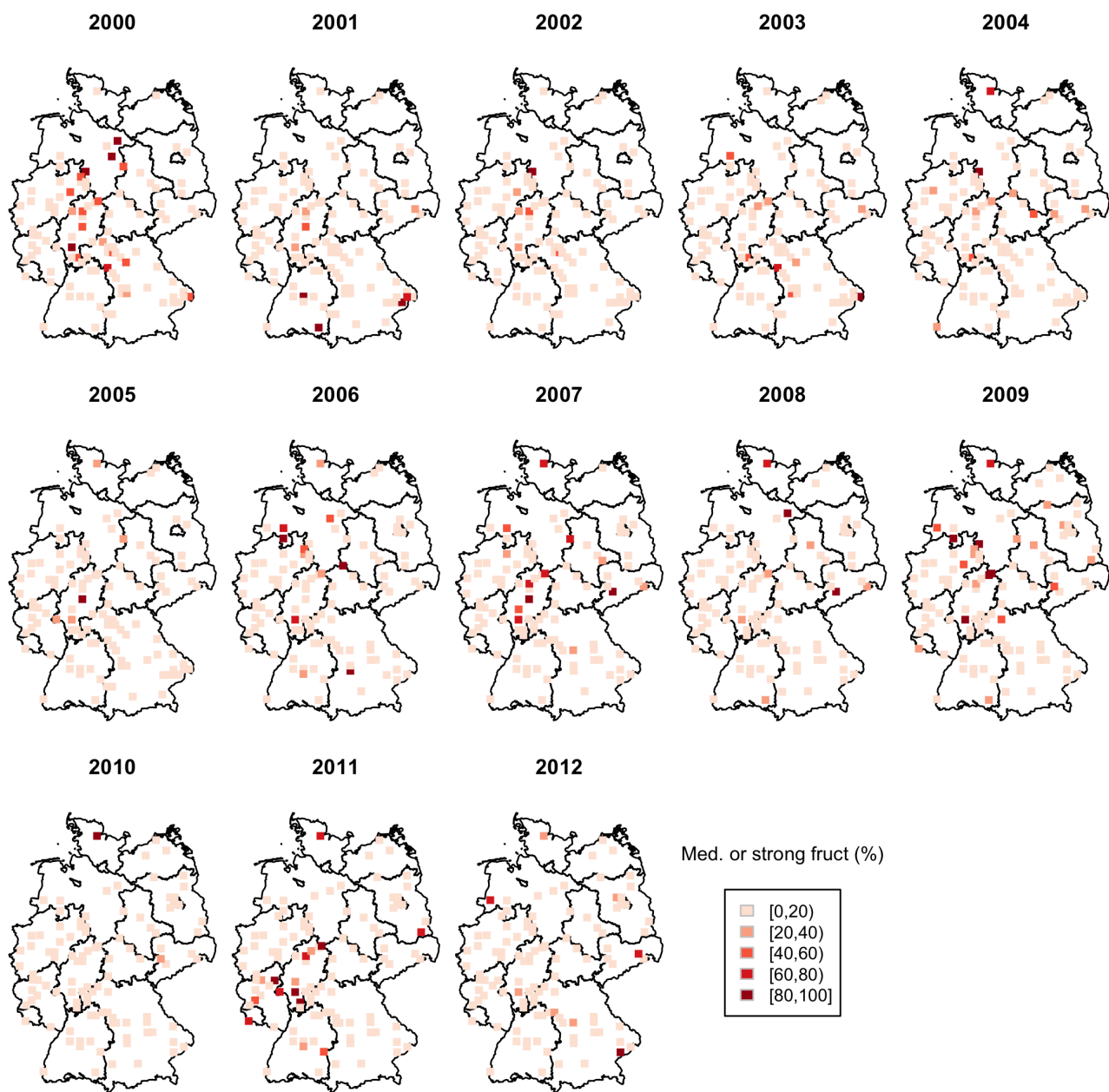


Figure 1.7: Fructification index of oak in Germany 2000–2012 at the observed sample points.

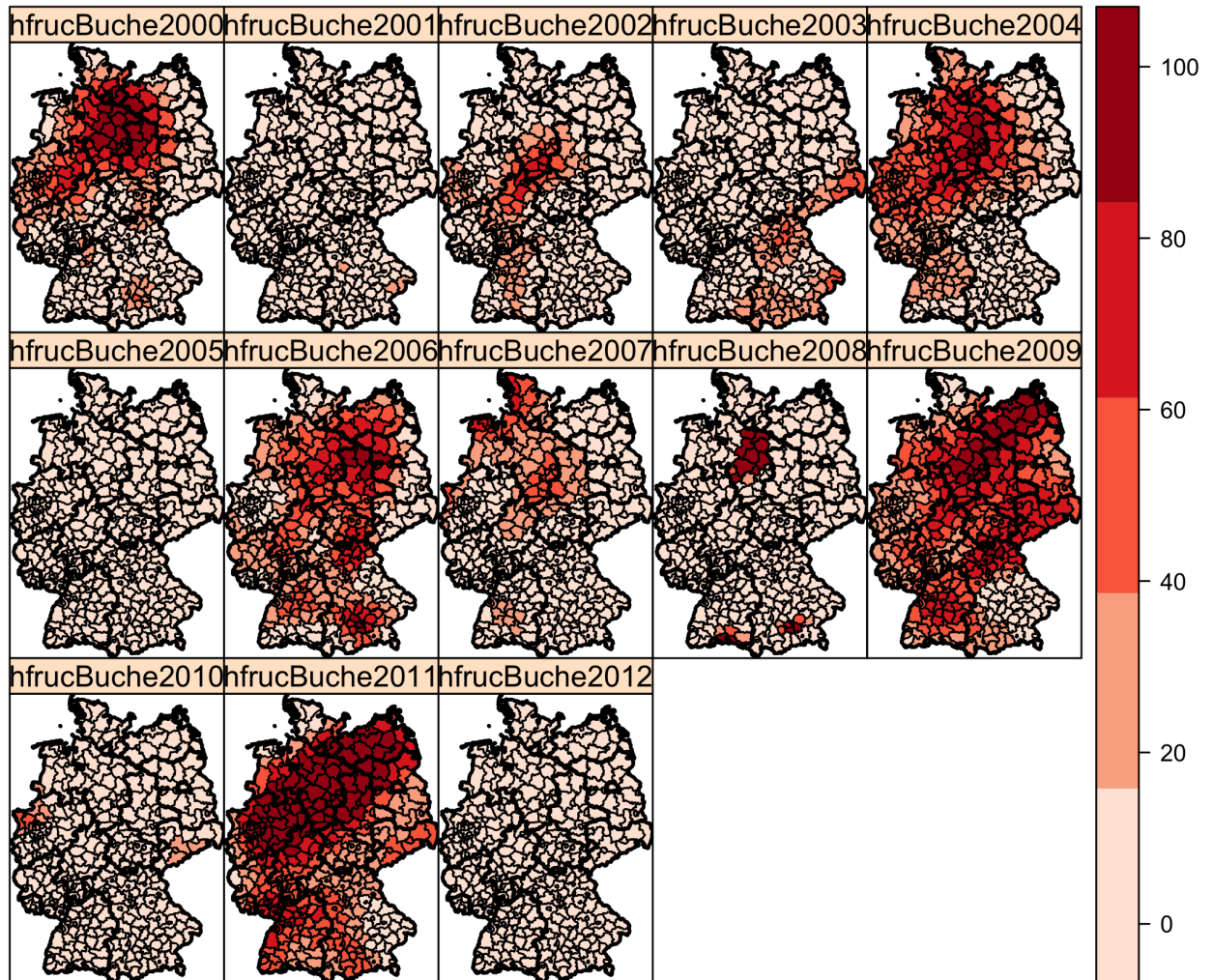


Figure 1.8: Extrapolated beech fructification for each of the 412 districts in Germany for the years 2000–2012.

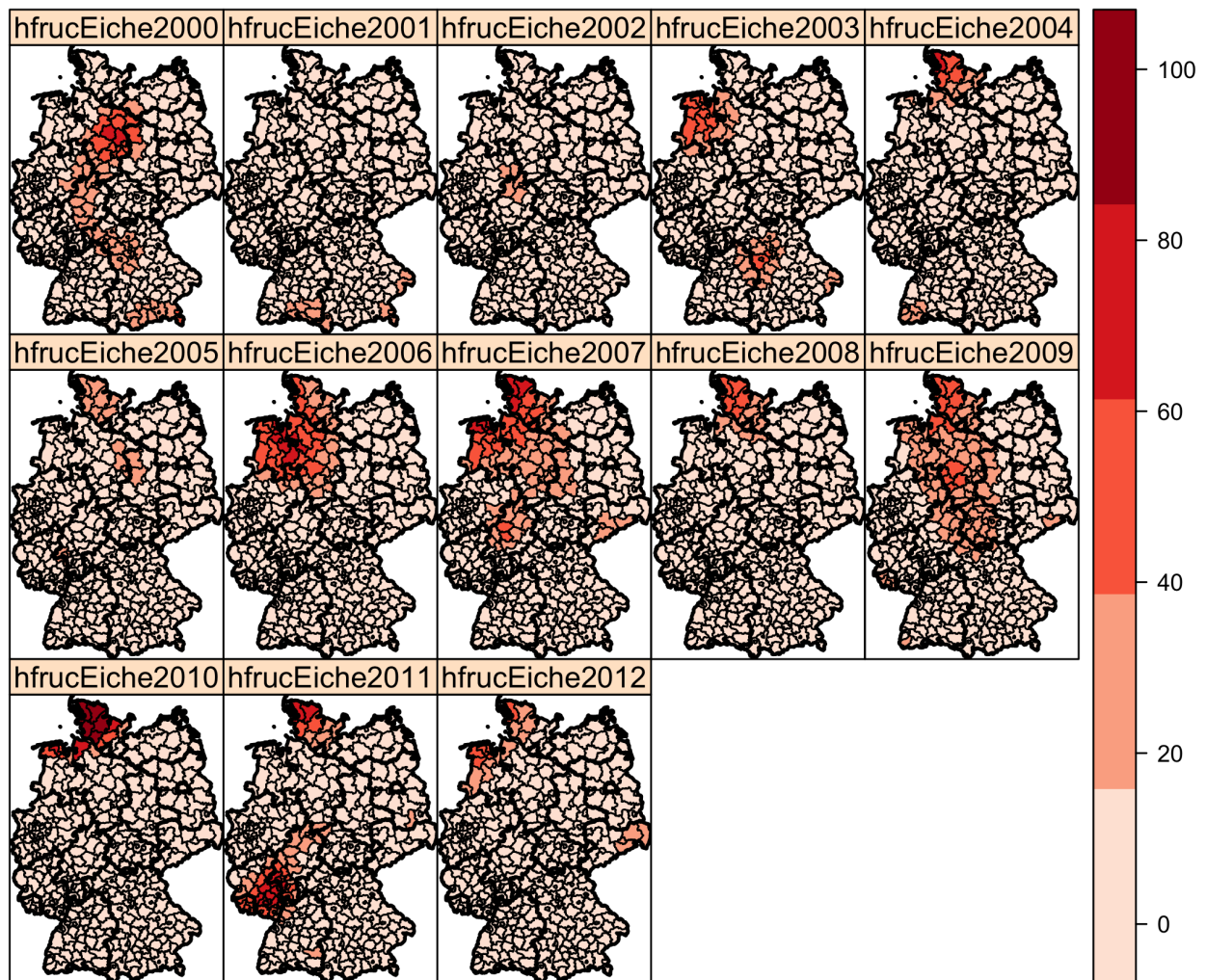


Figure 1.9: Extrapolated fructification of oak for each of the 412 districts in Germany for the years 2000–2012.

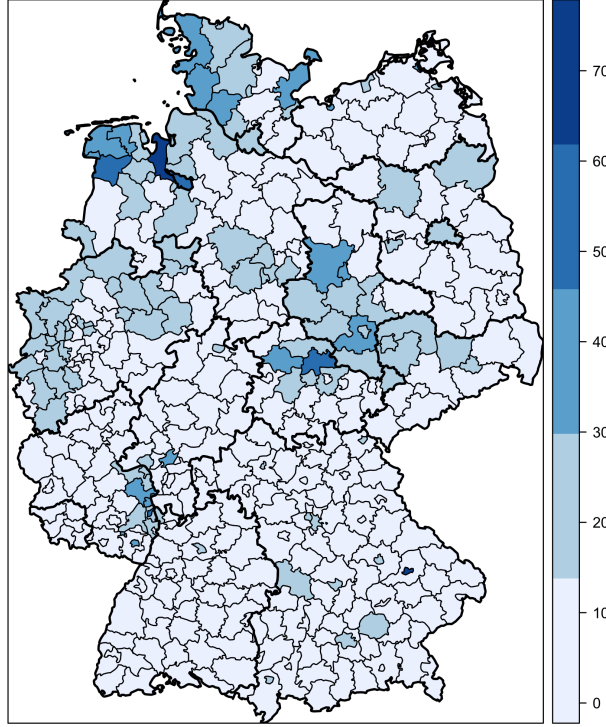


Figure 1.10: Mean distance (in km) of urban areas to forests, i.e. `u2fc` index, for each county.

1.1.5 Urban Proximity

The last covariate we include in the model is the proximity between urban and forest areas. To compute these distances, CORINE Land Cover data are used. Two types of areas that are comprised of several land cover classes each are considered, namely 'urban fabric' (i.e. continuous urban fabric, discontinuous urban fabric) and 'forests' (i.e. broad-leaved forest, coniferous forest, mixed forest).

A proximity variable for a county i , where $i \in \{1, \dots, 412\}$, can then be computed as the average distance of urban areas to forests in that county, i.e.

$$\text{u2fc}_i = \frac{1}{|U_i|} \sum_{s \in U_i} \text{dist}(s, F),$$

where U_i denotes the set of all 'urban' pixels in region i and $\text{dist}(s, F)$ denotes the Euclidean distance of pixel s to the nearest forest pixel. Since only German CORINE data are considered, small edge effects at the country border might appear. A map with the counties' average proximity is given in Figure 1.10. Lower values mean closer average proximity of urban areas to forests.

Chapter 2

Prerequisites for the Statistical Modeling

2.1 Bayesian Generalized Additive Models

In what follows we will assume a level of familiarity with Bayesian statistics as presented in Carlin and Louis (2011) and with generalized additive models (GAMs) as given in Fahrmeir et al. (2013). See for example Fahrmeir and Kneib (2011) for more details on GAMs or Hastie and Tibshirani (1990), where GAMs were first introduced. The expressions 'generalized additive model' and 'structured additive regression' will sometimes be used interchangeably in this thesis, see chapter 2 in Fahrmeir et al. (2013) for an overview. This section will provide a very brief recap of GAMs and how they will be used by INLA in chapter 3.

INLA deals with latent Gaussian models, a subclass of structured additive regression models. These models are formulated very similarly to generalized linear models (GLMs) but the linear predictor is replaced by a structured additive predictor which can include nonlinear effects. As in a GLM, we model our observations, the data $\mathbf{y} = (y_1, \dots, y_{n_d})^T$, through a marginal distribution of the one-parametric exponential family. In the model used in this thesis, this will be a Poisson distribution, i.e.

$$y_i | \eta_i \sim \text{Poi}(\mu_i),$$

where μ_i is connected to the structured additive predictor η_i , and therefore the covariates, through a link function. In our model we use the log function as a link function, which is the canonical link function for the Poisson distribution. We then have

$$\log(\mu_i) = \eta_i,$$

where η_i is the structured additive predictor. This predictor contains all covariates and can be written as

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i. \quad (2.1)$$

The variables that make up the predictor are an intercept α , nonlinear terms $f^{(j)}$, linear terms β_k and an error term ϵ_i . Therefore $\alpha, \{\beta_k\}_{k \in n_\beta}$ and $\{\epsilon_i\}_{i \in n_d}$ can be understood similarly to their GLM counterparts. The additional nonlinear terms $f^{(j)}$, however, can take a wide variety of forms. We will only be using Gaussian Markov random fields, see chapter 2.2, to account for spatial effects, but a wide variety of random effects, splines and other nonlinear functions are possible, see e.g. chapters 8 and 9 in Fahrmeir et al. (2013).

In a Bayesian setting $\alpha, \{\beta_k\}_k, f^{(j)}, \{\epsilon_i\}_i$ or equivalently $\alpha, \{\beta_k\}_k, f^{(j)}, \{\eta_i\}_i$ have prior distributions. As INLA deals mainly with latent Gaussian fields, we usually assign vague Gaussian priors, e.g. $N(0, 1000^2)$, to $\alpha, \{\beta_k\}_k, \{\eta_i\}_i$. Furthermore, in our case the prior of $\{f^{(j)}\}_j$ is a Gaussian Markov random field dependent on a hyperparameter θ whose hyperprior is a non-informative log-gamma distribution.

2.2 Gaussian Markov Random Fields

We now present the important concept of Gaussian Markov random fields (GMRFs). As they are used in many different situations, presenting them from two different viewpoints might be beneficial to their understanding. The first viewpoint sees GMRFs as random vectors with very efficient computational properties, whereas the second sees them as a nonlinear term f in a structured additive predictor that helps model temporal or spatial dependencies.

The viewpoint of computational efficiency, as presented in section 2.2.1, is taken in the theoretical part of this thesis and throughout Rue et al. (2009), where much of the superior speed of the INLA method relies on the computationally beneficial properties of GMRFs. The second viewpoint, section 2.2.2, as introduced by Besag et al. (1991) and as taken in the practical part of this thesis, in Fahrmeir et al. (2013) and in many other resources on temporal, spatial and spatio-temporal modeling, focuses on presenting the distribution of the random vector through the full conditionals as an intuitive and interpretable way of modeling.

2.2.1 Definition through Joint Density

A random vector \mathbf{x} follows a Markov Random Field if the marginal distribution of each element only depends on its immediate neighbors, where neighbors are defined according to a given problem, e.g. predecessor and successor in a time series or counties sharing a border in spatial applications. If we denote two elements i and j being neighbors by $i \sim j$ and the set of all neighbors of i by $N(i)$, then we can state the Markov property for x_i by

$$x_i \perp x_k | \{x_j\}_{j \in N(i)} \quad \text{for all } k, \text{ where } k \text{ is not a neighbor of } i. \quad (2.2)$$

Here, $x \perp y | z$ denotes the conditional independence of random variables x and y , i.e. x and y are independent given a third random variable z .

If \mathbf{x} also follows a multivariate Gaussian distribution we say that \mathbf{x} is a Gaussian Markov

random field. These GMRFs have several useful properties. As the entries in the precision matrix (inverse covariance matrix) of a multivariate Gaussian distribution imply the conditional independence structures, the precision matrices are usually sparse and computationally very efficient methods can be used, see Rue and Held (2005).

We define a random vector $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ as a Gaussian Markov random field with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a neighborhood structure, where \mathcal{V} is the set of nodes and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ the set of edges. The mean of \mathbf{x} is given by $\boldsymbol{\mu}$ and its precision matrix by \mathbf{Q} . Then \mathbf{x} is a GMRF if and only if its probability density function can be written as

$$\pi(\mathbf{x}) = (2\pi)^{n/2} \det(\mathbf{Q})^{1/2} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}) \right) \quad (2.3)$$

and

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathcal{E} \text{ for all } i \neq j.$$

We can also write the second condition as $Q_{ij} \neq 0 \Leftrightarrow i \sim j$. Note that a node cannot be a neighbor of itself and that the precision matrix \mathbf{Q} in this definition is positive definite.

2.2.2 Definition through Full Conditionals

Gaussian Markov random fields as part of the structured additive predictor to model spatial and temporal dependencies were introduced by Besag et al. (1991) and are now a common tool in spatial statistics, see for example Schrödle and Held (2011a) or Schrödle and Held (2011b). They are also referred to as conditionally autoregressive (CAR) priors. These GMRFs are usually *intrinsic* Gaussian Markov random fields, i.e. their precision matrix does not have full rank, see chapter 3 in Rue and Held (2005) for more details.

We use the notation $f(s_i) = \gamma_i$ to mean the spatial effect of region s_i . In the spatial case, with $i = 1, \dots, n_i$ spatial regions, we can specify the GMRF through the full conditional of γ_i

$$\pi(\gamma_i | \boldsymbol{\tau}) \propto \exp \left(-\frac{\tau}{2} \sum_{j \sim i} (\gamma_i - \gamma_j)^2 \right), \quad (2.4)$$

where τ is a precision parameter. It is also possible to choose a different loss function than the quadratic loss function and add symmetric weights between regions, e.g. common border length or euclidean distance of centroids. For more details on how to define valid GMRFs through full conditionals see chapter 2.2.4 in Rue and Held (2005).

We can then write the the distribution of the γ_i conveniently as

$$\gamma_i | \boldsymbol{\gamma}_{-i}, \boldsymbol{\tau} \sim N \left(\frac{1}{|N(i)|} \sum_{j \in N(i)} \gamma_j, \frac{1}{|N(i)| \cdot \tau} \right). \quad (2.5)$$

Here $\boldsymbol{\gamma}_{-i}$ denotes the set of all γ_j , with $j \neq i$. These γ_i have of course the spatial Markov property that the distribution of the γ_i only depends on γ_j , $j \in N(i)$. This way of formulating the GMRF through the full conditionals has an immediate interpretability, which allows their direct use in spatial modeling.

2.3 Model Selection

This section presents two common tools for model selection and evaluation that will be used in the applied part of this thesis.

2.3.1 Deviance Information Criterion

The deviance information criterion (DIC) was introduced by Spiegelhalter et al. (2002) and can, similarly to the AIC, be helpful in model selection as it provides a measure of both the fit and the complexity of the model. According to Fahrmeir and Kneib (2011) it can be used to evaluate the fit of a generalized additive model. The DIC is defined as

$$DIC = \bar{D} + p_D, \quad (2.6)$$

where \bar{D} is the posterior mean of the deviance, a measure of model fit and p_D is the number of effective parameters, a measure of model complexity. Ideally we would like both \bar{D} and p_D to be small and would therefore select the model with the smallest DIC value.

2.3.2 Probability Integral Transforms

Probability integral transforms (PIT) are a tool to check model validity. For a count data setting an adjusted version was introduced by Czado et al. (2009). PITs are based on a basic theorem in probability that states that a random variable $Y = F_X(X)$ is distributed uniformly on $[0, 1]$ for any continuous random variable X with cumulative distribution function F_X . This follows from

$$P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

Assume now that the \mathbf{y} come from a continuous distribution. If we now define the PITs as the values of the predictive cumulative distribution function at the observed values, i.e. $PIT_i = P(y_i^{new} \leq y_i | \mathbf{y}_{-i})$ (Martino and Rue, 2010), it follows that the PITs should be uniformly distributed if the observations were drawn from the predictive distribution. To assess model fit, a histogram of these PIT values can be plotted. It can then be visually inspected to check for deviations from uniformity.

Chapter 3

Integrated Nested Laplace Approximations

3.1 Outline

Integrated nested Laplace approximations (INLAs) were introduced by Rue et al. (2009) and expanded upon in Martins et al. (2013). They are used to directly approximate marginal posterior densities in Bayesian inference. As the marginal posterior density $\pi(x_i|\mathbf{y})$ can generally not be computed analytically, an approximation is necessary. This is nowadays most commonly done by Markov chain Monte Carlo methods (MCMC) (Held et al., 2010). INLAs offer a computationally efficient alternative for the calculation of marginal posterior densities.

INLAs work on latent Gaussian models, a subclass of structured additive regression models, see section 2.1. These models are hierarchical with the likelihood $\pi(\mathbf{y}|\mathbf{x})$, from the exponential family, as the first stage, the latent Gaussian field $\pi(\mathbf{x}|\boldsymbol{\theta})$ as the second stage and the (non-Gaussian) hyperpriors $\pi(\boldsymbol{\theta})$ as the third stage. Here, \mathbf{y} denotes the observations. The second stage, the latent Gaussian field, is a GMRF. The random vector \mathbf{x} consists of all terms of the structured additive predictor which have Gaussian priors, i.e.

$$\mathbf{x} = \left((\eta_i)_{i=1}^{n_d}, \alpha, (f^{(j)})_{j=1}^{n_f}, (\beta_k)_{k=1}^{n_\beta} \right)^T.$$

Assume that the vector \mathbf{x} has total length n .

Requirements for the INLA method are that the random vector with Gaussian priors \mathbf{x} can be very high dimensional but the vector of hyperparameters $\boldsymbol{\theta}$, should be low dimensional, practically feasible are dimensions up to around six. Integrated nested Laplace approximations are fast in comparison with MCMC because many direct approximations are used nested within each other but also because they use the computationally beneficial properties of Gaussian Markov random fields, i.e. the fact that \mathbf{Q} , the precision matrix of \mathbf{x} , is sparse and therefore computationally very efficient methods for sparse matrices can be used.

Note that we assume

$$\pi(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n_d} \pi(y_i|x_i)$$

or equivalently $\pi(y_i|\mathbf{x}) = \pi(y_i|x_i)$, meaning that every observation y_i in $\mathbf{y} = (y_1, \dots, y_{n_d})^T$ depends only on the one corresponding entry of the latent Gaussian field. Since the latent Gaussian field is usually larger than the number of data points, i.e. $n > n_d$, only the first n_d entries of \mathbf{x} are necessary to calculate the distribution of \mathbf{y} . These first n_d entries of \mathbf{x} correspond to $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{n_d})^T$. That means that \mathbf{y} is conditionally independent of $\{x_i\}_{i=n_d+1}^n$ given $\boldsymbol{\eta}$. More formally we have

$$\mathbf{y} \perp \{x_i\}_{i=n_d+1}^n | \boldsymbol{\eta}.$$

Using all this, we can now write the joint posterior of $\mathbf{x}, \boldsymbol{\theta}$ using Bayes' theorem

$$\begin{aligned} \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) &\propto \pi(\boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i=1}^{n_d} \pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) \det(\mathbf{Q}(\boldsymbol{\theta}))^{\frac{n}{2}} \exp \left(-\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n_d} \log\{\pi(y_i|x_i, \boldsymbol{\theta})\} \right). \end{aligned} \quad (3.1)$$

This is the joint posterior distribution of all model parameters and as such the aim in conducting Bayesian inference. Before we continue to introduce the INLA method we now present two approximations that are used extensively in the INLA method.

3.2 Gaussian Approximations

Gaussian approximations of posterior densities are computationally very fast, but might not be very accurate if the posterior density is not sufficiently Gaussian, see also Rue and Martino (2007). The density is approximated iteratively by matching the mode and the curvature at the mode. This equals using the Newton-Raphson procedure. The Gaussian approximations presented here are also referred to as GMRF approximations.

We now explain the method outlined in Rue and Held (2005). Assume that we have a GMRF \mathbf{x} with precision matrix \mathbf{Q} and mean zero. We then approximate a density in the form of (3.1),

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \exp \left(-\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i=1}^{n_d} \log\{\pi(y_i|x_i, \boldsymbol{\theta})\} \right). \quad (3.2)$$

We now expand the second term componentwise using a Taylor-expansion around $\mu_i^{(0)}$, the i -th entry of our starting point $\boldsymbol{\mu}^{(0)}$, which gives us

$$\log\{\pi(y_i|x_i, \boldsymbol{\theta})\} \approx \log\{\pi(y_i|\mu_i^{(0)}, \boldsymbol{\theta})\} + b_i x_i - \frac{1}{2} c_i x_i^2. \quad (3.3)$$

Inserting this we get

$$\tilde{\pi}(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \left(-\frac{1}{2} \mathbf{x}^T (\mathbf{Q} + \mathbf{c}) \mathbf{x} + \mathbf{b}^T \mathbf{x} \right), \quad (3.4)$$

which is normally distributed with precision matrix $\mathbf{Q} + \text{diag}(\mathbf{c})$ and mode $\boldsymbol{\mu}^{(1)}$ which is the solution of $\{\mathbf{Q} + \text{diag}(\mathbf{c})\} \boldsymbol{\mu}^{(1)} = \mathbf{b}$. This approximation relies of course on the starting value $\boldsymbol{\mu}^{(0)}$. It is more accurate the closer $\boldsymbol{\mu}^{(0)}$ is to the mode of $\pi(\mathbf{x})$. The process can then be iterated, with $\boldsymbol{\mu}^{(1)}$ as the new starting value, until an approximation of the required degree of accuracy is reached. Note that the precision matrix is only changed along its diagonal, which means that the Gaussian approximation retains the conditional independence structure of $\pi(\mathbf{x})$ and is a GMRF with all its computationally beneficial properties.

3.3 Laplace Approximations

Laplace approximations as an approximation of certain integrals through a Taylor series have been introduced by Laplace, see Laplace (1986). Laplace approximations were first applied to moments of probability densities and marginal posterior distributions by Tierney and Kadane (1986). We will use the name 'Laplace approximation' interchangeably for those two different concepts. When we talk about a Laplace approximation of a density we will mean the approximation of this density based on the use of the classical Laplace approximation as presented in this section.

Multivariate Laplace approximations for the integral of a scalar function $f(\mathbf{x})$, with \mathbf{x} a d -dimensional vector, have the following form

$$\int e^{nf(\mathbf{x})} d\mathbf{x} \approx \left(\frac{2\pi}{n} \right)^{d/2} \det(-H(f)(\mathbf{x}_0))^{-1/2} e^{nf(\mathbf{x}_0)}, \quad (3.5)$$

where \mathbf{x}_0 is the global maximum and $H(f)(\mathbf{x})$ is the Hessian of $f(\mathbf{x})$. The idea being that for big $n \in \mathbb{N}$, the integral has nearly all of its weight around the maximum.

Tierney and Kadane (1986) used this as an approximation for posterior densities $\pi(\boldsymbol{\theta} | \mathbf{y})$, where $\boldsymbol{\theta}$ is the parameter vector and \mathbf{y} the data. We state the derivations here in more detail: The parameter vector $\boldsymbol{\theta} = (\theta_1, \boldsymbol{\theta}_2)$ is split up into a 1-dimensional and a $d - 1$ -dimensional part. The marginal posterior density for θ_1 can then be written as

$$\pi(\theta_1 | \mathbf{y}^{(n)}) = \frac{\int \pi(\theta_1, \boldsymbol{\theta}_2) L(\mathbf{y} | \theta_1, \boldsymbol{\theta}_2) d\boldsymbol{\theta}_2}{\int \pi(\boldsymbol{\theta}) L(\mathbf{y} | \boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{\int e^{n \cdot g_1(\boldsymbol{\theta}_2)} d\boldsymbol{\theta}_2}{\int e^{n \cdot g(\boldsymbol{\theta})} d\boldsymbol{\theta}}, \quad (3.6)$$

where

$$g(\boldsymbol{\theta}) = (\log(\pi(\boldsymbol{\theta})) + \log(L(\mathbf{y} | \boldsymbol{\theta}))) / n$$

is just a rewrite to bring the likelihood $L(\mathbf{y} | \theta_1, \boldsymbol{\theta}_2)$ and the prior $\pi(\theta_1, \boldsymbol{\theta}_2)$ in the form necessary for the Laplace approximation, i.e. equation (3.5). Furthermore, g_1 is g with θ_1 held fixed, i.e.

$$g_1(\boldsymbol{\theta}_2) = (\log(\pi(\theta_1, \boldsymbol{\theta}_2)) + \log(L(\mathbf{y} | \theta_1, \boldsymbol{\theta}_2))) / n.$$

Let $\hat{\boldsymbol{\theta}}$ be the posterior mode which maximizes $\pi(\boldsymbol{\theta})L(\mathbf{y}|\boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_2^* = \hat{\boldsymbol{\theta}}_2^*(\theta_1)$ the posterior mode for a fixed θ_1 , which correspondingly maximizes $\pi(\theta_1, \boldsymbol{\theta}_2)L(\mathbf{y}|\theta_1, \boldsymbol{\theta}_2)$ for a given θ_1 . Let then $\mathcal{J}_1(\hat{\boldsymbol{\theta}}_2^*) = \mathcal{J}_1(\theta_1)$ denote the observed Fisher information matrix, i.e. the negative Hessian of g_1 evaluated at $\hat{\boldsymbol{\theta}}_2^*$. Note that we can see \mathcal{J}_1 as a function of $\hat{\boldsymbol{\theta}}_2^*$ or of θ_1 , which we will both use. Now apply the Laplace approximation (3.5) to the numerator of (3.6)

$$\left(\frac{2\pi}{n}\right)^{(d-1)/2} \det(\mathcal{J}_1(\hat{\boldsymbol{\theta}}_2^*))^{-1/2} e^{n \cdot g_1(\hat{\boldsymbol{\theta}}_2^*)} \quad (3.7)$$

and the denominator

$$\left(\frac{2\pi}{n}\right)^{d/2} \det(\mathcal{J}(\hat{\boldsymbol{\theta}}))^{-1/2} e^{n \cdot g(\hat{\boldsymbol{\theta}})} \quad (3.8)$$

with $\mathcal{J}(\hat{\boldsymbol{\theta}})$ being the observed Fisher information of g evaluated at $\hat{\boldsymbol{\theta}}$. This then gives us the Laplace approximation for the marginal density

$$\hat{\pi}(\theta_1|\mathbf{y}^{(n)}) = \left(\frac{n \det(\mathcal{J}_1^{-1}(\theta_1))}{2\pi \det(\mathcal{J}^{-1}(\hat{\boldsymbol{\theta}}))}\right)^{1/2} \frac{\pi(\theta_1, \hat{\boldsymbol{\theta}}_2^*)L(\mathbf{y}|\theta_1, \hat{\boldsymbol{\theta}}_2^*)}{\pi(\hat{\boldsymbol{\theta}})L(\mathbf{y}|\hat{\boldsymbol{\theta}})} \quad (3.9)$$

Laplace and Gaussian approximations of posterior densities are widely used in the INLA method and allow us to now introduce its main ideas.

3.4 Integrated Nested Laplace Approximations

We now explain the INLA method as described in Rue et al. (2009). Many details will be clarified in much greater detail and we present proofs omitted from the original paper. The model setup is taken from section 3.1 with \mathbf{y} the data, \mathbf{x} the latent Gaussian field, i.e. the parameters in the distribution of \mathbf{y} , and $\boldsymbol{\theta}$ the hyperparameters.

The goal in conducting Bayesian inference is to obtain the posterior density of the parameters, either the full posterior 3.1 or the marginal posterior densities. We will now focus on the latter, the marginal posterior density $\pi(x_i|\mathbf{y})$, which can be obtained by integrating out the hyperparameters

$$\pi(x_i|\mathbf{y}) = \int \int \pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x}_{-i} d\boldsymbol{\theta} = \int \pi(x_i|\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \quad (3.10)$$

where we used the notation \mathbf{x}_{-i} for the vector \mathbf{x} without the i -th element. Additionally, let m denote the dimension of $\boldsymbol{\theta}$. Expression (3.10) is almost always impossible to compute analytically. Instead, we try to approximate it by an approximate density $\tilde{\pi}(x_i|\mathbf{y})$. The idea of INLA is to do this approximation by using several approximations nested within each other. First, consider a usual multivariate quadrature approach where we approximate the m -dimensional integral on the right-hand side of expression (3.10) as a weighted sum by

$$\begin{aligned}
\pi(x_i|\mathbf{y}) &\approx \sum_{k=1}^{n_k} \pi(x_i|\boldsymbol{\theta}_k, \mathbf{y}) \pi(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k \\
&\approx \sum_{k=1}^{n_k} \tilde{\pi}(x_i|\boldsymbol{\theta}_k, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}_k|\mathbf{y}) \Delta_k =: \tilde{\pi}(x_i|\mathbf{y}),
\end{aligned} \tag{3.11}$$

using a number n_k of support points $\boldsymbol{\theta}_k$ with corresponding area weights Δ_k . Details on these support points are given in section 3.5.

In equation (3.11) we introduced two new approximations. The first one, $\tilde{\pi}(x_i|\mathbf{y}, \boldsymbol{\theta})$, is either a Gaussian, a Laplace or a simplified Laplace approximation of $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$. We will return to these in more detail in section 3.6.

The second one, $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$, is a Laplace approximation of $\pi(\boldsymbol{\theta}|\mathbf{y})$, which we will discuss now. To obtain this Laplace approximation we firstly see that from

$$\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta}|\mathbf{y}) \pi(\mathbf{y})$$

we can obtain

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})}. \tag{3.12}$$

The numerator of (3.12) is easily computed, since

$$\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) = \pi(\mathbf{y}|\mathbf{x}) \pi(\mathbf{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}).$$

All of which we know from the model formulation in section 3.1.

The denominator of (3.12), however, we do not know. We therefore replace $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ by its Gaussian approximation $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and evaluate the whole fraction at $\mathbf{x}^*(\boldsymbol{\theta})$, the posterior mode of $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ for a given $\boldsymbol{\theta}$. Taking all this together, we get

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \left. \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}. \tag{3.13}$$

As claimed in Rue et al. (2009), the approximation in expression (3.13) is equivalent to the Laplace approximation in Tierney and Kadane (1986) as we will show in the next subsection.

3.4.1 Equivalence of Laplace Approximations

We prove this by following analogously the idea proposed in a short comment from Leonard (1982).

The denominator of (3.13) has the following form

$$\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\mathcal{J}(\boldsymbol{\theta}, \mathbf{y})|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^T \mathcal{J}(\boldsymbol{\theta}, \mathbf{y}) (\mathbf{x} - \mathbf{x}^*) \right\} \tag{3.14}$$

where $\mathcal{J}(\boldsymbol{\theta}, \mathbf{y})$ is the observed Fisher information of $\log \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})$ evaluated at the posterior mode $\mathbf{x}^*(\boldsymbol{\theta})$. That is

$$\mathcal{J}(\boldsymbol{\theta}, \mathbf{y}) = \frac{\delta^2}{(\delta \mathbf{x})^2} \log \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}. \quad (3.15)$$

Expression (3.14) is now inserted back into (3.13). Note that the evaluation at the mode causes the exp-part in (3.14) to disappear. This gives

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto (2\pi)^{\frac{n}{2}} |\mathcal{J}(\boldsymbol{\theta}, \mathbf{y})|^{-\frac{1}{2}} \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})} \quad (3.16)$$

To show equivalence between this expression and the one in Tierney and Kadane (1986), we can compute a regular Laplace approximation of $\pi(\boldsymbol{\theta}|\mathbf{y})$ analogous to Tierney and Kadane (1986). Note that we will ignore the proportionality constant as we did above.

Rewrite $\pi(\boldsymbol{\theta}|\mathbf{y})$ as follows

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \int \pi(\boldsymbol{\theta}, \mathbf{x}, \mathbf{y}) d\mathbf{x} = \int \exp \{ \log \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) \} d\mathbf{x}, \quad (3.17)$$

which is the standard form for a Laplace approximation. Using the classical Laplace approximation 3.5 gives

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto (2\pi)^{\frac{n}{2}} |\mathcal{J}(\boldsymbol{\theta}, \mathbf{y})|^{-\frac{1}{2}} \pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y}) \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \quad (3.18)$$

which is equal to equation (3.16), showing the equivalence of the two methods.

Now that we have an approximation to $\pi(\boldsymbol{\theta}|\mathbf{y})$, section 3.6 will treat the remaining second approximation in 3.11, namely $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$, after evaluation points of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ are considered in section 3.5.

3.5 Finding Evaluation Points for $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$

For the numerical integration in 3.11 it is necessary to find several evaluation points of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$. This way a parametrical representation of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ is not necessary if evaluation points with according density values and weights are computed. As long as these points represent the distribution sufficiently well. Two approaches for finding evaluation points of $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ are presented in Rue et al. (2009), the GRID strategy and central composite design, the CCD strategy. The GRID strategy gives more accurate results but Rue et al. (2009) argue that CCD results are nearly as good for most practical applications with significantly reduced computing costs and it is therefore set as the default option in R-INLA. After evaluation points have been found, these can be used to obtain the posterior marginals $\pi(\theta_j|\mathbf{y})$.

The GRID strategy consists of computing the maximum and Hessian of $\log(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}))$, reparametrizing $\boldsymbol{\theta}$ with their help and then spanning an even grid around the maximum

to include values of sufficient probability mass. Since this grid is even, all evaluation points have the same weight Δ_k . These evaluation points can then also be used to create an interpolant for $\log(\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}))$, which in turn can be used for numerical integration to obtain approximations to the posterior marginals $\pi(\theta_j|\mathbf{y})$.

The CCD strategy uses a greatly reduced amount of evaluation points, allowing for higher m , the dimensionality of the hyperparameters. One point is placed on the origin and several more along each axis. Then, in a full factorial or fractional factorial design, evaluation points are placed on a hypersphere around the origin. The integration weights are equal for all points on the hypersphere and straightforward to calculate.

3.6 Approximations of $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$

Three different approximations of $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$ were presented in Rue et al. (2009) and implemented in R-INLA. The computationally fastest but least accurate is the Gaussian approximation. Much more accurate but also much more computationally expensive is the full Laplace approximation (FLA). The third alternative is the simplified Laplace approximation (SLA), introduced by Rue et al. (2009), which is supposed to be computationally less expensive than the FLA while still being much more accurate than the Gaussian approximation. The SLA is the default option in R-INLA.

3.6.1 Gaussian Approximation

The Gaussian approximation, see section 2.2, $\tilde{\pi}_G(x_i|\mathbf{y}, \boldsymbol{\theta})$ for $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$ is computationally efficient since the Gaussian approximation $\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ has already been computed when the space of $\pi(\boldsymbol{\theta}|\mathbf{y})$ has been searched for support points $\boldsymbol{\theta}_k$. Then, only the marginal variances need to be computed additionally to obtain $\tilde{\pi}_G(x_i|\mathbf{y}, \boldsymbol{\theta})$.

3.6.2 Full Laplace Approximation

We now consider the full Laplace approximation to $\pi(x_i|\mathbf{y}, \boldsymbol{\theta})$

$$\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})}, \quad (3.19)$$

where $\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ is a Gaussian approximation of $\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$ and the whole expression is evaluated at $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$, the mode of $\pi(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$.

Two adjustments are made to significantly speed up the computation. Firstly, the modal configuration is approximated by

$$\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta}) \approx E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i), \quad (3.20)$$

which is easy to compute since $\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ has already been computed. This is also advantageous since $E_{\tilde{\pi}_G}(\mathbf{x}_{-i}|x_i)$ is continuous in x_i , whereas a numerical optimization to find $\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})$ does not have this benefit. Secondly, it is assumed that only x_j "close" to x_i influence the marginal posterior $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$. Therefore, if suitably "far" x_j are neglected, only smaller matrices need to be factorized in the computation of $\tilde{\pi}_{GG}(\mathbf{x}_{-i}|x_i, \boldsymbol{\theta}, \mathbf{y})$.

3.6.3 Simplified Laplace Approximation

The simplified Laplace approximation $\tilde{\pi}_{\text{SLA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ is based on a series expansion of the full Laplace expansion $\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ up to third order. The inclusion of the third order term as well as the use of the skew normal distribution allow for more accurate approximations than the simple Gaussian approximation. For more details, see Rue et al. (2009).

3.6.4 Computing $\pi(x_i|\mathbf{y})$

Now that $\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})$ (Eq. 3.13) and $\tilde{\pi}(x_i|\mathbf{y}, \boldsymbol{\theta})$ (Eq. 3.19 or one of the other approximations) are available we can combine all these as given in 3.11 and compute an approximation to the marginal posterior $\pi(x_i|\mathbf{y})$.

3.7 Implementation in R-INLA

INLA has been implemented in C and has an R interface in the R-INLA package which is freely available as open-source at www.r-inla.org. It is based on the **GMRFLib** library accompanying Rue and Held (2005). Rue et al. (2009) provide some tests as to the accuracy of the method and later papers by e.g. Held et al. (2010), Schrödle et al. (2011) and Bivand et al. (2015) have found that the results obtained by INLA agree remarkably well with MCMC results.

R-INLA is intuitive to use as its syntax is very similar to that of the **glm** function in R. Firstly, the structured additive predictor needs to be specified as a **formula**, which is an extended version of the **glm formula**. As an example from the analysis in the following sections we have

```
f.st <- cases ~ 1 +                               #Intercept
  AEiche + ABuche + ATanne + ARest +             #Main effects
  hfrucBuche + hfrucEiche +                       #Main effects
  hfrucBuche:ABuche + hfrucEiche:AEiche +         #Interaction effects
  f(AGS.iid, model="iid") +                       #Unstructured error
  f(AGS.struc, model="besag", graph=graph)        #Structured error
```

The first lines look identical to the **formula** in a **glm**, whereas the last two lines are specific to R-INLA. The **f()** environment allows for the specification of a model for the latent Gaussian field. These specifications correspond with stage 2 of the model, as stated in section 3.1. In this case we have a spatial model, which we model in the way proposed by Besag et al. (1991), with a spatially structured and an unstructured error term. The call for the "besag" model requires an additional argument, **graph**, which defines the neighborhood structure.

This **formula** is then used in the call of the **inla** function.

```
m <- inla(f.st, family="poisson", E=Population, data=hanta,
  control.compute=list(dic=1))
```

Here, `family` is the distribution of our observational model, the likelihood. This is stage 1 of the model specification. `E` is the offset in a Poisson model and `control.compute=list(dic=1)` allows the specification of additional things to be computed, in this case the DIC as a later criterion for model selection. The output can then be accessed in the usual R fashion, for example by `summary(m)`.

Chapter 4

Analysis of the Hantavirus Data

We now analyze the hantavirus data described in section 1.1. Hantavirus is caused by transmission from bank voles to humans. Since we do not have any information on the number of bank voles, we use several proxy-variables, which we assume give us similar information on the number of bank voles but are possible to obtain. We consider the habitats of bank voles, e.g. forests and fields, as well as their proximity towards urban areas. We also consider the fructification of trees, as nuts and seeds are a food source to bank voles.

We fit a GLM and several GAMs to the data. The GLM will model the spatio-temporal data solely through the temporally and spatially varying covariates mentioned above, whereas the linear predictor of the GAMs will include a Gaussian Markov random field to account for the spatial structure that is clearly exhibited by the data. We fit the GLM as a reference against which to compare the later GAMs. We will use the R-INLA package as well as the `glm` function to fit these models.

4.1 Fitting a Generalized Linear Model

The data is first analyzed using as generalized linear model. To this end the `glm` function and then for comparison the `inla` function are used. As we are dealing with spatio-temporal count data, a Poisson distribution with spatio-temporally varying mean is assumed

$$y_{i,t} \sim \text{Po}(e_i \cdot \mu_{it}),$$

with $i = 1, \dots, 412$ and $t = 2002, \dots, 2012$, the counties and years respectively. The population of county i as an offset is given by e_i . We choose the canonical log-link

$$\log(\mu_{it}) = \eta_{it}$$

and can then specify the model through the linear predictor as

$$\begin{aligned}\eta_{it} = & \alpha + \beta_1 x_{i,AEiche} + \beta_2 x_{i,ABuche} + \beta_3 x_{i,ATanne} + \\ & \beta_4 x_{it,hfrucEiche} + \beta_5 x_{it,hfrucBuche} + \\ & \beta_6 x_{it,hfrucEiche} x_{i,AEiche} + \beta_7 x_{it,hfrucBuche} x_{i,ABuche} + \\ & \beta_8 x_{i,CAWaldstrauch} + \beta_9 x_{i,CAGrasland} + \beta_{10} x_{i,CAWiesenweiden} + \beta_{11} x_{i,u2fc}.\end{aligned}$$

The model contains an intercept, α , as well as main and interactions effects $\beta = \beta_1, \dots, \beta_{11}$. The main effects are the percentage of area covered by oak, beech and fir in a county, section 1.1.3, and the fructification index of that county for the previous year, section 1.1.4. This is due to our hypothesis that high fructification in one year is followed by a high case number in the next year. The variable $x_{it,hfrucBuche}$ denotes therefore the beech fructification in county i in year $t - 1$. The interaction effects of some of these variables are also considered. Lastly, the Corine land coverage variables for bushes, grassland and fields, section 1.1.2, and the distance between forests and urban areas, section 1.1.5 enter the model.

This predictor is specified as a `formula` in R by

```
f.st <- cases ~ 1 +                                     #Intercept
  AEiche + ABuche + ATanne + ARest +                   #time constant forest variables
  hfrucBuche + hfrucEiche +                             #time varying forest variables
  hfrucBuche:ABuche + hfrucEiche:AEiche +               #interaction effects
  CAWaldstrauch + CAGrasland + CAWiesenweiden +         #CORINE variables
  u2fc. #urbanization variables (derived from CORINE)
```

Note, that we will specify the offset separately using `E` in `inla` and using `offset` in the `glm` function. In what follows, we will present the predictors of other models in R-style as this is more compact.

The data can now be analyzed in a standard fashion by calling `glm`

```
m.st <- glm(f.st, offset=log(KreisPop), family=poisson,
  data=hanta.long.endemic)
```

or in a Bayesian framework using `inla`.

```
m.st.inla <- inla(f.st, E=KreisPop, family="poisson",
  data=hanta.long.endemic,
  control.compute=list(dic=TRUE, cpo=TRUE),
  control.inla=list(strategy="laplace", npoints=21,
    int.strategy = "grid", diff.logdens = 4))
```

We use the R-INLA default, which implies that vague Gaussian priors, i.e. $N(0, 1000^2)$, are assigned to intercept and covariates. Note that the `control.compute()` and `control.inla()` arguments are optional, and chosen here to allow for the accurate computation of model selection criteria in section 4.3.

4.2 Fitting Generalized Additive Models

We will now introduce several models that utilize Gaussian Markov random fields to account for spatial variation in the counties of Germany. The first model with regional effects, named *Pois*, is given by

$$y_{it} \sim \text{Po}(e_i \cdot \mu_{it})$$

$$\log(\mu_{it}) = \eta_{it} = \alpha + \mathbf{x}'_{it}\boldsymbol{\beta} + f(s_i) + u_i$$

where y_{it} is the number of hantavirus cases in county i during year $t = 2002, \dots, 2012$, e_i is the population as an offset and \mathbf{x}'_{it} is the vector of (time varying) covariates of i introduced in section 4.1. The discrete location variable with the set of all 412 counties as its domain is labeled s_i and $f(s_i)$ is the spatial effect of s_i . We will use a Gaussian Markov random field with precision parameter τ to account for the structural error, see section 2.2. A log-gamma prior distribution with parameters $(1, 5e-05)$ is assigned to $\log(\tau)$. The unstructured error is represented by the random effect u_i , where $u_i \sim N(0, \tau_u^{-2})$ are independent and identically distributed. The log precision τ_u^2 is assigned a log-gamma prior with parameters $(1, 5e-05)$. The `inla formula` for the linear predictor is then given by

```
f.st.inla
## cases ~ 1 + AEiche + ABuche + ATanne + ARest + hfrucBuche + hfrucEiche +
##      hfrucBuche:ABuche + hfrucEiche:AEiche + CAWaldstrauch + CAGrasland +
##      CAWiesenweiden + u2fc. + f(AGS.iid, model = "iid") + f(AGS.struc,
##      model = "besag", graph = graph)
```

In addition, a number of different models was also fitted, since *Pois* is a reasonable assumption about how the hantavirus cases can be modeled through the given covariates but not necessarily based on any model described in the literature. Since we hypothesize fructification to be a very important factor, we will consider several models with different combinations of the fructification main and interaction effects. We consider next a model that is similar to *Pois* but with the fructification main effects removed, which we will call *Pois-no-main*. Although it may look atypical to exclude the main effect of a variable while keeping its interaction term, the rationale behind this is that only areas that contain forests can be said to have a fructification. This model therefore only considers interaction effects between fructification and forest area. The predictor in `inla` is given by


```
f.st.inla2
```

```
## cases ~ 1 + AEiche + ABuche + ATanne + ARest + hfrucBuche:ABuche +  
##      hfrucEiche:AEiche + CAWaldstrauch + CAGrasland + CAWiesenweiden +  
##      u2fc. + f(AGS.iid, model = "iid") + f(AGS.struc, model = "besag",  
##      graph = graph)
```

Thirdly, we will consider model *Pois-no-interact*, which is similar to *Pois* but has no interaction effects between fructification and area covered by tree species. This might give a better fit since it reduces model complexity.

```
f.st.inla3
```

```
## cases ~ 1 + AEiche + ABuche + ATanne + ARest + hfrucBuche + hfrucEiche +  
##      CAWaldstrauch + CAGrasland + CAWiesenweiden + u2fc. + f(AGS.iid,  
##      model = "iid") + f(AGS.struc, model = "besag", graph = graph)
```

Model *Pois-interact* has interaction effects of the urban proximity variable *u2f*, exploring possible differences in how proximity is affected by fructification. Note that we use here the inverse of the proximity as we believe that smaller proximity and higher fructification both lead to increased cases numbers.

```
## cases ~ 1 + AEiche + ABuche + ATanne + ARest + hfrucBuche + hfrucEiche +  
##      hfrucBuche:ABuche + hfrucEiche:AEiche + CAWaldstrauch + CAGrasland +  
##      CAWiesenweiden + u2fc + I(1/u2fc):hfrucBuche + I(1/u2fc):hfrucEiche
```

Lastly, a negative binomial model, *NBinom*, was also considered. The predictor was equal to the one in model *Pois*, but a negative binomial distribution was used instead of the Poisson distribution.

4.3 Model Selection

In this section the fit of the different models to the data will be evaluated. The deviance information criterion and the probability integral transforms as introduced in sections 2.3.1 and 2.3.2 will mainly be used to evaluate model fit. The DIC and PIT histogram for each model can easily be obtained from *inla* by setting `control.compute=list(dic=TRUE, cpo=TRUE)` in the function call. To obtain reliable results for these measures of fit, it is necessary to choose the Full Laplace Approximation and the GRID strategy with additional evaluation points. Figure 4.1 shows the PIT histograms for all models and Table 4.1 shows the corresponding DICs.

As lower a DIC value is preferable, at first glance, the negative binomial model seems to be a very good fit. From the Poisson models, model *Pois-interact* seems to be the best compromise between model fit and complexity. The pure GLM without spatial effects

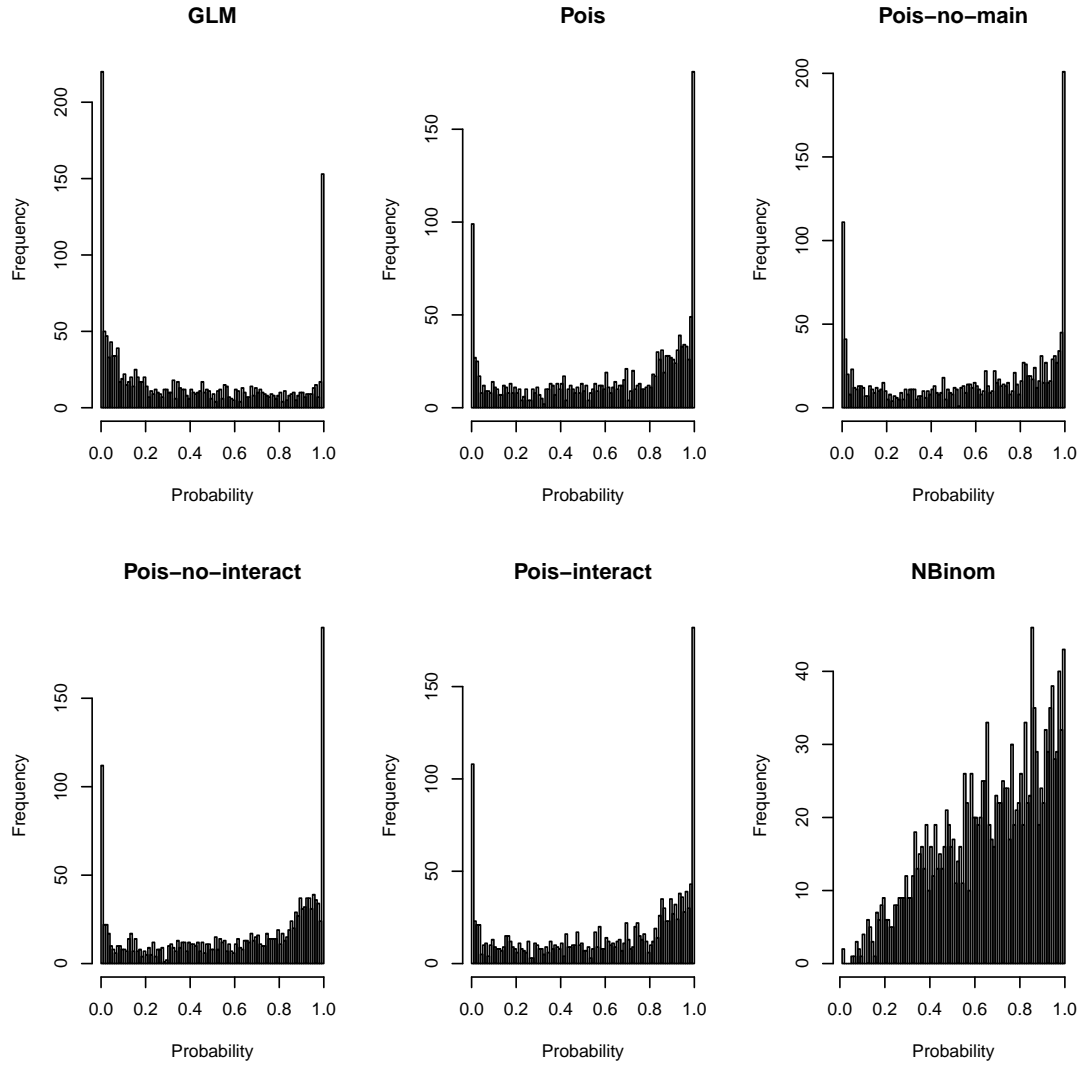


Figure 4.1: PIT histograms for all models. The U-shaped form indicates underdispersion. The diagonal shape of the negative binomial model indicates that central tendencies are biased.

GLM	<i>Pois</i>	<i>Pois-no-main</i>	<i>Pois-no-interact</i>	<i>Pois-interact</i>	<i>NBinom</i>
19269.40	9705.17	10364.58	9860.93	9643.58	5973.15

Table 4.1: DIC values of all models

seems to be the worst model, according to its DIC.

The PIT histograms reveal signs of underdispersion for all Poisson models, as indicated by the U-shape. The negative binomial model shows a skewed shape, which according to Czado et al. (2009) is a sign of central tendencies being biased.

Model *NBinom* has an excellent DIC but shows severe problems when we use PIT histograms as a diagnostic tool. The Poisson models on the other hand have a higher DIC, but show a less problematic behavior in the PIT histograms. We note that model *Pois-interact* has the best DIC value of the Poisson models followed by model *Pois*. The inclusion of many fructification interactions seems to give a better model fit.

In the following section we will consider the results obtained from the GLM and the model including spatial effects, *Pois*. We choose model *Pois* to allow for easier comparability between the GLM and the GAM but also because it has a very good DIC value and its linear predictor can be interpreted very well.

4.4 Results of the GLM

The results of the GLM fit with `glm` and `inla` are nearly identical when we consider estimates for the mean and for the 95% confidence and credibility intervals. This is not very surprising since we used weakly informative priors in the Bayesian case and a large 150×11 spatio-temporal data set for fitting. As such the posterior is dominated by the likelihood. The `inla` output for the fixed effects is presented in Table 4.2.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-10.937	0.084	-11.102	-10.937	-10.772	-10.937
AEiche	0.044	0.005	0.034	0.044	0.054	0.044
ABuche	0.055	0.004	0.047	0.055	0.062	0.055
ATanne	0.044	0.006	0.032	0.044	0.055	0.044
ARest	-0.017	0.002	-0.021	-0.017	-0.013	-0.017
hfrucBuche	0.026	0.001	0.025	0.026	0.028	0.026
hfrucEiche	0.007	0.002	0.003	0.007	0.011	0.007
CAWaldstrauch	0.131	0.011	0.109	0.131	0.154	0.132
CAGrasland	0.334	0.012	0.311	0.334	0.356	0.334
CAWiesenweiden	-0.009	0.002	-0.013	-0.009	-0.005	-0.009
u2fc	-0.082	0.004	-0.091	-0.082	-0.074	-0.082
ABuche:hfrucBuche	-0.000	0.000	-0.000	-0.000	-0.000	-0.000
AEiche:hfrucEiche	-0.001	0.000	-0.002	-0.001	-0.001	-0.001

Table 4.2: Summary statistics of the posterior distributions of the GLM fixed effects.

We can interpret these estimates in the standard fashion for Poisson regression. Take beech fructification as an example. Here we are dealing with an interaction effect of two continuous variables. Note that Table 4.2 only gives three digits. The full effect of `ABuche:hfrucBuche` is -3.1×10^{-4} . Since we are mainly interested in the effect of fructification, we will exemplarily interpret the unit increase of fructification with three fixed values for the forest area, its mean (6.831) and mean \pm standard deviation (1.602 and 12.06). Taking the mean forest area, a unit increase of beech fructification in the year $t-1$ in a county i increases the expected number of reported cases per population of i , i.e. the incidence in county i , by a factor of $\exp(0.026 + (-3.1 \times 10^{-4}) \cdot 6.831) = 1.025$ for the following year t . Inserting mean-standard deviation and mean+standard deviation for the forest area, we obtain factors of 1.026 and 1.023.

We get a 95% credibility interval for the increase of the expected incidence for a unit increase of beech fructification very easily by centering the value for `ABuche` and re-running the model. If we center `ABuche` at its mean and insert the mean forest value, then we can neglect the term for `ABuche:hfrucBuche` as it is zero. The credibility interval for `hfrucBuche` is then the credibility interval for the fixed effect plus the interaction effect for a mean forest value. This gives us the credibility interval (1.024, 1.025). The credibility intervals of the beech fructification for forest area values of mean-standard deviation and mean+standard deviation can be computed similarly and are given by (1.025, 1.027) and (1.022, 1.024) respectively.

Clearly, the fructification of trees, especially beeches, has a strong effect even after having adjusted for a multitude of other time-constant and time-varying factors. We also notice that none of the 95% credibility intervals for the covariates cover 0.

4.5 Results of the GAM

We first consider the fixed effects of the spatio-temporal model, *Pois*. We have the following summary statistics for their posterior distributions in Table 4.3.

In contrast to the GLM, it can be seen that the influence of the fructification of oaks and beeches is similar if we account for spatial structure. That the effects of fructification change is not surprising given their different spatial distributions in Figures 1.8 and 1.9. We can interpret the effect of a unit increase in beech fructification as we did for the GLM. Assuming the mean of the forest area, the expected incidence of the county increases by a factor of 1.033 (95% credibility interval (1.032, 1.034)) for a unit increase in beech fructification. Inserting mean-standard deviation and mean+standard deviation for the forest area gives us factors of 1.027 and 1.038 as well as credibility intervals (1.025, 1.028) and (1.037, 1.04) respectively.

Estimates for the precisions of the unstructured and the spatially structured effects are available in Table 4.4. As the unstructured effect is distributed as $u_i \sim N(0, \sigma^2)$, a very high precision, and therefore a very small variance, means that the unstructured effect is not important. On the other hand, the spatially structured effect, signifying the different county effects, is very important. These effects are shown separately in Figure 4.2 and added together in Figure 4.3. The last representation is preferable, since it is question-

	mean	sd	0.025quant	0.5quant	0.975quant	mode
(Intercept)	-13.067	0.534	-14.116	-13.066	-12.019	-13.066
AEiche	0.005	0.040	-0.073	0.005	0.083	0.005
ABuche	-0.014	0.026	-0.065	-0.014	0.037	-0.014
ATanne	-0.001	0.058	-0.115	-0.001	0.112	-0.001
ARest	0.003	0.013	-0.023	0.003	0.028	0.003
hfrucBuche	0.024	0.001	0.023	0.024	0.026	0.024
hfrucEiche	0.021	0.002	0.016	0.021	0.026	0.021
CAWaldstrauch	-0.078	0.117	-0.308	-0.077	0.152	-0.077
CAGrasland	0.215	0.121	-0.022	0.215	0.452	0.216
CAWiesenweiden	-0.002	0.014	-0.030	-0.002	0.026	-0.002
u2fc	0.007	0.024	-0.041	0.007	0.054	0.007
ABuche:hfrucBuche	0.001	0.000	0.001	0.001	0.001	0.001
AEiche:hfrucEiche	0.004	0.001	0.002	0.004	0.005	0.004

Table 4.3: Summary statistics of the posterior distributions of the *Pois* model fixed effects.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Precision for AGS.iid	19107.9	18696.8	1260.1	13611.7	68371.7	3397.5
Precision for AGS.struc	0.2	0.0	0.1	0.2	0.3	0.2

Table 4.4: Summary statistics of the posterior distributions of the *Pois* model random effects, i.e. the spatially structured and unstructured effects.

able whether the two effects can really be cleanly separated, see chapter 5 in Fahrmeir and Kneib (2011).

Figure 4.2 still offers several insights. Firstly, we see that the unstructured error is of the same size and always negligible for all endemic regions. This is in line with our observation that the unstructured effect has a very high precision. It is also in line with our observation through the PIT histogram that the model is underdispersed, i.e. we have less variation in the data than the model suggests and therefore the extra variation in the model possible through the unstructured effect is nearly non-existent.

Secondly, the left panel of Figure 4.2, which shows the spatially structured effects, illustrates the effect of the Gaussian Markov random field very well. Its effect is not contained to the endemic regions but does of course affect the regions around them as well, as we assume a spatially smooth distribution is underlying our observations 1.2. In Figure 4.3 we do, however, only consider the endemic regions. Lastly we notice again that we might have some border effects as we only consider cases in Germany but several regions with strong effects according to the GMRF lie on the border of Germany.

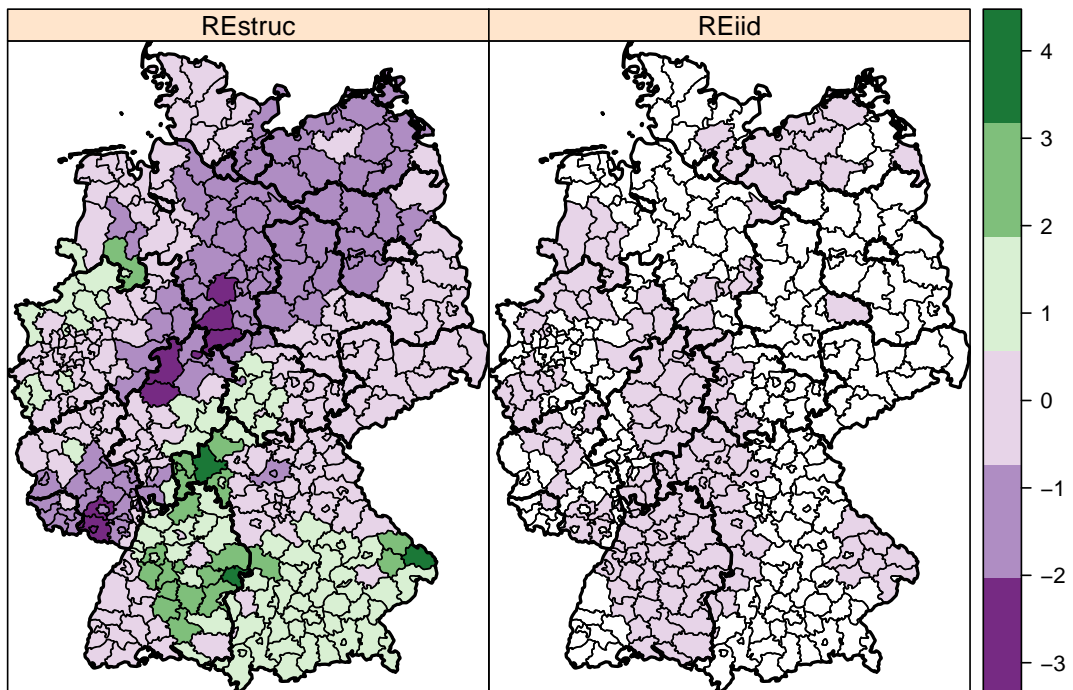


Figure 4.2: Spatially structured and unstructured error separately.

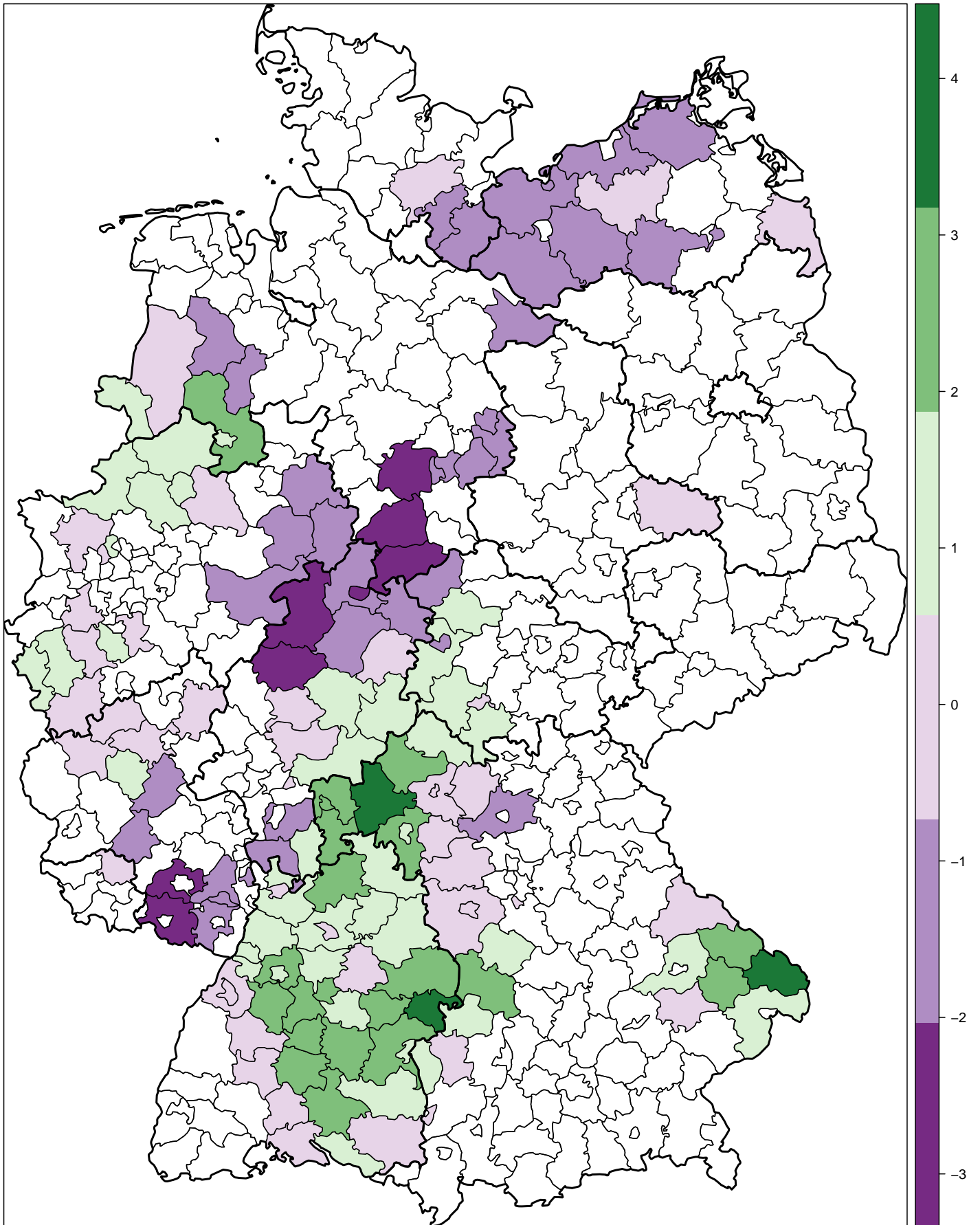


Figure 4.3: Spatially structured and unstructured error of model *Pois* added together.

Chapter 5

Conclusion

For this thesis a spatio-temporal analysis was conducted for an extensive dataset containing yearly hantavirus cases in Germany between 2002 and 2011 as well as a large number of environmental variables. After an initial exploratory analysis of the data, we performed our analysis using the state-of-the-art INLA inference method, see Rue et al. (2009), instead of the very common MCMC methods to avoid long runtimes and difficult implementation. MCMC in spatial applications requires careful implementation to avoid extremely long runtimes and convergence issues due to strong dependencies in the data. The R package R-INLA provides a universal interface for many different kinds of models which is fairly easy to use. Runtimes of around ten minutes for each model, even though the computationally most expensive approximations were chosen, are excellent and corroborate the claims of computational efficiency in Rue et al. (2009). Even the fully saturated model could be fitted in under two hours using simplified Laplace approximations (model omitted). We used the R-INLA package to model the reported number of human hantavirus cases in Germany in the years 2002 to 2012 in a spatio-temporal framework using generalized linear models and generalized additive models. The GAM contains a Gaussian Markov random field to account for the spatial structure in the data. This modeling of the spatial dependencies was found to be very important and lead to a much better fitting model.

As part of this thesis the theoretically intricate INLA method and the necessary advanced background material were explained. Important detailed steps in the proofs leading to an increased didactical presentation of the material were given. The INLA method, as the name suggests, relies on several (Laplace) approximations and numerical integrations nested within each other. This method of direct approximations proved to be very effective in practical use. Spatio-temporal models such as the ones we considered become quite complex very quickly so that interpretation and residual checking become extremely important. This issue was addressed by the use of probability integral transforms to check model fit. Alternatively, one-step-ahead predictions could have been used. An alternative to the use of the R-INLA package could have been the BayesX package, see Belitz et al. (2009), which could have fit our models using MCMC.

Even though no model produced a perfect fit by our model validation criteria, important

conclusions can still be drawn from the models. A variety of different data sources was combined using geographic information systems like preprocessing in order to investigate the relationship between the environmental variables and the reported case number. We found strong evidence for our research hypothesis that the fructification of especially beech but also oak trees in one year has a strong influence on the number of reported cases in the following year. This hypothesis was motivated by the literature, see for example Boone et al. (2012), and our explorative data analysis. A high fructification of beeches and oaks leads to an abundance of food for bank voles and therefore to a greater number of offspring as well as a higher chance of survival for the next winter. Bank voles are the main carrier of hantavirus and their excretions are the main source of infection for humans.

Furthermore, we produced a map, Figure 4.3, that gives a clear picture of the regions with bank vole populations that carry hantavirus and therefore pose a risk to humans. The expected increase of the incidence by a factor of 1.033 (95% credibility interval (1.032,1.034)) is an important finding that can for example be used in forecasting case numbers. These forecasts are done per county so that warnings after years of strong fructification can be issued to specific regions, enabling these regions to prepare and take measures for a coming year with high numbers of expected infections. These warnings can be given to health professionals but can also be issued to the general population and especially to people working on the countryside or in close proximity to forests.

Limitations of the analysis are that we only use German data, which can lead to errors close to the borders, since our models with spatial effects take a county's neighbors into account. Counties that are close to a border are therefore modeled as if there were for example no forests on the other side of the border. Further research could include an extended way of computing the fructification spline, e.g. including the forest area directly as we now assume an even spread of forest across Germany when calculating the fructification spline. A more complex and possibly better model could include the modeling of the fructification as a random walk to include the effects from several years, see for example Schrödle et al. (2011). The interaction between the two continuous variables forest area and fructification could be modeled using a smooth two-dimensional function.

Outlook

As more and more datasets with spatial information are being made publicly available by governments, health institutes and other sources, the need to be able to analyze these data increases as well. This requires skills in geographic information systems, computer science and statistics. This thesis presents such an analysis, but its results need to be interpreted carefully since it is an ecological regression. Results that were found on the county level can therefore not easily be generalized to individuals in these counties. It nonetheless provides important results on which further research can be based.

Bibliography

- Christiane Belitz, Andreas Brezger, Thomas Kneib, Stefan Lang, and N Umlauf. BayesX-Software for Bayesian inference in structured additive regression models. <http://www.bayesx.org>, 2009.
- Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.
- Roger Bivand, Virgilio Gómez-Rubio, and Håvard Rue. Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software, Volume 63, Issue 20*, 2015.
- I Boone, C Wagner-Wiening, D Reil, J Jacob, UM Rosenfeld, RG Ulrich, D Lohr, and G Pfaff. Rise in the number of notified human hantavirus infections since October 2011 in Baden-Württemberg, Germany. *Eurosurveillance*, 17(21):1–5, 2012.
- Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2011.
- Claudia Czado, Tilmann Gneiting, and Leonhard Held. Predictive model assessment for count data. *Biometrics*, 65(4):1254–1261, 2009.
- Mirko Faber and Michael Höhle. Ecological regression of hantavirus incidence in Germany 2001-2010. Technical report, Robert Koch-Institut, 2013.
- Mirko Faber, Rainer Ulrich, Christina Frank, Stefan Brockmann, Günter Pfaff, Jens Jacob, Detlev H Krüger, and Klaus Stark. Steep rise in notified hantavirus infections in Germany, April 2010. *Eurosurveillance*, 15(20), 2010.
- Ludwig Fahrmeir and Thomas Kneib. Bayesian smoothing and regression for longitudinal, spatial and event history data. *OUP Catalogue*, 2011.
- Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx. *Regression: Models, methods and applications*. Springer Science & Business Media, 2013.
- Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

- Leonhard Held, Birgit Schrödle, and Håvard Rue. Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. In *Statistical modelling and regression structures*, pages 91–110. Springer, 2010.
- Pierre Simon Laplace. Memoir on the probability of the causes of events. *Statistical Science*, pages 364–378, 1986.
- Tom Leonard. Comment on "A Simple Predictive Density Function," by M. Lejeune and G.D. Faulkenberry. *Journal of the American Statistical Association*, 77(379):657–658, 1982.
- Sara Martino and Håvard Rue. Case studies in Bayesian computation using INLA. In *Complex data modeling and computationally intensive statistical methods*, pages 99–114. Springer, 2010.
- Thiago G Martins, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.
- Håvard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.
- Håvard Rue and Sara Martino. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, 137(10): 3177–3192, 2007.
- Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 71(2):319–392, 2009.
- Birgit Schrödle and Leonhard Held. A primer on disease mapping and ecological regression using INLA. *Computational statistics*, 26(2):241–258, 2011a.
- Birgit Schrödle and Leonhard Held. Spatio-temporal disease mapping using INLA. *Environmetrics*, 22(6):725–734, 2011b.
- Birgit Schrödle, Leonhard Held, Andrea Riebler, and Jürg Danuser. Using integrated nested Laplace approximations for the evaluation of veterinary surveillance data from Switzerland: a case-study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(2):261–279, 2011.
- David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.
- Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

Antti Vaheri, Heikki Henttonen, Liina Voutilainen, Jukka Mustonen, Tarja Sironen, and Olli Vapalahti. Hantavirus infections in Europe and their impact on public health. *Reviews in Medical Virology*, 23(1):35–49, 2013.

Simon Wood. *Generalized additive models: an introduction with R*. CRC press, 2006.