Spatial Analysis of Exposure Coefficients with Applications to Stomach Cancer

Maria Martinho

St Hugh's College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Hilary Term 2007

Abstract

Spatial Analysis of Exposure Coefficients with Applications to Stomach Cancer

Maria Martinho
St Hugh's College, Oxford

D.Phil. Thesis
Hilary Term, 2007

Earlier ecological studies on the relation between *H. pylori* infection and stomach cancer have considered that the relation between these two variables, as estimated by the exposure coefficient, is constant. However, there is evidence to suggest that this relation changes geographically due to differences in strains of *H. pylori*. Since the prevalence of *H. pylori* varies with socio-economic status, the association between the latter and stomach cancer mortality may also vary geographically. This thesis studies stomach cancer by taking into account the geographical variability of the exposure coefficients.

The study proposes the use of regression mixtures, clustering models and spatially varying regressions for the study of varying exposure coefficients. The effect of transformations of variables in these models appears to have been little considered. We provide new necessary conditions for invariance under transformations of variables for mixed effect models in general, and for the proposed models in particular. In addition, we show that varying exposure coefficients may induce a varying baseline risk.

The regression mixtures and the clustering model are applied to a data set on stomach cancer incidence and *H. pylori* prevalence in 57 countries worldwide. We extend the clustering model to reflect any distance measure between the

geographical units, including the Euclidean distance, in the formation of clusters. We also show that the clustering model performs better than the regression mixture model when the aim is to identify connected clusters and the observations present large variance. The results obtained with the clustering model supported the existence of three clusters where the interaction between the human and *H. pylori* populations have similar characteristics.

Spatially varying regressions are applied to a data set of areal death counts of stomach cancer and spending power in 275 counties in continental Portugal. We provide an original strategy for implementing multivectorial intrinsic autoregressions as the distribution for the random effects. The results obtained with the application of this methodology were consistent with a varying exposure coefficient of spending power.

To my mother

Acknowledgements

I would like to express my thanks and appreciation to my supervisor, Dr John Bithell, for his support and guidance. Thanks are also due Dr Peter Clifford for providing helpful advice and supervision at the early stages of this thesis and to Dr Geoff Nicholls for his valuable suggestions. I have received kind support from the staff of the Department of Statistics and I would like to particularly thank David del Campo Hill and Susan Hutchinson for all their computer and programming tips.

The Portuguese data studied in this thesis were kindly provided by the Portuguese General Board of Health (DGS), the Portuguese Institute of National Statistics and the Portuguese Geographic Institute. In particular, I am indebted to Mário Carreira of the DGS for his invaluable help. The financial support of the Portuguese Foundation for Science and Technology is also gratefully acknowledged.

On a more personal level, I have been very fortunate to make some great friends in Oxford. The list of thanks would not be complete without mentioning Linda, Lorenzo, Maud, Natalia and, last but not least, Brais. Thanks also to my office mates James, Kristin and Lucas for the good companionship and stimulating conversations during the time we shared a common office and to the friends from abroad Antonio, Celia and Sofia for their continuous friendship. Finally, I would like to thank my mother for her unfailing support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The topic treated in this thesis was inspired by the study of two data sets on stomach cancer. The first data set consists of stomach cancer incidence rates and prevalence rates of the *H. pylori* infection in 57 countries worldwide. The second data set comprises death counts attributed to stomach cancer and spending power indexes for 275 counties in continental Portugal.

The methodology developed is meant to respond to the questions raised from the data analysis. In particular, we aimed at investigating the association between stomach cancer and the *H. pylori* infection. If the characteristics of the infectious agent and the susceptibility of the sub-population are unknown but believed to vary spatially, the relation between the exposure (as estimated by the prevalence of infection in the population) and the disease outcome will also vary spatially. In addition, exposure coefficients of variables associated with the prevalence of the *H. pylori* infection, like socio-economic status, may also display spatial variation. Therefore, the emphasis of this thesis is on (spatially) varying coefficient models, and their application to the analysis of varying exposure coefficients.

Since the *H. pylori* bacteria spread at a slow pace, the geographical distribu-

tion of the exposure can be assumed constant for the periods of time covered in the two data sets. Thus, we do not consider methods modelling the spread of the infection.

We propose to use techniques such as mixture regressions, clustering models and spatially varying regressions for investigating geographical variations in exposure coefficients. In such a way we are able to assess the importance of the exposure in explaining the disease, while accounting for spatial variations of the exposure-disease relation. It is hoped that the models produced from the two data sets studied in this thesis will be relevant to other areal health data.

This chapter introduces some key concepts in geographical epidemiology and provides an overview of the material included in this thesis.

1.1 Definitions

1.1.1 Epidemiology

Epidemiology can be defined as the study of the mortality, prevalence and incidence of disease and related factors to provide clues about the aetiology of disease. Since, for fatal diseases, routine mortality data are usually the only source of data, epidemiological studies are often based on the study of mortality. Alderson (1976) pointed out three main aims of epidemiological studies: to describe the distribution of disease incidence; to identify aetiological factors; to provide data to help in prevention, control and treatment of disease.

There are two distinct types of epidemiological studies: observational studies and intervention studies. The distinction relies on whether the factors involved are being controlled (*intervention studies*) or not (*observational studies*). Only

observational studies on disease are considered here with the aim of describing the distribution of disease and investigating hypotheses concerning disease aetiology.

1.1.2 Contagious and non-contagious diseases

The distinction between contagious and non-contagious diseases is an important one in a probabilistic sense. In non-contagious diseases, cases are generally considered to be independent, in other words, the occurrence of the disease in one individual does not cause the disease in another. In contagious diseases, this independence assumption does not hold.

Recently, however, the distinction between contagious and non-contagious diseases became blurred as viral and bacterial infections have been linked to some forms of cancer (Mueller, 1995; Parsonnet, 1995; Correa, 2003). Some infections spread very fast, causing continuous changes in the proportion and geographical distribution of the infected population. Others spread so slowly that the infected population can be assumed constant for certain periods of time.

This thesis will consider non-infectious disease, in its traditional sense, but considering the influence of infectious agents. From the point of view of the risk in a given period, most non-infectious diseases are rare.

1.1.3 Geographical epidemiological studies

Geographical epidemiology is the term generally used to define studies of the distribution of disease *in geographical space*.

The distribution of a disease typically varies geographically. Since, for a *purely* non-contagious disease, the cases are independent this must be a consequence of heterogeneity of environmental, ecological and genetically determined

risk factors among individuals. Disease risk is, in general, age and sex dependent. All epidemiological studies involve some kind of standardization to take this into account. However, the disease risk can still show some variation after this standardization. If this variability can not be explained by chance alone, then it must be the consequence of different exposures to risk and/or protective effects and/or genetic differences between the exposed sub-populations (*risk variation*). This is a completely different mechanism from the clustering observed in infectious diseases (*genuine clustering*), which is caused by contagion. In non-contagious diseases, any observed clustering can theoretically be "removed", if all genetic, ecological and environmental risk and protective factors are used to explain disease variability. In practice, the complexity of the biological mechanisms of chronic diseases makes the full knowledge of all relevant factors impossible. Consequently, there usually exists some unexplained variability.

In diseases whose aetiology involves infectious and non-infectious factors, clustering may arise in both ways: genuine clustering and risk variation.

1.1.4 Environmental, ecological and infectious factors

One advantage of geographical studies is that the geographic information can be useful to explain risk variability. In non-infectious diseases, the explanatory potential of geographic information can come from two distinct types of risk factors: *environmental* and *ecological* factors. Environmental factors are intrinsically geographical attributes (e.g. pollution source, radiation source) while ecological factors are surrogates for individual attributes (e.g. socio-economic levels and broadly based health indicators). For the latter case, the geography can be used to provide a surrogate when the individual attributes show a spatial

pattern. When infectious agents are considered, geographic information can be used for investigating the pattern of infection.

1.1.5 Areal and point data

Turning now to the data used in geographical epidemiology, it is important to distinguish between point and areal data. In point data, the precise location of each individual case is available. In areal data, the cases have been aggregated into areas, e.g. administrative regions, so that only the areal counts of disease cases are available. On one hand, areal data are easier to obtain owing to confidentiality issues. On the other hand, areal data pose statistical challenges of their own because a part of the geographical information is lost. Since areal studies implicitly assume that the disease risk is homogeneous within areas, low or high spatial concentration of cases within an area can be missed. The analysis of areal data depends on the areal partition chosen. Different areal partitions may result in different patterns emerging.

This thesis will only treat the analysis of areal data.

1.1.6 Cancer statistics

Age specific death rates, age standardized rates and standard mortality ratios are widely used to report cancer mortality rates.

The age specific death rate (AR) is calculated simply by dividing the number of cancer deaths observed during a given time period by the corresponding number of persons in the respective age-group at risk. The result is usually expressed as an annual rate per 100000 persons at risk. The age groups tend to be in five or ten year ranges, except for very young or old

groups where bigger ranges are used because of the small number of cases or small populations sizes.

The age standardized rate (ASR) is a weighted average of age-specific death rates of the population of study. The weight for each age category is the proportion of people in the age category in a standard population. The standard population is often chosen to be the world population. The weights are used to adjust for the difference in age distribution in comparing the death rates of two or more populations. The ASR is also expressed per 100000 persons per year. The distribution of the world population is given in Prokhorkas (1998) and it is reproduced in appendix B.

The standardized mortality ratio (SMR) is the ratio of the observed number of cases to the expected number of cases. The expected number of cases is the number of cases that would be expected if age and sex specific rates from a standard population were applied to the population of study.

The definitions above can also be used for reporting cancer incidence.

The ASRs and the SMRs are subject to random variability because their calculation relies on the observed number of cases, which is a random variable. Therefore, differences of these standardized measures can be observed among different geographical areas as a result of random variation.

1.2 Outline of the thesis

Chapter 2 introduces basic modelling of areal counts and studies the effect of within area variability of the risk of disease on the variance:mean ratio of the

areal counts. It also describes some aspects related to the study of exposure coefficients.

Chapter 3 reviews the current knowledge of stomach cancer aetiology, especially regarding the influence of the *H. pylori* infection on stomach cancer. The chapter discusses the impact of the virulence of the *H. pylori* infection on geographical studies of stomach cancer.

Chapter 4 proposes the use of varying coefficient models to account for the virulence of infectious risk factors. These models are a special case of mixed effect models. Some aspects of the modelling are discussed. In particular, the chapter shows that varying exposure coefficients may induce varying baseline risks. The chapter presents specific models, which will be discussed in detail in chapters 6, 7 and 8, and justifies the choice of a Bayesian framework.

Chapter 5 attempts to explain why the results obtained with certain mixed effect models change with the scale and origin of the covariates. This has been named in the literature as coding variance. The chapter reviews a known necessary condition to avoid coding variance and advances additional necessary conditions.

Aiming at the study presented in chapter 9, chapters 6 and 7 present two alternative models for identifying groups of areas with similar exposure coefficients. Chapter 6 proposes the use of regression mixture models for the study of varying exposure coefficients, and discusses some aspects of the implementation of these models. In particular, it provides an explanation for the trapping states found in the Markov chain Monte Carlo (MCMC) simulation. Chapter 7 adapts the model developed in Knorr-Held and Rasser (2000) for identifying clusters of similar exposure coefficient while taking into account the spatial structure of the

areas. The resulting model is a special case of regression mixtures but includes information on the adjacency and geographical distance between the areas of the study. The performance of this model vis-à-vis the regression mixtures is also evaluated.

Chapter 8 reviews the use of spatial intrinsic autoregressions in varying coefficient models. In order to account for the dependence between the intercept term and the regression slopes, the chapter proposes the use of the multivectorial intrinsic autoregressions which were presented in Gamerman et al. (2003). We show that these multivectorial intrinsic autoregressions can be obtained as linear combinations of the usual intrinsic autoregressions. This finding allows us to develop an original strategy to facilitate the MCMC implementation of these models. The chapter also assesses the performance of this model within the context of the application of chapter 10.

Chapter 9 contains an analysis of a world data set on stomach cancer incidence and prevalence of the *H. pylori* infection. It uses the regression mixtures and the clustering model, presented respectively in chapters 6 and 7, for assessing the relation between the two variables. The clustering model is applied to the data for estimating the *H. pylori* virulence in different countries as well as for calculating worldwide attributable risks of the *H. pylori* infection on stomach cancer.

Chapter 10 consists of an analysis of data on stomach cancer mortality in Portugal in 1985-1998. The chapter studies the association between stomach cancer mortality and spending power, used here as a surrogate for *H. pylori* prevalence. The model discussed in chapter 8 is used for investigating the geographical variability of that association. The results are compared with those obtained with

alternative models. We attempt to estimate the number of cases attributable to lower socio-economic status, on the basis of the selected model.

The thesis finishes with a general conclusion discussing some methodological aspects and the epidemiological findings of the two studies.

Chapter 2

Modeling areal counts

The first attempts to analyse areal heath data used transformations of standardized mortality ratios to approximate normality (Cressie, 1993). This approach, developed primarily for geostatistical data, failed to account for the intrinsic relation between the mean and variance of the Poisson variables (i.e. the counts). For this reason, current literature models the counts directly. Strictly speaking, the counts are sums of Bernoulli random variables, corresponding to the individual outcomes. The individual probabilities associated with these variables are not necessarily identical because the disease risk may not be the same for all individuals. Thus, conditional on the individual probabilities, the counts are usually not binomially distributed, and, theoretically, are less dispersed than the binomial distribution. This chapter explains why areal counts can nevertheless be approximated by a Poisson distribution, an approximation which is convenient because the Poisson distribution is easier to handle.

The chapter also studies the consequences of within area variability in the distribution of the areal counts. In particular, it shows that large within area variability may compromise the approximation to the Poisson distribution, but it

concludes that this does not happen in studies where the probabilities of disease are very small. The chapter further introduces the basic modelling for disease areal counts, and discusses several aspects related to the study of areal exposure coefficients.

2.1 The Poisson model

When, independently, all the individuals in an area have the same probability, p, of disease, then the areal counts of disease events are binomially distributed:

$$Y \sim \text{Binomial}(n, p),$$

where n is the number of individuals in the area.

It is well known that when $n \to \infty$ and $p \to 0$ so that np remains constant, the distribution of Y converges to a Poisson(np). Because of this, when n is large and p is small, the binomial distribution is approximately the same as the Poisson distribution. For rare diseases, the value of p is small, justifying the use of the Poisson distribution:

$$Y \sim \text{Poisson}(np)\,.$$

In particular, the variance:mean ratio ψ of the areal counts is approximately one.

When the disease risk is not equal for all individuals, the areal count is a sum of n independent Bernoulli variables:

$$\begin{aligned} Y &= \sum_{k=1}^{n} Y_k, \\ Y_k|p_k &\sim \text{Bernoulli}(p_k), \end{aligned}$$

where p_k is the probability of disease for the individual k.

It follows that, conditional on the vector of probabilities $\boldsymbol{p}_\bullet = (p_1, \ldots, p_n)$, the areal count satisfies:

$$\begin{aligned} \mathrm{E}[Y|\boldsymbol{p}_\bullet] &= \sum_{k=1}^{n} p_k = n\bar{p}, \\ \mathrm{Var}[Y|\boldsymbol{p}_\bullet] &= \sum_{k=1}^{n} p_k(1-p_k), \\ &= n\bar{p}(1-\bar{p}) - \sum_{k=1}^{n}(p_k-\bar{p})^2, \end{aligned} \tag{2.1}$$

i.e. the variance is smaller than what would be expected from a binomial distribution. The variance:mean ratio ψ of the areal counts is then

$$\psi = \frac{\mathrm{Var}[Y|\boldsymbol{p}_\bullet]}{\mathrm{E}[Y|\boldsymbol{p}_\bullet]} = (1-\bar{p}) - \frac{s^2}{\bar{p}}, \tag{2.2}$$

where

$$s^2 = \frac{\sum_{k=1}^{n}(p_k-\bar{p})^2}{n}.$$

The expression 2.2 above implies that the larger the quantity $\frac{s^2}{\bar{p}}$ the smaller the value of the variance:mean ratio ψ. Hence, it is natural to ask whether the Poisson approximation is adequate when considerable within area variability, leading to a large value of $\frac{s^2}{\bar{p}}$, is present. We will study this issue in the next section.

2.2 Study of underdispersion

Since the p_k are probabilities, they take values on the interval $[0, 1]$ and therefore s^2 must have a (finite) supremum for a given value of $\bar{p}$. We now investigate the supremum (i.e. the least upper bound) of the quantity $s^* = s^2/\bar{p}$ (with

$\bar{p} \neq 0$). This establishes how small the ratio ψ can be (see equation 2.2), and thus how underdispersed the model can be relative to the Poisson distribution. We consider for this analysis the supremum of s^* in the finite population case.

2.2.1 Supremum of s^* in the finite population case

In a population of n individuals, in which individual k has a probability p_k of being a case, what is the supremum of the population variance of these probabilities,

$$s^2 = \frac{1}{n}\sum_{k=1}^{n}(p_k - \bar{p})^2, \tag{2.3}$$

for a fixed value of the areal mean $\bar{p} = \frac{1}{n}\sum_{k=1}^{n} p_k$?

Formally, this problem consists in finding the supremum of the n-dimensional function

$$s^2(p_1, \ldots, p_n) = \frac{1}{n}\sum_{k=1}^{n}(p_k - \bar{p})^2, \tag{2.4}$$

subject to the constraint:

$$\bar{p} = \frac{1}{n}\sum_{k=1}^{n} p_k, \tag{2.5}$$

where $\bar{p}$ is fixed and

$$p_k \in [0, 1], \text{ for all } k = 1, \ldots, n. \tag{2.6}$$

Theorem 2.2.1 *If* $\bar{p} \in [\frac{m}{n}, \frac{m+1}{n})$, *with* $m \in \{0, \ldots, n-1\}$, *the point* $(p_1, \ldots, p_n)$ *defined by:*

- $p_1 = n\bar{p} - m$,
- $p_k = 1$, $k = 2, \ldots, m+1$,

- $p_k = 0,\ k = m+2, \ldots, n,$

maximizes s^2 under the constraints (2.5) and (2.6). The order of the p_k is obviously irrelevant. Therefore any permutation of the p_k above will maximize s^2 under the same constraints. At the maximum, the expression for s^2 is

$$s^2 \quad = \quad \frac{(n\bar{p} - m)^2 - n\bar{p}^2 + m}{n}. \tag{2.7}$$

The theorem can be restated in terms of a simple quadratic programmme. There is a substantial literature on quadratic programming: see for example, Barankin and Dorfman (1958). A self-contained proof of the theorem is given in appendix A.

Since

$$m/n \leq \bar{p} < (m+1)/n \Rightarrow n\bar{p} - 1 < m \leq n\bar{p}\,,$$

the parameter m in (2.7) can be replaced by $n\bar{p} - \theta$, with $\theta \in [0, 1)$, leading to

$$s^2 = \frac{\theta(\theta - 1)}{n} + \bar{p}(1 - \bar{p}).$$

Using this quantity in expressions 2.1 and 2.2, we obtain, respectively,

$$\mathrm{Var}(Y|\boldsymbol{p}_\bullet) = \theta(1 - \theta)$$

and

$$\psi = \frac{\theta(1 - \theta)}{n\bar{p}}.$$

Consequently, as θ varies from zero to one, the value of the maximum variance varies from 0 to $1/4$, and the variance:mean ratio ψ varies from zero to $1/(4n\bar{p})$.

2.2.2 Consequences for areal studies of disease

Equation 2.2 appears to show that when there is strong within area variability, ψ may not be close to one. By way of illustration, table 2.1 shows the values

	ψ				
$\bar{p}$	$s^2 = 0$	$s^2 = 0.1\bar{p}$	$s^2 = 0.25\bar{p}$	$s^2 = 0.5\bar{p}$	$s^2 = 0.75\bar{p}$
0.5	0.5	0.4	0.25	0	-
0.1	0.9	0.8	0.65	0.4	0.15
0.01	0.99	0.89	0.74	0.49	0.24
0.001	0.999	0.899	0.749	0.499	0.249
0.0001	0.9999	0.8999	0.7499	0.4999	0.2499
0.00001	0.99999	0.89999	0.74999	0.49999	0.24999

Table 2.1: Variance:mean ψ ratio for different values of $\bar{p}$ and s^2.

of the variance:mean ratio ψ of the areal counts for possible values of $\bar{p}$ and s^*. These values were calculated from equation 2.2. One entry of the table is empty because s^* is always smaller than $1 - \bar{p}$. This table shows that, when the ratio $s^2/\bar{p}$ is small, the variance:mean ratio is practically equal to one. As the ratio $s^2/\bar{p}$ gets larger, the value of ψ decreases substantially. For instance, if $s^2 = 0.1\bar{p}$, ψ will at most be 0.9. If $s^2 = 0.75\bar{p}$, ψ will at most be 0.25.

However, this simple conditional analysis is misleading. Under-dispersion relative to the Binomial distribution is only seen when the probabilities for every individual are fixed and given. In practice, such probabilities are unknown, although in some circumstances we might suppose that they are sampled from a known population with a known population mean, $\tilde{p}$. In this case, it is easy to show that the resulting marginal count distribution is Binomial$(n, \tilde{p})$, regardless of the shape of the population that is being sampled.

Furthermore, in epidemiological applications, the size of the population at risk is rarely known precisely. Suppose, for example, that demographic information suggests that there are 10,000 individuals at risk. Can we feel confident that this figure is accurate? Probably not. More realistically, we might suppose that the figure is an estimate accurate to within a hundred or so. In other words, we have an estimated population size with error comparable to that of a Poisson variable. Taking this uncertainty into account, and assuming a Poisson distribution, we

see that the count Y has marginal distribution given by

$$P(Y=k)=\sum_{n=k}^{\infty}\frac{e^{-\lambda}\lambda^n}{n!}\binom{n}{k}\tilde{p}^k(1-\tilde{p})^{n-k}=\frac{e^{-\lambda\tilde{p}}(\lambda\tilde{p})^k}{k!},$$

where λ is the true population size. In other words, $Y \sim \text{Poisson}(\lambda\tilde{p})$. For these reasons, under-dispersion of count data is, in practice, almost never observed.

2.3 Generalized linear models

Generalized linear models (GLM) are comprehensively discussed in McCullagh and Nelder (1989) and have been extensively used in geographical epidemiology (e.g. Breslow and Clayton (1993)).

In the GLM, the response variable $\boldsymbol{Y}=(Y_1,\ldots,Y_n)$ has a distribution in the exponential family:

$$f(\boldsymbol{y}|\boldsymbol{\eta}) \;=\; \prod_i \exp\left[\frac{(y_i\eta_i-g(\eta_i))}{\varphi}+c(y_i,\varphi)\right], \tag{2.8}$$

where $\boldsymbol{y}=(y_1,\ldots,y_n)$ are the values of the observed $\boldsymbol{Y}$. The distribution is parameterized with respect to the canonical parameters η_i and the dispersion parameter φ. Conditional on parameters η_i the observations Y_i are assumed independent.

The GLM permits the inclusion of covariates through a link function $h(.)$:

$$h(\boldsymbol{\mu}) \;=\; \boldsymbol{X}\boldsymbol{\beta}, \tag{2.9}$$

where $\boldsymbol{\mu}=E[\boldsymbol{Y}]$. The function $h(.)$ is called the canonical link function when $h(.)$ is such that $h(\boldsymbol{\mu})=\boldsymbol{\eta}$.

We assume throughout this thesis that the Poisson approximation holds for the distribution of the counts, $\boldsymbol{Y}$, of rare diseases (sections 2.1 and 2.2). The

values of the areal probabilities, $\bar{p}$, of disease are typically unknown. Assuming the areal counts are conditionally independent, they can be analysed in a GLM framework.

2.4 Multiplicative versus additive models

The choice between multiplicative and additive models is part of the more general topic of model uncertainty. It must be noted that we do not assume that there is a true model. We can only attempt to find a model that best explains the variability observed in the data.

In epidemiological models, the covariates are usually assumed to be multiplicative (Breslow and Day, 1987, p. 125). In addition, the multiplicative model is more consistent in the sense that additive models may generate negative rates. The choice of a multiplicative model corresponds to using a logarithmic link function in the Poisson GLM, this being the canonical link for this model (McCullagh and Nelder, 1989). Formally, the risk-exposure multiplicative model is given by:

$$\begin{aligned} Y_{is} &\sim \text{Poisson}(n_{is}p_{is}), \\ \log(p_{is}) &= \boldsymbol{\beta}^t \mathbf{x}_i, \end{aligned}$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{is,p-1})^t$ is a vector of covariates, $\boldsymbol{\beta}^t = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ is a vector of parameters, Y_{is}, n_{is} and p_{is} are respectively the number of cases, the number of individuals and the probability of disease in area i and in group s.

There is empirical evidence in a considerable number of epidemiological studies that age-sex effects are multiplicative (Breslow and Day, 1987, p. 125). Since estimates for age-sex specific rates are usually available, it is common practice to introduce them in the modelling of the p_{is}. Under the assumption of multiplica-

tive age-sex effects, the p_{is} can be factorized into two terms $p_{is} = \theta_i p_s$, which leads to the model:

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i\theta_i), \\ \log(\theta_i) &= \boldsymbol{\beta}^t \mathbf{x}_i, \end{aligned} \tag{2.10}$$

where the E_i are the areal expected numbers of cases. The parameter θ_i is usually named the *areal relative risk*.

2.5 Individual versus aggregated effects

When exposures of interest are included in the model, the aim is to know what is the effect of the exposures on the probability of an individual developing a disease. For areal health data, the exposures of interest are usually not available for each individual but rather as averages over the areas. If the exposures for all individuals in the same area are the same, there is no loss of information in considering areal averages rather than individual exposures. However, the geographical areas are typically not chosen to correspond to constant exposures, but are chosen for administrative reasons. When the multiplicative model is valid and the within area variabilities of the exposures of interest are not negligible, inference on the effects of the exposures may be biased. To ignore this bias in the interpretation of the statistical results leads to what is known in the literature as the *ecological fallacy* (Robinson, 1950). Wakefield and Salway (2001) discussed the use of complementary within-area information on exposures for improving analysis and interpretation of areal data.

Despite the danger of the ecological fallacy, areal studies may be useful for suggesting hypotheses for further individual studies. Areal studies can be prefer-

able to individual studies when the factors involved are difficult to measure individually but are more reliable as averages over groups (Plummer and Clayton, 1996).

2.6 Over-dispersion and spatial correlation

As discussed in section 1.1.3, after the inclusion of relevant covariates in the model, there might still exist unexplained variability. This unexplained variability is usually named *overdispersion* or *extra-Poisson variation* ($\mathrm{Var}[Y] > \mathrm{E}[Y]$) and might be, in some cases, a consequence of spatial correlation between the areal counts. Extra-Poisson variation should be accounted for to avoid incorrect inferences about the significance of the effects of the covariates (Cliff and Ord, 1981). In addition, the knowledge of the spatial correlation structure of the data can be of interest because the structure of this unexplained risk may suggest possible aetiological factors.

Generalized linear models can easily be extended to capture the spatial structure and the extra-Poisson variation of the data. A common strategy has been to introduce random intercepts to account for the extra variation and spatial pattern of the areal counts.

2.7 Constant versus varying exposure coefficient

Exposure coefficient refers to the effect of an exposure on the risk of disease. The exposure coefficient is measured by the regression coefficient β in modelling the effect of an exposure x on the disease outcome, as exemplified in the model 2.10 above.

In ecological studies, the exposure coefficient is usually assumed to be constant across the region of study. However, this assumption is not always appropriate because the exposure coefficient may be modified by other factors, such as diet, cultural habits or poor living conditions. When bacterial infections play a role in the aetiology of the disease, the exposure coefficient may change according to the virulence of the bacterial agent (chapter 4). If the bacterial infection is contagious, strains of similar virulence are more likely to be found in points close in space and the exposure coefficient variation may be suitable for spatial modelling. In this thesis, we focus on the variation of exposure coefficients that might result from varying virulence of bacterial infections.

Chapter 3

Stomach cancer

Clear declines in mortality due to stomach cancer have been observed throughout the industrialized world over the last few decades. Despite this improvement, stomach cancer still ranked among the ten top cancer sites for both incidence and mortality in the 1990s. The great majority of stomach cancers are still detected at a late stage (Everett and Axon, 1998; Hohenberger and Gretschel, 2003), and five-year relative survival rates are low (González et al., 2002). In Europe, the average survival rates are about 24% for females and 21% for males (Berrino et al., 1999). For this reason, there is great interest in improving prevention and, particularly, in identifying the aetiology of the disease.

Multiple factors have been linked to the risk of stomach cancer. The *Helicobacter pylori* infection is considered a risk factor, established as a result of consistent positive associations found in several epidemiological studies worldwide (IARC Working Group, 1994; Correa, 2003). The decrease in stomach cancer mortality over the last decades suggests that life-style factors might play a major aetiological role, but the influence of genetic factors is also hypothesized (González et al., 2002).

The purpose of this chapter is to present key results and ongoing hypotheses on the aetiology of stomach cancer. We will end by putting forward the hypothesis that the virulence of *H. pylori* as well as genetic differences among the exposed sub-populations may partially explain the inconsistent results obtained in ecological studies on the association between the *H. pylori* infection and stomach cancer.

3.1 The Helicobacter pylori infection

The infection with *H. pylori*, a bacterium that colonizes the human stomach, is classified as a human carcinogen by the International Agency of Research on Cancer (IARC Working Group, 1994). The link with stomach cancer was recognized in epidemiological case-control and cohort studies using individuals as the observation units (Correa, 2003). On the basis of an international study, the EUROGAST Study Group (1993) concluded that, after accounting for sex, the infection could explain one-fifth of the variation in gastric mortality rates worldwide ($R^2 = 0.2$), and about one-third of the variation of the incidence rates ($R^2 = 0.3$).

The *H. Pylori* infection is both chronic and very common. Not all persons infected develop stomach cancer (Yamaguchi and Kakizoe, 2001; Correa, 2003), but due to the high prevalence rates of *H. Pylori* infection worldwide, the infection could be responsible for a substantial proportion of the cases of stomach cancer. In Forman et al. (1991), the attributable risk was found to be high: between 35% and 55% of gastric tumours were attributed to the *H. pylori* infection.

It is known that the *H. pylori* bacteria are characterized by high genetic diversity. Studies conducted at the family and country levels provided some

insight into the genetic variability of the bacteria. Among members of the same family, the bacteria can be indistinguishable, or show minor differences, which is consistent with intrafamilial transmission (Pérez-Pérez et al., 1997; Achtman et al., 1999; Axon, 1999; Covacci et al., 1999). By contrast, major genetic differences have been identified at the country level (Van Der Ende et al., 1998; Van Doorn et al., 1999; Miwa et al., 2002).

Despite the differences found at the country level, the genetic structure of the bacteria reveals strong grouping according to the continent of origin (Achtman et al., 1999; Covacci et al., 1999; Van Doorn et al., 1999; Falush et al., 2003). In a recent study on the genealogical tree of *H. pylori*, Falush et al. (2003) showed that DNA sequences differ with the geographical source of isolation and described how *H. pylori* has accompanied human migrations since ancient times. Their study is based on samples of *H. pylori* from individuals from different countries, ethnicities and populations with different language families. According to the results obtained, modern populations of *H. pylori* appear to have derived their gene pools from ancestral populations that arose in Africa, Central Asia, and East Asia, which evolved to four populations as a result of prehistoric migrations of various human ethnic groupings.

As Covacci et al. (1999) explain, this common evolution of *H. pylori* and mankind is maybe due to the particular characteristics of transmission of *H. pylori*. When infectious diseases spread quickly, strains from different regions are fairly similar. By contrast, *H. pylori* spreads at a slow pace due to its family-linked mode of transmission. Until fairly recently, the exchanges among different human communities have been limited, resulting in clear geographical grouping of human genetic traits, and accordingly, of *H. pylori* strains.

Overall, these findings suggest a relation between geographical location and the populations of *H. pylori*. This geographical relation has two important aspects: (i) strong genetic similarities of the *H. pylori* strains are expected in locations close in space; (ii) clear groups of countries with similar *H. pylori* strains are expected.

Falush et al. (2003) also claimed that the populations of *H. pylori* are now merging more due to globalization. The study has shown coexistence of native populations of the bacteria with bacteria imported in the last few hundred years due to recent human migrations, such as the European colonization of the Americas, Africa and Australia and the slave trade with the Americas.

3.1.1 The virulence of Helicobacter pylori

There are indications that high prevalence of *H. pylori* infection does not always go together with high stomach cancer rates. Worldwide ecological studies searching for an association between the prevalence of the *H. pylori* infection and stomach cancer rates have not always found strong correlations between the two variables (Lunet and Barros, 2003; EUROGAST Study Group, 1993; Pérez-Pérez et al., 1997). It is also known that in some African and Asian countries, such as India, stomach cancer rates are low despite the high prevalence of *H. pylori* infection. It seems natural to ask why there is such a disparity of results.

Recent findings suggest that genetic differences among distinct bacterial populations of *H. pylori* may correlate with differences in virulence, i.e. not all strains seem to be equally related to stomach cancer (Ando et al., 2002; Atherton, 1997; Axon, 1999; Basso et al., 1998; Kidd et al., 1999; Miehke et al., 2000; Parsonnet et al., 1997). Consequently, the virulence of *H. pylori* should affect the values of

the bacteria exposure coefficient of stomach cancer.

As discussed at the beginning of section 3.1, similar strains are likely to be found in individuals living in locations which are close in space and different genetic strains of *H. pylori* with different virulence are likely to exist in different continents. In view of this, it can be argued that the association between the prevalence of the *H. pylori* infection and stomach cancer rates is likely to differ from one geographic region to another. The association between prevalence and disease will depend on the varying degree of virulence of the strain and the varying genetically determined susceptibility of the human sub-populations (section 3.2). Ecological studies on regions of different bacterial/human sub-populations may thus miss the impact of the infection if those differences are not taken into account.

3.2 Genetic susceptibility

González et al. (2002) presented a comprehensive review of published evidence on the influence of genetic susceptibility on stomach cancer in humans. The authors claim that gene-environment interactions could explain the variation of stomach cancer. They report inconsistent results, but argue that these may result from limitations in design and the presence of confounding factors.

You et al. (2000) reported results of a cross-sectional study in China, in which precancerous gastric lesions were found to be significantly associated with blood type A and parental history of cancer. Since blood type is part of genetic inheritance, this finding seems to support some influence of genetic factors.

A large number of studies have found strong evidence of family aggregation of cases of stomach cancer (Buiatti et al., 1991; Inoue et al., 1998; Dhillon

et al., 2001; Yatsuya et al., 2002; Kondo et al., 2003). These could suggest an aetiological role for genetic factors. However, results presented in Chang et al. (2002) support a role for *H.Pylori* infection in family aggregation of stomach cancer. Without other evidence, it is not possible to disentangle these two explanations.

3.3 Life-style factors

Some authors have argued that the major causes of stomach cancer are related to life-style factors. This hypothesis was mainly based on the following facts: (i) the rapid decrease in rates over recent decades; (ii) the remarkable geographical variability of the rates; (iii) an observed decrease in risk among migrants from high-risk areas to low risk areas; and (iv) higher stomach cancer rates in lower socio-economic classes.

Migrant studies in particular have been used to assess the predominance of life-style factors over genetic factors in the aetiology of cancer, but even these studies show inconsistent findings.

Wilkens and Lee (1998) described a study on the mortality from cancer of Japanese migrants in the United States, in which stomach cancer mortality rates showed steady reductions from those in the parent country, where rates are high, to those in the host country, where rates are low. They argued that environmental and lifestyle factors must be responsible for stomach cancer incidence because the migrant populations and the people in their country of origin shared the same genetic background.

On the other hand, Gregorio et al. (1992) observed that among the European communities in Connecticut (U.S), the Portuguese community presented the

highest stomach cancer mortality rate. Portugal being the country in Western Europe with highest stomach cancer mortality, this finding could support the importance of genetic factors in stomach cancer. However, the prevalence of the *H. pylori* infection in Portugal is also very high (82.8% according to Lunet and Barros (2003)). Since the bacteria tend to spread within families, the prevalence may be as high in the migrant communities as in the country of origin. This could explain the high stomach cancer rates in the Portuguese migrant community.

3.3.1 Diet

It seems reasonable to assume that diet must have an influence in stomach cancer incidence and mortality and several studies worldwide have supported this hypothesis (Hirohata and Kono, 1997; La Vecchia et al., 1997; Ahn, 1997; Kaaks et al., 1998; Ji et al., 1998; López-Carrillo et al., 1999; Kobayashi et al., 2002). Despite the difficulty in obtaining dietary information, it has been possible, in some cases, to assert consistent results. There is evidence to suggest that an increased risk of stomach cancer is associated with a dietary pattern characterized by low intake of animal fat and protein, vegetables and fresh fruit, and a high intake of complex carbohydrates, hard grains, salt and nitrites, smoked fish and meat. The role played by nutrients is more difficult to assess, but vitamin A, vitamin E, vitamin C and carotene have been suggested to have a protective effect, while risk has been associated with high intake of salt and carbohydrates, in association with N-nitroso compounds (Azevedo et al., 1999).

3.3.2 Alcohol drinking

The results on the relation between alcohol consumption and stomach cancer rates in case-control studies are not consistent. For example, Hoey et al. (1981), Destefani et al. (1990) and Azevedo et al. (1999) reported a positive association with wine consumption, but Davanzo et al. (1994) found a non-significant association for wine using data from a case-control study in Northern Italy. Boeing et al. (1991) reported a decrease in risk with wine and liquor consumption, but an increase in risk with beer consumption. Lagergren et al. (2000) and Rao et al. (2002) found no significant association. Jedrychowski et al. (1993) and Muñoz et al. (2001) found significant results, both reporting a risk about 3-times higher for regular drinkers when compared to never drinkers.

3.3.3 Smoking

Trédaniel et al. (1997) present a review of the evidence found in epidemiological studies on the relation between tobacco smoking and stomach cancer. In their review, all cohort studies showed a positive association between smoking and stomach cancer, while case-control studies reported inconsistent results. Kobayashi et al. (2002) argued that the high risk reported for current smokers in some studies is due to the relative deficit of vegetable and fruit consumption in smokers. However, recently, a large number of studies have reported significant increase in risk for smokers when adjusted for dietary items, including vegetables and fruit (Inoue et al., 1999; Lagergren et al., 2000; Chao et al., 2002; Sasazuki and Sasaki, 2002; González et al., 2003).

3.3.4 Socio-economic factors

Apart from dietary factors, lower socio-economic status has been associated with a higher risk of stomach cancer. Buiatti et al. (1991) and Muñoz et al. (2001) reported, from results of case-control studies, higher risk in lower socio-economic levels.

It has been suggested that poor living conditions favourable to *H. pylori* infection may explain the association between lower socio-economic status and risk of stomach cancer (Yamaguchi and Kakizoe, 2001).

3.4 Discussion

The association between the *H. pylori* infection and stomach cancer appears to be undeniable. In addition, other factors, usually associated with stomach cancer, may be so associated through their relation with the *H. pylori* infection. For instance, both the *H. pylori* infection and stomach cancer have been linked with life-style factors, such as socio-economic status. Previous conclusions on the aetiological role of genetic factors on stomach cancer can also be reinterpreted in the light of the *H. pylori* infection, especially the ones based on the clustering of stomach cancer within families.

Nonetheless, some geographical studies worldwide failed to find an association between stomach cancer rates and the prevalence of *H. pylori*. It can be argued that this is due to the existence of strains of *H. pylori* with different virulence in different countries or to the variation in the susceptibility to stomach cancer among populations of different origin and genetic heritage. It is important to recognise that disease susceptibility may be genetically determined and

differentially expressed in various human sub-populations. The spread of human populations on an evolutionary time-scale will produce geographical variation of cancer incidence. Effects that are attributed to varying "bacterial virulence" may therefore reflect varying human genetic susceptibility.

Chapter 4

Varying exposure coefficient

In the previous chapter, we decided that the virulence of *H. pylori* for particular human sub-populations might be important in the analysis of stomach cancer incidence and mortality worldwide. However, data on the virulence of *H. pylori* are inexistent. Although some potential virulence-associated genes have been identified (Basso et al., 1998; Kidd et al., 1999; Miehke et al., 2000; Parsonnet et al., 1997), there is not a complete knowledge of the genetic traits of the bacteria that are actually responsible for the disease - this undermines any attempt to obtain data on virulence.

In the absence of information on virulence, models attempting to assess the association between *H. pylori* infection and stomach cancer must allow for a varying exposure coefficient for *H. pylori* prevalence. The variation in the exposure coefficient will reflect the variation in the carcinogenic virulence of the strain for particular human sub-populations.

However, data may not even be available for the prevalence of the bacteria. In that case, surrogate variables may be considered: for example, several studies seem to indicate that the prevalence of *H. pylori* decreases with increasing socio-

economic status.

This chapter discusses the modelling of varying exposure coefficients within the framework of mixed effect models. It shows that in some circumstances the variation of the baseline risk may be caused, at least partly, by the variation of the exposure coefficient. The chapter further highlights some identifiability issues in the models presented. Finally, it proposes the use of a Bayesian approach to address the modelling of varying exposure coefficients.

4.1 The effects of virulence on the exposure coefficient

For purposes of illustration, consider that we are in a situation in which three *H. pylori* strains with different virulence exist in respectively three different areas, and that the human populations in these areas are genetically similar. Assume that the association between the prevalence of the *H. pylori* infection and the log relative risk of stomach cancer is approximately linear, although each areal line has its specific slope reflecting the area-specific virulence of the bacteria. This scenario is illustrated by the three lines shown in graph 4.1(a). The lines intersect at one point: when the prevalence of the bacterial infection is zero. At this point, the risk equals the baseline risk. The differences in virulence from area to area thus cause a variation in the areal exposure coefficient but not in the baseline risk.

However, if a surrogate variable for the prevalence of *H. pylori* infection is used, the interpretation of the areal regression lines is not so straightforward. For example, graph 4.1(b) illustrates a hypothetical relation between the log relative

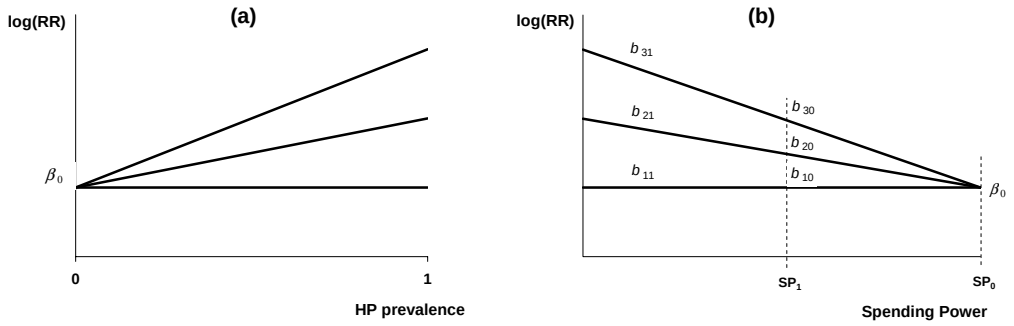


Figure 4.1: The hypothetical influence of *H. pylori* virulence on the association between stomach cancer risk and (a) *H. pylori* prevalence; (b) spending power.

risk, $\log(\mathrm{RR})$, and spending power, SP, on the basis of an inverse linear relation between the prevalence of the *H. pylori* infection and socio-economic status.

If it is believed that the prevalence of *H. pylori* decreases gradually to zero with spending power, perhaps the simplest way of dealing with this situation is to recode the surrogate variable to reflect the inverse relationship. For example, if socio-economic status is measured by spending power, then the inverse of spending power can be used, or indeed any other negative power of this quantity. In this way, high socio-economic status translates to low socio-economic poverty, and vice-versa. After such recoding, the relationship between log(RR) and the surrogate variable will be similar to that of log(RR) and HP prevalence, with a constant intercept at 0.

If however zero prevalence is attained at a fixed threshold of spending power, the problem of using an indicator such as spending power as a surrogate for *H. pylori* infection is that the corresponding level for *zero prevalence* is unknown. We would in general choose an arbitrary origin. However, with an arbitrary origin the intercept may not be constant any more. The baseline risk will vary from line to line (i.e. from area to area). This is illustrated in graph 4.1(b) when the

origin is moved from SP_0 to SP_1. Then, not only the exposure coefficient varies across the different lines but also the baseline risk.

Since all lines in the graph 4.1(b) satisfy

$$\log(\mathrm{RR}) \;=\; b_{i1}(\mathrm{SP} - \mathrm{SP}_0) + \beta_0, \tag{4.1}$$

a translation of the origin will lead to the equation,

$$\log(\mathrm{RR}) \;=\; b_{i1}(\mathrm{SP} - \mathrm{SP}_1) + b_{i1}(\mathrm{SP}_1 - \mathrm{SP}_0) + \beta_0, \tag{4.2}$$

thus indicating that the varying intercept b_{i0} and slope b_{i1} are related by

$$b_{i0} \;=\; b_{i1}(\mathrm{SP}_1 - \mathrm{SP}_0) + \beta_0. \tag{4.3}$$

Hence, the vector $\boldsymbol{b}_{\bullet 0} = (b_{10}, \ldots, b_{n0})^t$ is a linear transformation of the vector $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{n1})^t$, where n is the total number of areas. If the two vectors are assumed random, their distributions are related by a linear transformation of the variables. For example, if $\boldsymbol{b}_{\bullet 1}$ is multivariate Gaussian distributed, then $\boldsymbol{b}_{\bullet 0}$ will also be multivariate Gaussian distributed. In general, the parameter β_0 would be captured by the fixed intercept term of the model.

If there were not any other source of variation of the baseline risk across areas, the assumption $\boldsymbol{b}_{\bullet 0} = \phi \boldsymbol{b}_{\bullet 1} + \beta_0 \mathbf{1}_n$ would be appropriate. In practice this is unrealistic, as other unknown factors usually contribute to the development of disease. If the random intercept $\boldsymbol{b}_{\bullet 0}$ were modelled as $\phi \boldsymbol{b}_{\bullet 1} + \tilde{\boldsymbol{b}}_{\bullet 0} + \beta_0 \mathbf{1}_n$, with $\tilde{\boldsymbol{b}}_{\bullet 0}$ a random effect, the first term $\phi \boldsymbol{b}_{\bullet 1}$ would account for the variation induced by the variability of the exposure coefficient, and the second term $\tilde{\boldsymbol{b}}_{\bullet 0}$ would account for the variability in the baseline risk due to factors other than the variation of the exposure coefficient. Thus in theory we have a way of separating the two terms $\boldsymbol{b}_{\bullet 1}$ and $\tilde{\boldsymbol{b}}_{\bullet 0}$. However, the model has too many parameters and

not all of them will be identifiable unless additional constraints are introduced or a sufficient number of data points per area are available.

An alternative strategy for dealing with an *unknown origin* is to set the intercept constant but to consider an additional parameter δ that captures the true origin,

$$\log(\mathrm{RR}) \quad = \quad \beta_0 + \beta_1 + b_{i1}(\mathrm{SP} - \delta).$$

However, this model is equivalent to a model with a random intercept and a random slope, as a simple decomposition of the right hand side of the expression above will show.

As far as we know, none of the formulations above has been used in the literature. Some varying coefficient models assume that only the slope is varying; others assume that both the intercept and the slope are random but independent. Any of these assumptions clearly violate the expected behaviour of the random intercept in the case of the graph 4.1(b) above. We will study in later chapters other issues posed by these types of modelling.

A few studies in the literature have assumed multivariate distributions for the intercept and the slope, although implementation and identifiability issues have been recurring. We will discuss those issues later in this chapter and throughout this thesis as they arise in our applications.

4.2 Modelling varying exposure coefficients

We now discuss models for capturing varying exposure coefficients. As with many statistical studies, the models we propose are meant to be a useful simplification for understanding the processes underlying the data. Assume, without loss of

generality, n areas in which we observe random variables Y_i, $i = 1, \ldots, n$. In geographical epidemiology studies, the observations y_i are typically disease counts or log transformations of the disease rates, for which it is usual to assume a Poisson or a normal distribution, respectively. Both distributions belong to the exponential family (section 2.3, equation 2.8). They can thus be considered within the context of generalized mixed linear models. The canonical parameters $\boldsymbol{\eta}$ are related to the linear predictor through a canonical link function as follows:

$$\boldsymbol{\eta} \quad = \quad \boldsymbol{X\beta} + \boldsymbol{Zb}, \tag{4.4}$$

where $\boldsymbol{X}$ is the $n \times p$ design matrix of independent variables, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^t$ is the $p \times 1$ vector of parameters for the fixed effects, $\boldsymbol{Z}$ is the $n \times nq$ design matrix of predictor variables for the random effects and $\boldsymbol{b} = (b_{1\bullet}, \ldots, b_{n\bullet})^t$ is the $nq \times 1$ vector of random effects. We will consider models with arbitrarily many covariates, but total order up to 1 (i.e. we will exclude polynomial terms of degree superior to 1).

The model above has been explored in geographical epidemiology, mainly for estimating areal relative risks, in which case only the intercept was considered random (i.e. $\boldsymbol{Z} = \boldsymbol{I}$). Although it is fairly common in mixed effect models to assume $\boldsymbol{b}_{\bullet 0} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, in geographical studies of disease other assumptions have usually been considered:

- Clayton and Kaldor (1987) assumed a conditional autoregression for $\boldsymbol{b}_{\bullet 0}$;
- Breslow (1996) assumed an intrinsic autoregression for $\boldsymbol{b}_{\bullet 0}$;
- Besag et al. (1991) assumed a convolution distribution for $\boldsymbol{b}_{\bullet 0}$, defined as a sum of an independent Gaussian with an intrinsic autoregression.

In these examples, the aim was to obtain improved estimates of the relative risks in each area, and better inference on the fixed effects by accounting for extra variation due to unknown factors. We will now also consider the slope as a random effect, in an attempt to model the variation of the exposure coefficient.

We assume at this stage that the model only includes one covariate, z, the same for both fixed and random slopes,

$$\eta_i \;\; = \;\; \beta_0 + z_i\beta_1 + b_{i0} + z_i b_{i1}\,, \tag{4.5}$$

with $\sum_i b_{i0} = 0$ and $\sum_i b_{i1} = 0$. The parameter β_0 corresponds to the global baseline risk, and its value is assumed to be the same in all areas. The variation in the baseline risk caused by unknown factors or induced by the varying exposure coefficient is assumed to be accounted for by the parameter b_{i0}, which is area dependent. The parameter b_{i1} is also area dependent and reflects the varying exposure coefficient. One main purpose of this model is to obtain an estimate of the geographical variation of the exposure coefficient b_{i1}.

Due to the nature of the spread of *H. pylori* infection, spatial patterns are expected for the values of the corresponding exposure coefficient. Information on the expected spatial variation of the exposure coefficient can be introduced through a prior distribution for the random slopes $\boldsymbol{b}_{\bullet 1}$.

4.3 Discrete versus continuous exposure coefficient

Depending on our assumptions, it may be reasonable to consider the exposure coefficient as discretely or continuously varying. For the first, regression mix-

tures can be considered (chapters 6 and 7). For the second, spatially varying regressions using Markov random fields may be appropriate (chapter 8).

In particular, regression mixtures may be helpful when the presence of discontinuities and a patchy pattern of the exposure coefficients are expected. In standard regression mixture models, the variables Y_i are assumed to belong to one of M different groups. Each group has a different value of b_{i1}. The geographical location of the areas is ignored so that members of a mixture component can be spread over the whole region of study.

As *H. pylori* infections of the same strain are expected to be found in homogeneous ethnic groups in nearby areas, spatial constraints can be appropriate, by assuming for example the existence of M *connected* sets of areas with different values of b_{i1} (see section 7.1 for a definition of connected set of areas). With this aim, we propose in chapter 7 the use of a *geographical clustering model*. This model is a special case of the regression mixture model. The sample space is restricted to configurations of connected sets of areas.

Alternatively, the exposure coefficient may be expected to vary at the local level. Prior distributions introducing local dependence can then be useful. For example, the vector $\boldsymbol{b}_{\bullet 1}$ may be assumed to have as prior distribution an intrinsic autoregression. This prior promotes the similarity of the parameter b_{i1} in adjacent or nearby areas.

We will discuss these different modelling choices in more detail in the next chapters.

4.4 Random versus constant intercept

We have seen in section 4.1 that a varying exposure coefficient may induce a varying intercept. However, in some epidemiological studies, there may be strong evidence for considering the intercept to be constant. For example, in section 4.1, it was sensible to assume the intercept to be constant when the origin for the *H. pylori* prevalence was set at zero.

If there are no a priori reasons for considering the intercept to be constant, it is preferable to allow for a varying intercept, which may be due to unknown factors or induced by the varying exposure coefficient. The variation in the baseline risk can be modelled in a similar way to the variation of the exposure coefficient itself, by introducing, for example, a spatial or mixture prior distribution on this parameter. However, there are concerns about the non-identifiability of the parameters when both intercept and slope are random.

4.5 Parameter identifiability and informative data

Hierarchical models introducing random effects often yield parametric models of high dimension. For instance, the model displayed above in expression 4.5 includes more than one parameter per area. If only one observation is available per area, the data may not provide enough information for uniquely estimating all the parameters, and therefore some parameters will not be identifiable.

Poirier (1998); Gelfand and Sahu (1999) have provided formal definitions of identifiability and related concepts. Formally, a parameter $\theta \in \Theta$ is *nonidentifiable* when for any $\theta^{(1)} \in \Theta$ there exists $\theta^{(2)} \in \Theta$ with $\theta^{(2)} \neq \theta^{(1)}$ such that for any $\boldsymbol{y}$, $L(\boldsymbol{\theta}^{(1)}; \boldsymbol{y}) = L(\boldsymbol{\theta}^{(2)}; \boldsymbol{y})$, where $L(.; \boldsymbol{y})$ denotes the likelihood function.

A case in point is the non-identifiability of the parameters of mixture models (section 6.4.2).

Possible ways to deal with the lack of identifiability include constraints on the parameter space so that the likelihood is an injective function of $\boldsymbol{\theta}$ in the restricted parameter space. In a Bayesian setting, prior distributions can be used as an alternative. Since the posterior distribution of a parameter is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})},$$

the prior distribution $\pi(\boldsymbol{\theta})$ can solve the ambiguity in the likelihood. It is sufficient that $\pi(\boldsymbol{\theta}^{(1)}) \neq \pi(\boldsymbol{\theta}^{(2)})$ for any $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$ that yield $L(\boldsymbol{\theta}^{(1)};\boldsymbol{y}) = L(\boldsymbol{\theta}^{(2)};\boldsymbol{y})$.

If the prior does not help in identifying the non-identifiable parameters, the posterior distribution can be difficult to interpret as for any $\theta^{(1)} \in \Theta$ there will be $\theta^{(2)} \in \Theta$ and $\theta^{(2)} \neq \theta^{(1)}$ with, for any $\boldsymbol{y}$, $\pi(\boldsymbol{\theta}^{(1)}|\boldsymbol{y}) = \pi(\boldsymbol{\theta}^{(2)}|\boldsymbol{y})$. This property may raise difficulties in the posterior estimation of the parameters (e.g. mixture models, section 6.4.2).

However, even if the prior promotes the identifiability of the parameters, others issues can arise. If the likelihood does not depend on $\boldsymbol{\theta}$, the posterior and the prior will be identical, $\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \pi(\boldsymbol{\theta})$. In this case, we say that there is no *Bayesian learning* or that the *data is not informative about* $\boldsymbol{\theta}$ (Gelfand and Sahu, 1999). Lack of Bayesian learning can cause computational difficulties when the prior is improper because the posterior will also be improper. If the prior is proper, poor Bayesian learning can still lead to models which are too sensitive to the choice of the prior distributions.

In the case of the model given by expressions 2.8 and 4.5, all parameters are identifiable when there are more than two observations per area because all the regression parameters are identifiable if the model is applied separately for each

area. Priors on the random effects can still be useful for improving the reliability of the estimates. Spatial priors, which promote the use of the information from neighbouring areas, have been used with this aim in e.g. Assuncão (2003).

When only one data point is available per area, prior distributions may facilitate the estimation of the otherwise unidentifiable parameters. Gelfand et al. (2003) used a Gaussian spatial process prior for the random effects and claimed that the parameters were uniquely identified. On the other hand, other authors pointed out that spatial autoregressions may not yield all the parameters identifiable in the sense that the data is not informative about the parameters (Assuncão, 2003). As a result, the posterior distribution will be very dependent on the prior choice, which may lead to unreliable estimates (Richardson and Best, 2003).

In complex models, using complex priors, it is not easy to assess whether the prior will allow for Bayesian learning on the unidentifiable parameters. Simulations may be useful to shed light into the identifiability of the model parameters and into Bayesian learning. We are especially interested in this thesis in the case in which the prior on the random effects includes spatial information. Identifiability issues will be discussed for specific models in the next chapters.

4.6 Bayesian framework

We chose to use a Bayesian hierarchical framework to address the modelling of varying exposure coefficients. First, regarding the applications that we have in mind, the possibility of incorporating prior information constitutes an advantage. Secondly, Bayesian methods will facilitate the implementation of the models proposed here. Thirdly, the Bayesian framework is attractive in the proposed setting,

because we are interested in inference for the random effects. In particular, we obtain posteriors for all model parameters.

Choice of suitable priors is generally a contentious issue in the application of Bayesian methodology. Although the possibility to introduce complementary knowledge through the prior distributions is usually considered as one of the advantages of Bayesian approaches, much effort has been expended by Bayesians in the search for *non-informative* priors. In this thesis, we will assume that we have some prior information about the exposure coefficient. The challenge is then how to represent this knowledge mathematically as a prior distribution of the parameters, while also avoiding an over-informative prior.

Chapter 5

Invariant mixed effects models

In simple linear models, linear transformations of the covariates do not change the model class and assessments of model fit are unaffected. We refer to the models with this property as being *coding invariant*. However, other models are generally not coding invariant, like polynomial and mixed effects models, where transformed covariates can lead to different statistical results. For polynomial regression models, Peixoto (1990) showed that under general conditions well-formulated models (section 5.3) are invariant, and vice-versa. For mixed effect models, no general results have yet been established.

The problem of coding invariance was described by Longford (1993) and Morrel et al. (1997) in the context of the mixed effects models, but has been for the most part largely ignored in the literature. Morrel et al. (1997) have shown that the notion of the well formulated model is also crucial in mixed effects models, although sufficient conditions for invariance were not given by these authors. In particular, Morrel et al. (1997) stated that the model should be well formulated in both fixed and random effects for it to be invariant. This condition is sufficient for the fixed part of the model, but not for the random

part.

Longford (1993) noted that patterns in the covariance matrix of the random effects may change when the explanatory variables are transformed. This phenomenon was illustrated in Morrel et al. (1997) with a model including a random slope but without a random intercept. We will further show in this chapter that independence between the random intercept and the random slope also leads to models which are not coding invariant.

When the interest of the problem is on the study of the structure of the random effects, the lack of coding invariance constitutes a drawback. For example, some spatial regression models used in Assuncão (2003), Gamerman et al. (2003), and Gelfand et al. (2003) are not invariant under coding transformations. The spatial pattern of the spatially varying coefficients may indeed vary with the scale of the covariate.

We illustrate here the impact of the lack of coding invariance with the following example, in which the covariate has been centred about its mean.

Example 1

The data set in this example is the one analysed in chapter 10, fitted with the model described in expressions 2.8 and 4.4, adapted for Poisson responses, and with a constant intercept and a random slope distributed as an intrinsic autoregression (Mollié, 1996). Figure 5.1 shows the pattern of the estimated random slopes when we used the original covariate (figure 5.1(a)) and when the covariate was centred about its mean (figure 5.1(b)).

All maps presented here and throughout this thesis were produced with the package Mapinfo Professional 6.5, using the *natural break* option for determining

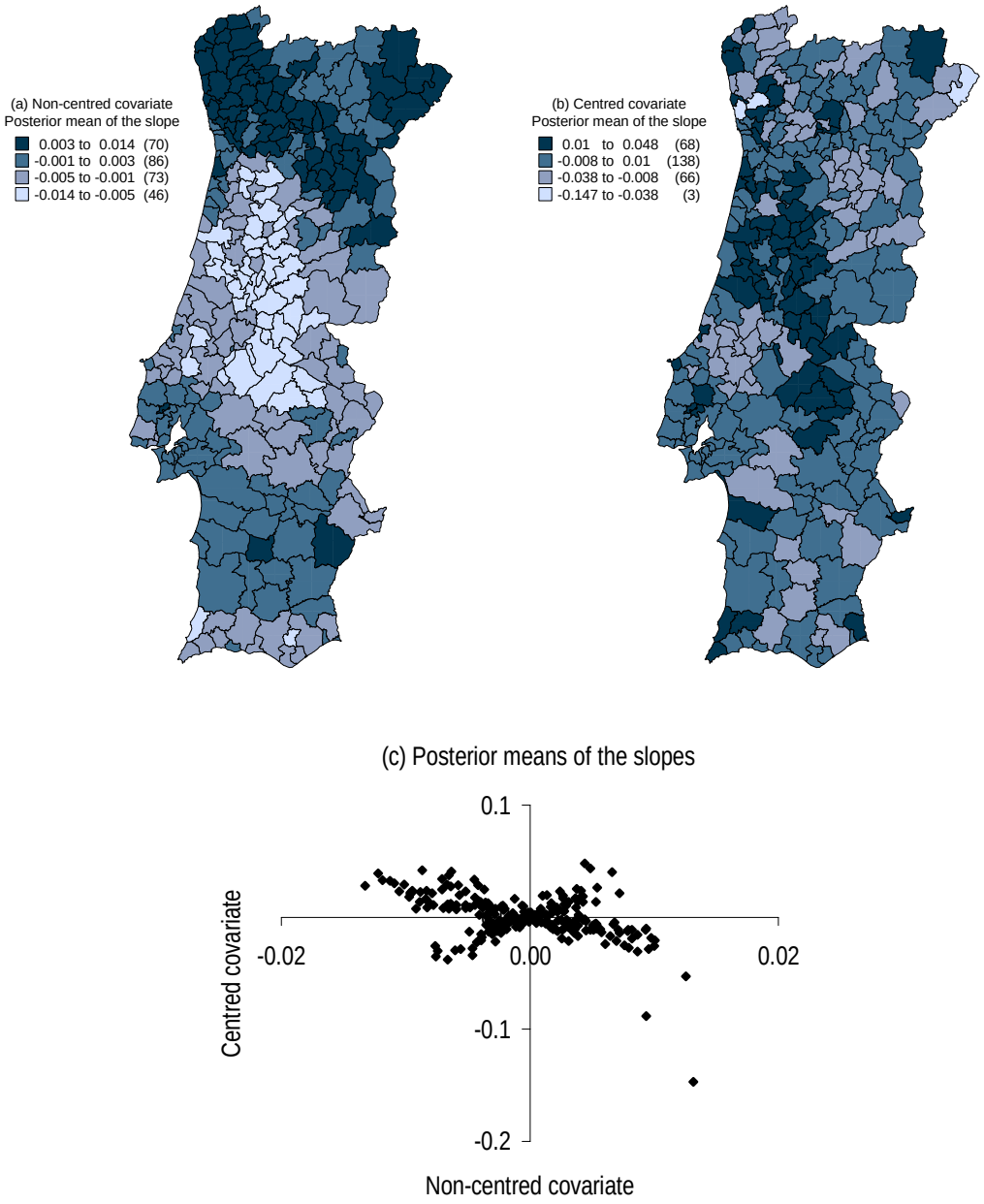


Figure 5.1: Example 1: Posterior means of the areal exposure coefficients measuring the association between stomach cancer mortality and: (a) spending power; (b) spending power centred about the mean. (c) Scatterplot of the posterior means obtained with the non-centred covariate versus those obtained with the centred covariate.

the range of the intervals (Jenks and Caspall, 1971; MapInfo Corporation, 2001). This option minimizes the sum of the absolute differences between each point and the centre of the interval it belongs to. This procedure ensures that the interval centres reliably represent the data points.

In figure 5.1(a), the analyst may conclude that there is a large cluster of lower and negative exposure coefficient in the centre of the country, whereas in figure 5.1(b), the analyst may think that there is a cluster of high positive exposure coefficient in the same area. Which one should he believe (if any)?

This example illustrates the ambiguity that may be encountered in estimating varying exposure coefficients. The relation between the slope estimates produced by the two model fits of example 1 is shown in the scatterplot of figure 5.1(c). One may ask whether there is a way to transform the estimates displayed in figure 5.1(a) into the estimates displayed in figure 5.1(b), thus establishing an equivalence between the two model fits. We argue in this chapter that there is not. The two model fits used to produce the estimates of figures (a) and (b), although apparently conceptually identical, are not equivalent.

We continue this chapter by formalizing the notion of transformation of covariates. We then discuss the impact of these transformations in mixed effects models and provide some necessary conditions for coding invariant models.

5.1 Coding transformations for fixed effects

Consider a $n \times p$ design matrix $\boldsymbol{X}$ of covariates for the fixed effects. Let $\boldsymbol{W}$ be a linear transformation of $\boldsymbol{X}$ of the form

$$\boldsymbol{W} \;=\; \boldsymbol{X}\boldsymbol{C} + \mathbf{1}_{\boldsymbol{n}}\boldsymbol{d}^t\,, \tag{5.1}$$

where $\boldsymbol{C} = \mathrm{diag}(c_0, c_1, \ldots, c_{p-1})$ is nonsingular, and $\boldsymbol{d}^t = (d_0, d_1, \ldots, d_{p-1})$. We will refer to this type of linear transformation of the covariate design matrix as a *coding transformation*. If $\boldsymbol{C} \neq \boldsymbol{I}_p$ and $\boldsymbol{d} = \mathbf{0}$, $\boldsymbol{W}$ is a *scale transformation* of $\boldsymbol{C}$. If $\boldsymbol{C} = \boldsymbol{I}_p$ and $\boldsymbol{d} \neq \mathbf{0}$, $\boldsymbol{W}$ is a *translation transformation* of $\boldsymbol{C}$.

Let us assume that the first column of $\boldsymbol{X}$ corresponds to the intercept term, i.e. the first column of $\boldsymbol{X}$ is equal to the unity vector $(1, \ldots, 1)^t$. Since it does not make much sense to rescale or translate an intercept term, the elements c_0 and d_0 are usually equal to 1 and 0, respectively. The transformation (5.1) is then equivalent to

$$\boldsymbol{W} \;=\; \boldsymbol{X}\boldsymbol{T}\,, \tag{5.2}$$

with

$$\boldsymbol{T} = \begin{bmatrix} 1 & d_1 & d_2 & \ldots & d_{p-1} \\ 0 & c_1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & c_{p-2} & 0 \\ 0 & \ldots & & 0 & c_{p-1} \end{bmatrix}.$$

The equivalence between equations 5.1 and 5.2 is easily established. Assuming equality between the right hand sides of the two equations, i.e.

$$\boldsymbol{X}\boldsymbol{T} = \boldsymbol{X}\boldsymbol{C} + \mathbf{1}_{\boldsymbol{n}}\boldsymbol{d}^t \iff$$

$$\iff \boldsymbol{X}(\boldsymbol{T}-\boldsymbol{C}) = \mathbf{1_n}\boldsymbol{d}^t \iff$$

$$\iff \boldsymbol{X}\begin{bmatrix} 0 & d_1 & d_2 & \dots & d_{p-1} \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & 0 \\ 0 & \dots & & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & d_1 & d_2 & \dots & d_{p-1} \\ 0 & d_1 & d_2 & \dots & d_{p-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & d_1 & d_2 & \dots & d_{p-1} \\ 0 & d_1 & d_2 & \dots & d_{p-1} \end{bmatrix},$$

we obtain a universal condition because all the elements in the first column of $\boldsymbol{X}$ are equal to 1.

Coding transformations had been formulated as in equation 5.2 in Morrel et al. (1997) for the $p = 2$ case. We proved in this section that the formulation of equation 5.2 can be used for coding transformations in any dimension, a result that will be particularly useful for studying coding invariance when the number of covariates is larger than $p = 2$ (section 5.4.1).

5.2 Coding transformations for random effects

Consider a $n \times nq$ design matrix $\boldsymbol{Z}$ of covariates for the random effects,

$$\boldsymbol{Z} = \begin{bmatrix} z_{10} & \dots & z_{1,q-1} & 0 & \dots & & & & 0 \\ 0 & \dots & 0 & z_{20} & \dots & z_{2,q-1} & 0 & \dots & 0 \\ & & & & & & & & \\ 0 & \dots & & & & 0 & z_{n0} & \dots & z_{n,q-1} \end{bmatrix}.$$

Let $\widetilde{\boldsymbol{W}}$ be a linear transformation of $\boldsymbol{Z}$ of the form

$$\widetilde{\boldsymbol{W}} \;=\; \boldsymbol{Z}\widetilde{\boldsymbol{C}} + \mathbf{1_n}\tilde{\boldsymbol{d}}^t\,, \tag{5.3}$$

where $\widetilde{\boldsymbol{C}}$ is a $nq \times nq$ block diagonal matrix, $\widetilde{\boldsymbol{C}} = \mathrm{diag}(\boldsymbol{C}, \ldots, \boldsymbol{C})$, with non-singular blocks $\boldsymbol{C} = \mathrm{diag}(c_0, c_1, \ldots, c_{q-1})$, and

$$\tilde{\boldsymbol{d}}^t = \begin{bmatrix} d_0 & \ldots & d_{q-1} & 0 & \ldots & & & & 0 \\ 0 & \ldots & 0 & d_0 & \ldots & d_{q-1} & 0 & \ldots & 0 \\ & & & & & & & & \\ 0 & \ldots & & & & 0 & d_0 & \ldots & d_{q-1} \end{bmatrix}$$

is a $n \times nq$ matrix.

We will use the same terminology as was used before for coding transformations of the design matrix of the fixed effects, and also refer to these types of linear transformations of the covariate design matrix $\boldsymbol{Z}$ as *coding transformations*. Further, if $\widetilde{\boldsymbol{C}} \neq \boldsymbol{I}_{nq \times nq}$ and $\widetilde{\boldsymbol{d}} = \boldsymbol{0}$, $\widetilde{\boldsymbol{W}}$ is a *scale transformation* of $\widetilde{\boldsymbol{C}}$. If $\boldsymbol{C} = \boldsymbol{I}_{nq \times nq}$ and $\widetilde{\boldsymbol{d}} \neq \boldsymbol{0}$, $\widetilde{\boldsymbol{W}}$ is a *translation transformation* of $\widetilde{\boldsymbol{C}}$.

Let us assume that the entries $(z_{10}, \ldots, z_{n0})$ of $\boldsymbol{Z}$ correspond to the random intercept term, and their value is therefore equal to one. Similarly to section 5.1, usually $c_0 = 1$ and $d_0 = 0$. The transformation (5.3) is then equivalent to $\widetilde{\boldsymbol{W}} = \boldsymbol{Z}\widetilde{\boldsymbol{T}}$, where

$$\widetilde{\boldsymbol{T}} = \begin{bmatrix} \boldsymbol{T} & 0 & \ldots & 0 \\ 0 & \boldsymbol{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \boldsymbol{T} \end{bmatrix}$$

is a $nq \times nq$ block diagonal matrix, with blocks

$$\boldsymbol{T} = \begin{bmatrix} 1 & d_1 & d_2 & \ldots & d_{q-1} \\ 0 & c_1 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ldots & 0 & c_{q-2} & 0 \\ 0 & \ldots & & 0 & c_{q-1} \end{bmatrix}.$$

The equivalence between the transformation $\widetilde{\boldsymbol{W}} = \boldsymbol{Z}\widetilde{\boldsymbol{T}}$ and the transformation 5.3 can be shown in a way similar to that used in the case of coding transformations of design matrices of fixed effects (section 5.1).

It is easy to show that, if $\boldsymbol{T}_1$ and $\boldsymbol{T}_2$ are coding transformations then $\boldsymbol{T}_1\boldsymbol{T}_2$ is also a coding transformation.

5.3 Invariant fixed effects models

In fixed effects regression models that include polynomial and interaction terms, the model is invariant under coding transformations if and only if it is *well-formulated*. This result was proved in Peixoto (1990), the main reference of the material of this section. By definition, in a well-formulated model,

(i) if a variable X^n is included in the model, then all terms of order less than n are also included in the model;

(ii) if an interaction term $X_1^n X_2^m$ is included in the model, then all terms $X_1^i X_2^j, 0 \leq i \leq n, 0 \leq j \leq m$ must also remain in the model.

When a model is invariant under coding transformations, the choice of the covariate origin does not affect measures of goodness of fit such as the coefficient of correlation R^2 or the mean square error. In that case, if $\boldsymbol{T}$ is a coding transformation and $\boldsymbol{X}_T = \boldsymbol{X}\boldsymbol{T}$ is the transformed design matrix, then there is a linear transformation relating the two sets of estimates, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, of the following models:

$$\boldsymbol{y} = \boldsymbol{X}_T\boldsymbol{\gamma} + \boldsymbol{e} \tag{5.4}$$

and

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{5.5}$$

5.4 Invariant mixed effects models

Morrel et al. (1997) have shown that in mixed effects models, the fixed effects part of the model is coding invariant if it is well-formulated. The authors have also shown that if the random effects part of the model is not well-formulated then the model is not coding invariant. To illustrate the latter point, we consider the example presented in the introduction of this chapter. The linear predictor is given by,

$$\eta_i = \beta_0 + x_i\beta_1 + x_i b_{i1}, \tag{5.6}$$

where β_0 and β_1 are the fixed effects and $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{n1})$ is the vector of random slopes.

Assume that expression 5.6 yields the true model. Then, when we use a transformed variable $x_{Ti} = cx_i + d$, we obtain

$$\eta_i = (\beta_0 - \beta_1\frac{d}{c}) + x_{Ti}\frac{\beta_1}{c} - b_{i1}\frac{d}{c} + x_{Ti}\frac{b_{i1}}{c}. \tag{5.7}$$

In addition to the fixed intercept $\beta_0 - d\beta_1/c$, the fixed slope β_1/c, and the random slope b_{i1}/c, the linear predictor in expression 5.7 has now an extra random intercept, $-db_{i1}/c$. Therefore, the model without a random intercept does not correspond any more to the true model. It is thus not surprising that the statistical results obtained with the linear predictor 5.6 are different according to whether we use the untransformed or the transformed covariate (Example 1).

The same reasoning can be applied to random terms of higher order than the slope.

Consequently, to ensure coding invariance it is usually necessary to include all random terms of order inferior to the highest order included in the random effects part of the model. In other words, it is necessary that the random effects part of the model is well-formulated.

On the other hand, a well-formulated random effects model is not necessarily coding invariant, as we now show. Consider as part of the true model the well-formulated linear predictor

$$\eta_i = \beta_0 + x_i\beta_1 + b_{i0} + x_i b_{i1}, \tag{5.8}$$

where the random intercept b_{i0} and the random slope b_{i1} are independent. Using again the coding transformation $x_{Ti} = cx_i + d$, we obtain

$$\eta_i = (\beta_0 - \beta_1\frac{d}{c}) + x_{Ti}\frac{\beta_1}{c} + (b_{i0} - b_{i1}\frac{d}{c}) + x_{Ti}\frac{b_{i1}}{c}. \tag{5.9}$$

There is one main difference between the linear predictors in expressions 5.8 and 5.9: in 5.9, the random intercept $b_{i0} - b_{i1}\frac{d}{c}$ is no longer independent of the random slope $\frac{b_{i1}}{c}$, as these two effects share a common term b_{i1}. Hence, the model assuming independence between the random intercept and the random slope is no longer the true model if the transformed covariate is used. This example shows the importance of allowing for non-zero correlations between the random intercept and the random slope. For instance, if the random effects are multivariate Gaussian and the covariance matrix is diagonal, a transformation of the covariates leads to different results.

Example 2

The impact of the lack of coding invariance caused by a postulated independence between the random intercept and the random slope is illustrated in figure 5.2, which displays the results obtained with the same data set and the same distribution for the response variables used in the previous example 1, but with the linear predictor 5.8. The two random effects are independent and their distributions are intrinsic autoregressions. The maps 5.2(a) and 5.2(b) correspond, respectively, to the use of a non-centred and centred covariate and show that different spatial patterns of the random slope are obtained in each case. In addition, the scatterplots (c) and (d) in figure 5.2 show the estimates obtained when using the non-centred covariates versus those obtained when using the centred covariates. Although a linear relation between the random slopes is apparent in comparing expressions 5.8 and 5.9, the independence assumption in the model 5.8 and the lack of it in the model 5.9 lead to a non-linear relation between the random slopes (figure 5.2(d)).

We now define *parametric family of distributions* as a set $\{p(.|\theta), \theta \in \Theta\}$ specified by a probability density function $p(.|\theta)$ and a parametric space Θ of permissible values of θ. For example, the distributions $\mathrm{N}(0, \sigma^2)$ and $\mathrm{N}(0, 2\sigma^2)$ belong to the same parametric family if $\sigma^2 \in (0, +\infty)$. On the contrary, if $\sigma^2 \in (1, +\infty)$ they do not belong to the same parametric family because the permissible values for the variance are not the same for both densities.

Expression 5.9 indicates that the distributions of the random effects have to be chosen carefully. First, the postulated distribution for the random slope has to be closed under scaling. In most applications, the random slope is assumed

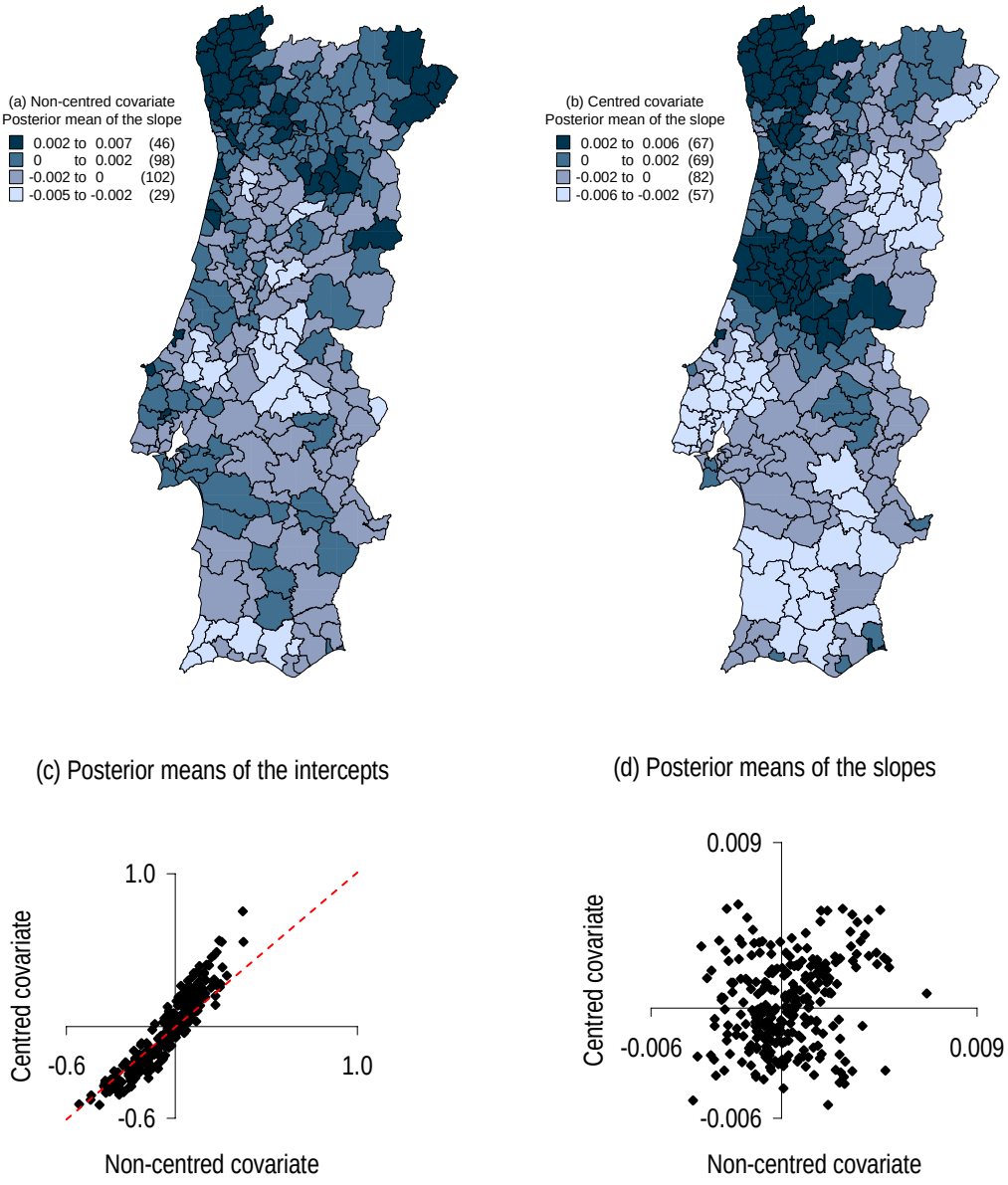


Figure 5.2: Example 2: Maps of the posterior means of the areal exposure coefficients measuring the association between stomach cancer mortality and: (a) spending power; (b) spending power centred about the mean. Scatterplots of the posterior means of (c) the random intercepts and (d) the random slopes, obtained with the non-centred covariate versus the centred covariate.

to be Gaussian distributed, a distribution which is closed under scaling. If other distributions are chosen, the random effects model may not be coding invariant. For instance, the Gamma distribution with unit scale parameter is a case in point that will lead to a model which is not coding invariant, as the scale parameter does not remain equal to 1 under scaling. Secondly, the random intercept distribution and the convolution of the distribution of the random intercept with the distribution of the random slope have to belong to the same parametric family.

Example 3

By way of illustration, we consider two models: in model 1, the random slope is defined as a convolution of the random intercept with another random variable, whereas in model 2, the random intercept is defined as a convolution of the random slope with another random variable. To put it formally,

Model 1

$$\begin{aligned} \boldsymbol{b}_{\bullet 0} &\sim \mathrm{IAR}(\sigma_0^2) \\ \boldsymbol{b}_{\bullet 1} &= \boldsymbol{\gamma} + \phi \boldsymbol{b}_{\bullet 0} \qquad (5.10) \\ \boldsymbol{\gamma} &\sim \mathrm{MVN}(\mathbf{0}, \sigma_1^2 \boldsymbol{I}) \qquad (5.11) \end{aligned}$$

Model 2

$$\begin{aligned} \boldsymbol{b}_{\bullet 0} &= \boldsymbol{\gamma} + \phi \boldsymbol{b}_{\bullet 1} \\ \boldsymbol{b}_{\bullet 1} &\sim \mathrm{IAR}(\sigma_1^2) \qquad (5.12) \\ \boldsymbol{\gamma} &\sim \mathrm{MVN}(\mathbf{0}, \sigma_0^2 \boldsymbol{I}) \qquad (5.13) \end{aligned}$$

where IAR stands for intrinsic autoregression and MVN for multivariate Gaussian distribution.

The posterior estimates obtained with model 1 are shown in figure 5.3. Although the differences between the two maps are not as striking as in the examples 1 and 2, the differences are still noticeable. These results support our previous conclusions, which indicated that the first model is not coding invariant.

Turning now to model 2, similar maps of the estimates of the random slopes are obtained when using centred covariates or not (figure 5.4). In addition, a linear relation can be observed between the posterior means of the random slopes obtained in the models with a centred and non-centred covariate.

5.4.1 Results for q covariates

In the model given by expressions 2.8 and 4.4, a coding transformation $\widetilde{\boldsymbol{T}}$ of the random effects design matrix $\boldsymbol{Z}$ will lead to:

$$\boldsymbol{\eta} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_T\boldsymbol{b}_T, \tag{5.14}$$

where $\boldsymbol{Z}_T = \boldsymbol{Z}\widetilde{\boldsymbol{T}}$ and $\boldsymbol{b}_T = \widetilde{\boldsymbol{T}}^{-1}\boldsymbol{b}$.

The matrix $\widetilde{\boldsymbol{T}}$ is block diagonal (section 5.2), and its inverse is,

$$\widetilde{\boldsymbol{T}}^{-1} = \begin{bmatrix} \boldsymbol{T}^{-1} & 0 & \dots & 0 \\ 0 & \boldsymbol{T}^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{T}^{-1} \end{bmatrix}.$$

Since $\boldsymbol{T}$ is upper-triangular, its inverse is also upper-triangular. It is indeed easy

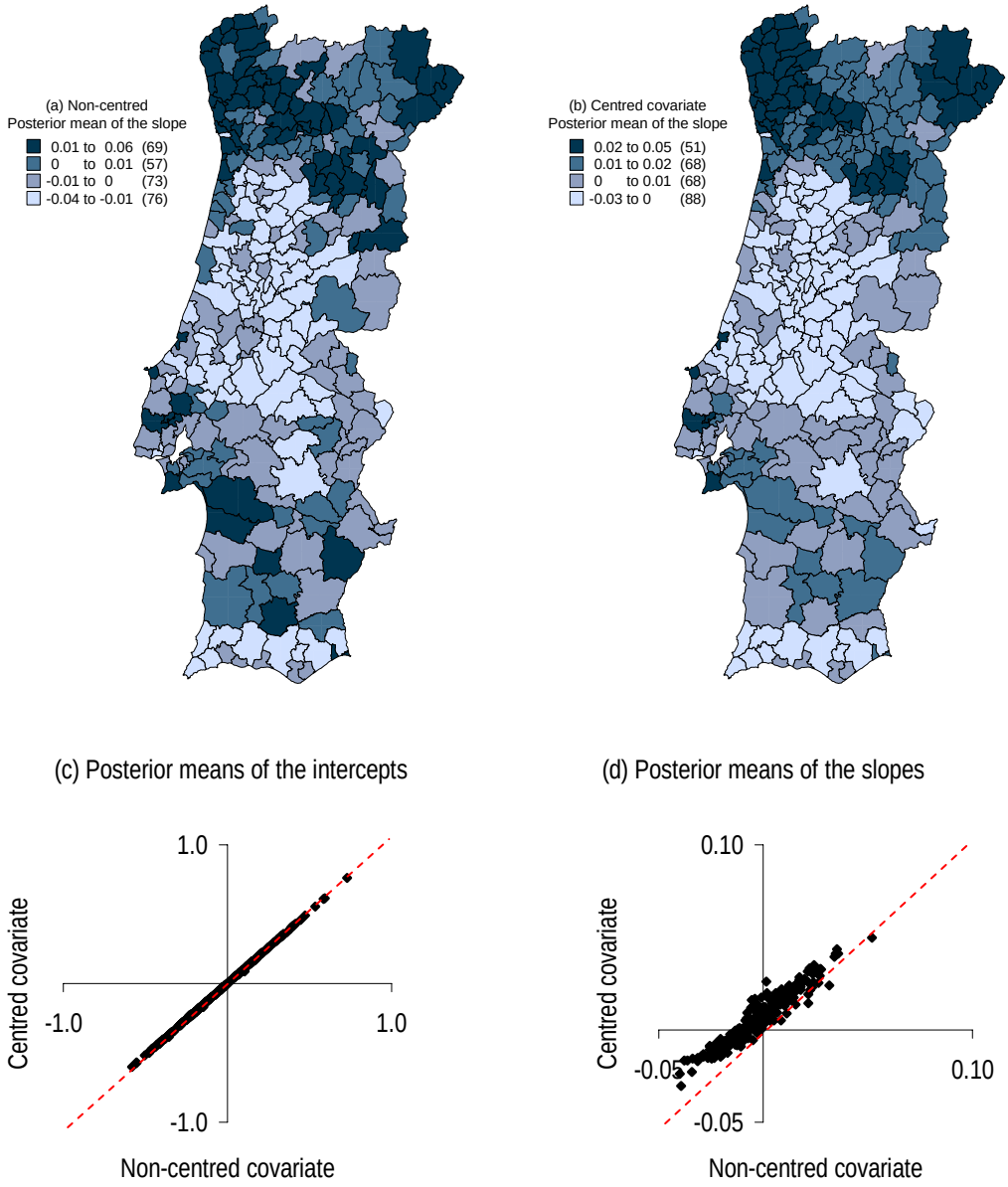


Figure 5.3: Example 3, model 1: Maps of the posterior means of the areal exposure coefficients measuring the association between stomach cancer mortality and: (a) spending power; (b) spending power centred about the mean. Scatterplots of the posterior means of (c) the random intercepts and (d) the random slopes, obtained with the non-centred covariate versus the centred covariate.

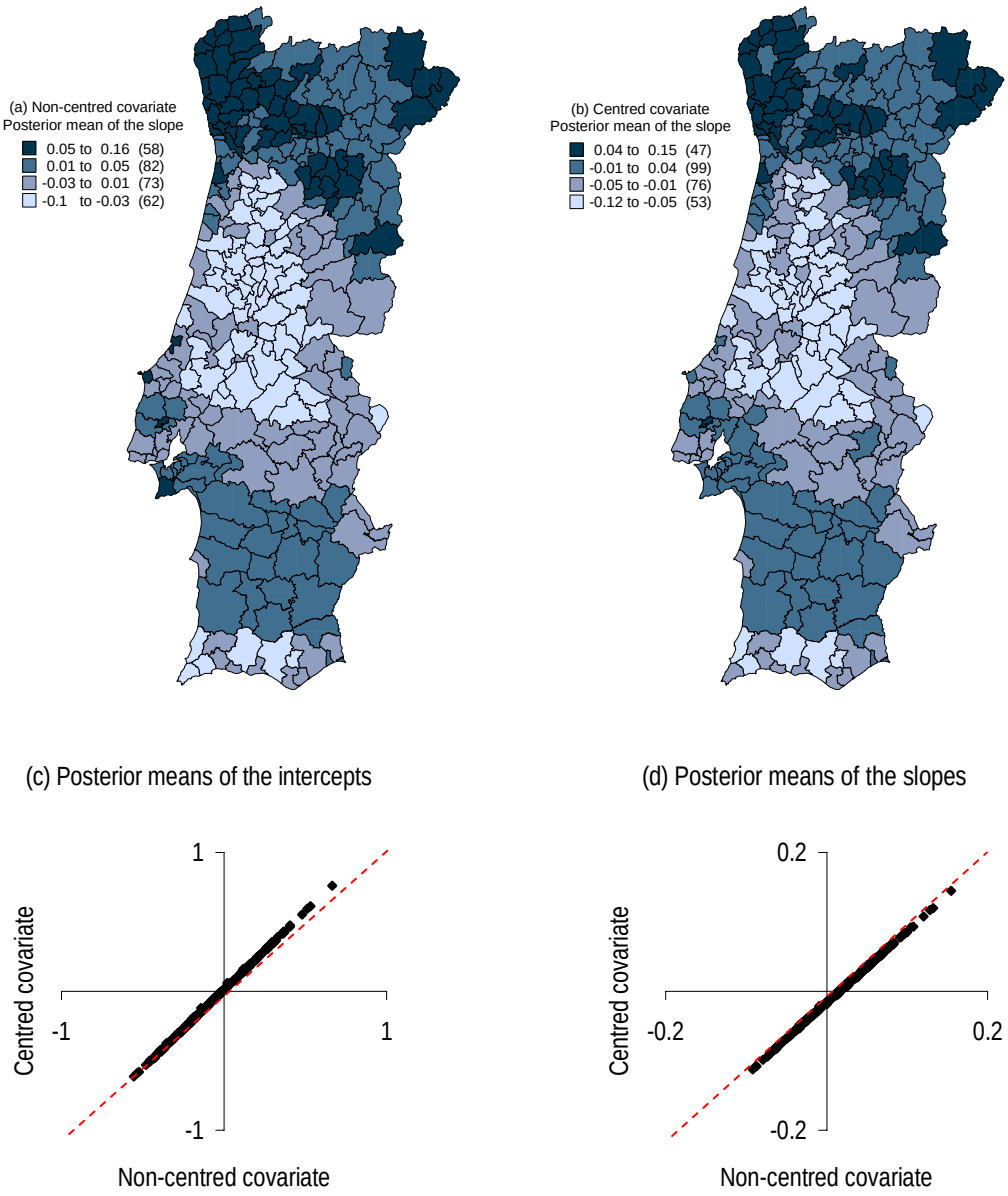


Figure 5.4: Example 3, model 2: Maps of the posterior means of the areal exposure coefficients measuring the association between stomach cancer mortality and: (a) spending power; (b) spending power centred about the mean. Scatterplots of the posterior means of (c) the random intercepts and (d) the random slopes, obtained with the non-centred covariate versus the centred covariate.

to see that,

$$\boldsymbol{T}^{-1} = \begin{bmatrix} 1 & -\frac{d_1}{c_1} & -\frac{d_2}{c_2} & \dots & -\frac{d_{q-1}}{c_{q-1}} \\ 0 & \frac{1}{c_1} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \frac{1}{c_{q-2}} & 0 \\ 0 & \dots & & 0 & \frac{1}{c_{q-1}} \end{bmatrix}.$$

Consequently, $\boldsymbol{b}_T = \widetilde{\boldsymbol{T}}^{-1}\boldsymbol{b}$ is given by,

$$\boldsymbol{b}_T = \begin{bmatrix} \boldsymbol{T}^{-1}\boldsymbol{b}_{1\bullet} & 0 & \dots & 0 \\ 0 & \boldsymbol{T}^{-1}\boldsymbol{b}_{2\bullet} & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \boldsymbol{T}^{-1}\boldsymbol{b}_{n\bullet} \end{bmatrix},$$

where $\boldsymbol{b}_{i\bullet} = (b_{i0}, \dots, b_{i,q-1})$ are the random effects associated with unit i. We use $\boldsymbol{b}_{\bullet j} = (b_{1j}, \dots, b_{nj}), j = 0, \dots, q-1$, to refer to the random effects associated with covariate j. Then,

$$\boldsymbol{b}_T = \begin{bmatrix} b_{10} - \sum_{s=1}^{q-1} b_{1s}\frac{d_s}{c_s} \\ \frac{b_{11}}{c_1} \\ \vdots \\ \frac{b_{1,q-1}}{c_{q-1}} \\ \vdots \\ \vdots \\ b_{n0} - \sum_{s=1}^{q-1} b_{ns}\frac{d_s}{c_s} \\ \frac{b_{n1}}{c_1} \\ \vdots \\ \frac{b_{n,q-1}}{c_{q-1}} \end{bmatrix}$$

or, equivalently,

$$\boldsymbol{b}_{T_{\bullet 0}} = \boldsymbol{b}_{\bullet 0} - \sum_{s=1}^{q-1} \boldsymbol{b}_{\bullet s} \frac{d_s}{c_s}, \tag{5.15}$$

$$\boldsymbol{b}_{T_{\bullet 1}} = \boldsymbol{b}_{\bullet 1} \frac{1}{c_1},$$

$$\ldots \tag{5.16}$$

$$\boldsymbol{b}_{T_{\bullet q-1}} = \boldsymbol{b}_{\bullet q-1} \frac{1}{c_{q-1}}.$$

Therefore, a similar reasoning to the one used in section 5.4 leads us to conclude that the model given by expressions 2.8 and 4.4 is coding invariant only if all the following conditions are satisfied:

1. the model contains an intercept term $\boldsymbol{b}_{T_{\bullet 0}}$;

2. the marginal distributions of the random intercept $\boldsymbol{b}_{\bullet 0}$ and the convolution of this distribution with the distributions of the random slopes $\boldsymbol{b}_{\bullet j}, j = 1, \ldots, q-1$, must belong to the same parametric family;

3. the marginal distributions of all the random slopes $\boldsymbol{b}_{\bullet j}, j = 1, \ldots, q-1$ are closed under scaling;

4. the random intercept $\boldsymbol{b}_{\bullet 0}$ is not assumed independent of the random slopes $\boldsymbol{b}_{\bullet j}, j = 1, \ldots, q-1$.

5.4.2 Necessary conditions for coding invariance

To sum up, the results discussed and shown in this chapter can be summarized in the following lemmas:

Lemma 5.4.1 *If the model given by expressions 2.8 and 4.4 is not well-formulated then the model is not coding invariant.*

Lemma 5.4.2 *In the model given by expressions 2.8 and 4.4, if the random intercept is assumed to be independent of the random slope then the model is not coding invariant.*

Lemma 5.4.3 *In the model given by expressions 2.8 and 4.4, if the distribution of the random intercept can not be expressed as the convolution of the distribution of the random slopes with another distribution then the model is not coding invariant.*

Lemma 5.4.4 *In the model given by expressions 2.8 and 4.4, if the distribution of the random slope is not closed under scaling then the model is not coding invariant.*

We will take into account these results when developing in the next chapters models for analysing varying exposure coefficients, in applications in which coding transformations are of interest.

For now, we show in figure 5.5 the estimates obtained with the model developed in chapter 8, which does not violate the four lemmas above and assumes in particular that the random effects may be correlated. The lack of coding invariance seems not to be present. The two maps of slope estimates are similar and the relation between the two sets of slope estimates is approximately linear. The small differences between the two maps in figure 5.5 are possibly due to the choice of priors and initial values. The priors and initial values were the same for the model with the non-centred and centred covariate, but they correspond to different prior information in each of the models. For example, a prior mode at zero for the exposure coefficient in the non-centred covariate model may not be equivalent to a prior mode at the same value and for the same parameter in the

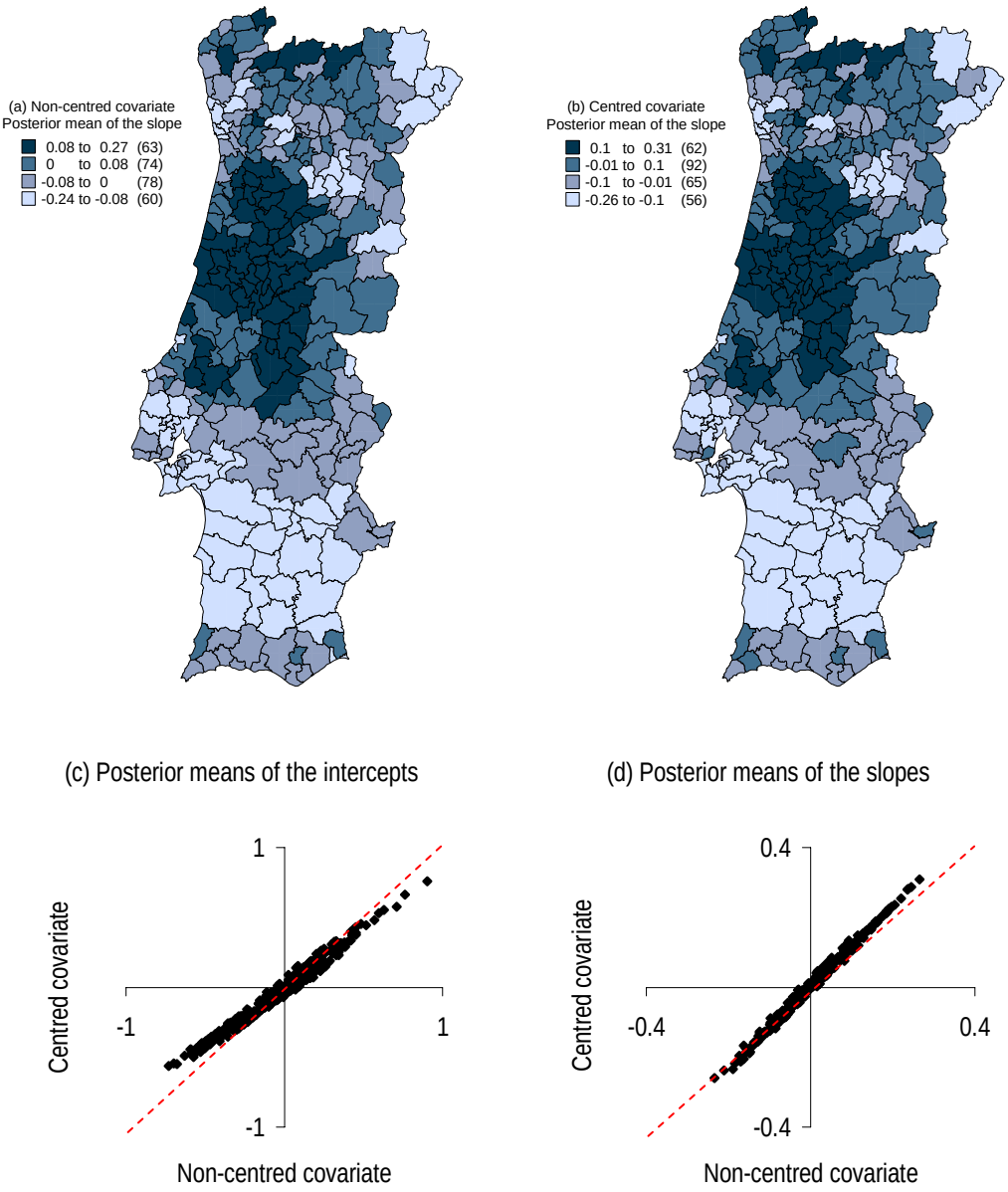


Figure 5.5: Maps of the posterior means, obtained from the model developed in chapter 8, of the areal exposure coefficients measuring the association between stomach cancer mortality and: (a) spending power; (b) spending power centred about the mean. Scatterplots of the posterior means of (c) the random intercepts and (d) the random slopes, obtained with the non-centred covariate versus the centred covariate.

centred covariate model. In addition, the differences between the two maps in figure 5.5 are also partly due to the sampling error in the MCMC simulation. To assess the importance of the sampling error, we used the Monte Carlo standard error of the mean, also known as Monte Carlo error, which is as an output of the Winbugs 1.4 software (Roberts, 1996; Spiegelhalter et al., 2003). The average Monte Carlo error for the slopes was about 0.004 in both models with the centred and non-centred covariate. The average difference in the posterior means of the slopes between the models with the centred and non-centred covariate was 0.02.

Coding invariance has usually been defined within a frequentist approach. In some problems, there may be reasons to believe that the choice of units should not affect the problem (for example, in some problems, it should not matter if we use centimetres or inches, the results obtained should be the same). The Bayesian approach has attempted to deal with these type of issues by developing invariant priors, such as location and scale invariant priors (Berger, 1985; Milne, 1983). However, if the model is not coding invariant, the posterior distributions obtained with the use of different units may not be equivalent even if appropriate invariant priors are used.

In a Bayesian perspective, the model should reflect the beliefs about the problem. If there are reasons to believe that the choice of units should not change the problem, then the model used should be coding invariant.

For coding invariant models, the likelihood function remains invariant under coding transformations, i.e.:

$$f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{Z}) \;\; = \;\; f(\boldsymbol{y}|\boldsymbol{b}_T, \boldsymbol{Z}_T),$$

where $\boldsymbol{b}_T$ and $\boldsymbol{Z}_T$ are defined as in equation (5.14).

Denoting by $\pi_B(.)$ and $\pi_B(.|\boldsymbol{y})$ the proper prior and posterior distributions, respectively, of $\boldsymbol{b}$, it is easy to show by a transformation of variables that $\pi_{B_T}(\boldsymbol{b}_T) = |\widetilde{\boldsymbol{T}}|\pi_B(\boldsymbol{b})_{\boldsymbol{b}=\widetilde{\boldsymbol{T}}\boldsymbol{b}_T}$ and $\pi_{B_T}(\boldsymbol{b}_T|\boldsymbol{y}) = |\widetilde{\boldsymbol{T}}|\pi_B(\boldsymbol{b}|\boldsymbol{y})_{\boldsymbol{b}=\widetilde{\boldsymbol{T}}\boldsymbol{b}_T}$.

Consequently, if the prior distribution $\pi_T(.)$, of the parameters $\boldsymbol{b}_T$ of the model using the transformed covariate $\boldsymbol{Z}_T$, is such that $\pi_T(\boldsymbol{b}_T) = |\widetilde{\boldsymbol{T}}|\pi_B(\boldsymbol{b})_{\boldsymbol{b}=\widetilde{\boldsymbol{T}}\boldsymbol{b}_T}$, i.e. the prior information is the same in the models using the original and the transformed covariate, then the posterior distributions of the original and transformed parameters, $\boldsymbol{b}$ and $\boldsymbol{b}_T$, will be equivalent, $\pi_T(\boldsymbol{b}_T|\boldsymbol{y}) = |\widetilde{\boldsymbol{T}}|\pi_B(\boldsymbol{b}|\boldsymbol{y})_{\boldsymbol{b}=\widetilde{\boldsymbol{T}}\boldsymbol{b}_T}$. In particular, there will be a correspondence between point posterior estimates: $E_T(\boldsymbol{b}_T|\boldsymbol{y}) = |\widetilde{\boldsymbol{T}}|^{-1}E_B(\boldsymbol{b}|\boldsymbol{y})$ and, if $\boldsymbol{b}^*$ is the posterior mode of the original parameters, then $\widetilde{\boldsymbol{T}}^{-1}\boldsymbol{b}^*$ will be the posterior mode of the transformed parameters.

Chapter 6

Regression mixtures

Mixture distributions are typically used for modelling data in which each observation is assumed to have arisen from one of a number of different groups, each group being modelled by a density from the same family. In this thesis, we will use mixture models in the context of what is usually called *cluster analysis*, i.e. identification of different groups in a data set.

We want to take into account the scenario envisaged in chapter 3, which indicated that groups of countries with similar *H. pylori* virulence and human susceptibility may exist. We assume that the population in each group of countries is characterised by a distinct response to *H. pylori* infection and that there are no distinguishable sub-populations. Regression mixtures, like the ones described in Congdon (2001), p. 215, can be used to identify these groups.

In view of the study we present in chapter 9, we consider the problem in the context of normal distributions, but it can be adapted to other distributions. Also, we assume that the variability across areas can be explained by only one factor, whose exposure coefficient is allowed to vary across groups.

Formally, the regression mixture model we consider is

$$Y_i \sim \sum_{m=1}^{M} p_m \mathrm{N}(\mu_{im}, \sigma^2), \tag{6.1}$$

$$\mu_{im} = \beta_0 + b_{m1} z_i,$$

where the variables z_i and Y_i correspond to, respectively, the risk factor and the log disease rate, in area $i, i = 1, \ldots, n$. The probabilities $p_m, m = 1, \ldots, M$ are the mixing proportions which are constrained to be non-negative and sum to unity. The parameters $\{b_{m1}, m = 1, \ldots, M\}$ are the group specific parameters and reflect the various values of the exposure coefficient. For notational convenience, we denote the vector of random slopes as $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{M1})$. The parameter β_0 corresponds to the global baseline risk, and its value is assumed to be the same in all areas. We know from chapter 5 that this assumption leads to a model which is not coding invariant, but the assumption is justified in situations where the regression lines are expected to cross at a given value of the exposure and this value is known (e.g. the study in chapter 9). The variation in rates caused by unknown factors or sampling variation is believed to be accounted by the parameter σ^2.

The components of the mixture have a physical interpretation and the mixture model analysis will aim at:

1. Estimating the parameters of the individual components of the mixture for inference on the exposure coefficient;

2. Estimating the *classification probabilities* for the observed data points;

3. Allocating the data points into sets of "similar" exposure coefficient.

We will carry out statistical analyses with various fixed values of M.

Problems associated with the maximum likelihood approach to mixture models have been reported in the literature, mostly resulting from the fact that the likelihood is unbounded (Robert, 1996; Stephens, 1997; Marin et al., 2005). The Bayesian approach with weakly informative proper priors is a viable alternative. In this chapter, we present the implementation of regression mixtures within a Bayesian approach. Most of the material presented in this chapter relies on Robert (1996) and on well known results about mixture models. We adapted most of the results to the context of regression mixtures, and we propose an explanation for the trapping states we encountered in the Markov chain Monte Carlo (MCMC) simulation of regression mixtures.

6.1 The Bayesian approach

We write the probability density function of a mixture model for the vector of observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ as,

$$f(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} \sum_{m=1}^{M} p_m f_m(y_i|\theta_m), \tag{6.2}$$

where $\boldsymbol{p} = (p_1, \ldots, p_M)$ is the vector of the *mixing probabilities* and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_M)$ that of *component parameters*. For the particular case of model 6.1, $\theta_m = (\beta_0, b_{m1}, \sigma^2)$ and $f_m(y_i|\theta_m)$ is the probability density function of a normal distribution with mean $\beta_0 + b_{m1} z_i$ and variance σ^2. In a Bayesian framework, the unknowns $\boldsymbol{p}$ and $\boldsymbol{\theta}$ are regarded as random quantities, drawn from an appropriate prior distribution $\pi(\boldsymbol{p}, \boldsymbol{\theta})$. It is possible to show that the posterior distribution is

given by a sum of M^n terms:

$$\begin{aligned}\pi(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{p},\boldsymbol{\theta})\prod_{i=1}^{n}\sum_{m=1}^{M} p_m f_m(y_i|\theta_m) \\ &= \pi(\boldsymbol{p},\boldsymbol{\theta})\sum_{(m_1,\ldots,m_n)\in\{1,\ldots,M\}^n}\Big[\prod_{m=1}^{M}\big(p_m^{c_m}\prod_{i:m_i=m} f_m(y_i|\theta_m)\big)\Big],\end{aligned} \tag{6.3}$$

where $c_m = \#\{m_i : m_i = m, i = 1,\ldots,n\}$.

If $f_m(.|.)$ belongs to the exponential family (expression 2.8) and the prior distributions are such that the vector $\boldsymbol{p}$ and the θs are drawn independently, with a Dirichlet prior for $\boldsymbol{p}$,

$$\pi(\boldsymbol{p}) \propto p_1^{\alpha_1-1}\ldots p_M^{\alpha_M-1}$$

and the conjugate prior of $f(.|.)$ for θ_m,

$$\pi(\theta_m|\kappa,\varsigma) \propto e^{\frac{\theta_m\kappa-\varsigma h(\theta_m)}{\varphi}},$$

the equation 6.3 can then be simplified to

$$\pi(\boldsymbol{p},\boldsymbol{\theta}|\boldsymbol{y}) \propto \sum_{(m_1,\ldots,m_n)\in\{1,\ldots,M\}^n}\prod_{m=1}^{M} p_m^{\alpha_m+c_m-1}\pi(\theta_m|\kappa+c_m\bar{y}_m,\varsigma+c_m), \tag{6.4}$$

where $\bar{y}_m = \sum_{i:m_i=m} y_i/c_m$.

Although the calculation of each individual term in the sum of expression 6.4 is straightforward, Robert (1996) indicated that the computing time required to evaluate the M^n terms is prohibitive, even in problems of moderate size. A viable alternative is provided by MCMC methods. As acknowledged by several authors, the Gibbs sampler is particularly suited for mixture models in the case where M is considered known (Robert, 1996; Stephens, 1997).

6.2 The missing data formulation

Model (6.2) can be written in terms of n unobserved *allocation variables* $V_i, i = 1, \ldots, n$, by assuming that the observation y_i arises from the component m with probability

$$P(V_i = m|\boldsymbol{p}) = p_m.$$

Therefore, the conditional likelihood of observation $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$f(\boldsymbol{y}|\boldsymbol{p}, \boldsymbol{v}, \boldsymbol{\theta}) \quad \propto \quad \prod_{i=1}^{n} \prod_{m=1}^{M} (f(y_i|\theta_m))^{v_{im}}, \tag{6.5}$$

where $v_{im} = I(v_i = m)$. Integrating out the variables $\boldsymbol{V}$ leads to model (6.2).

In a Bayesian context, missing data are considered as a random quantity drawn from an appropriate prior distribution. Since the Vs are unknown, we regard them as missing data within the Bayesian context. The posterior distribution can thus be written as,

$$\pi(\boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{v}|\boldsymbol{y}) \quad \propto \quad \prod_{i=1}^{n} \prod_{m=1}^{M} (p_m f_m(y_i|\theta_m))^{v_{im}} \pi(\boldsymbol{p}, \boldsymbol{\theta}).$$

6.3 Prior distributions

Although conjugacy is not necessary for using the MCMC techniques, in mixture models it is common practice to choose the priors in such a way that it is relatively easy to determine the expression for the full conditional and posterior distributions. Conjugate, or approximately conjugate, priors are often chosen. We use the same strategy for the regression mixtures used in this thesis.

6.3.1 Prior distributions for the mixing probabilities

The usual choice for the vector of group proportions $\boldsymbol{p} = (p_1, \ldots, p_M)$ is a Dirichlet prior

$$\pi(\boldsymbol{p}|\alpha_1, \ldots, \alpha_M) \propto p_1^{\alpha_1 - 1} \ldots p_M^{\alpha_M - 1}. \tag{6.6}$$

When the parameters of the Dirichlet prior are equal, $\alpha_1 = \ldots = \alpha_M = \alpha$, the prior information is the same for all the ps. Choosing $\alpha = 1$ gives a uniform distribution over the space $p_1 + \ldots + p_k = 1$. However, we will see in section 6.4.3 that this value of α can create difficulties in the MCMC simulation, in which case larger values may be considered.

6.3.2 Prior distributions for the regression parameters

In the case of a mixture of Gaussian distributions, we can not use independent improper priors on the component-specific parameters of the mixed model because it leads to improper posteriors. (Stephens, 1997, p.12) provides a proof of this statement, which is summarized here for the reader's convenience. Consider values of the vector $\boldsymbol{v}$ that assign no observation to the first group. The corresponding vector of completed data, $(\boldsymbol{y}, \boldsymbol{v})$, thus contains no information about the parameter b_{11}, and if the $b_{11}, \ldots, b_{M1}$ are *a priori* independent, then the posterior distribution $\pi(b_{11}|\boldsymbol{y}, \boldsymbol{v})$ will be the same as the prior distribution $\pi(b_{11})$. If this latter distribution is improper, $\pi(b_{11}|\boldsymbol{y}, \boldsymbol{v})$ will also be improper, and hence, $\pi(b_{11}|\boldsymbol{y})$ will be improper because

$$\pi(b_{11}|\boldsymbol{y}) = \sum_{\boldsymbol{v} \in \{1, \ldots, M\}^n} \pi(b_{11}|\boldsymbol{y}, \boldsymbol{v}) P(\boldsymbol{v}|\boldsymbol{y}).$$

Among possible proper prior distributions, it is common to choose the natural conjugate for the univariate Gaussian distribution, in which the mean and

variance are not independent (Diebolt and Robert, 1994; Robert, 1996). Other authors considered "approximate conjugate" distributions, in which the mean and the variance are drawn independently (Richardson and Green, 1997; Stephens, 1997). This is the approach we follow here and in the analysis in chapter 9. Consequently, Gaussian prior distributions are chosen for the regression coefficients and a gamma prior is chosen for the precision, i.e. the inverse of the variance. Modelling the precision instead of the variance makes it easier to attribute larger weights to large values of the variance. The Gaussian distribution, adopted for the regression coefficients, is symmetrically distributed - a natural prior assumption for these parameters. All these prior distributions can be taken rather flat with an appropriate choice of hyperparameters. Flat distributions reflect better the fact that little or no information is available *a priori* on the regression parameters.

6.4 Implementation

6.4.1 Gibbs sampler for regression mixtures

The Gibbs sampler is a commonly used MCMC algorithm for generating sequences of samples from the joint probability distribution (Smith and Roberts, 1993). The samples are drawn iteratively from the full conditional distributions. The full conditional distributions for each parameter can be deduced from the joint posterior distribution,

$$\pi(\sigma^2, \beta_0, \boldsymbol{b}_{\bullet 1}, \boldsymbol{p}, \boldsymbol{v}|\boldsymbol{y}) \quad \propto \quad f(\boldsymbol{y}|\boldsymbol{v}, \sigma^2, \beta_0, \boldsymbol{b}_{\bullet 1})\pi(\boldsymbol{v}|\boldsymbol{p})\pi(\boldsymbol{p})\pi(\tau^2)\pi(\beta_0)\pi(\boldsymbol{b}_{\bullet 1}), \tag{6.7}$$

by ignoring the terms which do not depend on the parameter of interest on the right hand side of the expression 6.7. Explicit derivation of full conditional distributions in the context of mixture models can be found in Robert (1996).

6.4.2 Label switching

A well known result states that the parameters of mixture models are not identifiable because the likelihood is invariant under permutation of the group labels. Similarly, in regression mixtures, the value of the likelihood

$$L(\boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2|\boldsymbol{y}) \;=\; \prod_{i=1}^{n}[p_1 f(y_i|\beta_0, b_{11}, \sigma^2) + \ldots + p_M f(y_i|\beta_0, b_{M1}, \sigma^2)]$$

is the same for all permutations of the labels of $\boldsymbol{p}$ and $\boldsymbol{b}_{\bullet 1}$, i.e. for every permutation $(\nu(1), \ldots, \nu(M))$ of $1, \ldots, M$,

$$L(\boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2|\boldsymbol{y}) \;=\; L(\boldsymbol{p}_\nu, \beta_0, \boldsymbol{b}_{\bullet 1\nu}, \sigma^2|\boldsymbol{y}) \tag{6.8}$$

where $\boldsymbol{p}_\nu = (p_{\nu(1)}, \ldots, p_{\nu(M)})$ and $\boldsymbol{b}_{\bullet 1\nu} = (\beta_{\nu(1)}, \ldots, \beta_{\nu(M)})$. In a Bayesian framework, the posterior distribution will be similarly invariant whenever the prior distribution of the parameters is also invariant under permutations of the parameters. Consequently, the posterior distribution will have $M!$ modes corresponding to $M!$ copies of the "genuine" mode.

It is thus not surprising that the applications of MCMC techniques to mixture models have often reported that the simulated chain moves between the different modes, causing the classification probabilities and the component-specific parameters to be the same for all components of the mixture. This problem has been referred to in the literature as *label-switching*.

Attempts to deal with the problem of label-switching have centred around the use of identifiability constraints on the parameter space, such as $p_1 < \ldots < p_M$

or $b_{11} < \ldots < b_{M1}$. These identifiability constraints impose an order on the parameters of expression 6.8, thus excluding permutations and allowing us to obtain sensible estimates on the individual components of the mixture. As we aim at studying the variation in the exposure coefficient, we opt for a constraint on the parameters $b_{11}, \ldots, b_{M1}$ because we expect to find groups with different slopes, while we have no reason to believe that the sizes of the groups, which are proportional to the ps, are different from each other. The constraint on the slopes is introduced with a reparametrization

$$b_{m+1,1} \quad = \quad b_{m1} + \delta_m, \delta_m > 0, m = 1, \ldots, M-1, \tag{6.9}$$

where the parameters $\delta_1, \ldots, \delta_{M-1}$ are given Gaussian priors, truncated to positive values.

6.4.3 Trapping states

Some previous analysis of normal mixture distributions using Gibbs sampling reported problems with "trapping states", i.e. states in which the chain would require an enormous number of iterations to escape from the state. In extreme cases, the probability of escape may be below the minimal precision of the computer and the trapping state is truly absorbing. These trapping states appear to be due to allowing the allocation of very few observations to one of the components of the mixture (Robert, 1996; Stephens, 1997), which causes the likelihood to tend to infinity as the mean of this component is set to be within the range of the observed data points and the variance of this component tends to zero. This problem has been avoided by either reparametrizing the Gaussian parameters (Robert, 1996, p.453-4) or constraining the component variances to be equal (Stephens, 1997, p.24). In the study presented in chapter 9, the component

variances are assumed to be equal and the trapping state problem just described was not observed.

Other authors indicated that there is a danger that, at some iteration, all the data will go to one component of the mixture, this state being difficult to escape from (Spiegelhalter et al., 2003). In our view, this problem seems to be due to two reasons: first, the use of priors for $\boldsymbol{p}$ that give a non-null probability to setting one of the p_i equal to 1 and the others to zero, so that the Markov chain is eventually allowed to enter this state at some point in the simulation; secondly, the use of the Dirichlet distribution as the prior for $\boldsymbol{p}$ (section 6.3.1), with all parameters $\alpha_m, m = 1, \ldots, M$, equal to 1. As a result, the full conditional distribution for $\boldsymbol{p}$ is given by (Robert, 1996),

$$\text{Dirichlet}(1 + n_1, 1 + n_2, \ldots, 1 + n_M),$$

where n_i is the number of observations allocated to group $i, i = 1, \ldots, M$. We now show that this choice of parameters is a cause of the absorbing state problem just described.

Assume, without loss of generality, that all the observations are allocated to group 1 at some point in the simulation, i.e. $n_m = 0$ for all $m = 2, \ldots, M$. The full conditional distribution of $\boldsymbol{p}$ is then proportional to p_1^n and does not depend on any of the other p_m (n being the number of observations). This distribution has a point of maximum at $p_1 = 1$ and even for moderate values of n the probability of sampling values of p_1 which are not near the value one is very small. Thus the chain will tend to stay in the same state. The state will be more absorbing the larger the value of n.

We propose avoiding this problem by setting the values of the hyperparameters $\alpha_1, \ldots, \alpha_M$ to be larger than or equal to two. When all observations are

attributed to the same component, the chain will still favour large values of p_1, but values near one will have very low probabilities because the full conditional of $\boldsymbol{p}$, which is proportional to $p_1^{\alpha_1+n-1}p_2^{\alpha_2-1}\ldots p_M^{\alpha_M-1}$ (Robert, 1996), tends to zero as any of the p_m tends to zero. The larger the value of the αs, the easier it will be for the chain to avoid the absorbing state.

Values of the $\alpha_m, m = 1, \ldots, M$, larger than or equal to two yield informative priors which attribute small weight to values of the p_ms near zero or one. While large values of the αs lead to highly informative priors, the choice $\alpha_m = 2, m = 1, \ldots, M$ provides the less informative prior and may be a reasonable choice in some cases. For these values of $\alpha_m = 2, m = 1, \ldots, M$, it is easy to show that the Dirichlet probability density function reaches a maximum at $p_1 = p_2 = \ldots = p_M = 1/M$. At all other points, the values of the Dirichlet probability density function are lower. For example, for $M = 2$, $\pi(1/2, 1/2) = 1.5$ whereas $\pi(0.1, 0.9) = 0.54$. For $M = 3$, $\pi(1/3, 1/3, 1/3) = 4.44$ while $\pi(0.1, 0.1, 0.8) = 0.96$; and for $M = 4$, $\pi(1/4, 1/4, 1/4, 1/4) = 19.69$ but $\pi(0.1, 0.1, 0.1, 0.7) = 3.53$. In general, the prior will favour values of $\boldsymbol{p}$ which are nearer to the point $p_1 = p_2 = \ldots = p_M = 1/M$, thus favouring groups of similar size.

We note that values of α_m between 1 and 2 yield extremely large probabilities for values of the p_m near zero and 1, and are thus not suitable for our analysis.

6.5 Bayesian group allocation or probabilities of membership

The determination of the posterior probabilities of membership is of interest, to classify the observations with respect to the components of the mixture. One

possible strategy for classifying the observations is assigning the ith observation to that group m for which it has the highest posterior probability of belonging,

$$\hat{v_i} = \text{argmax}_m\{P(v_i = m|\boldsymbol{y}, M)\}. \tag{6.10}$$

According to Richardson and Green (1997), this strategy corresponds to the Bayes classification of an existing observation y_i using the usual *percentage correctly classified* loss function.

In the maximum likelihood approach, the derivation of the probabilities of membership, also called *classification probabilities*, is done after estimates of the model parameters are obtained:

$$P(V_i = m|y_i, \hat{\boldsymbol{p}}, \hat{\beta}_0, \hat{\boldsymbol{b}}_{\bullet 1}, \hat{\sigma^2}) \quad \propto \quad \hat{p}_m P(y_i|V_i = m, \hat{\boldsymbol{p}}, \hat{\beta}_0, \hat{\boldsymbol{b}}_{\bullet 1}, \hat{\sigma}^2)\,.$$

This approach has also been followed within an empirical Bayes framework (Schlattman and Böhning, 1993).

In a full Bayesian approach, it is legitimate to take into account the uncertainty associated with the model parameters, by considering the whole posterior distribution $\pi(\boldsymbol{p}, \beta_0, \hat{\boldsymbol{b}}_{\bullet 1}, \sigma^2|\boldsymbol{y})$. The classification probabilities for the observations y_i are then given by

$$\begin{aligned} P(V_i = m|\boldsymbol{y}) \quad &= \quad \int P(V_i = m|\boldsymbol{y}, \boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)\pi(\boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{p}\, d\beta_0\, d\boldsymbol{b}_{\bullet 1}\, d\sigma^2 \\ &= \quad \int \frac{p_m f(y_i|V_i = m, \boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)}{\sum_s p_s f(y_i|V_i = s, \boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)} \pi(\boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{p}\, d\beta_0\, d\boldsymbol{b}_{\bullet 1}\, d\sigma^2\,. \end{aligned} \tag{6.11}$$

Instead of monitoring the V_i, one can use the average of the conditional expectations $P(V_i = m|\boldsymbol{y}, \ldots, \sigma^2)$. This method is known as Rao-Blackwellization

(Gelfand and Smith, 1990; Robert, 2001). Approximations to the $P(V_i = m)$ can thus be obtained through the monitoring of the quantities

$$\frac{p_m f(y_i|V_i = m, \boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)}{\sum_s p_s f(y_i|V_i = s, \boldsymbol{p}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)}, \; m = 1, \ldots, M, \; i = 1, \ldots, n,$$

during the MCMC run.

Chapter 7

Geographical clustering model

In the model discussed in chapter 6, the groups are not necessarily formed by areas near in space. We now propose a model similar to the *clustering model* developed in Knorr-Held and Rasser (2000), but fixing the number of groups and using a Gaussian distribution instead of the Poisson. In order to avoid confusion with the common use of the word *clustering* in the statistical literature, we will refer to this model as the *geographical clustering model*. The model is a special case of mixture regressions, with spatial dependence implicitly introduced at the level of the allocation variables $\boldsymbol{V}$. In this sense, it is similar to the one described in Green and Richardson (2002). Markov random fields have also been used to model spatial partitions in the context of image analysis (Besag, 1986).

The geographical clustering model was previously used in the context of Bayesian model averaging, in which a smooth estimate of the risk surface is obtained from the posterior mean risk surface. Here, we will use it in the context of clustering, and we would thus like to calculate classification probabilities. Knorr-Held and Rasser (2000) implemented the model using reversible jump MCMC because the number of groups was unknown. Here, the Gibbs sampler is

applicable because the number of groups is fixed.

The model was developed because of its capacity in detecting discontinuities. For our study, the main advantage of this model is its capacity to find solutions which are consistent with the expected spatial structure of the groups of areas. The support space of the geographical clustering prior is restricted to configurations in which the areas within each group are geographically close in some determined sense. Naturally, restricting the support space of the prior distribution results in an identical restriction in the support space of the posterior distribution. Basically, the prior model assumes that the region under study can be divided into several *connected clusters* of constant virulence. Here and throughout this thesis, we use the term *connected cluster* to refer to a connected set of areas (section 7.1).

In this chapter, we start by generalizing the procedure provided by Knorr-Held and Rasser (2000) for forming clusters, i.e we define a general process of forming clusters on the basis of any distance measure. We then develop distance measures which ensure that all the clusters are connected sets. Consequently, we are able to extend the model of Knorr-Held and Rasser (2000) to a model that depends not only on the adjacency but also on the the Euclidean distance between the areas. We consider that the number of clusters is known and implement the model using the Winbugs software. We further develop an algorithm to calculate the *geographical distance*, which reflects the adjacency and the Euclidean distance between the areas. We also examine whether all configurations of connected clusters can be found by the geographical clustering model (representability), and whether they all have the same prior probability (equiprobability). Finally, we compare the performance of the model with that of the regression mixture model

discussed in chapter 6. In particular, we show that the clustering model provides better results than the mixture model when the variance of the observations is large.

7.1 Paths and connected groups of areas

Let S be a region formed by n areas, $S = \{1, \ldots, n\}$. For notational convenience, we use from now on the notation $i \sim j$ to state that two areas i and j are contiguous. Assume that all areas are connected, in other words, for any two areas $i, j \in S$ there is a *path* $\xi(i, j)$ through contiguous areas from i to j,

$$\xi(i,j) = \{i_0, i_1, \ldots, i_{p-1}, i_p \,|\, i_0 = i, i_p = j, i_0 \sim i_1 \sim \ldots \sim i_p\}. \qquad (7.1)$$

In what follows, a *path* will always refer to a sequence of contiguous areas, in the way it is formally defined in expression 7.1 above. The set of all paths between two areas i and j is denoted by $\Xi(i, j)$. We define a set of areas C as *connected* when all areas of C are connected by a path in C. Formally, a set of areas C is connected when for all $i, j \in C$, there is a path $\xi(i, j) \subset C$.

7.2 Defining clusters through a distance measure

This section generalizes the procedure developed by Knorr-Held and Rasser (2000) to construct clusters. While these authors defined their procedure on the basis of the number of borders between areas, we consider here any distance measure between areas. First, we review the well known definition of distance measure.

Definition 7.2.1 (Distance measure) *A distance measure $d : S \times S \to \mathbb{R}$ is a function that satisfies the following conditions:*

1. $d(i,j) \geq 0$, *and* $d(i,j) = 0$ *if and only if* $i = j$ *(positivity),*

2. $d(i,j) = d(j,i)$ *(symmetry),*

3. $d(i,j) \leq d(i,s) + d(s,j)$ *(triangle inequality).*

Equality in 3. holds if and only if s lies on the segment $\bar{ij}$, i.e. on the minimal path from i to j.

A set of M clusters is defined by a distance measure $d(.,.)$ and by M areas, $(g_1, g_2, \ldots, g_M) \in S^M$. According to the terminology used in Knorr-Held and Rasser (2000), the areas $(g_1, g_2, \ldots, g_M)$ are referred to as *cluster centres*. The distance measure $d(.,.)$ is used to allocate every area $i \in S$ to one and only one cluster: an area i is assigned to cluster C_m when it is closer to g_j than to any other cluster centre. Once the cluster centres $\boldsymbol{g}$ are chosen, the distance d is thus used to assign the remaining $n - M$ areas to one of the clusters.

Definition 7.2.2 (Cluster) *Let $\boldsymbol{g} = (g_1, \ldots, g_M) \in S^M$ be a set of cluster centres. The cluster $C_j, j = 1, \ldots, M$ is formed by all areas i which have minimal distance to the corresponding cluster centre g_j, i.e.*

$$d(i, g_j) \leq d(i, g_l),\ l \in \{1, \ldots, M\},\ l \neq j$$

and for which g_j has the smallest index position among all the cluster centres with minimal distance to area i,

$$j = \min\{l \,|\, d(i, g_l) = \min(d(i, g_1), \ldots, d(i, g_M))\}. \tag{7.2}$$

According to expression 7.2, the areas are assigned to the cluster centre with the smallest index position, among all the cluster centres with minimal distance to area i. Since some areas may be equally distant to two or more cluster centres, expression 7.2 ensures that the vector $\boldsymbol{g}$ uniquely defines a cluster configuration, i.e. an assignment of all regions to one and only one of the clusters. Therefore, the order of selection of the cluster centres matters in the definition of the cluster partition. The same cluster centres may lead do different partitions according to the order at which their were selected. For example, $(g_1, g_2) = (i, j)$ may not lead to the same cluster partition as $(g_1, g_2) = (j, i)$. In the latter, areas which are equally distant to areas i and j are allocated to the cluster of area j while in the former they are allocated to the cluster of area i.

For some distance measures, the attribution rule 7.2 may be irrelevant as it will be unlikely that one area is equally distant to two other areas. Distances reflecting Euclidean distances between geographical locations are a case in point (section 7.3.1).

Finally, we note that the prior probabilities will be defined in such a way that the vectors $\boldsymbol{g}$ are equally probable (section 7.5.1), so that each area has equal probability of being a cluster centre.

7.3 Distance measure ensuring connected clusters

The clusters defined in definition 7.2.2 are not necessarily connected. In other words, a cluster can be comprised by two or more non-connected sets of areas. For example, usage of the Euclidean distance between the centroids of each area

would not guarantee that the clusters are always connected.

By extending a procedure developed by Knorr-Held and Rasser (2000), we now propose a family of distance measures leading to clusters that are always connected.

Since a path is formed by sequences of adjacent areas, its length can be defined by a sum of distances between adjacent areas. Formally, define the length λ of a path $\xi(i,j) = (i_0, ..., i_p)$ between areas i and j as

$$\lambda(\xi(i,j)) = \sum_{l=0}^{p-1} d_{adj}(i_l, i_{l+1})\,. \tag{7.3}$$

where $d_{adj}(.,.)$ is a function defined on the space $S_A \subset S^2$ of all pairs of adjacent areas, and satisfying the first two conditions in definition 7.2.1. For example, $d_{adj}(.,.)$ can be the distance between the centroids of two adjacent areas or the difference between the socio-economic level of two adjacent areas, or simply equal to one for all pairs of adjacent areas. In what follows, we will refer to the function $d_{adj}(.,.)$ as the *adjacency-distance*.

We also define the *minimal path* between any two areas i and j as the path $\xi_{\min}(i,j)$ with minimal length, i.e. the path that satisfies

$$\lambda(\xi_{\min}(i,j)) = \min\{\, \lambda(\xi(i,j)) \,|\, \xi(i,j) \in \Xi(i,j)\}\,.$$

The minimal path is not necessarily unique, in other words, there may be more than one minimal path between two areas.

In what follows, (ξ_1, ξ_2) denotes *path concatenation*, i.e the joining of two paths that share a common extremity, e.g. $\xi_1 = (i_1, \ldots, i_{s-1}, i_s)$ and $\xi_2 = (i_s, i_{s+1}, \ldots, i_{l-1}, i_l)$. When they are concatenated they form a new path $(\xi_1, \xi_2) = (i_1, \ldots, i_{s-1}, i_s, i_{s+1}, \ldots, i_{l-1}, i_l)$.

We now define a distance measure $d(.,.)$ between any two areas $i, j \in S^2$ as the length of the minimal path(s) between them,

$$d(i,j) \;=\; \lambda(\xi_{\min}(i,j)). \tag{7.4}$$

It can be easily shown that $d(.,.)$ is a distance measure according to definition 7.2.1. Positivity and symmetry result directly from the properties of the adjacency-distance function $d_{adj}(.,.)$. The triangle inequality can be easily shown by realizing that $d(i,s) + d(s,j)$ is the length of a path between areas i and j passing through area s and is therefore larger than or equal to the length of the minimal path between areas i and j, i.e.

$$\begin{aligned} d(i,s) + d(s,j) &= \lambda(\xi_{\min}(i,s)) + \lambda(\xi_{\min}(s,j)) \\ &= \lambda((\xi_{\min}(i,s), \xi_{\min}(s,j))) \\ &\geq \lambda((\xi_{\min}(i,j))) = d(i,j). \end{aligned} \tag{7.5}$$

Lemma 7.3.1 *Equality holds in expression 7.5 if and only if the area s belongs to (one of) the minimal path(s) between areas i and j.*

Proof. We first prove that if s belongs to a minimal path, $\xi_{\min}(i,j) = (i, \ldots, s, \ldots, j)$, then equality holds in expression 7.5. Since $s \in \xi_{\min}(i,j)$, then $\xi_{\min}(i,j) = (\xi(i,s), \xi(s,j))$. The paths $\xi(i,s)$ and $\xi(s,j)$ are minimal paths; otherwise, $\xi_{\min}(i,j)$ would not be a minimal path because there would exist a smaller path $\xi_{\min}(i,j) = (\xi_{\min}(i,s), \xi_{\min}(s,j))$. Therefore, by definition, the following equations hold:

$$\begin{aligned} d(i,j) &= \lambda(\xi_{\min}(i,j)) \\ &= \lambda(\xi(i,s),\, \xi(s,j)) \\ &= \lambda(\xi(i,s)) + \lambda(\xi(s,j)) \\ &= d(i,s) + d(s,j). \end{aligned}$$

We now prove that if equality holds in expression 7.5, then s belongs to the minimal path. Note that $(\xi_{\min}(i,s), \xi_{\min}(s,j))$ is a minimal path from i to j because $\lambda(\xi_{\min}(i,s), \xi_{\min}(s,j)) = d(i,j)$. Therefore, s belongs to a minimal path from i to j and this completes our proof.

Adapting the proof of Knorr-Held and Rasser (2000), we now prove that if the distance function $d(.,.)$ in expression 7.4 is used with definition 7.2.2, the clusters are always connected.

Theorem 7.3.1 (Connected clusters) *Let S be a connected set of areas, $d(.,.)$ a distance measure defined by equation 7.4, and $C_1, \ldots, C_M$ a set of clusters defined by definition 7.2.2. Then each cluster is a connected set.*

Proof. Suppose there is a cluster C_m that breaks down into at least two parts that are not connected, i.e. it can be decomposed into two sets

$$C_m = C_{m_1} \cup C_{m_2}$$

such that

$$\forall_{i_1 \in C_{m_1}} \ \forall_{i_2 \in C_{m_2}} \ \forall_{\xi \in \Xi(i_1,i_2)} \ \exists_{s \in \xi} \ s \notin C_m. \tag{7.6}$$

Assume, without loss of generality, that the cluster centre $g_m \in C_{m_1}$. Let $i \in C_{m_2}$ and $\xi_{\min}(i, g_m)$ be a minimal path between areas i and g_m. According to expression 7.6, it is possible to choose an area s which belongs to the minimal path but not to the geographical cluster m, i.e.

$$s \in \xi(i, g_m), \ s \notin C_m.$$

The area s belongs thus to another geographical cluster which we denote by C_l.

Since s belongs to the minimal path from i to g_m, lemma 7.3.1 ensures that:

$$d(i, g_m) \quad = \quad d(i, s) + d(s, g_m).$$

But $s \in C_l$, which implies that

$$d(s, g_m) > d(s, g_l)$$

and leads to

$$d(i, g_m) \quad > \quad d(i, s) + d(s, g_l).$$

Finally, using the triangle inequality,

$$d(i, g_m) \quad > \quad d(i, g_l)$$

and therefore $i_1 \in C_l$, which contradicts our initial assumption, and completes the proof.

For simplicity, we assumed that the minimal distances were uniquely defined. The inclusion of the rule of attribution (7.2) to the geographical cluster with lower order does not invalidate the proof.

7.3.1 Examples of distance measures

Boundary distance

As an example of a distance measure which produces connected clusters, we describe the distance function used in Knorr-Held and Rasser (2000). The distance $d(i_1, i_2)$ between two regions i_1 and i_2 is defined as the minimal number of boundaries that have to be crossed to move from area i_1 to area i_2. According to expressions 7.3 and 7.4, this distance can be obtained by defining the adjacency-distance as $d_{adj}(i, j) = 1$, where $i \sim j$. Here and throughout, we will call the corresponding distance measure d the *boundary distance*.

The boundary distance can be computed from the well known adjacency matrix (Knorr-Held and Rasser, 2000), $\boldsymbol{W} = [w_{ij}]_{i,j=1,\ldots,n}$, whose entries are $w_{ij} = 1$ if the areas i and j are adjacent and $w_{ij} = 0$ otherwise. The boundary distance $d(i_1, i_2)$ corresponds to the minimum value k that yields a positive value in the entry (i_1, i_2) of the k-th power of the neighbour matrix, W^k (see appendix C for the S-PLUS code to calculate the matrix of boundary distances from the adjacency matrix). To see why this procedure does provide the boundary distance, notice that the (i, j) entry of the matrix $\boldsymbol{W}^k$ is given by,

$$\sum_{s_1} \ldots \sum_{s_k} w_{is_1} w_{s_1 s_2} \ldots w_{s_k j},$$

where w_{ij} are the entries of the adjacency matrix $\boldsymbol{W}$. If this sum is equal to zero, then $\forall_{s_1,\ldots,s_k} w_{is_1} w_{s_1 s_2} \ldots w_{s_k j} = 0$, i.e. there are no paths from i to j in k steps. On the other hand, if there is a sequence $s_1 \ldots s_k$ which yields a non-zero product $w_{is_1} w_{s_1 s_2} \ldots w_{s_k j}$, then this sequence defines a path between i and j, and the sum above will be positive. By taking the minimum value of k that yields the sum above positive, we obtain the length of the minimum path (i.e. the minimum number of boundaries that have to be crossed to go from an area to the other).

Geographical distance

Another adjacency-distance that may be of interest is the one defined by $d_{adj}(i, j) = \mathrm{dist}(i, j)$, where $i \sim j$, and $\mathrm{dist}(i, j)$ is the usual Euclidean distance between the centroids of areas i and j. We will call the resulting distance measure d the *geographical distance*. Less formally, the geographical distance is the length of the path with minimal Euclidean length among all paths between areas i and j.

The geographical distance is calculated in a less straightforward way than the

boundary distance. First, we look for the shortest path among paths that cross the same number of boundaries. With this aim, we create a matrix with the distances between all pairs of adjacent areas:

$$\boldsymbol{D}(i,j) = \begin{cases} d(i,j) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}.$$

From the matrix $\boldsymbol{D}$, we define $\boldsymbol{D}_k$ by recurrence,

$$\boldsymbol{D}_1(i,j) = \boldsymbol{D}(i,j),$$

$$\boldsymbol{D}_k(i,j) = \begin{cases} \Delta_{i,j}^k & \text{if } \boldsymbol{D}^k(i,j) > 0 \\ 0 & \text{otherwise} \end{cases}, k = 2,3,4,\ldots,$$

where

$$\Delta_{i,j}^k = \min_{\{s:\boldsymbol{D}_{k-1}(i,s)>0, \boldsymbol{D}(s,j)>0, s=1,\ldots,n\}}(\boldsymbol{D}_{k-1}(i,s) + \boldsymbol{D}(s,j))$$

and $\boldsymbol{D}^k$ is the k-th power of $\boldsymbol{D}$ (not to be confused with $\boldsymbol{D}_k$).

Theorem 7.3.2 *The entry (i,j) of the matrix $\boldsymbol{D}_k$ contains the length of the shortest path that crosses k borders from i to j, or is equal to zero if such a path does not exist.*

For a proof of this theorem, see appendix E.

Secondly, by scanning $\boldsymbol{D}_k(i,j)$ for all possible values of k, we then choose the length of the shortest path, i.e. we choose the (i,j) entry in the matrix $\boldsymbol{D}_k$ where k is such that $\boldsymbol{D}_k(i,j) = \min_l D_l(i,j)$. Since the minimal number of boundaries between any two areas is at most $n-1$, we only have to consider values of k smaller than the number of areas n.

The S-PLUS code for the calculation of the geographical distance is displayed in appendix D. This algorithm is an alternative to the algorithm presented in Dijkstra (1959) for calculating the shortest path between two vertices.

7.4 Likelihood

The model we consider is similar to the regression mixture model described in chapter 6, except that the M groups are replaced by M connected clusters $C_1, \ldots, C_M$. Each connected cluster, $C_m \subset \{1, \ldots, n\}$, has constant virulence b_{m1}. The number of geographical clusters, M, is assumed known. The geographical clusters $C_1, \ldots, C_M$ cover the whole region and they do not overlap. Their sizes and locations are unknown.

In the regression mixture model, the groups were defined through the allocation variables $\boldsymbol{V}$. Here, the geographical clusters are defined by the vector of M cluster centres, $\boldsymbol{g} = (g_1, \ldots, g_M) \in \{1, \ldots, n\}^M$, using the geographical distance defined in section 7.3.1. In particular, g_m is the cluster centre of $C_m, m = 1, \ldots, M$.

The conditional likelihood for this model is given by

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{g}) \;=\; \prod_{m=1}^{M} \prod_{i \in C_m(\boldsymbol{g})} f_m(y_i|\theta_m), \tag{7.7}$$

where $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are defined as in section 6.1 and $(C_1(\boldsymbol{g}), \ldots, C_M(\boldsymbol{g}))$ is the set of geographical clusters.

7.5 Prior distributions

7.5.1 Prior distribution for the cluster centres

We have seen in section 7.2 that the order of selection of the cluster centres matters in the definition of the cluster partition. The number of possible permutations of the M cluster centres is $n!/(n-M)!$. We use a uniform distribution

over all possible permutations, i.e. we assume that each vector of cluster centres $\boldsymbol{g} = (g_1, \ldots, g_M)$ has equal probability (Knorr-Held and Rasser, 2000),

$$\pi(\boldsymbol{g}) \;=\; \frac{(n-M)!}{n!}\,. \tag{7.8}$$

7.5.2 Prior distribution for the regression parameters

The prior distributions for regression parameters are chosen in the same way as the ones chosen for the regression mixture model in chapter 6.

7.6 Label switching in the clustering model

The clustering model is subject to the same label-switching problem as the mixture model discussed in chapter 6. Consider for example the boundary distance. Figure 7.1 shows that the configuration defined by $g_1 = 1$ and $g_2 = 4$ is the same as the one defined by $g_2 = 1$ and $g_1 = 4$, i.e. both pairs of clusters centres lead to the clusters $\{1, 2\}$ and $\{3, 4\}$. Therefore, we introduce the constraint on the slopes that was considered in the mixture model (section 6.4.2, equation 6.9). However, due to expression 7.2, the set of cluster centres $\boldsymbol{g} = (g_1, \ldots, g_M)$ do not always define the same partition as the set resulting from a permutation of these cluster centres, $\boldsymbol{g} = (g_{\nu(1)}, \ldots, g_{\nu(M)})$, where $\nu(\boldsymbol{g})$ is a permutation of $(1, \ldots, M)$. This is illustrated by the configurations shown in figure 7.2. The clusters defined by $g_1 = 2$ and $g_2 = 4$ are $C_1 = \{1, 2, 3\}$ and $C_2 = \{4\}$, which are not the same as those defined by $g_2 = 2$ and $g_1 = 4$, namely $C_1 = \{3, 4\}$ and $C_2 = \{1, 2\}$. Therefore, a constraint $b_{11} < b_{21}$ will not necessarily yield the same results as the constraint $b_{21} < b_{11}$.

The problem of the non-exchangeable cluster centres is due to the use of

the attribution rule for allocating areas which are equally distant to two or more cluster centres (expression 7.2). The use of this attribution rule is likely to be common when the clusters are defined by the boundary distance, but the rule may never be used in practice when the clusters are defined by the geographical distance, which uses the Euclidean distance. In fact, in the analysis of chapter 9, we used the geographical distance and observed that the results were the same when we run the chain with the constraint $b_{11} < b_{21}$ or $b_{21} < b_{11}$. This is certainly a positive aspect of the use of the geographical distance for defining the clusters.

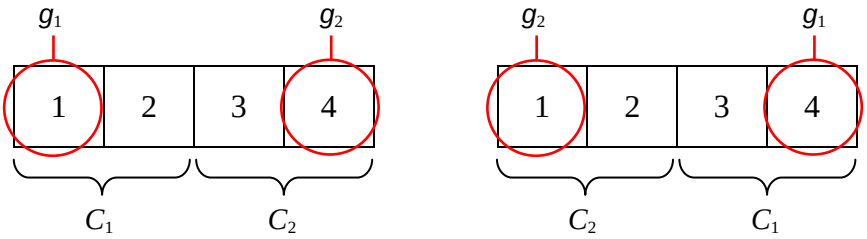


Figure 7.1: Exchangeable cluster centres.

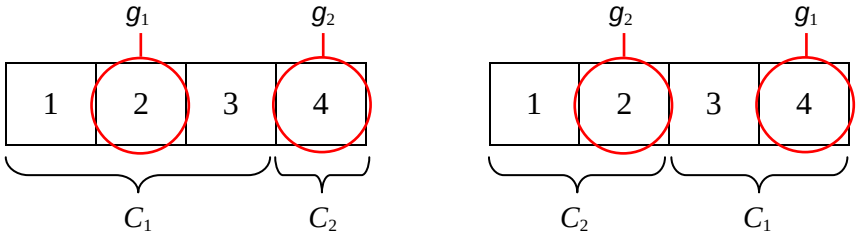


Figure 7.2: Non-exchangeable cluster centres.

7.7 Posterior distribution

The joint posterior distribution is given by

$$\pi(\boldsymbol{g}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2 | \boldsymbol{y}) \propto f(\boldsymbol{y} | \boldsymbol{g}, \beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2) \pi(\boldsymbol{g}) \pi(\beta_0) \pi(\boldsymbol{b}_{\bullet 1}) \pi(\sigma^2), \tag{7.9}$$

where $\pi(\boldsymbol{g})\pi(\beta_0)\pi(\boldsymbol{b}_{\bullet 1})\pi(\sigma^2)$ denote the prior distributions. Due to the complexity of the model, the posterior distribution is not directly tractable. In the same way as for the regression mixture model (section 6.2), we use the Gibbs sampler for inference on the posterior quantities of interest.

Since $E[\theta|\boldsymbol{y}] = E[E[\theta|\boldsymbol{g}, \boldsymbol{y}]]$, the posterior means of the regression parameters are weighted averages of the posterior means for the regression parameters for each configuration. The weighting is done through the posterior probability of all possible configurations. For a given configuration, the regression estimates are typically weighted averages of the classical estimators and the prior means of each parameter.

7.8 Full conditionals

As explained in section 6.4.1, the full conditional distributions can be determined from the joint distribution. Since we decided in section 7.5.2 to choose the same prior distributions for the regression parameters as those used in the regression mixture model, the joint posterior distribution for the parameters of the clustering model, given by expression 7.9, is similar to that of the regression mixtures (expression 6.7), the difference being in the use of $\boldsymbol{g}$ in the place of the vectors $\boldsymbol{v}$ and $\boldsymbol{p}$. In addition, the likelihood for both models is very similar, with the difference that the products $\prod_{i:v_i=m}$ in the regression mixture are replaced by $\prod_{i \in C_m(\boldsymbol{g})}$ in the clustering model (expressions 6.5 and 7.7). The full conditional

distributions for the regression parameters are then similar to the ones for the standard regression mixtures.

Turning now to the cluster centres $\boldsymbol{g}$, the full conditional distribution is given by

$$\pi(\boldsymbol{g}|\beta_0, \bullet\mathbf{1}, \sigma^2, \boldsymbol{y}) \propto \exp(-\frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{i\in C_m(\boldsymbol{g})}(y_i - \beta_0 - b_{m1}z_i)^2), \qquad (7.10)$$

that is, it privileges configurations $\boldsymbol{g}$ which minimize the sum,

$$\sum_{m=1}^{M}\sum_{i\in C_m(\boldsymbol{g})}(y_i - \beta_0 - b_{m1}z_i)^2 .$$

7.9 Representation of geographical clusters

While the procedure presented in section 7.3 for defining clusters always leads to connected clusters, there may exist partitions of connected clusters which can not be defined by a set of cluster centres. In other words, the function $\boldsymbol{g} \mapsto (C_1(\boldsymbol{g}), \ldots, C_M(\boldsymbol{g}))$ of cluster centres into configurations of connected sets is not necessarily surjective.

(a)

1	2	3
4	**5**	6

(b)

1	2	3	4
5	**6**	7	8
9	**10**	11	12
13	14	15	16

Figure 7.3: Non-representable partitions.

By way of illustration, we present a simple example. Consider a lattice of 6 units such as the one represented in figure 7.3, partition (a), where any two

adjacent areas are at the same distance. The two clusters $C_1 = \{5\}$ and $C_2 = \{1, 2, 3, 4, 6\}$ are obviously connected. However, no pair of cluster centres (g_1, g_2) could lead to such a partition. It is obvious that $g_1 = 5$ or $g_2 = 5$. In either case, it is not possible to choose the second cluster centre in such a way that all other areas are closer to it than to area number 5.

In the same way, there is no pair of cluster centres (g_1, g_2) which could lead to the partition represented in figure 7.3, partition (b), $C_1 = \{5, 6, 9, 10\}$ and $C_2 = \{1, 2, 3, 4, 7, 8, 11, 12, 13, 14, 15, 16\}$.

7.10 Equiprobability of cluster configurations

In distinction to the standard regression mixtures described in Chapter 6, the prior probabilities of allocation, $P(V_i = m)$, $i = 1, \ldots, n$, $m = 1, \ldots, M$, are not independent of i and m in the clustering model. In fact, some partitions are more probable than others. This happens because the function $\boldsymbol{g} \rightarrow C_1, \ldots, C_M$ is not injective, i.e. two or more different sets of cluster centres $\boldsymbol{g}$ may lead to the same partition $C_1, \ldots, C_M$. Since all sets of cluster centres $\boldsymbol{g}$ are taken to have equal prior probabilities (section 7.5.1), the probability of a partition $C_1, \ldots, C_M$ depends on the number of $\boldsymbol{g}$s that lead to that partition.

For instance, in the 4×4 square lattice of Figure 7.4, partition (a), there is only one pair of cluster centres (g_1, g_2) which leads to the partition $C_1 = \{4\}$ and $C_2 = \{1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16\}$, namely $(g_1, g_2) = (7, 4)$. However, the partition $C_1 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ and $C_2 = \{9, 10, 11, 12, 13, 14, 15, 16\}$ (figure 7.4, partition (b)), can be represented by several pairs of cluster centres, for example, $(g_1, g_2) = (5, 9)$, or $(g_1, g_2) = (6, 10)$ or $(g_1, g_2) = (2, 14)$, etc.

We will study the impact of non-representability and non-equiprobability on

(a)

1	2	3	**4**
5	6	7	8
9	10	11	12
13	14	15	16

(b)

1	2	3	4
5	6	7	8
9	**10**	**11**	**12**
13	**14**	**15**	**16**

Figure 7.4: Two partitions with different prior probability.

the estimation of probabilities of classification through a simulation study in section 7.13.

7.11 Bayesian cluster allocation or probabilities of membership

The determination of the posterior probabilities of membership is of interest, to assign the observations to connected clusters of identical exposure-coefficient.

In the clustering model, the allocation of areas to clusters is done through the vector of cluster centres $\boldsymbol{g}$. Hence, the values of the prior probabilities of membership $P(V_i = m)$, i.e. the prior probability that area i is allocated to cluster m, depend on the prior distribution of $\boldsymbol{g}$ and on the spatial configuration of the areas. Since it is not straightforward to calculate the prior distribution of the Vs, it is not easy to use expression 6.11, provided in chapter 6, to calculate the posterior probabilities of membership. We can alternatively monitor the proportion of iterations in which each observation i is allocated to the group m during an MCMC run. The posterior probabilities of allocation are then estimated

by the relative amount of simulation time in which area i was allocated to group m. For this monitoring, we defined for each group m and each observation i an indicator variable $v_{im}^{(t)}$ which equals one when the observation i is assigned to group m at the t-th iteration and zero otherwise.

In the MCMC run, the distribution of states in the chain will converge towards an equilibrium distribution which is the joint posterior distribution. Hence, running the chain for a long time, we expect $v_{im}^{(t)}$ to be approximately distributed according to $\mathrm{Bernoulli}(P(V_i = m|\boldsymbol{y}))$. The ergodic theorem on Markov chains tell us that the average of the values sampled between iterations t_1 and t_2,

$$\frac{\sum_{t=t_1}^{t_2} V_{im}^{(t)}}{t_2 - t_1 + 1},$$

will converge almost surely to $E[V_{im}|\boldsymbol{y}] = E[P(V_i = m|\boldsymbol{y})]$ (Robert and Casella, 1999, p.140-1). Under general conditions, the Gibbs-sampler generates Markov chains. Therefore, we can monitor the quantity $\frac{\sum_{t=t_1}^{t_2} V_{im}^{(t)}}{t_2 - t_1 + 1}$ in order to estimate the posterior means of the probabilities of membership.

7.12 Predictive probabilities

Predictive classification addresses the question of classifying a future observation y^*. If the allocation variable corresponding to this is v^*, then we are in principle interested in the posterior probabilities of membership of the new observation y^* to each cluster m, $P(v^* = m|\boldsymbol{y}, y^*, P_1^*, \ldots, P_M^*, M)$, where P_j^* is the prior probability that the new observation belongs to cluster j. Although it is common in academic examples to choose P_j^* as the membership probabilities $P_j = P(V_i = j|\boldsymbol{y})$ of the original set of observations, in some cases this may not be a reasonable choice. As we will see in chapter 9, when the mechanism under-

lying the generation of the new data is not exactly the same as for the original set and/or the original data were not a random sample of the population, it may be more appropriate to choose uniformly distributed probabilities of membership.

In a maximum likelihood approach, the $P(v^* = j|\mathbf{y}, y^*, P_1, \ldots, P_M, M)$ can be estimated by

$$\frac{P_m^* f(y^*|\hat{\boldsymbol{\theta}}_m)}{\sum_l P_l^* f(y^*|\hat{\boldsymbol{\theta}}_l)},$$

where $\hat{\boldsymbol{\theta}}_m$ are the estimates of the regression parameters for cluster m; and the sum is over the set of clusters to which the new observation y^* might belong.

In a fully Bayesian framework, we can monitor the $P(v^* = j|\mathbf{y}, y^*, P_1, \ldots, P_M)$ during the MCMC run to obtain an estimate of

$$\int \frac{P_m^* f(y^*|\boldsymbol{\theta}_m)}{\sum_l P_l^* f(y^*|\boldsymbol{\theta}_l)} \pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M|\boldsymbol{y}),$$

where $\boldsymbol{y}$ is the set of the original observations. This quantity reflects better the uncertainty in the estimation of $\boldsymbol{\theta}$.

Using the percentage correctly classified loss function (section 6.5), the Bayes classification of a new observation $y*$ is then given by

$$\widehat{v^*} \quad = \quad \text{argmax}_j\{P(v^* = j|\mathbf{y}, y^*, P_1, \ldots, P_M)\}. \tag{7.11}$$

In the application presented in Chapter 9, we know there are M populations of *H. pylori*, found on the basis of the expected geographical distribution of the bacterium and the relation between the infection prevalence of that bacterium and stomach cancer. We would like to estimate the probabilities of each strain in another country, which may have been eventually contaminated by one or several of the strains from the original set. For this, we will assume in section 9.6 that the prior probabilities of membership $P_1^*, \ldots, P_M^*$ of the new country are uniformly distributed among the different clusters of countries.

7.13 Performance of the clustering model

In this section, we briefly discuss the performance of the clustering model described in this chapter in synthetic data sets. As an illustration, we consider three different partitions, each formed by two clusters ($M = 2$):

- Partition 1 consists two clusters of approximately the same size;
- Partition 2 contains one cluster which is about one third the size of the other;
- Partition 3 can not be represented by a pair of cluster centres (one cluster about one third the size of the other).

These three partitions were designed to test the performance of the clustering model in different situations. Comparing the results of partitions (1) and (2) will allow us to assess the influence of the non-equiprobability of the partitions. For partition (3), the aim is to investigate how the cluster model behaves when the partition is not representable by cluster centres.

In all the examples, we analysed a regular square lattice of size 7×7. Note that in square lattices the boundary distance is equivalent to the geographical distance (as defined in section 7.3.1), i.e. the distance between any pair of adjacent areas is the same. For simplicity, we only attempt to estimate areal means. The existence of covariates is an addition to the complexity of estimation.

We compare the clustering model with the standard mixture model. We incrementally increased the variance in order to assess the performance of the clustering model and the standard mixture model when the observations have large variance but the clusters are known to be connected.

All computations were conducted in Winbugs 1.4. We used the code shown in appendix G, adapted for two clusters and for Gaussian responses. In order to assess convergence, we inspected the sampled values of the means and variance for several MCMC runs and the corresponding Brooks-Gelman-Rubin convergence statistics (Brooks and Gelman, 1998). Different initial values were used for each of the MCMC runs.

The Brooks-Gelman-Rubin convergence statistics is the ratio of the length of the 80% interval for the pooled runs over the average length of the 80% intervals for the individual runs. As convergence is approached, the Brooks-Gelman-Rubin convergence statistics are expected to be near one and the pooled and the average interval lengths should stabilise. In general, stationarity occurred rather quickly in the clustering model (often quicker than in the standard mixtures). The Brooks-Gelman-Rubin convergence statistics were immediately drawn to one even with very dispersed initial values across runs. A length of 10000 iterations was judged reasonable for the burn-in period. The 10000 iterations after the burn-in period yielded reasonable values of the Monte Carlo error, which estimates the sampling error of the posterior mean in the MCMC run (Roberts, 1996; Spiegelhalter et al., 2003). Therefore, the total number of iterations was 20000.

We only present the results for a *single* simulated sample. Replications of these experiments were conducted and provided similar results.

For each example, we display the posterior mean and standard deviation of the group/cluster parameters; the number of misclassified areas; and the *averaged posterior mean squared error* (MSE) of the areal means (Green and Richardson, 2002), defined as

$$\frac{1}{n}\sum_{i=1}^{n} E[(\mu_{v_i} - r_i)^2|\boldsymbol{y}], \tag{7.12}$$

where μ_{v_i} and r_i are respectively the estimated and the true areal means, $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the vector of observations, and the expectation is with respect to the posterior distribution of $(\mu_{v_1}, \ldots, \mu_{v_n})$.

Example 1 (partition 1)

We analyse two clusters of approximately the same size in a 7×7 squared lattice (represented in normal font and bold font in the figure 7.5(a)). We set the true areal means equal to 1 in the first cluster (normal font); and equal to 2 in the second cluster (bold font). Three different values for the true standard deviation σ were used: 0.1, 0.3 and 0.5.

(a)

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	**23**	**24**	**25**	**26**	**27**	**28**
29	**30**	**31**	**32**	**33**	**34**	**35**
36	**37**	**38**	**39**	**40**	**41**	**42**
43	**44**	**45**	**46**	**47**	**48**	**49**

(b)

	Averaged posterior MSE	
σ	Clustering model	Mixture model
0.1	6.52×10^{-4}	6.43×10^{-4}
0.3	0.018	0.094
0.5	0.095	0.286

Figure 7.5: Example 1: (a) partition 1, 7×7 squared lattice; and (b) averaged posterior mean squared error (MSE) for different values of σ, for the clustering model and the mixture model.

In figure 7.5(b), we display for each of the models the averaged posterior MSE calculated according to expression 7.12, for each value of σ. Table 7.1 displays, for each of the models and for three different values of σ, the Bayesian estimates of the posterior means of the model parameters and the number of misclassified areas.

Both models give comparable results for low values of σ: the averaged posterior MSE, the number of misclassified areas and the estimates of the model

(a) $\sigma = 0.1$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.01	2.01	0.11	0
	(0.025)	(0.022)	(0.012)	
Mixture model	1.01	2.01	0.11	0
	(0.025)	(0.022)	(0.012)	

(b) $\sigma = 0.3$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.01	2.10	0.30	0
	(0.069)	(0.058)	(0.032)	
Mixture model	1.09	2.12	0.36	1
	(0.16)	(1.53)	(0.09)	

(c) $\sigma = 0.5$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	0.97	1.92	0.59	4
	(0.14)	(0.13)	(0.06)	
Mixture model	1.13	2.61	0.66	11
	(0.27)	(5.25)	(0.11)	

Table 7.1: Example 1 (partition 1), 7×7 squared lattice, with (a) $\sigma = 0.1$, (b) $\sigma = 0.3$ and (c) $\sigma = 0.5$: posterior means and posterior standard deviations (in brackets) of the regression parameters μ_1, μ_2 and σ, and number of misclassified areas (NMA), for the clustering and the mixture model.

parameters are practically identical for $\sigma = 0.1$.

However, as the variance increases, the clustering model gives more accurate marginal inference on the true areal means than the regression mixtures. In particular, the averaged posterior MSE of the clustering model and the number of misclassifications are smaller than those of the mixture model. Regarding the estimates of the model parameters, the disparity between the estimates of the clustering and the mixture models increases as the variance increases, with the estimates of the clustering model being closer to the true values.

Example 2 (partition 2)

In a 7×7 square lattice, we analyse a partition in which one cluster is about half the size of the other (the larger and the smaller clusters are represented in normal and bold font, respectively, in the figure 7.6(a)). We set the true areal means of the larger cluster equal to 1; and those of the smaller cluster equal to 2. As in example 1, three different values for σ are considered: 0.1, 0.3 and 0.5.

(a)

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	**23**	**24**	**25**	26	27	28
29	**30**	**31**	**32**	33	34	35
36	**37**	**38**	**39**	40	41	42
43	**44**	**45**	**46**	47	48	49

(b)

	Averaged posterior MSE	
σ	Clustering model	Mixture model
0.1	4.40×10^{-4}	4.25×10^{-4}
0.3	0.019	0.15
0.5	0.022	0.20

Figure 7.6: Example 2: (a) partition 2, 7×7 squared lattice; and (b) averaged posterior mean squared error (MSE) for different values of σ, for the clustering model and the mixture model.

For each of the models, we calculated the averaged posterior MSE of the areal means for each value of σ. The values obtained are displayed in figure 7.6(b) and lead to similar conclusions to those obtained in the previous example 1. Again, the clustering model performs better than the mixture model for large values of σ, and shows comparable performance for small values of σ. The summary measures in table 7.2 also support this fact. Although the configuration in example 2 is less probable than that in example 1, the performance of the clustering model is similar in both examples. This suggests that non-equiprobability does not have a major effect in the identification of the clusters.

(a) $\sigma = 0.1$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	0.99	1.98	0.083	0
	(0.014)	(0.021)	(0.008)	
Mixture model	0.99	1.98	0.080	0
	(0.014)	(0.020)	(0.008)	

(b) $\sigma = 0.3$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.01	1.94	0.33	0
	(0.058)	(0.087)	(0.035)	
Mixture model	1.11	3.14	0.36	4
	(0.13)	(7.06)	(0.10)	

(c) $\sigma = 0.5$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.05	1.99	0.47	0
	(0.08)	(0.13)	(0.05)	
Mixture model	1.10	2.37	0.47	7
	(0.15)	(3.66)	(0.10)	

Table 7.2: Example 2 (partition 2), 7×7 squared lattice, with (a) $\sigma = 0.1$, (b) $\sigma = 0.3$ and (c) $\sigma = 0.5$: posterior means and posterior standard deviations (in brackets) of the regression parameters μ_1, μ_2 and σ, and number of misclassified areas (NMA), for the clustering and the mixture model.

Example 3 (partition 3)

We analyse a partition of clusters which can not be represented by a pair of cluster centres, with one cluster about half the size of the other (figure 7.7(a)).

The results displayed in figure 7.7(b) and table 7.3 show that the clustering model performs worse in this example than in the two previous examples. This was expected because the cluster configuration in this example can not be represented by a pair of cluster centres. Despite the non-representability of the configuration, for large values of σ, the performances of the mixture and the

(a)

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	**16**	**17**	**18**	19	20	21
22	**23**	**24**	**25**	26	27	28
29	**30**	**31**	**32**	33	34	35
36	**37**	**38**	**39**	40	41	42
43	44	45	46	47	48	49

(b)

	Averaged posterior MSE	
σ	Clustering model	Mixture model
0.1	0.069	4.30×10^{-4}
0.3	0.13	0.14
0.5	0.13	0.20

Figure 7.7: Example 3: (a) partition 3, 7×7 squared lattice; and (b) averaged posterior mean squared error (MSE) for different values of σ, for the clustering model and the mixture model.

(a) $\sigma = 0.1$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.00	1.77	0.29	4
	(0.054)	(0.066)	(0.032)	
Mixture model	0.99	1.98	0.080	0
	(0.014)	(0.020)	(0.008)	

(b) $\sigma = 0.3$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.12	1.64	0.45	4
	(0.10)	(0.12)	(0.05)	
Mixture model	1.08	1.96	0.33	4
	(0.11)	(2.10)	(0.08)	

(c) $\sigma = 0.5$, $\mu_1 = 1$, $\mu_2 = 2$

Model	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}$	NMA
Clustering model	1.07	1.76	0.55	5
	(0.12)	(0.16)	(0.07)	
Mixture model	1.11	2.59	0.48	8
	(0.15)	(4.69)	(0.11)	

Table 7.3: Example 3 (partition 3), 7×7 squared lattice, with (a) $\sigma = 0.1$, (b) $\sigma = 0.3$ and (c) $\sigma = 0.5$: posterior means and posterior standard deviations (in brackets) of the regression parameters μ_1, μ_2 and σ, and number of misclassified areas (NMA), for the clustering and the mixture model.

clustering models are comparable. Actually, for $\sigma = 0.5$, the clustering model seems to perform better: the averaged posterior MSE and the number of misclassified areas are smaller, and most parameter estimates are closer to the true values. These results indicate that although some cluster configurations can not be found with a clustering model, the model may still be preferable to the mixture model when connected clusters are expected and the variance of the observations is large.

Chapter 8

Spatially varying regressions

Recently, some authors have aimed at assessing and modelling the spatial variation of the regression coefficient $\boldsymbol{b}_{\bullet 1}$. Brunsdon et al. (1998) used kernel functions to explore the spatial variation of regression coefficients, with a technique known as *geographically weighted regression*. The parameters of the kernels were estimated on the basis of data from areas within a circle of radius R, by minimizing the sum of squared errors. The authors applied this model to data on the prevalence of long-term illness in the north of England. Assuncão et al. (2002) assumed that the prior distributions of the random slopes as well as the random intercept were intrinsic autoregressions. Their hierarchical model, which also assumes independence between the random slopes and the random intercept, was used for estimating fertility rates by age in Brazil. Assuncão (2003) and Gamerman et al. (2003) extended the model of Assuncão et al. (2002) to incorporate within area dependence among the random intercept and the random slopes. Gelfand et al. (2003) used a p-variate Gaussian spatial process model to accommodate spatial variation of both the intercept and the slopes, as well as within location dependence among the random intercept and the random slopes.

The purpose of their modelling was to explain log selling price of single-family houses.

In several of the applications above (Brunsdon et al., 1998; Assuncão, 2003; Gamerman et al., 2003; Gelfand et al., 2003), the data consisted of only one observed value of the response and one observed value of each of the covariates per location. Richardson and Best (2003) discussed the modelling of varying exposure coefficients and pointed out that the parameters may not always be identifiable and the estimates may not always be reliable, particularly when only one data point per area is available. Identifiability issues were also referred to in Assuncão (2003), but unfortunately none of the papers above presented any formal or simulation results. An exception is Gamerman et al. (2003), although the simulation study provided is relatively limited in scope. A few authors also mentioned implementation problems, especially for non-Gaussian responses (Gelfand et al., 2003; Gamerman et al., 2003).

Our approach in this chapter for modelling spatially varying exposure coefficients is to assume that the regression coefficients are a realization from a spatial intrinsic autoregression, as in Assuncão (2003). We use the findings of chapter 5 to assess coding-invariance in these models and derive a distribution for the random effects satisfying the necessary conditions of lemmas 5.4.1 to 5.4.4. We will refer to the resulting distribution, which was presented in Gamerman et al. (2003), as the multivectorial intrinsic autoregression because it is based on two or more random vectors which are distributed as (univectorial) intrinsic autoregressions (section 8.1).

We prove in this chapter that bivectorial intrinsic autoregressions can be found as a limiting case of two correlated conditional autoregressions, in the same way

as univectorial intrinsic autoregressions, which are limiting cases of conditional autoregressions.

This result can be generalized to any dimension. We also show that multivectorial intrinsic autoregressions can be expressed as a linear transformation of independent univectorial intrinsic autoregressions. This will inspire the development of a hierarchical modelling suitable for easy implementation of an MCMC sampling scheme. Finally, we discuss the identifiability of the model proposed in this chapter.

8.1 Univectorial intrinsic autoregressions for varying exposure coefficients

There is a vast amount of literature on the use of intrinsic autoregressions to estimate areal means (e.g. Mollié (1996)). Their use in modelling geographically varying slopes is not so common, though it can be found in Assunção et al. (2002).

The slope parameter b_{i1} in the model of expression 4.5 is assumed here to have a spatial structure induced by a neighbourhood graph where each site i has a set of neighbours indicated by ∂i. Following the notation of chapter 6, $i \sim s$ means that site i is a neighbour of site s. The prior distribution of $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{n1})$ is defined as an intrinsic autoregression,

$$f(\boldsymbol{b}_{\bullet 1}|\tau) \quad \propto \quad \exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}\sum_{s>i} w_{is}(b_{i1} - b_{s1})^2\right), \tag{8.1}$$

where $w_{is} = 1$ if site i is a neighbour of site s, and 0 otherwise. The conditional

distribution is then given by

$$b_{i1}|\{b_{s1}, s \neq i\}, \tau \sim \mathrm{N}\left(\bar{b}_{i1}, \frac{1}{\tau w_{i+}}\right), \tag{8.2}$$

where $w_{i+} = \sum_{s \neq i} w_{is}$ and $\bar{b}_{i1}$ is the mean of the set $\{b_{s1} : s \in \partial i\}$.

It is easy to see that the joint distribution (8.1) is invariant under translation of $(b_{11}, \ldots, b_{n1})$ by a single constant added to all elements, and it is thus improper. This distribution can be made proper by adding a constraint such as $\sum_i b_{i1} = C$, where C is a constant (Best et al., 1999; Assuncão, 2003).

8.2 Coding invariance

If only the slope is random but the intercept is constant, the model is not well-formulated (section 5.3). Therefore, we know from lemma 5.4.1 that the model is not coding invariant. According to lemma 5.4.2, if the random intercept and the random slope are independent then the model is also not coding invariant. Hence, if the exposure coefficient is modelled as a random effect as in section 8.1, the model should also include a random intercept term and allow for a non-zero correlation between the random intercept and the random slope. Finally, the family of the distribution of the vector of random effects should be closed under coding transformations (lemmas 5.4.3 and 5.4.4).

In section 8.4, we present a distribution satisfying these characteristics, the *bivectorial intrinsic autoregression*, which is the two-dimensional version of the multivectorial intrinsic autoregression (Gamerman et al., 2003). The *bivectorial intrinsic autoregression* is a limiting case of the *bivectorial conditional autoregression*. As the name indicates, the latter distribution is based on two vectors which are distributed as (univectorial) conditional autoregressions.

8.3 Bivectorial conditional autoregressions

For purposes of illustration, we consider the model given in equations 2.8 and 4.5. Let $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{n1})$ be distributed as a univectorial conditional autoregression (Richardson, 1992), i.e.

$$\boldsymbol{b}_{\bullet 1} \sim \mathrm{N}(\boldsymbol{0}, (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{T}), \tag{8.3}$$

where $\boldsymbol{T} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and $(\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{T}$ is symmetric and positive definite. Equivalently, the conditional distributions of the random slopes are:

$$b_{i1}|b_{j1}, j \neq i \sim \mathrm{N}(\sum_{j=1}^{n} C_{ij} b_{j1}, \sigma_i^2), \tag{8.4}$$

where C_{ij} is the entry (i, j) of the matrix $\boldsymbol{C}$.

Several authors (Besag et al., 1991; Richardson, 1992; Besag and Koperberg, 1995; Mollié, 1996) have suggested, in the context of intrinsic autoregressions, that the conditional variance for irregular lattices should depend on the number of neighbours of each area (as for example in equation 8.2). For the conditional autoregression presented in equation 8.3, this can be achieved by defining $\boldsymbol{C} = \rho_S \boldsymbol{D}\boldsymbol{W}$ and $\boldsymbol{T} = \sigma^2 \boldsymbol{D}$, where $\boldsymbol{W}$ is the neighbourhood matrix, with entries $w_{is} = 1, \text{ if } i \sim s$ and 0 otherwise; $\boldsymbol{D} = \mathrm{diag}(1/w_{1+}, \ldots, 1/w_{n+})$; and w_{i+} is the sum of the entries on row i of the matrix $\boldsymbol{W}$. This parameterisation leads to the following model:

$$\boldsymbol{b}_{\bullet 1} \sim \mathrm{N}(\boldsymbol{0}, (\boldsymbol{I} - \rho_S \boldsymbol{D}\boldsymbol{W})^{-1}\boldsymbol{D}\sigma^2). \tag{8.5}$$

Since the entry (i, s) of the precision matrix, $\boldsymbol{D}^{-1}(\boldsymbol{I} - \rho_S \boldsymbol{D}\boldsymbol{W})$, is given by

$$\begin{cases} w_{i+} & \text{for } i = s \\ -\rho_S w_{is} & \text{for } i \neq s \end{cases},$$

the variance matrix in equation 8.5 is symmetric. When ρ_S is smaller than one, the matrix is also positive definite (Besag and Koperberg, 1995). The parameter σ^2 determines the order of magnitude of the variances and covariances. Without loss of generality for the results in the remainder of this chapter, we set $\sigma^2 = 1$. Then

$$E(b_{i1}|b_{j1}, j \neq i) = \sum_{j=1}^{n} \rho_S \frac{w_{ij}}{w_{i+}} b_{j1},$$
$$V(b_{i1}|b_{j1}, j \neq i) = \frac{1}{w_{i+}}.$$

In what follows, $\rho(i, s)$ denotes the covariance between sites i and s created by such a conditional autoregression, i.e. $\rho(i, s)$ corresponds to the entry (i, s) of the covariance matrix in (8.5), with $\sigma^2 = 1$.

We assume at this stage that the random intercept $\boldsymbol{b}_{\bullet 0} = (b_{10}, \ldots, b_{n0})$ is also distributed as a conditional autoregression, with the same covariance matrix as that used for the random slope, i.e.

$$\boldsymbol{b}_{\bullet 0} \sim \mathrm{N}(\mathbf{0}, (\boldsymbol{I} - \rho_S \boldsymbol{D}\boldsymbol{W})^{-1}\boldsymbol{D}).$$

Then the covariance between the random intercepts at two different sites, for example b_{i0} and b_{s0}, is the same as the covariance between the random slopes at the same two sites, which we have denoted as $\rho(i, s)$.

We want to allow for non-zero correlations between the two vectors $\boldsymbol{b}_{\bullet 0}$ and $\boldsymbol{b}_{\bullet 1}$. With this aim, and inspired by a construction used in Gelfand et al. (2003) in the context of covariance functions for spatial processes, we define the distribution of a new vector $\boldsymbol{b} = (b_{10}, b_{11}, \ldots, b_{n0}, b_{n1})^t$ as

$$\boldsymbol{b} \sim \mathrm{N}(\mathbf{0}, ((\boldsymbol{I} - \rho_S \boldsymbol{D}\boldsymbol{W})^{-1}\boldsymbol{D}) \otimes \boldsymbol{\Psi}_c^{-1}), \tag{8.6}$$

with $\otimes$ denoting the Kronecker product and

$$\boldsymbol{\Psi}_c^{-1} = \begin{bmatrix} \sigma_0^2 & \rho_c \sigma_0 \sigma_1 \\ \rho_c \sigma_0 \sigma_1 & \sigma_1^2 \end{bmatrix}.$$

Therefore, the covariance between b_{iu} and b_{st} is given by

$$C(i,s)_{ut} \;\; = \;\; \rho(i,s)\,\psi^{*}_{u+1,t+1}, u,t=0,1; i,s=1,\ldots,n; \qquad (8.7)$$

where $\psi^{*}_{u+1,t+1}$ is the $(u+1,t+1)$ entry of the matrix $\mathbf{\Psi}_c^{-1}$. For instance, the covariance between the random intercept b_{i0} in location i and the random slope b_{s1} in location s is given by $C(i,s)_{01} = \rho_c\sigma_0\sigma_1\rho(i,s)$.

Using one of the properties of the Kronecker product, $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$, it follows that the distribution in expression 8.6 corresponds to

$$\pi(\boldsymbol{b}) \;\; \propto \;\; \exp\big(-\frac{1}{2}\boldsymbol{b}^t(\mathbf{\Psi}_S \otimes \mathbf{\Psi}_c)\boldsymbol{b}\big), \qquad (8.8)$$

where $\mathbf{\Psi}_S = \boldsymbol{D}^{-1}(\boldsymbol{I} - \rho_S\boldsymbol{DW})$. We will refer to this distribution as a bivectorial conditional autoregression.

8.4 Bivectorial intrinsic autoregressions

By letting $\rho_S \rightarrow 1$ in expression 8.6, the matrix $\mathbf{\Psi}_S = \boldsymbol{D}^{-1}(\boldsymbol{I} - \rho_S\boldsymbol{DW})$ becomes singular (the sum of the entries of each row of the matrix $\mathbf{\Psi}_S$ is $1-\rho_S$, which equals zero when $\rho_S = 1$) and the resulting distribution in expression 8.8 becomes a bivectorial intrinsic autoregression (see appendix H for a proof of this statement). The bivectorial intrinsic autoregression was formulated in Gamerman et al. (2003) as:

$$\pi(\boldsymbol{b}|\mathbf{\Psi}_c) \;\; \propto \;\; |\mathbf{\Psi}_c|^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{s=i+1}^{n} w_{is}(\boldsymbol{b}_{i\bullet}-\boldsymbol{b}_{s\bullet})^T\mathbf{\Psi}_c(\boldsymbol{b}_{i\bullet}-\boldsymbol{b}_{s\bullet})\right\}, \qquad (8.9)$$

where $\boldsymbol{b}_{i\bullet} = (b_{i0}, b_{i1}), i = 1,\ldots,n$.

Univectorial intrinsic autoregressions are improper unless appropriate constraints are introduced. Likewise, constraints are also appropriate for the multivectorial case because the right-hand side of expression 8.9 is invariant under translation of $\boldsymbol{b}_{\bullet u}$ by a constant added to all the elements $b_{1u}, \ldots, b_{nu}$, and thus the distribution 8.9 is improper. Constraints such as $\sum_i b_{iu} = C$, where C is a constant, lead to a proper distribution (Assunção, 2003).

The conditional prior distribution of (b_{i0}, b_{i1}) given all the other (b_{s0}, b_{s1}) and the hyperparameters is a bivariate normal,

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} \mid \begin{bmatrix} b_{s0} \\ b_{s1} \end{bmatrix}_{s \in \partial i} \sim \mathrm{N}\left(\begin{bmatrix} \bar{b}_{i0} \\ \bar{b}_{i1} \end{bmatrix}, \frac{1}{w_{i+}} \begin{bmatrix} \sigma_0^2 & \rho_c \sigma_0 \sigma_1 \\ \rho_c \sigma_0 \sigma_1 & \sigma_1^2 \end{bmatrix} \right),$$

where w_{i+} is again the number of neighbours of area i, and $\bar{b}_{i0}$ and $\bar{b}_{i1}$ are respectively the average values of the sets $\{b_{s0} : s \sim i\}$ and $\{b_{s1} : s \sim i\}$. This result was presented in Assunção (2003) for the case in which $\boldsymbol{\Psi}_c$ is diagonal.

The matrix $\boldsymbol{\Psi}_c^{-1}/w_{i+}$ represents the covariance matrix associated with the vector (b_{i0}, b_{i1}) in area i. In particular, the parameter ρ_c defines the conditional correlation between the random intercept b_{i0} and the random slope b_{i1}. Hence, the matrix $\boldsymbol{\Psi}_S$ introduces correlation across areas, while the matrix $\boldsymbol{\Psi}_c$ introduces correlation between random effects of the same area. Here and throughout this thesis we will refer to $\boldsymbol{\Psi}_S$ as the *spatial precision matrix*, and to $\boldsymbol{\Psi}_c$ as the *conditional precision matrix*. Accordingly, we will refer to the parameter ρ_c as the *conditional correlation*.

As discussed in chapter 5, if $\boldsymbol{\Psi}_c$ is diagonal the model is not invariant. We will thus consider the case in which this matrix is not diagonal.

The results of this section can be easily generalized for multivectorial intrinsic autoregressions of any dimension.

8.5 Implementation of spatially varying coefficient models

The computational cost of the implementation of the model 8.9 for Gaussian responses has been mentioned in the literature (Assuncão et al., 2002; Gamerman et al., 2003), as well as the computational difficulties associated with other types of responses (Gamerman et al., 2003).

We propose here a strategy for easily implementing this model. It is based on the fact that any linear transformation of a multivectorial intrinsic autoregression is still a multivectorial intrinsic autoregression, with the same neighbourhood matrix $\boldsymbol{W}$. To put it formally, assume that $\widetilde{\boldsymbol{T}}$ is a block diagonal $nq \times nq$ matrix, $\widetilde{\boldsymbol{T}} = (\boldsymbol{T}, \ldots, \boldsymbol{T})$, where $\boldsymbol{T}$ is a $q \times q$ matrix. The distribution of $\boldsymbol{b}_T = \widetilde{\boldsymbol{T}}\boldsymbol{b}$ is,

$$\pi(\boldsymbol{b}_T|\boldsymbol{\Psi}_{cT}) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\sum_{s=i+1}^{n} w_{is}(\boldsymbol{b}_{Ti\bullet} - \boldsymbol{b}_{Ts\bullet})^T \boldsymbol{\Psi}_{cT}(\boldsymbol{b}_{Ti\bullet} - \boldsymbol{b}_{Ts\bullet})\right\},$$

where $\boldsymbol{b}_{Ti\bullet} = \boldsymbol{T}\boldsymbol{b}_{i\bullet}$ and $\boldsymbol{\Psi}_{cT} = (\boldsymbol{T}^{-1})^T \boldsymbol{\Psi}_c \boldsymbol{T}^{-1}$, and therefore satisfies the distributional form of the multivectorial intrinsic autoregression (expression 8.9).

If $\boldsymbol{\Psi}_c$ is diagonal, the $\boldsymbol{b}_{\bullet j} = (b_{1j}, \ldots, b_{nj}), j = 0, \ldots, q-1$, are a set of independent univectorial intrinsic autoregressions (expression 8.1). We can use the transformation $\boldsymbol{T}$ to obtain a non-diagonal matrix $\boldsymbol{\Psi}_{cT}$, and consequently a set of non-independent $\boldsymbol{b}_{T\bullet j} = (b_{T1j}, \ldots, b_{Tnj}), j = 0, 1, \ldots, q-1$. Hence, a set of q independent univectorial intrinsic autoregressions can be transformed into a q-dimensional multivectorial intrinsic autoregression in which the components are not any more independent.

Since independent univectorial intrinsic autoregressions can be easily implemented with existing software, this suggests a procedure for implementing the model 8.9 for any positive definite $q \times q$ matrix $\boldsymbol{\Psi}_{Tc}$. It remains only to prove

that for any positive definite $q \times q$ matrix $\mathbf{\Psi}_{Tc}$, there exist a matrix $\boldsymbol{T}$ and a diagonal matrix $\mathbf{\Psi}_c$ such that $\mathbf{\Psi}_{cT} = (\boldsymbol{T}^{-1})^T \mathbf{\Psi}_c \boldsymbol{T}^{-1}$. To address this issue, we present in the next section the similar and well-known procedure which is used to generate a bivariate Gaussian distribution from two independent univariate Gaussian distributions. This was the procedure that inspired our work in the context of autoregressions.

8.5.1 Bivariate Gaussian

It is well known that any bivariate Gaussian can be written as a linear transformation of two independent Gaussian distributions. The bivariate normal distribution with zero mean is given by,

$$f(b_0, b_1) = \frac{1}{2\pi\sigma_0\sigma_1\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[(\frac{b_0}{\sigma_0})^2 - \frac{2\rho\, b_0\, b_1}{\sigma_0\, \sigma_1} + (\frac{b_1}{\sigma_1})^2\right]\right\}. \tag{8.10}$$

The distribution has a bell shape: it has a peak at the origin and it then falls symmetrically along two orthogonal directions. The contours of equal probability will be ellipses and if we rotate the axes in such a way that they coincide with those orthogonal directions, the cross term in $b_0 b_1$ in expression 8.10 disappears (figure 8.1).

In two dimensions, a clockwise rotation through an angle ω of the plane about the origin, where (b_0, b_1) is mapped to (z_0, z_1), is given by:

$$\begin{aligned} z_0 &= b_0 \cos\omega - b_1 \sin\omega, \\ z_1 &= b_0 \sin\omega + b_1 \cos\omega. \end{aligned}$$

Simple calculations show that the symmetry axes are at an angle

$$\omega = \frac{1}{2}\tan^{-1}\left[\frac{2\rho\sigma_0\sigma_1}{\sigma_1^2 - \sigma_0^2}\right], \tag{8.11}$$

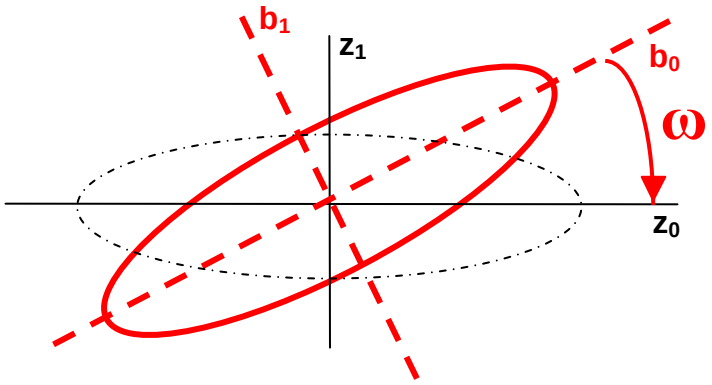


Figure 8.1: Elipse rotation.

as the corresponding rotation eliminates the cross term $b_0 b_1$ in the probability density function (8.10).

Therefore, any zero mean bivariate Gaussian distribution can be expressed as a rotation of two independent univectorial Gaussian distributed variables, z_0 and z_1. The inverse of the transformation of coordinates above can be expressed as

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \varrho & \sqrt{1-\varrho^2} \\ -\sqrt{1-\varrho^2} & \varrho \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \end{bmatrix},$$

where $\varrho = \cos\omega \in (-1, 1)$, z_0 and z_1 are independent zero mean univectorial Gaussian distributions.

Other transformations can be used to generate any bivariate Gaussian distribution. For instance,

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \phi\rho & \phi\sqrt{1-\rho^2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \end{bmatrix},$$

where z_0 and z_1 are independent univectorial Gaussian distributions with zero mean and equal variance σ^2, $\rho \in (-1, 1)$ and $\phi \in (0, \infty)$. In the latter trans-

formation, the parameters are directly interpretable: $V(b_0) = \phi^2\sigma^2$, $V(b_1) = \sigma^2$ and $\mathrm{Cov}(b_0, b_1) = \rho$.

8.5.2 Bivectorial intrinsic autoregressions

Inspired by the example in section 8.5.1, we show that a bivectorial intrinsic autoregression can also be expressed as a linear transformation of two independent univectorial intrinsic autoregressions. If $\boldsymbol{z}_{\bullet 0} = (z_{10}, \ldots, z_{n0})$ and $\boldsymbol{z}_{\bullet 1} = (z_{11}, \ldots, z_{n1})$ are two independent intrinsic autoregressions with the same precision τ^2 (expression 8.1), the joint distribution of $(\boldsymbol{z}_{\bullet 0}, \boldsymbol{z}_{\bullet 1})$ will have the form of a multivectorial intrinsic autoregression (expression 8.9) with conditional precision matrix $\boldsymbol{\Psi}_c$ given by

$$\boldsymbol{\Psi}_c = \begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix}.$$

Applying a rotation of coordinates

$$\begin{bmatrix} b_{i0} \\ b_{i1} \end{bmatrix} = \begin{bmatrix} \phi\rho_c & \phi\sqrt{1-\rho_c^2} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} z_{i0} \\ z_{i1} \end{bmatrix},$$

the joint distribution of $\boldsymbol{b}_{\bullet 0}$ and $\boldsymbol{b}_{\bullet 1}$ keeps the form of a multivectorial intrinsic autoregression (expression 8.9) but the transformed conditional precision matrix $\boldsymbol{\Psi}_{cT}$ will be given by

$$\boldsymbol{\Psi}_{cT} = \frac{1}{\phi^2(1-\rho_c^2)} \begin{bmatrix} \tau^2 & -\rho_c\tau^2\phi \\ -\rho_c\tau^2\phi & \tau^2\phi^2 \end{bmatrix}$$

with corresponding conditional covariance,

$$\boldsymbol{\Psi}_{cT}^{-1} = \begin{bmatrix} \phi^2\sigma^2 & \rho_c\phi\sigma^2 \\ \rho_c\phi\sigma^2 & \sigma^2 \end{bmatrix},$$

where $\tau^2 = 1/\sigma^2$. When $\rho_c \neq 0$, the two random vectors $\boldsymbol{b}_{\bullet 0}$ and $\boldsymbol{b}_{\bullet 1}$ are not independent.

The matrix $\boldsymbol{\Psi}_{cT}^{-1}$ is uniquely defined by the parameters (ϕ, ρ_c) of the transformation

$$\boldsymbol{T} = \begin{bmatrix} \phi\rho_c & \phi\sqrt{1-\rho_c^2} \\ 1 & 0 \end{bmatrix}$$

and by the variance σ^2 of the distributions of $\boldsymbol{z}_{\bullet 0}$ and $\boldsymbol{z}_{\bullet 1}$. Therefore, for each bivectorial intrinsic autoregression $(\boldsymbol{b}_{\bullet 0}, \boldsymbol{b}_{\bullet 1})$ with a given neighbourhood matrix $\boldsymbol{W}$, there is one and only one $\boldsymbol{T}$ with the form above that transforms two independent univectorial intrinsic autoregressions $(\boldsymbol{z}_{\bullet 0}, \boldsymbol{z}_{\bullet 1})$, with variance σ^2 and neighbourhood matrix $\boldsymbol{W}$, into the bivectorial intrinsic autoregression $(\boldsymbol{b}_{\bullet 0}, \boldsymbol{b}_{\bullet 1})$.

8.5.3 Hierarchical structure for implementation

A mixed effects model whose random intercept $\boldsymbol{b}_{\bullet 0}$ and random slope $\boldsymbol{b}_{\bullet 1}$ are distributed as a bivectorial intrinsic autoregression may thus be fitted by considering an additional hierarchical level with the two independent univectorial intrinsic autoregressions vectors $\boldsymbol{z}_{\bullet 0}$ and $\boldsymbol{z}_{\bullet 1}$ and by defining:

$$\begin{aligned} \boldsymbol{b}_{\bullet 0} &= \phi\rho_c \boldsymbol{z}_{\bullet 1} + \phi\sqrt{1-\rho_c^2}\boldsymbol{z}_{\bullet 0}, \\ \boldsymbol{b}_{\bullet 1} &= \boldsymbol{z}_{\bullet 1}. \end{aligned} \tag{8.12}$$

As explained in section 8.4, the parameter ρ_c represents the conditional correlation, which can take values in $(-1, 1)$. Possible hyperprior distributions for this parameter include a Uniform$(-1, 1)$. The vectors $\boldsymbol{z}_{\bullet 0}$ and $\boldsymbol{z}_{\bullet 1}$ are independent univectorial intrinsic autoregressions, with identical variance. The parameter ϕ, which represents the ratio between the conditional standard deviations of the

intercept and the slope, takes positive values. Gamma distributions can be considered for this parameter.

In the expressions 8.12, the random intercept is a weighted average of the random slope $\boldsymbol{b}_{\bullet 1}$ and an additional random term $\boldsymbol{z}_{\bullet 0}$. Therefore, according to the discussion in section 4.1, the quantity $\phi^2\rho_c^2/(\phi^2\rho_c^2 + \phi^2(1-\rho_c^2))$, which is equal to ρ_c^2, can be interpreted as the proportion of the variability of the random intercept induced by the varying exposure coefficient.

8.6 Evaluation of the proposed methodology and implementation

We carried out a simulation study for assessing the performance of the bivectorial intrinsic autoregression in modelling random effects. In particular, we would like to investigate whether the performance of a bivectorial autoregressive model depends on: (i) the correlation between the slope and the intercept and (ii) the ratio between the variability of the slope and the variability of the intercept. A higher correlation between the intercept and the slope is likely to provide better results, in the sense that the estimation process produces estimates that are closer to the postulated means, because the model has more information about the random effects. The variability of the slope relative to the intercept may also influence the results.

8.6.1 Data simulation

We simulated a set of county-level disease counts from a Poisson model with mean $E_i e^{\eta_i}$, where the E_i correspond to the expected number of cases in 275

counties, taken from the Portuguese data set of stomach cancer mortality studied in chapter 10. We also included the spending power in Portugal in 1993 as the covariate (figure 10.1).

On the basis of the spatial structure of the 275 counties, we simulated the η_i as $\eta_i = \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i$, with $\beta_0 = -0.35$ and $\beta_1 = 0$. Spatial patterns of the exposure coefficient and baseline risk were simulated with a bivectorial conditional autoregression, defined in expression 8.6, with values of ρ_S close to one ($\rho_S = 0.999$). The algorithm used for the simulation of a bivectorial conditional autogression is as follows:

1. Generate $\boldsymbol{e}_{i\bullet} = (e_{i0}, e_{i1}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$, for $i = 1, \ldots, n$, where n is the number of areas and
$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & r \\ r & \sigma_2^2 \end{bmatrix},$$
with $r = \rho_c \sigma_1 \sigma_2$.

2. Determine the matrix $\boldsymbol{L}$ that satisfies $\boldsymbol{L}^T\boldsymbol{L} = \Sigma_{CAR}$, where $\Sigma_{CAR} = (I - \rho_S \boldsymbol{DW})^{-1}\boldsymbol{D}$ is the covariance matrix of a univectorial conditional autogression (as defined in expression 8.5 with $\sigma^2 = 1$).

3. Calculate the vectors $\boldsymbol{b}_{\bullet 0} = \boldsymbol{L}^T\boldsymbol{e}_{\bullet 0}$ and $\boldsymbol{b}_{\bullet 1} = \boldsymbol{L}^T\boldsymbol{e}_{\bullet 1}$, where $\boldsymbol{e}_{\bullet 0} = (e_{10}, \ldots, e_{n0})^T$ and $\boldsymbol{e}_{\bullet 1} = (e_{11}, \ldots, e_{n1})^T$.

4. Define $\boldsymbol{b} = (b_{10}, b_{11}, b_{20}, b_{21}, \ldots, b_{n0}, b_{n1})$.

It is easy to prove that the vector $\boldsymbol{b}$ is distributed as a bivectorial conditional autoregression (expression 8.6).

Proof:

Since $b_{i0} = \sum_{j=1}^{n} l_{ij} e_{j0}$ and $u_{i1} = \sum_{j=1}^{n} l_{ij} e_{j1}$, where l_{ij} is the (i,j)-entry of the matrix $\boldsymbol{L}^T$,

- $\mathrm{Cov}(b_{i0}, b_{k1}) = \sum_{j=1}^{n} l_{ij} l_{kj} r$;
- $\mathrm{Cov}(b_{i0}, b_{k0}) = \sum_{j=1}^{n} l_{ij} l_{kj} \sigma_1^2$;
- $\mathrm{Cov}(b_{i1}, b_{k1}) = \sum_{j=1}^{n} l_{ij} l_{kj} \sigma_2^2$.

As c_{ik}, the (i,k)-entry of the matrix Σ_{CAR}, is equal to $c_{ik} = \sum_{j=1}^{n} l_{ij} l_{kj}$, the covariance of $\boldsymbol{b}$ is,

$$V(\boldsymbol{b}) = \begin{bmatrix} c_{11}\Sigma & \dots & c_{1n}\Sigma \\ \vdots & \ddots & \vdots \\ c_{n1}\Sigma & \dots & c_{nn}\Sigma \end{bmatrix} = \Sigma_{CAR} \otimes \Sigma,$$

and the vector $\boldsymbol{b}$ is therefore distributed as a bivectorial conditional autoregression (expression 8.6).

The S-PLUS code for generating a bivectorial conditional autoregression according to the algorithm above is available in appendix I.

For the other parameters of the bivectorial conditional autoregression, we considered three values for the conditional correlation ρ_c: 0.9, 0.5 and 0.1, and two values for the conditional variances: $(\sigma_1, \sigma_2) = (0.1, 0.3)$ and $(\sigma_1, \sigma_2) = (0.3, 0.1)$. Hence, we considered cases in which the slope and the intercept are highly, moderately and weakly conditionally correlated, and in which the conditional variance of the slope is higher or smaller than the variance of the intercept.

8.6.2 Implementation

Fully Bayesian estimation was carried out using MCMC methods due to the intractable nature of the marginal posterior distributions. Using the strategy

described in section 8.5.3, the implementation of a bivectorial intrinsic autoregression reduces to that of univectorial intrinsic autoregressions. Fully Bayesian estimation of the model can be implemented using the Winbugs software. The Winbugs code developed for this purpose is provided in appendix J. It was developed for fitting the following model:

$$\begin{aligned} y_i &\sim \text{Poisson}(E_i\theta_i), i = 1, \ldots, n \\ \log(\theta_i) &= \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i \\ b_{i0} &= \phi\rho_c z_{i1} + \phi\sqrt{1-\rho_c^2}z_{i0} \\ b_{i1} &= z_{i1} \\ \boldsymbol{z}_{\bullet 0} &\sim \text{IAR}(\sigma^2) \\ \boldsymbol{z}_{\bullet 1} &\sim \text{IAR}(\sigma^2) \end{aligned} \tag{8.13}$$

where IAR stands for univectorial intrinsic autoregression.

8.6.3 Results

The Winbugs software was used to run the model in the simulation study. In each case, a burn-in period of 100000 iterations was sufficient to achieve convergence of the regression coefficients, although mixing was poor for some parameters. By way of illustration, sampled values of selected model parameters are shown in figures 8.2 and 8.3, for three MCMC runs and the simulated data set with $\rho_c = 0.9$, $\sigma_1 = 0.3$ and $\sigma_2 = 0.1$. Each MCMC run was initiated with different starting values. The Brooks-Gelman-Rubin convergence statistics (section 7.13) for the three MCMC runs are presented in figure 8.4.

Due to the practical constraints of repeating this process for all values of ρ_c, σ_1 and σ_2, a single further 50000 iterations were run, yielding acceptable

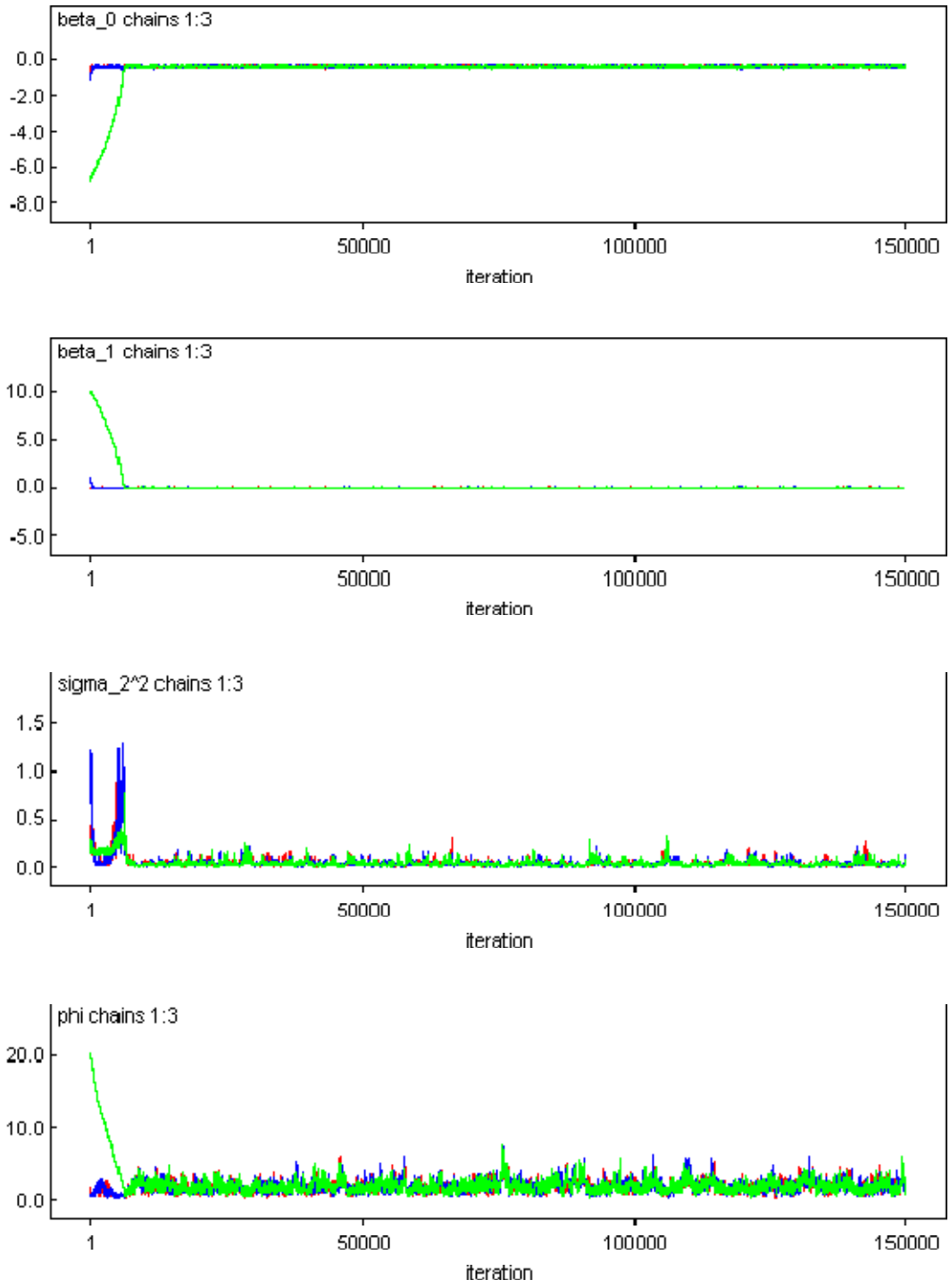


Figure 8.2: Sampled values of the parameters β_0, β_1, σ_2^2 and ϕ when fitting a spatial regression to simulated data with $\rho_c = 0.9$, $\sigma_1 = 0.3$ and $\sigma_2 = 0.1$ in three MCMC runs.

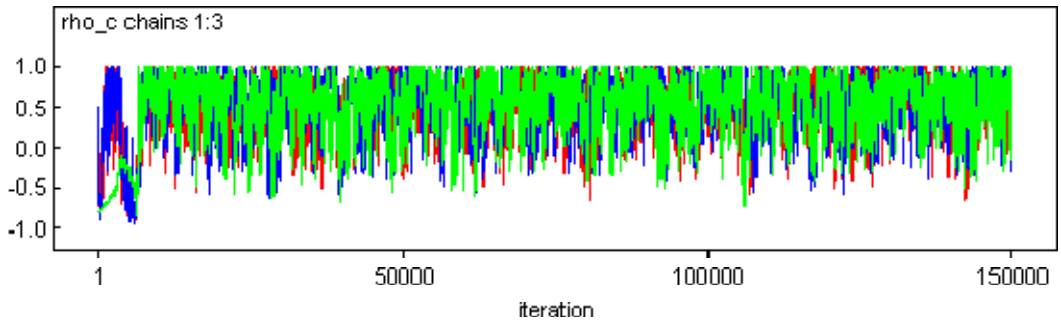


Figure 8.3: Sampled values of the parameter ρ_c when fitting a spatial regression to simulated data with $\rho_c = 0.9$, $\sigma_1 = 0.3$ and $\sigma_2 = 0.1$ in three MCMC runs.

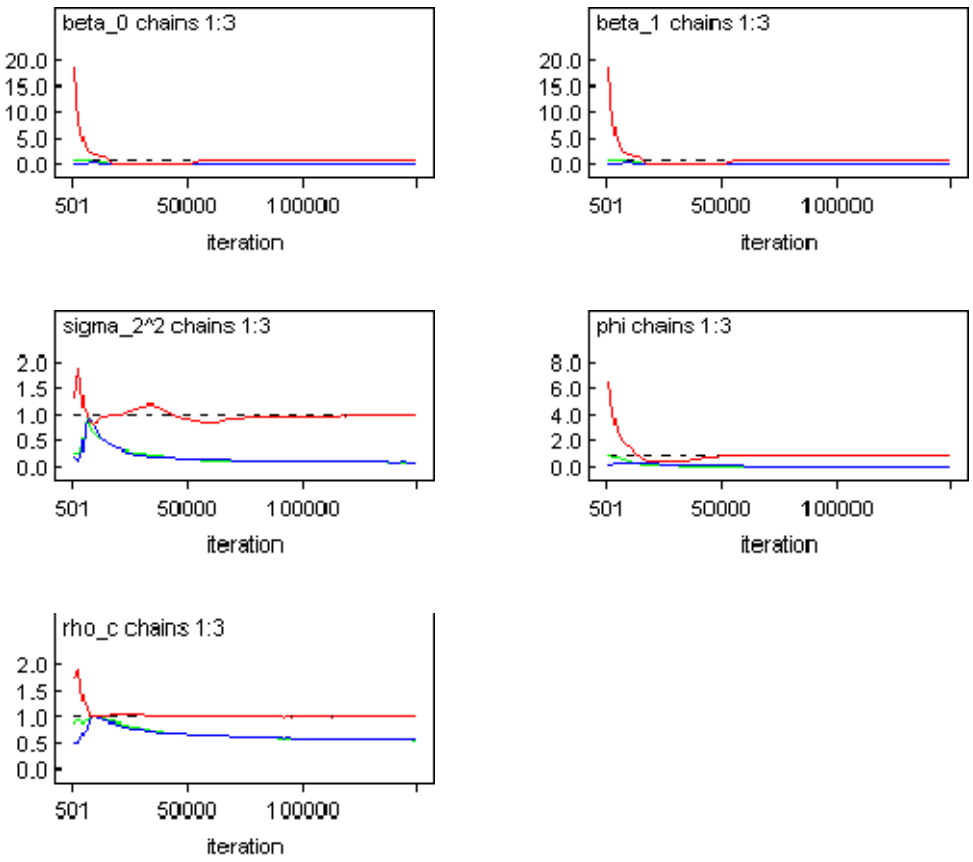


Figure 8.4: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for the parameters β_0, β_1, σ_2^2, ϕ and ρ_c when fitting a spatial regression to simulated data with $\rho_c = 0.9$, $\sigma_1 = 0.3$ and $\sigma_2 = 0.1$.

Monte Carlo errors. Posterior summaries were based on those 50000 iterations. We present in detail the results for a *single* simulated sample (section below and table 8.1). Results of other replications of these experiments are also summarized below (table 8.2).

We calculated the correlation between the simulated intercept and the simulated exposure coefficient. For the same value of ρ_c, the values of this correlation varied substantially. For the first set of simulated data (table 8.1), we chose simulated datasets for which this correlation was similar to the conditional correlation ρ_c. The objective of this practice is to restrict the potential sources of variability that may affect the results. For comparison purposes, this practice was not carried out in the additional replications (table 8.2).

It must be noted that the results are likely to depend on the underlying geography and spatial pattern of the covariate, and do not necessarily indicate any general results. The findings presented in this section are however useful for the interpretation of the results obtained in chapter 10 using the same covariate and underlying geography, but with real data.

Summary measures comparing the estimated with the true values

In order to assess the difference between the true values of the random effects and the posterior mean estimates, we used three measures. First, we calculated the sample mean of the true values as well as the sample mean of the posterior means of the random effects. The ratio of the latter to the former is displayed in table 8.1, under the column heading *mean ratio*. This measure aims at assessing how well the means of the random effects were captured. Secondly, we calculated the sample variance of the true values as well as the sample variance of the

posterior means of the random effects. Their ratio is displayed in table 8.1 under the column heading *variance ratio*, with the purpose of showing how well the spread of the random effects has been estimated. Thirdly, we calculated the correlation between the true values of the random effects and the estimated posterior means. These correlations are presented as well in table 8.1 under the column heading *Correlation*. Values near one indicate that the relation between the estimated and the true values is approximately a linear relation. All three measures were calculated for the intercept ($\beta_0 + b_{i0}$) and the slope ($\beta_1 + b_{i1}$).

Correlations between the true and estimated relative risks were high in all sets (results not shown). However, correlations between the true and estimated regression coefficients were somewhat smaller, especially as the conditional correlation ρ_c decreases and for the regression coefficient with smallest variance. On the other hand, some of the worst estimates of the mean and variance of the random effects were obtained when the correlation between the estimated and true values is actually not too far from one.

	Comparison with true values					
	Intercept			Slope		
	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
$\sigma_1 = 0.1$, $\sigma_2 = 0.3$						
$\rho_c = 0.9$	1.1	0.2	0.9	4.5	1.3	0.9
$\rho_c = 0.5$	1.1	0.6	0.8	1.4	0.6	0.7
$\rho_c = 0.1$	1.1	3.1	0.4	0.8	0.2	0.7
$\sigma_1 = 0.3$, $\sigma_2 = 0.1$						
$\rho_c = 0.9$	0.9	0.6	0.9	4.2	1.0	0.9
$\rho_c = 0.5$	1.0	0.8	0.9	1.5	1.4	0.6
$\rho_c = 0.1$	1.0	0.9	0.9	1.0	1.1	0.4

Table 8.1: Comparison between county-level posterior means and true values (one simulated data point per area).

We observed high positive correlations between the posterior means of the random intercept and those of the random slope in the data sets with $\sigma_1 = 0.1$ and $\sigma_2 = 0.3$. This may indicate that the regression coefficients are only weakly

identified and also explain why the values of *Correlation* in table 8.1 are better for larger values of ρ_c.

Turning now to the hyperparameters, the posterior means do not generally concentrate around the generated values, even when there is clear evidence of Bayesian learning on those parameters. By way of illustration, figure 8.5 displays the estimated posterior densities for the hyperparameters ρ_c and σ_2^2 for the fit to the simulated data set with $\rho_c = 0.9$, $\sigma_1 = 0.1$ and $\sigma_2 = 0.3$. These densities differ considerably from the prior densities for ρ_c and σ_2^2 and indicate that there was Bayesian learning on the respective hyperparameters.

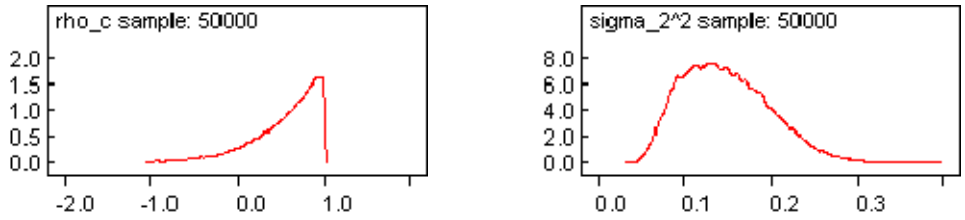


Figure 8.5: Estimates of the posterior density of selected hyperparameters ρ_c and σ_2^2 when fitting a spatial regression to the simulated data set with $\rho_c = 0.9$, $\sigma_1 = 0.1$ and $\sigma_2 = 0.3$.

Spatial pattern of the estimated versus the true values

Maps displaying the county-level posterior mean estimates of the regression coefficients are shown in figures 8.6 to 8.11, along with the respective true values. The maps of the regression coefficient estimates generally reflect the geographical pattern of the underlying values. However, there is a tendency for maps to show spurious patterns when the correlation between the two regression coefficients is small. Although the spatial patterns are broadly captured, the mean

and range of the values are sometimes not accurate. Replications of this study have lead to similar conclusions. The results of these replications are summarized in table 8.2. The findings of this section seem to indicate that while a spatial regression analysis can be useful for suggesting spatial patterns, the magnitude of the exposure and baseline risk estimates should be handled with care.

			Comparison with true values					
			Intercept			Slope		
σ_1	σ_2	ρ_c	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
0.1	0.3	0.9	0.8	0.6	0.8	0.7	0.9	0.9
			0.9	1.5	0.8	0.6	0.7	0.9
		0.5	1.1	0.4	0.9	0.7	0.9	0.8
			1.1	0.2	0.5	0.5	0.7	0.8
		0.1	1.0	0.6	0.5	1.1	0.5	0.8
			1.0	0.6	0.7	0.9	0.5	0.7
0.3	0.1	0.9	1.1	0.6	0.9	1.1	2.5	0.9
			1.1	1.1	1.0	2.1	0.4	0.8
		0.5	1.0	0.7	0.9	1.1	1.1	0.6
			0.9	0.8	0.9	1.4	0.5	0.6
		0.1	1.0	0.5	0.9	1.2	0.7	0.0
			1.0	1.1	1.0	1.4	0.1	-0.4

Table 8.2: Comparison between county-level posterior means and true values for replicated simulated data sets (one simulated data point per area).

8.6.4 Conditional and intrinsic autoregressions

In this chapter, bivectorial intrinsic autoregressions were fitted to data that are actually distributed as a bivectorial conditional autoregression. The bivectorial conditional autoregression (expression 8.6), which is a proper distribution, is certainly not the same as the improper bivectorial intrinsic autoregression (expression 8.9). However, the latter is a limiting form of the former (section 8.4). If the parameter ρ_S of the bivectorial conditional autoregression is very close to one, the generated data will be a reasonable approximation. Since ρ_S was chosen equal to 0.999, fitting intrinsic autoregressions to conditional autoregressive data should not have a considerable impact on the results. To illustrate this point, we

present here results obtained for univectorial autoregressions.

Data were simulated as described in section 8.6, but with predictors: (I) $\eta_i = \beta + b_i$; (II) $\eta_i = (\beta + b_i)x_i$ and (III) $\eta_i = (\beta + b_i)(1/x_i)$, where $\boldsymbol{b} = (b_1, \ldots, b_n)$ is distributed as a conditional autoregression (expression 8.5). The S-PLUS algorithm used for the simulation of the conditional autoregressive vector $\boldsymbol{b}$ is presented in Kaluzny *et al.* (1998). Four replications of the vector $\boldsymbol{b}$ were generated, which we will refer to as sets A, B, C and D. We use these four replications in each of the predictors I, II and III. We have thus a total of 12 simulated sets.

The three models fitted to the data are presented below. Models 1, 2 and 3 were fitted to data generated with predictors I, II and III, respectively. Since there are four replications of $\boldsymbol{b}$, each model was fitted to four data sets.

Model 1

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= \beta_0 + b_{i0} \\ b_{i0} &\sim \text{IAR}(\sigma^2) \end{aligned} \tag{8.14}$$

Model 2

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= (\beta_1 + b_{i1})x_i \\ b_{i1} &\sim \text{IAR}(\sigma^2) \end{aligned} \tag{8.15}$$

Model 3

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= (\beta_1 + b_{i1})(1/x_i) \\ b_{i1} &\sim \text{IAR}(\sigma^2) \end{aligned} \tag{8.16}$$

The results are displayed in table 8.3. The mean ratio varies between 1.0 and 1.5, the variance ratio varies between 0.7 and 1.0 and the correlation between the true and the estimated values varies between 0.9 and 1.0. In general the estimated regression coefficients are close to the true values. There is only a noticeable discrepancy between the true and estimated mean for model 2-predictor II with set C. Apart from this discrepancy, the estimations of the random intercept and the random slope in respectively models I and II were similar. The estimation of the random slope slightly improved when an inverse function of the spending power was used as a covariate (predictor III, model 3). This is possibly due to the fact that the inverse function is less skewed than the spending power itself.

These results show that the intrinsic autoregressive model captures well the conditional autoregressive data.

Set	Model 1, predictor I Intercept			Model 2, predictor II Slope			Model 3, predictor III Slope		
	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
A	1.0	0.9	0.9	1.0	0.9	0.9	1.0	0.9	1.0
B	1.0	0.9	1.0	1.0	0.8	0.9	1.0	1.0	1.0
C	1.1	0.8	0.9	1.5	0.7	0.8	1.0	0.9	1.0
D	1.0	0.7	0.9	1.0	0.8	0.9	1.0	0.9	0.9

Table 8.3: Results of the fit of models 1, 2 and 3.

8.6.5 Comparison with another model

Data were simulated using the predictor $\eta_i = \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i$, where $\boldsymbol{b}_{\bullet 0} = (b_{10}, \ldots, b_{n0})$ and $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{n1})$ are independently distributed as conditional autoregressions (expression 8.5). The S-Plus algorithm of Kaluzny et al. (1998) was used for the simulation of the conditional autoregressive vectors $\boldsymbol{b}_{\bullet 0}$ and $\boldsymbol{b}_{\bullet 1}$. Only one replication of the vector $\boldsymbol{b}_{\bullet 0}$ was generated. Four replications of the vector $\boldsymbol{b}_{\bullet 1}$ were used, which are the same as the sets A, B, C

and D of section 8.6.4.

We fitted two different models: in the first model, the intercept and the slope are assumed to be distributed as a bivectorial intrinsic autoregression (i.e. the model described in expressions 8.13 was applied); in the second model, the intercept and the slope are assumed to be independently distributed as intrinsic autoregressions:

$$\begin{aligned} Y_i &\sim \text{Poisson}(E_i \exp(\eta_i)) \\ \eta_i &= \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i \\ b_{\bullet 0} &\sim \text{IAR}(\sigma_1^2) \\ b_{\bullet 1} &\sim \text{IAR}(\sigma_1^2) \\ b_{\bullet 0} \text{ and } b_{\bullet 1} &\text{ are independent} \end{aligned} \tag{8.17}$$

	Model in which the intercept and the slope are assumed to be distributed as two independent intrinsic autoregressions					
	Intercept			Slope		
Set	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
A	1.0	1.1	1.0	0.8	0.2	0.7
B	1.0	0.8	0.9	0.9	0.8	0.8
C	1.0	0.8	0.9	5.3	0.2	0.6
D	1.0	0.7	0.9	0.8	1.2	0.7

	Model in which the intercept and the slope are assumed to be jointly distributed as a bivectorial intrinsic autoregression (model 8.13)					
	Intercept			Slope		
Set	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
A	1.0	0.9	1.0	0.8	0.4	0.8
B	1.0	0.9	0.9	0.9	0.8	0.8
C	1.1	0.8	0.9	6.2	0.3	0.6
D	1.0	0.7	0.9	0.8	1.6	0.7

Table 8.4: Results of the fit of two models: a model in which the intercept and the slope are assumed to be distributed as two independent intrinsic autoregressions, and a model in which the intercept and the slope are assumed to be jointly distributed as a bivectorial intrinsic autoregression (model 8.13).

Since there are four replications of $\boldsymbol{b}_{\bullet 1}$, each model was fitted to four data sets. Table 8.4 displays the results obtained. The two models yielded comparable

results. Therefore, the findings of this section indicate that the model 8.13 provides reasonable results, comparable to the results obtained with other well-established models using univectorial intrinsic autoregressions.

8.6.6 The role of identifiability

Identifiability should not be an issue when there are more than two data points per area. In order to study the role of identifiability, we repeated the study of section 8.6.3, table 8.1, using three and ten data points per area. The covariate values in each area were chosen to be equally spaced and within the range of the real values of the covariate. For example, for the set with three data points per area the covariate values were 0.1, 1.3 and 2.5; for the set with ten data points per area the covariate values also ranged from 0.1 to 2.5 with a difference of 0.3 between consecutive covariate values.

			Comparison with true values - 3 data points per area					
			Intercept			Slope		
σ_1	σ_2	ρ_c	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
0.1	0.3	0.9	1.0	0.9	1.0	0.4	1.0	1.0
		0.5	1.0	0.6	0.9	1.0	1.0	1.0
		0.1	1.1	0.4	0.6	1.0	0.9	1.0
0.3	0.1	0.9	1.1	1.0	1.0	0.5	0.8	1.0
		0.5	1.0	0.9	1.0	1.0	0.9	0.9
		0.1	1.1	0.9	1.0	1.0	0.7	0.8

			Comparison with true values - 10 data points per area					
			Intercept			Slope		
σ_1	σ_2	ρ_c	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
0.1	0.3	0.9	1.0	1.2	1.0	0.8	1.0	1.0
		0.5	1.0	0.8	0.9	1.0	1.0	1.0
		0.1	1.1	0.5	0.7	1.0	0.9	1.0
0.3	0.1	0.9	1.0	1.0	1.0	0.7	1.0	1.0
		0.5	1.0	1.0	1.0	1.0	0.9	0.9
		0.1	0.9	0.9	1.0	1.0	0.8	0.9

Table 8.5: Comparison between county-level posterior means and true values (three and ten simulated data points per area).

The results, which are shown in table 8.5, improved in comparison with the

results displayed in table 8.1, especially for the set with ten data points per area. This supports the view defended by some authors (Richardson and Best, 2003; Assuncão, 2003) who considered that for models in which both the intercept and the slope vary across areas, the parameters may not be identifiable when only one data point is available per area.

8.6.7 Transformation of the covariate

The data on the spending power shows a positive skew. i.e. an elongated tail at the right. For this reason, the analysis of section 8.6.3 was repeated but using the inverse function of the spending power as a covariate. The results are displayed in table 8.6. In comparison with the results presented in table 8.1 (with the spending power used directly as a covariate), the estimation of the random coefficients has improved in some instances, but is worse in others; the estimation is particularly worse for the random intercept. Therefore, the remaining analyses of this thesis will use the spending power directly as the covariate.

			Comparison with true values					
			Intercept			Slope		
			Mean	Variance		Mean	Variance	
σ_1	σ_2	ρ_c	ratio	ratio	Correlation	ratio	ratio	Correlation
0.1	0.3	0.9	0.8	0.7	0.9	-0.8	1.1	1.0
		0.5	0.9	0.4	0.8	0.9	0.9	0.9
		0.1	0.0	0.1	0.2	1.2	0.9	0.9
0.3	0.1	0.9	1.4	0.9	0.9	-0.9	1.1	0.9
		0.5	1.6	0.8	0.8	0.6	0.9	0.7
		0.1	2.0	0.7	0.9	1.5	0.3	0.5

Table 8.6: Results obtained using the inverse of the spending power as the covariate.

8.6.8 Comparison with non-spatial models

We fitted a Poisson GLM to each area using the simulated data of the section 8.6.6 above, with ten data points per area. The results are displayed in

table 8.7. A comparison with the results presented in table 8.5 above shows that the spatial model 8.13 provided better estimates of the regression coefficients than the Poisson GLMs fitted to each area.

			Poisson GLMs - Non-spatial model Comparison with true values - 10 data points per area					
			Intercept			Slope		
σ_1	σ_2	ρ_c	Mean ratio	Variance ratio	Correlation	Mean ratio	Variance ratio	Correlation
0.1	0.3	0.9	1.0	2.5	0.7	0.8	1.0	1.0
		0.5	1.1	1.8	0.8	1.0	1.1	1.0
		0.1	1.2	2.8	0.5	1.0	1.0	1.0
0.3	0.1	0.9	1.1	1.1	0.9	0.5	1.6	0.8
		0.5	1.0	1.1	1.0	1.0	1.4	0.9
		0.1	0.9	1.1	1.0	0.9	1.8	0.8

Table 8.7: Results of the fit of a Poisson GLM to each area; ten data points available per area.

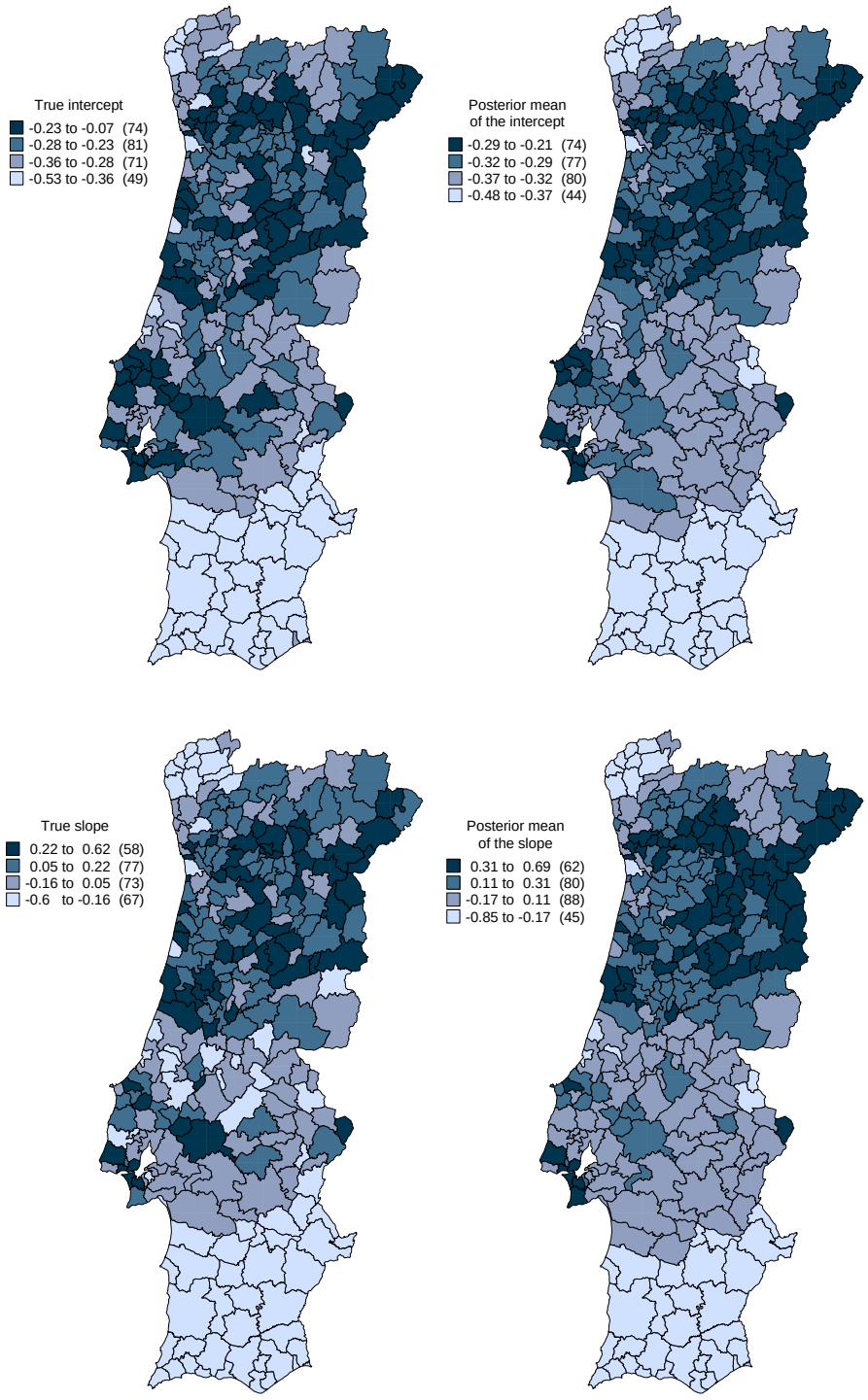


Figure 8.6: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.1, \sigma_2 = 0.3, \rho_c = 0.9$).

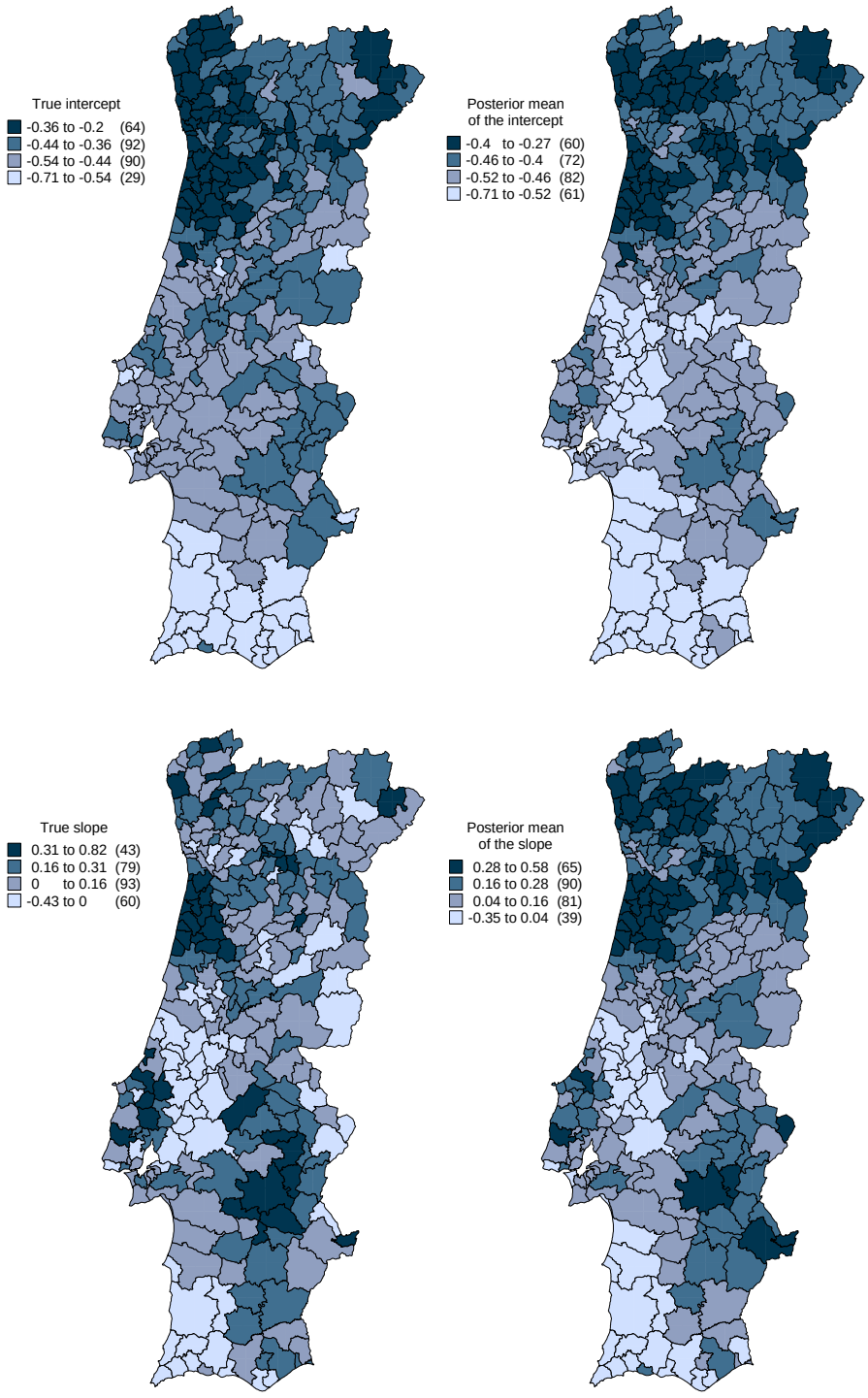


Figure 8.7: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.1, \sigma_2 = 0.3, \rho_c = 0.5$).

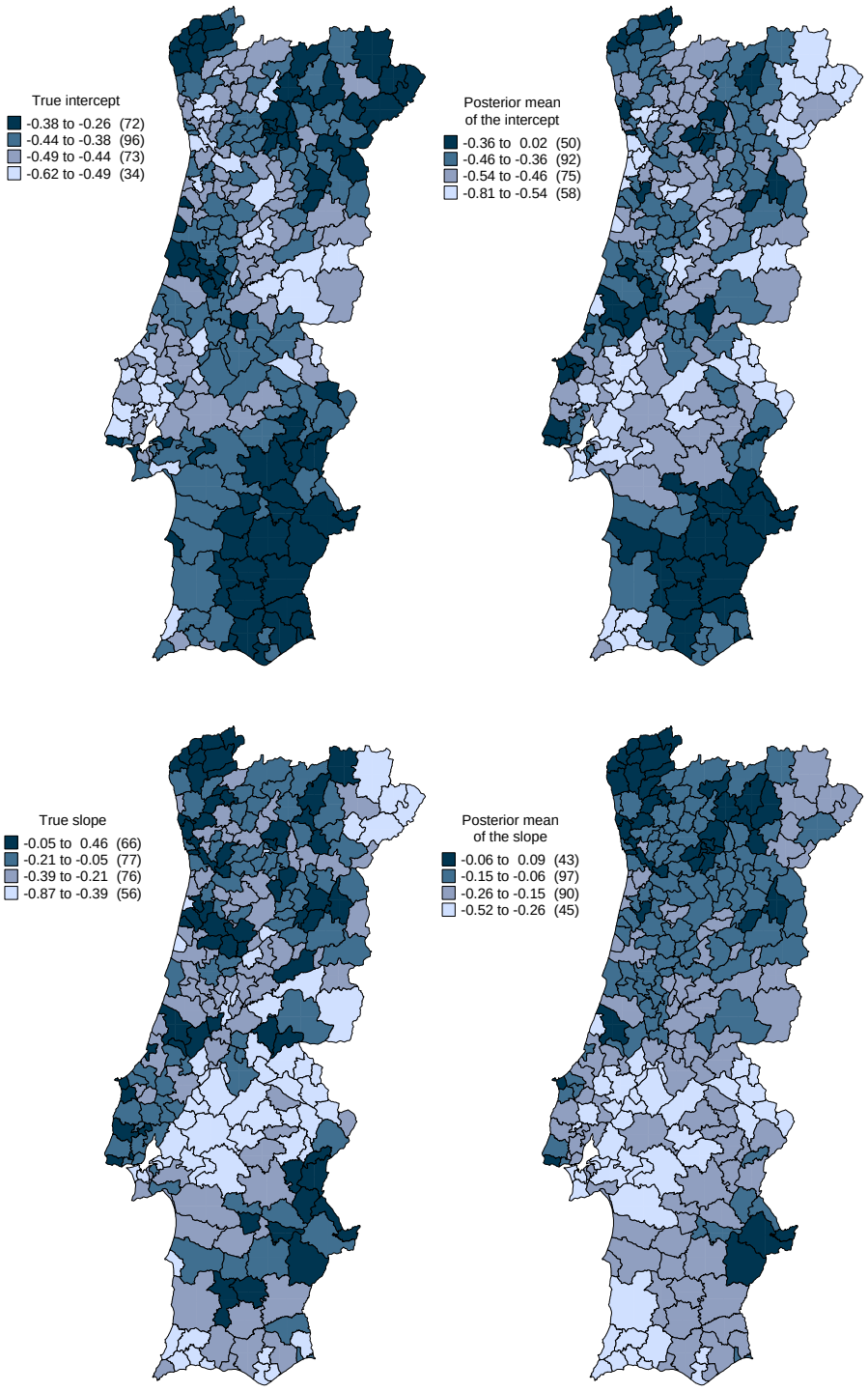


Figure 8.8: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.1, \sigma_2 = 0.3, \rho_c = 0.1$).

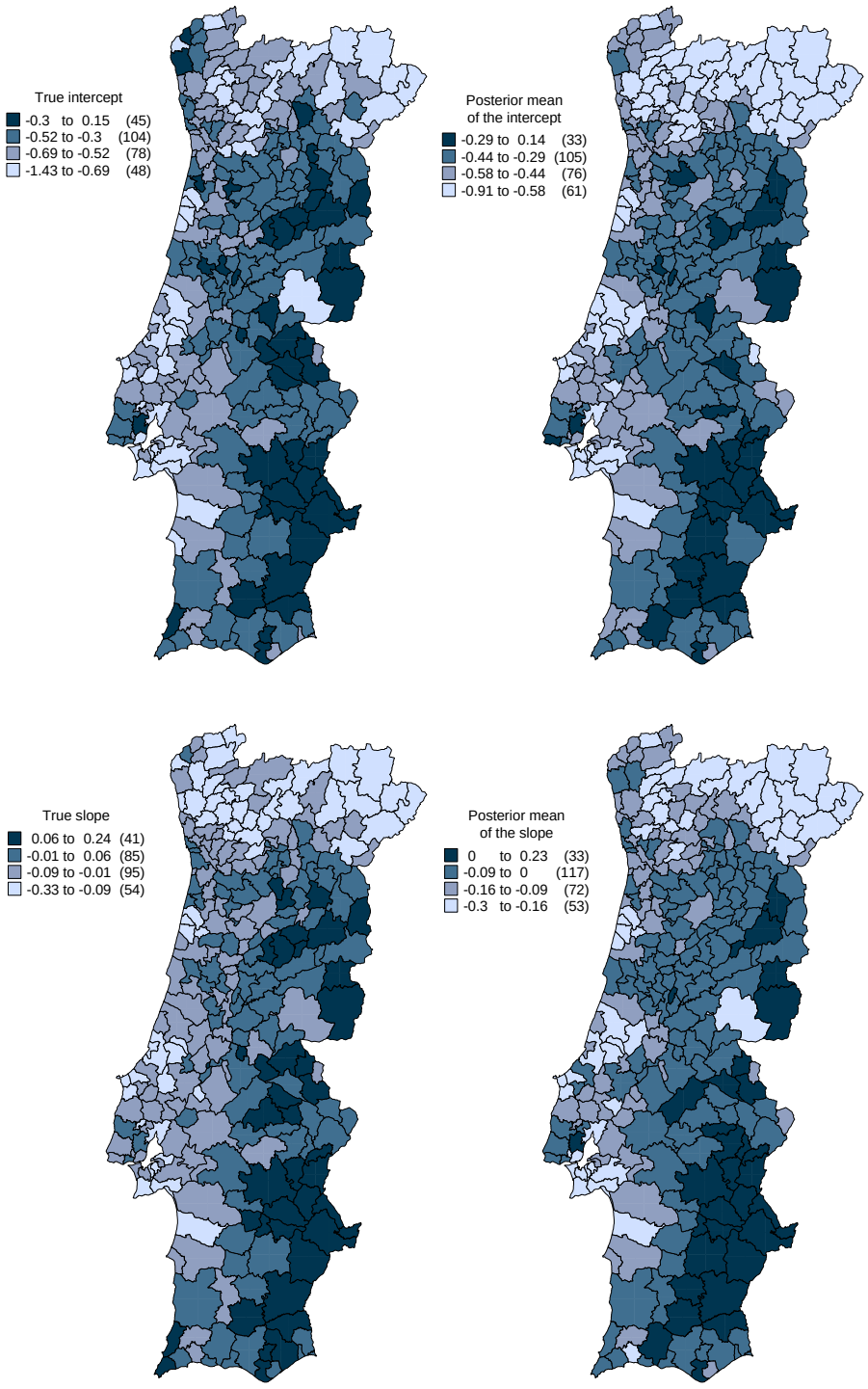


Figure 8.9: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.3, \sigma_2 = 0.1, \rho_c = 0.9$).

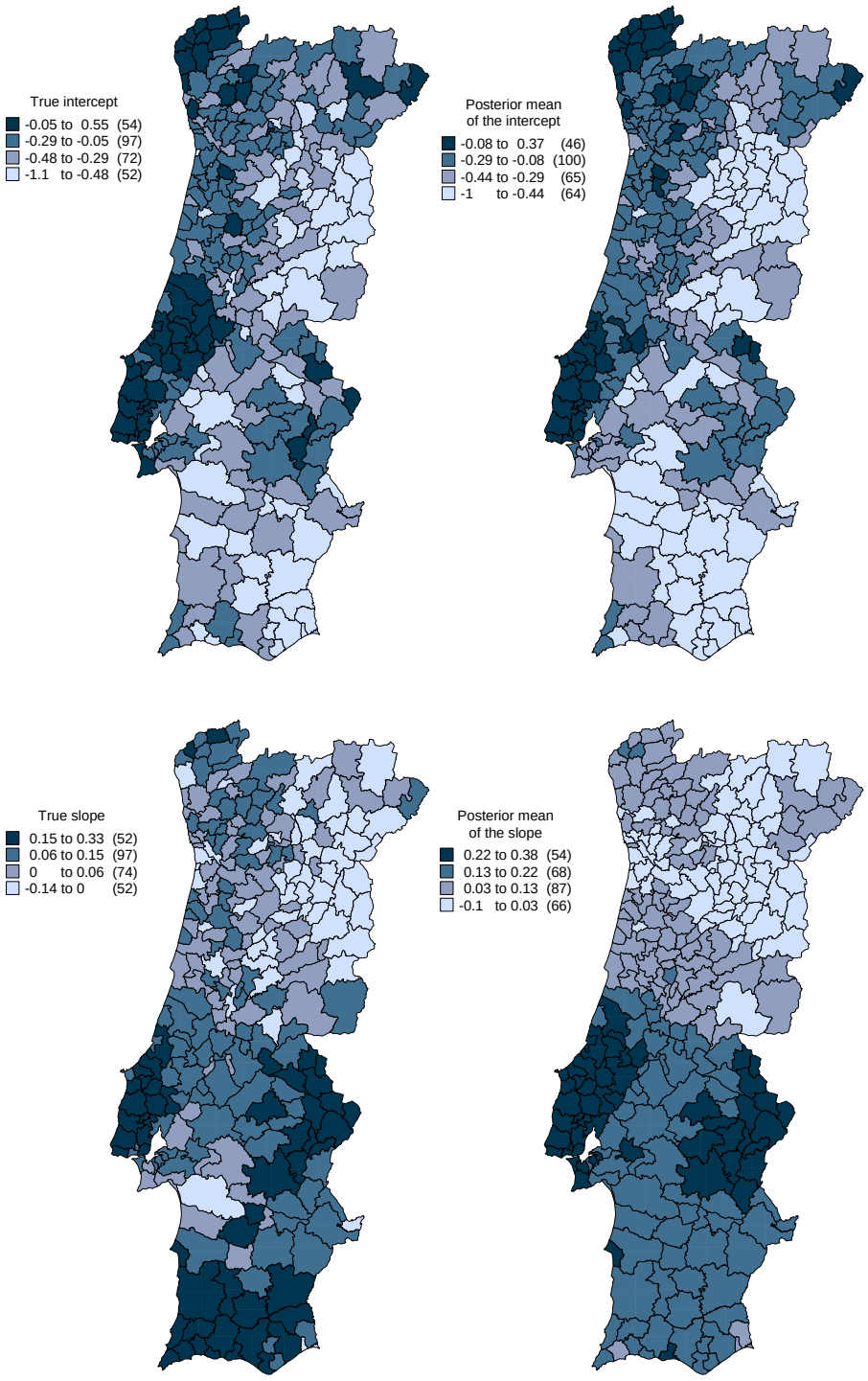


Figure 8.10: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.3, \sigma_2 = 0.1, \rho_c = 0.5$).

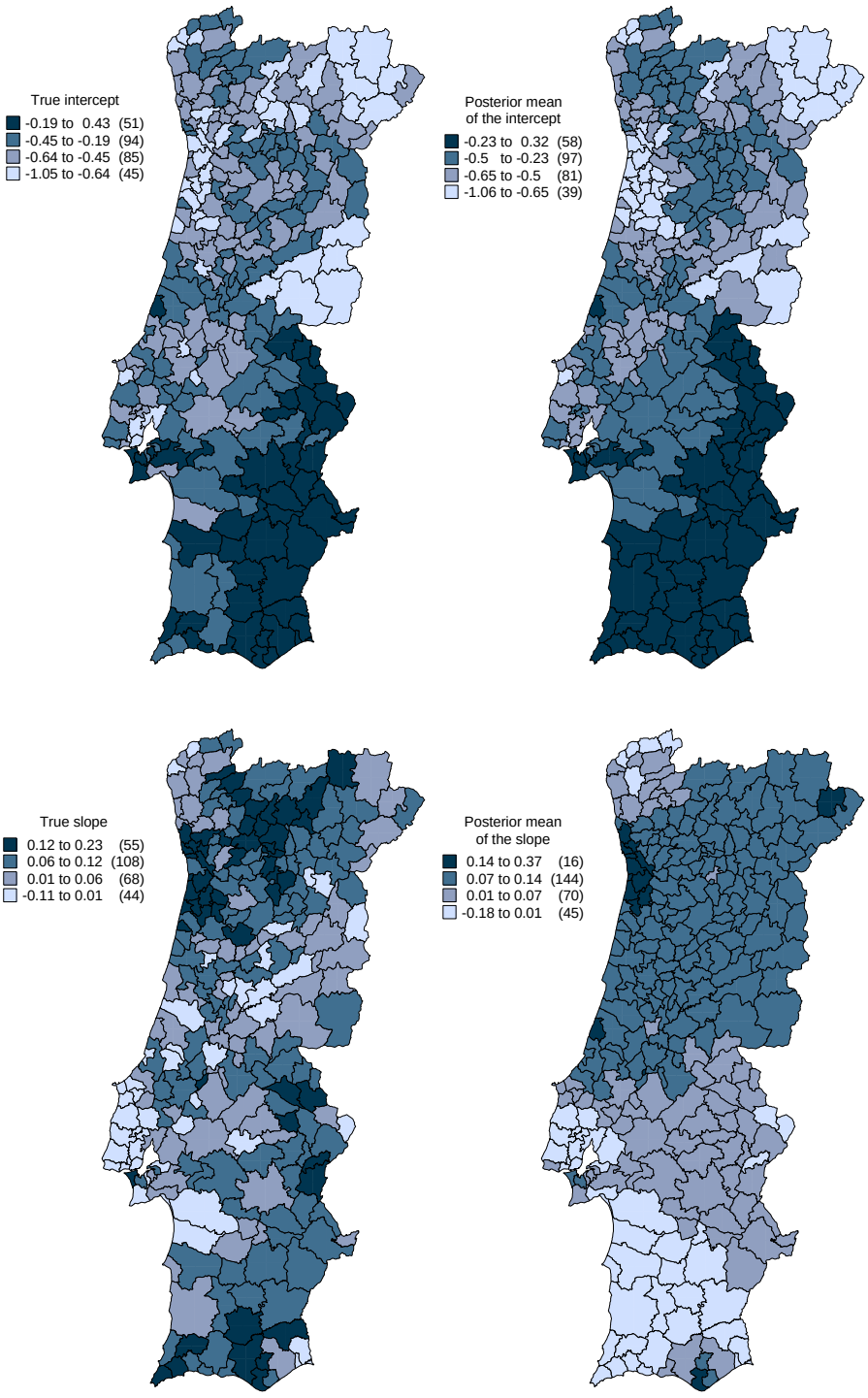


Figure 8.11: True values (left) and posterior means (right) of the regression coefficients ($\sigma_1 = 0.3, \sigma_2 = 0.1, \rho_c = 0.1$).

Chapter 9

Worldwide variation of Helicobacter pylori virulence

9.1 Introduction

The purpose of this chapter is to model the worldwide variation in the association between the *H. pylori* infection and stomach cancer rates. To accomplish this goal, worldwide national data on the prevalence of the *H. pylori* infection and stomach cancer incidence rates are analysed.

Studies conducted on the genetic variability of *H. pylori* suggest that significant differences exist in the strains between countries (chapter 3). Thus provided we can assume that all human populations are equally susceptible to both *H. pylori* infection and its putative carcinogenic effects, the relation between the prevalence of the *H. pylori* infection and the incidence or mortality of stomach cancer in each country may provide estimates of the virulence of the bacteria worldwide (chapter 4). The assumption of equal susceptibility is perhaps not very plausible, but this affects only the interpretation of the analyses. In the interests

of simplicity of description, we assume that the assumption is true and discuss the variation of virulence, when in fact we may wish to interpret the results as variation in the effect of different strains of virus on different ethnic groups with possibly different susceptibilities.

However, other factors are likely to influence the national stomach cancer rates. If these factors are unknown or unavailable, the resulting extra variability of the stomach cancer rates can mask the relation between the prevalence of the *H. pylori* infection and the stomach cancer rates. Since the virulence of the bacteria appears to cluster geographically, the geographical structure of countries can be used to improve the estimates of the virulence of *H. pylori*.

In chapter 7, we discussed statistical methodology which may be used to find clusters of similar exposure coefficients. Geographical location was used to define the prior distribution of the clusters. In this chapter, we further discuss the applicability of this methodology to the data under study.

The models should be able to describe the dependence of stomach cancer rates on the prevalence of *H. pylori* as well as assessing the geographical patterns of the association between these two variables. Once a model is selected it can be used to estimate the virulence of *H. pylori*, under the assumptions of homogeneous susceptibility.

9.2 Description of the data and initial analysis

Estimates of national prevalence of *H. pylori* infection and estimates of world age- and sex-standardized incidence rates (WASR) for 58 countries worldwide were extracted from a published study (Lunet and Barros, 2003). The data set is reproduced in appendix K. The two variables are lagged in time to allow for a

delay between putative exposure and disease. Most of the data on the prevalence of *H. pylori* infection were collected in the 1990s. The estimates of WASRs are projections for the year 2000. The cancer data used in the estimation are from periods 3-10 years earlier. Details on the calculation of these estimates can be found in Parkin et al. (2001) and Lunet and Barros (2003).

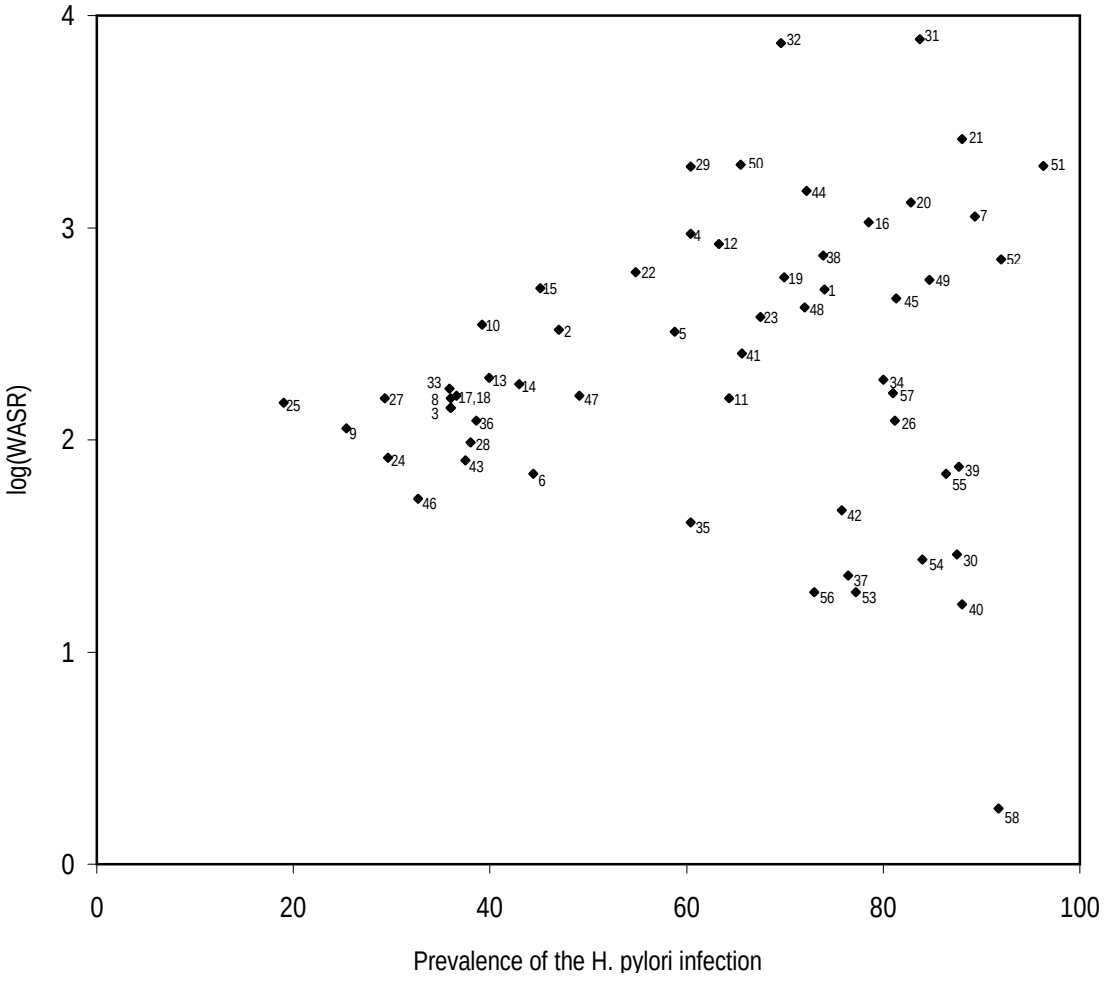
The rates of incidence of stomach cancer incidence show great differences worldwide, with national WASRs in the data set ranging from 1.30 to 48.9. The prevalence of *H. pylori* varies from 19% to 96.3%.

Figure 9.1 displays a scatterplot of the log-WASRs against the prevalence of the *H. pylori* infection. Despite the large variability of the points in this graph, the observation of Bangladesh seems to be an outlier. A check on the source of data indicated that the *H. pylori* infection prevalence was estimated from a very specific sample of the population, men undergoing a check-up to work as immigrant workers. There is a possibility that this group is not representative of the overall socio-economic status in the country. Since the prevalence of the *H. pylori* infection depends on the socio-economic status (chapter 3), we excluded this observation from our study.

An analysis of the whole data set was conducted in Lunet and Barros (2003), in which a linear regression of the log transformed stomach cancer incidence rates on the prevalence of the *H. pylori* infection yielded weak correlations ($r^2 = 0.01$).

Model	Intercept	Risk coefficient	r^2	Residual standard deviation	Number of observations
Europe	1.6	$0.017\,(< 0.01)$	0.69	0.23	28
America	1.3	$0.020\,(< 0.01)$	0.65	0.35	9
57 countries worldwide	2.0	$0.006\,(0.13)$	0.04	0.62	57

Table 9.1: Linear regression with log-WASR for stomach cancer incidence as the response variable and prevalence of the *H. pylori* infection as the explanatory variable (p-values are indicated in brackets).



1 Albania	13 Iceland	25 Switzerland	37 Thailand	49 Brazil
2 Austria	14 Ireland	26 Turkey	38 Vietnam	50 Chile
3 Belgium	15 Italy	27 UK	39 Algeria	51 Colombia
4 Croatia	16 Lithuania	28 Australia	40 Egypt	52 Venezuela
5 Czech Rep.	17 Netherlands	29 China	41 Israel	53 Côte d'Ivoire
6 Denmark	18 Norway	30 India	42 Saudi Arabia	54 Nigeria
7 Estonia	19 Poland	31 Japan	43 Canada	55 South Africa
8 Finland	20 Portugal	32 Korea, Rep. of	44 Jamaica	56 Sudan
9 France	21 Russian Fed.	33 Malaysia	45 Mexico	57 Zambia
10 Germany	22 Slovenia	34 Myanmar	46 United States	58 Bangladesh
11 Greece	23 Spain	35 Nepal	47 Argentina	
12 Hungary	24 Sweden	36 New Zealand	48 Barbados	

Figure 9.1: Scatterplot of stomach cancer log-WASR against the prevalence of *H. pylori* infection.

We tried an analysis by continent, which showed that stronger associations between *H. pylori* and stomach cancer exist in Europe and in America (table 9.1). The fact that the regressions in some continents yielded strong correlations supports the hypothesis that geography is indeed playing a role.

In addition, the range of the log-WASR appears to increase with an increasing prevalence of the *H. pylori* infection (figure 9.1). This visual display is typical of a regression mixture in which the regression lines have different slopes and cross at some determined point. By way of illustration, figure 9.2 shows 116 data points simulated according to a mixture of two regression lines with different slopes and common intercept. In the same way as in figure 9.1, the response variable seems to increase with an increasing value of the covariate.

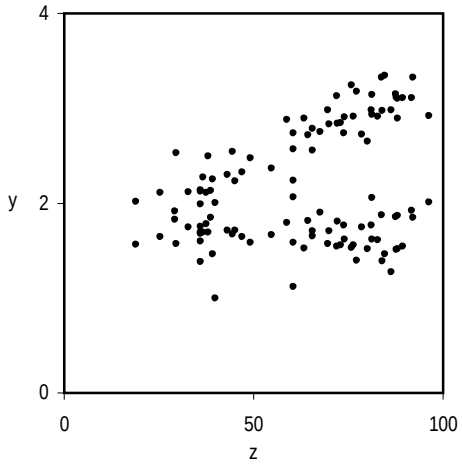


Figure 9.2: Data simulated from a mixture of two regressions (intercept, 1.64; slopes: 0.0019 and 0.0172; standard error, 0.2; covariate, prevalence of *H. pylori* infection given in appendix K).

To sum up, the preliminary analysis presented hitherto in this section suggests the presence of a mixture of regression lines with different slopes in figure 9.1. As a result, we decided that a linear model was inappropriate for these data and we went on to investigate a mixture model. Initially, we did not incorporate any

spatial structure in the model (section 9.3). Subsequently, we introduced spatial information via a prior distribution (section 9.4).

9.3 Regression mixtures

Mixtures of Gaussian distributions are used here to assess the geographical variation of the virulence of *H. Pylori*. We postulate the Gaussian model for the log-WASR described in chapter 6 and given in detail in expression 6.1. The responses $Y_1, \ldots, Y_n$ are the observed log-WASRs, and the covariates $z_1, \ldots, z_n$ are the observed prevalences of the *H. pylori* infection. Each component in the mixture is identified by a different value of the covariate coefficient b_{m1}. Thus, the vector of covariate coefficients $\boldsymbol{b}_{\bullet 1} = (b_{11}, \ldots, b_{M1})$ models the virulence of *H. pylori*. Since we aim at identifying groups of constant virulence, we will assume that M is known and fit a model for different values of M.

Prior distributions

Taking into account the discussions of sections 6.3 and 6.4, we use the following prior distributions for the regression parameters,

$$\beta_0 \sim \mathrm{N}(0, 10^3);$$

$$b_{11} \sim \mathrm{N}(0, 10^3) I(0, +\infty);$$

$$b_{1,m+1} = b_{1,m} + \delta_m, \delta_m > 0, m = 1, \ldots, M-1;$$

$$\tau = \frac{1}{\sigma^2} \sim \mathrm{Gamma}(10^3, 10^3).$$

The choice of the parameters of the prior distributions aimed at providing the least informative priors possible. The prior distributions of the slopes are

truncated to non-negative values because the *H. pylori* is an established risk factor. The prior for the parameters δ_m is chosen to be a normal distribution with mean 0 and variance 10^3, truncated to the interval $(0, +\infty)$.

9.3.1 Results of the regression mixtures

Bayesian estimation of mixtures can be carried out using the Gibbs sampler, which we use in fitting a regression mixture to the world data set. The computation is undertaken in Winbugs 1.4 (the code is provided in appendix F). The Gibbs sampler was run for 20000 iterations for $M = 2$ and $M = 3$.

$M = 2$

To obtain a well identified solution for $M = 2$, causing no numerical problems as we iterate to a run length of 20000, prior parameters for the Dirichlet distribution with $\alpha_1 = \alpha_2 = 2$ were taken (expression 6.6). The first 10000 sample points were discarded as burn-in, and estimates were made using the remaining 10000 sample points. The sampled values of $(\beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2, \boldsymbol{p})$ are shown in Table 9.2. Different slopes are clearly identified.

		Standard	Quantiles		
	Mean	deviation	2.5%	Median	97.5%
β_0	1.64	0.16	1.33	1.64	1.96
b_{11}	0.002	0.002	0.000	0.002	0.006
b_{21}	0.017	0.002	0.012	0.017	0.022
σ	0.39	0.06	0.30	0.38	0.53
p_1	0.33	0.10	0.17	0.32	0.57
p_2	0.67	0.10	0.43	0.68	0.83

Table 9.2: Summary measures of the posterior probabilities of the regression parameters when applying a Gibbs sampler to fit a regression mixture of two normal distributions.

We monitored the posterior mean of the probabilities of membership given by equation 6.11. We also monitored the proportions of iterations in which each

observation is allocated to each group (section 7.11). The difference between these two sets of estimates of the probabilities of membership was at most 0.006. For consistency with the results of the forthcoming application of the clustering model (section 9.4), we present in table 9.3 the latter set, i.e. the proportions of iterations in which each observation is allocated to each group.

$p(1,1)$	0.054	$p(1,31)$	0.001	$p(2,1)$	0.946	$p(2,31)$	0.999
$p(1,2)$	0.083	$p(1,32)$	0.001	$p(2,2)$	0.917	$p(2,32)$	0.999
$p(1,3)$	0.223	$p(1,33)$	0.172	$p(2,3)$	0.777	$p(2,33)$	0.828
$p(1,4)$	0.017	$p(1,34)$	0.553	$p(2,4)$	0.983	$p(2,34)$	0.447
$p(1,5)$	0.096	$p(1,35)$	0.931	$p(2,5)$	0.904	$p(2,35)$	0.069
$p(1,6)$	0.564	$p(1,36)$	0.263	$p(2,6)$	0.437	$p(2,36)$	0.737
$p(1,7)$	0.015	$p(1,37)$	0.995	$p(2,7)$	0.985	$p(2,37)$	0.005
$p(1,8)$	0.196	$p(1,38)$	0.027	$p(2,8)$	0.804	$p(2,38)$	0.973
$p(1,9)$	0.243	$p(1,39)$	0.975	$p(2,9)$	0.757	$p(2,39)$	0.025
$p(1,10)$	0.079	$p(1,40)$	0.998	$p(2,10)$	0.921	$p(2,40)$	0.002
$p(1,11)$	0.400	$p(1,41)$	0.175	$p(2,11)$	0.600	$p(2,41)$	0.825
$p(1,12)$	0.022	$p(1,42)$	0.978	$p(2,12)$	0.979	$p(2,42)$	0.022
$p(1,13)$	0.152	$p(1,43)$	0.410	$p(2,13)$	0.848	$p(2,43)$	0.590
$p(1,14)$	0.171	$p(1,44)$	0.009	$p(2,14)$	0.830	$p(2,44)$	0.991
$p(1,15)$	0.049	$p(1,45)$	0.083	$p(2,15)$	0.951	$p(2,45)$	0.917
$p(1,16)$	0.014	$p(1,46)$	0.509	$p(2,16)$	0.986	$p(2,46)$	0.491
$p(1,17)$	0.188	$p(1,47)$	0.240	$p(2,17)$	0.813	$p(2,47)$	0.760
$p(1,18)$	0.217	$p(1,48)$	0.083	$p(2,18)$	0.783	$p(2,48)$	0.917
$p(1,19)$	0.039	$p(1,49)$	0.065	$p(2,19)$	0.961	$p(2,49)$	0.935
$p(1,20)$	0.012	$p(1,50)$	0.006	$p(2,20)$	0.989	$p(2,50)$	0.994
$p(1,21)$	0.004	$p(1,51)$	0.005	$p(2,21)$	0.996	$p(2,51)$	0.995
$p(1,22)$	0.034	$p(1,52)$	0.048	$p(2,22)$	0.966	$p(2,52)$	0.952
$p(1,23)$	0.086	$p(1,53)$	0.996	$p(2,23)$	0.915	$p(2,53)$	0.004
$p(1,24)$	0.327	$p(1,54)$	0.996	$p(2,24)$	0.673	$p(2,54)$	0.004
$p(1,25)$	0.204	$p(1,55)$	0.978	$p(2,25)$	0.797	$p(2,55)$	0.022
$p(1,26)$	0.848	$p(1,56)$	0.995	$p(2,26)$	0.152	$p(2,56)$	0.005
$p(1,27)$	0.186	$p(1,57)$	0.679	$p(2,27)$	0.814	$p(2,57)$	0.321
$p(1,28)$	0.338			$p(2,28)$	0.662		
$p(1,29)$	0.007			$p(2,29)$	0.993		
$p(1,30)$	0.997			$p(2,30)$	0.003		

Table 9.3: Posterior probabilities of membership when applying a Gibbs sampler to fit a regression mixture of two normal distributions.

Using the allocation rule given in expression 6.10, 15 and 42 observations were attributed respectively to the groups with the lowest and the highest exposure coefficient. However, five of these observations were not very clearly allocated (i.e. the probabilities of membership were close to 0.5). The larger group has the highest value of the exposure coefficient ($b_{21} = 0.017$), which is almost ten times as large as the exposure coefficient of the smaller group.

We now decribe in detail the allocation of countries to the two groups. Observations 30, 35, 37, 39, 40, 42, 53, 54, 55 and 56 (India, Nepal, Thailand, Algeria, Egypt, Saudi Arabia, Côte d'Ivoire, Nigeria, South Africa, Sudan) all have a probability over 0.9 of belonging to the group with the lower exposure coefficient (group 1). Observation 26 (Turkey) has a probability of 0.85 of belonging to this group, and observation 57 (Zambia) a probability of 0.68. Observations 6, 11, 34, 43 and 46 (Denmark, Greece, Myanmar, Canada and the United States), with probabilities between 0.4 and 0.6, are not clearly associated with either group. As might be expected, these observations are on the borderline between the two groups (figure 9.3).

Observations 1, 2, 4, 5, 7, 10, 12, 15, 16, 19, 20, 21, 22, 23, 29, 31, 32, 38, 44, 45, 48, 49, 50, 51 and 52 (Albania, Austria, Croatia, Czech Republic, Estonia, Germany, Hungary, Italy, Lithuania, Poland, Portugal, Russia, Slovenia, Spain, China, Japan, Republic of Korea, Vietnam, Jamaica, Mexico, Barbados, Brazil, Chile, Colombia and Venezuela) all have probabilities over 0.9 of belonging to the group with the higher exposure coefficient (group 2). Observations 3, 8, 9, 13, 14, 17, 18, 24, 25, 27, 28, 33, 36, 41 and 47 (Belgium, Finland, France, Iceland, Ireland, Netherlands, Norway, Sweden, Switzerland, United Kingdom, Australia, Malaysia, New Zealand, Israel, Argentina) all have probability between 0.6 and 0.9 of belonging to the group with the higher exposure coefficient.

Label switching and trapping states in practice

The analysis so far involved fitting a mixture of two normal distributions. First, we have run the chain without any constraint on the slopes. Throughout the simulations, the parameters of both groups were interchanged, and yielded the

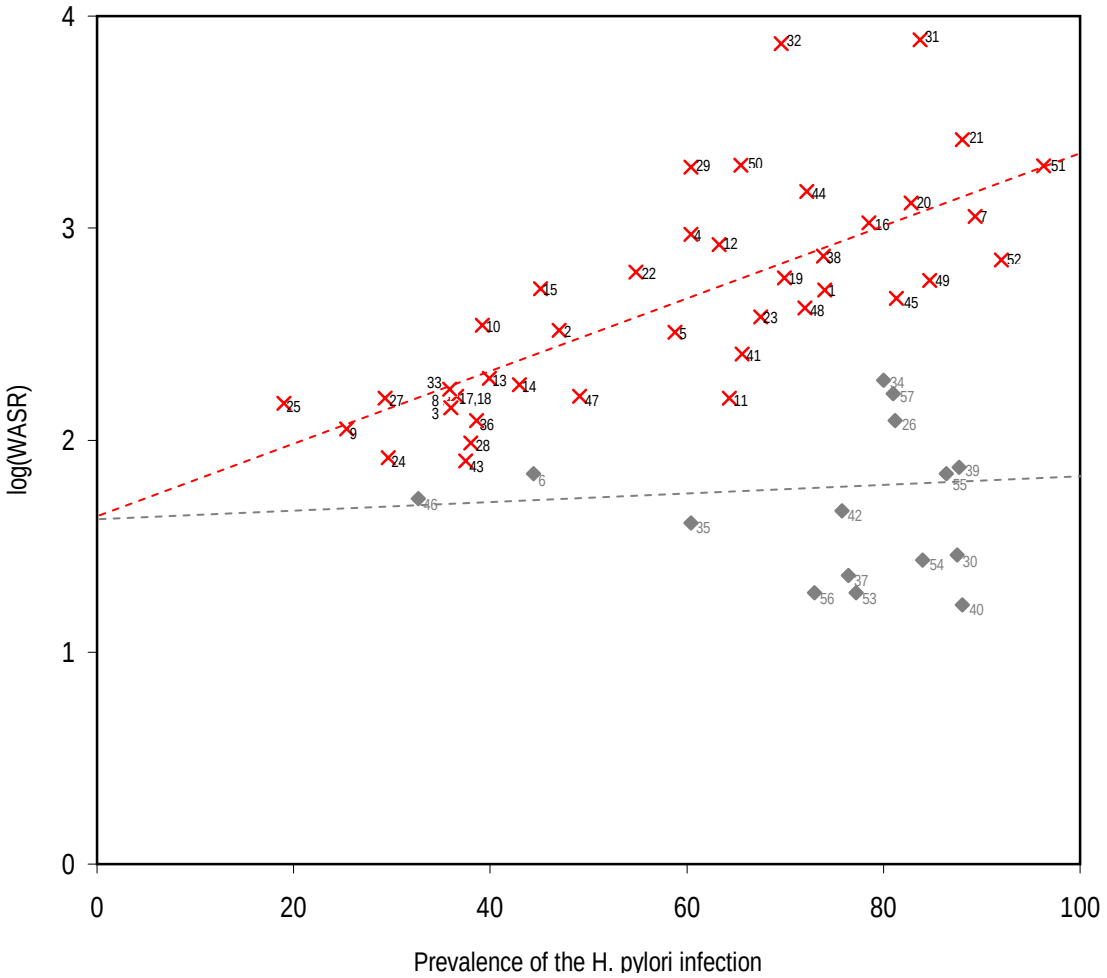


Figure 9.3: Scatterplot of stomach cancer log-WASR against the prevalence of *H. pylori* infection showing the regression lines estimated by a regression mixture of two normals and identifying observations in each group (group 1 in grey, group 2 in red, and country codes as in figure 9.1).

summary measures of the posterior distributions invalid. The constraint on the slopes described in expression 6.9 solved this problem.

Secondly, we have tried to run the simulation with the identifiability constraint and the non-informative Dirichlet$(1, 1)$ prior on the allocation probabilities. However, the chain moved at some point to a trapped state, in which all

observations were allocated to one group. As expected from the analysis described in section 6.4.3, we managed to avoid this by putting more information on the prior of the p_i. Instead of a non-informative Dirichlet$(1, 1)$, we chose a Dirichlet $(2, 2)$ thus giving density zero to the extremes of the interval $(0, 1)$.

$M = 3$

We tried to increase the number of groups to $M = 3$. To avoid the label switching problem we used the constraint on the slopes (expression 6.9). We initially set the prior distribution for the probabilities of membership $\boldsymbol{p}$ as in section 6.3.1, with $\alpha_1 = \alpha_2 = \alpha_3 = 2$. However, under this assumption, the observations fell into only two groups. By increasing the parameters of the Dirichlet distribution to $\alpha_1 = \alpha_2 = \alpha_3 = 5$, the prior probability of very small values of the ps decreases. As a result, three groups emerged, but a considerable number of observations were not clearly associated with any group. The sampled values of $(\beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2, \boldsymbol{p})$ are summarized in Table 9.4. The probabilities of membership of the groups for each observation are displayed in table 9.5. Again, we displayed the relative simulation time in the interests of consistency, the differences between these and the values obtained from equation 6.11 being never greater than 0.013.

		Standard	Quantiles		
	Mean	deviation	2.5%	Median	97.5%
β_0	1.58	0.15	1.30	1.58	1.87
b_{11}	0.002	0.002	0.000	0.001	0.006
b_{21}	0.015	0.003	0.007	0.016	0.020
b_{31}	0.022	0.004	0.015	0.022	0.030
σ	0.34	0.06	0.25	0.33	0.48
p_1	0.28	0.07	0.16	0.28	0.44
p_2	0.41	0.13	0.16	0.42	0.64
p_3	0.30	0.12	0.11	0.29	0.57

Table 9.4: Summary measures of the posterior probabilities of the regression parameters when applying a Gibbs sampler to fit a regression mixture of three normal distributions.

Figure 9.4 displays the fitted regression lines and the allocation of each coun-

try to a group. Group 1 consists of 14 countries, group 2 of 32 countries, and group 3 of 11 countries. The group with the lowest exposure coefficient coincides approximately with the group with the lowest exposure coefficient identified when $M = 2$. The exposure coefficient of the second group is also almost ten times the exposure of group 1, as seen for $M = 2$. Further comparing the results with the case $M = 2$, we also observe that the group 3 is a subset of the group 2. The difference between the exposure coefficients of groups 2 and 3 is also not as pronounced as the difference between the exposure coefficients of groups 1 and 2.

In the groups formed, it is possible to visualize some tendency for geographical proximity, but the groups are not fully consistent with the existing knowledge on the geographical variation of the strains and virulence of *H. pylori*. For example, it would be difficult to find an explanation for the allocation of Denmark into a different group from the rest of the European countries.

9.4 Clustering models

Since the similarity of populations of *H. pylori* appear to depend on geographical proximity, we decided to apply a clustering model.

In the clustering model, the clusters are formed on the basis of a distance between any two areas. One option would be to use the *boundary distance*, which depends only on adjacency (section 7.3.1). However, for the countries included in the data set, the size of the countries is so variable that the notion of adjacency may be meaningless. We used instead the *geographical distance* defined in section 7.3.1. This distance corresponds to the minimal length, in kilometres, of all paths from a region to another. As defined in section 7.1, a

$p(1,1)$	0.02	$p(1,30)$	0.99	$p(2,1)$	0.71	$p(2,30)$	0.01	$p(3,1)$	0.27	$p(3,30)$	0.00
$p(1,2)$	0.03	$p(1,31)$	0.00	$p(2,2)$	0.52	$p(2,31)$	0.11	$p(3,2)$	0.45	$p(3,31)$	0.89
$p(1,3)$	0.14	$p(1,32)$	0.00	$p(2,3)$	0.52	$p(2,32)$	0.09	$p(3,3)$	0.34	$p(3,32)$	0.91
$p(1,4)$	0.01	$p(1,33)$	0.10	$p(2,4)$	0.37	$p(2,33)$	0.51	$p(3,4)$	0.62	$p(3,33)$	0.39
$p(1,5)$	0.04	$p(1,34)$	0.39	$p(2,5)$	0.64	$p(2,34)$	0.55	$p(3,5)$	0.32	$p(3,34)$	0.06
$p(1,6)$	0.51	$p(1,35)$	0.93	$p(2,6)$	0.38	$p(2,35)$	0.06	$p(3,6)$	0.11	$p(3,35)$	0.01
$p(1,7)$	0.00	$p(1,36)$	0.18	$p(2,7)$	0.70	$p(2,36)$	0.54	$p(3,7)$	0.29	$p(3,36)$	0.28
$p(1,8)$	0.11	$p(1,37)$	0.99	$p(2,8)$	0.53	$p(2,37)$	0.01	$p(3,8)$	0.36	$p(3,37)$	0.00
$p(1,9)$	0.16	$p(1,38)$	0.01	$p(2,9)$	0.47	$p(2,38)$	0.64	$p(3,9)$	0.37	$p(3,38)$	0.35
$p(1,10)$	0.03	$p(1,39)$	0.96	$p(2,10)$	0.43	$p(2,39)$	0.03	$p(3,10)$	0.54	$p(3,39)$	0.00
$p(1,11)$	0.28	$p(1,40)$	1.00	$p(2,11)$	0.60	$p(2,40)$	0.00	$p(3,11)$	0.12	$p(3,40)$	0.00
$p(1,12)$	0.01	$p(1,41)$	0.10	$p(2,12)$	0.46	$p(2,41)$	0.70	$p(3,12)$	0.53	$p(3,41)$	0.20
$p(1,13)$	0.09	$p(1,42)$	0.97	$p(2,13)$	0.54	$p(2,42)$	0.02	$p(3,13)$	0.37	$p(3,42)$	0.00
$p(1,14)$	0.10	$p(1,43)$	0.35	$p(2,14)$	0.56	$p(2,43)$	0.47	$p(3,14)$	0.33	$p(3,43)$	0.19
$p(1,15)$	0.02	$p(1,44)$	0.00	$p(2,15)$	0.39	$p(2,44)$	0.38	$p(3,15)$	0.59	$p(3,44)$	0.61
$p(1,16)$	0.01	$p(1,45)$	0.04	$p(2,16)$	0.58	$p(2,45)$	0.78	$p(3,16)$	0.41	$p(3,45)$	0.19
$p(1,17)$	0.11	$p(1,46)$	0.48	$p(2,17)$	0.53	$p(2,46)$	0.38	$p(3,17)$	0.36	$p(3,46)$	0.15
$p(1,18)$	0.14	$p(1,47)$	0.15	$p(2,18)$	0.54	$p(2,47)$	0.62	$p(3,18)$	0.33	$p(3,47)$	0.23
$p(1,19)$	0.01	$p(1,48)$	0.03	$p(2,19)$	0.64	$p(2,48)$	0.72	$p(3,19)$	0.35	$p(3,48)$	0.24
$p(1,20)$	0.00	$p(1,49)$	0.02	$p(2,20)$	0.58	$p(2,49)$	0.78	$p(3,20)$	0.42	$p(3,49)$	0.19
$p(1,21)$	0.00	$p(1,50)$	0.00	$p(2,21)$	0.41	$p(2,50)$	0.22	$p(3,21)$	0.58	$p(3,50)$	0.78
$p(1,22)$	0.01	$p(1,51)$	0.00	$p(2,22)$	0.44	$p(2,51)$	0.64	$p(3,22)$	0.55	$p(3,51)$	0.35
$p(1,23)$	0.04	$p(1,52)$	0.02	$p(2,23)$	0.70	$p(2,52)$	0.80	$p(3,23)$	0.27	$p(3,52)$	0.19
$p(1,24)$	0.25	$p(1,53)$	0.99	$p(2,24)$	0.48	$p(2,53)$	0.01	$p(3,24)$	0.27	$p(3,53)$	0.00
$p(1,25)$	0.13	$p(1,54)$	0.99	$p(2,25)$	0.44	$p(2,54)$	0.01	$p(3,25)$	0.43	$p(3,54)$	0.00
$p(1,26)$	0.78	$p(1,55)$	0.97	$p(2,26)$	0.20	$p(2,55)$	0.03	$p(3,26)$	0.02	$p(3,55)$	0.00
$p(1,27)$	0.10	$p(1,56)$	0.99	$p(2,27)$	0.48	$p(2,56)$	0.01	$p(3,27)$	0.42	$p(3,56)$	0.00
$p(1,28)$	0.25	$p(1,57)$	0.53	$p(2,28)$	0.51	$p(2,57)$	0.43	$p(3,28)$	0.23	$p(3,57)$	0.04
$p(1,29)$	0.00			$p(2,29)$	0.20			$p(3,29)$	0.80		

Table 9.5: Posterior probabilities of membership when applying a Gibbs sampler to fit a regression mixture of three normal distributions.

path is a set of segments from centroid to centroid of adjacent regions.

The values of the centroids were taken from the world data provided with the package Mapinfo Professional 6.5, and using the functions *centroidX* and *centroidY* of this package. These functions return the x and y coordinates of the centre of each country's *minimum bounding rectangle*, which is always oriented to the x and y axes. The x and y coordinates correspond to longitude and latitude values (MapInfo Corporation, 2001). Overall, the geographical distance takes into account adjacencies and Euclidean distances.

A considerable number of countries in this study do not have any adjacent countries, either because they correspond to islands (Iceland, Japan, UK, etc.) or due to the sparse data in some continents, such as Africa. Among the five

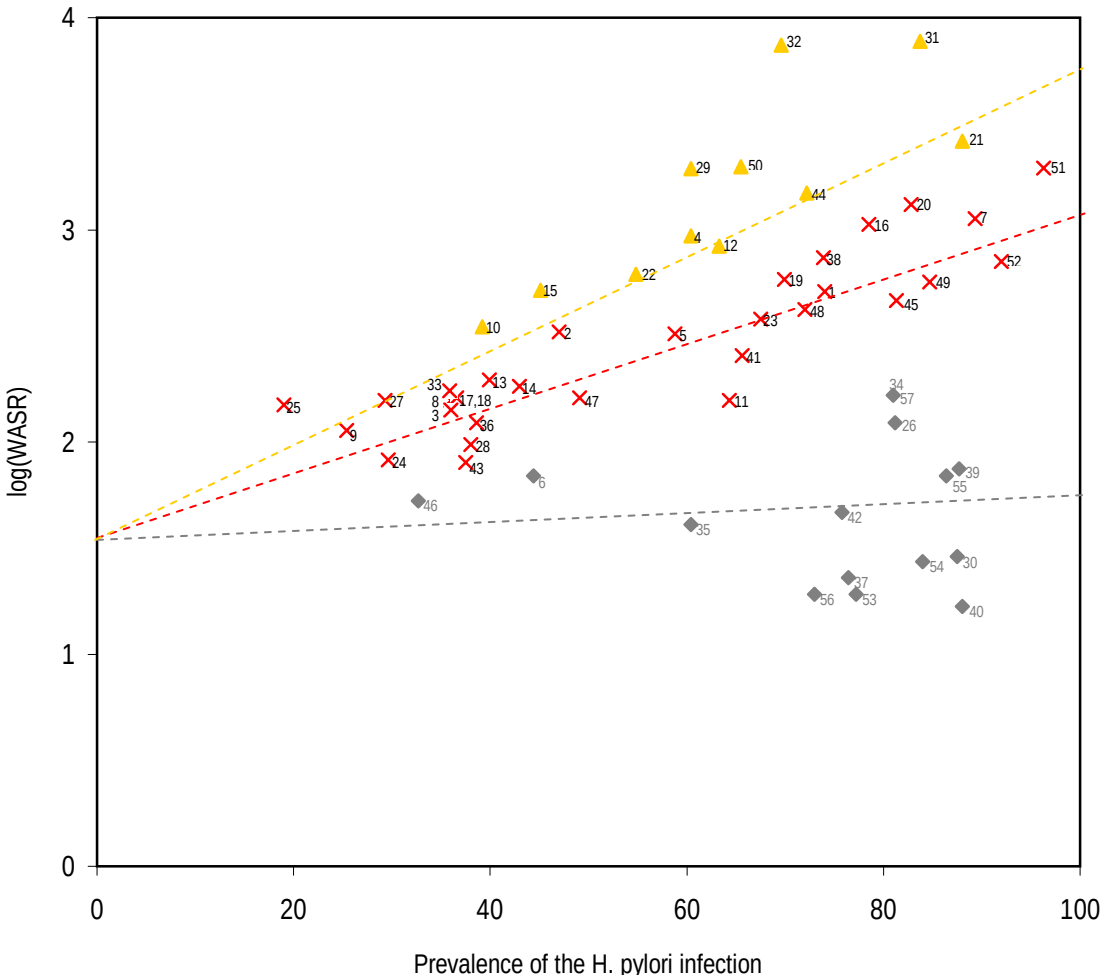


Figure 9.4: Scatterplot of stomach cancer log-WASR against the prevalence of *H. pylori* infection showing the regression lines estimated by a regression mixture of three normals and identifying observations in each group (group 1 in grey, group 2 in red, group 3 in yellow and country codes as in figure 9.1).

African countries included in this study, there is not even one pair of adjacent countries. To deal with these cases, we generalized the definition of geographical distance.

While the sea may have been an obstacle for the movement of populations, small distances in the sea were likely to be crossed easily. Hence, two countries

separated by small distances in the sea may well have common strains of *H. pylori*. Our first generalization of the notion of adjacency is thus to consider that two countries are *linked* when they are adjacent or the minimum distance in sea between their borders is less than 150 miles. The geographical distance between any two linked countries is defined as the Euclidean distance between their centroids, as if the countries were adjacent. An area is thus linked to:

1. all its adjacent countries,

2. all countries at distance by sea less than 150 miles.

In countries for which not all adjacent countries are included in the study, it is important to account for all the directions through which the *H. pylori* bacteria may have spread. Otherwise, the analysis will be biased, as the direction towards the adjacent countries will be the preferred one for the formation of clusters. Therefore, when adjacent countries were absent, we ensured that there was at least one link to a country in each of four quadrants, defined by the crossing of the directions north-south, east-west (figure 9.5).

To put it formally, for every country s we first checked whether all adjacent countries and all countries at distance by sea less than 150 miles of the country s are part of the study. If yes, all directions through which the *H. pylori* infection may have spread from and into the country s are already taken into account and the full set of *linked* countries is fully defined by rules 1 and 2 above. If not, we checked whether there is already a linked country in every quadrant among the linked countries obtained by applying the rules 1 and 2 above. If there are linked countries in every quadrant, we considered that all directions had been taken into account. If not, we needed to identify a linked country in the quadrants without one. With this aim, we identified the country at the shortest Euclidean distance

to the country s in the quadrants where there was not yet a linked country. The country at the shortest Euclidean distance is then linked to country s only if there is a path from the latter to the former that does not pass by any country chosen by rules 1 and 2 above. If the only paths between the two countries pass by countries which were already linked then the direction through which the *H. pylori* infection may have spread has already been taken into account and there is no need to establish another link.

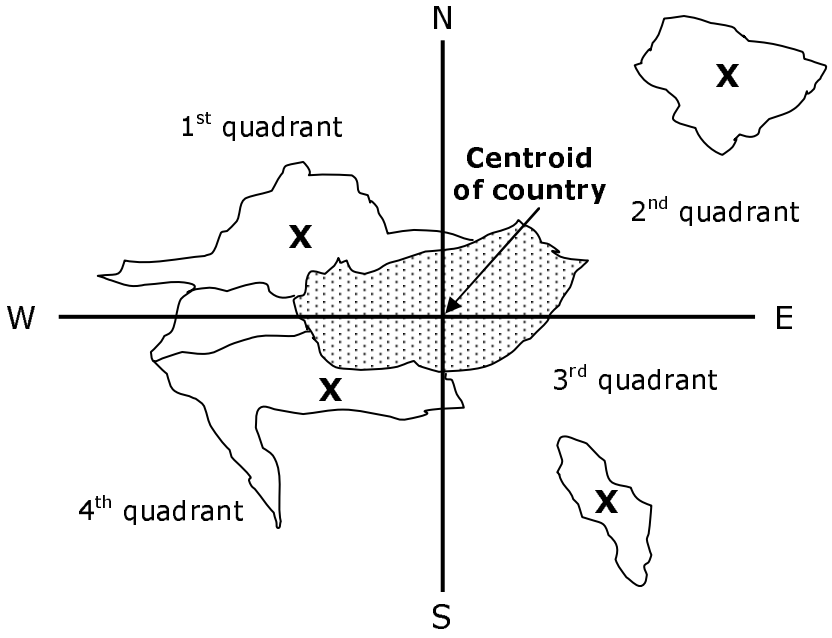


Figure 9.5: Illustration of quadrants in relation to a country's centroid. Linked countries in each quadrant are marked with an 'X'.

For pairs of linked countries, chosen either by rules 1 and 2 or by the procedure just described, the geographical distance between the two countries is the Euclidean distance between their centroids. In addition, paths are defined here as sequences of linked countries as opposed to the definition given in expression 7.1, in which paths were defined as sequences of adjacent countries.

The resulting structure of links is displayed in Figure 9.6. For example, Sweden and Denmark, as well as UK and France, are linked because the minimum sea distance between them is less than 150 miles. Zambia is linked to Senegal, as the latter is the nearest country in the first quadrant of the former. Similarly, Saudi Arabia is linked to India, which is the nearest country in the third quadrant of Saudi Arabia.

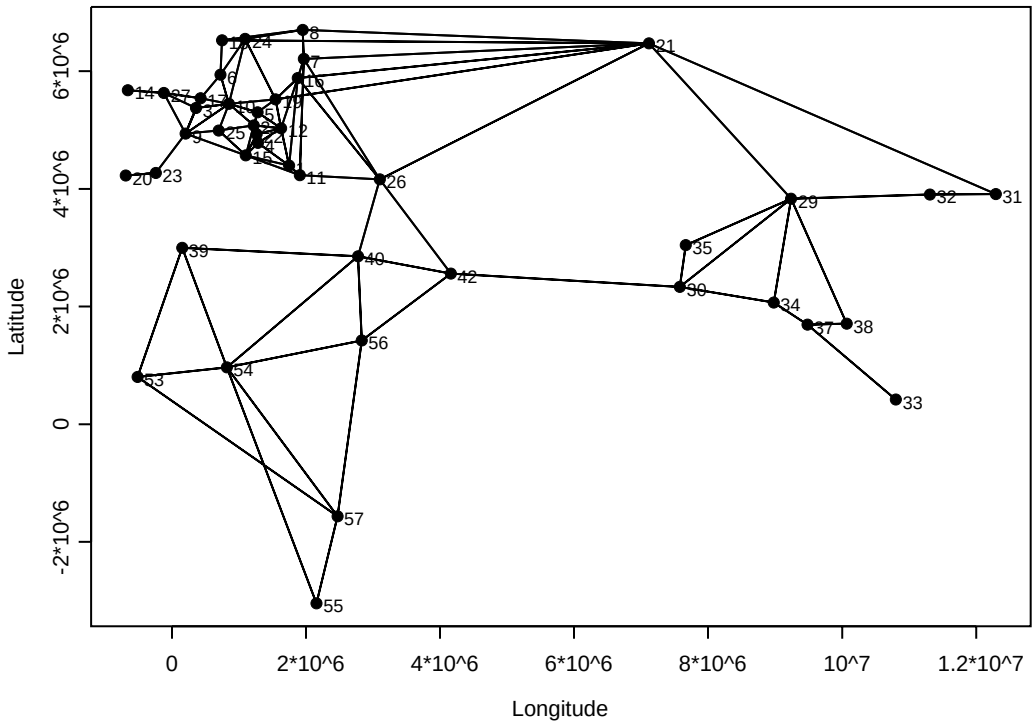


Figure 9.6: Neighbourhood structure for 43 countries in Africa, Asia and Europe. Country codes as in figure 9.1

We omitted Iceland because it is too far from any country in our study, and no links could be established to other countries using the rules described above. Other countries, such as Australia, New Zealand, Israel and the American

countries, which were targets of large migration movements from other continents in a recent past, also have to be dealt with in a different way because the *H. pylori* strains may have been largely influenced by those of the immigrant populations. We adopt the following strategy. First, we omit these countries from the fitting and analyse only the remaining 43 countries. We will thus have a model in which geographical distance and adjacencies are the only factors taken into account. Predictive probabilities of membership can then be calculated for the omitted countries, assuming uniform prior weights among all groups.

Prior distributions

Once the pairs of linked countries are defined (figure 9.6), the prior model for clustering is the one defined in section 7.5.

As explained in sections 6.4.1 and 7.5, the prior distributions for the slope components b_{11} and $\delta_m, m = 1, \ldots, M-1$ are assumed to be Gaussian, with mean zero and a large variance, but truncated to $(0, +\infty)$ to reflect the established knowledge of the *H. pylori* infection as a risk factor. The parameters of the prior Gamma distribution for the precision were also chosen in such a way that the variance of the prior distribution is very large, in order to obtain a vague prior.

9.4.1 Results of the clustering model

Three chains of the Gibbs sampler with different starting values were run in order to assess convergenge. This procedure was repeated for $M = 2$, $M = 3$ and $M = 4$. In general, the chains seemed to have already converged after 5000 iterations. By way of illustration, the sampled values for selected parameters

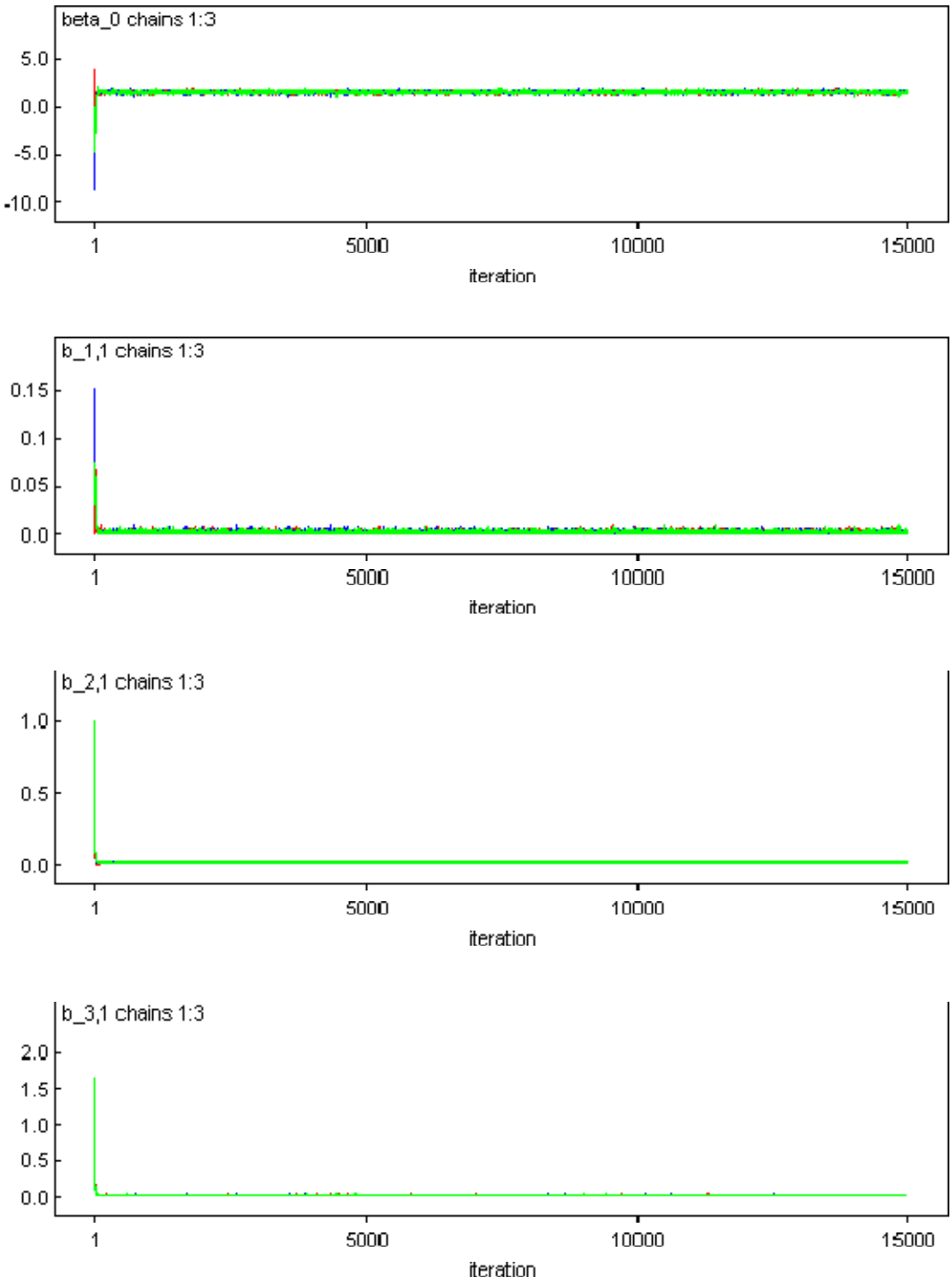


Figure 9.7: Sampled values of the intercept β_0 and the slopes $\boldsymbol{b}_{\bullet 1}$ when fitting the clustering model for $M = 3$ in three MCMC runs.

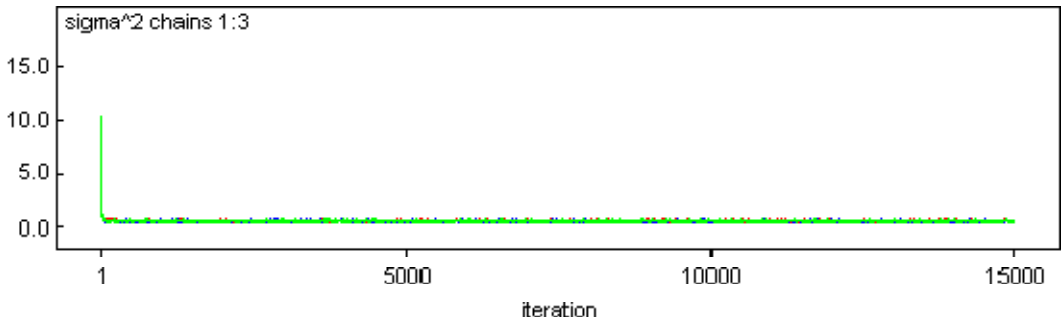


Figure 9.8: Sampled values of the variance σ^2 when fitting the clustering model for $M = 3$ in three MCMC runs.

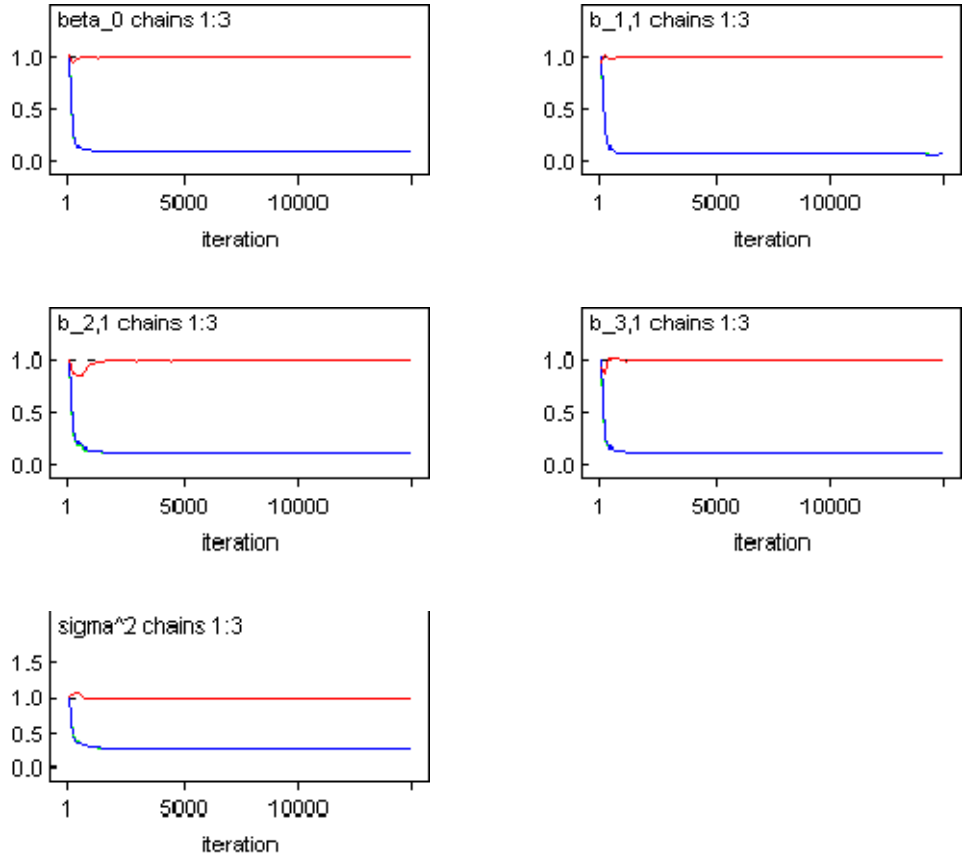


Figure 9.9: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for the intercept β_0, the slopes $\boldsymbol{b}_{\bullet 1}$ and the variance σ^2 when fitting the clustering model for $M = 3$.

for $M = 3$ are displayed in figures 9.7 and 9.8 and the Brooks-Gelman-Rubin convergence statistics (section 7.13) are shown in figure 9.9. The values of the Brooks-Gelman-Rubin convergence statistics converged immediatly to one despite the large difference in the initial values chosen for the MCMC runs.

The Gibbs sampler was run for 15000 iterations for $M = 2$, $M = 3$ and $M = 4$ with the first 5000 iterations discarded as burn-in. Estimates were made using the remaining 10000 sample points, which provided acceptable Monte Carlo errors. The computation was undertaken in Winbugs 1.4 (the code is given in appendix G).

The response variables were first assumed to be Gaussian distributed. We have also used as the distribution for the responses the double exponential distribution, which can be seen as a robust version of the Gaussian distribution. The results obtained using the two types of responses were very similar. We have opted to present here the results obtained with the double exponential distribution.

$M = 2$

The sampled values of $(\beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)$ are shown in Table 9.6. Different slopes are clearly identified. We monitored the posterior probabilities of membership, which are shown in Table 9.7 and denoted as $p(m, i) = P(Z_i = m|...)$.

		Standard	Quantiles		
	Mean	deviation	2.50%	Median	97.50%
β_0	1.52	0.12	1.28	1.53	1.74
b_{11}	0.002	0.002	0.000	0.002	0.006
b_{21}	0.019	0.002	0.015	0.019	0.024
σ	0.57	0.05	0.49	0.57	0.67

Table 9.6: Summary measures of the posterior probabilities of the regression parameters when applying a Gibbs sampler to fit a clustering model with two clusters.

$p(1,1)$	0.00	$p(1,24)$	0.00	$p(2,1)$	1.00	$p(2,24)$	1.00
$p(1,2)$	0.00	$p(1,25)$	0.00	$p(2,2)$	1.00	$p(2,25)$	1.00
$p(1,3)$	0.00	$p(1,26)$	0.00	$p(2,3)$	1.00	$p(2,26)$	1.00
$p(1,4)$	0.00	$p(1,27)$	0.00	$p(2,4)$	1.00	$p(2,27)$	1.00
$p(1,5)$	0.00	$p(1,29)$	0.00	$p(2,5)$	1.00	$p(2,29)$	1.00
$p(1,6)$	0.00	$p(1,30)$	1.00	$p(2,6)$	1.00	$p(2,30)$	0.00
$p(1,7)$	0.00	$p(1,31)$	0.00	$p(2,7)$	1.00	$p(2,31)$	1.00
$p(1,8)$	0.00	$p(1,32)$	0.00	$p(2,8)$	1.00	$p(2,32)$	1.00
$p(1,9)$	0.00	$p(1,33)$	1.00	$p(2,9)$	1.00	$p(2,33)$	0.00
$p(1,10)$	0.00	$p(1,34)$	1.00	$p(2,10)$	1.00	$p(2,34)$	0.00
$p(1,11)$	0.00	$p(1,35)$	1.00	$p(2,11)$	1.00	$p(2,35)$	0.00
$p(1,12)$	0.00	$p(1,37)$	1.00	$p(2,12)$	1.00	$p(2,37)$	0.00
$p(1,14)$	0.00	$p(1,38)$	1.00	$p(2,14)$	1.00	$p(2,38)$	0.00
$p(1,15)$	0.00	$p(1,39)$	1.00	$p(2,15)$	1.00	$p(2,39)$	0.00
$p(1,16)$	0.00	$p(1,40)$	1.00	$p(2,16)$	1.00	$p(2,40)$	0.00
$p(1,17)$	0.00	$p(1,42)$	1.00	$p(2,17)$	1.00	$p(2,42)$	0.00
$p(1,18)$	0.00	$p(1,53)$	1.00	$p(2,18)$	1.00	$p(2,53)$	0.00
$p(1,19)$	0.00	$p(1,54)$	1.00	$p(2,19)$	1.00	$p(2,54)$	0.00
$p(1,20)$	0.00	$p(1,55)$	1.00	$p(2,20)$	1.00	$p(2,55)$	0.00
$p(1,21)$	0.00	$p(1,56)$	1.00	$p(2,21)$	1.00	$p(2,56)$	0.00
$p(1,22)$	0.00	$p(1,57)$	1.00	$p(2,22)$	1.00	$p(2,57)$	0.00
$p(1,23)$	0.00			$p(2,23)$	1.00		

Table 9.7: Posterior probabilities of group membership for 43 countries when applying a Gibbs sampler to fit a clustering model with two clusters.

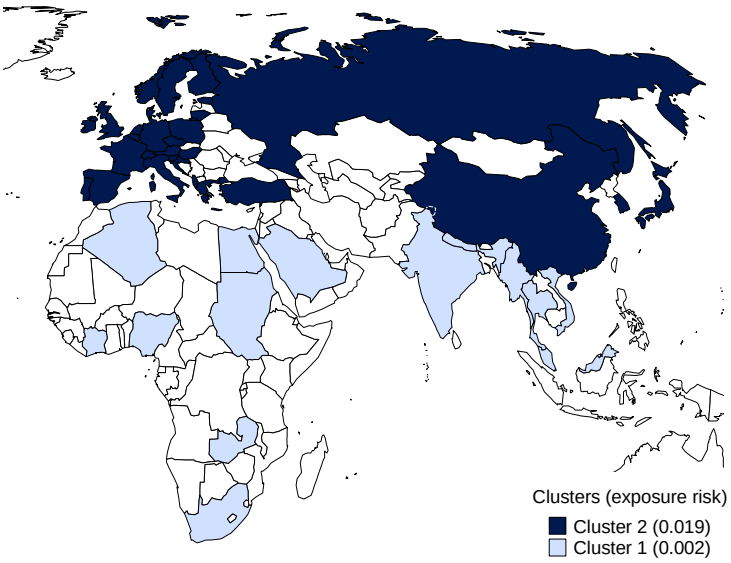


Figure 9.10: Clusters obtained with a clustering model with two clusters.

The countries were clearly allocated to one of the clusters. Figure 9.10 displays a map with the clusters. A total of 14 observations were allocated to the group with the lowest exposure coefficient (group 1): observations 30, 33,

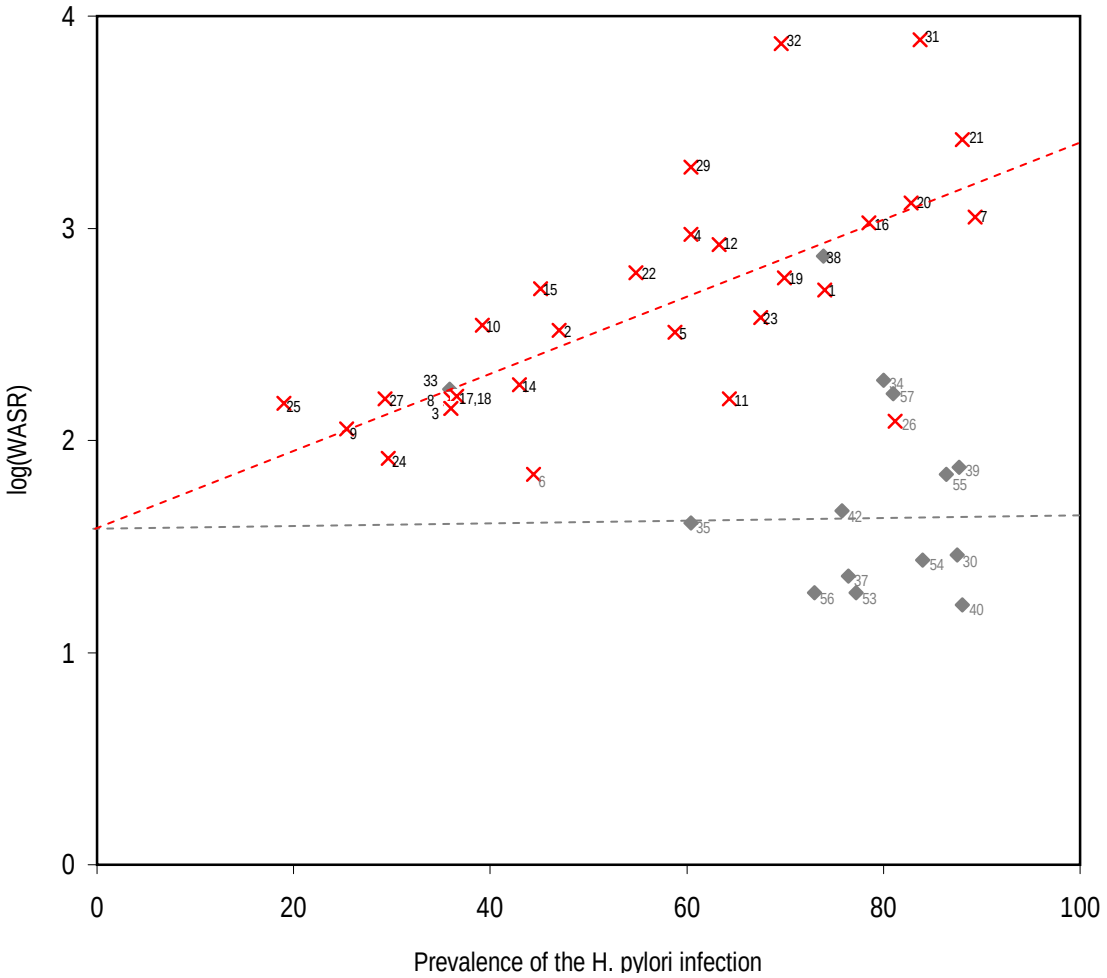


Figure 9.11: Scatterplot of stomach cancer log-WASR against the prevalence of *H. pylori* infection showing the regression lines for the two clusters identified by the clustering model and identifying observations in each group (group 1 in grey, group 2 in red, and country codes as in figure 9.1).

34, 35, 37, 38, 39, 40, 42, 53, 54, 55, 56 and 57 (India, Malaysia, Myanmar, Nepal, Thailand, Vietnam, Algeria, Egypt, Saudi Arabia, Côte d'Ivoire, Nigeria, South Africa, Sudan, Zambia). All these have a probability near 1 of belonging to group 1. The group comprises all African countries, countries from the Middle East and countries from South East Asia.

The remaining 29 observations were clearly allocated to the group with highest exposure coefficient (group 2). The group comprises the countries from Europe and Eastern Asia (China, Japan, Korea and Vietnam).

Figure 9.11 displays the regression lines obtained along with the observations allocated to each group. There are a few differences between these results and those obtained with the regression mixture (figure 9.3) - figure 9.12 displays copies of figures 9.3 and 9.11 to aid comparison between the results of these two models. Contrary to the results obtained with the regression mixture, observations 33 and 38 (Malaysia and Vietnam) were now assigned to the cluster with the lowest exposure coefficient (group 1), and observation 26 (Turkey) was allocated to the cluster with the highest exposure coefficient (group 2). The countries of observations 26 and 38 are on the border of the clusters found. The allocation of observations on the border is likely to be influenced by the number of neighbours in each of the surrounding clusters. The allocation of observation 33 (Malaysia) to group 1 is due to the fact that the country is surrounded by countries assigned to the group 1.

$M = 3$

The sampled values of $(\beta_0, \boldsymbol{b}_{\bullet 1}, \sigma^2)$ are shown in Table 9.8. Different slopes were identified. We estimated the posterior probabilities of membership as shown in Table 9.9. These probabilities indicate clear allocation of each observation to one of the clusters.

The three clusters of countries are presented in the map of figure 9.13. The regression lines obtained along with the observations allocated to each group are shown in figure 9.14.

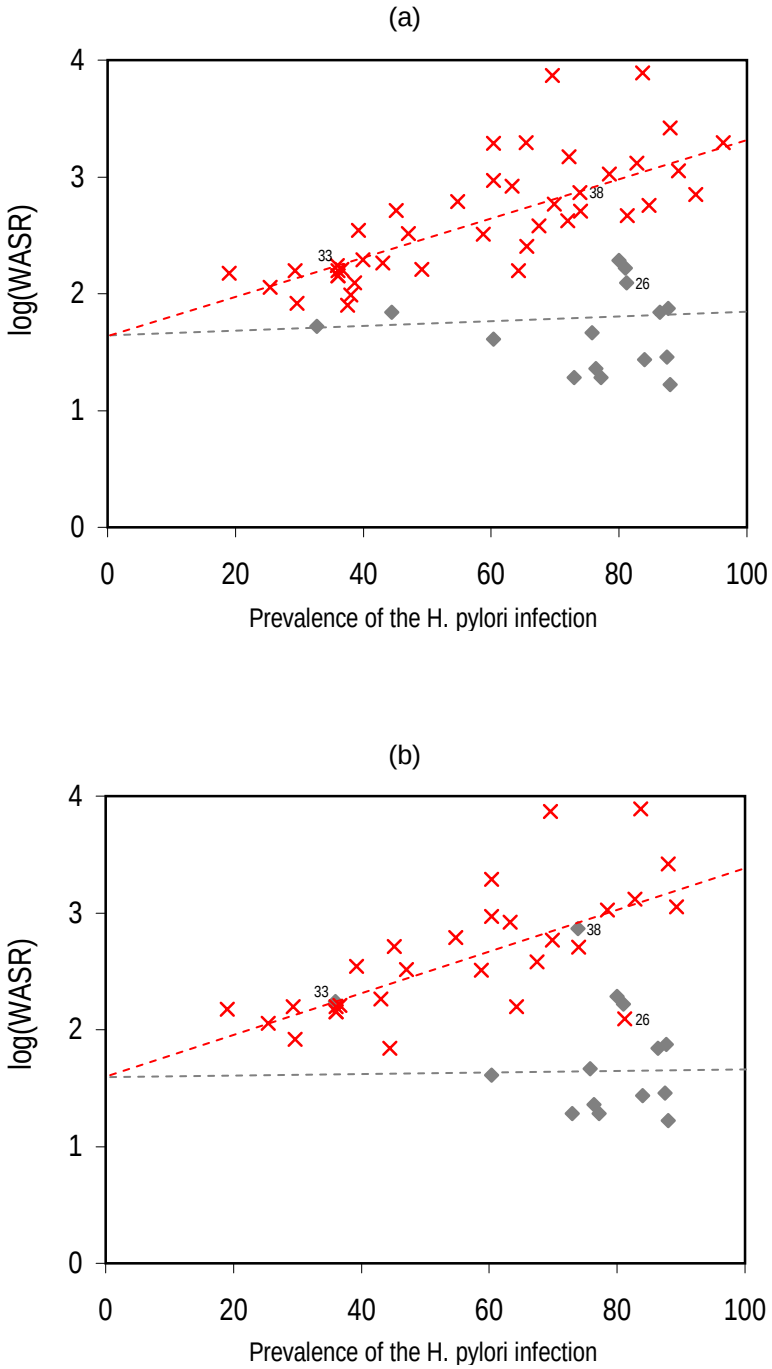


Figure 9.12: Copies of: (a) figure 9.3 and (b) figure 9.11, to aid comparison between them.

	Mean	Standard deviation	Quantiles 2.5%	Median	97.5%
β_0	1.58	0.11	1.35	1.58	1.78
b_{11}	0.002	0.001	0.000	0.001	0.005
b_{21}	0.017	0.002	0.013	0.017	0.022
b_{31}	0.025	0.003	0.019	0.026	0.031
σ	0.51	0.04	0.44	0.51	0.60

Table 9.8: Summary measures of the posterior distributions of the regression parameters when applying a Gibbs sampler to fit a clustering model with three clusters.

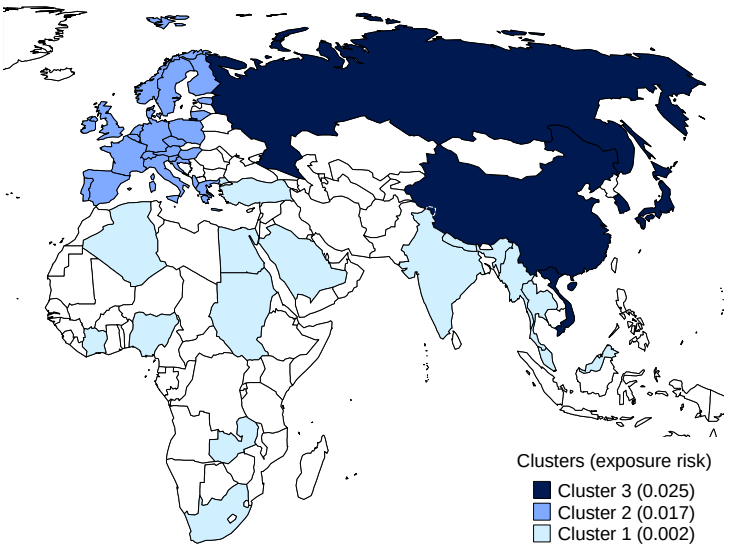


Figure 9.13: Clusters obtained with a clustering model with three clusters.

The countries from Africa, South East Asia and the Middle East all have a probability over 0.9 of belonging to the group with the lowest exposure coefficient (group 1). The group includes observations 26, 30, 33, 34, 35, 37, 39, 40, 42, and 53-57 (Turkey, India, Malaysia, Myanmar, Nepal, Thailand, Algeria, Egypt, Saudi Arabia, Côte d'Ivoire, Nigeria, South Africa, Sudan and Zambia).

The 23 European countries of our study were assigned to the group with the second highest exposure coefficient (observations 1-12, 14-20, 22-25 and 27).

All other observations, i.e. observations 21, 29, 31, 32, 38 (Russian Federation, China, Japan, Republic of Korea and Vietnam), belong to the group with

$p(1,1)$	0.00	$p(2,1)$	0.97	$p(3,1)$	0.03
$p(1,2)$	0.00	$p(2,2)$	0.97	$p(3,2)$	0.03
$p(1,3)$	0.00	$p(2,3)$	0.98	$p(3,3)$	0.02
$p(1,4)$	0.00	$p(2,4)$	0.97	$p(3,4)$	0.03
$p(1,5)$	0.00	$p(2,5)$	0.97	$p(3,5)$	0.03
$p(1,6)$	0.00	$p(2,6)$	0.97	$p(3,6)$	0.03
$p(1,7)$	0.00	$p(2,7)$	0.97	$p(3,7)$	0.03
$p(1,8)$	0.00	$p(2,8)$	0.97	$p(3,8)$	0.03
$p(1,9)$	0.00	$p(2,9)$	0.98	$p(3,9)$	0.02
$p(1,10)$	0.00	$p(2,10)$	0.97	$p(3,10)$	0.03
$p(1,11)$	0.04	$p(2,11)$	0.93	$p(3,11)$	0.03
$p(1,12)$	0.00	$p(2,12)$	0.97	$p(3,12)$	0.03
$p(1,14)$	0.00	$p(2,14)$	0.98	$p(3,14)$	0.02
$p(1,15)$	0.00	$p(2,15)$	0.97	$p(3,15)$	0.03
$p(1,16)$	0.00	$p(2,16)$	0.97	$p(3,16)$	0.03
$p(1,17)$	0.00	$p(2,17)$	0.97	$p(3,17)$	0.03
$p(1,18)$	0.00	$p(2,18)$	0.97	$p(3,18)$	0.03
$p(1,19)$	0.00	$p(2,19)$	0.97	$p(3,19)$	0.03
$p(1,20)$	0.00	$p(2,20)$	0.98	$p(3,20)$	0.02
$p(1,21)$	0.00	$p(2,21)$	0.02	$p(3,21)$	0.98
$p(1,22)$	0.00	$p(2,22)$	0.97	$p(3,22)$	0.03
$p(1,23)$	0.00	$p(2,23)$	0.98	$p(3,23)$	0.02
$p(1,24)$	0.00	$p(2,24)$	0.97	$p(3,24)$	0.03
$p(1,25)$	0.00	$p(2,25)$	0.97	$p(3,25)$	0.03
$p(1,26)$	0.96	$p(2,26)$	0.03	$p(3,26)$	0.01
$p(1,27)$	0.00	$p(2,27)$	0.98	$p(3,27)$	0.02
$p(1,29)$	0.00	$p(2,29)$	0.02	$p(3,29)$	0.98
$p(1,30)$	1.00	$p(2,30)$	0.00	$p(3,30)$	0.00
$p(1,31)$	0.00	$p(2,31)$	0.02	$p(3,31)$	0.98
$p(1,32)$	0.00	$p(2,32)$	0.02	$p(3,32)$	0.98
$p(1,33)$	1.00	$p(2,33)$	0.00	$p(3,33)$	0.00
$p(1,34)$	1.00	$p(2,34)$	0.00	$p(3,34)$	0.00
$p(1,35)$	1.00	$p(2,35)$	0.00	$p(3,35)$	0.00
$p(1,37)$	1.00	$p(2,37)$	0.00	$p(3,37)$	0.00
$p(1,38)$	0.02	$p(2,38)$	0.02	$p(3,38)$	0.96
$p(1,39)$	1.00	$p(2,39)$	0.00	$p(3,39)$	0.00
$p(1,40)$	1.00	$p(2,40)$	0.00	$p(3,40)$	0.00
$p(1,42)$	1.00	$p(2,42)$	0.00	$p(3,42)$	0.00
$p(1,53)$	1.00	$p(2,53)$	0.00	$p(3,53)$	0.00
$p(1,54)$	1.00	$p(2,54)$	0.00	$p(3,54)$	0.00
$p(1,55)$	1.00	$p(2,55)$	0.00	$p(3,55)$	0.00
$p(1,56)$	1.00	$p(2,56)$	0.00	$p(3,56)$	0.00
$p(1,57)$	1.00	$p(2,57)$	0.00	$p(3,57)$	0.00

Table 9.9: Posterior probabilities of group membership when applying a Gibbs sampler to fit a clustering model with three clusters.

the highest exposure coefficient, with probability higher than 0.9. This group includes mainly Far East Asian countries.

$M = 4$

We increased the number of groups to $M = 4$ and estimated the posterior probabilities of membership. All observations were allocated to only three groups,

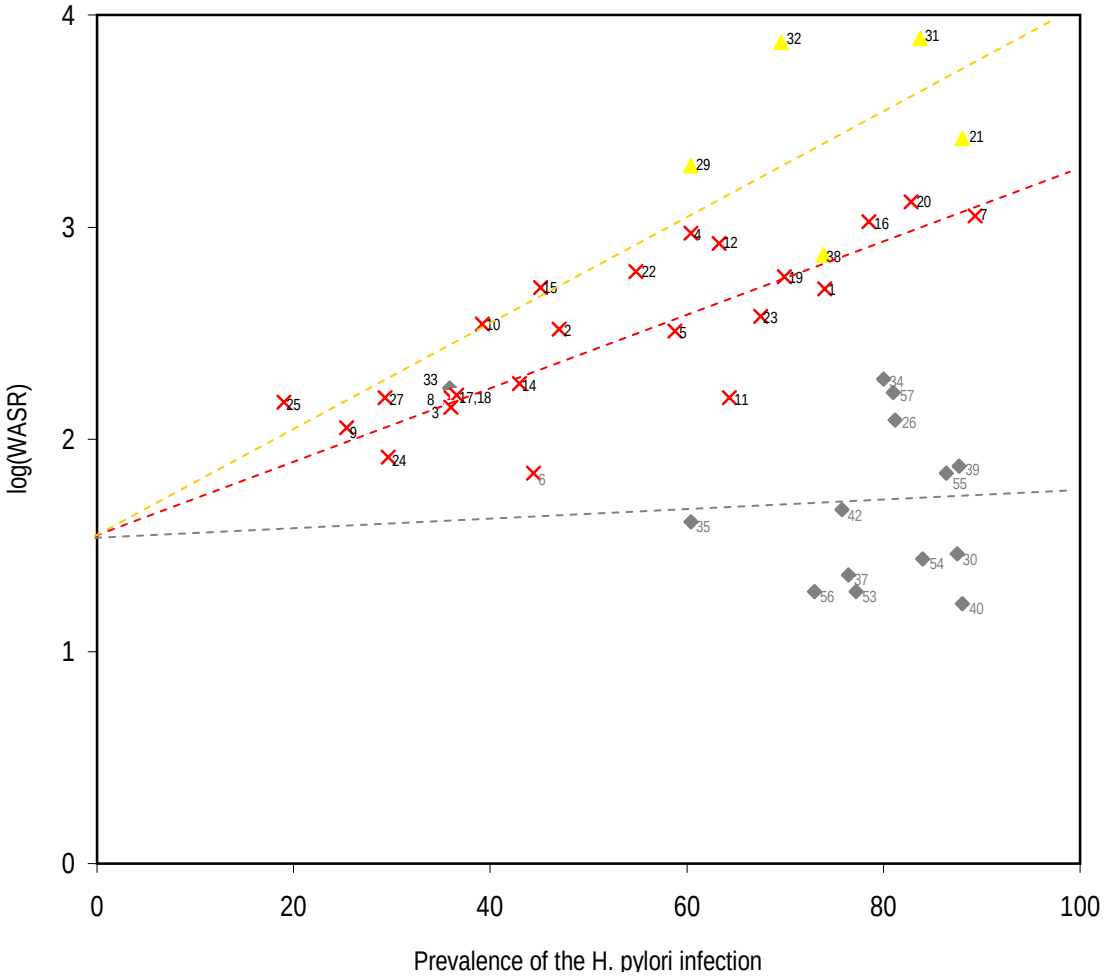


Figure 9.14: Scatterplot of stomach cancer log-WASR against the prevalence of *H. pylori* infection showing the regression lines for the three clusters identified by the clustering model and identifying observations in each group (group 1 in grey, group 2 in red, group 3 in yellow and country codes as in figure 9.1).

the same groups that were identified with $M = 3$.

9.5 The choice of M

To choose the number of groups M that best fits the data with the clustering model, we calculated the values of the averaged posterior mean squared error (MSE), defined as

$$\frac{1}{n}\sum_{i=1}^{n} E[(y_i - \mu_{v_i})^2|\boldsymbol{y}]$$

where y_i stands for the log-WASR and μ_{v_i} for the estimated posterior mean of the i-th country.

	Averaged posterior MSE		
	Posterior	CI	
M	mean	2.5%	97.5%
1	0.46	0.43	0.57
2	0.18	0.17	0.21
3	0.10	0.09	0.13
4	0.11	0.09	0.15

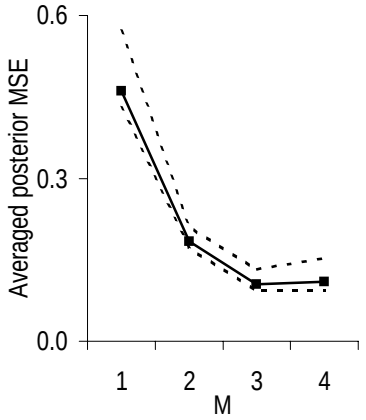


Figure 9.15: Averaged posterior mean squared error (bold line) and credible intervals (dashed lines) when applying a Gibbs sampler to fit a linear regression (M=1) and clustering models with two (M=2), three (M=3) and four clusters (M=4).

Figure 9.15 displays the averaged posterior MSE for various choices of the number of groups, M. The fall of the averaged posterior MSE and the fact that clear groups only appear until $M = 3$ seem to support the existence of three groups of different virulence of *H. Pylori*. The analyses presented from here onwards are thus based on the results obtained for $M = 3$.

9.6 Predictive probabilities

We calculated the probability of membership for the 14 remaining countries, according to the expressions provided in section 7.12. We assumed that all countries have equal prior probability of belonging to each of the three groups. The resulting probabilities of membership are displayed in table 9.10. The results suggest a predominant influence of European *H. pylori* strains in Israel and in Latin American countries (Argentina, Brazil, Colombia, Mexico and Venezuela). In the North American countries, i.e. Canada and the United States, both the African and the European strains seem to be important, with a slight predominance of the African strains. The strains in the two countries from Oceania, Australia and New Zealand, included in this analysis appear to be more mixed, with a slight predominance of the European strain. The results for Iceland also indicate a mix of the original strains, with a slight predominance of the European one.

Country	$P(V=1\|...)$	$P(V=2\|...)$	$P(V=3\|...)$
Iceland	0.20	0.44	0.36
Australia	0.35	0.40	0.25
New Zealand	0.30	0.42	0.28
Israel	0.25	0.57	0.18
Canada	0.41	0.38	0.21
Jamaica	0.01	0.49	0.50
Mexico	0.15	0.70	0.15
United States	0.47	0.34	0.19
Argentina	0.29	0.47	0.24
Barbados	0.14	0.65	0.21
Brazil	0.12	0.73	0.16
Chile	0.01	0.35	0.64
Colombia	0.01	0.73	0.26
Venezuela	0.09	0.77	0.13

Table 9.10: Posterior probabilities of membership to the three clusters for the countries for which the probabilities of membership were estimated by predictive probabilities.

For some countries, these results are consistent with the available information on the presence of ethnic groups in these countries. According to United Nations records, the population in Israel in 1983 was 4.038 million. Among these, 1.231

million were of European origin (30%), and 727,000 were born in Europe (18%). This may explain why the European strain seems to be predominant in this country. This finding is also consistent with the results on the isolates of *H. pylori* from Israel analysed in Falush et al. (2003), which were indeed similar to the European ones.

Both New Zealand and Australia have large proportions of their population with European origin: 86% in New Zealand and 89% in Australia (Table 9.11). However, our analysis pointed towards a mix of strains of *H. pylori* in these countries, including the European strain. This may suggest that different strains of *H. pylori* spread differently, with some strains presenting predominance over others.

Country, Year	New Zealand, 1981 census	Australia, 1976 census
Total population	3,175,737	13,548,449
Population of European origin (%)	2,723,223 (86%)	12,037,151 (89%)

Table 9.11: Population of European origin in New Zealand and Australia. Source: United Nations.

9.7 Estimating the virulence of H. pylori

The exposure coefficient of the regression lines fitted to the log-WASRs can be seen as an estimate of the virulence of *H.pylori*. The values of the exposure coefficient for the 43 countries included in the clustering model were shown in the map of figure 9.13. For all these 43 countries, the allocation to one cluster was very clear because the probability of membership was very high for one of the clusters. For the other 14 countries that were not included in the clustering model, cluster membership is not so clear, with membership being distributed among two and even three clusters (table 9.10). It is then natural to consider a

weighted averaged exposure coefficient b_i for each country, where the weights are determined by the posterior probabilities of membership displayed in table 9.10. To put it formally,

$$b_i \;=\; b_{11}p(1,i) + b_{21}p(2,i) + b_{31}p(3,i)\,, \tag{9.1}$$

where $p(j,i)$ is the predictive probability of allocation of country i to cluster j and b_{j1} is the posterior mean of the exposure coefficient of group j. The weighted averaged exposure coefficients are displayed in table 9.12.

The map of figure 9.16 shows the estimated virulence of *H.pylori*, i.e the country specific estimated exposure coefficients, for the 57 countries included in the study.

			Relative risk		
		Exposure	Posterior	CI	
Code	Country	coefficient	mean	2.5%	97.5%
13	Iceland	0.017	5.6	3.6	8.8
28	Australia	0.014	4.0	2.7	6.1
36	New Zealand	0.015	4.5	3.0	6.9
41	Israel	0.015	4.5	3.0	6.9
43	Canada	0.013	3.6	2.5	5.4
44	Jamaica	0.021	8.4	4.9	13.8
45	Mexico	0.016	5.1	3.4	7.9
46	United States	0.012	3.2	2.3	4.8
47	Argentina	0.015	4.4	3.0	6.8
48	Barbados	0.017	5.4	3.6	8.5
49	Brazil	0.017	5.4	3.6	8.4
50	Chile	0.022	9.5	5.4	15.8
51	Colombia	0.019	6.9	4.3	11.0
52	Venezuela	0.017	5.5	3.6	8.6

Table 9.12: Weighted averaged exposure coefficient (posterior means) and estimated relative risk (posterior means and 95% credible intervals) of the countries for which the cluster membership was estimated with predictive probabilities.

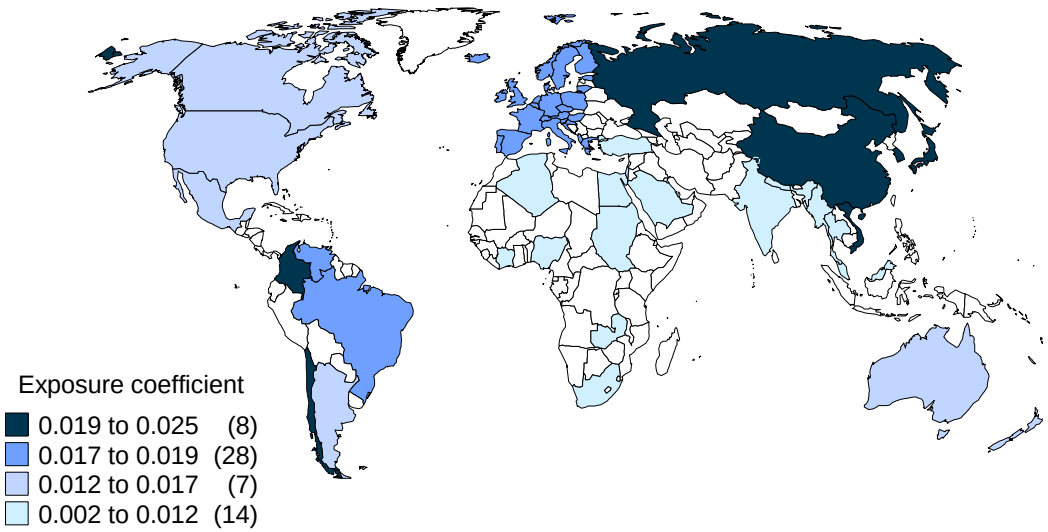


Figure 9.16: Estimated exposure coefficient of the *H. pylori* infection in 57 countries worldwide. The estimates were calculated on the basis of a clustering model with three groups.

9.8 Estimating the relative risk

An estimate of the relative risk (RR) of *H. pylori* infection can be obtained as the ratio between the absolute risk in a population with 100% prevalence and that in a population with no infection. This in turn is provided by e^{100b_1} where b_1 is the exposure coefficient or slope of the regression line fitted to the log-WASRs.

Estimation of the RR relies on an extrapolation to the values of prevalence 0% and 100%, which is not without its dangers, since it assumes that the fitted regression lines are valid outside the range of the covariates available in this study.

On the basis of the results displayed in table 9.8, an estimate of the RR for the *H. pylori* infection was obtained from the selected clustering model with

three clusters,

$$\begin{aligned} \text{Cluster 1: } RR &= e^{0.002\times 100} \approx 1.2, \\ \text{Cluster 2: } RR &= e^{0.017\times 100} \approx 5.5, \\ \text{Cluster 3: } RR &= e^{0.025\times 100} \approx 12.2. \end{aligned} \tag{9.2}$$

The 95% credible intervals for the slopes were estimated as (table 9.8),

$$\begin{aligned} CI_{95\%}(b_{11}) &= (0.000, 0.005), \\ CI_{95\%}(b_{21}) &= (0.013, 0.022), \\ CI_{95\%}(b_{31}) &= (0.019, 0.031), \end{aligned}$$

which leads to the following credible intervals for the relative risks:

$$\begin{aligned} \text{Cluster 1: } CI_{95\%}(RR) &= (1.0, 1.6), \\ \text{Cluster 2: } CI_{95\%}(RR) &= (3.8, 8.8), \\ \text{Cluster 3: } CI_{95\%}(RR) &= (6.6, 22.3). \end{aligned} \tag{9.3}$$

Expressions 9.2 provided the relative risk for the 43 countries that were included in the clustering model. For the other 14 countries that were not included in the clustering model, the relative risks were estimated as

$$RR_i = e^{100 b_i},$$

where b_i is the weighted averaged exposure coefficient given in expression 9.1 of section 9.7. The country specific relative risks and the corresponding 95% credible intervals are displayed in table 9.12.

9.9 Attributable risk

The attributable risk is defined as the excess proportion of incidence among the population that is associated to a risk factor. Formally, the attributable risk corresponds to the ratio

$$AR = (I_P - I_{NE})/I_P, \tag{9.4}$$

where I_P is the incidence in the population and I_{NE} is the incidence of the non-exposed population. The estimated value of the relative risk (RR_i), together with the available estimates of the prevalence (l_i) of the infection, can also be used to estimate the attributable risk (AR_i) in each country, using the expression

$$AR_i = \frac{l_i(RR_i - 1)}{1 + l_i(RR_i - 1)}. \tag{9.5}$$

Armitage et al. (2002), p.682, shows that this formulation is equivalent to that of equation 9.4 by setting $RR_i = I_E/I_{NE}$ and $I_P = l_i I_E + (1 - l_i)I_{NE}$. Expression 9.5 corresponds to the well known expression for the attributable risk (e.g. Armitage et al. (2002), p. 682), also known as *population attributable risk* (Benichou, 1998).

On the basis of the lower and upper limits of the 95% credible interval for the relative risks (expressions 9.3 and table 9.12), we calculated 95% credible intervals (AR_l, AR_u) for the attributable risk,

$$AR_l = \frac{p}{\frac{1}{RR_{iL}-1} + p}, \tag{9.6}$$

$$AR_u = \frac{p}{\frac{1}{RR_{iU}-1} + p}, \tag{9.7}$$

where RR_{iL} and RR_{iU} are respectively the lower and upper limits of the 95% credible interval for the relative risk of country i. The posterior means, calculated

			Attributable risk		
Code	Country	Group	2.5% CI	Posterior mean	97.5% CI
1	Albania	2	0.68	0.78	0.85
2	Austria	2	0.57	0.69	0.79
3	Belgium	2	0.50	0.63	0.74
4	Croatia	2	0.63	0.74	0.83
5	Czech Republic	2	0.62	0.73	0.82
6	Denmark	2	0.56	0.68	0.78
7	Estonia	2	0.72	0.81	0.87
8	Finland	2	0.50	0.63	0.74
9	France	2	0.42	0.54	0.67
10	Germany	2	0.52	0.65	0.75
11	Greece	2	0.64	0.75	0.83
12	Hungary	2	0.64	0.75	0.83
14	Ireland	2	0.55	0.67	0.77
15	Italy	2	0.56	0.68	0.78
16	Lithuania	2	0.69	0.79	0.86
17	Netherlands	2	0.51	0.63	0.74
18	Norway	2	0.50	0.63	0.74
19	Poland	2	0.66	0.77	0.85
20	Portugal	2	0.70	0.80	0.87
21	Russian Federation	3	0.83	0.91	0.95
22	Slovenia	2	0.61	0.72	0.81
23	Spain	2	0.66	0.76	0.84
24	Sweden	2	0.45	0.58	0.70
25	Switzerland	2	0.35	0.47	0.60
26	Turkey	1	0.01	0.13	0.34
27	United Kingdom	2	0.45	0.58	0.70
29	China	3	0.77	0.88	0.93
30	India	1	0.01	0.14	0.36
31	Japan	3	0.82	0.91	0.95
32	Korea, Rep.	3	0.80	0.89	0.94
33	Malaysia	1	0.00	0.06	0.19
34	Myanmar	1	0.00	0.13	0.34
35	Nepal	1	0.00	0.10	0.28
37	Thailand	1	0.00	0.12	0.33
38	Vietnam	3	0.81	0.90	0.94
39	Algeria	1	0.01	0.14	0.36
40	Egypt	1	0.01	0.14	0.36
42	Saudi Arabia	1	0.00	0.12	0.32
53	Côte d'Ivoire	1	0.00	0.12	0.33
54	Nigeria	1	0.01	0.13	0.35
55	South Africa	1	0.01	0.13	0.35
56	Sudan	1	0.00	0.12	0.32
57	Zambia	1	0.01	0.13	0.34

Table 9.13: Estimated *H. pylori* attributable risk for stomach cancer (posterior means and 95% credible intervals) in countries where the estimates of the relative risks were obtained from a clustering model with three clusters.

with expression 9.5, are shown in Tables 9.13 and 9.14 and displayed in the map of figure 9.17. The 95% credible intervals obtained are presented in Tables 9.13 and 9.14.

		Attributable risk		
Code	Country	2.5% CI	Posterior mean	97.5% CI
13	Iceland	0.51	0.65	0.76
28	Australia	0.40	0.53	0.66
36	New Zealand	0.44	0.57	0.70
41	Israel	0.57	0.69	0.79
43	Canada	0.36	0.49	0.62
44	Jamaica	0.74	0.84	0.90
45	Mexico	0.66	0.77	0.85
46	United States	0.29	0.41	0.55
47	Argentina	0.49	0.63	0.74
48	Barbados	0.65	0.76	0.84
49	Brazil	0.68	0.79	0.86
50	Chile	0.74	0.85	0.91
51	Colombia	0.76	0.85	0.91
52	Venezuela	0.71	0.80	0.87

Table 9.14: Estimated *H. pylori* attributable risk for stomach cancer (posterior means and 95% credible intervals) in countries where the estimates of the relative risk were calculated using predictive probabilities.

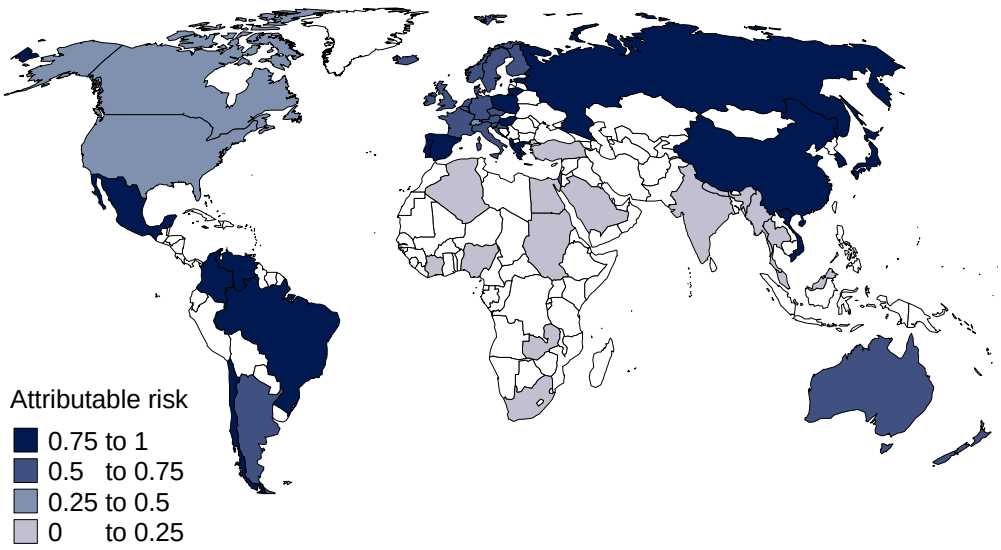


Figure 9.17: Estimated *H. pylori* attributable risk for stomach cancer in 57 countries worldwide. The estimates were calculated on the basis of a clustering model with three groups.

9.10 Discussion

Comparing regression mixtures with the clustering model

The estimated regression coefficients obtained with the standard regression mixtures (section 9.3.1) and with the clustering model (section 9.4.1) were very

similar, in spite of the fact that the clustering model used only a subset of the 57 countries whereas the regression mixture model was applied to the whole set. The allocation of the 43 countries into groups of similar exposure coefficient was almost the same in both models, with some differences resulting from the inclusion of geographical constraints through a prior distribution in the clustering model. In problems in which geographical continuity is expected, this feature of the clustering model constitutes an advantage.

Model choice

The issue of model choice is an important one in any statistical study. It is generally accepted that the choice of the "best" model depends largely on the aim of the study. We have shown that there is evidence to suggest that geographical information is relevant in the analysis of the data studied in this chapter, and in this sense, the choice of the clustering model seems justified. We have conducted simple linear regressions on each of the groups found by the clustering model, and found that the prevalence of *H. pylori* seems to explain a large proportion of the variability of the log transformed stomach cancer rates in the groups with the highest exposure coefficient (results not shown).

Coding-invariance

None of the models used in this chapter is coding invariant because they are not well-formulated, i.e. the models include a random slope but not a random intercept (section 5.4). However, the choice of models with a constant intercept is conceptually justified for the data studied in this chapter because there is obviously no effect of *H. pylori* infection on stomach cancer when the prevalence

is equal to zero (section 4.1). At this value of the prevalence, the baseline risk can thus be assumed common for all groups, which justifies the assumption of a common intercept when the origin of the prevalence is taken at zero.

Normalization

In the MCMC simulations we did not normalize the covariate values. Normalization is commonly used because it tends to accelerate the mixing of the chain. In the context of our analysis, if the covariates are not normalized, the intercept corresponds to the world baseline risk, when the prevalence is equal to zero. We have seen in sections 4.1 and 5.4 that a change of the origin of the prevalence would induce a varying intercept. Therefore, the models considered in this chapter, in which the intercept was assumed constant, would no longer be valid. Thus, normalization would complicate the parameterization of the model and we preferred not to use it.

Label switching

We chose to use constraints on the parameters $\boldsymbol{b}_{\bullet 1}$ to avoid label switching during the MCMC run (sections 6.4.2 and 7.6). Some authors have suggested that these constraints may push the estimates apart and lead to biased estimates (McLachlan and Peel (2000)). Label switching occurred rarely in the model used in this chapter. In the few cases in which label switching was observed, we adopted the following strategy. We have run the MCMC with and without constraints. In the simulation without constraints, different modes were clearly identified in the monitoring of the MCMC output. We were thus able to separate the output from one of the modes, which we compared with the MCMC output

obtained with the constrains on $\boldsymbol{b}_{\bullet 1}$. The results were practically identical. This is of course a very empirical and somewhat simplistic method of assessment, but we believe it suggests that any bias in the estimates arising from the contraints on the parameters $\boldsymbol{b}_{\bullet 1}$ is negligible.

African and Asian enigma

We briefly mentioned in section 3.1.1 the African and Asian enigma, which refers to the puzzling fact that African and some Asian countries, such as India, have high prevalence of the *H. pylori* infection, but low rates of stomach cancer.

Studies aiming at explaining the African and Asian enigma have tried to introduce other covariates, such as diet and environmental factors, that could explain the enigma. In this thesis, we propose a different explanation based on the geographical variation of the virulence of *H. pylori*. The countries of Africa and South-East Asia, including India, were assigned to the cluster with the smaller exposure coefficient, thus suggesting that the virulence of *H. pylori* in these countries is lower than in other parts of the world.

H. pylori and confounding factors

Apart from the differences in the strains of *H. pylori*, geography can also reflect host genetic and cultural characteristics. The identification of different groups with different exposure coefficients in this chapter supports the existence of different virulent populations of *H. pylori*. However, these associations could also be related to distinct responses of the host, as well as to differences in the life-style and environmental characteristics, which may be confounding factors.

Country sizes and sampling error

In the models presented in this chapter, the variance of the observations was assumed to be constant. However, different variances arise from distinct country sizes. We also applied a model in which the variances were assumed to be different (the variance was modelled as in Pocock et al. (1981)), but the sampling variance accounted for less than one per cent of the total variance and the results were similar to those provided here (results not shown).

Mortality versus incidence data

Mortality data are usually considered to be more reliable than incidence data. However, data on mortality due to stomach cancer were not available for the majority of countries in Africa and Asia. This is the reason why we used data on incidence.

Ethnic versus geographic data

Although ethnicity is likely to reflect in some degree the *H. pylori* spread among humans, ethnicity data are available only for a few countries worldwide. The use of different definitions across countries and throughout time also compromises the comparability of the data. In addition, sometimes the ethnic groups for which data are available do not necessarily reflect the whole range of ethnic groups present in the country. The ethnic groups can also be mixed, which will potentially entail the coexistence of several strains of *H. pylori*. The behaviour of competing strains of *H. pylori* is uncertain. This justifies why we did not use ethnic data. We preferred to use geographical distance and location as a surrogate of the spread of the *H. pylori* infection.

Chapter 10

Association between socio-economic level and stomach cancer in Portugal

In chapter 9, Portugal was allocated to a group with considerable virulence of the *H. pylori* infection. We know from the same study that the prevalence of the bacteria in this country is high. Since the stomach cancer mortality shows considerable variation across the country (figure 10.2), it is reasonable to ask whether the virulence of the bacteria is constant within the country or not. However, the prevalence of this infection in the Portuguese population is not known at a small geographical level. A surrogate measure is used in the analysis.

The choice of the surrogate measure was based on the following evidence: the prevalence of the *H. pylori* infection tends to be higher in populations with lower socio-economic level. Socio-economic level can thus be used as a surrogate for the prevalence of the *H. pylori* infection. If the virulence of *H. pylori* varies across

the region of study, the association between the risk of stomach cancer and *H. pylori* prevalence will vary across the region. As a result, the association between the risk of stomach cancer and socio-economic level will also vary across the region. For instance, areas with similar socio-economic level (and, consequently, similar *H. pylori* prevalence) may have different stomach cancer risks because of differences in the *H. pylori* virulence among the areas.

From the literature, it can be argued that the difference in strains of *H. pylori* tends to increase with geographical distance and that the spatial correlation between geographical areas could be a consequence of areas having similar strains (chapter 3). By allowing the coefficient of socio-economic status to vary with location, we can model the foregoing patterns. Therefore, spatially varying regression models can be used to identify areas with similar exposure coefficients of socio-economic status.

This chapter describes a study of mortality of stomach cancer in Portugal in the period 1985-1998. The chapter starts with a description of the data, and proceeds with a study of the association between socio-economic status and stomach cancer mortality.

10.1 Description of the Data

10.1.1 Study Region

The study region consists of continental Portugal. The geographical partition considered consists of 275 counties. The population sizes of the counties vary from 2052 to 663394 inhabitants (1991 Census). Socio-economic data are available at this scale.

10.1.2 Stomach Cancer Mortality Data

Mortality registries of stomach cancer were obtained from the General Board of Health in Portugal (DGS) for the period 1985-1998. These death records were coded according to the 9th International Classification of Diseases (ICD), which assigns a 3 digit code to every major disease, injury or cause of death (Rose and Barker, 1986, p. 31). The data set provided by DGS consists of all deaths coded as ICD 151, the code for the malignant neoplasms of the stomach. For each death, age at death, sex and code of the administrative unit of residence (county) were recorded.

Over the 14 year period there were a total of 37965 deaths caused by stomach cancer. Mortality data are vulnerable to errors in registries as well as to misreporting (Rodrigues et al., 1992). According to the World Health Organization, the estimated coverage of death registries in Portugal is 100% and the percentage of deaths attributed to ill-defined diseases is 9%. This indicates that misclassification, i.e. the wrong attribution or the lack of attribution of the cause of death in a death certificate, is a larger problem than misreporting. The geographical variability of misclassification across Portugal is unknown. In the study of this chapter, we implicitly assume that misclassification occurs at random.

Incidence data for stomach cancer in the period considered and for the required geographical precision were not available at the time of this study. However, the results of a study on the mortality should not differ considerably from a study on the incidence because the survival rate for stomach cancer is low.

10.1.3 Population Estimates

Population counts stratified by age and sex are available by county from the 1991 Census. However, the population sizes change in inter-censal years due to births, deaths, migration and ageing of the population. Unfortunately, the 1981 Census data are not available in an exploitable format and the results of the 2001 Census were not yet published at the time of this study. Therefore, 1991 Census data were used to estimate population sizes. Since the mortality data relate to deaths from 1985 to 1998, the year 1991 corresponds approximately to the middle year. Thus the data from the 1991 census is expected to provide reasonable estimates of the population sizes.

10.1.4 Spending Power Index

A spending power index, based on 12 variables reflecting income and consumption, is available at county level for the year 1993. This index is provided by the Portuguese Institute of National Statistics (INE, 1993) and is standardized so that the average is 100. Its value in each county is shown in figure 10.1.

10.2 Standardized Mortality Ratios

The age-sex specific risk rates for stomach cancer mortality were calculated using the numbers of registered deaths in 1985-1998 and the population sizes given in the 1991 Census data. These reference rates are presented in appendix L.

Using the internal method, the expected numbers of cases for each geographical unit were based on those age-sex specific rates and on the population sizes given in the 1991 Census data. Therefore, the total expected number of deaths

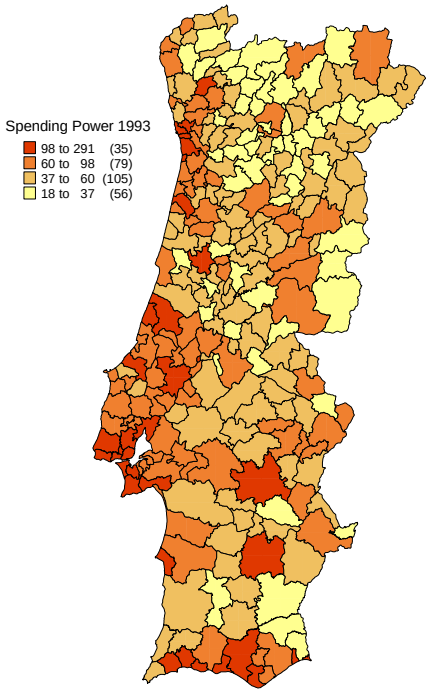


Figure 10.1: Spending power index, 1993.

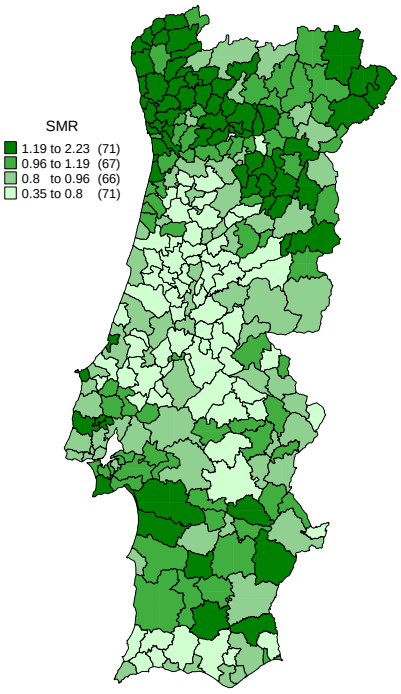


Figure 10.2: Stomach cancer SMR, 1985-1998.

equals the total number of registered deaths. Maximal and minimal expected numbers of cases per year at the county level are respectively 284 and 0.85. The observed counts and the expected number of cases were used to calculate standard mortality ratios (section 1.1.6). Maximal and minimal SMRs are respectively 0.35 and 2.23. The SMRs in each county are displayed in figure 10.2.

10.3 Socio-economic level exposure coefficient

We are specifically interested in investigating quantitatively the association between the mortality of stomach cancer and spending power. The aim of the analysis is to assess the existence or not of spatial patterns of this association. We use the spatially varying regression model presented in chapter 8 for this assessment (section 10.3.1). We compare the results with those obtained in models in which the exposure coefficient is constant or absent (section 10.3.2). In all models, the observed areal counts are assumed to follow a Poisson distribution, as in expression 2.10,

$$Y_i \sim \text{Poisson}(E_i\theta_i),$$

where Y_i is the stomach cancer death count and E_i is the expected number of deaths in county i. We consider different predictors reflecting different model assumptions.

To facilitate the presentation of the numerical results, we rescaled the spending power index, described in section 10.1.4, by dividing all values by 100. The average value of the resulting covariate is thus one.

10.3.1 Spatially varying regression

In this section, the results of applying the spatially varying regression model developed in chapter 8 to the stomach cancer mortality data are discussed. Formally,

the model fitted assumed the following linear predictor:

$$\begin{aligned} \text{Model 1: } \log(\theta_i) &= \beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i \\ b_{i0} &= \phi\rho_c z_{i1} + \phi\sqrt{1-\rho_c^2}z_{i0} \\ b_{i1} &= z_{i1} \\ z_{\bullet 0} &\sim \text{IAR}(\sigma^2) \\ z_{\bullet 1} &\sim \text{IAR}(\sigma^2) \\ z_{\bullet 0} &\amalg z_{\bullet 1}, \end{aligned} \tag{10.1}$$

where x_i corresponds to the spending power index in county i, and IAR stands for intrinsic autoregression.

The model was implemented in Winbugs (appendix J). This program imposes the constraints $\sum_i z_{i0} = 0$ and $\sum_i z_{i1} = 0$, thus making the intrinsic autoregression priors proper. This is equivalent to imposing the constraints $\sum_i b_{i0} = 0$ and $\sum_i b_{i1} = 0$. We used a weakly informative prior, $\text{Gamma}(0.05, 0.00005)$, for the precision parameters (inverse of the variance σ^2) of the intrinsic autoregressions, and vague Gaussian priors for the fixed effects β_0 and β_1. A non-informative $\text{Uniform}(-1, 1)$ prior was chosen for ρ_c. To avoid numerical difficulties, a more informative prior, $\text{Gamma}(1, 1)$, was used for ϕ. We also run the simulation with two alternative priors for ϕ, a $\text{Gamma}(0.5, 1)$ and a $\text{Gamma}(1, 0.5)$, favouring respectively larger and smaller variances for the random intercept $\boldsymbol{b}_{\bullet 0}$. The results obtained with all these priors were similar. Another Gamma prior was also considered for the precision $1/\sigma^2$, a $\text{Gamma}(0.001, 0.001)$, but this choice lead to minor differences on the results. Only the results using a $\text{Gamma}(1, 1)$ prior for ϕ and a $\text{Gamma}(0.05, 0.00005)$ for the precision parameter are presented in this section.

Three MCMC simulations with different initial values were run. Figure 10.3 displays the Brooks-Gelman-Rubin statistics (section 7.13) for selected parameters. Figures 10.4 and 10.5 show the sampled values for the same selected parameters. The chains appear to have converged after 50000 iterations.

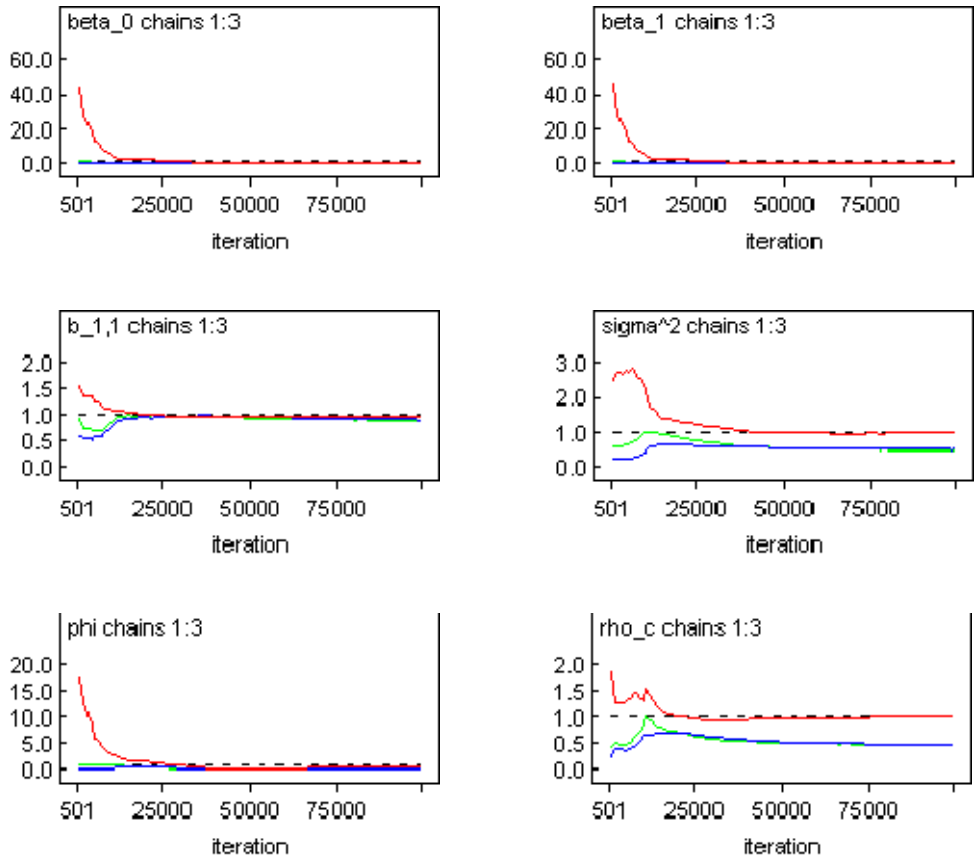


Figure 10.3: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for the parameters β_0, β_1, $b_{1,1}$, σ^2, ϕ and ρ_c when fitting a spatial regression (model 1).

One MCMC simulation was then run for 100000 iterations, with the first 50000 discarded as the burn-in period. Summary measures of the posterior distributions of the model parameters are presented in table 10.1. Maps of the

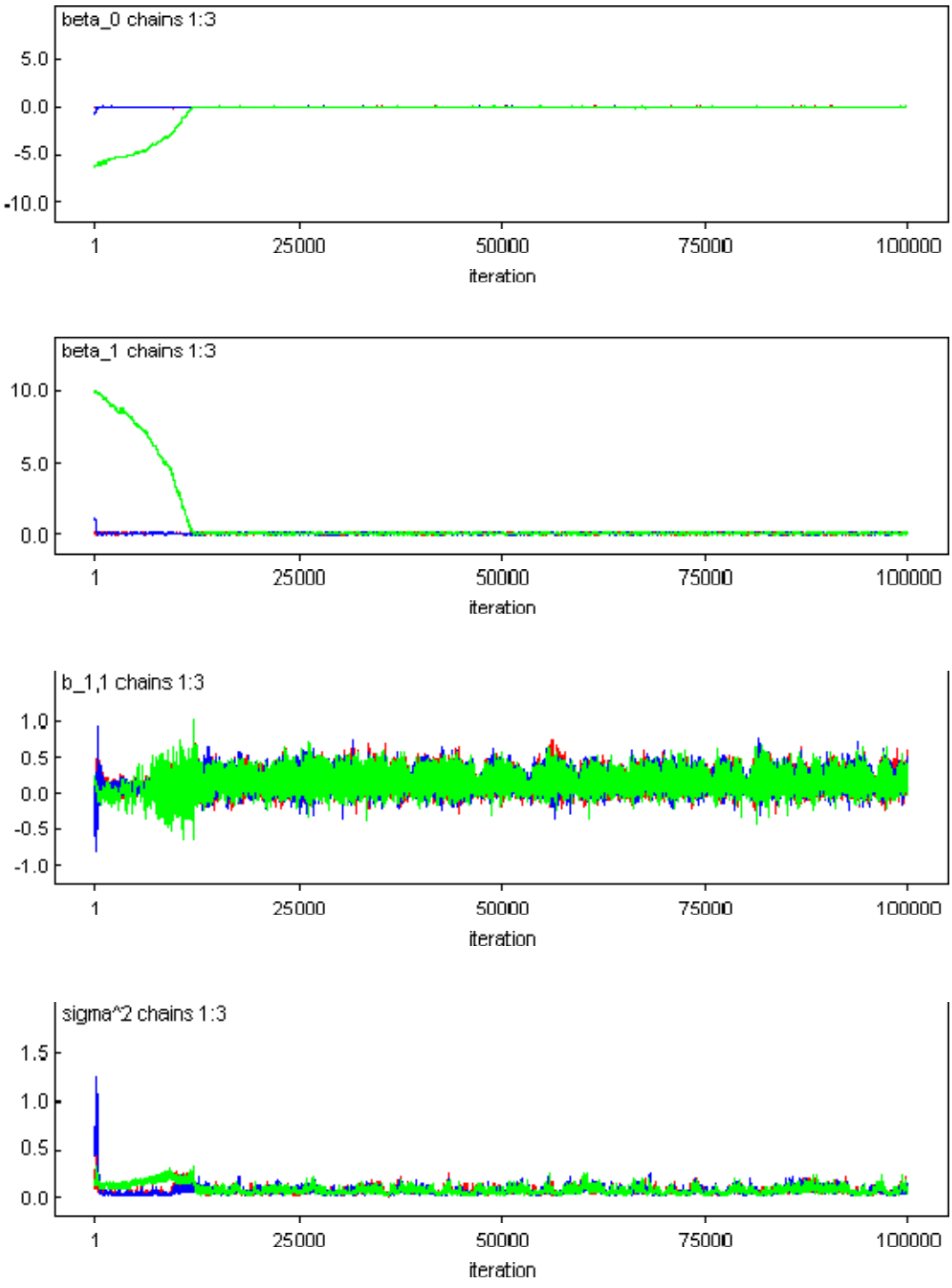


Figure 10.4: Sampled values of the parameters β_0, β_1, $b_{1,1}$ and σ^2 when fitting a spatial regression (model 1) in three MCMC runs.

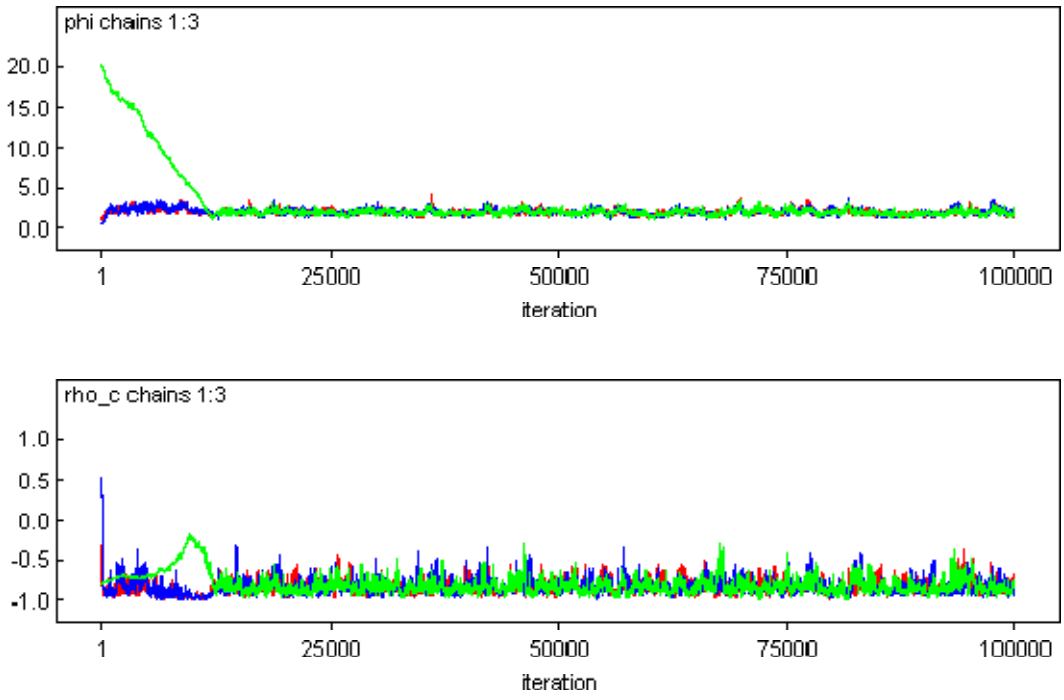


Figure 10.5: Sampled values of the parameters ϕ and ρ_c when fitting a spatial regression (model 1) in three MCMC runs.

posterior means of the regression coefficients are displayed in figure 10.6. The estimated exposure coefficients take both positive and negative values. In some areas in the south and the north-east, there is a strong negative relationship whereas in most areas in the centre the relationship is positive. The former is perhaps the relation that one would intuitively expect to observe: spending power as a protective factor for *H. pylori* infection and consequently for stomach cancer. It is possibly not by chance that the areas with negative exposure coefficient coincide with those with higher stomach cancer risk. These are perhaps the areas in which the virulence of the strain is higher, and where, as a result, the protective factor given by a higher spending power may have a bearing on the stomach cancer rates.

	Mean	Standard deviation	Quantiles 2.5%	Median	97.5%
β_0	-0.039	0.027	-0.092	-0.039	0.013
β_1	0.003	0.040	-0.073	0.002	0.085
ϕ	2.4	0.9	1.4	2.2	5.0
ρ_c	-0.84	0.15	-0.98	-0.88	-0.45
σ^2	0.04	0.03	0.01	0.04	0.12

Table 10.1: Summary measures of the posterior distributions of the parameters of the spatial regression (model 1).

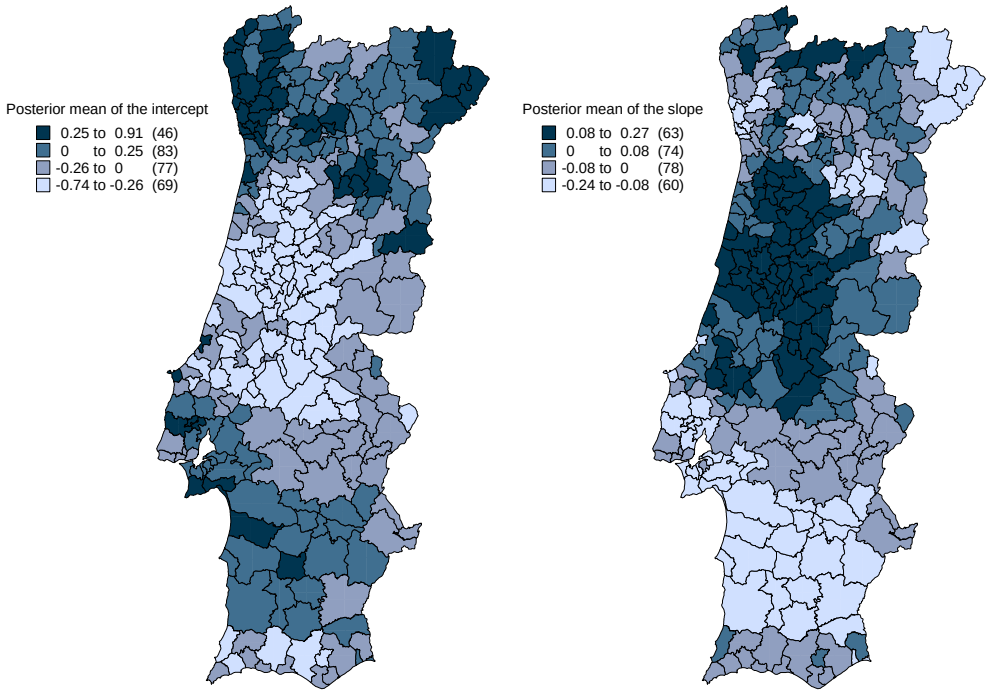


Figure 10.6: Posterior means of the areal baseline risks (left) and the areal exposure coefficients of the association between stomach cancer mortality and spending power (right) when fitting a spatial regression (model 1).

Induced variation of the intercept

According to section 8.5.3, the proportion of the variability of the random intercept induced by the varying exposure coefficient can be measured by ρ_c^2, whose posterior mean was estimated as 0.7.

10.3.2 Alternative models

First, we considered a model that accounts for overdispersion and spatial correlation of the areal risks but does not include any covariate. The model includes a random effect u for unstructured heterogeneity and a random effect v for structured heterogeneity,

$$\begin{aligned} \text{Model 2: } \log(\theta_i) &= \beta_0 + b_{i0} \\ b_{i0} &= v_i + u_i \\ u_i &\sim \mathrm{N}(0, \sigma_u^2) \\ v_\bullet &\sim \mathrm{IAR}(\sigma_v^2), \end{aligned} \tag{10.2}$$

where IAR stands for intrinsic autoregression and $v_\bullet = (v_1, \ldots, v_n)$.

Secondly, we used a link function which contains only fixed effects. In particular, the model assumes a constant exposure coefficient:

$$\text{Model 3: } \log(\theta_i) = \beta_0 + \beta_1 x_i, \tag{10.3}$$

where x_i is the spending power index in area i.

Thirdly, we consider a predictor including a constant exposure coefficient and a random intercept accounting for structured and unstructured variability. This corresponds to the model

$$\begin{aligned} \text{Model 4: } \log(\theta_i) &= \beta_0 + b_{i0} + \beta_1 x_i \\ b_{i0} &= v_i + u_i \\ u_i &\sim \mathrm{N}(0, \sigma_u^2) \\ v_\bullet &\sim \mathrm{IAR}(\sigma_v^2), \end{aligned} \tag{10.4}$$

where IAR stands again for intrinsic autoregression, $v_\bullet$ for $(v_1, \ldots, v_n)$ and x_i for the spending power index.

In all models 2 to 4, the fixed effects were given vague priors. A vague prior, $\mathrm{Gamma}(0.001, 0.001)$, was used for the precision parameters of the prior distributions of the random effects u and v, when these terms were included.

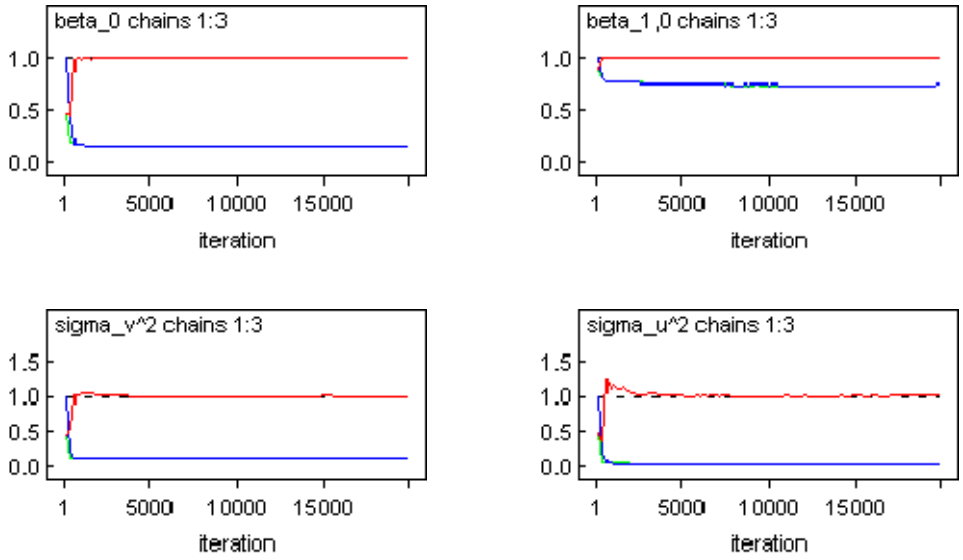


Figure 10.7: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for selected intercept parameters β_0 and $b_{1,0}$ and variances σ_v^2 and σ_u^2 of model 2.

The Winbugs software was used to run all simulations. Convergence of the MCMC runs was assessed by inspecting the sampled values and the Brooks-Gelman-Rubin convergence statistics (section 7.13) in three MCMC chains, each starting with a different set of initial values. Figures 10.7 to 10.12 present the sampled values and Brooks-Gelman-Rubin convergence statistics of these three MCMC runs for selected parameters of models 2 to 4. The Brooks-Gelman-Rubin convergence statistics, whose values are plotted in red colour, were drawn immediately to one despite the disparities in the initial values. The lengths of

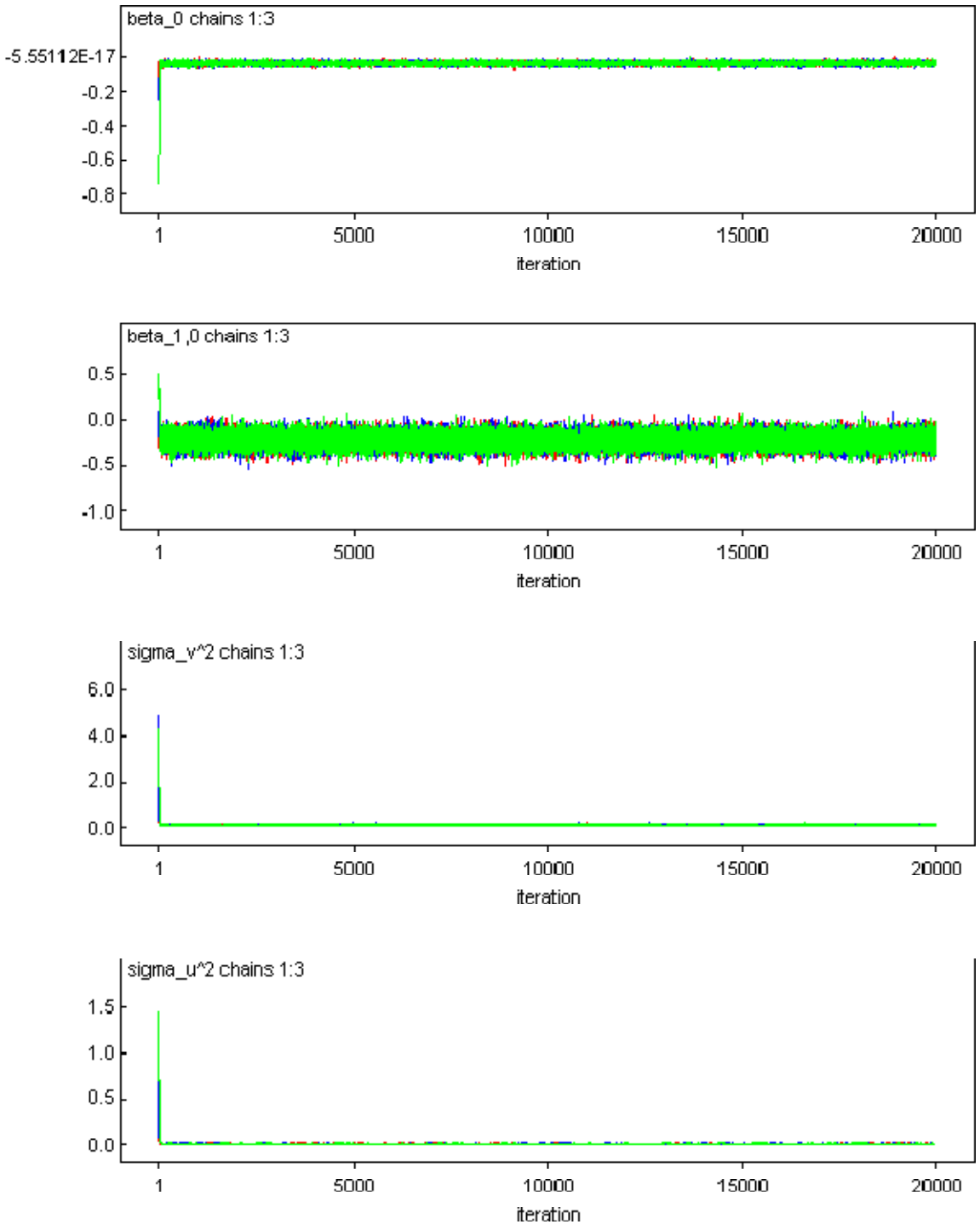


Figure 10.8: Sampled values of selected intercept parameters β_0 and $b_{1,0}$ and variances σ_v^2 and σ_u^2 of model 2 in three MCMC runs.

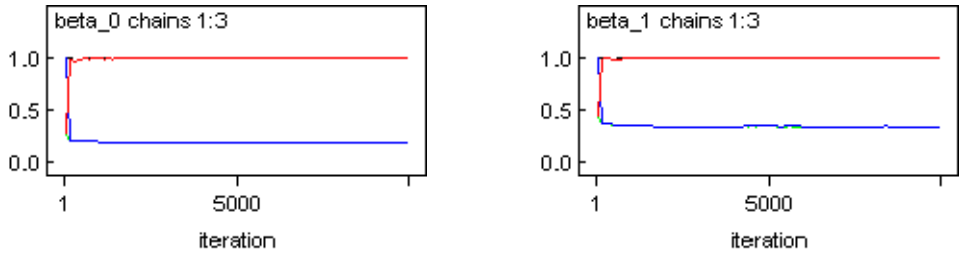


Figure 10.9: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for the intercept β_0 and the slope β_1 of model 3.

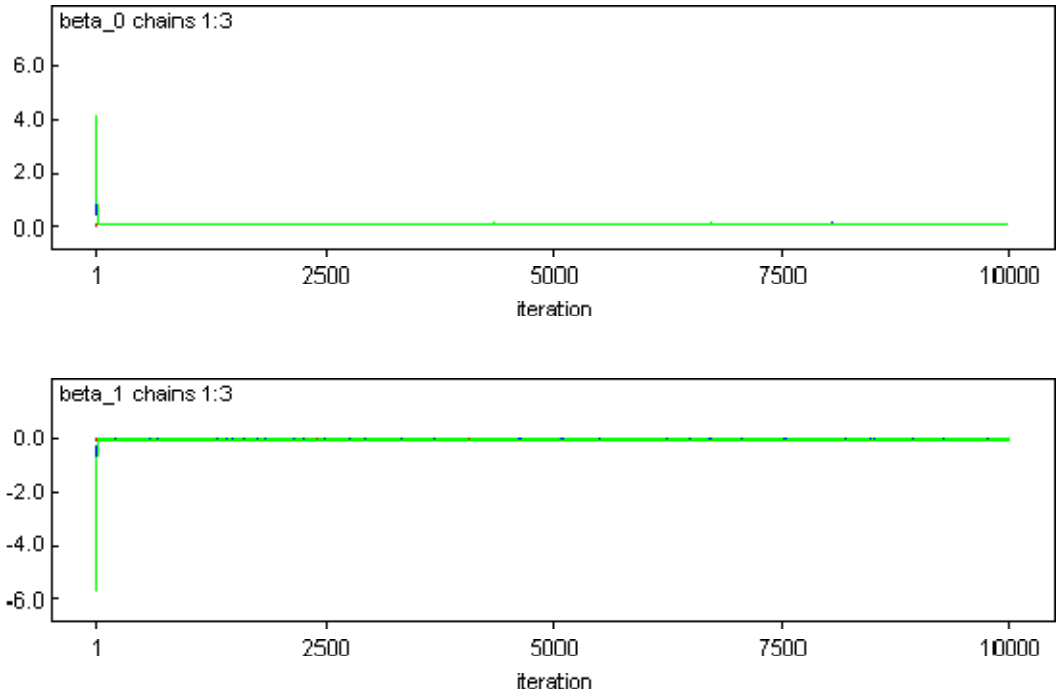


Figure 10.10: Sampled values of the intercept β_0 and the slope β_1 of model 3 in three MCMC runs.

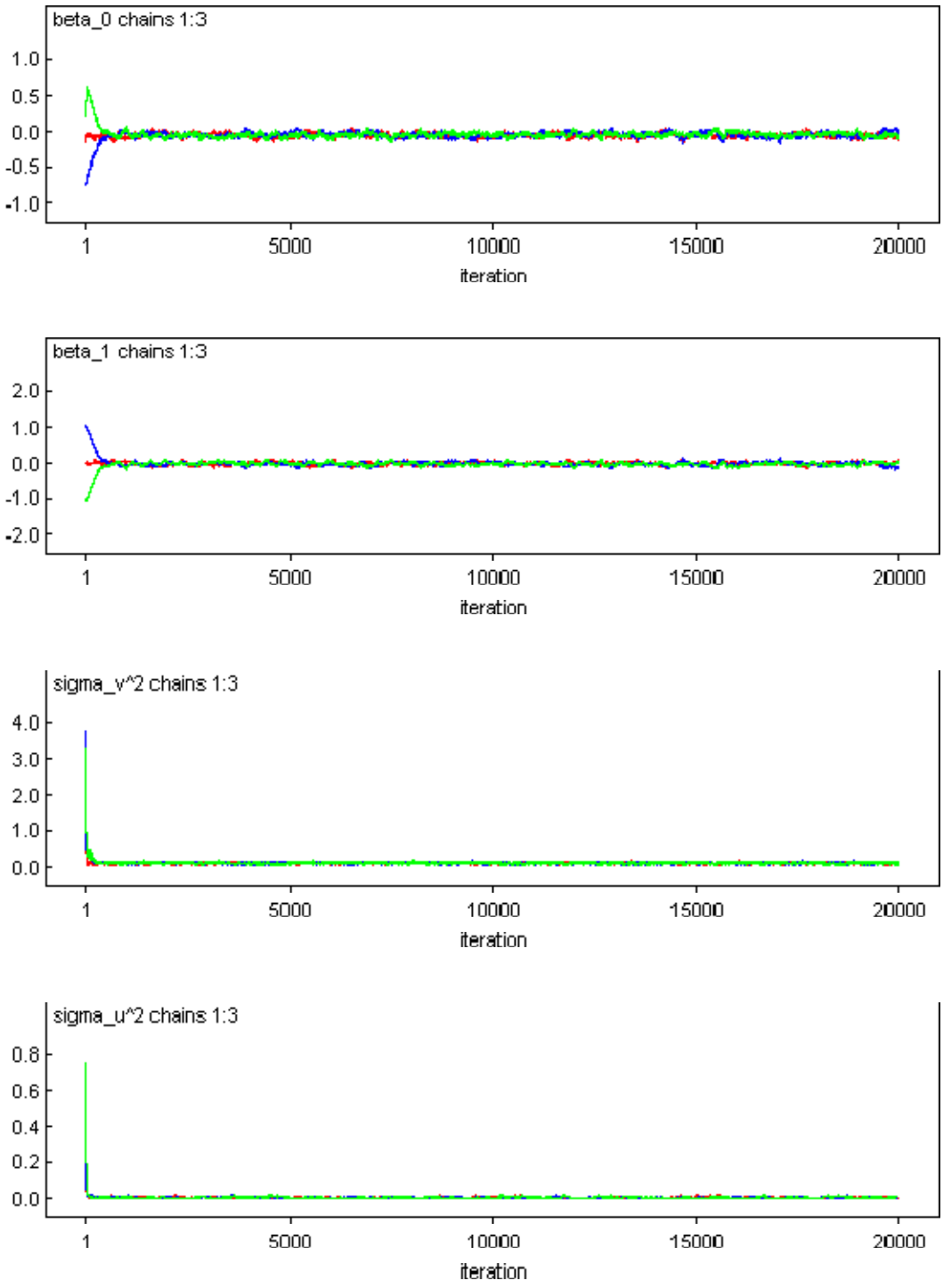


Figure 10.11: Sampled values of the intercept β_0, the slope β_1, the variances σ_v^2 and σ_u^2 of model 4 in three MCMC runs.

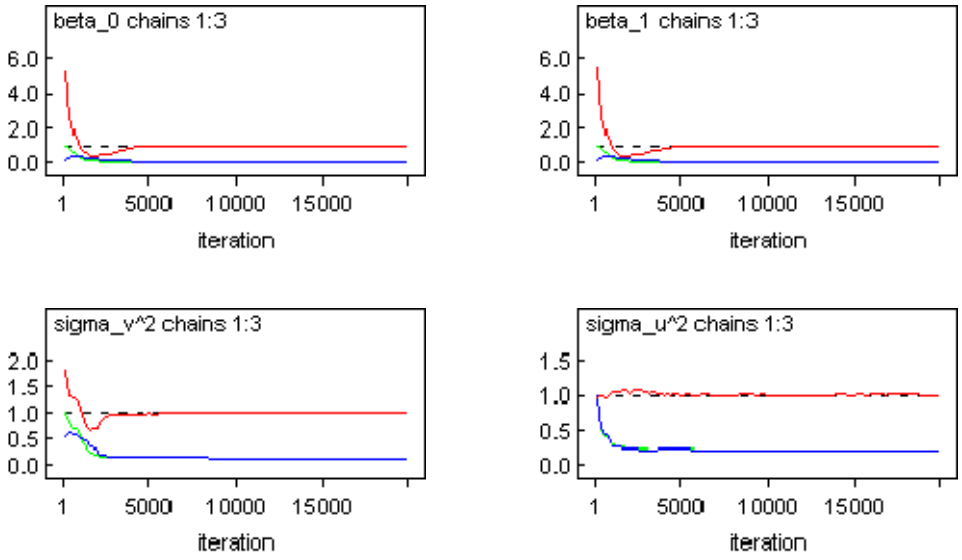


Figure 10.12: Brooks-Gelman-Rubin convergence statistics (red), length of the empirical central 80% interval of the pooled runs (green) and average length of the empirical central 80% intervals of the individual runs (blue) for the intercept β_0, the slope β_1, the variances σ_v^2 and σ_u^2 of model 4.

the central 80% intervals for the pooled runs (in green) and the average lengths of the central 80% intervals for the three MCMC runs (in blue) appeared to have stabilized at most after 10000 iterations in all cases. The Monte Carlo errors showed reasonable values in the following 10000 iterations. Therefore, the MCMC simulation was run for 20000 iterations, with the first 10000 discarded as the burn-in period.

Results

Table 10.2 presents posterior means and credible intervals for the model with the random intercept and no covariate (model 2), the model of fixed effects (model 3) and the model with a fixed slope and a random intercept (model 4). To

summarize the relative contributions of the spatial random effect v to the total overdispersion $u + v$ in models 2 and 4, we calculated the posterior distribution of the quantity (Mollié, 1996),

$$\psi \;=\; \frac{\sigma_v^2/n_\partial}{(\sigma_v^2/n_\partial) + \sigma_u^2},$$

where n_∂ is the average number of neighbours throughout the region.

Without the presence of random effects the spending power seems to have a protective effect, in line with what is expected from general disease patterns. However, there is an attenuation in the coefficient of the covariate when the random effects are incorporated. Further, the variance components are practically the same in models 2 (without the covariate) and 4 (with the covariate). This suggests that the covariate is not able to fully explain the variability of the stomach cancer mortality rates.

In the two models that include random effects (models 2 and 4), the value of ψ was found to have a posterior mean of about 0.9 and a 95% credible interval given by $(0.61, 0.98)$. Therefore, spatial variation appears to have a greater influence on the relative risks than unstructured heterogeneity does.

Model	Intercept	Exposure coefficient	σ_u	σ_v	ψ
2	-0.042	-	0.05	0.32	0.86
	(-0.058, -0.026)		(0.02, 0.10)	(0.26, 0.37)	(0.61, 0.98)
3	0.072	-0.072	-	-	-
	(0.055, 0.089)	(-0.086, -0.058)			
4	-0.046	0.006	0.06	0.32	0.85
	(-0.101, 0.013)	(-0.082, 0.09)	(0.02, 0.10)	(0.26, 0.37)	(0.61, 0.98)

Table 10.2: Posterior mean and 95% credible interval (in brackets) for the fixed intercept, the fixed slope, the parameters σ_u and σ_v, and the fraction of conditional excess variation attributable to spatial effects (ψ).

10.4 Model choice

Spiegelhalter et al. (2002) proposed the use of the *deviance information criterion* (DIC) for comparing Bayesian hierarchical models. The criterion adapts the well known Akaike information criterion (AIC) (Akaike, 1974), which is a combined measure of complexity and fit.

AIC uses the number of parameters as a measure of complexity, but in Bayesian hierarchical models this number may not entirely reflect the complexity of the model. The prior distribution can introduce dependencies among parameters thus decreasing the dimension of the model. Spiegelhalter et al. (2002) suggested an alternative measure of complexity, p_D, defined as the posterior expectation of the deviance minus the deviance evaluated at the posterior mean of the model parameters. The DIC was then defined as:

$$DIC = p_D + \bar{D}, \tag{10.5}$$

where $\bar{D}$ is the posterior expectation of the deviance, which in turn is defined as minus twice the logarithm of the likelihood. Smaller values of $\bar{D}$ correspond to better model fits. The minimum DIC estimate indicates a preferred model.

Although the interpretation of the DIC criterion is not yet fully understood, it has produced consistent results in previous applications with exponential family likelihoods (Spiegelhalter et al., 2002).

We calculated the DIC for comparing the models applied so far in this chapter to the Portuguese data set. The values obtained are displayed in table 10.3. The model including spatial exposure coefficient and baseline risk (model 1) seems to be superior, but only with a slight difference from the DIC of models 2 and 4, which account for overdispersion but not for a varying exposure coefficient.

Model		DIC
1	Spatial exposure coefficient and baseline risk	2151
2	Overdispersion	2190
3	Constant exposure coefficient	4500
4	Constant exposure coefficient and overdispersion	2191

Table 10.3: Deviance information criterion (DIC) over all models.

We continue the analysis of this chapter using model 1 (section 10.3.1), which includes a spatial baseline risk and a spatial exposure coefficient and shows the lowest value of the DIC.

10.5 Attributable risk

The posterior mean of the exposure coefficient was negative for 138 counties. For these counties, the model therefore estimated that the spending power is a protective factor for stomach cancer. The highest spending power observed in these counties is 291, and the lowest 36. We estimate in this section how many cases could be avoided if the spending power were increased to 291 in all these counties.

With this aim, for each county i, we compared the estimated risk, as given by $\theta_i = \exp(\beta_0 + b_{i0} + (\beta_1 + b_{i1})x_i)$, with the estimated risk if the spending power level was 291, i.e. $\theta_{SP=291} = \exp(\beta_0 + b_{i0} + 2.91(\beta_1 + b_{i1}))$.

The attributable risk was defined in section 9.9, equation 9.4. The incidence in the population I_P corresponds to $E_i\theta_i$, while that of the population non-exposed to lower spending power level I_{NE} can be estimated as $E_i\theta_{SP=291}$. The risk attributable to spending power lower than 291 can then be estimated by,

$$AR = \frac{e^{\beta_0+b_{i0}+(\beta_1+b_{i1})x_i} - e^{\beta_0+b_{i0}+2.91(\beta_1+b_{i1})}}{e^{\beta_0+b_{i0}+(\beta_1+b_{i1})x_i}}.$$

A map of the estimated attributable risk is given in figure 10.13. The attributable

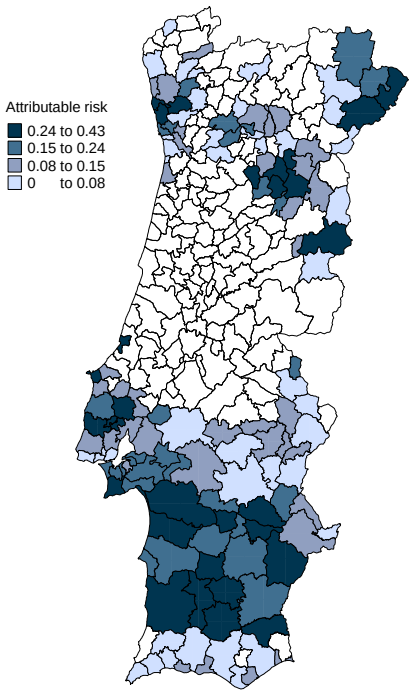


Figure 10.13: Estimated attributable risk due to lower spending power for the 138 counties for which spending power was estimated to have a protective effect.

risk varies from 0.0 to 0.42. By multiplying each county attributable risk by the observed number of cases in each county, we can obtain an estimate of the excess number of cases associated with lower levels of spending power. The total number of excess deaths for the 138 counties is 3310, which correspond to 13 per cent of the total number of deaths observed in these counties in the fourteen years considered in the study.

10.6 Discussion

The results obtained with the four models applied in this chapter were very different, thus indicating that the conclusions depend substantially on the assumptions taken. Nevertheless, spatial correlation seems to be a distinctive characteristic of the data analysed, with models accounting for this phenomenon showing much better fits.

The model including a spatially varying exposure coefficient showed a superior fit. Attributable risks calculated on the basis of this model suggested that more than ten per cent of the cases may be avoided by improving the socio-economic level, as reflected by spending-power, of the counties in which an increasing spending power seems to act as a protective effect.

Since socio-economic status is usually a protective factor for stomach cancer (chapter 3), we restricted the calculation of the attributable risks to counties in which the posterior mean of the exposure coefficient was estimated as negative.

We have tried to restrict all exposure coefficients to negative values, by using transformations of the prior distribution. However, the resulting models were difficult to implement with an MCMC approach. Future research may wish to focus on spatial models for negative variables.

Many factors usually contribute to the development of the disease. While we believe that the findings of this chapter provide enriching clues for the potential role of socio-economic level in stomach cancer, other confounding factors may also play a role in the associations between socio-economic factors and stomach cancer.

In the study presented in chapter 2, it was suggested that under-dispersion may occur when the within-area variability is large. Under-dispersion was not

observed in the data studied in this chapter, which on the contrary shows evidence of considerable over-dispersion (large values of the residual deviance in the fixed effects model presented in section 10.3.2 - results not shown).

Finally, we note that the models considered in this chapter do not violate the necessary conditions for coding invariance which were established in section 5.4.2.

Chapter 11

Conclusion

In this thesis we considered the use and implementation of regression mixtures, clustering models and spatially varying regressions for assessing geographical patterns of the exposure coefficient in studies of stomach cancer. While developing the methodology for these studies, we provided necessary conditions for coding invariance in models which include varying exposure coefficients and explained why certain models are not coding invariant. In particular, we showed that models which assume independence between the random intercept and the random slopes are not coding invariant. We also explained that spatial variation of the exposure coefficient can induce spatial variation of the baseline risk.

In chapter 7, we adapted the clustering model of Knorr-Held and Rasser (2000) for identifying clusters of similar exposure coefficients on the basis of any distance measure between adjacent areas. In particular, we applied the model using the Euclidean distance. For this purpose, we developed an algorithm for calculating the length of the shortest path between any two areas. The simulation study indicated that the clustering model is preferable to the regression mixture model when the aim is to identify connected sets and the observations present

large variance.

In chapter 8, we proposed the use of the multivectorial intrinsic autoregressions of Gamerman et al. (2003) for modelling the random effects of spatially varying regressions. We showed that the bivectorial intrinsic autoregressions can be obtained as a limit of two correlated conditional autoregressions. This result can be generalized for multivectorial intrinsic autoregressions of any dimension. In addition, we proved that multivectorial intrinsic autoregressions can be expressed as a linear transformation of independent univectorial intrinsic autoregressions, and we used this result for developing an original strategy for implementing the model. This strategy is less complex than other methods. The results of our simulation study indicated that the model can be useful for suggesting spatial patterns of the exposure coefficients. However, many questions still remain open regarding the reliability of the estimates obtained with the model.

In chapters 9 and 10, we analysed two data sets on stomach cancer. In the first data set, we investigated the influence of the prevalence of *Helicobacter pylori* infection on stomach cancer risk while accounting for the consequences of potential differences in *H. pylori* virulence and in human sub-populations. Clustering models were used for identifying groups of countries with similar behaviour. Three clusters were found. The ratio of stomach cancer risk to *H. pylori* infection seems to be highest in Far East Asia, relatively high in Europe and very low in Africa and South-East Asia. In the latter two regions, there is very little evidence of an association between infection and the incidence of stomach cancer, even in countries where the prevalence of *H. pylori* infection is high. In the other regions, the evidence for association is strong. It is hoped that this study represents a step forward on the mapping of host-pathogen interaction. The results obtained

can be used as a basis for choosing countries for future genetic studies of *H. pylori* infection and human populations. Such studies are usually carried out by comparing the strains in two, or at most, three countries. By choosing countries that appear to have distinct human-*H. pylori* interaction it may be easier to identify the responsible genes, in the human and bacterial sub-populations.

In chapter 10, we assessed the variation of the association between spending power - a usually protective factor for *H. pylori* infection - and stomach cancer deaths in Portugal. Spatial regression models were used to investigate the spatial variation of the exposure coefficient. The results were compared with alternative models in which the exposure coefficient was constant or excluded. This comparison selected the spatial varying exposure coefficient model as the one with the best fit. Although these results seem to support the existence of a non-constant exposure coefficient, the nature of the data analysed do not permit drawing firm conclusions. More reliable results could be achieved if more data were available per area.

Overall, the results presented in this thesis suggest that geographic information can be useful in the study of varying exposure coefficients, when these risks are expected to vary according to a certain spatial pattern. In addition, the analyses showed that ignoring the existence of different risk-disease associations in aetiological studies can lead to very different conclusions. If there is background knowledge to support the existence of different values of the risk-disease association, this hypothesis should be investigated.

Future work could consider more closely the evaluation of the models studied in this thesis and focus on ways of assessing the contribution of the exposure coefficient variation to the total areal risk variation. Another issue which would

benefit from more research concerns the identifiability of the model parameters in varying exposure coefficient models.

The models studied in this thesis may be extended in useful ways. First, the number of clusters could be included as one of the unknown parameters in the Bayesian analysis of the geographical clustering model. This would lead to a more comprehensive assessment of the number of clusters. Secondly, the spatial models for the exposure coefficients could be restricted to negative values. These models would be appropriate for cases in which the exposure studied is known to have a protective effect.

Appendix A

Proof of theorem 2.2.1

First of all, since $\bar{p} \in [0,1]$, there is always a unique $m \in \{0,1,\ldots,n-1\}$ such that $\bar{p} \in [\frac{m}{n}, \frac{m+1}{n})$. Secondly, the solution presented in theorem 2.2.1 satisfies the constraints (2.5):

$$\frac{1}{n}\sum_{k=1}^{n} p_k \;\; = \;\; \frac{n\bar{p} - m + m}{n} = \bar{p}$$

and (2.6):

$$m \leq n\bar{p} < m+1 \;\; \Rightarrow \;\; 0 \leq n\bar{p} - m < 1.$$

The proof will continue as follows. First, it is proved that the maximum always lies on the borders of the set defined by the constraints of this problem. Then, the result will be proved by induction.

The constraint (2.5) defines a hyperplane H_{n-1} of dimension $n-1$ in a space of n dimensions and perpendicular to the n-dimensional unity vector $(1,\ldots,1)$. The constraint (2.6) defines an n dimensional cube, which we denote by C_n. Therefore, we are looking for maxima in the bounded hyperplane H^*_{n-1} defined by the intersection of H_{n-1} and C_n.

The method of Lagrange multipliers gives one stationary point in H_{n-1}, $p_k = \bar{p}$, for all k. This stationary point clearly is a minimum of s^2 because $s^2 \geq 0$ and $s^2 = 0$ at this stationary point.

Since there is no other stationary point, and this point is a minimum, it follows that s^2 will take a maximum value at the border of H^*_{n-1}. Otherwise, the method of Lagrange multipliers would have indicated other stationary points in H_{n-1}.

The borders of H^*_{n-1} are always on the hyperplanes defined by $p_k = 0$ or $p_k = 1$ because these are the borders of C_n. Note that:

- $\bar{p} < 1/n$ (i.e. $m = 0$) implies that all p_k are smaller than 1, and therefore the maximum is on the hyperplanes $p_k = 0$;
- $\bar{p} > (n-1)/n$ (i.e. $m = n-1$) implies that all p_k are larger than 0, and therefore the maximum is on the hyperplanes $p_k = 1$.

We now prove by induction that the point given in theorem 2.2.1 is the maximum of s^2 over the borders of H^*_{n-1}. The first part of the proof by induction proves that the result is true for $n = 2$, and then that if the result is true for n then it is true for $n + 1$.

Result for $n = 2$

When $n = 2$, the border is defined by two points, the extremities of a finite line. Two cases must be considered: when $\bar{p} \in [0, 1/2]$, the border points are $(0, 2\bar{p})$ and $(2\bar{p}, 0)$ and the value of s^2 at these points is $s^2 = \bar{p}^2$; when $\bar{p} \in [1/2, 1]$, the border points are $(1, 2\bar{p} - 1)$ and $(2\bar{p} - 1, 1)$ and the value of s^2 at these points is $s^2 = (1 - \bar{p})^2$. Both values of s^2 satisfy expression 2.7 in theorem 2.2.1.

Result for n and $n+1$

The problem for $n+1$ dimensions is to maximize

$$s^2 \;=\; \frac{1}{n+1}\sum_{k=1}^{n+1}(p_k-\bar{p})^2$$

subject to

$$\bar{p}=\frac{1}{n+1}\sum_{k=1}^{n+1}p_k \;=\; \bar{p}$$

Consider the hyperplanes $p_k=0$ and the values $m=0,\ldots,n-2$. Due to the symmetry of the problem it is enough to find the maximum value at one of the hyperplanes, e.g. $p_{n+1}=0$. Over this hyperplane, the maximization problem reduces to

$$\max\ \frac{1}{n+1}[\sum_{k=1}^{n}(p_k-\bar{p})^2+\bar{p}^2]$$

subject to

$$\frac{1}{n+1}\sum_{k=1}^{n}p_k=\bar{p}\ .$$

Since

$$\sum_{k=1}^{n}(p_k-\bar{p})^2=\sum_{k=1}^{n}(p_k-\frac{(n+1)}{n}\bar{p})^2+\frac{\bar{p}^2}{n}, \qquad \text{(A.1)}$$

this problem is equivalent to:

$$\max\ \frac{1}{n}\sum_{k=1}^{n}(p_k-\frac{n+1}{n}\bar{p})^2$$

subject to

$$\frac{\sum_{k=1}^{n}p_k}{n} \;=\; \frac{(n+1)\bar{p}}{n}.$$

Noting that

$$\frac{m}{n+1}\le\bar{p}\le\frac{m+1}{n+1} \;\Rightarrow\; \frac{m}{n}\le\frac{(n+1)\bar{p}}{n}\le\frac{m+1}{n}$$

and using the result from the induction hypothesis, the solution obtained is

- $p_1 = (n+1)\bar{p} - m$,
- $p_k = 1, k = 2, \ldots, m+1$,
- $p_k = 0, k = m+2, \ldots, n$.

Since $p_{n+1} = 0$, this solution is the same as that given in the theorem 2.2.1.

Now consider the case $m = 1, \ldots, n-1$ and the hyperplanes $p_k = 1$. Again, it is only necessary to consider one of the hyperplanes because of the symmetry of the problem. Let's choose the hyperplane $p_{n+1} = 1$. Since

$$(n+1)\bar{p} = \sum_{k=1}^{n+1} p_k = 1 + \sum_{k=1}^{n} p_k \iff \frac{\sum_{k=1}^{n} p_k}{n} = \frac{(n+1)\bar{p} - 1}{n}$$

and

$$\sum_{k=1}^{n} (p_k - \bar{p})^2 \propto \sum_{k=1}^{n} (p_k - \frac{(n+1)\bar{p} - 1}{n})^2 + \text{constant}, \tag{A.2}$$

the problem is then to maximize

$$\frac{1}{n} \sum_{k=1}^{n} (p_k - \frac{(n+1)\bar{p} - 1}{n})^2$$

subject to

$$\frac{\sum_{k=1}^{n} p_k}{n} = \frac{(n+1)\bar{p} - 1}{n} \ .$$

Since

$$\frac{m}{n+1} \leq \bar{p} \leq \frac{m+1}{n+1} \quad \Rightarrow \quad \frac{m-1}{n} \leq \frac{(n+1)\bar{p} - 1}{n} \leq \frac{m}{n}$$

the result from the induction hypothesis can be applied for $m-1$ and provides the solution

- $p_1 = (n+1)\bar{p} - m$,
- $p_k = 1, k = 2, \ldots, m$,

- $p_k = 0, k = m+1, \ldots, n.$

As $p_{n+1} = 1$, and the order of the ps is irrelevant, this solution is the same as the one obtained when we started with $p_{n+1} = 0$.

This solution is thus the maximum over the borders of H_n^*, and so the theorem has been proved by induction.

Appendix B

World standard population

Age group (years)	World population	Age group (years)	World population
0-4	12 000	**45-49**	6 000
5-9	10 000	**50-54**	5 000
10-14	9 000	**55-59**	4 000
15-19	9 000	**60-64**	4 000
20-24	8 000	**65-69**	3 000
25-29	8 000	**70-74**	2 000
30-34	6 000	**75-79**	1 000
35-39	6 000	**80-84**	500
40-44	6 000	**85+**	500
Total	100 000		

Table B.1: World standard population.

Appendix C

Calculating the boundary distance

Here we give one of the programs written for S-PLUS for calculating the boundary distance between any two areas (i.e. the path crossing the smallest number of boundaries).

```
N<-49                                          # Number of areas
Nbound <- matrix(0,N,N)
Nweights0 <- spatial.weights(sim.ng$nhbr)      # Adjacency matrix
Nweights1 <- Nweights0
nr.bound <- 1                                  # Number of boundaries
count <- N*N
for (i in 1:N)
{ for (j in 1:N)
     {
if (Nbound[i,j]==0 && Nweights1[i,j]>0)
{
     Nbound[i,j] <- nr.bound
     count <- count-1
}
}
}

while (count>0)
     {
          Nweights1 <- Nweights0 %*%    Nweights1
          nr.bound <- nr.bound + 1
          for (i in 1:N)
          { for (j in 1:N)
               {
                    if (Nbound[i,j]==0 && Nweights1[i,j]>0)
                    {
                         Nbound[i,j] <- nr.bound
                         count <- count-1
                    }
}
}

}

for (i in 1:N)
{
     Nbound[i,i] <- 0
}
```

Appendix D

Calculating the geographical distance

Here we give one of the programs written for S-PLUS to calculate the geographical distance between any two areas. The correspondence between the variables used in the program and those in theorem 7.3.2 is as follows:

S-Plus program	Theorem 7.3.2
AdjDist	$\boldsymbol{W}$
D	$\boldsymbol{D}$
Dkminus1	$\boldsymbol{D}_{k-1}$
Dk	$\boldsymbol{D}_k$
Dpowerk	$\boldsymbol{D}^k$
k	k

```
N<-43                                              # Number of areas
FinalDist <- matrix(0, N, N)                       # Output matrix
Nweights <- spatial.weights(world43.ng$nhbr)       #  Adjacency matrix
D <- matrix(0, N, N)                               #  Distances between adjacent areas
Dkminus1 <- matrix(0, N, N)
Dpowerk <- matrix(0, N, N)
Dk <- matrix(0, N, N)

for (i in 1:N)
{ for (j in 1:N)
     {
     if (Nweights[i,j]>0)
     {
          D[i, j] <- ((Latitude[i]-Latitude[j])^2+ (Longitude[i]-Longitude[j])^2)^(1/2)
          FinalDist[i,j] <- D[i,j]
          Dkminus1[i,j] <- D[i,j]
          Dpowerk[i,j] <- D[i,j]
}}}

k <- 1                    #  Number of boundaries

while (k < N)
     {
          k <- k + 1
          Dpowerk <- Dpowerk %*% D

          for (i in 1:N)
          { for (j in 1:N)
               {
                    if (Dpowerk[i,j]==0) {Dk[i,j]=0}
                    if (Dpowerk[i,j]>0) {Dk[i,j] <- min((Dkminus1[i,]+D[,j])[(Dkminus1[i,]>0 & D[,j]>0)])}
          }}
          for (i in 1:N)
          { for (j in 1:N)
               {
                    if (Dk[i,j]>0)
                    {
                         if (FinalDist[i,j]==0){FinalDist[i,j] <- Dk[i,j]}
                         if (FinalDist[i,j]>0){FinalDist[i,j] <- min(Dk[i,j], FinalDist[i,j])}
          }}}
          Dkminus1 <- Dk
     }

for (i in 1:N) { FinalDist[i,i] <- 0 }        # Output matrix with the geographical distance
```

Appendix E

Proof of theorem 7.3.2

Theorem

Let $\boldsymbol{D}$ be the matrix with the distances between all pairs of adjacent areas:

$$\boldsymbol{D}(i,j) = \begin{cases} d(i,j) & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}$$

From the matrix $\boldsymbol{D}$, define $\boldsymbol{D}_k$ by recurrence,

$$\boldsymbol{D}_1(i,j) = \boldsymbol{D}(i,j),$$

$$\boldsymbol{D}_k(i,j) = \begin{cases} \Delta_{i,j}^k & \text{if } \boldsymbol{D}^k(i,j) > 0 \\ 0 & \text{otherwise} \end{cases}, k = 2,3,4,\ldots,$$

where

$$\Delta_{i,j}^k = \min_{\{s:\boldsymbol{D}_{k-1}(i,s)>0, \boldsymbol{D}(s,j)>0, s=1,\ldots,n\}}(\boldsymbol{D}_{k-1}(i,s) + \boldsymbol{D}(s,j))$$

and $\boldsymbol{D}^k$ is the k-th power of $\boldsymbol{D}$.

The entry (i,j) of the matrix $\boldsymbol{D}_k$ then contains the length of the shortest path that crosses k borders from i to j, or is equal to zero if such a path does not exist.

Proof (by induction)

We will argue by induction, but first note that the non-zero entries of the matrix $\boldsymbol{D}^k$ are the same ones as the non-zero entries of the matrix k-th power of the adjacency matrix, $\boldsymbol{W}^k$ (section 7.3.1).

- By definition of $\boldsymbol{D}$, the result is trivial for k=1.
- Now assume that the proposition is true for k. By definition of $\boldsymbol{D}$, if $\boldsymbol{D}^{k+1}(i,j) = 0$, then there is no path from i to j crossing $k+1$ boundaries, and therefore the proposition is true for all zero entries of $\boldsymbol{D}_{k+1}$. If $\boldsymbol{D}^{k+1}(i,j) > 0$, then there is at least one path from i to j crossing $k+1$ boundaries. If $\boldsymbol{D}_k(i,s) > 0$, by the induction hypothesis, $\boldsymbol{D}_k(i,s)$ is the length of the shortest path from i to s crossing $k-1$ boundaries. If $\boldsymbol{D}(s,j) > 0$, by definition of $\boldsymbol{D}$, s and j are adjacent, and $\boldsymbol{D}(i,s)$ is the distance between their centroids. Then, $\boldsymbol{D}_k(i,s) + \boldsymbol{D}(s,j)$ is the length of the shortest path from i to j that passes by s and crosses $k+1$ boundaries. By taking the minimum over all possible areas s for which $D(s,j) > 0$ - i.e. over all adjacent areas to the area j - we get the shortest path from i to j crossing $k+1$ boundaries.

Hence the result follows by induction.

Appendix F

Winbugs program - Regression mixtures

Here we give one of the programs written for Winbugs 1.4 to fit a regression mixture of two normal distributions.

```
model
	{
		for( i in 1 : N ) {
			y[i] ~ dnorm(lambda[i], tau)
			lambda[i] <- mu[i,T[i]]

		# Allocation variables
			T[i] ~ dcat(P[])

		for( j in 1 : J )
		{
			Post.proba[j,i] <- equals(T[i],j)
			mu[i,j] <- beta0+beta1[j]*(x[i])
		}
	}

	for (j in 1:J){    alpha[j]  <- 2;}
	P[1:J] ~ ddirch(alpha[])
	beta0 ~ dnorm(0, 1.0E-3)
	theta1 ~ dnorm(0, 1.0E-3)I(0.0, )
	beta1[2] <- beta1[1] + theta1
	beta1[1] ~ dnorm(0, 1.0E-3)I(0.0, )
	tau ~ dgamma(0.001, 0.001)
	sigma <- 1 / sqrt(tau)
	}

# Example of initial values
list(beta0=0, beta1=c(100, NA), tau=0.1, theta1=5, T = c(1, 1, 1, 1, 1, 1, 1, 1, 2, 2,2, 1, 1, 1, 1, 1,
1, 2, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1,1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1,
1, 2))

# Data
list(N=57, J=2, y = c( 2.71, 2.52, 2.15, 2.97, 2.51, 1.84, 3.05, 2.20, 2.05, 2.54, 2.20, 2.92, 2.29,
2.26, 2.71, 3.03, 2.21, 2.15, 2.77, 3.12, 3.42, 2.79, 2.58, 1.92, 2.17, 2.09, 2.20, 1.99, 3.29, 1.46,
3.89, 3.87, 2.24, 2.28, 1.61, 2.09, 1.36, 2.87, 1.87, 1.22, 2.41, 1.67, 1.90, 3.17, 2.67, 1.72, 2.21,
2.62, 2.75, 3.30, 3.29, 2.85, 1.28, 1.44, 1.84, 1.28, 2.22),
x = c(74, 47, 36, 60.4, 58.8, 44.4, 89.3, 36, 25.4, 39.2, 64.3, 63.3, 39.9, 43, 45.1, 78.5, 36.6, 36,
69.9, 82.8, 88, 54.8, 67.5, 29.6, 19, 81.2, 29.3, 38, 60.4, 87.5, 83.7, 69.6, 35.9, 80, 60.4, 38.6,
76.4, 73.9, 87.7, 88, 65.6, 75.8, 37.5, 72.2, 81.3, 32.7, 49.1, 72, 84.7, 65.5, 96.3, 92, 77.2, 84,
86.4, 73, 81))
```

Appendix G

Winbugs program - Clustering model

Here we give one of the programs written for Winbugs 1.4 to fit the clustering model with three clusters.

```
model
{
        for( i in 1 : N )
        {
                y[i] ~ ddexp(lambda[i], tau)
                lambda[i] <- mu[T[i],i]

                post.proba[1,i] <- equals(T[i],1)          # Posterior probabilities of classification
                post.proba[2,i] <- equals(T[i],2)
                post.proba[3,i] <- equals(T[i],3)

                for( j in 1 : J ) { mu[j,i] <- beta0+b1[j]*x[i] }
        }

        # Priors
        for (s in 1:N){p1[s]<- 1/N}
        for (s in 1:(N-1)){p2[s]<-1/(N-1)}
        for (s in 1:(N-2)){p3[s]<-1/(N-2)}

        aux1 ~ dcat(p1[])
        aux2~dcat(p2[])
        aux3 ~ dcat(p3[])

        g[1] <- aux1
        g[2] <- aux2 + step(aux2-g[1])   # they are the clusters centers
        g[3] <- aux3 + step(aux3-min(g[1],g[2]))+ step(aux3+ step(aux3-min(g[1],g[2]))-
                max(g[1],g[2]))

        for ( i in 1:N)
        {
                c1[i] <- min(min(Nbound[i,g[1]],Nbound[i,g[2]]), Nbound[i,g[3]])

                # allocating observations to each of the clusters defined by the cluster centers g
                T[i] <- equals(Nbound[i,g[1]],c1[i]) + 2*equals(Nbound[i,g[2]],c1[i])*(1-
                        equals(Nbound[i,g[1]],Nbound[i,g[2]])) + 3*equals(Nbound[i,g[3]],c1[i])*(1-
                        equals(Nbound[i,g[1]],Nbound[i,g[3]]))*(1-
                        equals(Nbound[i,g[2]],Nbound[i,g[3]]))
        }

        # Prior for the regression parameters
        theta2 ~ dnorm(0, 1.0E-3)I(0.0, )
        theta1 ~ dnorm(0, 1.0E-3)I(0.0, )
        b1[3] <- b1[2] + theta2
        b1[2] <- b1[1] + theta1
        b1[1] ~ dnorm(0, 1.0E-3)I(0.0, )
        beta0 ~ dnorm(0, 1.0E-3)
        tau ~ dgamma(0.01, 0.01)
        sigma2 <- 1 / sqrt(tau)
}

# Example of initial values
list(beta0=10, b1=c(10, NA, NA), theta1=5, theta2=5, tau=0.1, aux1=12, aux2=17, aux3=12)

# Data
list(N=43, J=3, y = c(2.71, …, 2.22 ), x=c(74, …, 81), Nbound=structure(.Data=c(0,908,…,
3011,0),.Dim=c(43,43)))
```

Appendix H

Bivectorial intrinsic autoregression as a limit of a bivectorial conditional autoregression

The probability density function of a bivectorial intrinsic autoregression is given in equation 8.9 and corresponds to the limit of the probability density function in equation 8.8 when $\rho_S \to 1$. This can be easily shown. First, note that the term $\boldsymbol{b}^t(\boldsymbol{\Psi}_S \otimes \boldsymbol{\Psi}_c)\boldsymbol{b}$ in the exponent of equation 8.8 can be written as

$$\begin{aligned} \boldsymbol{b}^t(\boldsymbol{\Psi}_S \otimes \boldsymbol{\Psi}_c)\boldsymbol{b} &= \sum_{i=1}^{n}\sum_{s=1}^{n} \Psi_{S,is}\boldsymbol{b}_{i\bullet}^t\boldsymbol{\Psi}_c\boldsymbol{b}_{s\bullet}\,, \\ &= \sum_{i=1}^{n} \Psi_{S,ii}\boldsymbol{b}_{i\bullet}^t\boldsymbol{\Psi}_c\boldsymbol{b}_{i\bullet} + \sum_{i=1}^{n}\sum_{s=1,s\neq i}^{n} \Psi_{S,is}\boldsymbol{b}_{i\bullet}^t\boldsymbol{\Psi}_c\boldsymbol{b}_{s\bullet}\,, \quad \text{(H.1)} \end{aligned}$$

where $\Psi_{S,is}$ is the (i,s) entry of the matrix $\boldsymbol{\Psi}_S$.

When $\rho_S = 1$, the rows of the matrix $\boldsymbol{\Psi}_S$ sum to zero, i.e. $\sum_{s=1}^{n}\Psi_{S,is} = 0$, and thus $\Psi_{S,ii} = -\sum_{s=1}^{i-1}\Psi_{S,is} - \sum_{s=i+1}^{n}\Psi_{S,is}$. The first term in equation H.1

can then be written as

$$-\sum_{i=1}^{n}\sum_{s=1}^{i-1}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}-\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}\,.$$

Due to the symmetry of $\boldsymbol{\Psi}_S$, a change in the order of summation in the first term of the latter equation easily shows that

$$\sum_{i=1}^{n}\sum_{s=1}^{i-1}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}\;=\;\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{s\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\,.$$

To sum up, the first term in equation H.1 is equivalent to

$$-\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}-\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{s\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\,. \tag{H.2}$$

Turning now to the second term of equation H.1,

$$\sum_{i=1}^{n}\sum_{s=1,s\neq i}^{n}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\;=\;\sum_{i=1}^{n}\sum_{s=1}^{i-1}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}+\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\,.$$

The first term on the right-hand side of the equation above can be written as

$$\sum_{i=1}^{n}\sum_{s=1}^{i-1}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\;=\;\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{s\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}$$

because of the symmetry of $\boldsymbol{\Psi}_S$ and by changing the order of summation. As a result, the second term of equation H.1 is equivalent to

$$\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{s\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{i\bullet}+\sum_{i=1}^{n}\sum_{s=i+1}^{n}\Psi_{S,is}\boldsymbol{b}_{i\bullet}^{t}\boldsymbol{\Psi}_{c}\boldsymbol{b}_{s\bullet}\,. \tag{H.3}$$

Therefore, equation H.1 is equivalent to the sum of equations H.2 and H.3. Since $\Psi_{S,is}=-w_{is}$, for all $i\neq s$, this sum is equal to:

$$\sum_{i=1}^{n}\sum_{s=i+1}^{n}w_{is}(\boldsymbol{b}_{i\bullet}-\boldsymbol{b}_{s\bullet})^{T}\boldsymbol{\Psi}_{c}(\boldsymbol{b}_{i\bullet}-\boldsymbol{b}_{s\bullet})\,,$$

which corresponds to the expression for a bivectorial intrinsic autoregression given in equation 8.9.

Appendix I

Simulating a bivectorial conditional autoregression

Here we give the programs written for S-PLUS to simulate data from a bivectorial conditional autoregression.

```
Id <- diag(rep(1,275))    # Identity matrix

conctable <- tabulate(conc.ng$nhbr$row.id)  # Vector of the number of neighbours per area

conc.nhbr <- conc.ng$nhbr  # Neighbourhood structure of the areas
conc.nhbr$weights <- 1/conctable[conc.nhbr$row.id]

deltaN <- spatial.weights(conc.nhbr, parameters=c(0.999))    # Matrix (rho_S D W)

e.norm <- rmvnorm(275, sd=c(0.1,0.3), rho=0.9) # Bivariate normal

#e.norm <- rmvnorm(275, sd=c(0.1,0.3), rho=0.5)     # Other parameter choices
#e.norm <- rmvnorm(275, sd=c(0.1,0.3), rho=0.1)
#e.norm <- rmvnorm(275, sd=c(0.3,0.1), rho=0.9)
#e.norm <- rmvnorm(275, sd=c(0.3,0.1), rho=0.5)
#e.norm <- rmvnorm(275, sd=c(0.3,0.1), rho=0.1)

conc.carcov <-  solve(Id-deltaN) %*% diag(1/conctable)
# Covariance matrix of the univectorial CAR

conc.carcovL <- chol((conc.carcov+t(conc.carcov))/2)
# Cholesky decomposition of the covariance matrix
# Due to rounding the inverse of the covariance matrix may not be symmetric;
# the sum with the transpose ensures that the matrix is symmetric.

conc.carsim1 <- t(conc.carcovL) %*% e.norm[,1]
conc.carsim2 <- t(conc.carcovL) %*% e.norm[,2]
indice1 <- rep(1:275,rep(2,275))
indice2 <- rep(c(0,275),275)
indice <- indice1 + indice2
conc.carsim <- c(conc.carsim1,conc.carsim2)[indice]
# Vector of a simulated bivectorial conditional autoregression
```

Appendix J

Winbugs program - Spatially varying regression

Here we give one of the programs written for Winbugs 1.4 to fit the spatially varying regression discussed in chapter 8, i.e. a mixed effects model in which the prior distribution of the random effects is a bivariate intrinsic autoregression.

```
model
{
    for( i in 1 : N )
    {
        O[i]  ~ dpois(mu[i])
        log(mu[i]) <- log(E[i])+mu.beta[1]+beta1[i]+(mu.beta[2]+beta2[i])*(0.01*x[i])
        RR[i] <- exp(mu.beta[1]+beta1[i] + (mu.beta[2]+beta2[i])*(0.01*x[i]))
         beta1[i] <- rho*phi*beta2[i]+phi*sqrt(1-pow(rho,2))*aux[i]
    }

   # IAR prior distribution for the spatial slope - improper distribution
   beta2[1:N] ~ car.normal(adj[],weights[], num[], tau2)
   for(k in 1:sumNumNeigh) {weights[k] <- 1}

   aux[1:N] ~ car.normal(adj[],weights[], num[], tau2)

   # Other priors:
   mu.beta[1:2] ~ dmnorm(mean[],prec[ , ])
   tau2 ~ dgamma(0.5, 0.0005)
   sigma2 <- 1/tau2
   phi ~ dgamma(1, 1)
   rho ~ dunif(-1,1)
   sigma1 <- pow(phi,2)*sigma2
}

# Example of initial values
list(tau2=1, phi=1, rho=0, mu.beta = c(0,0), aux= c(0,0,0,0,0,0,0,0,0,NA,0,… ,0),
beta2=c(0,0,0,0,0,0,0,0,0,0,0,NA,0,0,…,0))

# Data
list(N=275, mean = c(0,0), prec = structure(.Data = c(1.0E-6, 0, 0, 1.0E-6), .Dim =
c(2, 2)), O   = c(   123,…,53), E = c( 162.69, …, 70.42), x=c(71.12,…,37.3),
num=c(9,6,…,5), sumNumNeigh = 1464, adj=c(2, 3, …274))
```

Appendix K

The world dataset

Here we give a listing of the full dataset used for the world analysis of stomach cancer in chapter 9. The country codes correspond to the coding used in the figures throughout this thesis.

Code	Country	World age standardized incidence rate	Prevalence of the *H. pylori* infection (%)
1	Albania	15.0	74.0
2	Austria	12.4	47.0
3	Belgium	8.6	36.0
4	Croatia	19.5	60.4
5	Czech Republic	12.3	58.8
6	Denmark	6.3	44.4
7	Estonia	21.2	89.3
8	Finland	9.0	36.0
9	France	7.8	25.4
10	Germany	12.7	39.2
11	Greece	9.0	64.3
12	Hungary	18.6	63.3
13	Iceland	9.9	39.9
14	Ireland	9.6	43.0
15	Italy	15.1	45.1
16	Lithuania	20.6	78.5
17	Netherlands	9.1	36.6
18	Norway	8.6	36.0
19	Poland	15.9	69.9
20	Portugal	22.6	82.8
21	Russian Federation	30.5	88.0
22	Slovenia	16.3	54.8
23	Spain	13.2	67.5
24	Sweden	6.8	29.6
25	Switzerland	8.8	19.0
26	Turkey	8.1	81.2
27	United Kingdom	9.0	29.3
28	Australia	7.3	38.0
29	China	26.8	60.4
30	India	4.3	87.5
31	Japan	48.9	83.7
32	Korea, Rep.	47.9	69.6
33	Malaysia	9.4	35.9
34	Myanmar	9.8	80.0
35	Nepal	5.0	60.4
36	New Zealand	8.1	38.6
37	Thailand	3.9	76.4
38	Vietnam	17.6	73.9

Code	Country	World age standardized incidence rate	Prevalence of the *H. pylori* infection (%)
39	Algeria	6.5	87.7
40	Egypt, Arab Rep.	3.4	88.0
41	Israel	11.1	65.6
42	Saudi Arabia	5.3	75.8
43	Canada	6.7	37.5
44	Jamaica	23.9	72.2
45	Mexico	14.4	81.3
46	United States	5.6	32.7
47	Argentina	9.1	49.1
48	Barbados	13.8	72.0
49	Brazil	15.7	84.7
50	Chile	27.0	65.5
51	Colombia	26.9	96.3
52	Venezuela	17.3	92.0
53	Côte d'Ivoire	3.6	77.2
54	Nigeria	4.2	84.0
55	South Africa	6.3	86.4
56	Sudan	3.6	73.0
57	Zambia	9.2	81.0
58	Bangladesh	1.3	91.7

Source: Lunet and Barros (2003)

Appendix L

Age-sex specific death rates in Portugal

Males AR (per 10^5 persons per year)								
Age	0-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
AR	0.023	0.071	0.27	1.2	3.5	5.9	11	21
Age	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
AR	34	54	84	125	197	245	308	376
Females AR (per 10^5 persons per year)								
Age	0-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
AR	0	0.091	0.34	1.2	2.1	4.4	6.8	11
Age	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
AR	14	22	35	54	95	128	183	217

Table L.1: Age-sex specific risks (AR) for stomach cancer mortality in continental Portugal for the period 1985-1998 (population sizes taken from the 91 census).

Bibliography

Achtman, M., Azuma, T., Berg, D. E., Ito, Y., Morelli, G., and Pan, Z.-J. (1999). Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Molecular Microbiology*, 32(3):459–470.

Ahn, Y.-O. (1997). Diet and stomach cancer in Korea. *International Journal of Cancer*, Suppl 10:7–9.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alderson, M. (1976). *An Introduction to Epidemiology*. The MacMillan Press, Ltd.

Ando, T., Wassenaar, T., Peek, R., Aras, R., Tschumi, A., Van Doorn, L.-J., Kusugami, K., and Blaser, M. (2002). A *Helicobacter pylori* restriction endonuclease-replacing gene, *hrgA*, is associated with gastric cancer in Asian strains. *Cancer Research*, 62(8):2385–2389.

Armitage, P., Berry, G., and Matthews, J. (2002). *Statistical Methods in Medical Research*. Blackwell Publishing.

Assuncão, R. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14(5):453–473.

Assuncão, R., Potter, J., and Cavenaghi, S. (2002). A Bayesian space varying parameter model applied to estimating fertility schedules. *Statistics in Medicine*, 21(14):2057–2075.

Atherton, J. (1997). The clinical relevance of strain types of *Helicobacter pylori*. *Gut*, 40(6):701–703.

Axon, A. (1999). Are all helicobacters equal? Mechanisms of gastroduodenal pathology and their clinical implications. *Gut*, 45(Suppl 1):I1–I4.

Azevedo, L., Salgueiro, L., Claro, R., Teixeira-Pinto, A., and Costa-Pereira, A. (1999). Diet and gastric cancer in Portugal - a multivariate model. *European Journal of Cancer Prevention*, 8:41–48.

Barankin, E. and Dorfman, R. (1958). *On quadratic programming*. University of California Press.

Basso, D., Navaglia, F., Brigato, L. Piva, M., Toma, A., Greco, E., Di Mario, F., Galeotti, F., Roveroni, G., Corsine, A., and Plebani, M. (1998). Analysis of *Helicobacter pylori vacA* and *cagA* genotypes and serum antibody profile in benign and malignant gastroduodenal diseases. *Gut*, 43(2):182–186.

Benichou, J. (1998). Attributable risk. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*. Springer, 2nd edition.

Berrino, F., Capocaccia, R., Estve, J., Gatta, G., Hakulinen, T., Micheli, A., Sant, M., and Verdecchia, A., editors (1999). *Survival of Cancer Patients in Europe: the EUROCARE-2 Study*. Number 151 in IARC Scientific Publication. IARC Press, Lyon.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B, Methodological*, 48:259–279.

Besag, J. and Koperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):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):1–59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6*, pages 131–156.

Boeing, H., Frentzelbeyme, R., Berger, M., Berndt, V., Gores, W., Korner, M., Lohmeier, R., Menarcher, A., Mannl, H., Meinhardt, M., Muller, R., Ostermeier, H., Paul, F., Shwemmle, K., Wagner, K., and Wahrendorf, J. (1991). Case-control study on stomach-cancer in Germany. *International Journal of Cancer*, 47(6):858–864.

Breslow, N. and Day, N. (1987). *Statistical Methods in Cancer Research - Vol. 2 - The Analysis of Cohort Studies*. IARC.

Breslow, N. E. (1996). Generalized linear models: Checking assumptions and strengthening conclusions (italian). *Statistica Applicata*, 8:23–41.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Brooks, S. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression - modelling spatial non-stationarity. *The Statistician*, 47(3):431–443.

Buiatti, E., Palli, D., Bianchi, S., Decarli, A., Amadori, D., Avellini, C., Cipriani, F., Cocco, P., Giacosa, A., and Lorenzini, L. (1991). A case-control study of gastric-cancer and diet in Italy.iii. Risk patterns by histologic type. *International Journal of Cancer*, 48(3):369–374.

Chang, Y.-W., Han, Y.-S., Lee, D., Kim, H.-J., Lim, H.-S., Moon, J.-S., Dong, S.-H., Kim, B.-H., Lee, J.-I., and Chang, R. (2002). Role of *Helicobacter Pylory* infection among offspring or siblings of gastric cancer patients. *International Journal of Cancer*, 101(5):469–474.

Chao, A., Thun, M., Henley, S., Jacobs, E., McCullough, M., and Calle, E. (2002). Cigarette smoking, use of other tobacco products and stomach cancer mortality in US adults: the cancer prevention study II. *International Journal of Cancer*, 101(4):380–389.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671–681.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion.

Congdon, P. (2001). *Bayesian Statistical Modelling*. John Wiley & Sons, Ltd.

Correa, P. (2003). Bacterial infections as a cause of cancer. *Journal of the National Cancer Institute*, 95(7):3.

Covacci, A., Telford, J., Del Giudice, G., Parsonnet, J., and Rappuoli, R. (1999). *Helicobacter pylori* virulence and genetic geography. *Science*, 284(5418):1328–1333.

Cressie, N. (1993). *Statistics for Spatial Data - Revised Edition*. Wiley.

Davanzo, B., Lavecchia, C., and Franceschi, S. (1994). Alcohol consumption and the risk of gastric-cancer. *Nutrition and Cancer - An International Journal*, 22(1):57–64.

Destefani, E., Correa, P., Fierro, L., Carzoglio, J., Deneopellegrini, H., and Zavala, D. (1990). Alcohol drinking and tobacco smoking in gastric-cancer - a case-control study. *Revue Epidemiologique de Santé Publique*, 38(4):297–307.

Dhillon, P., Farrow, D., Vaughan, T., Chow, W.-H., Risch, H., Gammon, M., Mayne, S., Stanford, J., Schoenberg, J., Ahsan, H., Dubrow, R., West, A., Rotterdam, H., Blot, W., and Fraumeni, J. (2001). Family history of cancer and risk of esophageal and gastric cancers in the United States. *International Journal of Cancer*, 93(1):148–152.

Diebolt, J. and Robert, C. (1994). Estimation of finite mixture distributions

through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56(2):363–375.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271.

EUROGAST Study Group (1993). An international association between *Helicobacter pylori* infection and gastric cancer. *The Lancet*, 341(8857):1359–1362.

Everett, S. and Axon, A. (1998). Early gastric cancer: disease or pseudo-disease? *The Lancet*, 351(9112):1350–1352.

Falush, D., Wirth, T., Linz, B., Pritchard, J., Steffens, M., Kidd, M., Blaser, M., Graham, D., Vacher, S., Perez-Perez, G., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M., and Suerbaum, S. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science*, 299(5612):1582–1585.

Forman, D., Newell, D., Fullerton, F., Yarnell, J., Stacey, A., Wald, N., and Sitas, F. (1991). Association between infection with *Helicobacter pylori* and risk of gastric cancer: evidence from a prospective investigation. *British Medical Journal*, 302(6788):1302–1305.

Gamerman, D., Moreira, A., and Rue, H. (2003). Space varying models: specifications and simulation. *Computational Statistics and Data Analysis*, 42(3):513–533.

Gelfand, A., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling

with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelfand, A. and Sahu, S. (1999). Identifiability, improper priors, and gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, 94(445):247–253.

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

González, C., Pera, G., Agudo, A., Palli, D., Krogh, V., Vineis, P., Tumino, R., Panico, S., Berglund, G., Simán, H., Nyrén, O., Agren, A., Martinez, C., Dorronsoro, M., Barricarte, A., Tormo, M., Quiros, J., Allen, N., Bingham, S., Day, N., Miller, A., Nagel, G., Boeing, H., Overvad, K., Tjonneland, A., Bueno-De-Mesquita, H., Boshuizen, H., Peeters, P., Numans, M., Clavel-Chapelon, F., Helen, I., Agapitos, E., Lund, E., Fahey, M., Saracci, R., Kaaks, R., and Riboli, E. (2003). Smoking and the risk of gastric cancer in the European prospective investigation into cancer and nutrition (EPIC). *International Journal of Cancer*, 107(4):629–634.

González, C., Sala, N., and Capellá, G. (2002). Genetic susceptibility and gastric cancer risk. *International Journal of Cancer*, 100(3):249–260.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070.

Gregorio, D., Flannery, J., and Hansen, H. (1992). Stomach cancer patterns

in European immigrants to Connecticut, United-States. *Cancer Causes & Control*, 3(3):215–221.

Hirohata, T. and Kono, S. (1997). Diet/nutrition and stomach cancer in Japan. *International Journal of Cancer*, 71(Suppl 10):34–36.

Hoey, J., Montvernay, C., and Lambert, R. (1981). Wine and tobacco: risk factors for gastric cancer in France. *American Journal of Epidemiology*, 113(6):668–674.

Hohenberger, P. and Gretschel, S. (2003). Gastric cancer. *The Lancet*, 362(9380):305–315.

IARC Working Group (1994). Schistosomes, liver flukes and *Helicobacter pylori*. In *IARC monographs on the evaluation of carcinogenic risk to humans.*, volume 61. IARC.

INE (1993). *Estudo sobre o poder de compra concelhio*. Instituto Nacional de Estatistica.

Inoue, M., Tajima, K., and Yamamura, Y. (1999). Influence of habitual smoking on gastric cancer by histologic subtype. *International Journal of Cancer*, 81(1):39–43.

Inoue, M., Tajima, K., Yamamura, Y., Hamajima, N., Hirose, K., Kodera, Y., Kito, T., and Tominaga, S. (1998). Family history and subsite of gastric cancer: data from a case-referent study in Japan. *International Journal of Cancer*, 76(6):801–805.

Jedrychowski, W., Boeing, H., and Wahrendorf, J. (1993). Vodka consumption,

tobacco smoking and risk of gastric cancer in Poland. *International Journal of Epidemiology*, 22(4):606–613.

Jenks, G. and Caspall, F. (1971). Error on choroplethic maps: Definition, measurement, reduction. *Annals of the Association of American Geographers*, 61(2):217–244.

Ji, B.-T., Chow, W.-H., Yang, G., McLaughlin, J., Zheng, W., Shu, X.-O., Jin, F., Gao, R.-N., Gao, Y.-T., and Fraumeni, J. (1998). Dietary habits and stomach cancer in Shanghai, China. *International Journal of Cancer*, 76(5):659–664.

Kaaks, R., Tuyns, A., Haelterman, M., and Riboli, E. (1998). Nutrient intake patterns and gastric cancer risk: a case-control study in Belgium. *International Journal of Cancer*, 78(4):415–420.

Kaluzny, S., Vega, S., Cardoso, T., and Shelly, A. (1998). *S+SpatialStats, User's Manual for Windows and UNIX*. Springer-Verlag, New York.

Kidd, M., Lastovica, A., J.C., A., and Louw, J. (1999). Heterogeneity in the *Helicobacter pylori vacA* and *cagA* genes: association with gastroduodenal disease in South Africa? *Gut*, 45(4):499–502.

Knorr-Held, L. and Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21.

Kobayashi, M., Tsubono, Y., Sasazuki, S., Sasaki, S., and Tsugane, S. (2002). Vegetables, fruit and risk of gastric cancer in Japan: a 10-year follow-up of the JPHC study Cohort I. *International Journal of Cancer*, 102(1):39–44.

Kondo, T., Toyoshima, H., Tsuzuki, Y., Hori, Y., Yatsuya, H., Tamakoshi, K., Tamakoshi, A., Ohno, Y., Kikuchi, S., Sakata, K., Hoshiyama, Y., Hayakawa,

N., Tokui, N., Mizoue, T., and Yoshimura, T. (2003). Aggregation of stomach cancer history in parents and offspring in comparison with other sites. *International Journal of Epidemiology*, 32(4):579–583.

La Vecchia, C., Muñoz, S., Braga, C., Fernandez, E., and Decarli, A. (1997). Diet diversity and gastric cancer. *International Journal of Cancer*, 72(2):255–257.

Lagergren, J., Bergström, R., Lindgren, A., and Nyrén, O. (2000). The role of tobacco, snuff and alcohol use in the aetiology of cancer of the oesophagus and gastric cardia. *International Journal of Cancer*, 85(3):340–346.

Longford, N. (1993). *Random Coefficient Models*. Clarendon Press.

López-Carrillo, L., López-Cervantes, M., Ward, M., Bravo-Alvarado, J., and Ramírez-Espitia, A. (1999). Nutrient intake and gastric cancer in Mexico. *International Journal of Cancer*, 83(5):601–605.

Lunet, N. and Barros, H. (2003). *Helicobacter Pylori* infection and gastric cancer: facing the enigmas. *International Journal of Cancer*, 106(6):953–960.

MapInfo Corporation (2001). *MapInfo Professional User's Guide - version 6.5*. Available at the URL adress: www.mapinfo.com/docs.

Marin, J., Mengersen, K., and Robert, C. (2005). Bayesian modelling and inference on mixtures of distributions. In Dey, D. and Rao, C., editors, *Handbook of Statistics*, volume 25. Elsevier.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapmand and Hall, 2nd edition.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley-Interscience.

Miehke, S., Kirsh, C., Agha-Amiri, K., Günter, T., Lehn, N., Malfertheiner, P., Stolte, M., Ehninger, G., and Bayerdörfer, E. (2000). The *Helicobacter pylori vacA, s1, m1* genotype and *cagA* is associated with gastric carcinoma in Germany. *International Journal of Cancer*, 87(3):322–327.

Milne, P. (1983). A note on scale invariance. *The British Journal for the Philosophy of Science*, 43(1):49–55.

Miwa, H., Go, M., and Sato, N. (2002). *H. pylori* and gastric cancer: the Asian enigma. *The American Journal of Gastroenterology*, 97(5):1106–1112.

Mollié, A. (1996). Bayesian mapping of disease. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 359–379. Chapman & Hall, London.

Morrel, C., Pearson, J., and Brant, L. (1997). Linear transformations of linear mixed-effects models. *The American Statistician*, 51(4):338–343.

Muñoz, N., Plummer, M., Vivas, J., Moreno, V., De Sanjosé, S., Lopez, G., and Oliver, W. (2001). A case-control study of gastric cancer in Venezuela. *International Journal of Cancer*, 93(3):417–423.

Mueller, N. (1995). Overview: viral agents and cancer. *Environmental Health Perspectives*, 103(Suppl 8):259–262.

Parkin, D., Bray, F., Ferlay, J., and Pisani, P. (2001). Estimating the world cancer burden: Globocan 2000. *International Journal of Cancer*, 94(2):153–156.

Parsonnet, J. (1995). Bacterial infection as a cause of cancer. *Environmental Health Perspectives*, 103(Suppl 8):263–8.

Parsonnet, J., Friedman, G., Orentreich, N., and Vogelman, H. (1997). Risk of gastric cancer in people with *CagA* positive or *CagA* negative *Helicobacter pylori* infection. *Gut*, 40(3):297–301.

Peixoto, J. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, 44(1):26–30.

Pérez-Pérez, G., Bhat, N., Gaensbauer, J., Fraser, A., Taylor, D., Kuipers, E., Zhang, L., You, W., and Blaser, M. (1997). Country-specific constancy by age in $cagA^+$ proportion of *Helicobacter pylori* infections. *International Journal of Cancer*, 72(3):453–456.

Plummer, M. and Clayton, D. (1996). Estimation of population exposure in ecological studies. *Journal of the Royal Statistical Society, Series B*, 58(1):113–126.

Pocock, S., Cook, D., and Lambert, R.Beresford, S. (1981). Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics*, 30(3):286–295.

Poirier, D. (1998). Revising beliefs in nonidentified models. *Econometric theory*, 14(4):483–509.

Prokhorkas, R. (1998). Mortality, international comparisons. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley.

Rao, D., Ganesh, B., and Dinshaw, K. (2002). A case-control study of stomach cancer in Mumbai, India. *International Journal of Cancer*, 99(5):727–731.

Richardson, S. (1992). Statistical methods for geographical correlation studies. In Elliott, P., Cuzick, J., English, D., and Stern, R., editors, *Geographical*

and environmental epidemiology : methods for small-area studies, chapter 17, pages 181–204. Oxford University Press.

Richardson, S. and Best, N. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, 14(2):129–147.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59(4):731–758.

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag, New York.

Robert, G. (1996). Mixtures of distributions: inference and estimation. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 441–464. Chapman & Hall, London.

Roberts, G. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov chain Monte Carlo in practice*, pages 45–57. Chapman & Hall, London.

Robinson, W. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15(3):351–357.

Rodrigues, V., Veiga, F., and Cardoso, S. (1991-1992). Mortalidade por cancro e ambiente: um exercício epidemiológico. *Arquivos do Instituto Nacional de Saúde*, 16-17:89–168.

Rose, G. and Barker, D. (1986). *Epidemiology for the Uninitiated*. BMJ Books, 2nd edition.

Sasazuki, S. and Sasaki, S. (2002). Cigarette smoking, alcohol consumption and subsequent gastric cancer risk by subsite and histologic type. *International Journal of Cancer*, 101(6):560–566.

Schlattman, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12:1943–1950.

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

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). Winbugs user manual version 1.4 - examples vol. II. Available at the URL adress: http://www.mrc-bsu.cam.ac.uk/bugs/.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–640.

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. PhD thesis, University of Oxford.

Trédaniel, J., Boffeta, P., Buiatti, E., Saracci, R., and Hirsch, A. (1997). Tobacco smoking and gastric cancer: review and meta-analysis. *International Journal of Cancer*, 72(4):565–573.

Van Der Ende, A., Pan, Z., Bart, A., Van Der Hulst, R., Feller, M., Xiao, S.-D., Tytgat, G., and Dankert, J. (1998). *cagA*-Positive *Helicobacter pylori*

populations in China and The Netherlands are distinct. *Infection and Immunity*, 66(5):1822–1826.

Van Doorn, L.-J., Figueiredo, C., Sanna, R., Blaser, M., and Quint, W. (1999). Distinct variants of *Helicobacter pylori cagA* are associated with *vacA* subtypes. *Journal of Clinical Microbiology*, 37(7):2306–2311.

Wakefield, J. and Salway, R. (2001). A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A*, 164(1):119–137.

Wilkens, L. and Lee, J. (1998). Nutritional epidemiology. In Armitage, P. and Colton, T., editors, *Encyclopedia of Biostatistics*. Wiley.

Yamaguchi, N. and Kakizoe, T. (2001). Synergistic interaction between *Helicobacter pylori* gastritis and diet in gastric cancer. *The Lancet Oncology*, 2(2):88–94.

Yatsuya, H., Toyoshima, H., Mizoue, T., Kondo, T., Tamakoshi, K., Hori, Y., Tokui, N., Hoshiyama, Y., Kikuchi, S., Sakata, K., Hayakawa, N., Tamakoshi, A., Ohno, Y., and Yoshimura, T. (2002). Family history and the risk of stomach cancer death in Japan: differences by age and gender. *International Journal of Cancer*, 97(5):688–694.

You, W.-C., Ma, J.-l., Liu, W., Gail, M., Chang, Y.-s., Zhang, L., Hu, Y.-R., Fraumeni, J., and Xu, G.-W. (2000). Blood type and family cancer history in relation to precancerous gastric lesions. *International Journal of Epidemiology*, 29(3):405–407.