Spatial Multilevel Modelling of Cancer Mortality in Europe

Carolyn A Davies

A Dissertation submitted to the
University of Glasgow
for the degree of
Doctor of Philosophy

Department of Statistics
January 2005

Abstract

The high number of cancer deaths in the western world is of growing concern. There were an estimated 2.6 million new cases of cancer in Europe in 1995, representing over one-quarter of the world burden of cancer despite Europe's inhabitants comprising only one-eighth of the world's population. The corresponding number of deaths from cancer was around 1.6 million. These figures demonstrate the very substantial burden of cancer in Europe, and since a lot of cancer mortality is, at least in theory, highly preventable, more research is needed into where cancer mortality poses the strongest threat and what preventative interventions are needed.

This thesis attempts to provide accurate estimates of cancer mortality rates throughout Europe by using appropriate modelling methods to produce smoothed disease maps of the true distribution of the disease. Since this study is on a large scale geographically, mortality data and risk factor exposure data are not available at an individual level. However, different forms of aggregated data are available allowing this ecological study to be carried out. Incorporating differing levels of exposure to various risk and protective factors reduces the variation in the disease risk estimates and allows the effects such factors have on European cancer mortality patterns to be quantified.

Poisson spatial multilevel modelling techniques are used to explore the distribution of all cancer mortality. The models are fitted using both empirical Bayes and fully Bayesian approaches and the effects of incorporating a spatial factor and of adding a higher geographical level are explored. The Poisson spatial multilevel model with correlated covariance structure was compared to other models which are commonly used for disease mapping purposes; the conditional autoregressive (CAR) model and the multiple-membership multiple-classification (MMMC) model. The most suitable method for modelling aggregated mortality

data proved to be the spatial multilevel model which allows for correlated random effects and is fitted using Monte Carlo Markov Chain (MCMC) methods.

This model is then used to explore the spatial mortality patterns of specific cancers, namely, lung, colorectal and oesophageal cancer, in the EU. Variability within and between countries was evident for each of the cancer groups examined. Much of this variation could be accounted for by risk factors such as smoking and diet, which had strong yet differing effects on each of the cancers. Accounting for spatial patterning of the disease was also shown to reduce variation in cancer mortality substantially across the EU. Cancer 'hotspots' were identified providing evidence that urgent preventative public health intervention on tobacco and diet modification is required in many European regions.

Acknowledgements

Firstly I would like to thank the Medical Research Council Social and Public Health Sciences unit who funded this PhD and provided a friendly and helpful working environment. I would also like to thank the Statistics Department for their support throughout my PhD and for their training and encouragement in the field of statistics.

Thank you to my supervisors, Dr Alastair Leyland for his constant help, guidance and support, and Professor Mike Titterington for his knowledgeable help and input.

I would like to thank Ruth, Helen, Anne and Margaret for being great office-mates and friends and especially appreciate Ruth's proof reading and helpful comments on my PhD. Thanks to Lesley for also proof reading my PhD and who has been a great friend and to another very close friend, Sharon. Both have given me much needed escapism and laughs over recent years.

I owe a special amount of gratitude to my parents, Sheila and Jeff, for their love, support and encouragement in everything I do. My dad's encouragement in my academic career has been a major influence in the path I've taken and my mum's overwhelming financial support and kindness has made the journey a much easier one.

My greatest thanks go to Andrew who has always put me first and his love, friendship and understanding has made doing my PhD a happy time.

I would like to say a final thank you to my sister, Sharon, who has always been my inspiration and, I know, would be very proud.

Declaration

Material from Section 4.2.2 has been published *(Leyland AH, Davies CA. Empirical Bayes methods for disease mapping. Statistical Methods in Medical Research, 2005. **14**(1): p. 17-34).*

Material covered in Chapter 2 and results from Chapter 5 have been published as a book chapter *(Davies CA, Leyland AH. Spatial patterns of cancer mortality in Europe. In: Kirch W, editor. Public Health in Europe: 10 years EUPHA. 1st ed. Berlin: Springer Verlag; 2003. p. 227-243).*

Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 Motivation

The World Health Organisation recently published an atlas of mortality in Europe (1). Population and mortality information was collected for the periods 1980/1981 and 1990/1991, co-ordinating with two recent censuses. The atlas was attempting to identify differences in trends in mortality from various causes at the sub-national level in Europe and also to indicate areas in which more study is needed to determine both reasons for these differences and the most appropriate action to reduce them. The data collected by WHO have been made available and provided the initial motivation and main source of information for this study.

Using subsets of this data set and developing existing spatial modelling methods, the aim of this thesis is to provide an accurate account of the spatial patterning of cancer mortality across Europe.

1.1.1 Cancer Mortality

The high number of cancer deaths in the western world is of growing concern. There were an estimated 2.6 million new cases of cancer in Europe in 1995, representing over one-quarter of the world burden of cancer despite Europe's inhabitants comprising only approximately one-eighth of the world's population. The corresponding number of deaths from cancer was around 1.6 million (2). It has been predicted that in 2020, 3.4 million new cases of cancer will occur in Europe (3). These figures demonstrate the very substantial burden of cancer in

Europe, and the scope for prevention and has motivated the research for this thesis.

Studies of continuity of mortality rates across national borders and an evaluation of the distribution between national and regional variations in health status have revealed large spatial mortality variation both within and between countries in Europe (1). This thesis aims to show such patterns but specifically in relation to cancer mortality. More accurate estimates of mortality rates are obtained by incorporating additional information that has been gathered with regards to potential risk and protective factors for cancer mortality. Variation in cancer mortality often reflects differences in demographic structure, socio-economic conditions or lifestyle factors. It is of particular interest to quantify the effect of such factors on cancer mortality patterns and also to examine any relationships that exist after taking these factors into account.

1.1.2 Disease Mapping

A common method used to investigate the spatial variability of a disease such as cancer is to map mortality rates. Producing accurate disease maps of cancer mortality is of key importance in the field of public health as it draws experts in the area closer to understanding the true geographical distribution of a potentially fatal disease which is becoming ever more common.

Disease mapping has been a growing area over recent years (4) with the main aim being to produce a map 'clean' of random noise and any natural variation in the human population, allowing the identification of areas with high or low rates. This research aims to develop existing disease mapping methods to produce accurate maps of cancer mortality allowing the assessment of the true underlying distribution of the disease.

1.1.3 Spatial Multilevel Modelling

To produce accurate disease maps of cancer mortality, accurate estimates of disease rates have to be calculated. The method proposed in this thesis is to provide estimates by Poisson spatial multilevel modelling techniques.

Geographically distributed data such as counts of cancer deaths within regions within countries take on a natural hierarchical structure and exemplifies the type of data appropriate for a multilevel analysis. The observations within these clustered units are likely to be more similar than observations from different clusters, due to shared social and geographical environments. It is therefore important to take account of this underlying structure and the correlation that exists between observations from the same cluster. Along with taking into account the multilevel structure of the data this method gives a convenient and efficient way to overcome problems which arise from traditional methods of modelling and mapping disease rates. It allows the extra-Poisson variation, that often exists in observed counts of cancer deaths, to be modelled. Also, areas geographically close to one another share similar disease rates and common factors which influence the incidence and outcome of a disease, and this spatial patterning of disease can be taken into account through multilevel modelling. Using this type of modelling to estimate risks of disease mortality allows potential ecological covariates to be incorporated into the models, hopefully giving a more accurate picture of disease patterns across the map. Modelling these ecological covariates also gives the opportunity to quantify the associations they have with cancer mortality across Europe.

The research that exists in Poisson spatial multilevel modelling concentrates mainly on fitting models with two levels. This research extends existing models (5, 6) to incorporate further levels such as country at a higher level.

Earlier disease mapping applications tended to focus on fitting models using empirical Bayes approaches. However, more recently, with the growing availability of Markov Chain Monte Carlo methods in packages such as WinBUGs (7), there has been much development in fully Bayes approaches. Both

methods are explored in this thesis in terms of fitting Poisson spatial multilevel models with a correlated variance structure. These methods are compared, along with two other MCMC methods that have been proposed as reliable approaches to modelling such multilevel data; the multiple membership model and the conditional autoregressive model.

1.2 Objectives

The main objectives of this thesis are to explore the patterns of cancer mortality in Europe and to develop Poisson spatial multilevel models to provide accurate estimates of mortality risks. Since this study is on such a large scale geographically, mortality data and risk factor exposure data are not available at an individual level. However, different forms of aggregated data are available allowing this 'ecological study' to be carried out.

Despite this type of analysis being fairly crude it can play an important role in epidemiology. For example, detecting areas or clusters with extreme disease rates could influence geographical assessment of health resource allocation; or significant relationships that are shown to exist between rates of a disease and the prevalence of a risk factor in given populations while analytically controlling for the prevalence of other confounding factors and for spatial autocorrelation may lead to the formulation of aetiological hypothesis which can then be examined further on an individual level.

1.3 Summary of Thesis

In the following Chapter, a summary is given of existing literature that has explored cancer mortality risk and protective factors. This review identifies the relationships that are known to exist between four specific cancers and lifestyle factors such a diet, smoking and socio-economic status.

In Chapter 3 a description and summary of the European population and cancer mortality data is given. The problems associated with data that has been collected on such a wide scale, such as data quality and missingness, are discussed. Data was also obtained that represents exposure to the various risk factors discussed in Chapter 2. Descriptions of these data sets, along with their summaries and problems are also given.

Chapter 4 reviews existing methods of disease mapping. The main focus is on modelling Poisson distributed counts of deaths and the development of the methods used to model relative risks of disease are explored. This involves looking at more recent spatial methods from both empirical Bayes and fully Bayesian approaches.

In Chapter 5, the European cancer mortality data are modelled using multilevel models. Variance components models are fitted and estimated using iterative generalised least squares procedures and quasi-likelihood methods. This model is extended to incorporate a spatial component and the resulting estimates are discussed. Disease maps of the relative risks from the models are examined.

Chapter 6 explores fitting the spatial multilevel model to the European cancer mortality data using Markov chain Monte Carlo. Aspects of using this fully Bayes method are discussed including fitting the model using WinBUGS, choice of priors, different sampling methods and convergence properties. Again, results are discussed and estimates explored through disease maps. The model is then extended to incorporate a further country level and differences in the results are discussed.

In Chapter 7 two other approaches to fitting spatial multilevel models are explored. Using MCMC, a multiple membership model and a correlated autoregression model are fitted to the cancer mortality data. Results from these methods are compared to the empirical Bayes' variance components and spatial multilevel models fitted in Chapter 5 and to the MCMC spatial multilevel model fitted in Chapter 6. Point estimates, confidence intervals and the diagnostic information criterion are used to compare the models. One model is then chosen

to explore specific cancer mortality rates further. In Chapter 8 lung, colorectal and oesophageal cancer mortality rates are examined separately using the MCMC three-level spatial multilevel model. The focus in this Chapter is mainly on identifying any 'hot spots' of cancer mortality from the disease maps and to quantify the causal relationships that exist between the specific cancer mortality rates and risk factors.

Finally, Chapter 9 discusses the limitations of the research carried out, both from a statistical modelling perspective and from a public health point of view. General conclusions about the thesis as a whole are drawn.

Chapter 2

2 Cancer Mortality Risk Factors

2.1 Introduction

Prevention of cancer is an increasingly important and integral part of public health. The first steps in preventing this disease are to understand its causes and attempt to quantify the proportion of cases due to each cause. Previous studies of geographic variation in cancer rates have provided important clues to the role of lifestyle factors and cancer risk.

2.2 Diet

The relationship between dietary components and cancer is not fully established; however, the overall impact of diet on cancer mortality appears to be significant. Evidence that diet is a determinant of cancer risk comes from several sources, including the following: correlation between national and regional food consumption data and the incidence of cancer in the population; studies on the changing rates of cancer as they migrate from a region or country of one dietary culture to another; case-control studies of dietary habits of individuals with and without cancer; prospective studies; intervention studies. While it is not yet possible to provide quantitative estimates of the overall risks, it has been estimated that 35 percent of cancer deaths may be related to dietary factors (8). The association between dietary components and cancers differs among different cancers or groups of cancers.

A recent study into the estimates of cancer incidence and mortality in Europe (2) showed that the most common cause of death in Europe in 1995 among cancers was lung cancer with 330,000 deaths, representing about one fifth of the total number of cancer deaths. Deaths from cancer of the colon and rectum (189,000) ranked second. Breast cancer was the most common cause of cancer in females, representing 17% of all female cancer deaths and lung cancer the most common in males (29%). Most countries in Europe have shown a rising trend in oesophageal cancer over the last thirty years, especially in males (9). In 1990, oesophageal cancer accounted for 3% of male cancer deaths (10). The rising trend and poor prognosis (five-year survival in Europe is less than 10 percent (11)), makes this disease of growing concern in cancer research. Further details of the specific relationships in these cancer groups (breast, colorectal, oesophageal and lung) are summarised in Table 2.1. Note that the blank areas in the table indicate that there is insufficient evidence of a relationship.

Table 2.1 Association between dietary components and selected cancers

Dietary Component	Site of Cancer			
	Breast	**Colorectal**	**Oesophageal**	**Lung**
Animal Fats	+	+		+
Vegetables	-	-	-	-
Fruit	-	-	-	-
Fish/Fish Oil	-	-		
Alcohol	+	+	+	+
Coffee		-		
Cheese	+			

\+ = positive association; increased intake with increased cancer mortality

\- = negative association; increased intake with decreased cancer mortality

2.2.1 Breast Cancer and Diet

Breast cancer is a common cause of death in woman throughout Europe. The EUROCARE II study recently estimated survival rates of incident cases between 1985 and 1989 in 17 countries in Europe (11) in which breast cancer was shown to have a five-year survival rate of 75%. One of the most often studied cancer associations is breast cancer risk and dietary fat. Earlier studies have supported the causal association between dietary fat and breast cancer; however, many recent studies have been uncertain. Research suggests that these conflicting results are due to concentration on dietary fat as opposed to a more specific focus on content of saturated fat in the diet along with fruit and vegetable consumption (12). Saturated fat, found mainly in animal fats, has been found to be positively related to risk of breast cancer. A recent ecologic study conducted using breast cancer mortality rates and dietary supplement data confirmed results from other studies showing that animal products are associated with risk for breast cancer and that fish intake and vegetable consumption are associated with risk reduction (13). There is strong and consistent evidence that increased consumption of fruit and vegetables is associated with reduced risks of many common forms of cancer including breast cancer (14, 15). A meta-analysis of studies on breast cancer risk and diet confirmed the association between intake of vegetables and, to a lesser extent, fruits and breast cancer risk (16). There has been much research focusing on dairy foods specifically. However, the only conclusive results regarding breast cancer risk appears to be a positive association with cheese intake (17). Many case-control and cohort studies (18, 19) and meta-analyses (20) have found a positive association between alcohol use and breast cancer. A study found risk of breast cancer was increased by 40-45% for woman ever drinking versus never drinking (21). The ecologic study, mentioned above, that examines breast cancer mortality rates (13) found alcohol to be an associated risk factor. It has been suggested that fibre intake decreases the risk of breast cancer. However, epidemiological studies have not consistently supported such a relationship. A meta-analysis of ten case-control studies (22) showed dietary fibre was inversely associated with the role of breast cancer, but, Willet at al (23) discussed that, in a

large number of prospective studies, inverse associations have generally not been shown to exist.

2.2.2 Colorectal Cancer and Diet

Colon cancer is the third most common form of cancer in Europe with particularly high incidence rates in Western Europe. Colorectal cancer is a leading cause of cancer mortality in the industrialised world. More literature is available examining relationships between diet and colon cancer incidence rather than mortality, but incidence should be a resonable measure of colorectal cancer mortality rates as survival of cancer of the colon is fairly poor because most cases are diagnosed at an advanced stage (24). The EUROCARE II project (11) showed colorectal cancer to have a five-year survival rate of 50%. Almost all the specific risk factors of colorectal cancer are of dietary origin. International comparisons indicate diets low in dietary fibre (low in vegetable and fruit consumption) and high in animal fat increase the risk of colon cancer (25). The large majority of studies in humans has found a protective effect of fibre from vegetables and possibly fruits (26, 27). Willett et al showed that animal fat from red meat intake was positively associated with colon cancer but dairy foods which contributed to total animal fat intake were not significantly related to the risk of colon cancer (28). High fat intake has been shown to be positively associated with the risk of colorectal cancer (26, 29). Fernandez et al suggested that fish consumption has a protective effect against colon cancer (30). Tavani et al confirmed coffee to be a risk factor showing an inverse association between coffee intake and risk of cancer of the colon (31). A prospective study of cancers of the colon and rectum confirmed a positive association between alcohol use and cancer of the colon (32).

2.2.3 Oesophageal Cancer and Diet

Cancer of the oesophagus is generally characterised by relatively low mortality rates in Europe with incidence in most Western European countries also being low (33). Several studies have demonstrated a positive association between

oesophageal cancer and several dietary factors including low intakes of vitamin A, C, riboflavin, nicotinic acid, calcium and zinc (34). In dietary terms the associations are with low intakes of lentils, green vegetables, fresh fruits and animal protein. Fernandez et al found a significant negative relationship with fish consumption and oesophageal cancer (30). Alcohol appears to play a major risk in many of the epidemiological studies examining risk factors for oesophageal cancer with intake of alcohol appearing to be an independent risk factor (34). A fairly recent study looking at oesophageal cancer mortality in Spain found similar results to other studies, supporting a role of alcohol in causation of oesophageal cancer (35).

2.2.4 Lung Cancer and Diet

Lung cancer is the most common cause of cancer mortality among men in Europe, and is becoming an increasingly important cause of cancer mortality among women (2). During the last 25 years it has become apparent that diet is the other major cause of cancer, but theories have moved steadily from a search for causal agents (eg, too much fat) to protective agents (eg, too little fruit and vegetables) (36, 37). Mortality studies have shown excess risk to be associated with consumption of saturated fats (red meat) (38, 39) and protective effects associated with the intake of vegetables (40) and fruit (41). Lung cancer incidence has also been shown to have an inverse association with high intakes of plant foods (42). A comprehensive search of the literature available on relationships between diet and lung cancer showed that a diet rich in fruit and vegetable reduces the incidence of lung cancer by approximately 25% (43). Evidence of dairy fats having a particular effect on lung cancer is inconclusive (17). Many of the suggested relationships between alcohol and lung cancer are actually confounded, or at least can be explained, by smoking. However, overall, the existing evidence suggests a small increase in the risk of lung cancer from alcohol drinking that does not appear to be fully explained by tobacco smoke (44).

2.3 Lifestyle Factors

Other lifestyle factors such as smoking and socio-economic status are commonly associated with the risk of cancers. It has been shown that people with high socio-economic status are at a greater risk from some cancers; however, evidence of an association with cancer mortality is sometimes inconclusive. Smoking is well known to be a major risk factor for many common cancers. The relationship between such lifestyle factors and the four cancers previously discussed are shown in Table 2.2.

Table 2.2 Association between other lifestyle factors and cancer mortality

Lifestyle Factor	Site of Cancer			
	Breast	**Colon**	**Oesophageal**	**Lung**
Smoking			+	+
High socio-economic status	-	-		-

2.3.1 Breast Cancer and Lifestyle Factors

Earlier evidence that breast cancer risk is unlikely to be affected by cigarette smoking continues to be challenged by recent findings. Although no definite conclusions can be made, several recent studies have reported findings that strongly hinted at such a relationship. A recent survival study found smoking to be a significant risk factor for breast cancer mortality (45). However, a recent collaborative reanalysis of individual data from 53 epidemiological studies (46) showed smoking has little or no independent effect on the risk of developing breast cancer. Some recent studies have established passive exposure to environmental tobacco smoke (ETS) as a risk factor for breast cancer (47), yet another study looking at breast cancer mortality found no relationship with ETS

(48). It appears that results from studies into the effect cigarette smoking has on breast cancer mortality are equivocal and need further examination.

Socio-economic status is thought to influence the risk of breast cancer (49). A study of breast cancer in young women showed that, compared with controls, cases were significantly more educated (50), with breast cancer incidence being most frequently reported in those with greater than thirteen years education. Few studies have shown such a relationship with breast cancer mortality, and many conflicting results have appeared. However, a recent study showed a clear gradient in survival, with better survival for women with higher socio-economic status (51). There is also an increasing amount evidence to suggest a positive association between risk of breast cancer and obesity (52-54).

2.3.2 Colorectal Cancer and Lifestyle Factors

Colorectal cancer also appears to be clearly affected by socio-economic status. Tavani et al showed that the number of years of education was strongly associated with colorectal cancer incidence with a significant trend in risk when comparing those with the highest level of education to those with less than seven years' education (50). There is little evidence to suggest that this association is also apparent with colorectal cancer mortality but incidence, again, may be viewed as an indicator of mortality due to the poor survival rates. There also appears to be little evidence of an association between cigarette smoking and colorectal cancer mortality. However, an increased risk of colorectal cancer with early onset and a long history of cigarette smoking has been suggested (55), but no association was seen in a large case-control study (56) or in a study of male construction workers in Sweden (57).

2.3.3 Oesophageal Cancer and Lifestyle Factors

Many studies have consistently identified smoking as a risk factor for oesophageal cancer. A recent mortality study confirmed this by showing the role of cigarette consumption on causation of oesophageal cancer (35). Nyrén and Adami (58) give

a comprehensive review of the evidence suggesting a positive association between tobacco smoking and oesophageal cancer. There is little evidence of an association between socio-economic status and deaths from oesophageal cancer. In recent years obesity has emerged as a major risk factor for this disease with a positive association being shown to exist between BMI or relative weight and oesophageal cancer (59-63).

2.3.4 Lung Cancer and Lifestyle Factors

It is well known and accepted that tobacco smoking is the main risk factor for lung cancer. The risk among smokers relative to the risk among never-smokers lies between 8 to 15 in men and 2 to 10 in woman (44). These risks reflect the contribution of the different aspects of tobacco smoking, namely, average consumption, duration of smoking, time since quitting, age at start, type of tobacco product, and inhalation pattern. Evidence of a causal relationship between cigarette smoking and lung cancer mortality has been accumulating since the 1950s. In recent years more attention has been focused on the potential health effects of ETS. Numerous studies have now led to the expectation that exposure to ETS also entails some increase in lung cancer risk (64). There is little evidence of a relationship between socio-economic status and lung cancer mortality. The prognosis of lung cancer is generally poor due to the carcinomas being diagnosed at an advanced stage. In Europe, the 5-year survival from 1985-1989 is reported to be 10% for woman and 9% for men (65), and therefore there is less scope, in terms of time, for inequalities in survival by socio-economic group to occur. However, due to the poor prognosis, relationships with lung cancer incidence and socio-economic status are likely to reflect mortality rates as well. In most countries, lung cancer incidence in men and woman shows a social class gradient, with those from higher social classes having lower incidence than lower social classes (66).

Chapter 3

3 Data

3.1 Mortality and Population Data

One of the main goals of cancer control is to reduce mortality from the disease. Hence, examining mortality as an outcome is an important method to use when attempting to quantify the burden of cancer in a given population.

There are various measures of disease burden, namely incidence, cumulative incidence, prevalence, survival, life-years lost and disability-adjusted life years. However mortality is widely considered the most important indicator of the burden of cancer. Cancer mortality rates measure, at a given population level, the risk of dying from specific cancers or from all cancers.

3.1.1 Source of Data

The World Health Organisation recently published an atlas of mortality in Europe (1). Along with providing national averages for all the main causes of death within the WHO European Region, it aims to provide a geographical presentation of variations in gender and cause specific mortality across the WHO European Region. Information has been collected for the periods 1980/1981 and 1990/1991 to attempt to identify differences in trends in mortality at the sub-national level in Europe. The atlas also indicates areas in which more study is needed to determine both the reasons for these differences in mortality rates and the most appropriate action to reduce them. The information provided by WHO attempts to serve as a background against which to generate hypotheses and to formulate programmes

for epidemiological studies to explain the differences found within and between countries. The data collected by WHO have therefore been made available and will provide the main source of information for this project.

Data were only available from official national sources in European Member States. Collection of data was finalised in 1995 and up until this date there were 49 countries who were member states of the WHO European region. Data for the Federal Republic of Yugoslavia are included but this country is not a WHO Member State. A list of these countries and the years they became member states is given in Table 3.1. A map of the areas covered by the WHO European region is given in Figure 3.1

3.1.2 Population Data

Population data were requested from time points that co-ordinated with two recent censuses, providing information on the number of residents, according to midyear estimates, per region of residence for the years 1980, 1981, 1990 and 1991. The data have been broken down into 5-year age bands (up to the 80-84 age band and then all aged 85 and over) for both sexes. Region of residence was defined according to the EUROSTAT nomenclature of administrative units in the countries of the European Union (EU) or corresponding administrative equivalents in other countries. Where available, the units of analysis were level II NUTS (standard nomenclature of territorial units for statistics) or an equivalent level (sub-national) just below the country level (such as county in the UK, *département* in France, voivodship in Poland and *oblast* in the Russian Federation).

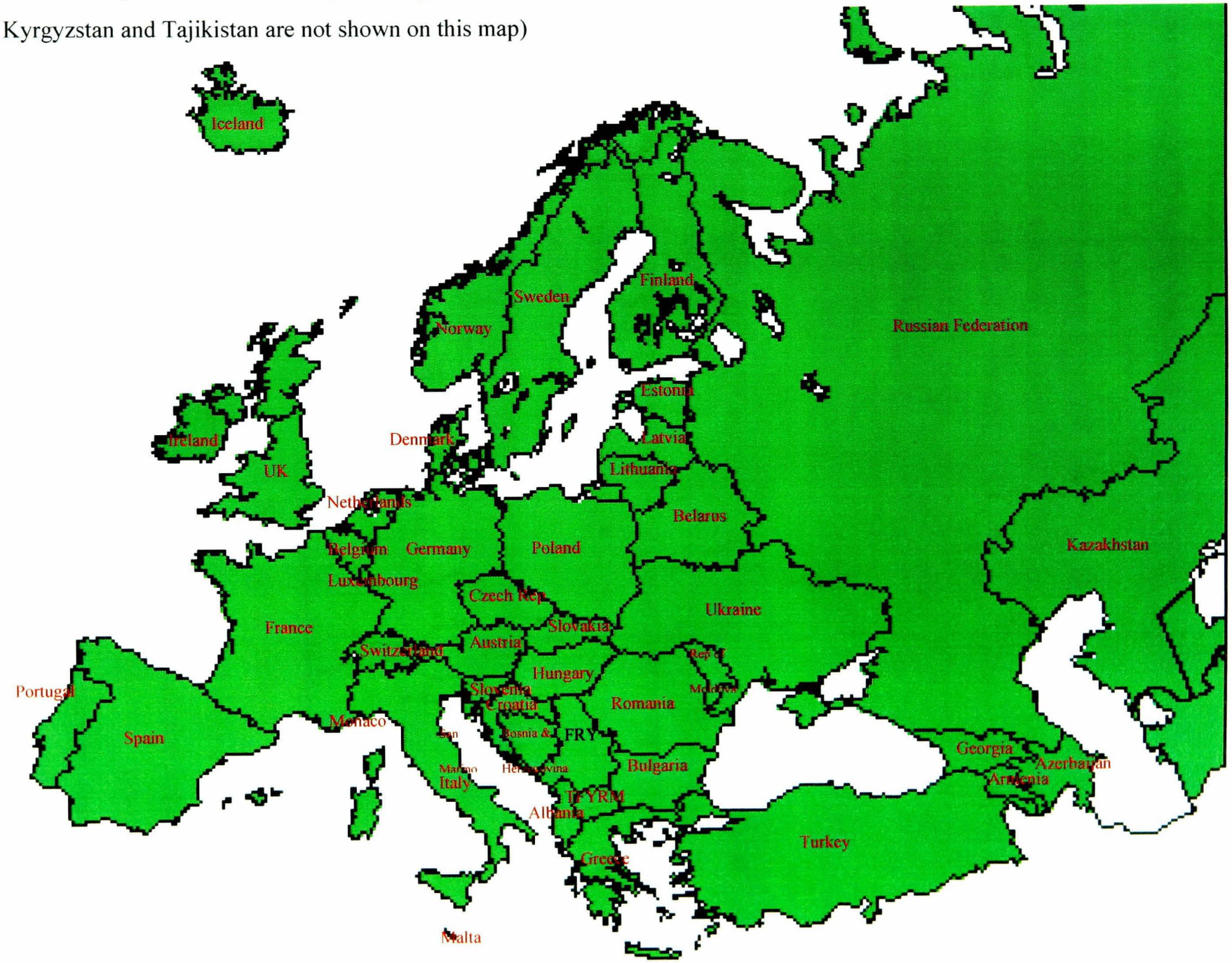
Population data are available for 36 countries in the WHO European Region for the period 1980/81 and for 44 countries in 1990/91. The total population for these 36 countries in 1980/81 (averaged over the 2 timepoints) is 677,057,315 with a mean of 18,807,148 and a range from 229,482 (Iceland) to 139,221,495 (Russian Federation). In 1990/91 the total population for the same 36 countries is 722,727,899 with a mean of 20,075,775 and a range from 256,377 (Iceland) to

Table 3.1 Countries in the WHO European Region

Country	Date of becoming party to WHO Constitution
Albania	26 May 1947
Armenia	4 May 1992
Austria	30 June 1947
Azerbaijan	2 October 1992
Belarus	7 April 1948
Belgium	25 June 1948
Bosnia & Herzegovina	10 September 1992
Bulgaria	9 June 1948
Croatia	11 June 1992
Czech Republic	22 January 1993
Denmark	19 April 1948
Estonia	31 March 1993
Finland	7 October 1947
France	16 June 1948
Georgia	26 May 1992
Germany	29 May 1951
Greece	12 March 1948
Hungary	17 June 1948
Iceland	17 June 1948
Ireland	20 October 1947
Israel	21 June 1949
Italy	11 April 1947
Kazakhstan	19 August 1992
Kyrgyzstan	29 April 1992
Latvia	4 December 1991
Lithuania	25 November 1991
Luxembourg	3 June 1949
Malta	1 Feb 1965
Monaco	8 July 1948
Netherlands	25 April 1947
Norway	18 August 1947
Poland	6 May 1948
Portugal	13 February 1948
Republic of Moldova	4 May 1992
Romania	8 June 1948
Russian Federation	24 March 1948
San Marino	12 May 1980
Slovakia	4 February 1993
Slovenia	7 May 1992
Spain	28 May 1951
Sweden	28 August 1947
Switzerland	26 March 1947
Tajikistan	4 May 1992
The Former Yugoslav Republic of Macedonia	22 April 1993
Turkey	2 January 1948
Turkmenistan	2 July 1992
Ukraine	3 April 1948
United Kingdom	22 July 1946
Uzbekistan	22 May 1992
Yugoslavia, Federal Republic of	-

Figure 3.1 Map of the WHO European Region

(Israel, Kyrgyzstan and Tajikistan are not shown on this map)



149,749,900 (Russian Federation). The total population in these countries has had, on average, a 7% increase during this ten year time period. Data were available at a sub-national level for 24 countries in 1980/81 and 31 countries in 1990/91. The mean population for a region (using only the 24 countries available over both time periods) is 1,545,474 with a range from 12,807 (Appenzell-Inner Rhoden in Switzerland) to 1,7044,864 (North Rhine-Westphalia in Germany) in 1980/81. In 1990/91, the mean is 1,558,570, ranging from 13,593 to 17,328,180 (same regions as 1980/81). Table 3.2 summarises the population data for each country.

3.1.3 Mortality Data

Cause-specific death data by gender and region of residence were provided for all ages and for the groups aged 0, 1-14, 15-34, 35-64, 65-74, 75-79, 80-84 and 85 and over. Again the data were provided for the time periods 1980/1981 and 1990/1991. The causes of deaths were based on the Ninth Revision of the International Classification of Diseases (ICD-9). There are 19 specific diagnostic categories within 8 main groups of causes of deaths (see Table 3.3). These causes of death categories account for around 80% of all deaths and constitute the most important causes of death in the European population.

Mortality data are available for 36 countries in 1980/81 and 43 in 1990/91. In the 35 countries with data available over both time periods, there were a total of 6,988,156 deaths recorded in 1980/81 and 7,576,797 in 1990/91. Total deaths have, on average, increased by 8% over the ten year period. Taking into account population size at a regional level, crude death rates were calculated for each country and region. A summary of these is given in Table 3.4. Thirty-four countries have the crude death rates available for both time periods. For these countries, the mean crude death rate for 1980/81 is 1012 (per 100,000), ranging from 540 (Armenia) to 1354 (Hungary). For 1990/91 the mean crude death rate is 1017, with a range from 633 (Armenia) to 1404 (Hungary). Regional crude death rates are available within 21 countries in 1980/81, ranging from 620 (Madrid, Spain) to 1850 (Copenhagen and Frederiksberg (city), Denmark). In 1990/91

Table 3.2 Summarised population data

Country	Year	Total	If regional data available: Mean for all regions	regions with extreme values: Min (region)	regions with extreme values: Max (region)	No. of regions
Albania	-	-	-	-	-	-
Armenia	1980/81	3126741	-	-	-	-
	1990/91	3578200	-	-	-	
Austria	1980/81	7557031	839670	269958 (AT01)	1535279 (AT09)	9
	1990/91	7771755	863529	270571 (AT01)	1522251 (AT09)	
Azerbaijan	1980/81	-	-	-	-	3
	1990/91	7097008	2365669	334753 (AZ03)	5028272 (AZ01)	
Belarus	1980/81	9663343	1380478	1134454 (BY02)	1619560 (BY03)	7
	1990/91	10217277	1459611	1182179 (BY02)	1650042 (BY0401)	
Belgium	1980/81	9859111	3286371	1000983 (BE3)	5632209 (BE1)	3
	1990/91	9967450	3322484	962350 (BE3)	5753850 (BE1)	
Bosnia & Herzegovina	1980/81	-	-	-	-	-
	1990	4480790	-	-	-	
Bulgaria	1980/81	8876326	986259	697860 (BG04)	1241679 (BG05)	9
	1990/91	8688071	965341	642515 (BG04)	1238950 (BG05)	
Croatia	1980/81	4599213	-	-	-	-
	1990/91	4779427	-	-	-	
Czech Republic	1980/81	10447972	1289375	690496 (CZ02)	2042959 (CZ03)	8
	1990/91	10335711	1291964	699487 (CZ02)	2054631 (CZ03)	
Denmark	1980/81	5122300	341487	47534 (DK023)	624780 (DK012)	15
	1990/91	5147618	343175	45705 (DK023)	601853 (DK012)	
Estonia	1980/81	-	-	-	-	-
	1990/91	1570665	-	-	-	
Finland	1980/81	4789756	399147	22789 (FI01)	1128844 (FI11)	12
	1990/91	5000199	416683	24580 (FI01)	1248909 (FI11)	
France	1981	54284652	2467484	240038 (FR83)	10053563 (FR01)	22
	1990/91	56895248	2586148	250156 (FR83)	10737070 (FR01)	
Georgia	1980/81	5067173	-	-	-	-
	1990	5417600	-	-	-	
Germany	1980/81	61616978	5601543	693543 (DE4)	17044864 (DE5)	11
	1990/91	79723463	4982717	680175 (DE4)	17328180 (DE5)	16
Greece	1980/81	9686279	745099	182325 (GR22)	3343654 (GR3)	13
	1990/91	10144347	780335	193305 (GR22)	3451335 (GR3))	
Hungary	1980/81	10711481	535574	239724 (HU13)	2061095 (HU05)	20
	1990/91	10355441	517772	225732 (HU13)	2017200 (HU05)	
Iceland	1980/81	229482	-	-	-	-
	1990/91	256377	-	-	-	
Ireland	1980/81	3424609	-	-	-	-
	1990/91	3521754	-	-	-	
Israel	-	-	-	-	-	-
Italy	1980/81	56433883	2821694	112329 (IT12)	8872566 (IT2)	20
	1990	57576429	2878821	115270 (IT12)	8911995 (IT2)	
Kazakstan	1980/81	-	-	-	-	20
	1990/91	8334889	416745	153028 (KZ16)	927897 (KZ03)	
Kyrgystan	1980/81	-	-	-	-	7
	1990/91	4391366	627338	197739 (KG07)	1315872 (KG06)	
Latvia	1980/81	2515561	-	-	-	-
	1990/91	2666542	-	-	-	
Lithuania	1980/81	3423077	-	-	-	-
	1990/91	3726225	-	-	-	
Luxembourg	1980/81	364688	-	-	-	-
	1990/91	384475	-	-	-	
Macedonia	-	-	-	-	-	-
Malta	1980/81	319000	-	-	-	-
	1990/91	355900	-		-	
Monaco	-	-	-	-	-	-
Netherlands	1980/81	14199897	1183325	351383 (NL74)	3105220 (NL73)	12
	1990/91	15013980	1154922	221893 (NL25)	3246364 (NL73)	

(Continued over page)

Norway	1980/81	4092657	227370	78331 (NO03)	821676 (NO10)	18
	1990/91	4251601	236200	74629 (NO03)	880318 (NO10)	
Poland	1980/81	35739989	729388	230989 (PL05)	3741807 (PL13)	49
	1990/91	38181654	779218	247169 (PL05)	3988987 (PL13)	
Portugal	1980/81	9858400	-	-	-	7
	1990/91	9982037	1426005	240596 (PT2)	3510567 (PT11)	
Republic of Moldova	1980/81	4033407	-	-	-	-
	1990/91	4361105	-	-	-	
Romania	1980/81	22277011	543342	215365 (RO16)	2060237 (RO10)	41
	1990/91	23195902	565754	236249 (RO16)	2328932 (RO10)	
Russian Federation	1980/81	139221495	1907144	272972 (RU1005)	8298614 (RU0314)	79
	1990/91	149749900	1874418	158592 (RU1109)	8899639 (RU0314)	
San Marino	-	-	-	-	-	-
Slovakia	1980/81	5000682	1250171	381957 (SK01)	1687363 (SK04)	4
	1990/91	5290589	1322648	443240 (SK01)	1722077 (SK04)	
Slovenia	1980/81	1889871	-	-	-	-
	1990/91	1999929	-	-	-	
Spain	1980/81	37750844	2097269	118893 (ES63)	6464996 (ES61)	18
	1990	38924600	2162478	124800 (ES63)	6903000 (ES61)	
Sweden	1980/81	8320485	346687	55485 (SE05)	1531870 (SE19)	24
	1990/91	8641322	360055	57343 (SE05)	1655755 (SE19)	
Switzerland	1980/81	6372691	245104	12807 (CH03)	1124037 (CH26)	26
	1990/91	6752248	259702	13593 (CH03)	1151264 (CH26)	
Tajikistan	1980/81	-	-	-	-	6
	1990/91	5367875	894647	168000 (TJ02)	1637532 (TJ05)	
Turkey	-	-	-	-	-	-
Turkmenistan	1980/81	-	-	-	-	5
	1990/91	3703631	740726	413334 (TM01)	923170 (TM05)	
Ukraine	1980/81	49960794	1921569	904387 (UA0103)	5180225 (UA0302)	26
	1990/91	51663235	1987048	941343 (UA0103)	5324226 (UA0302)	
United Kingdom	1980/81	56340950	1006089	110330 (UK914)	6828083 (UK55)	56
	1990/91	57680993	1030018	117698 (UK914)	6871241 (UK55)	
Uzbekistan	1980/81	-	-	-	-	13
	1990/91	10210082	785393	286827 (UZ11)	1176162 (UZ09)	
Yugoslavia	1980/81	9879486	4939743	582706 (YU01)	9296781 (YU03)	2
	1990	10529295	5264648	644302 (YU01)	9884993 (YU03)	

Table 3.2 gives the WHO codes for the regions with extreme rates within each country. Appendix A1.2 gives the corresponding region names (and also gives region names for Tables 3.4 and 3.5)

regional crude death rates are available for 30 countries, ranging from 302 (Dahalal-Abad, Kyrgystan) to 1775 (Mikhaylovgrad, Bulgaria).

For the purpose of this thesis, we are interested in examining a subset of these data. Since the main focus is on European spatial patterns of cancer mortality, all malignant neoplasms (ICD-9, 140- 208) are summarised in Table 3.5. Cancer mortality data are available for 41 countries in 1980/81 and 40 in 1990/91. Taking account of population size at a regional level, crude death rates were calculated for each country and region. Thirty-one countries have the crude death rates available for both time periods. For these countries, the mean crude death rate for 1980/81 is 191 (per 100,000), ranging from 75 (Armenia) to 309 (Denmark). For 1990/91 the mean crude death rate is 214, with a range from 99 (Armenia) to 348 (Denmark). Regional crude death rates are available for 19 countries in 1980/81, ranging from 73 (Montenegro, Federal Republic of Yugoslavia) to 541 (Copenhagen and Frederiksberg (city), Denmark). In 1990/91 regional crude death rates are available for 31 countries, ranging from 21 (Dahalal-Abad, Kyrgystan) to 487 (Copenhagen and Frederiksberg (city), Denmark).

3.2 Data Issues

When collecting data on such a wide scale globally, inevitably certain problems will occur resulting in limitations to the final dataset.

3.2.1 Missing Data

The full data set should ideally contain information on the sub-national level from all Member States of the WHO European Region. This was not possible for many reasons resulting in deviations in the availability of data, including not only lack of data of a certain type or from a certain year but also cases in which data were supplied for different years or with different age aggregations from those requested.

Table 3.3 Specific cause mortality data

Causes of death according to ICD-9	code
Infectious and parasitic diseases	001 - 138
Malignant Neoplasms	**140 - 208**
Malignant Neoplasm of oesophagus	150
Malignant Neoplasm of colon, rectum rectosigmoid junction and anus	153 - 154
Malignant Neoplasm of liver, specified as primary	155
Malignant Neoplasm of trachea, bronchus and lung	162
Malignant Neoplasm of female breast	174
Malignant Neoplasm of bladder	188
Leukaemia	204 - 208
Diseases of the circulatory system	**390 - 459**
Ischaemic heart disease	410 - 414
Diseases of the pulmonary circulation and other forms of heart disease	415 - 429
Cerebrovascular disease	430 - 438
Atherosclerosis	440
Diseases of the respiratory system	**460 - 519**
Pneumonia	480 - 486
Chronic obstructive pulmonary disease and allied conditions	490 - 496
Diseases of the digestive system	**520 - 579**
Chronic liver disease and cirrhosis	571
Disease of the urinary system	**580 - 599**
Congenital anomalies	**740 - 759**
Accidents, injury and poisoning	**E800 – E999**
Motor vehicle traffic accidents	E810 – E819
Other transport and self-inflicted injury	E800 – E807, E826 – E845
Suicide and self-inflicted injury	E950 – E959

Data collection involved making individual requests from national statistical offices. This process coincided with major political, economic and social changes in the region. For example, during the period of data collection, 16 new Member States were acquired, most of them newly independent states of the former USSR, together with the unification of Germany. This sometimes led to a difficulty in access to information and also could have complicated communication with the national statistical offices. To reduce complications, data for the same administrative areas have been used over the decade, but sometimes under new names and within the borders of new countries.

Also, due to the wide range of countries, data collection was constrained by the availability of databases as well as their quality and completeness. The requested resolution of the region of residence was not always obtained. Sometimes data were only available on a national level or a level lower than NUTS II. For such countries the higher level of data has been used.

Often data for certain years were missing, with some countries only being able to provide data for one decade rather than two. For example, in countries where population data are collected only by census, data may be available only for the years when the census was made. Trends can still be examined as long as one year is available out of each of the two consecutive years. In these instances the data summaries use one year instead of the mean over the two years.

The number of residents (midyear estimates) and data for all-cause mortality by gender and region of residence were collected for all ages and for specific age groups. Not all countries, however, could supply the data in the requested format. For example in some countries such age divisions could not be achieved for elderly people or sometimes groups aged under 1 year were aggregated with those for the group ages 1-4 years. When population and mortality data were obtained in age categories that were slightly different from those requested, account was taken of these exceptions in the standardisation of rates by grouping the age categories of the standard population similarly to the categories of the index population.

Table 3.4 Summarised mortality data

Country	Year	Total Deaths	Crude Death Rate (per 100000)	If regional data available: regions with extreme values Min (region)	Max (region)	No. of regions
Albania	-	-	-	-	-	-
Armenia	1980/81	16892	540	-	-	-
	1990/91	22707	633	-	-	
Austria	1980/81	92568	1225	781 (AT08)	1648 (AT09)	9
	1990/91	83190	1070	753 (AT08)	1369 (AT09)	
Azerbaijan	1980/81	-	-	-	-	3
	1990/91	43735	616	508 (AZ03)	670 (AZ04)	
Belarus	1980/81	94325	976	-	-	7
	1990/91	112083	1098	686 (BY0401)	1415 (BY06)	
Belgium	1980/81	113012	1146	1038 (BE1)	1325 (BE3)	3
	1990/91	-	-	-	-	
Bosnia & Herzegovina	1980	13048	-	-	-	-
	1990	15174	678	-	-	
Bulgaria	1980/81	96696	1090	-	-	9
	1990/91	109516	1262	1050 (BG07)	1775 (BG04)	
Croatia	1980/81	50760	1104	-	-	-
	1990/91	53512	1120	-	-	
Czech Republic	1980/81	132972	1274			8
	1990/91	126728	1226	1105 (CZ05)	1385 (CZ06)	
Denmark	1980/81	56149	1096	741 (DK014)	1850 (DK011)	15
	1990/91	60254	1171	864 (DK014)	1751 (DK011)	
Estonia	1980/81	-	-	-	-	-
	1990/91	19585	1247	-	-	
Finland	1980/81	44521	929	804 (FL06)	1105 (FL07)	12
	1990/91	49697	994	857 (FL08)	1198 (FL07)	
France	1981	554823	1022	833 (FR1)	1430 (FR63)	22
	1990/91	525443	924	731 (FR1)	1324 (FR63)	
Georgia	1980/81	43654	861	-	-	-
	1990	46473	858	-	-	
Germany	1980/81	718155	1166	1008 (DE8)	1826 (DEBW)	11
	1990/91	916345	1149	992 (DE8)	1399 (DEE)	16
Greece	1980/81	86752	896	799 (GR3)	1435 (GR41)	13
	1990/91	94825	935	834 (GR12)	1356 (GR41)	
Hungary	1980/81	145056	1354	1175 (HU12)	1526 (HU15)	-
	1990/91	145237	1404	1230 (HU07)	1513 (HU15)	
Iceland	1980/81	1597	696	-	-	-
	1990/91	1750	684	-	-	
Ireland	1980/81	33201	869	-	-	-
	1990	31457	893	-	-	
Israel	-	-	-	-	-	-
Italy	1980/81	549901	974	780 (IT91)	1296 (IT13)	20
	1990	543708	944	733 (IT91)	1324 (IT13)	
Kazakstan	1980/81	-	-	-	-	20
	1990/91	64309	772	672 (KZ03)	1177 (KZ06)	
Kyrgystan	1980/81	-	-	-	-	7
	1990/91	30684	699	302 (KG03)	960 (KG04)	
Latvia	1980/81	32095	1276	-	-	-
	1990/91	34781	1034	-	-	
Lithuania	1980/81	-	-	-	-	-
	1990	39713	1070	-	-	
Luxembourg	1980/81	4109	1128	-	-	-
	1990/91	3759	978	-	-	
Macedonia	-	-	-	-	-	-
Malta	1980/81	3219	1009	-	-	-
	1990/91	2791	784	-	-	
Monaco	-	-	-	-	-	-
Netherlands	1980/81	114897	809	676 (NL51)	939 (NL74)	12
	1990/91	129391	862	495 (NL25)	968 (NL11)	
Norway	1980/81	41618	1017	847 (NO03)	1161 (NO04)	18
	1990/91	45432	1069	837 (NO12)	1269 (NO04)	

(Continued over page)

Poland	1980/81	339563	950	714 (PL19)	1178 (PL36)	49
	1990/91	396196	1038	817 (PL19)	1361 (PL21)	
Portugal	1980/81	95432	968	-	-	7
	1990/91	103738	1039	913 (PT11)	1376 (PT14)	
Republic of Moldova	1980/81	40974	1016	-	-	-
	1990/91	44138	1012	-	-	
Romania	1980/81	228256	1025	799 (RO25)	1484 (RO02)	41
	1990/91	249423	1075	830 (RO08)	1455 (RO02)	
San Marino	-	-	-	-	-	-
Russian Federation	1980/81	1525021	1095	-	-	79
	1990/91	1670959	1128	387 (RU1109)	1522 (RU0203)	
Slovakia	1980/81	50106	1002	904 (SK01)	1104 (SK04)	4
	1990/91	54619	1032	882 (SK01)	1128 (SK04)	
Slovenia	1980/81	18777	994	-	-	-
	1990/91	18940	947	-	-	
Spain	1980/81	291365	772	620 (ES3)	931 (ES43)	18
	1990	333142	856	680 (ES7)	1012 (ES12)	
Sweden	1980/81	91917	1105	967 (SE20)	1372 (SE07)	24
	1990/91	95182	1101	899 (SE20)	1337 (SE07)	
Switzerland	1980/81	59484	933	684 (CH04)	1278 (CH05)	26
	1990/91	63187	936	693 (CH25)	1420 (CH05)	
Tajikistan	1980/81	-	-	-	-	6
	1990/91	33001	615	541 (TJ07)	684 (TJ01)	
Turkey	-	-	-	-	-	-
Turkmenistan	1980/81	-	-	-	-	5
	1990/91	26536	716	683 (TM05)	759 (TM04)	
Ukraine	1980/81	568516	1138	834 (UA0114)	1370 (UA0304)	26
	1990/91	649413	1258	898 (UA0114)	1533 (UA0102)	
United Kingdom	1980/81	661619	1174	866 (UK522)	1658 (UK531)	56
	1990/91	645942	1120	832 (UK522)	1505 (UK562)	
Uzbekistan	1980/81	-	-	-	-	13
	1990/91	61713	604	506 (UZ03)	792 (UZ13)	
Yugoslavia	1980/81	90118	912	634 (YU01)	930 (YU03)	2
	1990	97665	928	611 (YU01)	948 (YU03)	

Table 3.5 Summarised cancer mortality data

Country	Year	Total Deaths	Crude Death Rate (per 100000)	If regional data available: regions with extreme values Min (region)	Max (region)	No. of regions
Albania	-	-	-	-	-	-
Armenia	1980/81	2330	75	-	-	-
	1990/91	3526	99	-	-	
Austria	1980/81	19224	254	170 (AT08)	352 (AT09)	9
	1990/91	19324	249	190 (AT08)	309 (AT09)	
Azerbaijan	1980/81	-	-	-	-	3
	1990/91	4804	68	55 (AZ03)	88 (AZ04)	
Belarus	1980/81	12155	126	-	-	7
	1990/91	17771	174	147 (BY0401)	231 (BY06)	
Belgium	1980/81	26537	269	263 (BE1)	310 (BE3)	3
	1990/91	-	-	-	-	
Bosnia & Herzegovina	-	-	-	-	-	-
Bulgaria	1980/81	12839	145	120 (BG02)	192 (BG04)	9
	1990/91	30551	175	142 (BG06)	222 (BG04)	
Croatia	1980/81	8538	186	-	-	-
	1990/91	10611	222	-	-	
Czech Republic	1980/81	26415	256	-	-	8
	1991	28102	273	249 (CZ05)	310 (CZ01)	
Denmark	1980/81	15811	309	226 (DK013)	541 (DK011)	15
	1990/91	17899	348	288 (DK013)	487 (DK011)	
Estonia	1980/81	-	-	-	-	-
	1990/91	3331	212	-	-	

(Continued over page)

Finland	1980/81	8998	188	155 (FI06)	235 (FI01)	12
	1990/91	9718	194	161 (FI08)	221 (FI10)	
France	1981	126632	233	210 (FR51)	318 (FR63)	22
	1990/91	138629	244	203 (FR1)	323 (FR63)	
Georgia	1980/81	-	-	-	-	-
	1990	5645	104	79 (GE02)	126 (GE05)	
Germany	1980/81	157662	256	218 (DE8)	360 (DEBW)	11
	1990/91	207843	261	201 (DEC)	314 (DE2)	16
Greece	1980/81	16418	169	136 (GE43)	232 (GE41)	13
	1990/91	19689	194	162 (GE42	230 (GE22)	
Hungary	1980/81	27720	259	193 (HU16)	334 (HU05)	20
	1990/91	31139	301	253 (HU07)	364 (HU05)	
Iceland	1980/81	349	152	-	-	-
	1990/91	450	176	-	-	
Ireland	1980/81	6254	183	-	-	-
	1990	7217	205	-	-	
Israel	-	-	-	-	-	-
Italy	1980/81	122776	218	125 (IT93)	323 (IT33)	20
	1990	145036	252	152 (IT93)	365 (IT33)	
Kazakstan	1980/81	-	-	-	-	20
	1990	22595	136	93 (KZ03)	214 (KZ06)	
Kyrgystan	1980/81	-	-	-	-	7
	1990/91	3361	77	21 (KG03)	146 (KG01)	
Latvia	1980/81	-	-	-	-	-
	1990/91	5535	208	-	-	
Lithuania	1980/81	-	-	-	-	-
	1990	6956	187	-	-	
Luxembourg	1980/81	955	262	-	-	-
	1990/91	963	250	-	-	
Macedonia	-	-	-	-	-	-
Malta		-	-	-	-	-
Monaco	-	-	-	-	-	-
Netherlands	1980/81	31134	219	159 (NL51)	258 (NL74)	12
	1990/91	35409	236	163 (NL25)	268 (NL11)	
Norway	1980/81	8809	215	147 (NO03)	247 (NO10)	18
	1990/91	9819	231	185 (NO03)	256 (N011)	
Poland	1980/81	60269	169	130 (PL16)	218 (PL21)	49
	1990/91	73436	192	152 (PL24)	255 (PL21)	
Portugal	1980/81	14266	145	-	-	7
	1990/91	18204	182	154 (PT11)	234 (PT14)	
Republic of Moldova	1980/81	3989	99	-	-	-
	1990/91	5751	132	-	-	
Romania	1980/81	28782	129	88 (RO04)	194 (RO10)	41
	1990/91	32853	142	87 (RO39)	207 (RO02)	
San Marino	-	-	-	-	-	-
Russian Federation	1980/81	225689	162	-	-	79
	1990/91	287362	194	68 (RU1109)	270 (RU0204)	
Slovakia	1980/81	8600	172	-	-	-
	1990/91	10377	196	-	-	
Slovenia	1980/81	3506	186	-	-	-
	1990/91	4206	210	-	-	
Spain	1980/81	58862	156	120 (ES3)	192 (ES12)	18
	1990	76823	197	157 (ES63)	244 (ES12)	
Sweden	1980/81	19659	236	184 (SE14)	270 (SE07)	24
	1990/91	20369	236	194 (SE14)	284 (SE04)	
Switzerland	1980/81	14887	234	116 (CH22)	348 (CH05)	26
	1990/91	16473	244	159 (CH25)	392 (CH05)	
Tajikistan	1980/81	-	-	-	-	6
	1990/91	2637	49	28 (TJ04)	97 (TJ01)	
Turkey	-	-	-	-	-	-
Turkmenistan	1980/81	-	-	-	-	5
	1990/91	2325	63	49 (TM04)	96 (TM01)	
Ukraine	1980/81	76210	153	-	-	26
	1990	101352	196	125 (UA0112)	225 (UA0308)	
United Kingdom	1980/81	146712	260	200 (UK511)	357 (UK531)	56
	1990/91	162026	281	217 (UK521)	371 (UK562)	
Uzbekistan	-	-	-	-	-	-
Yugoslavia	1980/81	12385	125	73 (YU01)	129 (YU03)	2
	1990	15568	148	94 (YU01)	151 (YU01)	

The data sometimes had missing mortality counts for specified ICD groups or the data were aggregated otherwise than defined. This had minimal effect on total mortality counts. If countries had specific mortality groups with different definitions in terms of ICD codes they were classified as missing.

3.2.2 Data Quality

Data reliability is obviously of concern when collecting mortality statistics on such a large scale. This depends somewhat on how accurately causes of death have been classified. For instance, often a patient has a known cancer diagnosis, however, it is not always simple to classify whether the disease was irrelevant, an underlying cause or a contributing cause of death. In less developed countries, where there is less public awareness of early cancer signs, poor access to health care and/or low autopsy rates, there will be many more opportunities for misclassification of cause of death (67).

Despite the rather large amount of data requested by WHO, 90% of the approached Member States returned data. When a considerable amount of data are collected from several sources and almost 50 countries, inevitably some errors and misunderstandings will occur. In order to ensure a high standard of data quality, numerous checks on quality and plausibility were carried out. Most data sets were forwarded to NCBS (Netherlands Central Bureau of Statistics) for checks. Some data were sent directly to the ECEH (European Centre for Environment and Health). They were checked by the ECEH and RIVM (International Environmental Data Service). If the analysis of final data sets revealed inconsistencies, these were corrected after consulting with NCBS.

The procedures for data quality and plausibility included verification of the completeness of the data received, arithmetic checks and comparison of crude mortality rates (comparing regional level data with a specified reference line). If data appeared incomplete or contained errors, additional or revised data were requested from Member States. If further requests to countries did not yield improvement, the section of the data was classified as missing. Also, population

and mortality data at a national level were verified with statistics published in the editions of the World Health Statistics Annual for the years 1980, 1981, 1990 and 1991. The final verification of all national level data did not reveal major differences and the final survey of the contents of the regional level data gave a good overall impression of the plausibility of the data.

3.2.3 Sources of Bias

Mortality data are widely considered to have the highest degree of international comparability in developed countries because deaths are mostly reported in accordance with international reporting standards and therefore counts are fairly accurate. However, as with most data collection and processing, the risks of bias and random error are still highly probable. Reasons for the bias occurring in this type of data are that countries, and sometimes regions within countries, differ in their training of medical staff and their use of diagnostic technologies and autopsies to confirm causes of deaths. Also, the accuracy of death certificates differs between countries in Europe. This could be due to the level of health care services, numbers of specialists or use of screening tests for detection of diseases. International differences in mortality data also depend on methods of data collection and procedures of coding (eg the procedure used for nationals or residents dying abroad).

Migration in and out of regions can also affect mortality rates as this often changes the size and composition of the population dwelling in particular areas. However, most member states should have avoided this affecting the accuracy of their rates due to population censuses being carried out every ten years and with birth and death counts being carried out in between censuses.

Certain cancers are not meaningfully represented by mortality rates. It should be noted that the burden of cancers with favourable prognosis, such as the common non-melanoma skin cancer and endometrial cancer or the more rare thyroid and testis cancer, are not reflected by mortality rates. For those cancers

that remain almost inevitably fatal such as lung, oesophagus, liver and pancreas, mortality rates are actually good approximations of incidence rates.

The high standards of the World Health Organisation should reflect the quality of the data. All data collection and processing procedures have been carried out thoroughly so the reliability and plausibility of the data should be rated highly. In view of factors outwith the control of WHO that may have affected the data, some caution should be made when making international comparisons of cause-specific mortality rates. To minimise the influence of random fluctuations and give clear pictures of patterns of mortality, modelling techniques are necessary to analyse these data.

3.2.4 Mortality Data Advantages

Despite the use of population-level mortality data having drawbacks, they do have overriding advantages. As previously mentioned and as can be seen from the vast amount of cancer mortality literature, cancer mortality rates are widely accepted as important measures of the burden of cancer.

One of the main advantages of examining mortality data, and the probable reason for its common usage is that it is the most widely available measure of cancer burden. This is because the compilation of cancer mortality statistics is a simpler task than that of cancer incidence, prevalence or survival. In economically developed countries, such as many of those within the WHO European region, the coverage of mortality statistics collection was close to 100% in 1990 (68).

3.3 Cancer Mortality Data in the EU

In subsequent Chapters a subset of the main mortality dataset will be explored. Malignant neoplasms (ICD-9, 140- 208) are examined for the EU countries in the time-period 1991. These data are summarised in Table 3.6.

Taking account of population size at the regional level, crude death rates have been calculated for each country and region. These are given for each country in Table 3.6 and corresponding extreme rates are given for regions within these countries. The mean crude cancer death rate in 1991 is 257 (per 100,000) ranging from 157 (Norte, Portugal) to 487 (Copenhagen and Frederiksberg (city), Denmark). Data are missing for 4 of the 15 EU countries in 1991 and regional crude death rates are available for 10 of the remaining 11 countries.

Table 3.6 Summarised cancer mortality data in the EU in 1991

Country	**Total Deaths**	Crude Death Rate **(per 100000)**	If regional data available: regions with extreme values **Min** (region)	**Max** (region)	No. of regions
Austria	19317	247	193 (Vorarlberg)	304 (Vienna)	9
Belgium	*missing*				
Denmark	17764	345	290 (Sønderjylland)	487 (Copenhaen and Frederiksberg (city))	15
Finland	9626	192	165 (Oulu)	234 (Fahvenanmaa)	12
France	138778	243	203 (Île de France)	322 (Limousin)	22
Germany	210537	263	207 (Brandenburg)	320 (Hamburg)	11
Greece	19945	196	168 (Aegean North)	236 (Ionian Islands)	13
Ireland	*missing*				
Italy	*missing*				
Luxembourg	957	247	-	-	-
Netherlands	35640	237	169 (Flevoland)	273 (Groningen)	12
Portugal	18203	185	157 (Norte)	2323 (Alentejo)	7
Spain	*missing*				
Sweden	20406	235	201 (Norrbotten)	305 (Gävleborg)	24
UK	161555	280	219 (Northern Ireland)	377 (Isle of Wight)	56

3.4 Risk Factor Data

Ecological studies aim to ascertain associations between a disease and exposure to risk or protective factors for groups or populations. It is therefore necessary to analyse data that reflect levels of exposure to such factors along with population and mortality data.

In Chapter 2 we discussed existing literature on cancer mortality risk and protective factors. Evidence suggests that the main factors affecting some of the most common cancer groups are diet, specifically fruit, vegetable and animal fat consumption, alcohol intake, smoking level and socio-economic status.

3.4.1 Source of Data

Data were obtained from various sources to reflect exposure levels to these risk and protective factors.

3.4.1.1 Diet

Data reflecting levels of consumption of fruit, vegetables and animal fat were obtained from the Food and Agriculture Organisation of the United Nations. FAOSTAT is their online statistical database (69) and they compile information and data on various aspects of food and agriculture from all countries. In this instance, country level information was used from food balance sheets. This comprises the average amounts of fruit, vegetables and animal fat available for human consumption during the period 1991, and are measured in kilograms. Alcohol consumption for each European country was also available from FAOSTAT in kilograms for the year 1991.

3.4.1.2 Smoking

Data to reflect levels of smoking within European countries were obtained from a World Health Organisation global status report. The Tobacco or Health report (70) was carried out in 1997 and the details are provided on the internet as a

service by the Office on Smoking and Health of the National Centre for Chronic Disease Prevention and Health Promotion (71). The annual adult (age over fifteen) consumption of manufactured cigarettes per capita was obtained for each country for the time period 1990-1992.

3.4.1.3 Socio-Economic Status

Socio-economic status (SES) is some description of a person's position in society such as income, educational level attained, occupation or value of dwelling place. A good indicator of SES on a population level is Gross Domestic Product (GDP) which is a relative measure of wealth of individual regions or countries. GDP data were obtained from the World Health Organisation (72) and are available for each country in Europe. It is measured in US dollars per inhabitant in 1995. GDP was available at a sub-national level for EU countries only from Eurostat (73); these data will be summarised, along with the other EU risk factors data in Chapter 5.

3.4.2 Data Summaries

Risk factor data are summarised in Table 3.7; for each risk or protective factor the median, minimum and maximum and the countries with extreme levels of these factors are given.

3.4.3 Data Issues

Various risk factor data have been collected on a large global scale, which again may cause occasional discrepancies.

Table 3.7 Summarised risk factor data

Risk/protective factor	Median	Minimum	Maximum
Fruit kg/year/capita	74.5 (UK)	24.6 (Estonia)	142.6 (Greece)
Vegetables kg/year/capita	88.2 (UK)	34.5 (Israel)	300.4 (Greece)
Animal Fat kg/year/capita	14 (Russian Federation)	1.3 (Armenia)	29 (Hungary)
Alcohol kg/year/capita	76.2 (Sweden)	9.3 (Armenia)	173.9 (Germany)
Smoke cigarettes/year/adult	1920 (Italy)	910 (Azerbaijan)	3620 (Poland)
GDP (region level) US$/inhabitant	12670 (Spain)	2180 (Tajikistan)	21780 (Switzerland)

3.4.3.1 Time Periods

The time periods used for each of the risk/protective factors might not be ideal. It is difficult to determine at which time period each exposure should be measured, for example at which period in a person's life their diet most affects their risk of cancer. For a population it is clearly impossible to consider accumulated lifetime exposure to such factors. Also, there is sometimes a lack of availability of risk/protective factor data at specific time periods. For these reasons, data that exist for approximately the same period as our mortality data have been obtained and are used in subsequent Chapters for the modelling in an attempt to reflect the cultures of the different countries or regions. However, it is important to remain aware that, if the change over time in patterns of exposure to risk or protective factors has differed substantially between these regions or countries, then this will have an impact upon both the estimated relationship with such risk factors and the estimated adjusted risk of mortality in these countries.

3.4.3.2 Country Level

All the data collected to reflect the levels of risk or protective factors are at the country level except for GDP in the EU. It would obviously be more informative for all the data to be at the less aggregated level. However, since it was necessary to obtain different types of data from different countries this was not possible. These data are still very useful as they should represent the country's dietary and lifestyle habits. When we are interpreting the results of analysing such data it should be clear that differing levels of exposure to risk factors within countries have not been taken into account.

3.4.3.3 Data Quality

Similar issues of data quality and reliability arise with the risk factor data as did with the mortality and population data. For example, the country-level consumption data were collected by various methods: tailored questionnaires sent to member countries, magnetic tapes, diskettes, FTP transfers and accessing websites of the countries, national and international publications, country visits made by the Food and Agriculture organisation (FAO) statisticians and reports of FAO representatives in member countries. Therefore maximum effort was made to gather accurate statistics on factors such as food consumption, but FAOSTAT comment that data reliability, especially in developing countries, may in some cases be questionable. These problems cannot be overcome and one should just remain aware when drawing any conclusions that the type of data under examination means that any form of analysis is fairly crude.

Chapter 4

4 Disease Mapping Review

4.1 Introduction to Disease Mapping

Understanding the geographical distribution of disease mortality or incidence is of key importance in the field of public health. Mapping incidence or mortality rates from diseases such as cancer is one of the primary tools used to investigate the spatial variability of risk from specific diseases. The main aim in doing so is to produce a map 'clean' of random noise and any natural variation in the human population, and which allows identification of areas with high or low rates. These maps play an important role in epidemiology allowing, for example, the geographical assessment of health resource allocation and the formulation of aetiological hypotheses.

4.1.1 Standardised Mortality Ratio

Disease mapping initially involves the choice of an epidemiological measure which shall be displayed on the map. Traditionally the measure used is a standardised mortality/morbidity ratio (SMR). This gives a geographical picture of the disease rates expected in a given area compared to the observed rate. For example, if we consider a population of regions, i, $i = 1, \ldots, I$, with observed (O_i) and expected (E_i) counts of deaths. The E_i are often calculated through standardisation based on the number of deaths in the population N_i. This standardisation is conducted for age and sex strata k, and if stratum k has O_{ik} out of a population N_{ik}, the expected number of events is given by

$$E_{ik} = N_{ik} \frac{\sum_i O_{ik}}{\sum_i N_{ik}},$$

$$E_i = \sum_k E_{ik}.$$

The standardised mortality ratio for the disease of interest is therefore

$$\mathrm{SMR}_i = \frac{O_i}{E_i} \times 100.$$

To produce maps of the SMRs, the chloropleth method (74) is often used which involves specifying the SMRs into class intervals and assigning to each interval a specific colour, shade or pattern. The choice of interval categorisation can be of importance as it can fundamentally change the geographical picture of the disease. A common approach is that of equal-interval classification in which the range of data values is divided into a fixed number of classes whereby each class represents an equal range of SMRs. A problem with this method is that, if we have a highly skewed data distribution, the majority of areas will probably fall into 1 or a small number of classes. The map will therefore show little spatial variation with the majority of SMRs being classified as similar when in fact they are not. Another traditional approach is based on the percentiles of the SMR distribution, such as quartiles, quintiles or sixtiles. This method ensures that each class is equally represented on the map. However, this method can also be misleading because areas with similar SMRs may be assigned to different classes. Some areas will appear to be very heterogeneous on the map when in fact they are not. Another method used to determine class intervals is to search for natural divisions in the distribution of SMRs. These classes would represent clusters of SMRs and the approach differs from the two previous methods, as it does not arbitrarily assign observations that have similar values into different classes. To

use this method some form of cluster analysis would probably be performed so clusters are determined statistically that minimise within-class variation. Since the map is often used to determine groups of areas with similar disease, this method may over-emphasise any clusters that exist. The choice of classification method is important but does depend on specific data distributions. Muehrcke et al (75) gives a fuller discussion of ways of defining disease mapping class intervals.

Once the SMRs are displayed on a map the intention is to allow interpretation of the geographical variation in the disease. However, this method has limitations and has been criticised by various authors, particularly Clayton and Kaldor (76). When examining SMRs, the ratios for small or sparsely populated areas have large variability, and therefore often dominate the map, as there are few or no observations. An alternative method that has been commonly used, but is also open to criticism, is to display the statistical significance levels for a test of the difference between the risk in a specific area and that from the overall rate on the map. This type of map ignores the size of the effect in that two areas with identical SMRs may appear to be quite different if they are of unequal population sizes, and the most prominent areas may simply be those with the largest populations. Also, risk factors often play an important role when studying a disease's geographical variation, and traditional methods do not allow potential risk factors to be taken into account.

4.1.2 Poisson Model

To overcome such problems modern approaches use interpolation methods to improve on the raw rates. A multiplicative model originally proposed by Breslow and Day (77) allows the standardised rates to be mapped based on Poisson inference. Estimating the parameters using maximum likelihood and mapping the modelled standardised rates has the advantage of providing estimates of the parameters, namely the disease rates (78). This assumes that the observed cases O_i follow a Poisson distribution (79) with

$$O_i \sim Poisson(\theta_i E_i), \tag{4.1}$$

where θ_i are the unknown area-specific relative risks of mortality from the disease. The likelihood of the relative risk θ_i is

$$g(O_i|\theta_i) = \frac{e^{-(\theta_i E_i)}(\theta_i E_i)^{O_i}}{O_i!}, \tag{4.2}$$

where again E_i denotes the expected number of cases in region *i*. The maximum likelihood estimate (MLE) of θ_i is the SMR for the *i*th area:

$$\hat{\theta}_i = O_i / E_i,$$

with estimated standard error

$$\hat{s}_i = \sqrt{\hat{\theta}_i} / E_i.$$

However, this again leads to the most extreme SMRs tending to be based on counts from small or sparsely populated areas. On the other hand, *p*-values can be calculated under the null hypothesis $\theta = 1$ or the adjusted null hypothesis which is based on

$$\hat{\theta} = \sum_{i=1}^{n} O_i \Big/ \sum_{i=1}^{n} E_i,$$

where n is the number of areas (80), which compares the SMR to other areas on the map. In the event of no deaths in a region, which commonly occurs in small populations and for rare diseases, the probability is estimated as 0 (81). On the contrary, areas with extreme *p*-values are often only identifying areas with large populations. Although modelling methods have been used, the maps of estimates of SMRs and *p*-values can still be difficult and often misleading.

For sparsely populated areas and rare diseases, the observed counts fluctuate about the mean within each area more than would be expected from a Poisson

distribution. If this extra-Poisson heterogeneity (82) is ignored, the overdispersion can create the impression of artificial geographic variation in the disease rates. Another problem in using conventional Poisson based methods is that they do not take account of any spatial pattern of disease, i.e. they ignore the fact that areas geographically close to one another share similar disease rates and common factors which influence the incidence and outcome of disease. A further disadvantage of methods discussed so far is that they do not allow the inclusion of ecological covariates. Inclusion of such variables would allow the construction of disease maps that take into account risk factors that are known to affect a specific disease. Methods that overcome these problems produce more accurate maps of disease mortality (and incidence) and allow a closer assessment of the true underlying distribution of a disease.

4.2 Bayesian Methods for Disease Mapping

Bayesian methods are now widely used to overcome such problems and there have been vast developments in the area recently. Lawson et al (4) discusses in detail many of the recent methodologies for the statistical evaluation of disease mapping. Such methods allow the production of smoothed estimates of relative risks and maps 'clean' of random noise, influential spatial factors and any natural variation in the human population.

4.2.1 Bayesian Approaches to Relative Risks

The various Bayesian statistical smoothing techniques assume that the relative risks θ_i are random effects that arise from a probability distribution of risks (76, 83). A random effects model can be fitted in which the relative risks have prior distributions and these prior distributions have hyper-parameters, which can have hyper-distributions also. The fully Bayesian approach bases inference on sampled parameters from the joint posterior distribution. The empirical Bayes approach

involves estimating the parameters of the prior distribution and inference can then be made conditional on these estimated parameters (78).

Mollié (79) outlines the approaches to Bayesian inference for relative risks. Disease mapping data combine two types of information, namely the information provided in each area by the observed deaths described by the Poisson likelihood $g(\boldsymbol{\theta}|\mathbf{O})$, and prior information on the relative risks specifying their variability in the overall map, summarised by their prior distribution ($\boldsymbol{\theta}$).

Bayesian inference about the unknown relative risks $\boldsymbol{\theta}$ is based on the marginal posterior distribution

$$g(\boldsymbol{\theta}|\mathbf{O},\boldsymbol{\gamma}) \propto g(\mathbf{O}|\boldsymbol{\theta})\, f(\boldsymbol{\theta}|\boldsymbol{\gamma}), \tag{4.3}$$

where $\mathbf{O}$ is the observed data and $\boldsymbol{\gamma}$ are the hyper-parameters. The prior probability density function of the relative risks is given by $f(\theta_i)$.

The likelihood function of the relative risks $\boldsymbol{\theta}$ for the data (observed number of deaths) $\mathbf{O}$ is the product of n independent Poisson distributions and can be written as

$$g(\mathbf{O}|\,\boldsymbol{\theta}) = \prod_i g(O_i|\,\theta_i), \tag{4.4}$$

$$f(\boldsymbol{\theta}) = \prod_i f(\theta_i). \tag{4.5}$$

The prior distribution ($\boldsymbol{\theta}$) reflects prior belief about variation in relative risks over the map and should be parameterised by hyperparameters γ and denoted $f(\boldsymbol{\theta}|\,\boldsymbol{\gamma})$. Equation (4.3) can be used as an approximation of the marginal posterior distribution of the relative risks given the observed data $g(\boldsymbol{\theta}|\mathbf{O})$. Empirical Bayes methods use estimates of the hyper-parameters and typically these are maximum likelihood estimates derived from the marginal likelihood of $\boldsymbol{\gamma}$,

$$g(\mathbf{O}|\,\boldsymbol{\gamma}) = \int g(\mathbf{O}\,|\,\boldsymbol{\theta})\, f(\boldsymbol{\theta}\,|\,\boldsymbol{\gamma}) d\boldsymbol{\theta}\,. \tag{4.6}$$

If areas are independent then the marginal posterior distribution is also independent and can be written as

$$g(\theta_i|O_i,\boldsymbol{\gamma}) \propto g(O_i\,|\,\theta_i)\, f(\theta_i\,|\boldsymbol{\gamma}). \tag{4.7}$$

A point estimate of the relative risks is given by a measure of location of this distribution, typically the posterior mean $E(\boldsymbol{\theta}|\mathbf{O})$ or the posterior median. However, direct evaluation of these parameters through analytical or numerical integration is not generally possible.

Another measure of location of this posterior is the posterior mode or maximum a posteriori (MAP) estimate that maximises $g(\boldsymbol{\theta}|\mathbf{O},\boldsymbol{\gamma})$. MAP estimation can be performed using penalised likelihood maximisation (84) and has been applied to disease mapping (81, 85).

Standard Bayesian analysis, considering a completely specified prior distribution $f(\boldsymbol{\theta}|\,\boldsymbol{\gamma})$ with known hyper-parameters $\boldsymbol{\gamma}$, is seldom used in practice. The empirical Bayes (EB) approach assumes that hyper-parameters are unknown and are drawn from an unspecified distribution. The fully Bayesian formulation comprises a three-stage hierarchical model in which the hyper-prior distribution ($\boldsymbol{\gamma}$) is specified.

4.2.2 Empirical Bayes

The empirical Bayes method was the first approach to disease mapping that attempted to overcome some of the problems encountered when examining SMRs and simple SMR models. Marshall (86) provides a good thorough review of various methods for statistical analysis of patterns of disease, including a review of spatial empirical Bayes methods. The early approaches that have been used to map geographical distributions of specific diseases have been shown to be

unsatisfactory at clearly identifying extreme rates, especially for rare events or for small populations. To overcome such problems, empirical Bayes estimation was developed in the area. Efron and Morris (87) appear to be the first to have used such methods to pool information across areas, reducing the total mean-square error. These methods were further developed by Tsutakawa (88) and Clayton and Kaldor (76) by taking into account the variation in the estimation precision across the map. The stability of rare estimates is increased by combining the Poisson variation in each area with a global model of the rates. Tsutakawa (88) initially derived improved estimates of mortality rates using an EB approach that treats true rates as samples from an unknown prior distribution that needs estimation. A normal distribution for the logit of the probability of disease in each geographical area was used in this early exploratory analysis. This approach is similar to Leonard's method for estimating binomial proportions (89).

As previously shown, our likelihood function derives from the Poisson distributed number of observed cases occurring within the geographical areas, and classical statistical analyses of relative risks are representative of the data. Also, we stated that we know the distribution of relative risks conditional on the distribution within each area. Therefore, the prior beliefs to be incorporated into Bayes theorem is the information we have on the relative risks eg that small populations are more likely to have extreme relative risks and the probability that relative risks obtained from larger populations are more reliable (76). The posterior distribution of risks can then be calculated using prior beliefs. Empirical Bayes methods always seek to approximate the posterior distribution. Any other method used to calculate the posterior distribution would be classified as full Bayes (90).

4.2.2.1 Poisson-Gamma Model

Along with incorporating an estimate of reliability through prior beliefs, a prior distribution of the overall relative risks can be specified, reflecting the distribution of relative risks between areas. The Gamma distribution was originally suggested by Clayton and Kaldor (76) for this purpose and has been further used and

developed by others (91-96). They considered a Gamma distribution for the relative risks described by the hyper-parameters, α and v, where α is a scale parameter and v is a shape parameter.

$$f(\theta_i \mid \gamma) = \frac{\alpha^{v} \theta_i^{v-1} e^{-\alpha\theta_i}}{\Gamma(v)}. \tag{4.8}$$

This choice of model involves describing the probability of disease mortality or incidence occurring within given areas by the Poisson distribution, and the Gamma distribution describes the denominator populations required for the observed cases to occur (94). Therefore, as shown by Clayton and Kaldor (76), conditioning on the true relative risk (θ_i), the number of deaths (or incidence counts) (O_i) in the ith geographical unit follows a Poisson model where E_i is the expected number of deaths.

The relative risks are assumed to follow a Gamma distribution with the marginal distribution of the O_i being negative binomial. The shape and scale parameter of the negative binomial model are estimated by maximum likelihood. It follows that the empirical Bayes estimate of the posterior expectation takes the form:

$$E(\theta_i \mid O_i, \hat{\gamma}) = \frac{O_i + \hat{v}}{E_i + \hat{\alpha}},$$

where the distribution of θ_i conditional on the observed count is also Gamma. This is a compromise between the observed SMR (O_i/E_i) and the general mean ($\hat{v}/\hat{\alpha}$). Since O_i have a negative binomial distribution, the unconditional expectation of the O_i is given by

$$E(O_i) = E_i \frac{v}{\alpha}.$$

In the model suggested by Clayton and Kalder (76) the adjustment for age has been achieved internally from overall age-specific rates over all the areas. Manton et al (97) suggested that age adjustment can be accomplished from a set of external age-specific standard rates.

Manton also suggested that this modelling can account for covariates by making the prior mean a function of the covariates. Clayton and Kaldor (76) suggest extending the distribution of θ_i to allow for covariates, $\mathbf{z}_i$. They give an example suggesting that the estimate for an area with relatively few observed or expected cases should not be drawn towards the overall mean relative mortality, but towards an estimated value consistent with the area's level of the given covariate. Clayton and Bernardinelli (84) differentiate between covariates at an area-level and those measured on individuals. Area-level or ecological covariates, z_i, can be included by modelling the logarithm of the relative risks as a linear function of these covariates. This is effectively a separate scale parameter for the prior $f(\theta_i \mid \alpha_i, \nu_i)$:

$$E[\log(\theta_i)] = \log\left(\frac{\nu}{\alpha_i}\right) = \mathbf{z}_i^{\,T}\boldsymbol{\beta},$$

then the prior in equation (4.8) can be written as

$$f(\theta_i \mid \alpha_i, \nu) = \frac{\alpha_i^{\,\nu}\theta_i^{\nu-1}e^{-\alpha_i\theta_i}}{\Gamma(\nu)}.$$

Marshall (98) extended the Gamma-Poisson model and proposed a non-iterative-distribution free approach using weighted moments to estimate a prior mean and variance. Marshall points out the difficulties arising in iterative methods of estimation and non-iterative ANOVA style estimators for the prior mean and variance. Tsutakawa (91) used the Poisson likelihood and Gamma framework to estimate relative risks for geographic regions with an additional random effects component.

4.2.2.2 Log-Normal Model

Various other forms of prior distribution have been considered. Clayton and Kaldor (76) also proposed that the rates follow a multivariate log-normal distribution. Here the normal prior hyper-parameters μ and σ_u^2 were specified for the logarithm of the relative risks:

$$f(\log(\theta_i) \mid \gamma) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{1}{2} \frac{[\log(\theta_i) - \mu]^2}{\sigma_u^2} \right). \qquad (4.9)$$

The posterior distribution for the relative risks is no longer tractable under such a model and parameters can be estimated using the EM algorithm.

Tsutakawa (91) proposed a log-normal model which is an alternative to standardising the data:

$$O_{ik} \sim Poisson(\pi_{ik} E_{ik}),$$

where the observed data in stratum k and area i are assumed to follow a Poisson distribution. Then the area and stratum specific relative risks π_{ik} were modelled in terms of area-specific relative risks θ_i and stratum-specific risks ϕ_k such that

$$\log(\pi_{ik}) = \log(\phi_k) + \log(\theta_i) + u_{ik}.$$

Here, the residuals u_{ik} are distributed with mean 0 and variance σ_u^2. This has the structure of a classic mixed effects model with fixed effects ϕ_k and random effects θ_i. The θ_i are assumed to come from a one parameter inverted Gamma-1 distribution with hyper-parameter α:

$$f(\theta_i \mid \alpha) = \frac{\alpha^\alpha}{\Gamma(\alpha)} \frac{e^{-\alpha/\theta_i}}{\theta_i^\alpha}.$$

Maiti (99) develops this log-normal EB method into a hierarchical Bayes estimation procedure for estimating mortality rates for disease maps. Bernardinelli and Montomoli (85) compared parametric (normal prior) EB estimates with Bayes estimates. Penalised log-likelihood maximisation was used for the EB estimation of the relative risks.

4.2.2.3 Non-Parametric Model

Clayton and Kaldor (76) suggested estimating a prior distribution non-parametrically using a method suggested by Laird (100). This assumes that θ_i are iid random variables with density, $f(\theta)$, of unspecified parametric form. $f(\theta)$ can be estimated nonparametrically using an EM algorithm. This method has the advantage of imposing few constraints. Heisterkamp et al (93) developed the non-parametric approach further. The prior distribution is replaced by a number of support points α_h such that

$$\theta_i = f_\theta(\alpha_1,...,\alpha_H),$$

and the probability of each value can be written as

$$\Pr[\theta = \alpha_h] = \pi_h .$$

The actual number of support points, H, gives an indication of the number of different values for the true SMR in the data, whereby H = 1 would suggest that all areas have the same true SMR up to a maximum of H = n unique SMRs. The conditional likelihood for a mixture of Poisson distributions (80) can be written as

$$l(\theta_i \mid \alpha_h, h = 1,...,H) = \sum_{h=1}^{H} \frac{\pi_h e^{-\alpha_h E_i} (\alpha_h E_i)^{O_i}}{O_i!}.$$

The posterior empirical Bayes estimates are a weighted average of the estimated support points $\hat{\alpha}_h$:

$$\mathrm{E}(\theta_i \mid O_i, E_i) = \frac{\sum_{h=1}^{H} \hat{\alpha}_h \hat{\pi}_h e^{-\hat{\alpha}_h E_i} (\hat{\alpha}_h E_i)^{O_i}}{\sum_{h=1}^{H} \hat{\pi}_h e^{-\hat{\alpha}_h E_i} (\hat{\alpha}_h E_i)^{O_i}}.$$

Schlattman (101) suggests using mixture models to identify population heterogeneity and map construction within an empirical Bayes framework. This assumes that O_i comes from a nonparametric mixture density. Estimation is done using a maximum likelihood approach.

4.2.2.4 Multilevel Models

Multilevel models can also be used to provide empirical Bayes estimates for disease mapping (6, 102-104). The simplest Poisson multilevel model fits one "level" to model the Poisson variation and the extra Poisson variation is fitted at the higher level. Referring to any covariates as $\mathbf{z}_i$, the logarithm of the relative risks can be written as

$$\log(\theta_i) = z_i^T \boldsymbol{\beta} + u_i \,, \tag{4.10}$$

with an assumed normal distribution for the residuals u_i:

$$u_i \sim N(0, \sigma_u^2). \tag{4.11}$$

Changing the prior for $\log(\theta_i)$ in (4.9) to one dependent on an area specific mean u_i gives

$$f(\log(\theta_i) \mid \mu_i, \sigma_u^2) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(-\frac{1}{2} \frac{[\log(\theta_i) - \mu_i]^2}{\sigma_u^2} \right), \tag{4.12}$$

where

$$\mu_i = \mathbf{z}_i^T \boldsymbol{\beta}. \tag{4.13}$$

This model can be extended to include higher levels of geographical aggregation (105). To fit a model with two levels, region i nested within country j, say, equation (4.1) becomes

$$O_{ij} \sim Poisson(\theta_{ij}E_{ij}), \tag{4.14}$$

then equation (4.9) can be written as

$$\log(\theta_{ij}) = \mathbf{z}_{ij}^{T}\beta + u_{ij} + v_j. \tag{4.15}$$

Here the logarithm of the relative risks for region i is modelled in terms of residuals at the levels of region and country. Each of these random effects are normally distributed:

$$u_{ij} \sim N(0,\sigma_u^2) \text{ and } v_j \sim N(0,\sigma_v^2). \tag{4.16}$$

There is now non-independence in the relative risks since relative risks for two regions within the same country share the same country level random effect v_j. The prior in equation (4.12) can now be written as a joint prior for n_j regions in country j:

$$f(\{\log(\theta_{ij})\} \mid \mu_j, \Sigma_j) =$$
$$(2\pi)^{-n_j/2}\left|\Sigma_j\right|^{-1/2}\exp\left(-\frac{1}{2}\left[\{\log(\theta_{ij})\}-\mu_j\right]^{T}\Sigma_j^{-1}\left[\{\log(\theta_{ij})\}-\mu_j\right]\right). \tag{4.17}$$

The dispersion matrix for country j, $\mathbf{\Sigma}_j$, has dimensions $n_j \times n_j$ with elements $\sigma_u^2 + \sigma_v^2$ on the diagonal and σ_v^2 off-diagonal.

Further levels can be added and is straightforward. Langford et al (106) showed that the model described in equation (4.14) and (4.15) can be extended to include random covariates. Langford et al (6) described how to include spatial correlation between the relative risks in the multilevel model. This will be discussed further in Chapters 5 and 6.

4.2.2.5 Spatial Models

Ignoring spatial configuration of areas is often unjustifiable as areas close to one another in geographical space often share similar environmental, social or demographic factors which influence disease rates. An EB estimate which incorporates a spatial component can be thought of as a weighted average between the SMR, a local mean rate, and the global mean rate.

Models of exchangeability (76) have been discussed so far, where area-specific estimates are more or less displaced depending on the mean value. This displacement depends on the intrinsic stability of the estimates and not the areas' locations on the map (85). However, often the relative risk estimates are strongly influenced by the estimates of geographical areas, and only indirectly by the estimates from the rest of the map. Incorporating the geographical structure of the map results in a more complex prior model. This sets a conditional independence structure on the relative risks whereby each relative risk is conditionally independent of all other relative risks, given a small set of geographically adjacent areas. Most of the attention in the literature in this area has focused on this type of method, modelling spatial dependence of the prior θ_i eg by modelling the mean of θ_i conditional on the θ_i's of its neighbours.

Another, less popular, approach for accounting for spatial location space is to make the prior on θ position dependent whereby a trend is placed on the prior mean $E(\theta)$. Marshall (98) proposed a method for empirical estimation of $E(\theta_i)$ for each i. It is assumed that the $E(\theta_i)$ do not vary greatly within neighbourhoods. An estimator shrunk towards a weighted neighbourhood average is obtained via a non-iterative distribution-free approach by a method of moments. The resulting degree of shrinkage depends on the local variability of the neighbourhood rates. Any of the EB methods results in estimates of relative risks being more displaced towards a local mean rather than a global value.

Various different prior distributions have been developed for the alternative, more commonly used approach. A prior distribution that naturally allows for the possibility of spatial dependence between rates is the multivariate log-normal

prior. This was developed by Clayton and Kaldor (76) and has been a popular technique for mapping disease (96, 99, 107). The log-normal model was described previously and a spatial extension to this is to suppose that the log relative risks are correlated, where the correlation is dependent on geographical proximity. The dispersion matrix, Σ, may be expressed as a function of a small number of parameters. Clayton and Kaldor (76) used the conditional autoregression (CAR) procedure to model the mean conditional on its neighbours. This model has been further developed (84, 108-110) such that the logarithm of the relative risks is modelled as

$$\log(\theta_i) = u_i + v_i \tag{4.18}$$

where the v_i are unstructured heterogeneous effects such that

$$v_i \sim N(0, \sigma_v^2) \tag{4.19}$$

and the u_i are spatially structured effects through an intrinsic Gaussian autoregression

$$u_i | u_{i'}, i' \neq i \sim N(\bar{u}_i, \sigma_{u_i}^2) \tag{4.20}$$

where $\bar{u}_i$ is the mean of areas bordering area i:

$$\bar{u}_i = \frac{\sum_{i'} w_{ii'} u_{i'}}{\sum_{i'} w_{ii'}}, \tag{4.21}$$

$$\sigma_{ui}^2 = \frac{\sigma_u^2}{\sum_{i'} w_{ii'}}, \tag{4.22}$$

and **W** is the adjacency matrix of the map, defined by

$$w_{ii'} = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are contiguous} \\ 0 & \text{otherwise.} \end{cases}$$

The v_i and u_i are assumed to be independent and if the u_i dominate then the relative risks show spatial structure.

Yasui (96) develops this model slightly using a Gaussian spatial autoregression process which assumes that $\log(\theta_1),\ldots,\log(\theta_i)$ follows a multivariate Gaussian distribution with conditional moments. Yasui then goes on to evaluate empirically the various priors used in the EB estimation of small area disease risk by comparing mean squared errors and weighted mean squared errors.

Mollié and Richardson (109) consider both the CAR and a simultaneous autoregression (SAR) prior model to attempt to smooth cancer mortality rates and discuss the differences between the models. The covariance structure of a SAR model implies that autocorrelations at a larger distance than for a CAR model are taken into consideration, without estimation of extra parameters. Besag (111) showed that a SAR model can be written as a CAR model of higher order.

Bernardinelli and Montomoli's (85) basic model is similar to that proposed by Clayton and Kaldor (76), but they go on to use a penalised log-likelihood maximisation for the EB estimation of relative risks. θ_i is modelled as the sum of the global mean, denoted by μ, that expresses the overall level of the log-relative risks throughout the map, and an area specific effect, denoted by ϕ, representing the area risk effects, that is the difference between the log-relative risk for an area i and the global mean. For the ith area $\log(O_i) = \eta_i = \mu + \phi_i$ and O_i follows the Poisson distribution given by

$$[O_i \mid E_i, \mu, \phi_i] \sim \text{Poisson}(O_i; E_i \exp(\mu + \phi_i)].$$

The aim is to combine the 'prior' belief about the risks, which is embodied in the prior $[\phi|\lambda]$, $[\lambda]$ and $[\mu]$, with the 'new' information contained in the data {Y, E}. Here $[\phi|\lambda]$ denotes the area effects and is viewed as a spatial process as it expresses the prior beliefs concerning the collection of unknown area effects; $[\lambda]$ denotes the prior for an unknown parameter which represents the geographical variability and which controls the amount of variation in risk distribution throughout the map; and $[\mu]$ denotes the prior for the global mean. The EB estimates of ϕ and μ are obtained by maximising the posterior distribution $[\phi, \mu \,|\, O, E, \hat{\lambda}]$, which is equivalent to

$$l^*(\phi, \mu) = \log[Y \,|\, E, \phi, \mu] + \log[\phi \,|\, \hat{\lambda}]$$

whereby $\hat{\lambda}$ is a suitable estimate of the parameter λ. The first term on the right hand side for fixed Y and E is the log-likelihood of the data while the second term can be interpreted as a penalty function that penalises departure of ϕ from the prior model. Therefore, the above function can be interpreted as the penalised log-likelihood. If we adopt the CAR prior model this becomes

$$l^*(\phi, \mu) = l(\phi, \mu) - \frac{\hat{\lambda} n_i}{2} \sum_{i=1}^{N} \phi_i (\phi_i - \bar{\phi}_i).$$

The log-area effects that differ greatly from their respective neighbours receive high penalties. The penalties cause these estimates to shrink towards a local mean. The quantity $\hat{\lambda}$ acts as a smoothing parameter and when it is large the penalty function receives more weight leading to smoother estimates of the effects $\{\phi\}$. When this parameter is equal to 0, the penalty is given no weight, which leads to the ML estimates of the effects.

4.2.3 Fully Bayesian

The fully Bayesian approach incorporates the distribution of the hyper-parameters $f_\gamma(\gamma)$ into the modelling. The joint posterior distribution of the relative risks θ and the hyper-parameters γ given the data **O** is

$$g(\theta,\gamma \mid \mathbf{O}) \propto g(\mathbf{O}|\theta) f_\theta(\theta|\gamma) f_\gamma(\gamma). \tag{4.23}$$

The marginal distribution for θ given the data is then obtained by integrating out the hyper-parameters:

$$g(\theta|\ \mathbf{O}) = \int g(\theta,\gamma \mid \mathbf{O}) d\gamma\,. \tag{4.24}$$

Fully Bayesian procedures, as with empirical Bayes, provide methods of variance reduction through the borrowing of information. The hierarchical structure leads to Bayes point estimates that are shrunk towards a value that is related to the distribution of all the parameters in the hierarchical structure. It is assumed that the prior structure is close to the 'true model', and consequently a different choice of priors will lead to different shrinkage.

Various fully Bayesian disease mapping models are compared using goodness of fit criteria, by Lawson et al (112) and a comprehensive review of the main classes of spatial priors that have been proposed for fitting fully Bayesian disease mapping models is given by Best at al (113). A fully Bayesian approach that is commonly used (114, 115) as a model for disease mapping aggregated count data is a three-level hierarchical model. Again, E_i are the expected number of deaths, β_i represents the log relative risks ($\log(\theta_i)$) and O_i represents the observed number of deaths in area *i*. The O_i follow a Poisson distribution:

$$O_i \sim Poisson(\theta_i E_i)\ \ , i = 1,\ldots,n$$

$$\beta_i \sim p(.|\lambda) \tag{4.25}$$

$$\lambda \sim \pi(). \tag{4.26}$$

where $p(\cdot \mid \lambda)$ is an appropriate second-stage prior distribution for β_i and λ are the hyper-parameters of this second stage model with hyper-distributions $\pi()$.

4.2.3.1 Multivariate Normal Priors

Correlated random variables are often represented by the multivariate normal distribution. The dependence structure can be structured in terms of an $n \times n$ covariance matrix $\boldsymbol{\Sigma}$, and the second-stage prior can be written as

$$\boldsymbol{\beta} \sim MVN(\mu, \Sigma) \tag{4.27}$$

where $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_n\} = \{\log(\theta_1), \ldots, \log(\theta_n)\}$ and represents the vector of area specific random effects from equation (4.24). $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Omega}$ and Ω_{ij} is the correlation between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_j$ (116). The elements of the correlation matrix are chosen to be a function of the relationship between the areas. This should be chosen so that the covariance matrix $\boldsymbol{\Sigma}$ remains positive definite.

An alternative approach that follows on from this was proposed by Kesall and Wakefield (117) and is based on specifying a Gaussian random field for the underlying distribution of the log relative risk at the second stage of the hierarchical model. Integrating this model over i areas gives a multivariate normal model for $\log(\theta_i)$.

Mollié (79) uses Gaussian Markov random fields (similar to CAR model) for mapping SMRs, which is a second stage model that was originally proposed by Besag et al (110), and discusses a compromise between a spatially structured prior and an unstructured prior (108). A convolution Gaussian prior is an intermediate distribution in the log-relative risks that ranges from prior independence to prior local dependence. The log relative risks in the prior are the sum of v, a normal variable with zero mean and variance λ^2, and u, an intrinsic Gaussian autoregression with conditional variances proportion to κ^2. Again, v

describes the unstructured heterogeneity of the relative risks and u represents local spatially structured variation, so $\log(\theta_i) = u + v$. The conditional variance is

$$\text{var}[\beta_i \mid \beta_j, j \neq i, \gamma] = \text{var}[\beta_i \mid \beta_j, j \in \partial_i, \kappa, \lambda] = \frac{\kappa^2}{w_{i+}} + \lambda^2,$$

where ∂_i denotes the set of areas adjacent to area i, the hyper-parameter $\lambda = \sigma$, and κ^2 corresponds to a total independence of the risks whereas $\lambda^2 = 0$ leads to purely local dependence modelled by the intrinsic Gaussian autoregression. A small κ^2/λ^2 reflects unstructured heterogeneity whereas a large κ^2/λ^2 indicates that a spatially structured variation dominates.

For the multivariate normal priors for the log relative risks, a more general class of hyper-priors for the inverse variance is the conjugate Gamma distribution with specified parameters (79). Vague Gamma hyper-priors are often assumed for u and v when there is a lack of information about the importance of each component. It has been suggested (85) that it is reasonable to choose vague gamma priors with means

$$\frac{2}{\text{var}(\log(\theta_i)} \quad \text{for} \quad \lambda^{-2}$$

and

$$\frac{2}{\overline{w}\,\text{var}((\log(\theta_i))} \quad \text{for} \quad \kappa^{-2}$$

where $\overline{w}$ is the mean of w_{i+} , $w_{i+} = \sum_j w_{ij}$ and w_{ij} is the ijth element of a symmetric $n \times n$ weight matrix **W** similar to that described for equations (4.21) and (4.22).

The Gaussian Markov random fields are the most commonly used second stage model. This model, which was further developed by Besag et al (110), can be written in general terms to coordinate with equations (4.25) and (4.26):

$$\beta_i = V_i + U_i,$$

$$V_i \sim N(0, \sigma_v^2), \tag{4.28}$$

$$U_i | \boldsymbol{u}_{(-i)} \sim N\left(\frac{\sum_j w_{ij} u_j}{w_{i+}}, \frac{\sigma_u^2}{w_{i+}}\right). \tag{4.29}$$

By setting the autocorrelation parameter, which reflects the overall strength of spatial dependence between locations, to its upper limiting value of 1 means the $\{U_i\}$ follow an intrinsic autoregression. As previously discussed, this can be described as the spatial component of between-area variation of disease risk and the $\{V_i\}$ represent the geographically unstructured heterogeneity effect. Posterior inference about the amount of spatial dependence that exists is based on the marginal variance of the U_i's; this can be estimated using MCMC methods.

Maiti (99) develops the CAR model for log-relative risks and assumes a hierarchical model. Non-informative priors for hyperparameters are used. This method was developed because unlike the EB method, the hierarchical Bayes method accounts for the uncertainty involved in the estimation of mean and variance of prior parameters by assigning the distributions of prior parameters. In the case of these models, the approximation to the variance of the MLE of parameters of the prior distribution is intractable. Mollié and Richardson suggest that an indication of the precision of these estimates can be based on the conditional expected information for the prior of the log rates. This is calculated from the expectation of the log of the prior density conditional on the observations and the current values of the parameter estimates, given in the EM algorithm. The CAR model appears to be a popular choice of model for examining spatially distributed disease data (79, 95). The Gaussian Markov random fields model is equivalent to specifying a multivariate normal model for the joint distribution of the area specific random effects but with the dependence structure parameterised in terms of the precision matrix, $\boldsymbol{P}$, rather than the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{P}^{-1}$.

Another multivariate normal model was proposed by McNab (118) which also parameterises in terms of the precision matrix. This model assumes

$$\boldsymbol{\beta} \sim \text{MVN}(\mathbf{0}, \sigma^2 \boldsymbol{D}^{-1}),$$

where

$$\boldsymbol{D} = \rho \boldsymbol{P} + (1-\rho)\boldsymbol{I} \ .$$

As discussed, $\boldsymbol{P}$ represents the precision matrix and $\boldsymbol{I}$ is the $n \times n$ identity matrix. $\rho \in [0\ ,\ 1]$ and can be interpreted as the measure of spatial dependence whereby if $\rho = 0$, the model reduces to the Gaussian independence prior (4.28) or if $\rho = 1$ it reduces to the intrinsic autoregression (4.29).

4.2.3.2 Mixture Models

Lawson and Clarke (119) propose an extension to the Gaussian Markov random fields model that includes a mixture of Gaussian and non-Gaussian conditional autoregressive components. The non-Gaussian is median based and aims to pick up discrete jumps in the relative risk structure which should hopefully avoid over-smoothing the distribution of relative risks.

Some of the models described in this section run the risk of oversmoothing the distribution of relative risks. Mixture models have been proposed in an attempt to prevent this happening and are becoming more popular for disease mapping. They take a different approach, in that, instead of assuming that the structure of the map can be described with a global structure, they assume that the map consists of a number of components and the aim is to identify these components. Schlattmann and Bohning (101) classify each area of the map as belonging to one component but other models, which are not so specific, are available.

Green and Richardson (120) proposed a mixture model where the allocation of β_i's for each area to a risk category follows a correlated process. They extend hidden Markov models to the spatial domain. It makes use of Potts model, which is frequently used in image processing, and involves fitting an interaction

parameter that controls the degree of spatial dependence and the number of 'states' of the mixture is estimated as part of the model. This model can be written as

$$\exp(\beta_i) = \tau_{a_i} ,$$

where a_i is the allocation variable and $a_i \in \{1,...,c\}$; $c \sim$ Unif(1, c_{max}) and represents the number of components in the mixture. Then,

$$p(a|\psi, c) = e^{\psi U(c) - \delta_c(\psi)}$$

$$\tau_j \sim \text{Gamma}(\alpha, \beta) , \quad j = 1,...,c .$$

The interaction parameter ψ is to be estimated and $U(z)$ is the number of pairs of neighbouring areas. An area i is allocated to component j based on ψ; larger values of ψ suggests area i has more neighbours in component j and will therefore be favoured. This model uses a reversible jump algorithm because the number of components in the mixture, c, is uncertain. Fernandez et al (121) also suggested mixture models for spatially distributed data. Transformations of autoregressive Gaussian processes are proposed for the spatially dependent weights and reversible jump Markov Chain Monte Carlo algorithms for posterior inference are developed.

4.3 Use of Disease Mapping Models

In this chapter, the techniques available for fitting empirical Bayes and fully Bayesian models have been discussed. In the following chapters, iterative generalised least squares procedures and the linearising approximation, penalised quasilikelihood, will be described as a means of obtaining empirical Bayes estimates of the spatial multilevel model. Other methods, such as the EM algorithm and Fisher scoring can also be used to fit the models described in section 4.2.2. Estimation procedures for fitting the fully Bayesian models will also

be discussed. Markov Chain Monte Carlo (MCMC) procedures will be described in terms of fitting spatial multilevel models. However, this technique can be applied to all of the fully Bayesian models described in this chapter.

Chapter 5

5 Spatial Multilevel Modelling

5.1 Introduction to Multilevel Modelling

Multilevel or hierarchically structured data are found in many areas of application and it is important that these structures are taken into account. An example which provides a clear case of such a structure is in education whereby we have a response from pupils who are assigned to level 1, these pupils are clustered within classes at level 2, which are clustered within schools at level 3, and these can be grouped within authorities or boards at level 4. Multilevel models take into account how this structure affects the measurements of interest and ignoring a hierarchical structure, if present, has consequences. For example, if analysing a school dataset similar to that described above, the relationship between two pupil level variables could be modelled using a simple linear regression. The variation between schools could be taken into account by incorporating separate terms for each school but this is very inefficient as it involves estimating many more coefficients than the multilevel procedure. It also does not treat schools as a random sample and therefore provides no information on the proportion of the variation that is attributable to differences between schools in the population. A further disadvantage of ignoring clustering is that it will generally cause standard errors of regression coefficients to be underestimated. Multilevel modelling also provides a useful tool to develop more complex models and this will be discussed in this thesis, mainly by looking at how to model geographically distributed data and explore the effects of incorporating the spatial structure of the data as a level.

For the purpose of this thesis, the background to multilevel modelling will be described from the view of modelling Poisson distributed data. Goldstein et al (122) give details of how to model normally distributed multilevel data.

5.2 Multilevel Modelling of Geographically Distributed Data

As previously discussed, problems associated with using conventional Poisson based methods to model geographically distributed data can be overcome by using multilevel modelling techniques based on iterative generalised least squares procedures (6). The data that will be used for our preliminary analysis and model exploration are geographically distributed in that we have counts of cancer deaths within regions within countries and they exemplify the type of data appropriate for a multilevel analysis.

5.2.1 European Cancer Mortality Data

To begin modelling, cancer mortality in 540 regions in 38 countries in the period 1991 will be used. The total population for the 38 countries was 690,308,858, ranging from 257,965 (Iceland) to 148,244,835 (Russian Federation). Population data are available at a sub-national level for 27 of the European countries. The mean population for a region was 1,295,139 with a range from 13,645 (Appenzzell-Inner Rhoden, Switzerland) to 17,429,759 (North Rhine-Westphalia, Germany). The causes of deaths were based on the Ninth Revision of the International Classification of Diseases (ICD-9) (123) and the causes of death considered here are malignant neoplasms (ICD-9 140-208). The total number of recorded deaths from cancer in these countries was 1,614,293, ranging from 446 (Iceland) to 290385 (Russian Federation). The mean number of deaths for a region was 2989, ranging from 29 (Appenzell-Inner Rhoden, Switzerland) to 49137 (North Rhine-Westphalia, Germany). If population or mortality data were not available for the timepoint 1991, data were used from the closest available timepoint. Further summaries of the data were presented in Chapter 3.

5.2.2 European Risk Factor Data

Since there is strong evidence of relationships between diet and other lifestyle factors and cancer mortality, it is important that we should adjust for these factors as necessary when examining spatial patterns of cancer mortality. We are also interested in quantifying what effect these factors have on mortality rates, and so need measures that reflect the population's diet and lifestyle.

The risk factor data were summarised in Table 3.7. Average consumption data for animal fats, alcoholic beverages, fruit and vegetables were obtained for the period 1991. These cover the dietary components that appear to be related to the most common cancer mortalities. The data are measured in kilograms per year per head of population for each of the European countries being examined. Annual average consumption of manufactured cigarettes per adult (+15) is provided at a country level for the time period 1990-1992 and should reflect the average level of smoking in each of the European countries. To take into consideration the differences in socio-economic status between countries, gross domestic product (GDP) is available and is measured in US dollars per inhabitant in 1995.

5.2.3 Variance Components Model

Firstly, we consider a population of regions, i, $i = 1,\ldots, 540$, with observed (O_i) and expected (E_i) counts of deaths. The expected number of deaths in the ith region, E_i, has been calculated based on the 1990 European age and sex specific cancer mortality rates and on the number of deaths in the population N_i. This standardisation is conducted for age and sex bands k and calculated as

$$E_{ik} = N_{ik} \frac{\sum O_{ik}}{\sum N_{ik}}, \tag{5.1}$$

$$E_i = \sum E_{ik} \,. \tag{5.2}$$

The relative risk of the mortality from the disease of interest is then

$$\theta_i = \frac{O_i}{E_i} \,. \tag{5.3}$$

A single level Poisson model including all six explanatory variables can be written as (105):

$$O_i \sim Poisson(\mu_i),$$

$$\log(\mu_i) = \log(E_i) + \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + u_i \tag{5.4}$$

$$[u_i] \sim N(0, \Omega_u), \quad \Omega_u = [\sigma_u^2]$$

where log (E_i) is treated as an offset and is included to account for the different populations at risk of death from cancer in each area. α is a constant and $x_{1i},\ldots,x_{6i}$ are the explanatory variables with coefficient $\beta_1,\ldots,\beta_6$, representing the mean (fixed) effects of the factors fruit consumption, vegetable consumption, animal fat consumption, alcohol consumption, cigarette consumption and gross domestic product respectively. We assume that the number of counts within each region follows a Poisson distribution. The u_i represent heterogeneity effects between areas (94) which can be viewed as having extra Poisson variation caused by the variation among underlying populations at risk in the regions considered.

A variance components model describes the random variation in the data by a set of variances. Looking at this model it can be seen that all of the variance is at a single level, i, which is region. There is only one random parameter, σ_u^2. Potentially, and more appropriately, this model can be expanded by partitioning the variance into that which is attributable to random variation between countries and that which arises due to differences between regions within countries, i.e.

adding a further level. Equation (5.4) can be extended to include a second level, country, indexed by *j*, and can be written as follows:

$$O_{ij} \sim Poisson(\mu_{ij}),$$

$$\log(\mu_{ij}) = \log(E_{ij}) + \alpha + \beta_1 x_{1ij} + \beta 2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + \beta_5 x_{5ij} + \beta_6 x_{6ij} + u_{ij} + v_j \qquad (5.5)$$

$$\lfloor v_j \rfloor \sim N(0, \Omega_v), \quad \Omega_v = \lfloor \sigma_v^2 \rfloor,$$

$$\lfloor u_{ij} \rfloor \sim N(0, \Omega_u), \quad \Omega_u = \lfloor \sigma_u^2 \rfloor.$$

The main difference between the models is that the variance has now been partitioned into σ_v^2, that which is attributable to differences between countries, and σ_u^2, that which is attributable to differences between regions.

5.2.4 European Cancer Mortality Results

The results of the variance components models are shown in Table 5.1. Model A represents the null model with no explanatory variables included, model B is the full single level model (5.4) and model C incorporates country as a level (5.5).

Firstly, looking at model B, the overall intercept, α, represents the logarithm of the average number of cancer deaths in all regions included in the study in addition to the (centred) logarithm of the expected cases, when all other fixed coefficients are zero. The estimates of the intercept are not particularly informative in these models as they reflect an unlikely situation whereby an area has zero exposure to any of the risk or protective factors. Before examining the other fixed parameter estimates it should be noted that for the purpose of interpretability the explanatory variables were re-based. The consumption variables were modelled in terms of 1000 kgs consumed per person per year, smoking was scaled to how many 1000 cigarettes were smoked per person per

Table 5.1 Results from variance components models: WHO European region

	Model A Estimate (95% CI)	**Model B** Estimate (95% CI)	**Model C** Estimate (95% CI)
Fixed Part			
α	7.38 (7.36 , 7.39)	6.99 (6.92 , 7.06)	7.03 (6.85 , 7.21)
β_1 *(FRU)*		-3.63 (-4.42 , -2.83)	-2.90 (-5.33 , -0.47)
β_2 *(VEG)*		0.99 (0.46 , 1.53)	0.36 (-0.99 , 1.71)
β_3 *(ANF)*		9.39 (6.18 , 12.60)	6.59 (-3.55 , 16.73)
β_4 *(ALC)*		1.39 (0.90 , 1.88)	1.99 (0.51 , 3.48)
β_5 *(SMO)*		0.05 (0.02 , 0.08)	0.09 (-0.01 , 0.18)
β_6 *(GDP)*		16.85 (12.85, 21.85)	5.91 (-1.86 , 13.68)
Random Part			
σ^2_u	0.049 (0.043 , 0.055)	0.027 (0.023 , 0.030)	0.013 (0.011 , 0.015)
σ^2_v			0.014 (0.006 , 0.022)

year, and GDP was examined as millions of dollars per inhabitant per year. The estimate of β_1 is the mean, or fixed slope for the explanatory variable fruit and it can be seen the estimate is negative and, judging significance by approximate 95 per cent confidence intervals, significant. This implies that when taking the other variables into account, an increase in fruit consumption decreases cancer mortality on average in the EU. The parameter estimate of -3.63 is a log relative risk of cancer mortality for each 1000 kg increase in fruit consumption per person per head. This suggests that every 10 kg increase in fruit consumption per person per year is associated with a decrease in the risk of cancer mortality of about 4% (RR=exp{-0.0363} = 0.964). It can be seen that all of the other variables appear to significantly affect cancer mortality; vegetable consumption surprisingly has a positive association with cancer mortality in Europe, animal fat consumption and

alcohol are also shown to be significant risk factors, smoking is shown to significantly heighten the risk of cancer mortality, and a higher GDP appears to increase the risk of cancer mortality.

The actual effect size of these variables will be discussed further on. There has been no partitioning of variance yet so all the variation here is due to differences between regions ($\sigma_u^2 = 0.027$). If we compare this value to the variance under model A ($\sigma_u^2 = 0.049$) it can be seen there has been a 45% reduction in variation after adding the explanatory variables. Due to the significant explanatory variables being at both region and country level, this shows that these factors are helping to explain the differences in cancer mortality between regions and countries. To show these differences visually, the estimated relative risks from models A and B have been mapped and are shown in figure 5.1 and 5.2.

There is definite clustering evident in Figure 5.1 and clusters of high cancer mortality rates can be identified in areas such as France, Denmark , parts of Russia, Hungary, UK and the Czech Republic. Cancer mortality rates in these regions are between 20% and 62% higher than expected if the European age and sex specific mortality rates had been applied to that area. Clusters of low mortality rates are most evident in regions Southern Europe, in particular Greece, Romania, Bulgaria and Ukraine. Overall, there is very high variability within and between many European countries. Figure 5.2, the map of the relative risks associated with living in these regions over and above the given risk factors, shows less variability; the relative risks now range from 0.59 to 1.51 (they ranged from 0.41 to 1.62 in Figure 5.1). As was evident from the random parameter estimates, the risk factors have, therefore, explained some of the variation in cancer mortality rates in the Europe, suggesting that even these initial models have been effective at smoothing geographic variation. However, the clustering within countries suggests it may be useful to take account of the higher geographical level.

Taking account of the fact that regions within a country are more alike than those from different countries, changes the parameter estimates somewhat (Table

5.1: Model C). The parameter estimates now appear more sensible with an increase in vegetable consumption and GDP not being associated with a significant increase in the risk of cancer mortality anymore. The variance has now been partitioned and it can be seen that a similar proportion can be attributed to σ_u^2, differences between regions, as to σ_v^2, differences between countries. From examining the map of relative risks from this model (Figure 5.3), it can be seen that there is much more clustering evident. There is a high amount of country level clustering as estimates for each region are drawn closer to the relative risks of the overall country. There are much more areas with very high (and very low) estimates which means the map is less useful for identifying regions as disease 'hotspots' (or areas of very low risk) but is useful in identifying countries at high risk as compared to other European countries.

Since the actual values of the fixed parameter estimates are not very informative, relative risks have been calculated which compare regions with high and low levels of exposure to each of the risk factors. The relative risks have been calculated for each of the variables from models B and C and are given in Table 5.2. Firstly, for the consumption of fruit as estimated by model C, the table shows that Greece has the highest level of consumption in Europe and Estonia has the lowest. The relative risk of 0.71 shows that a population consuming, on average, the same amount of fruit as Greece has a risk of cancer mortality that is 29% lower than if consumption was on the same level as Estonia. The other significant covariate from model C is alcohol consumption; consuming the same level of alcohol as Germany leads to a risk of cancer mortality that is 1.4 times as high as if consumption was on the same level as Armenia. For model B, it can be seen that all the variables are having significant effects on cancer mortality in Europe and vegetable consumption and GDP are having an opposite effect than would be expected. Populations across Europe lead very varied lifestyles and the relative risks, before accounting for the non-independence of regions within countries, are likely to be inaccurate. It makes more sense to concentrate on the estimates from the model taking account of some of the geographical structure in the data.

Table 5.2 Variable effect size of risk factors from model B and C: WHO European Region

Variables	Area with Highest Consumption	Area with Lowest Consumption	Relative Risk Model B	Relative Risk Model C
Fruit	Greece	Estonia	**0.65 (1.13 , 1.50)**	**0.71 (0.53 , 0.95)**
Vegetables	Greece	Israel	**1.30 (1.13 , 1.50)**	1.10 (0.77 , 1.58)
Animal Fat	Hungary	Armenia	**1.30 (1.19 , 1.42)**	1.20 (0.91 , 1.59)
Alcohol	Germany	Armenia	**1.26 (1.16 , 1.36)**	**1.39 (1.09 , 1.77)**
Cigarettes	Poland	Azerbaijan	**1.15 (1.06 , 1.24)**	1.28 (0.97 , 1.63)
GDP	Switzerland	Kyrgyzstan	**1.73 (1.52 , 2.04)**	1.21 (0.94 , 1.56)

To give a graphical alternative to the maps, a plot of the predicted relative risks from model C is given in Figure 5.4. A measure of longitude (east to west positioning) has been plotted against the predicted relative risks for each region in Europe. This is an alternative to mapping the results and prevents the loss of information that occurs when grouping mortality rates in ranges. A relative risk of 1 on the plot indicates a region where the number of cancer deaths is as expected. There is an indication of a negative slope in this plot, suggesting the farther west a region is, the higher the risk of cancer mortality. There is also the possibility of two separate slopes in the data (see lines imposed on plot), suggesting that analysing Europe as a whole is perhaps not the correct method. The lifestyles of populations across Europe are also very different as many eastern European countries are a lot less developed than Western Europe. It may make more sense to split the European data somehow before carrying out any further analysis.

Figure 5.1 Estimated Relative Risks from model A: all cancer mortality in the WHO European region

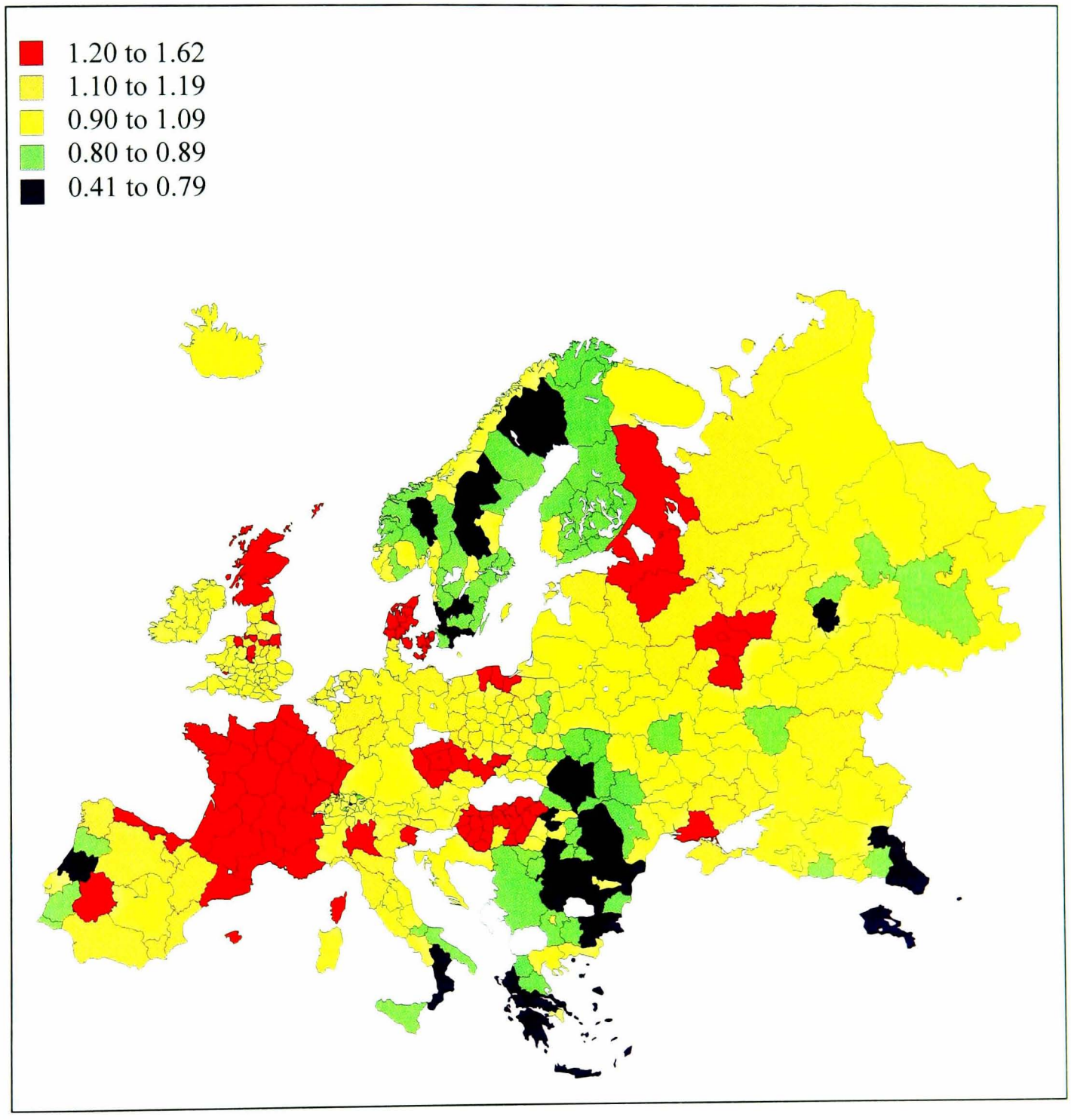


Figure 5.2 Estimated Relative Risks from model B: all cancer mortality in the WHO European region

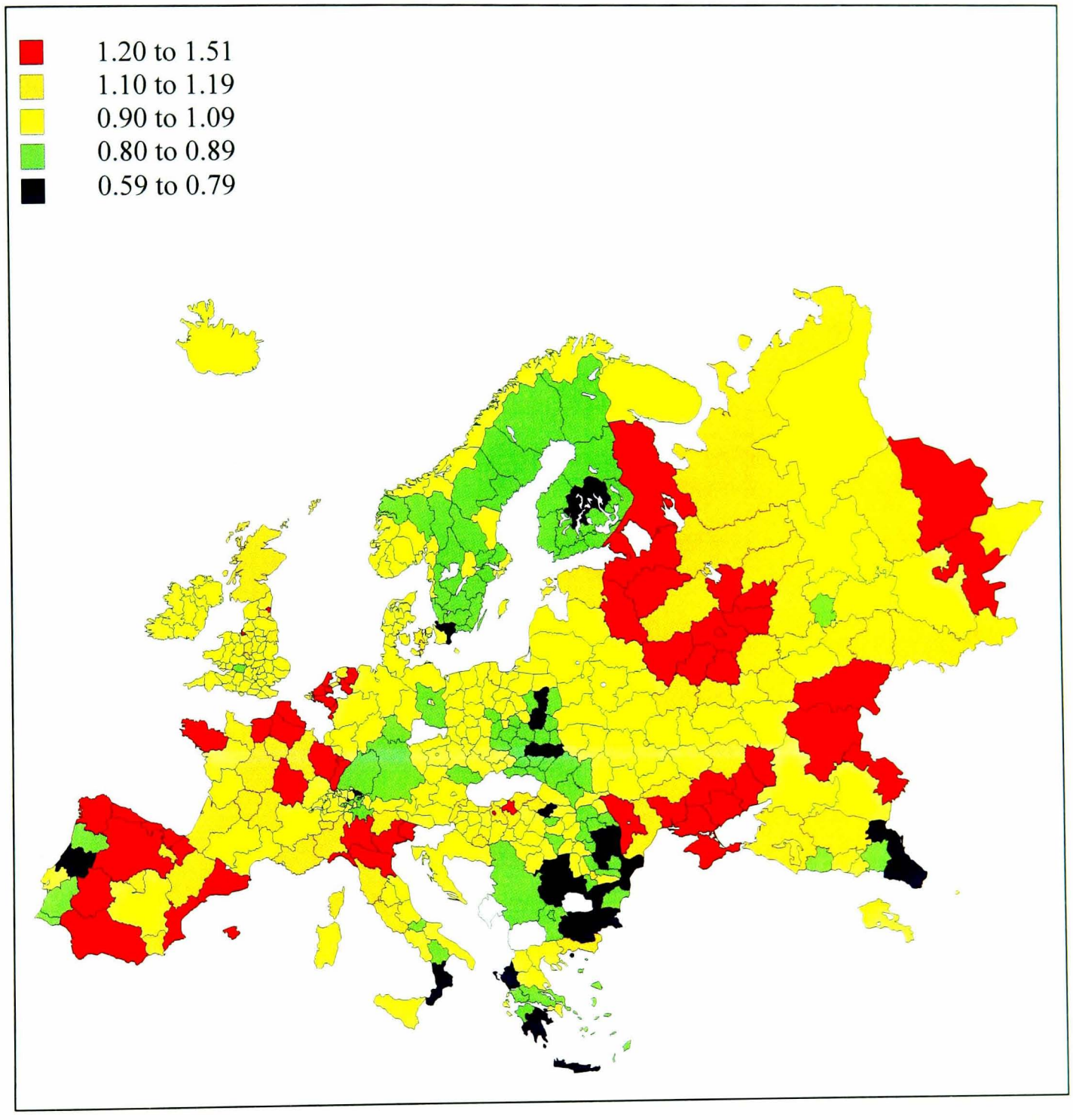


Figure 5.3 Estimated Relative Risks from model C: all cancer mortality in the WHO European region

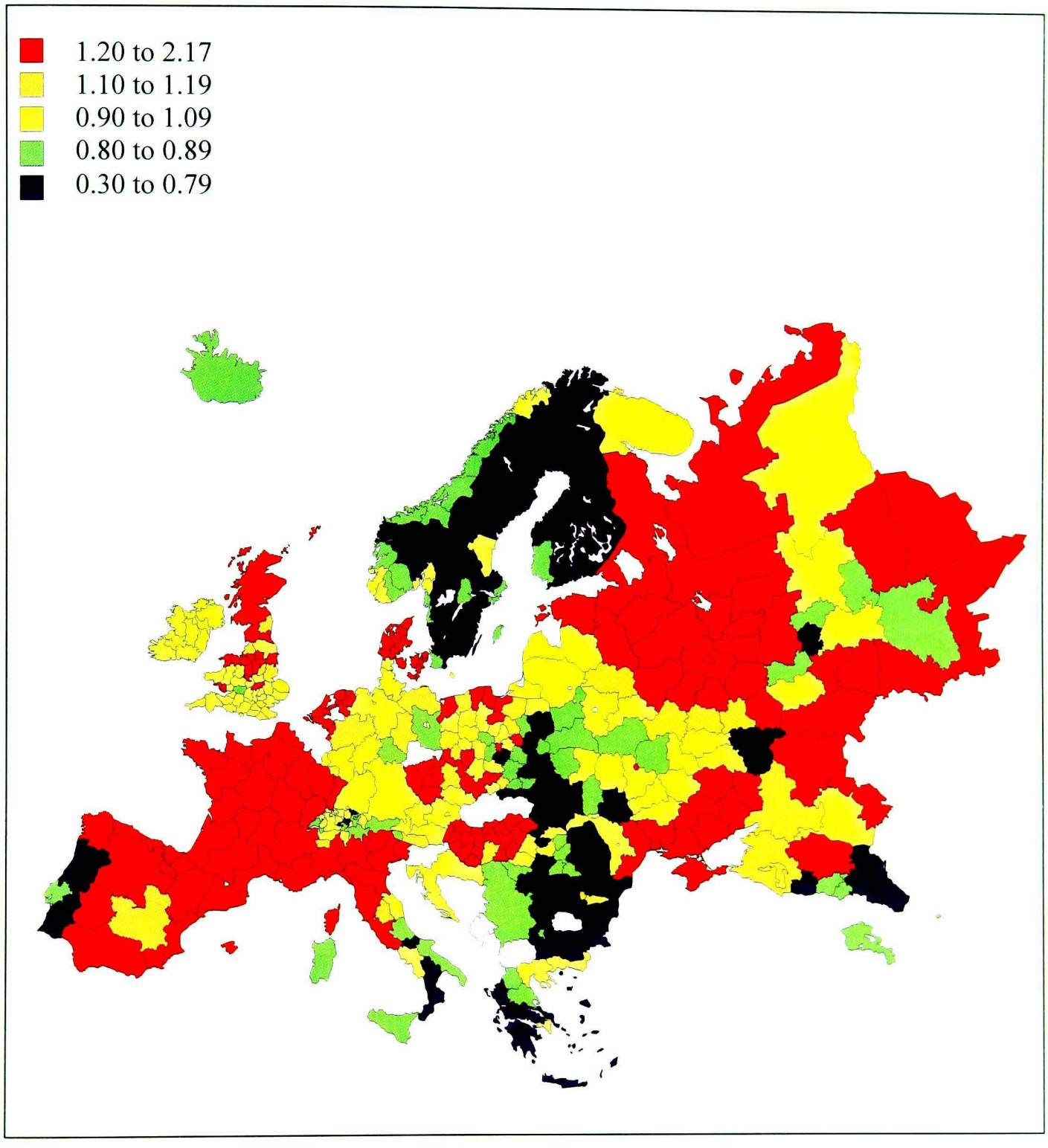
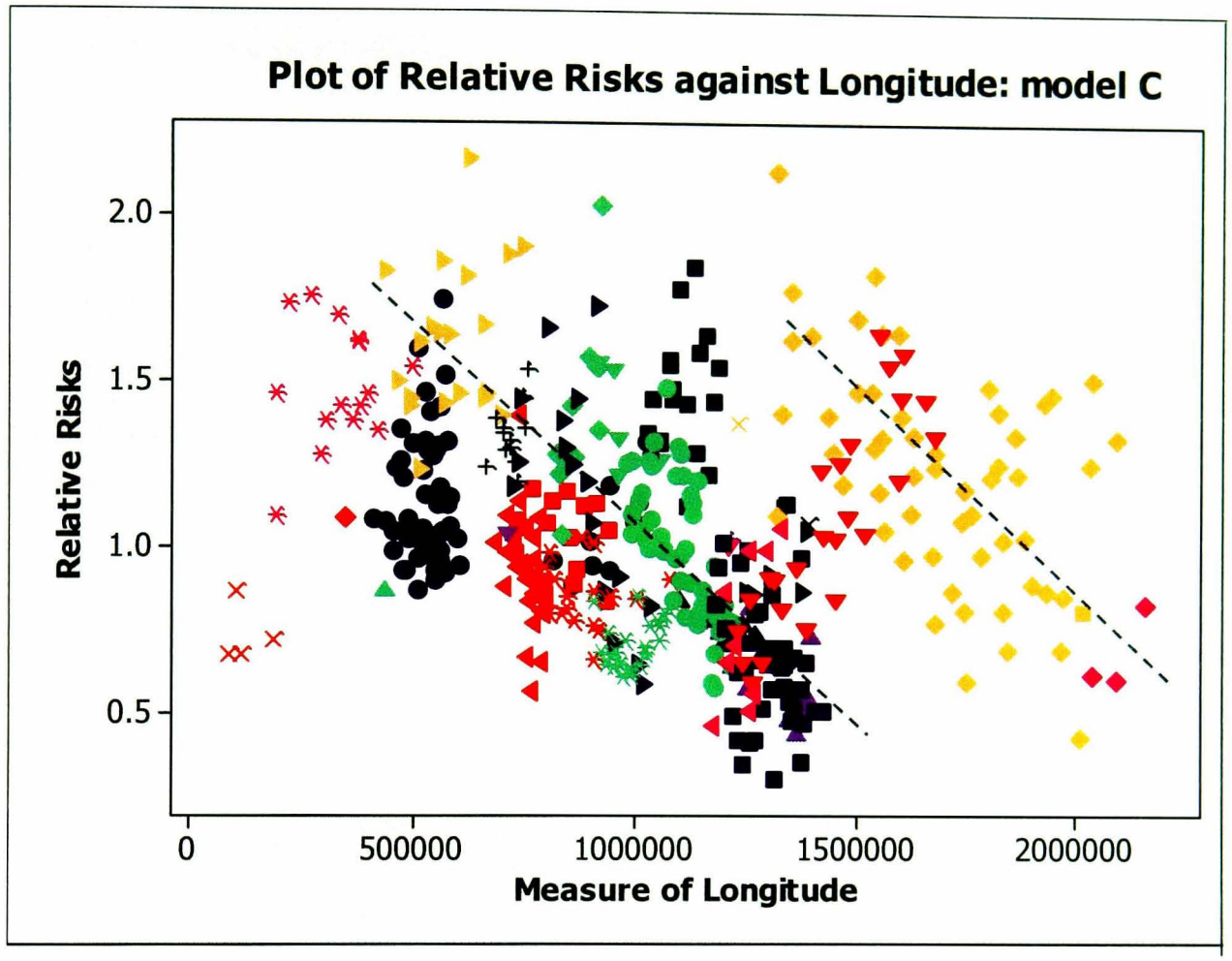


Figure 5.4 Plot of Relative Risks of cancer mortality against longitude: model C (WHO European region)



5.2.5 EU Cancer Mortality Data

For subsequent modelling, a subset of the full dataset will be used. Only the EU will be examined as all its countries are in Western Europe and thought to be fairly comparable in terms of development and lifestyle habits. All of the countries in the EU contain regional data, with the exception of Luxembourg, which makes the dataset more suitable for exploring spatial multilevel models. The dataset only contains areas which have both population and mortality data available for the time period 1991. Therefore, we have all cancer mortality in 187 regions in 11 EU countries. It should be noted that this dataset is based on the EU before the 10 new member states joined in 2004 (see Appendix A1.1) and that 4 of these 15 EU countries have missing data due to either the mortality or population data not being available for the time point being examined here. The total population for the 11 EU countries (Austria, Denmark, Finland, France,

Germany, Greece, Luxembourg, Netherlands, Portugal, Sweden and UK) was 257,075,105 with a mean of 23,370,464, ranging from 387,100 (Luxembourg) to 80,013,896 (Germany). The mean population for a region was 2,734,842 with a range from 24,734 (Ahvenanmaa in Finland) to 17,429,759 (North Rhine-Westphalia in Germany). Again malignant neoplasms (ICD9 140-208) are consdiered. In 1991 the total number of recorded deaths from cancer in these 11 countries was 654,126. The directly standardised mortality rates were calculated based on the standard European age and sex specific population. A summary of these is given in Table 5.3. The average standardised death rate for all countries is 297 per 100,000, ranging from 163 (Finland and Sweden) to 267 (Denmark). At the regional level standardised death rates range from 125 (Epirus in Greece) to 320 (Copenhagen and Frederiksberg city in Denmark).

Table 5.3 Standardised death rates in the EU

Country	Total deaths	Standardised Death Rate (per 100000)	If regional data available: Min (region)	Max (region)	No. of regions
Austria	19317	196	181 (Tirol)	224 (Burgenland)	9
Denmark	17764	267	228 (Sonderjylland)	320 (Copenhagen and Frederiksberg city)	15
Finland	9626	163	151 (Kuopio)	172 (Ahvenanmaa)	12
France	139310	212	182 (Midi-Pyrénées)	251 (Nord-Pas-de-Calais)	22
Germany	210537	206	187 (Brandenburg)	221 (Bremen)	16
Greece	19945	164	125 (Epirus)	192 (Macedonia East and Thrace)	13
Luxembourg	957	208	-	-	-
Netherlands	35645	209	197 (Friesland)	223 (Groningen)	12
Portugal	18230	165	150 (Centro)	189 (Azores)	7
Sweden	20406	163	148 (Kristianstad)	196 (Gävleborg)	24
UK	162389	218	186 (Gloucestershire)	264 (Tyne and Wear)	56

5.2.6 EU Risk Factor Data

We are using subsets of the risk factor datasets obtained for the whole of Europe and these are summarised in Table 5.4. The risk factors are the same as those examined for the whole of Europe except GDP, which is available at the regional level (NUTS II) within each country in the EU.

Table 5.4 Summarised EU risk factor data

Risk/protective factor	Median	Minimum	Maximum
Fruit kg/year/capita	100.8 (Sweden)	74.5 (UK)	142.6 (Greece)
Vegetables kg/year/capita	80.6 (Austria)	58.8 (Finland)	300.4 (Greece)
Animal Fat kg/year/capita	16.4 (France)	2.3 (Greece)	26.8 (Luxembourg)
Alcohol kg/year/capita	123.2 (UK)	60.0 (Greece)	173.9 (Germany)
Smoke cigarettes/year/adult	2120 (France)	1550 (Sweden)	3590 (Greece)
GDP (region level) ECU/inhabitant	17136 (Lorraine – France)	5611 (Ipeiros – Greece)	44711 (Copenhagen and Frederikberg city – Denmark)

5.2.7 EU Cancer Mortality Results

The results of the variance components models are shown in Table 5.5. Model A shows the null model with no explanatory variables included, model B is the results from the full model and model C is the full model with country added as a higher level.

Table 5.5 Results from variance components models: EU data

	Model A Estimate (95% CI)	**Model B** Estimate (95% CI)	**Model C** Estimate (95% CI)
Fixed Part			
β_o	7.53 (7.50 , 7.56)	6.83 (6.68 , 6.99)	6.87 (6.14 , 7.46)
β_1 *(FRU)*		-7.58 (-8.65 , -6.50)	-6.43 (-11.08 , -1.78)
β_2 *(VEG)*		-1.12 (-1.69 , -0.55)	-1.06 (-3.09 , 0.97)
β_3 *(ANF)*		26.36 (19.28 , 33.42)	22.58 (0.94 , 44.22)
β_4 *(ALC)*		0.64 (-0.17 , 1.45)	1.25 (-1.93 , 4.42)
β_5 *(SMO)*		0.49 (0.39 , 0.59)	0.44 (0.09 , 0.78)
β_6 *(GDP)*		0.20 (-4.61 , 5.01)	0.90 (-2.231 , 4.02)
Random Part			
σ^2_u	0.040 (0.032 , 0.048)	0.014 (0.011 , 0.017)	0.006 (0.004 , 0.007)
σ^2_v			0.019 (0.002 , 0.035)

The parameter estimates can be interpreted as before, therefore using model B, the covariates significantly affecting cancer mortality in the EU are fruit and vegetable consumption which both have an inverse association with the risk of mortality, when taking the other variables into account, and animal fat consumption and smoking which are shown to have a positive association. Again, the actual effect size of these variables will be discussed further on. There has been no partitioning of variance yet so all the variation here is due to differences between regions ($\sigma^2_u = 0.014$). If we compare this value to the variance under model A ($\sigma^2_u = 0.040$) it can be seen there has been a 66% reduction in variation after adding the explanatory variables. Due to the significant explanatory variables being at country level, this shows that these factors are helping to explain the differences in cancer mortality between regions and countries. The total variance

has been partitioned in model C and it can be seen that 76% of this is attributable to differences between countries.

To examine the distribution of disease risk across the EU, the estimates of relative risks from models A, B and C have been mapped and are shown in Figures 5.5 to 5.7. The main features of Figure 5.5 indicate that France and Denmark have particularly high cancer mortality rates. Cancer mortality rates in these regions are between 20% and 54% higher than expected if the European age and sex specific mortality rates had been applied to that area. Greece, Portugal, Finland and Sweden appear to have the lowest rates. The pattern suggests that there is lower cancer mortality in the east and higher in the west with the exception of Portugal. However, data are missing for 4 EU countries, illustrated in white here, so it is difficult on the basis of these data to tell if this pattern is consistent across the whole EU. Figure 5.6 then shows the relative risks after taking into consideration the effects of the risk and protective factors that have previously been shown to have a significant effect on cancer mortality. This map shows less variability with the relative risks now only ranging from 0.77 to 1.36 (they ranged from 0.59 to 1.54 in Figure 5.5). The risk factors have, therefore, explained some of the variation in the cancers mortality rates in the EU, suggesting that even these initial models have been effective at smoothing geographic variation. The map for model B is a lot smoother, hence allows easy identification of regions that could be labelled disease 'hotspots'. The disease map from model C is very similar to model B's. However, France noticeably has more high risk areas due to the country level relative risks for France being high and influencing the estimates for its regions.

Figure 5.5 Estimated Relative Risks from model A: all cancer mortality in the EU

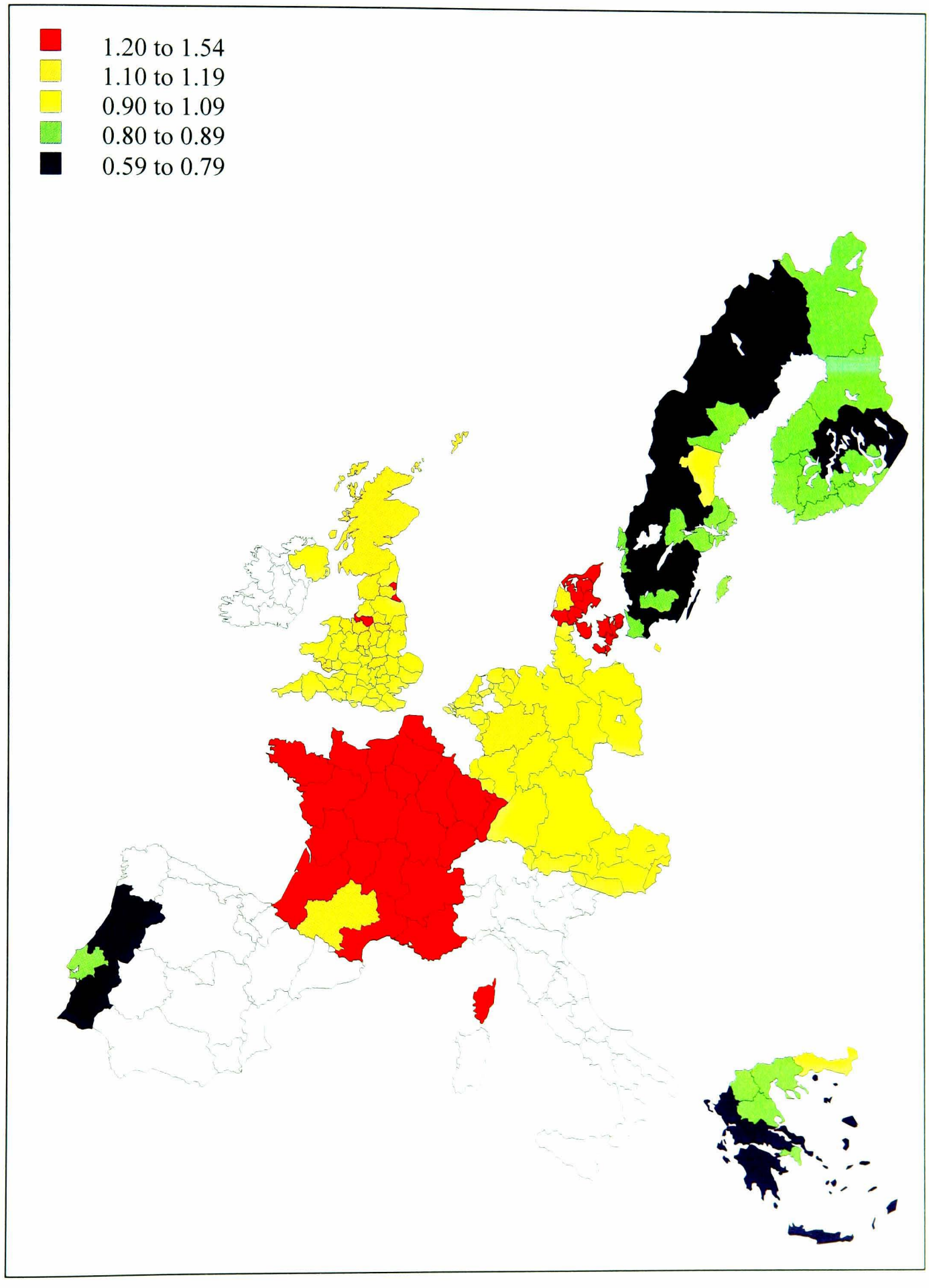


Figure 5.6 Estimated Relative Risks from model B: all cancer mortality in the EU

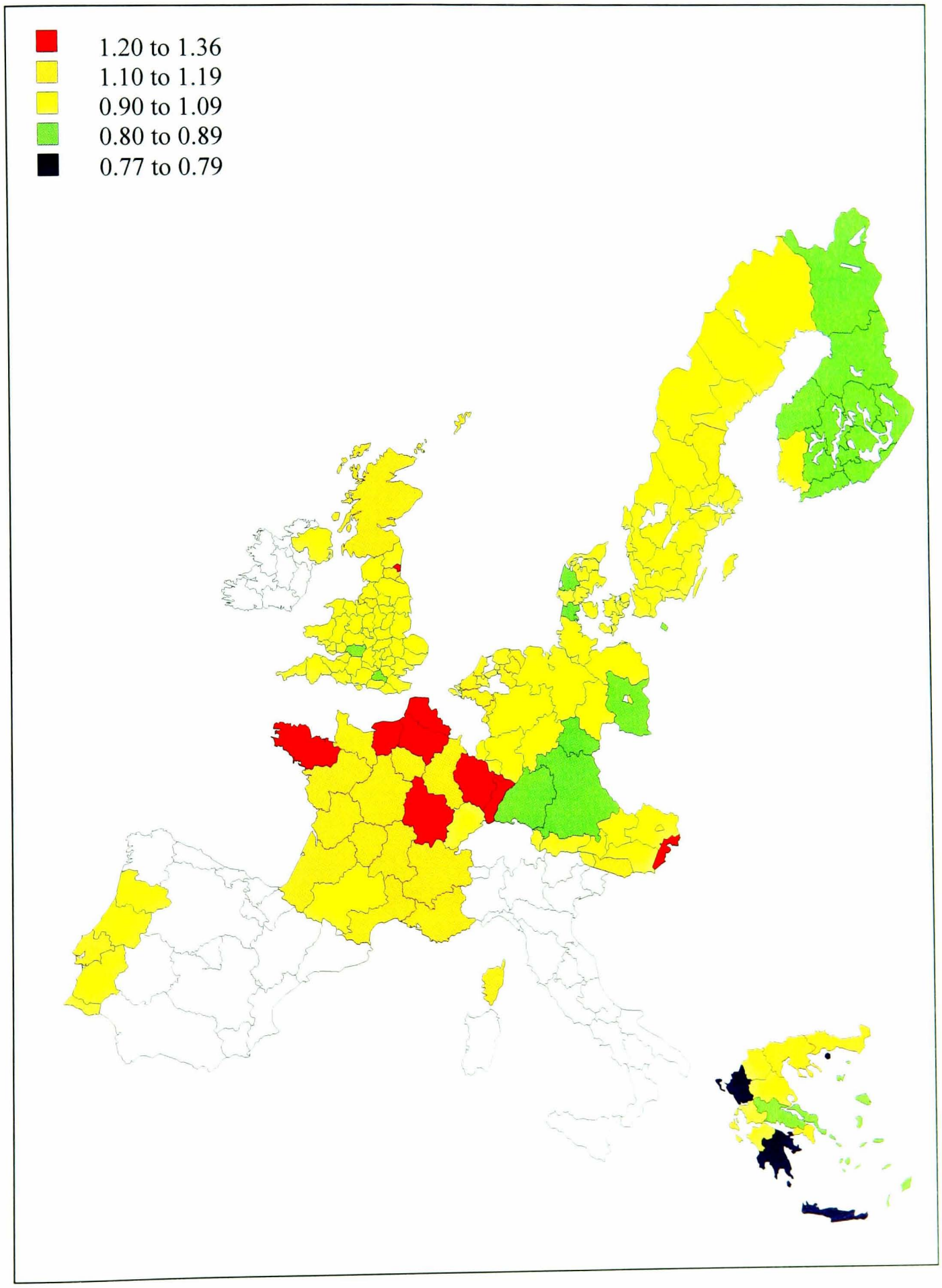
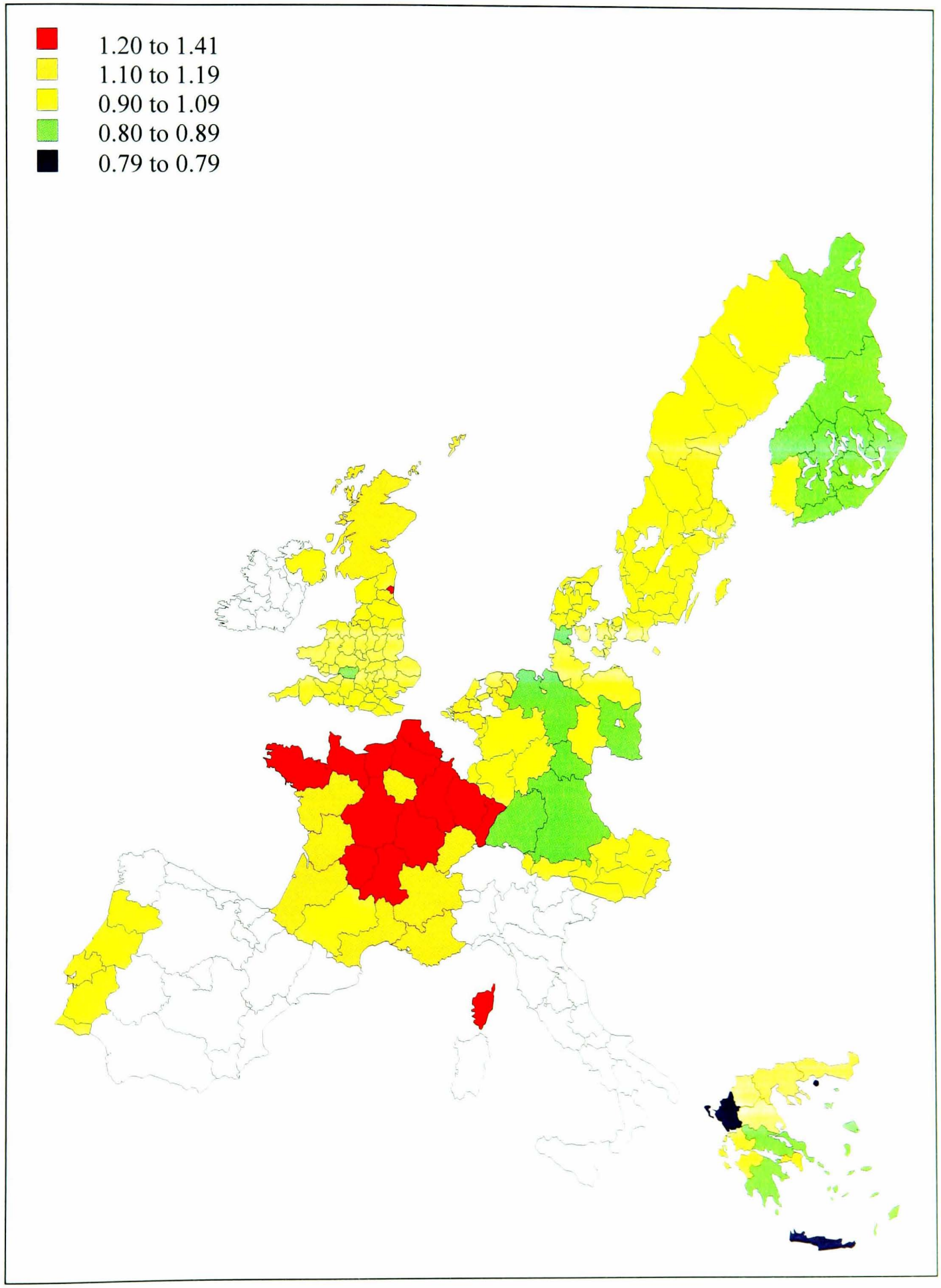


Figure 5.7 Estimated Relative Risks from model C: all cancer mortality in EU



5.3 Estimating the Models

The variance components models and subsequent models in this Chapter were fitted using a quasi-likelihood (frequentist) approach. This method involves finding pseudo maximum likelihood estimates for the unknown parameters in the model. The algorithm used here to find such estimates is known as iterative generalised least squares (IGLS) estimation or a restricted version (RIGLS).

5.3.1 Iterative Generalised Least Squares

The iterative generalised least squares method was introduced by Goldstein (124) to estimate normal response multilevel model. This procedure is based on a generalised least squares estimation that produces maximum likelihood (ML) estimates. Full details of how the algorithm fits multilevel models can be found in Goldstein (122).

This two-stage process involves estimating fixed and random parameters (variances and covariances of the random coefficients) in successive iterations using IGLS. Goldstein (122) described the basic model of fixed and random effects. Considering this in vector notation,

$$Y = X\beta + Z\theta,$$

where Y is a vector of observations being modelled by explanatory variables X and associated fixed parameters β, and explanatory variables Z with random coefficients θ. The fixed and random part design matrices X and Z need not be the same. θ is assumed to contain a set of random error terms along with other random effects.

Firstly, the fixed parameters are estimated using ordinary least squares regression, assuming higher level variance in the model to be zero. The vector of residuals produced from this model can then be used to construct initial values for the dispersion matrix V. The estimation procedure is iterative and it is firstly

applied using a generalised least squares estimation procedure to obtain the estimator for the fixed coefficients

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y . \tag{5.6}$$

The residuals are calculated again,

$$\tilde{Y} = Y - X\hat{\beta}$$

and by forming the matrix product of these residuals, $Y^* = \tilde{Y}\tilde{Y}^T$ and then stacking them into a vector produces $Y^{**} = \text{vec}(\tilde{Y}\tilde{Y}^T)$. The variance of the random coeffients θ, $\gamma = \text{cov}(\theta)$, can then be estimated as,

$$\hat{\gamma} = (Z^{*T} V^{*-1} Z^*)^{-1} Z^{*T} V^{*-1} Y^{**} . \tag{5.7}$$

Z^* is the appropriate design matrix for the random parameters and V^* is the Kronecker product of V, namely $V^* = V \otimes V$, where

$$V = E(\tilde{Y}\tilde{Y}^T) = E(Y^*) . \tag{5.8}$$

Now, assuming multivariate normality, the estimated covariance matrix for the fixed parameters can be written as

$$\text{cov}(\hat{\beta}) = (X^T V^{-1} X)^{-1} .$$

Goldstein and Rasbash (125) showed that covariance for the associated random parameters can be written as

$$\mathrm{cov}(\hat{\gamma}) = 2(Z^{*T} V^{*-1} Z^{*})^{-1}.$$

These random parameters and the associated variances can then be estimated from the model in a similar manner as the fixed parameters and their variances. In comparison to ordinary least squares regression, the random part is modelled taking into account the structure of the data allowing us to estimate a set of parameters rather than having a single residual error term (6).

5.3.1.1 Penalised Quasilikelihood Estimation

As previously discussed, a Poisson distribution is being modelled, therefore the non-linear (logarithmic) relationship between the outcome variable and the predictor part of the model has to be taken into account. This has been done by making a linearising approximation to estimate the random parameters. Looking at a simple case of heterogeneity effects only (as in VC model equation (5.4)), the residuals $\hat{u}_i$ can be estimated from the model using penalised quasi-likelihood (PQL) estimation with a second order Taylor series approximation (122, 125). After each iteration *t,* predictions H_t are made from the model where

$$H_t = X_i \hat{\beta}_t + \hat{u}_i .$$

Then these are used to calculate new predictions for iteration $t + 1$, so that

$$\begin{aligned} f(H_{t+1}) = f(H_t) + x_i(\hat{\beta}_{t+1} - \hat{\beta}_t) f'(H_t) \\ + \hat{u}_i f'(H_t) + \hat{u}_i^2 f''(H_t)/2' \end{aligned} \qquad (5.9)$$

where $f(\bullet)$ is a link function. An updating function is provided for the fixed part of the model by the first two terms on the right hand side of equation (5.9). The third

term is made up of a linear random component that is created by multiplying the first differential of the predictions by the random part of the model while the next term in the Taylor expansion about H_t makes up the fourth term in (5.9). Following on from this, the Poisson distribution takes the form

$$f(H) = f'(H) = f''(H) = \exp(X_t\hat{\beta}_t + \hat{u}_i),$$

and at each iteration, estimates are made about the fixed part of the model plus the residuals.

5.3.1.2 Marginal Quasilikelihood

This PQL method described above generally gives better estimates than the MQL method but is more prone to convergence problems or with the estimates 'blowing up' if the residuals are too large. In such cases, a MQL model can be estimated by choosing H_t to be the current value of the fixed part parameter only, that is omitting the estimated residuals from the linear component of the nonlinear function

$$H_t = X_i\hat{\beta}_t.$$

The MQL procedure does, however, tend to underestimate the values of both the fixed and random parameters, especially when the sample size is small (122).

Also, to help overcome convergence problems the term involving the second derivative in (5.9) can be omitted giving a first order approximation. However, it is expected that its inclusion, in general, improves estimates, so a typical procedure, when facing convergence problems, is to estimate using MQL first order initially and if this succeeds attempt estimating using second order and PQL procedures. The differences between the estimation procedures have been illustrated (126).

5.3.2 Restricted Iterative Generalised Least Squares

As previously discussed, the IGLS procedure uses the current estimates of the fixed and random parameters to iterate between equations (5.6) and (5.7). This procedure often produces biased estimates so Goldstein (127) shows how a simple modification can lead to restricted iterative generalised least squares (RIGLS) which produces estimates that are unbiased.

Using estimates of $\hat{\beta}$ to rewrite (5.8) gives

$$E(Y^*) = V_2 - X\,\mathrm{cov}(\hat{\beta})X^T = V_2 - X(X^T V_2^{-1} X)^{-1} X^T, \qquad (5.10)$$

where

$$V_2 = V_{2(1)} + V_{2(2)},$$

given that, in the two-level model, the residual matrices, θ_1 and θ_2, have expectation as follows:

$$E(\theta_1\theta_1^T) = V_{2(1)},\ E(\theta_2\theta_2^T) = V_{2(2)} \text{ and } E(\theta_1\theta_2^T) = 0.$$

Equation (5.10) has taken account of the sampling variation of $\hat{\beta}$ and subsequently allows an unbiased estimate of V_2 by adding the 'hat' matrix to Y^* at each iteration until convergence. This technique is similar to restricted maximum likelihood (REML) in normal response models.

5.4 Spatial Multilevel Modelling

The modelling so far does not take into account the geographical structure. Fitting the spatial model to the data takes into account the fact that areas close to each other in geographical space may share common factors that influence cancer mortality.

5.4.1 Spatial Model

A spatial multilevel model can be formed by extending the variance components models. This involves adding further random effects and the full spatial model for the EU cancer morality data can be written as follows:

$$O_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \log(E_i) + \alpha_i + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i} + u_i + v_i \quad , \qquad (5.11)$$

$$v_i = \sum_{j \neq i} z_{ij} v_j^* \, , \qquad (5.12)$$

$$\begin{bmatrix} u_i \\ v_j^* \end{bmatrix} \sim N(0, \Omega_{uv}), \quad \Omega_{uv} = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}. \qquad (5.13)$$

The model is as before but now there is the added parameter v_i. The v_i are spatially dependent random effects, and may have any one of a number of structures describing adjacency or nearness in space (128). The spatial effects v_i are considered to be the weighted sum of a set of independent random effects v_i^*. The v_i^* can be considered to be the effect of area upon other areas, moderated by a measure of proximity of each pair of areas z_{ij}. There are many ways in which z_{ij} can be formulated, in general it is written (129)

$$z_{ij} = w_{ij} / w_{i+} \, . \qquad (5.14)$$

In this case the w_{ij} are either 1's or 0s representing an adjacency matrix and

$$w_{i+} = \sum_{j \neq i} w_{ij} \, .$$

The binary adjacency matrix is the most common spatial weight matrix; however, others can be used to define the spatial structure in the model. The adjacency

matrix cannot differentiate the strength of spatial effects between contiguous locations, therefore more complex spatial weight matrices are often proposed for more precise spatial control.

The v_i* can then be estimated directly from the model and these are the residuals due to their independence. Therefore this multilevel model has within-area effects which are modelled with a Poisson distribution and relative risks between regions which are considered as having a lognormal distribution.

There is a more complex covariance structure present now. The variance has been partitioned with σ_u^2 referring to the variance that arises due to heterogeneity between regions and σ_v^2 referring to that which arises due to the spatial structure. The random effect has now been written as the sum of the heterogeneity effect, u_i, and a correlated spatially structured component, v_i. The model takes on similar structure to a simple autocorrelation model. The weights, z_{ij} can be thought of as spatial explanatory variable, and represent a measure of the relevance of area j to area i.

When fitting the model parameters, the spatial effects are more complex than the heterogeneity effects where there was simply a variance-covariance matrix with a variance term on the diagonal. Adding spatial effects requires off-diagonal terms in the variance-covariance matrix.

To consider the structure of the spatial part equations (5.11) to (5.13) can be rewritten in matrix notation

$$Y = \{\log(E_i) \quad 1 \quad x_{1i}\ldots x_{6i}\} \begin{bmatrix} 1 \\ \alpha \\ \beta_1 \\ \ldots \\ \beta_6 \end{bmatrix} + \begin{bmatrix} Z_u Z_v^* \end{bmatrix} \begin{bmatrix} \theta_u \\ \theta_v^* \end{bmatrix}, \tag{5.15}$$

where Z_u is the identity matrix and $Z_v^* = \{z_{ij}\}$. The variance structure can be written as

$$\text{var}\left(\begin{bmatrix}\theta_u \\ \theta_v^*\end{bmatrix}\right) = \begin{bmatrix}\sigma_u^2 I & \sigma_{uv} I \\ \sigma_{uv} I & \sigma_v^2 I\end{bmatrix}, \tag{5.16}$$

$$\text{where } \theta_u = \{u_i\} \text{ and } \theta_v^* = \{v_i\},$$

which is equivalent to

$$\text{var}\left(\begin{bmatrix}u_i \\ v_j^*\end{bmatrix}\right) = \begin{bmatrix}\sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2\end{bmatrix}.$$

The overall variance, conditional on the fixed parameters, can be written as

$$\text{var}(Y \mid X\beta) = Z\Sigma_\theta Z^T,$$

where Σ_θ is the variance of the random terms in θ. Using the partitions from the spatial model (5.15) defined in θ and Z and the variance structure of (5.16), it follows that:

$$\text{var}(Y \mid X\beta) = \sigma_u^2 Z_u Z_u^T + \sigma_{uv}(Z_u Z_v^{*T} + Z_v^* Z_u^T) + \sigma_v^2 Z_v^* Z_v^{*T}.$$

As mentioned above, the z_{ij} can be formulated in different ways. In this example, as shown in (5.14), the variance of an area decreases as the number of neighbours increases.

Finally, using RIGLS estimation, the random effects for heterogeneity and the spatial effects are specified within a generalised linear modelling framework. Weights matrices associated with the random effects are constructed and fitted into the model. The model can now be expressed in terms of 3 design matrices and is generalisable to the non-linear model.

5.5 Results from Spatial Model

The Spatial model was fitted to the EU cancer mortality data following the estimation procedure above. In this case model D is the model incorporating the spatial structure of the data and is fitted using second order PQL and RIGLS; this is shown alongside the results from models A and B in Table 5.6.

Table 5.6 Results from variance components and spatial models: EU data

	Model A Estimate (95% C.I.)	**Model B** Estimate (95% C.I.)	**Model D** Estimate (95% C.I.)
Fixed Part			
β_o	7.53 (7.50 , 7.56)	6.83 (6.68 , 6.99)	6.83 (6.53 , 7.13)
β_1 *(FRU)*		-7.58 (-8.65 , -6.50)	-7.51 (-9.67 , -5.36)
β_2 *(VEG)*		-1.12 (-1.69 , -0.55)	-1.11 (-2.25 , 0.04)
β_3 *(ANF)*		26.36 (19.28 , 33.42)	25.98 (11.72 , 40.23)
β_4 *(ALC)*		0.64 (-0.17 , 1.45)	0.71 (-0.91 , 2.33)
β_5 *(SMO)*		0.49 (0.39 , 0.59)	0.49 (0.29 , 0.69)
β_6 *(GDP)*		0.20 (-4.61 , 5.01)	0.33 (-9.40 , 10.06)
Random Part			
σ^2_u	0.040 (0.032 , 0.048)	0.014 (0.011 , 0.017)	0.002 (-0.014 , 0.017)
σ_{uv}			0.011 (-0.001 , 0.023)
σ^2_v			0.044 (0.011 , 0.077)

Firstly, comparing the fixed parts results from the variance components model and this model, the estimates change very little after including the spatial component. However, looking at the standard errors, they almost double for each term in the spatial model. The probable reason behind this is that areas

geographically close tend to display positive dependence or positive autocorrelation, but when this is ignored as in the variance components model, incorrect inference will result, in particular standard errors of explanatory variables will be too small. Fitting the spatial model here has improved the accuracy of the model. Using the variance terms, the total variance can be calculated. As can be seen from equation (5.12), σ_v^2 is based on the residual, v_j^*. To obtain the weighted residual, v_i, v_j^* must be divided by the number of neighbours region i has, n_i. Therefore, to obtain a weighted estimate of σ_v^2, it must be divided by the average number of neighbours a region has, $\overline{n}$. Hence, the total variance is $\sigma_u^2 + \sigma_v^2 / \overline{n} = 0.002 + 0.044 / 4.128 = 0.0127$. 85% of this variance now arises from spatial effects, confirming the importance of the spatial part in the model.

We are also interested in quantifying the effect size of the risk or protective factors when examining cancer mortality. The relative risks have been calculated for each of the significant variables from model D and are given in Table 5.7. Firstly, for the consumption of fruit, the table shows that Greece has the highest level of consumption in these 11 countries and the UK has the lowest. The relative risk of 0.58 shows that a population consuming, on average, the same amount of fruit as Greece has a risk of cancer mortality that is 42% lower than if consumption was on the same level as the UK. Also, consuming the same level of vegetables as Greece leads to a risk of cancer mortality that is 28% less than if consumption was at the same level as Finland or Sweden. Luxembourg consumes the most animal fat in the EU countries being examined and this level of consumption leads to a RR over 2 times as high as if consumption was on the same level as Greece. Finally, the inhabitants of Greece, on average, smoke the most cigarettes. This level of smoking leads to a RR of cancer mortality 3 times as high as if cigarette consumption was at the same level as Sweden (where cigarette consumption is the lowest).

Table 5.7 Variable effect sizes: EU data

Variables	Area with Highest Consumption	Area with Lowest Consumption	Relative Risk
Fruit	Greece	UK	0.58
Vegetables	Greece	Finland & Sweden	0.72
Animal Fat	Luxembourg	Greece	2.03
Cigarettes	Greece	Sweden	2.99

To examine visually the apparent effect of risk factors and also to compare the different models fitted, plots of the predicted relative risks from all four models are given in Figure 5.8. Looking at the plot for model A, the variance components model with no explanatory variables included, it is obvious that there is definite clustering, with mortality rates for regions within each country being very similar. There also appears to be evidence of a negative slope, with the exception of Greece and Portugal, suggesting that rates tend to be highest in Western Europe and lowest in Eastern Europe. Examining the plot of model B, the variance component model including the explanatory variables, it can be seen that the spread decreases. The relative risks for countries with low relative risks are higher than in model A and those that previously had high relative risks are lower. This is because the risk or protective factors included are measured at the level of country and are therefore explaining differences between countries. The key features from fitting this model are that regions in Portugal, which were previously shown to have low cancer mortality rates, now appear to have a risk of cancer mortality that is higher that expected given the lifestyle factors. Portugal has fairly high consumption of fruit and vegetables and very low animal fat consumption, so after taking these into consideration they have above average mortality. Also, looking at Denmark, the cancer mortality risks are high under model A but under model B for most regions they are below average. So, taking into account the fact that Denmark has high animal fat consumption and low fruit and vegetable consumption, the rates are actually lower than expected. Moving

onto the estimates obtained from model D, there is less variability overall. In plot B they were clustered but showed more variation within each country and more overlap between countries. With the spatial model the relative risks are more tightly clustered within countries. This is due to the mean for each area being centred on the mean of its neighbours. Effectively this model is providing greater smoothing to a map. Finally, comparing the plot from model D to that of model C, which takes account of country as a level, it can be seen that model D provides the greatest amount of smoothing and reduces the variability the most.

It should be noted here that there are some islands present in the data which, due to the level of data aggregation, have no regions bordering them. These regions have no neighbours so it may be assumed that they do not share common environmental and social factors with nearby areas, therefore island estimates would not be spatially smoothed. It is perhaps false to assume this, so each island was allocated two or three 'neighbours' based on the regions that were closest (ie smallest centroid to centroid distance). These regions were not necessarily all from the same country as the island but hopefully they provided some sort of smoothing based on the regions sharing similar spatial factors which may be influencing the rates. Also, as discussed in Chapter 3, missing data are inevitable in this type of study and when examining the EU, three countries have missing data for the time point of interest. This resulted in the regions which bordered areas with missing data being subject to a lower level of smoothing than regions which had information available for each of their neighbours. One should remain aware that the estimates for these areas are shrunk much less towards estimates from regions that share similar environmental and social factors. Also, areas at the boundary of the map have been smoothed to a lower degree.

Further, a slightly different form of disease map can be used to identify the risk of mortality after including actual variable effects. This is done by mapping each region's relative risks plus the fixed effects from the model. Figure 5.9 does so for Model D's (spatial model) estimates. As can be seen, this map differs from those previously examined. The relative risks of cancer mortality are all very high

($\geq$ 1.4). This is because the risk factors have stronger effects that the protective factors, hence, they are having more overall influence on the relative risks and pulling them higher. There is a high amount of country level clustering evident due to the covariates being at the country level and drawing the risks within countries towards similar values. For this reason, this type of disease map is most useful for identifying differences in country level relative risks. However, it is noticeable that two regions in France have risk of mortality higher than the rest of the country (and most of the EU) after the addition of the effect size of the risk factors. Country level differences indicate that Denmark, overall, has the highest risk of cancer mortality and Portugal, Greece and Finland have the lowest after adding the fixed effects.

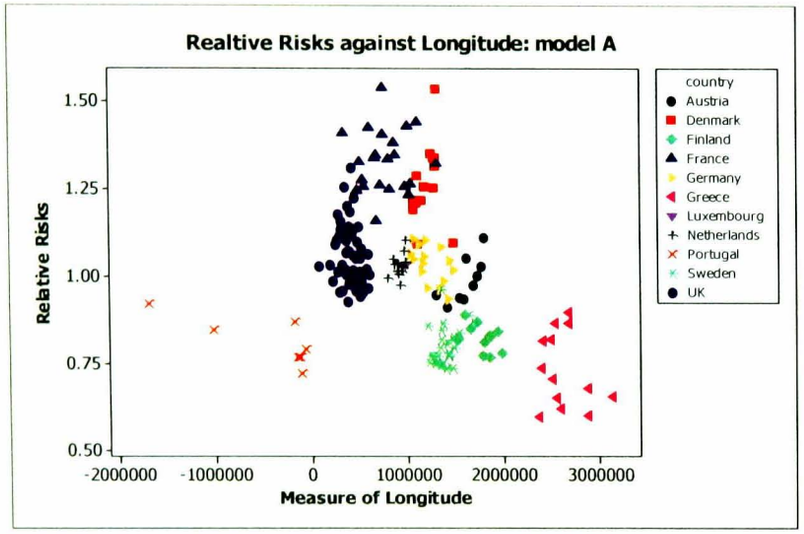
Maps from four types of models have been explored in this chapter. In theory, further maps could be investigated that examine the risks that can be attributed to different parts of the model, eg regional, spatial, country level risks. Figure 5.3 combines the region and country effects to obtain overall estimates of relative risks; however these could have been mapped separately. For the purpose of producing smoothed disease maps and 'hotspot' identification, the overall relative risks are more appropriate and further disease mapping will focus on using this method.

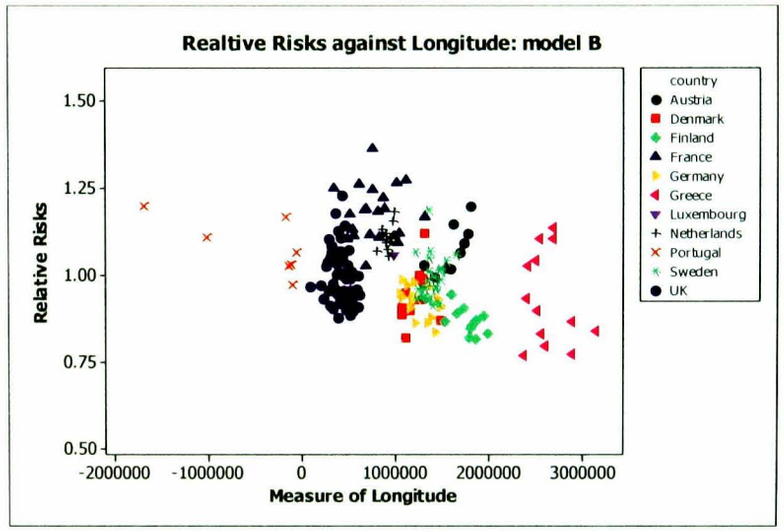
5.6 Further Spatial Modelling

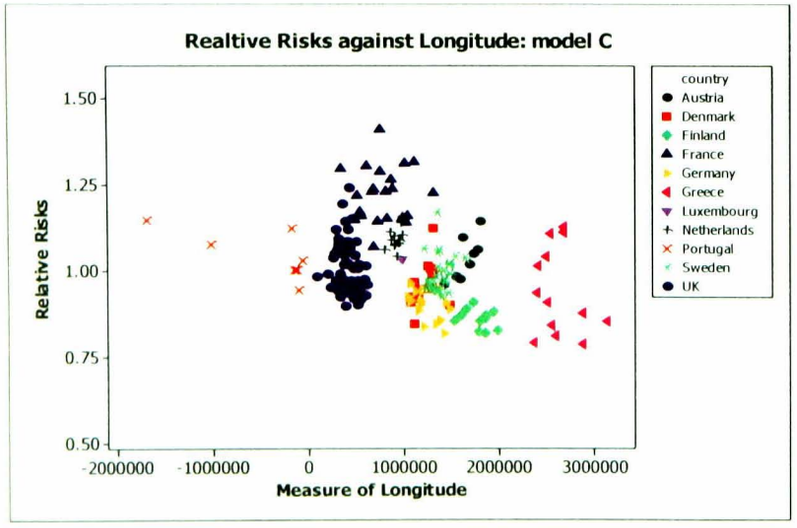
Empirical Bayes methods provide useful tools to fit spatial multilevel models. So far we have looked at two examples that incorporate the geographical structure of the data in to the model; one takes account of the non-independence of regions within countries by including a further hierarchical level and the other takes into account the fact that regions geographically close are more likely to share similar mortality rates and exposure to risk factors than regions far away from one another. The model with the spatial random effects appears to be more useful as it provides more smoothing and produces a disease map from which disease 'hotspots' can be easily identified. However, it would useful to explore the effects

of adding country as a higher level to the spatial multilevel model. This will be carried out in the next chapter by fitting the models using fully Bayesian methods.

Figure 5.8 Plots of Relative Risks against Longitude from models A, B, C and D for EU cancer mortality







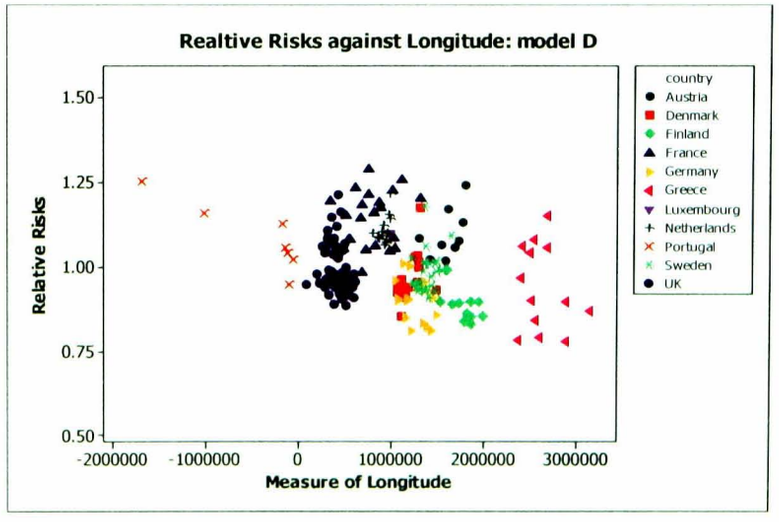
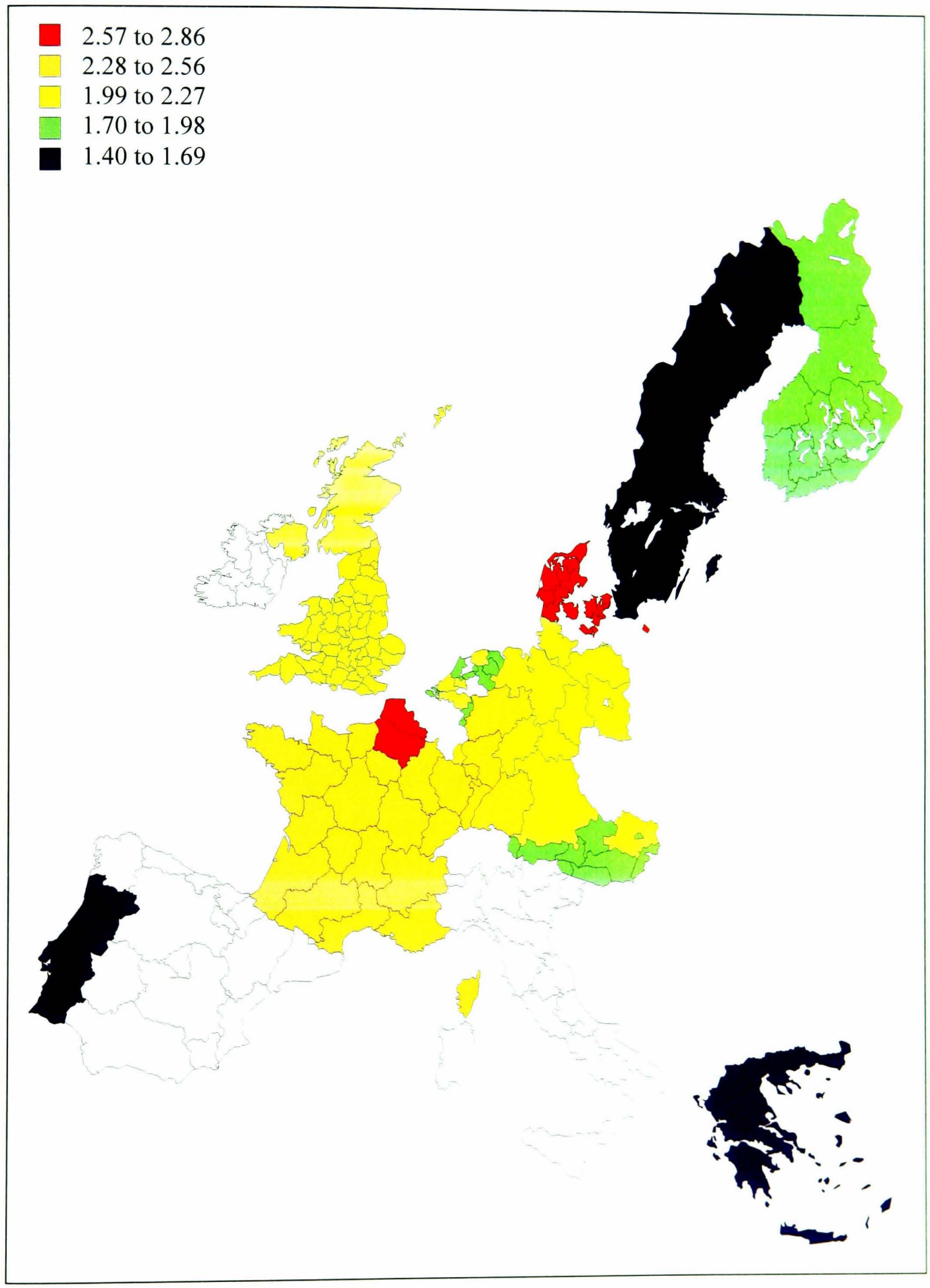


Figure 5.9 Estimated Relative Risks from model D plus fixed effects



Chapter 6

6 Fully Bayesian Modelling

6.1 Bayesian Estimation for Relative Risks

There has been much development recently of Bayesian methods in relation to disease mapping applications. In simple terms, Bayesian approaches to disease mapping of mortality rates combine two types of information. The first for each region is the number of observed deaths, described by the Poisson likelihood $g(\boldsymbol{O} \mid \boldsymbol{\theta})$, and the second is the prior information on the relative risks specifying their variability in the overall map, summarised by their prior distribution ($\boldsymbol{\theta}$) (78). As discussed in Chapter 4, a common assumption made when examining counts of deaths within an area is that $O_i \sim Poisson(e_i\theta_i)$, and that $\theta_i \sim Gamma(\alpha, \beta)$. The joint distribution can then be given by the product of the Poisson likelihood and the gamma distribution. This joint density is proportional to the posterior distribution for the parameters of interest.

6.1.1 Posterior Distribution

Prior beliefs about parameters of interest are combined with sample information to create updated, or posterior beliefs about the parameters. In the fully Bayesian approach, inference is based on the posterior distribution of $\boldsymbol{\theta}$ given the data and with the empirical Bayes approach the posterior distribution is approximated in some way. This chapter will focus mainly on the fully Bayesian approach, which has recently become widely available mainly because of the increased use of Markov chain Monte Carlo methods of posterior sampling.

Bayesian approaches to relative risks were discussed in section 4.2.1; the likelihood function of $\boldsymbol{\theta}$, the relative risks, were described in equations (4.4) and (4.5) and how to obtain the marginal posterior distribution was explained in (4.6) and (4.7), using empirical Bayes estimation, and in (4.23) and (4.24) using fully Bayesian methods. It was explained that, in general, the marginal posterior distribution cannot be solved analytically.

6.1.2 Empirical Bayes

The empirical Bayes approach was used to fit the spatial multilevel models in Chapter 5. This method assumed that the hyper-parameters are unknown and are drawn from an unspecified distribution. The parameters were then estimated using a technique such as iterative generalised least squares. As discussed, this is a two-stage process which involves estimating fixed and random parameters in successive iterations.

The spatial multilevel model (see equations (5.11) to (5.13)) can be written as

$$O_i \sim Poisson(\theta_i E_i) \tag{6.1}$$

$$\log(\theta_i) = z_i^T \boldsymbol{\beta} + u_i + v_i \,, \tag{6.2}$$

where the random effects have a joint normal distribution

$$\begin{bmatrix} u_i \\ v_i \end{bmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix}. \tag{6.3}$$

The dispersion matrix Σ is a linear function of hyper-parameters $\boldsymbol{\gamma}$ such that

$$\Sigma = f(\boldsymbol{\gamma}).$$

Therefore the algorithm alternates between an estimate of the fixed parameters conditional on the estimates of the hyper-parameters $[\hat{\beta} \mid \gamma = \hat{\gamma}]$ and an estimate of the hyper-parameters conditional on the fixed parameters and estimates of the hyper-parameters after x iterations $[\hat{\gamma} \mid \beta = \hat{\beta}]$.

6.2 Fully Bayesian

As previously discussed the empirical Bayes approach involves assuming the hyper-parameters are unknown and are drawn from an unspecified distribution, and estimates of the hyper-parameters are plugged in. Another method that was used to fit spatial multilevel models to the European cancer mortality data set was a fully Bayesian approach. This involves fitting a hierarchical model where the distribution of the hyper-prior [γ] is specified.

6.2.1 Prior Distributions

The prior distribution for the relative risks $\boldsymbol{\theta}$ is a probability distribution that explains all the information that is known about $\boldsymbol{\theta}$ before the data have been collected. An informative prior distribution is used when information that is available before data collection is included in the analysis. A non-informative prior distribution is commonly used and expresses no knowledge about $\boldsymbol{\theta}$ before data collection. They are also referred to as diffuse or flat priors and a common example is the uniform distribution over the range of sample space for $\boldsymbol{\theta}$.

6.2.1.1 Spatially Structured Priors

Prior knowledge of mortality rates indicates that areas geographically close to each other tend to have similar relative risks. To express the prior knowledge that there exists a local spatially structured variation in relative risks of cancer mortality in Europe, a spatially structured prior was used. Here the relative risks

have a locally dependent prior probability structure, whereby the conditional distribution of the relative risk in area i, given the values of the relative risks in all other areas $j \neq i$, depends only on the relative risks of area i's neighbouring areas.

Here we used a log-normal model which has a normal prior with hyper-parameters μ and Σ specified for the logarithm of the relative risks:

$$f(\log(\theta)|\ \mu, \Sigma) = |2\pi\Sigma|^{-1/2} \exp[-{}^1/_2[\log(\theta) - \mu]^T \Sigma^{-1} [\log(\theta) - \mu])$$

where

$$\mu_i = z_i^T \beta$$

and Σ is the dispersion matrix given in equation (6.2).

The prior model fitted results in the log relative risks being the sum of two independent components:

$$\log(\theta_i) = u_i + v_i^* \quad . \tag{6.4}$$

The random terms, u_i and v_i, have a joint prior distribution as described in equation (6.3) and the variance of $\log(\theta_i)$ is then dependent on the number $\bar{n}_i$ of neighbours of that area,

$$Var[\log(\theta_i)] = \frac{\sigma_v^2}{\bar{n}_i} + \sigma_u^2 \quad ,$$

and the covariance between areas i and i' depends on the number $\bar{n}_{ii'}$ of neighbours they have in common,

$$\mathrm{Cov}[\log(\theta_i), \log(\theta_{i'})] = \begin{cases} \dfrac{\bar{n}_{ii'}}{\bar{n}_i \bar{n}_{i'}} \sigma_v^2 + \left(\dfrac{1}{\bar{n}_i} + \dfrac{1}{\bar{n}_{i'}} \right) \sigma_{uv} & \text{if } i \text{ and } i' \text{ are contiguous} \\ \dfrac{\bar{n}_{ii'}}{\bar{n}_i \bar{n}_{i'}} \sigma_v^2 & \text{otherwise} \end{cases}$$

so that there is no covariance between areas with no common neighbours.

6.2.2 Hyperpriors

The fully Bayesian approach was used to model the data and involved specifying the hyperprior distribution [γ] and basing inference about the relative risks on the marginal posterior distribution (see equations (4.23) and (4.24)). Specifying the hyperprior distribution and basing inference on the marginal posterior distribution allows the variability in the hyperparameters γ to be incorporated.

6.2.2.1 Hyperpriors for Fixed Effects

A prior distribution for fixed parameters has to be defined over the whole real line allowing there to be no constraints and to take on any real value. Here, a flat prior has been used which is basically a uniform prior across the whole real line $(-\infty, \infty)$ for each fixed effect coefficient.

6.2.2.2 Hyperpriors for Random Effects

Variance parameters must be constrained to have positive values. Here, a variance matrix is considered, as opposed to a single variance, and the prior used is a multivariate Wishart prior. The inverse variance matrix, Σ^{-1}, is given a Wishart distribution with 2 degrees of freedom and precision $2\hat{\Sigma}$ where $\hat{\Sigma}$ is an estimate of the dispersion Σ:

$$\Sigma^{-1} \sim Wishart\left(2\hat{\Sigma}\ ,\ 2\right).$$

The Wishart distribution is used as prior for the inverse covariance matrix by giving it degrees of freedom equal to the order of the matrix, 2. Reasonable values for the precisions matrix were then chosen based on previous estimates from similar models.

In this example, the marginal posterior distribution (4.24) is not tractable. To overcome this problem Monte Carlo methods are used which involve drawing

samples from the joint posterior distribution $g(\theta,\gamma|\mathbf{O})$ and hence from the marginal posteriors $g(\theta \mid \mathbf{O})$ and $g(\gamma \mid \mathbf{O})$.

6.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo is a common method used for posterior sampling and consists of a range of algorithms designed for the iterative simulation of joint posterior distributions found in Bayesian modelling (130). Since the posterior distribution of the spatial multilevel model is complex and therefore intractable, MCMC simulation provides a useful method of posterior sampling.

In simple terms, using current values of parameters, MCMC involves proposing new values, then a comparison of the posterior probability of the new and current values is made. These proposed values are generated from given distributions and subsequently new values are accepted based on a certain probability criterion. If these new values are accepted they replace the current values. Since these are simulation-based procedures, instead of simply producing point estimates, these methods are run for many iterations and at each iteration an estimate for each unknown parameter is produced. The aim is then to generate a sample of values from the posterior distribution of the unknown parameters.

6.3.1 Sampling Methods

Two algorithms are widely used to aid MCMC simulation; these are Gibbs sampling and the Metropolis Hastings algorithm. The aim is to generate a sample of points from the joint posterior distribution of the unknown parameters of interest. The spatial multilevel model would involve generating samples from the distribution $[\boldsymbol{x}, \gamma| \mathbf{O}]$, where $\boldsymbol{x}$ is the log of the relative risks $\boldsymbol{\theta}$:

$$[\boldsymbol{x},\boldsymbol{\gamma} \mid \mathbf{O}] \propto [\mathbf{O} \mid \boldsymbol{x}][\boldsymbol{x} \mid \boldsymbol{\gamma}][\boldsymbol{\gamma}] = \prod_{i=1}^{n} [y_i \mid x_i][\boldsymbol{x} \mid \boldsymbol{\gamma}][\boldsymbol{\gamma}]$$

Although it is difficult to simulate from the joint posterior distribution, the conditional posterior distributions for the unknown parameters often have forms from which simulation is easy. Iterative sampling from these conditional posterior distributions is then equivalent to sampling from the joint posterior distribution, in the limit.

6.3.1.1 Gibbs Sampling

A common approach to this type of posterior sampling is Gibbs Sampling. Basically, this involves taking each parameter in turn and simulating a new value from its conditional distribution assuming the other parameters are true values. Starting values are then needed for each parameter and these are then updated in turn.

For instance, since

$$[x \mid \gamma] = [x_i \mid x_{j,j\neq i}, \gamma][x_{j,j\neq i}, \gamma],$$

new values of x_i can be drawn, given the current values

$$x'_{j,j\neq i} \text{ and } \gamma' ,$$

from the full conditional distribution

$$[x_i \mid x'_j, j \neq i, \gamma', \mathbf{O}] \propto [y_i \mid x_i][x_i \mid x_j', j \in \partial i, \gamma'].$$

Here the conditional distribution of the relative risk in area i, given values for relative risks in all other areas $j \neq i$, depends only on the relative risk values in the neighbouring areas ∂i of area i. A new value of γ is drawn given the current values of x' from the full conditional distribution

$$[\gamma \mid x'] \propto [x' \mid \gamma][\gamma] .$$

The joint distribution of the sample values of (x, γ) should then converge to the joint posterior distribution $[x, \gamma \mid \mathbf{O}]$. Therefore the distribution of the sample

values of x, respectively of γ, converges to the marginal posterior distribution $[x|\mathbf{O}]$, respectively of $[\gamma|\ \mathbf{O}]$ (79).

The Gibbs sample works well if the conditional posterior distributions have simple forms and are easy to simulate from but with the Poisson hierarchical model this is not the case.

6.3.1.2 Metropolis Hastings Algorithm

When the conditional posterior distributions do not have simple forms another MCMC method, Metropolis Hastings sampling, can be used. As previously discussed MCMC involves proposing new values and accepts or rejects this value as the new estimate for the next iteration. The conditional distribution is used as the proposal value for the Gibbs sampler. This is in fact a special case of the Metropolis Hastings sampler whereby every proposed value is accepted. This MH algorithm was used to fit the spatial hierarchical model, as easily-simulated full conditional distributions were not available.

To explain the procedure in simple terms, a draw is made from a full conditional distribution of the kth component, say $f(\boldsymbol{x}_k \mid \boldsymbol{x}_{-k}, \mathbf{O})$, which is not easy to draw from. Then, a Metropolis step can be used to carry out the update.

Firstly select a proposal distribution, say $q(. \mid \delta)$, where δ is its parameters fixed by the user. Theoretically this can be any compatible distribution eg it makes sense to choose a Normal distribution if your parameter can be any real number.

Then, to sample a candidate point from $f(\boldsymbol{x}_k \mid \boldsymbol{x}_{-k}, \mathbf{O})$, draw $\boldsymbol{Y} \sim q(. \mid \delta)$ and compute

$$\tau = \frac{f(\boldsymbol{Y} \mid \boldsymbol{x}_{-k}, \mathbf{O}) q(\boldsymbol{x}_k, \delta)}{f(\boldsymbol{x}_k \mid \boldsymbol{x}_{-k}, \mathbf{O}) q(\boldsymbol{Y}, \delta)} \quad .$$

The new value of $\boldsymbol{x}_k$ is then set to be $\boldsymbol{Y}$ with probability minimum (τ , 1); otherwise we retain the current value of $\boldsymbol{x}_k$. Many iterations are then run until the chain converges to a stationary distribution, which should be the required joint posterior distribution.

6.3.1.3 Other Sampling Methods

Often the distribution of interest, $f(x)$ say, cannot be easily sampled from. However, if there exists a distribution $g(x)$ such that

$$f(x) < Mg(x) \forall x ,$$

where M is a positive number, Ripley (131) described a technique where $g(x)$ can be sampled from without problems. This method is called rejection sampling and involves thinking of $g(x)$ as an envelope function that completely bounds the required distribution. An extension to this is used when there is a non-standard but log concave distribution of interest. The technique that can be used here is called adaptive rejection sampling (132).

6.4 Fitting the Spatial Multilevel Model

A similar model can be fitted to that described in the previous chapter (see equations (5.11) to (5.13)) and is written as follows:

$$\log(\mu_i) = \log(\mathrm{E}_i) + \alpha + \sum_m \beta_m x_{mi} + u_{a[i]} + \sum_{j \in \partial[i]} w_{i,j} v_j , \qquad (6.5)$$

where m is 1,...,6 and β_m represents each of the fixed effects covariates. $a[i]$ is the area from which the observed count was taken from, and $\partial[i]$ is the set of neighbouring areas to the area from which the count is taken from. The weights in this model are such that $\sum_{j \in \partial[i]} w_{i,j} = 1 \forall i$ and all neighbours are given equal weights so that $w_{i,j} = 1/n_i$, where n_i is the number of neighbours to $a[i]$. Since there is a

one-to-one correspondence between the sets of area and neighbour residuals (u_i and v_j), a joint multivariate Normal distribution is fitted to these two sets of residuals.

$$\begin{bmatrix} u_i \\ v_j \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix} \right). \tag{6.6}$$

The fully Bayesian spatial multilevel model was fitted in WinBUGS v1.3 and the code is shown on the next page. Line 5 of the model specifies the response, namely the counts of cancer deaths, as Poisson distributed, and has a mean and variance μ_i. The constant N used has a value that is input with the data and is the number of regions in the data. Lines 6-12 specify $\log(\mu_i)$ as a linear additive function of the offset, which is the logarithm of the expected deaths, the intercept, β_1, and 6 covariates, $\beta_2,\ldots,\beta_7$. Line 13 adds the random effects that account for difference between regions. Region i is the region identifier (1 ... 187) for count of deaths i and is multiplied by u2[,1] (or u), which is a random variable. Lines 14-25 add the set of random effects to take account of spatial variability, u2[,2] (or v). The nearest neighbours are available and a region has at most 12 nearest neighbours. A set of random terms is given for each neighbouring region identifier (neigh1[i],...,neigh12[i]). Each of these random terms is multiplied by a spatial weight (weight1[i],...,weight12[i]). Lines 28-29 specify the prior for our random terms which, as previously discussed, follow a multivariate normal distribution. Line 32 specifies the priors for the fixed effects to be flat. Lines 34 and 35 define the hyper-prior for the variance matrix τ to be Wishart distributed with 2 degrees of freedom and precision matrix R2. The values for the precision matrix are declared with the data input and in this case, the values used were close to $2\hat{\Sigma}$, where $\hat{\Sigma}$ is an estimate of the dispersion matrix Σ obtained from the variance matrix when the model is run using IGLS.

$$R2 = \begin{bmatrix} 0.006 & 0.002 \\ 0.002 & 0.055 \end{bmatrix}.$$

The last part of the WinBUGS code (lines 36-40) generates the inverse of τ. This was required to obtain Σ since the precision matrix is the inverse of the variance. An example of the WinBUGS data file is given in Appendix A2.3.

```
#----MODEL Definition----------

model                                                                    (1)
{                                                                        (2)
# Level 1 definition                                                     (3)
for(i in 1:N) {                                                          (4)
deaths[i] ~ dpois(mu[i])                                                 (5)
log(mu[i]) <- offs[i] + beta[1]                                          (6)
+ beta[2] * smoke[i]                                                     (7)
+ beta[3] * fruit[i]                                                     (8)
+ beta[4] * veg[i]                                                       (9)
+ beta[5] * animal[i]                                                    (10)
+ beta[6] * alcohol[i]                                                   (11)
+ beta[7] * gdp[i]                                                       (12)
+ u2[region[i],1]                                                        (13)
+ weight1[i] * u2[neigh1[i],2]                                           (14)
+ weight2[i] * u2[neigh2[i],2]                                           (15)
+ weight3[i] * u2[neigh3[i],2]                                           (16)
+ weight4[i] * u2[neigh4[i],2]                                           (17)
+ weight5[i] * u2[neigh5[i],2]                                           (18)
+ weight6[i] * u2[neigh6[i],2]                                           (19)
+ weight7[i] * u2[neigh7[i],2]                                           (20)
+ weight8[i] * u2[neigh8[i],2]                                           (21)
+ weight9[i] * u2[neigh9[i],2]                                           (22)
+ weight10[i] * u2[neigh10[i],2]                                         (23)
+ weight11[i] * u2[neigh11[i],2]                                         (24)
+ weight12[i] * u2[neigh12[i],2]                                         (25)
}                                                                        (26)
# Higher level definitions                                               (27)
for (j in 1:N) {                                                         (28)
u2[j,1:2] ~ dmnorm(zero[1:2],tau.u2[1:2,1:2])                            (29)
}                                                                        (30)
# Priors for fixed effects                                               (31)
for (k in 1:7) { beta[k] ~ dflat() }                                     (32)
# Priors for random terms                                                (33)
for (i in 1:2) {zero[i] <- 0 }                                           (34)
tau.u2[1:2,1:2] ~ dwish(R2[1:2,1:2],2)                                   (35)
det <- tau.u2[1,1]*tau.u2[2,2] - tau.u2[1,2]*tau.u2[2,1]                 (36)
sigma2.u2[1,1] <- tau.u2[2,2]/det                                        (37)
sigma2.u2[1,2] <- -tau.u2[1,2]/det                                       (38)
sigma2.u2[2,1] <- -tau.u2[2,1]/det                                       (39)
sigma2.u2[2,2] <- tau.u2[1,1]/det                                        (40)
}                                                                        (41)
```

6.5 Convergence

The aim of using the sampling methods described is to simulate a Markov chain whose equilibrium distribution is the desired distribution (130). The desired outcome is for the joint distribution of the sample values to converge to the joint posterior distribution. It is therefore common practice to run a sufficiently long burn-in of samples, which can then be discarded, and further dependent samples that are obtained can be assumed to come from the joint posterior distribution. Being interested in the log relative risks $\boldsymbol{x}$, the marginal posterior distribution $[\boldsymbol{x}|\boldsymbol{O}]$ was approximated (ignoring the γ values). Then, for each region, point estimates can be obtained from the simulated values, for example the posterior mean from the sample mean, and interval estimation can be made by calculating Bayesian credible intervals. Therefore, it is important to be able to check when the Markov chain has reached a stationary distribution and so convergence to the posterior distribution has been obtained.

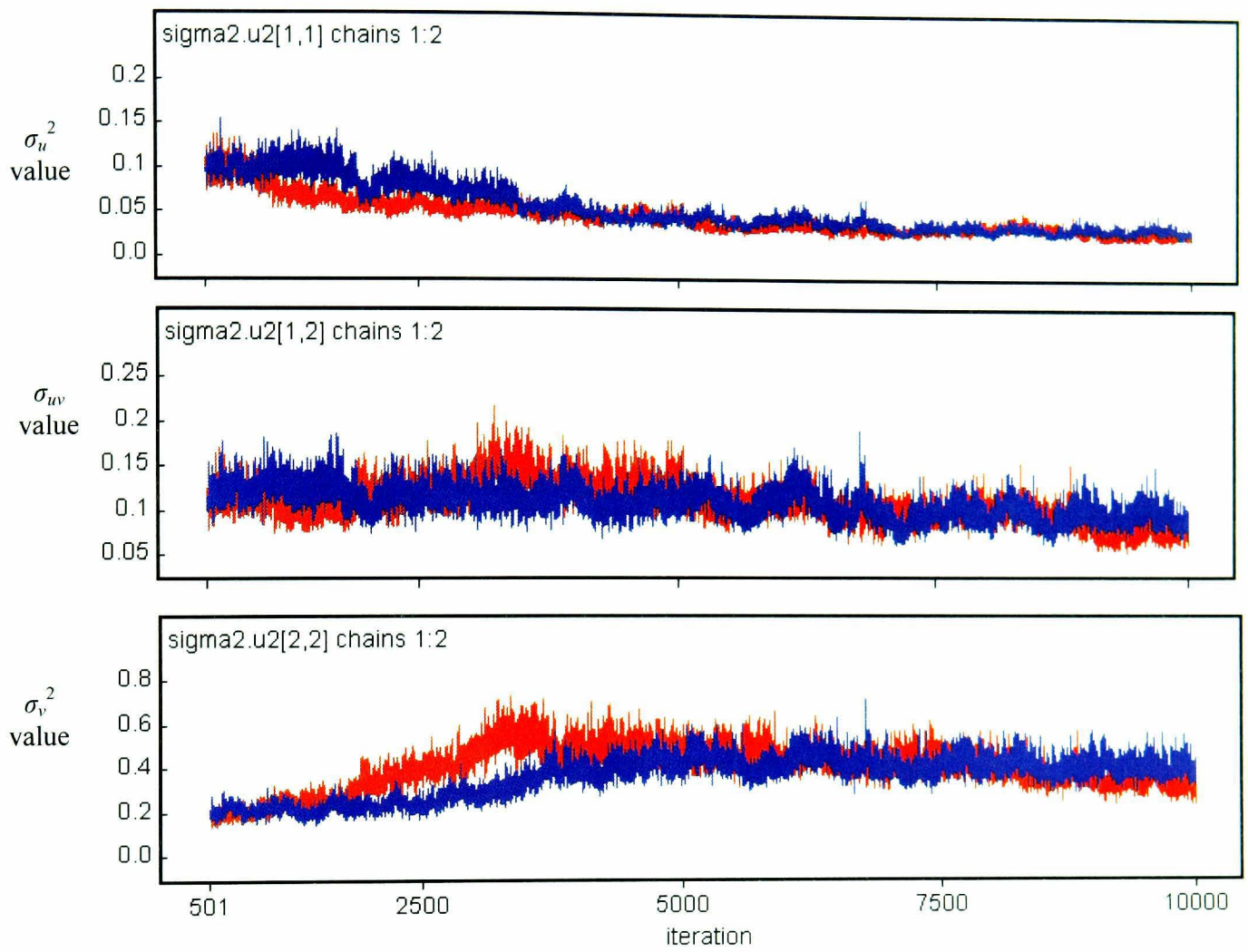
6.5.1 Multiple Chain Monitoring

The method used to assess convergence was to replicate the algorithm with different starting points and check that they show similar convergence behaviour. The reason for monitoring convergence in this way is that often a single sequence of Markov chain simulation appears to have converged, but when replicating the sequence independently it showed that within-sequence changes took too long to detect. Gelman and Rubin (133) give examples of applying Markov chain simulations of Bayesian posterior distributions where a single simulated sequence appears to have reached convergence, but when examining multiple independent simulations, poor convergence was evident. It has been suggested that with Bayesian posterior simulation, the added information obtained from replicating chains outweighs any additional costs required in multiple simulations (134).

6.5.1.1 Chain Trace Plots

It makes sense to use a combination of convergence diagnostics and visual inspection of trace plots to help determine when the simulation appears to have stabilised. For models with many parameters it is often impractical to check convergence for every parameter. In this case the relevant parameters must be chosen to be monitored, and convergence can then be examined with reasonable confidence. One method is to examine trace plots or time-series plots and compare the parallel sequences. When the patterns from parallel chains start to overlap and stabalise, it can be assumed convergence has been reached. Figure 6.1 gives an example of sample traces of chains from the spatial multilevel model. It shows plots of the random terms, σ_u^2, σ_{uv} and σ_v^2, where two parallel chains have been run from different initial values for 10,000 iterations. Looking at the plot for σ_v^2, that is sigma2.u2[2,2], it can be seen that by 2500 iterations the chains (represented by separate blue and red lines) are not yet overlapping. This suggests that the sample traces of the Markov chains of the random terms have not yet reached convergence at this point. Assessing the chains for further iterations, it can be seen that by around 5000 iterations, the sample chains are overlapping somewhat and by 7500 they have almost completely merged together. The other 2 plots, for σ_u^2 and σ_{uv}, have also converged by this stage and, in fact, both converge much more quickly. This suggests that convergence has been reached for all of the random terms, and hence 7500 simulations would be a suitable burn-in period. However, it doesn't make sense to only examine the random terms. Assessing convergence of the fixed terms is also of importance.

Figure 6.1 Trace plots of random terms



6.5.1.2 Gelman-Rubin Test

The Gelman and Rubin diagnostic test (135) is another method used to check the convergence of the MCMC algorithm. This method was initially inspired by the analysis of variance and the basic idea is to form an overestimate and an underestimate of the variance of the target distribution. When both these estimates are roughly equal, it can then be assumed convergence has been reached.

The Gelman-Rubin test is again based on running parallel chains from different starting values. Its convergence condition is that the empirical distribution of simulations that are obtained from each sequence separately is approximately equal to the distribution that is obtained by combining all the sequences together. Before convergence the samples collected within each single

sequence will be less variable than the samples collected from the combined sequences.

Consider any parameters of interest and predictive quantities as separate scalar summaries. At a time point, consider a single summary ω, and assume m parallel simulations, each of length n. For each of these a numerical equivalent of simply visually comparing sample traces of chains (figure 6.1) is wanted. For each scalar summary ω, label the m parallel sequences of length n as ω_{ij}, where j = 1,...,n and i = 1,...,m. Then two quantities are computed; the between-sequence variance B

$$\mathrm{B} = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\omega}_{i.} - \bar{\omega}_{..})^2,$$

$$\text{where } \bar{\omega}_{i.} = \frac{1}{n}\sum_{j=1}^{n}\bar{\omega}_{ij} \quad \text{and} \quad \bar{\omega}_{..} = \frac{1}{m}\sum_{i=1}^{m}\bar{\omega}_{i.}$$

and the within-sequence variance W

$$\mathrm{W} = \frac{1}{m}\sum_{i=1}^{m}s_i^2,$$

$$\text{where } s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\omega_{ij} - \bar{\omega}_{i.})^2.$$

The between-sequence variance B contains a factor of n because it is based on the variance of the within-sequence means, $\bar{\omega}_{i.}$. Each $\bar{\omega}_{i.}$ is an average of n values of ω_{ij}.

Using these two variance components, two estimates of the variance of ω in the posterior distribution are constructed. Firstly

$$\hat{\text{var}}(\omega) = \frac{n-1}{n}\text{W} + \frac{1}{n}\text{B}$$

is an estimate of the variance that is unbiased under stationarity, but is an overestimate under the more likely situation that the starting points are overdispersed. Therefore $\hat{\text{var}}(\omega)$ is a conservative estimate of the variance of ω under overdispersion.

On the other hand, the variance of ω is likely to be underestimated by the within-sequence variance W, for any finite n. This is because the individual sequences do not have time to range over all of the target distribution resulting in them having less variability. As n increases, $\hat{\text{var}}(\omega)$ and W approach the true value of ω, but from opposite directions.

The scale reduction factor, or Gelman-Rubin statistic, can then be calculated to monitor the convergence of the Markov chain. This involves calculating the ratio between the estimated upper and lower bounds for the standard deviation of ω, denoted by R:

$$\sqrt{\hat{\text{R}}} = \sqrt{\frac{\hat{\text{var}}(\omega)}{\text{W}}}.$$

As the scale reduction factor reduces to 1, this indicates that the parallel Markov chains are overlapping therefore a sufficient burn-in period has been reached. If the Gelman-Rubin statistic is greater than 1 then convergence hasn't been reached and further simulations need to be run.

In practice, the simulations are generally run until the values of $\hat{\text{R}}$ are less than 1.1 or 1.2 (134) for all the parameters. The $\hat{\text{R}}$ values can be examined numerically or a time series plot of them can be examined. The plots shown in Figure 6.2 are for the parameters β_1 to β_7, and show the Gelman-Rubin scale reduction factor (in red) up to 20,000 iterations. Also shown is the estimate of the variance of ω in blue and the average of the within-sequence variances, W, in

green. When the red line converges to around 1 and the green and blue lines are consistently overlapping, this indicates that the parameter has converged to stability. By 20,000 iterations it appears that convergence has not yet been reached as some of the fixed parameters' $\hat{R}$ values are still greater than 1, eg β_4 and β_6. However, by 150,000 iterations (Figure 6.3) all the parameters appear to have converged to stability. Definite convergence was evident by 200,000 iterations so this was chosen to be the burn-in period when running these models.

6.6 Further Iterations

Once a suitable burn-in period has been determined, it is then necessary to run further iterations that will then allow accurate posterior estimates to be obtained from the resulting samples. The method that was used here to determine how many iterations were needed after convergence was to assess the Monte Carlo error for each parameter. The Monte Carlo error is an estimate of the difference between the mean of the sampled values and the true posterior mean. To enable accurate posterior inference, the Monte Carlo error should be small in relation to the posterior standard deviation. A rule of thumb that has been suggested (136) is that iterations should be run until the Monte Carlo error for each parameter of interest is less than 5% of the sample standard deviation.

Table 6.1 reports the Monte Carlo error, sample standard deviation and the Monte Carlo error as a percentage of the standard deviation at 50,000 and 100,000 iterations after a burn in of 200,000. As can be seen, at 50,000 iterations all of the fixed and random terms have a percentage less than 5%. This indicates that running 50,000 iterations after convergence gives a suitable set of samples from which accurate posterior inference can be made.

Table 6.1 Monte Carlo error as percentage of posterior standard deviation

Param-eters	50,000 iterations			100,000 iterations		
	MC error	SD	*MC error as % of SD*	MC error	SD	*MC error as % of SD*
β_0	0.00522	0.1108	4.7%	0.0042	0.1067	4.0%
β_1	2.49e-06	0.0001	4.7%	2.47e-06	0.0001	4.0%
β_2	0.00003	0.0006	4.7%	0.0000	0.0007	3.9%
β_3	0.00002	0.0004	4.6%	0.0000	0.0004	3.9%
β_4	0.00014	0.0030	4.6%	0.0001	0.0036	3.9%
β_5	0.00002	0.0005	4.7%	2.13e-05	0.0005	3.9%
β_6	8.09e-08	0.0000	4.5%	6.64e-08	1.77e-06	3.8%
σ_u^2	0.00003	0.0009	3.3%	2.46e-05	0.0009	2.8%
σ_{uv}	0.00005	0.0016	3.0%	3.91e-05	0.0016	2.5%
σ_v^2	0.00018	0.0055	3.3%	0.0001	0.0054	2.5%

6.7 Model Results

As previously discussed, numerical summaries of the posterior samples are used to examine the model results. The posterior samples can be assumed to be from the desired joint posterior distribution, and each parameter has a set of samples that can be thought to have arisen from the individual marginal distributions.

6.7.1 Fixed and Random Effects Estimates

Some results from running the spatial multilevel model defined by equations (6.5) and (6.6) using MCMC are given in Table 6.2. The point estimates for each parameter from the null model and the full model are given. These are calculated by assuming the chain values are a sample from the posterior distributions and calculating the mean of these in the usual way. Credible intervals are also given for each of these point estimates, again for the null model and the full model. A credible interval is the Bayesian equivalent to the frequentist confidence interval.

Figure 6.2 Gelman-Rubin plots for fixed effects (at 20,000 iterations)

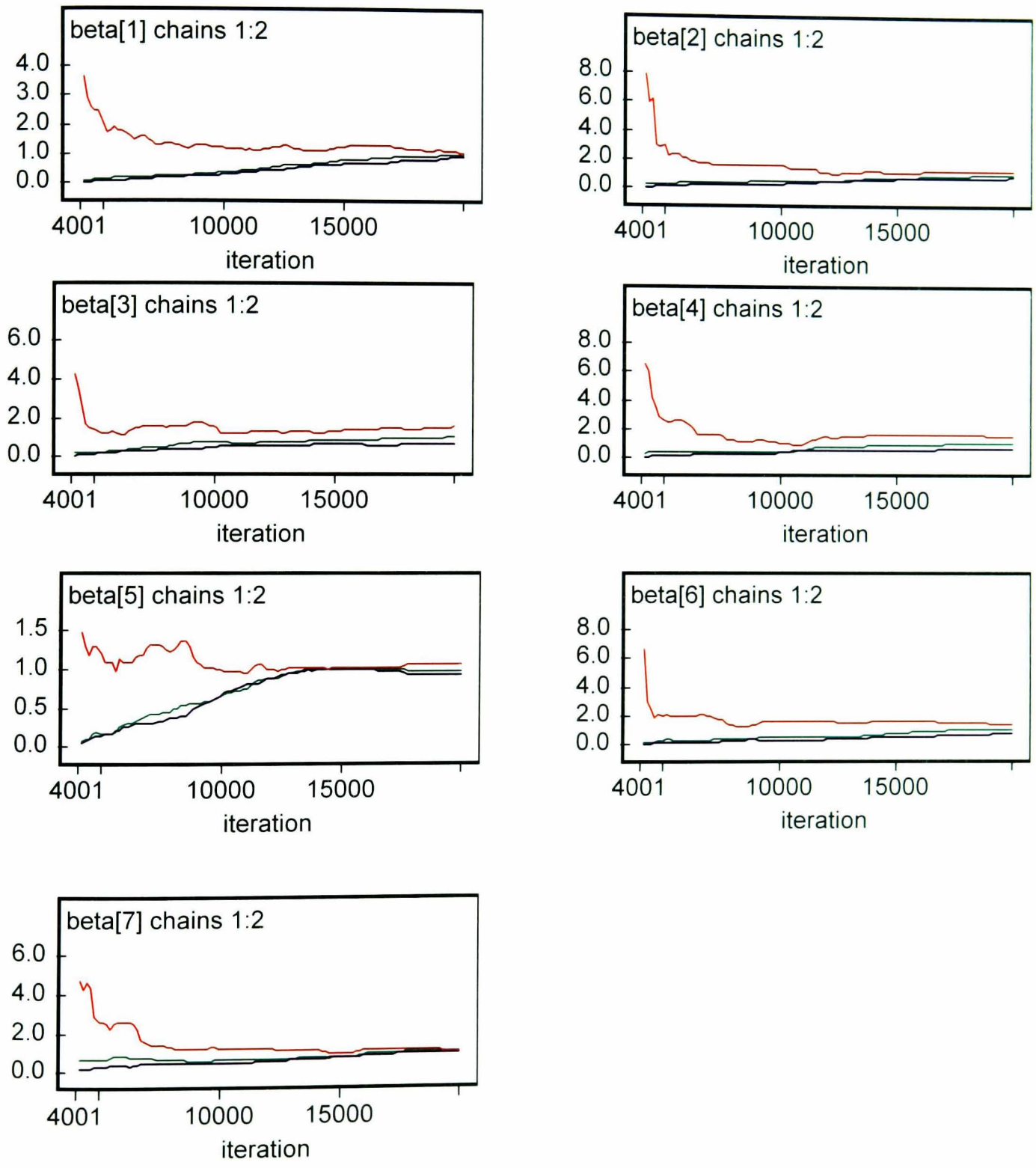
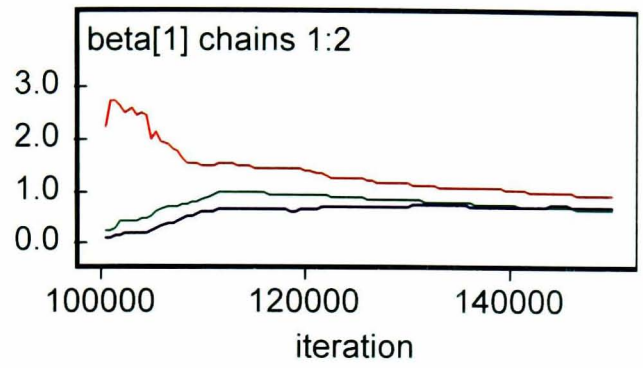
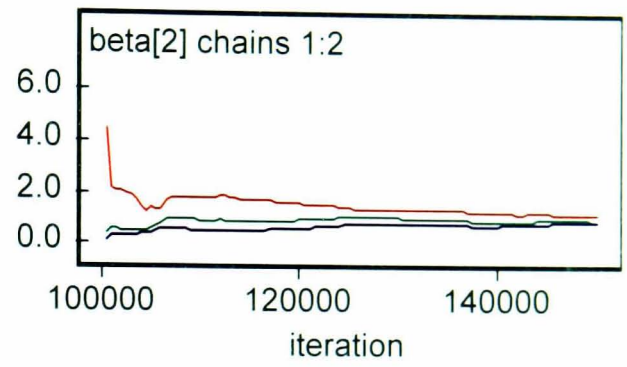
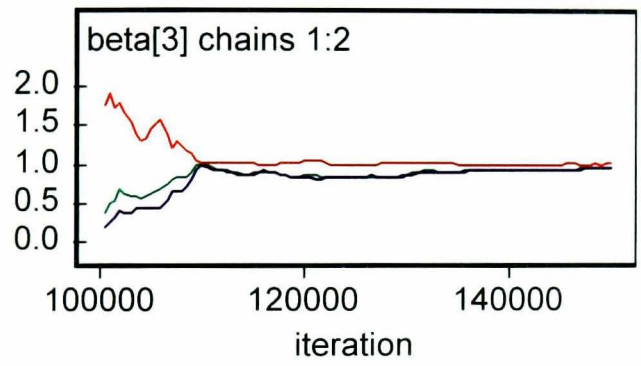
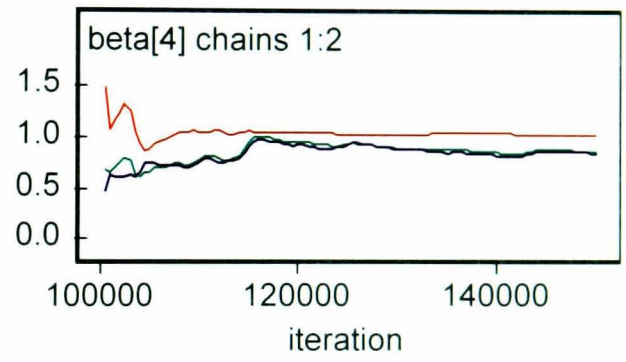


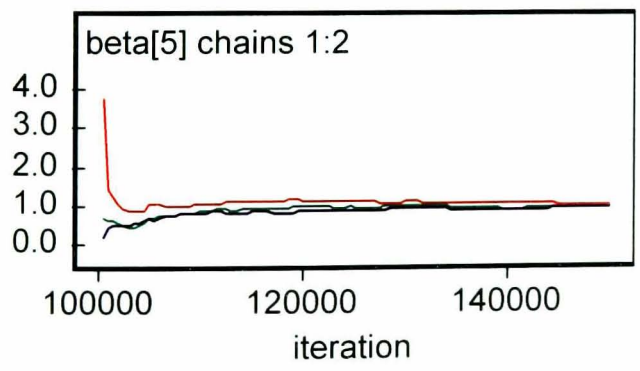
Figure 6.3 Gelman-Rubin plots for fixed effects (at 150,000 iterations)

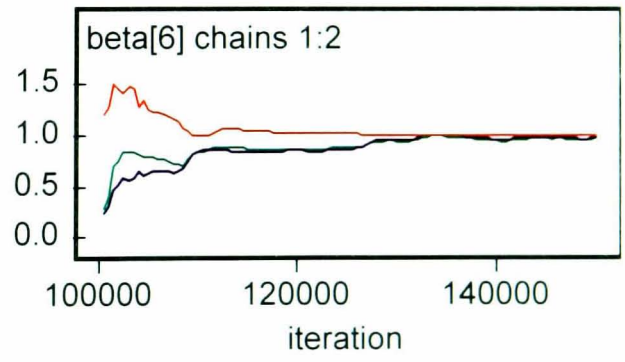


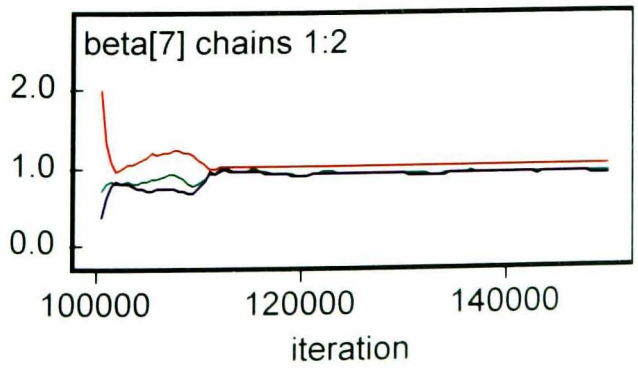












The quantiles of the samples can be used to produce credible intervals. The intervals given in Table 6.2 are 95% central Bayesian credible intervals, that is, if Q_i is the ith quantile, the given intervals for each parameter are $(Q_{0.025}, Q_{0.975})$.

The estimates are similar to those produced from the empirical Bayes method. A comparison of the estimates obtained from different modelling methods is given in Chapter 7. Interpretation of the fixed and random estimates is the same as in Chapter 5. Therefore, looking at the results from the fully Bayesian model (Table 6.2) we can see that significant covariates affecting risk of cancer mortality in Europe are fruit, vegetable, animal fat and alcohol consumption and cigarette smoking, while taking the others into account. Fruit, vegetable and alcohol consumption have an inverse association with cancer mortality rates while cigarette smoking and animal fat consumption show a positive relationship. Adding these explanatory variables has explained 50% of the variation in cancer mortality in Europe. It can be seen that 66% of the remaining variation is due to spatial patterning of disease.

Table 6.2 Results from fully Bayesian spatial multilevel models

Parameters	**Null Model**		**Full Model**	
	Estimate	**Credible Interval**	**Estimate**	**Credible Interval**
β_0	7.53	(7.49 , 7.56)	7.09	(6.92 , 7.26)
β_1 **(smoke)**		,	**0.0005**	**(0.0005 , 0.0006)**
β_2 **(fruit)**		,	**-0.0083**	**(-0.0089 , -0.0080)**
β_3 **(veg)**		,	**-0.0013**	**(-0.0016 , -0.0010)**
β_4 **(animal)**		,	**0.0293**	**(0.0225 , 0.0355)**
β_5 **(alcohol)**		,	**-0.0013**	**(-0.0020 , -0.0006)**
β_6 (Gdp)		,	-1.97E-6	(-4.33E-6 , 8.73E-7)
σ_u^2	0.0058	(0.0033 , 0.0092)	0.0035	(0.0022 , 0.0051)
σ_{uv}	0.0131	(0.0080 , 0.0194)	0.0050	(0.0023 , 0.0077)
σ_v^2	0.0593	(0.0422 , 0.0804)	0.0275	(0.0175 , 0.0378)

6.7.2 SMR Disease Map

To gain a visual picture of the results of modelling cancer mortality using full Bayes estimation, relative risks have been mapped and are given in Figures 6.5 and 6.6 (null and full model respectively). For comparison purposes, the standardised mortality ratios are mapped for each region in Figure 6.4.

Recall that there were 652,728 deaths (254 per 100,000) in total from cancer in the 187 regions under investigation in 1991. The SMRs (Figure 6.4) vary around their mean of 1.02 (sd = 0.20) from 0.58 in Ipeiros, Greece to 1.54 in Copenhagen and Frederiksberg city, Denmark. Both these regions have fairly small population sizes and numbers of deaths. The ten highest SMRs and the ten lowest SMRs are presented in Table 6.3. The standard errors are also given in this table. Note here that the maximum likelihood estimate of the area-specific relative risks, θ_i, is the standardised mortality ratio (SMR) for the ith area: $\hat{\theta}_i = \text{SMR} = \text{O}_i / \text{E}_i$ with estimated standard error $s_i = \sqrt{O_i} / E_i$. Due to the high variation in population sizes, these standard errors also have high variation. They range from 0.005 in North Rhine-Westphalia in Germany which has the largest population in the study to 0.111 in Ahvenanmaa, Finland which has the smallest population. Table 6.4 is similar but presents estimates for selected regions in order of increasing population size (ten smallest and ten largest given). 95% confidence intervals based on Poisson distribution have been computed for each SMR and are also presented in Table 6.3. Those that exclude unity have been presented in bold.

Also included in the table are the posterior mean relative risks, as modelled in equation (6.4), and 95% credible interval estimates from the fully Bayes spatial multilevel model firstly with no covariates then with all the covariates included.

Tables 6.3 and 6.4 include i, the number of the region; O_i, the observed number of deaths in the ith region; E_i, the expected number of deaths in the ith region; SMR, the standardised mortality ratio in the ith region; S_i, the standard error of SMR in the ith region; $CI_{95\%}$(SMR), confidence interval of the SMR in the ith region based on Poisson distribution; Mean, the posterior mean relative

risks from the fully Bayesian estimates of the spatial multilevel model; and $PI_{95\%}$, 95% posterior (credible) interval estimated from the spatial multilevel model.

Figure 6.4 is a map of the SMRs, with regions displayed in red indicating areas with high SMRs and those in blue having low SMRs. Despite the majority of regions' population sizes being fairly high, there are still some regions with high or low SMRs displayed on the map which are based on the least reliable data. Since these regions have lower population sizes their SMRs do not differ significantly from 1, but this is not apparent from the map. For example, regions 119 and 91 are numbered on the map of SMRs and, as can be seen from the size of these regions and from Table 6.4, they have small populations. The blue shade indicates they have relatively low SMRs, lying somewhere between 0.80 and 0.89. However, Table 6.4 shows that despite having low SMRs these regions do not differ significantly from unity, demonstrating that using SMRs alone to map mortality can be somewhat misleading. It can also be seen that region 452 has a fairly high SMR but, since it also has a low population size, the confidence interval again includes 1. The lower half of this table contains estimates for the regions with the highest populations and it can be seen some SMRs are a lot closer to 1 but are actually significantly different from 100 (ie regions 75 and 76). It is clear these estimates are more reliable but this is not evident when examining SMRs alone or maps of SMRs. Therefore, to produce more reliable estimates models are fitted that take into account extra Poisson variation, spatial patterning of the data and covariates that may be affecting the risk of mortality.

6.7.3 Relative Risk Disease Maps

The fully Bayesian estimates of relative risks are given on the maps shown in Figure 6.5 and 6.6. Figures 6.5 shows the relative risks from the null model and it is evident that they show less variability than the SMRs (Figure 6.4). They vary from 0.63 in region 321, Norte in Portugal to 1.48 in region 85, Copenhagen and Frederiksberg city in Denmark with a mean of 1.01 (sd = 0.11). It can be seen that the extreme estimates (Table 6.3) are all closer to unity after modelling the data

with the effect of smoothing the map overall. The 95% credible intervals that contain unity are in bold in Table 6.3 and it can be seen that only five of the twenty regions with extreme SMRs have RRs that are significant. These are regions 133, Picardie in France and 161, 163, 167 and 168 which are all in Greece and are numbered on the map (Figure 6.5). It can be seen that these areas with significantly high and low risks of cancer mortality also stand out on the map with many other regions taking on the purple colour (estimates close to 1). This type of modelling has smoothed the map because the mean for each region is centred on the mean of its neighbours. Therefore areas with extreme relative risks will tend to be shrunk towards a local mean, which is evident from Table 6.3.

Looking at the estimates of the relative risks from the fully Bayesian spatial model including covariates (Figure 6.6), again it can be seen that there is less variability than in the previous two maps. The relative risks now vary from 0.85 in region 168 Aegean North in Greece to 1.29 in region 327 in Madeira in Portugal with a mean of 1.00 (sd = 0.07). The map is therefore a lot more smooth with the majority of the regions being purple (ie RR range from 0.90-1.09). Since the risk and protective factors included in the model have been measured at a country level (except GDP) the reduction in variability is due to the explanatory variables explaining some of the differences between countries. Only two of the twenty regions that had the most extreme SMRs are now significantly different from unity, regions 133, Picardie in France, and 168, Aegean South in Greece. These are again pointed out in the map (Figure 6.6) and it can clearly be seen that these regions stand out as having extreme areas of risk from cancer mortality.

6.7.4 Disease Map Alternative

There is an obvious disadvantage to disease mapping in that a loss of information occurs when grouping mortality rates in ranges. An alternative that overcomes this is to plot the relative risks against a measure of latitude, south to north positioning, or a measure of longitude, east to west positioning. In doing so, exact

values of relative risks can be examined and regions and or countries can be identified by their geographical positioning.

The standardised mortality ratios and relative risks that were mapped in Figures 6.4-6.6 are plotted in Figure 6.7-6.10. Comparing Figures 6.7 and 6.8, again it is evident that the variability reduced somewhat. RRs in general move closer to unity as many extreme rates are shrunk toward a local mean based on their nearest neighbours' estimates. Figure 6.9 displays the full model estimates and it can be seen that the variability in the estimates is reduced even more, mainly between countries. This shows visually that the covariates are explaining a large amount of the variability, about 50% (see Table 6.2), at a country level. This plot is equivalent to a smoothed map and from this the extreme rates are more easily and reliably identified. Finally, Figure 6.10 is the same as Figure 6.9 but on a more natural scale. There appears to be evidence of a negative slope which in this case suggests that rates tend to be highest in the south and lowest in northern EU. Note here that, after taking into account the levels of exposure to the various risk factors, the region with the most extreme risk from cancer mortality is Maderia in Portugal (1.29). This island is very small and is not identifiable on the map (Figure 6.6). Other regions with particularly high relative risks after taking into account spatial patterning of cancer mortality and potential risk and protective factors are Copenhagen and Frederiksberg city and Nord-Pas-de-Calais. At the other end of the scale, regions that stand out with particularly low relative risks after adjusting for covariates and spatial autocorrelation are again various regions in Greece. Interestingly Norte in Portugal has a very low relative risk of cancer mortality showing that despite Portugal being a fairly small country it displays a very high amount of variability in cancer mortality rates.

Table 6.3 Estimates of SMRs/relative risks of mortality from cancer (selected regions shown, ordered by decreasing SMR)

i	Region	Country	O_i	E_i	SMR	S_i	$CI_{95\%}$(SMR)	$Mean_N$	$PI_{95\%(N)}$	$Mean_F$	$PI_{95\%(F)}$
85	Copenhagen *	Denmark	2683	1743.1	**1.54**	**0.030**	**(1.48 , 1.60)**	1.48	(0.98 , 2.27)	1.20	(0.88 , 1.72)
138	Nord-Pas-de-Calais	France	9992	6509.8	**1.53**	**0.015**	**(1.50 , 1.57)**	1.36	(0.82 , 2.32)	1.20	(0.89 , 1.76)
140	Alsace	France	3975	2762.0	**1.44**	**0.023**	**(1.39 , 1.48)**	1.17	(0.97 , 1.45)	1.08	(0.92 , 1.23)
139	Lorraine	France	5730	4008.5	**1.43**	**0.019**	**(1.39 , 1.47)**	1.22	(1.06 , 1.44)	1.12	(1.00 , 1.25)
134	Haute-Normandie	France	4253	2988.1	**1.42**	**0.022**	**(1.38 , 1.47)**	1.11	(0.94 , 1.32)	1.06	(0.92 , 1.22)
143	Bretagne	France	7689	5456.6	**1.41**	**0.016**	**(1.38 , 1.44)**	1.06	(0.79 , 1.43)	1.07	(0.85 , 1.30)
133	Picardie	France	4389	3122.0	**1.41**	**0.021**	**(1.36 , 1.45)**	**1.23**	**(1.07 , 1.41)**	**1.15**	**(1.02 , 1.27)**
137	Bourgogne	France	4570	3310.3	**1.38**	**0.020**	**(1.34 , 1.42)**	1.03	(0.87 , 1.22)	1.05	(0.91 , 1.18)
89	Vestsjaelland	Denmark	1049	771.5	**1.36**	**0.042**	**(1.28 , 1.44)**	1.21	(0.91 , 1.62)	1.04	(0.81 , 1.35)
88	Roskilde	Denmark	638	472.2	**1.35**	**0.053**	**(1.25 , 1.46)**	1.09	(0.93 , 1.29)	0.95	(0.83 , 1.08)
458	Kristianstad	Sweden	669	928.4	**0.72**	**0.028**	**(0.67 , 0.78)**	0.89	(0.73 , 1.06)	0.92	(0.80 , 1.06)
449	Blekinge	Sweden	344	478.6	**0.72**	**0.039**	**(0.64 , 0.80)**	0.87	(0.68 , 1.10)	0.94	(0.76 , 1.15)
322	Centro	Portugal	3298	4602.2	**0.72**	**0.013**	**(0.69 , 0.74)**	0.88	(0.76 , 1.03)	1.03	(0.90 , 1.19)
163	Greece West	Greece	1286	1845.5	**0.70**	**0.020**	**(0.66 , 0.74)**	**0.82**	**(0.69 , 0.97)**	0.88	(0.77 , 1.01)
167	Aegean South	Greece	434	658.0	**0.66**	**0.032**	**(0.60 , 0.72)**	**0.79**	**(0.64 , 0.98)**	0.90	(0.77 , 1.04)
164	Greece Central	Greece	1104	1722.5	**0.64**	**0.019**	**(0.60 , 0.68)**	0.89	(0.77 , 1.01)	0.92	(0.82 , 1.03)
168	Aegean North	Greece	431	678.1	**0.64**	**0.031**	**(0.58 , 0.70)**	**0.80**	**(0.66 , 0.97)**	**0.85**	**(0.74 , 0.98)**
165	Peloponese	Greece	1282	2097.2	**0.61**	**0.017**	**(0.58 , 0.65)**	1.09	(0.82 , 1.44)	1.10	(0.88 , 1.36)
169	Crete	Greece	933	1585.0	**0.59**	**0.019**	**(0.55 , 0.63)**	1.01	(0.72 , 1.39)	1.05	(0.80 , 1.38)
161	Epirus	Greece	588	1014.9	**0.58**	**0.024**	**(0.53 , 0.63)**	**0.82**	**(0.69 , 0.97)**	1.05	(0.80 , 1.38)

N – estimates from null model, F – estimates from full model *Copenhagen and Frederiksberg (city)

Table 6.4 Estimates of SMRs/relative risks of mortality from cancer (selected regions shown, ordered by increasing population)

i	Region	Country	pop	O_i	E_i	SMR	S_i	$CI_{95\%}$(SMR)	$Mean_N$	$PI_{95\%(N)}$	$Mean_F$	$PI_{95\%(F)}$
119	Ahvenanmaa	Finland	24734	58	68.6	0.85	0.111	(0.64 , 1.09)	0.93	(0.76 , 1.14)	0.96	(0.80 , 1.15)
452	Bornholm	Denmark	45616	154	138.3	1.11	0.090	(0.94 , 1.30)	0.89	(0.66 , 1.21)	0.99	(0.78 , 1.29)
91	Gotland	Sweden	57578	150	172.3	0.87	0.071	(0.74 , 1.02)	0.95	(0.74 , 1.23)	1.01	(0.83 , 1.24)
454	Powys	UK	118590	344	361.0	0.95	0.051	(0.86 , 1.06)	0.98	(0.85 , 1.13)	0.93	(0.83 , 1.04)
127	Isle of Wight	UK	126338	476	448.6	1.06	0.049	(0.97 , 1.16)	1.07	(0.84 , 1.35)	1.01	(0.85 , 1.24)
124	Jamtland	Sweden	135910	328	446.3	**0.74**	**0.041**	**(0.66 , 0.82)**	0.89	(0.72 , 1.10)	0.95	(0.80 , 1.12)
559	Blekinge	Sweden	151266	344	478.6	**0.72**	**0.039**	**(0.65 , 0.80)**	0.87	(0.68 , 1.10)	0.94	(0.76 , 1.15)
449	Pohjois-Karjala	Finland	177152	330	432.1	**0.76**	**0.042**	**(0.68 , 0.85)**	0.93	(0.76 , 1.13)	0.94	(0.80 , 1.09)
246	Kronoberg	Sweden	178961	441	554.4	**0.80**	**0.038**	**(0.72 , 0.87)**	0.86	(0.71 , 1.05)	0.95	(0.82 , 1.09)
327	Ionian Islands	Greece	194754	460	638.0	**0.72**	**0.034**	**(0.66 , 0.79)**	0.77	(0.56 , 1.08)	0.86	(0.68 , 1.10)
...	...	...	...	...	...	...	...	...	...	...	...	...
82	Saxony	Denmark	4721588	12630	12427.0	1.02	0.009	(1.00 , 1.03)	1.03	(0.85 , 1.28)	0.98	(0.87 , 1.12)
564	Scotland	UK	5107000	14876	12636.0	**1.18**	**0.010**	**(1.16 , 1.20)**	1.02	(0.88 , 1.18)	1.01	(0.89 , 1.13)
148	Rhone-Alpes	France	5418045	12075	9638.1	**1.25**	**0.010**	**(1.23 , 1.28)**	1.06	(0.91 , 1.24)	0.99	(0.87 , 1.09)
73	Hessen	Denmark	5800320	16132	15289.0	**1.06**	**0.008**	**(1.04 , 1.07)**	0.98	(0.83 , 1.15)	0.97	(0.87 , 1.08)
536	Greater London	UK	6889948	17648	16391.0	**1.08**	**0.008**	**(1.06 , 1.09)**	1.07	(0.93 , 1.24)	1.05	(0.94 , 1.19)
70	Lower Saxony	Denmark	7431517	20009	19743.0	1.01	0.007	(1.00 , 1.03)	1.02	(0.94 , 1.12)	1.01	(0.96 , 1.10)
75	Baden-Wurttemberg	Denmark	9911934	23636	24463.0	**0.97**	**0.006**	**(0.95 , 0.98)**	1.01	(0.81 , 1.26)	1.00	(0.89 , 1.14)
131	Ile de France	France	10781499	21915	17399.0	**1.26**	**0.009**	**(1.12 , 1.28)**	1.05	(0.90 , 1.22)	1.01	(0.91 , 1.11)
76	Bavaria	Denmark	11522397	28859	29274.0	**0.99**	**0.006**	**(0.97 , 1.00)**	0.94	(0.82 , 1.07)	**0.93**	**(0.88 , 0.99)**
72	North Rhine-Westphalia	Denmark	17429759	49137	44408.0	**1.11**	**0.005**	**(1.10 , 1.12)**	1.06	(0.92 , 1.24)	1.02	(0.94 , 1.12)

Figure 6.4 Map of SMRs

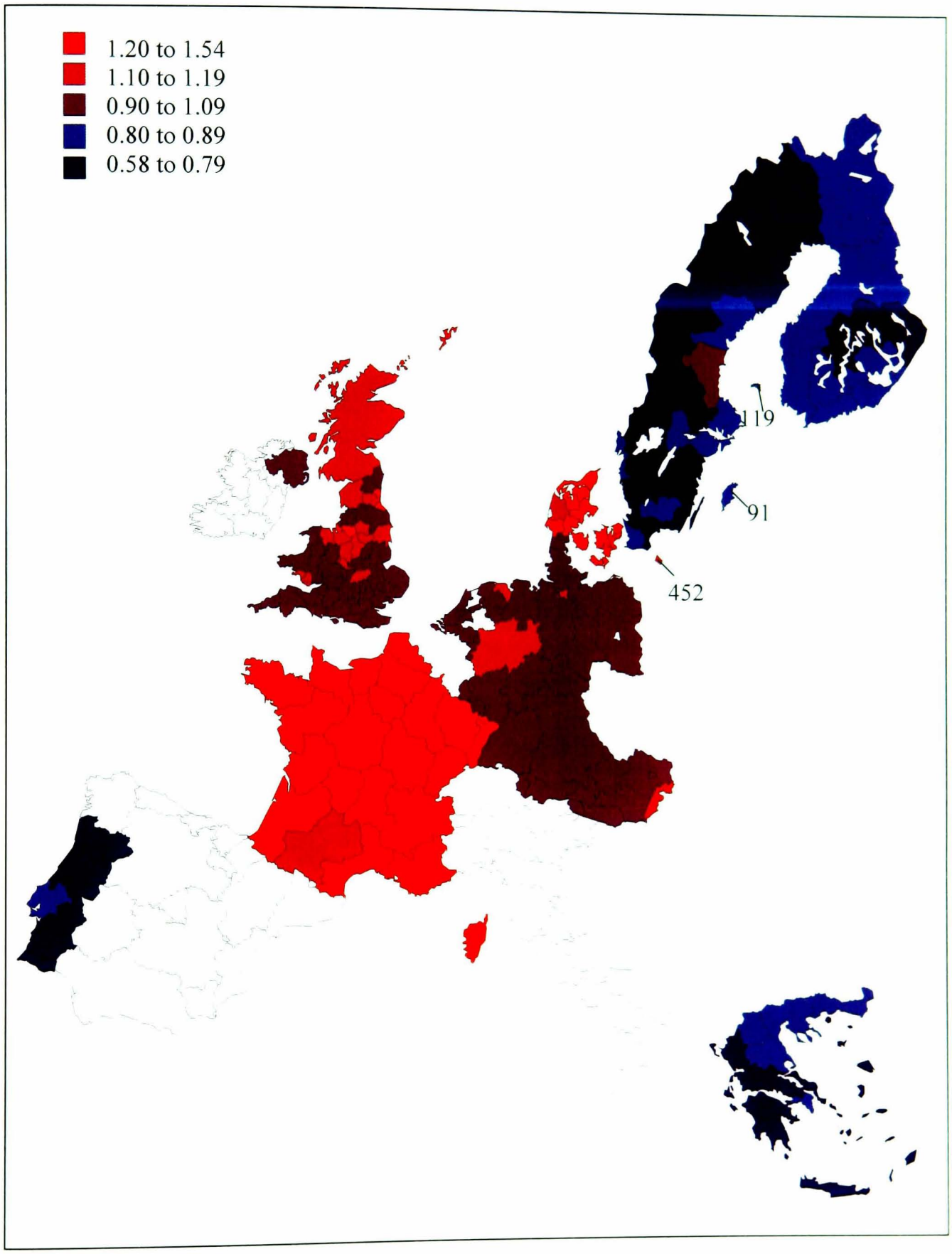


Figure 6.5 Map of RRs from fully Bayesian model (no covariates)

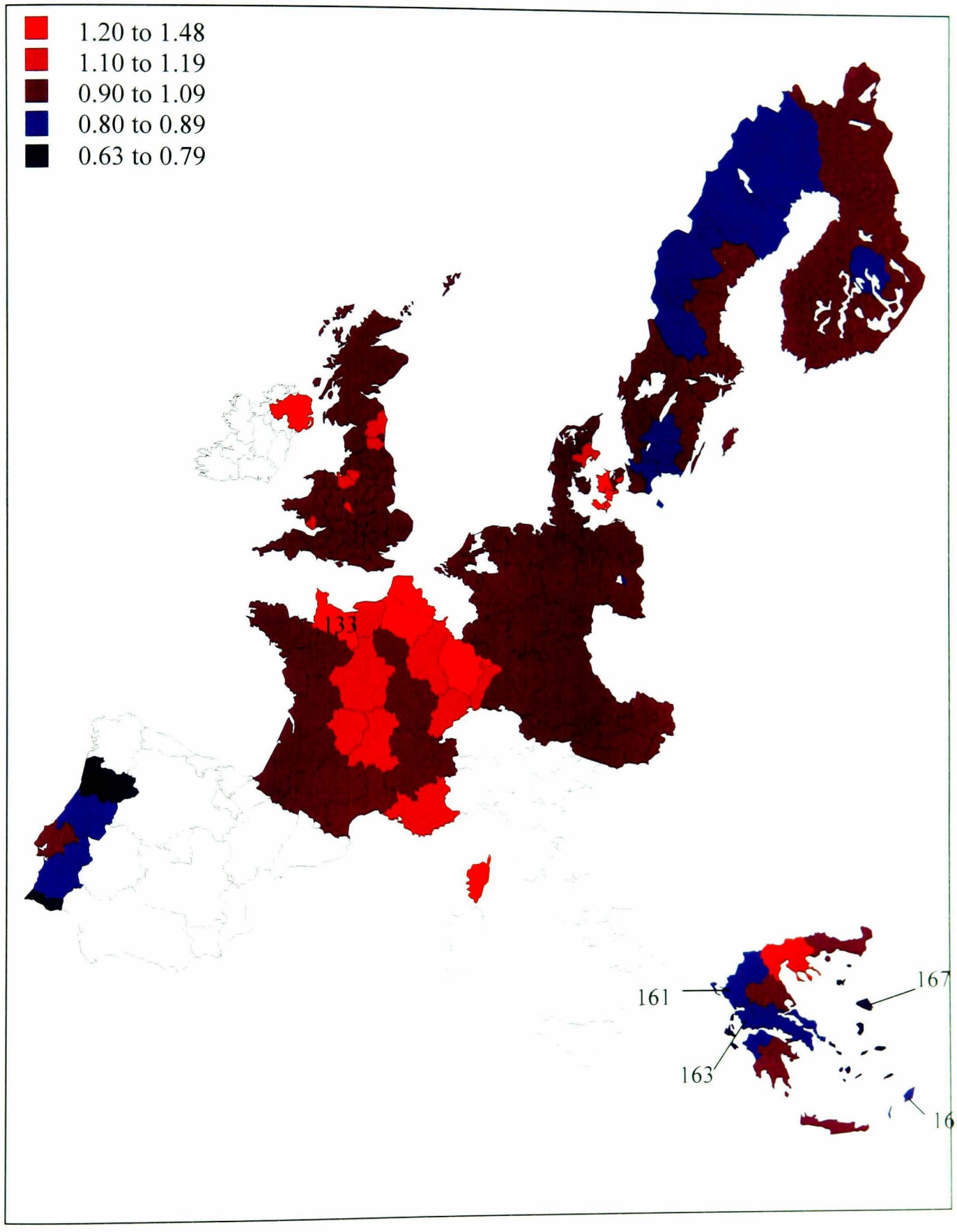


Figure 6.6 Map of RRs from fully Bayesian model (all covariates)

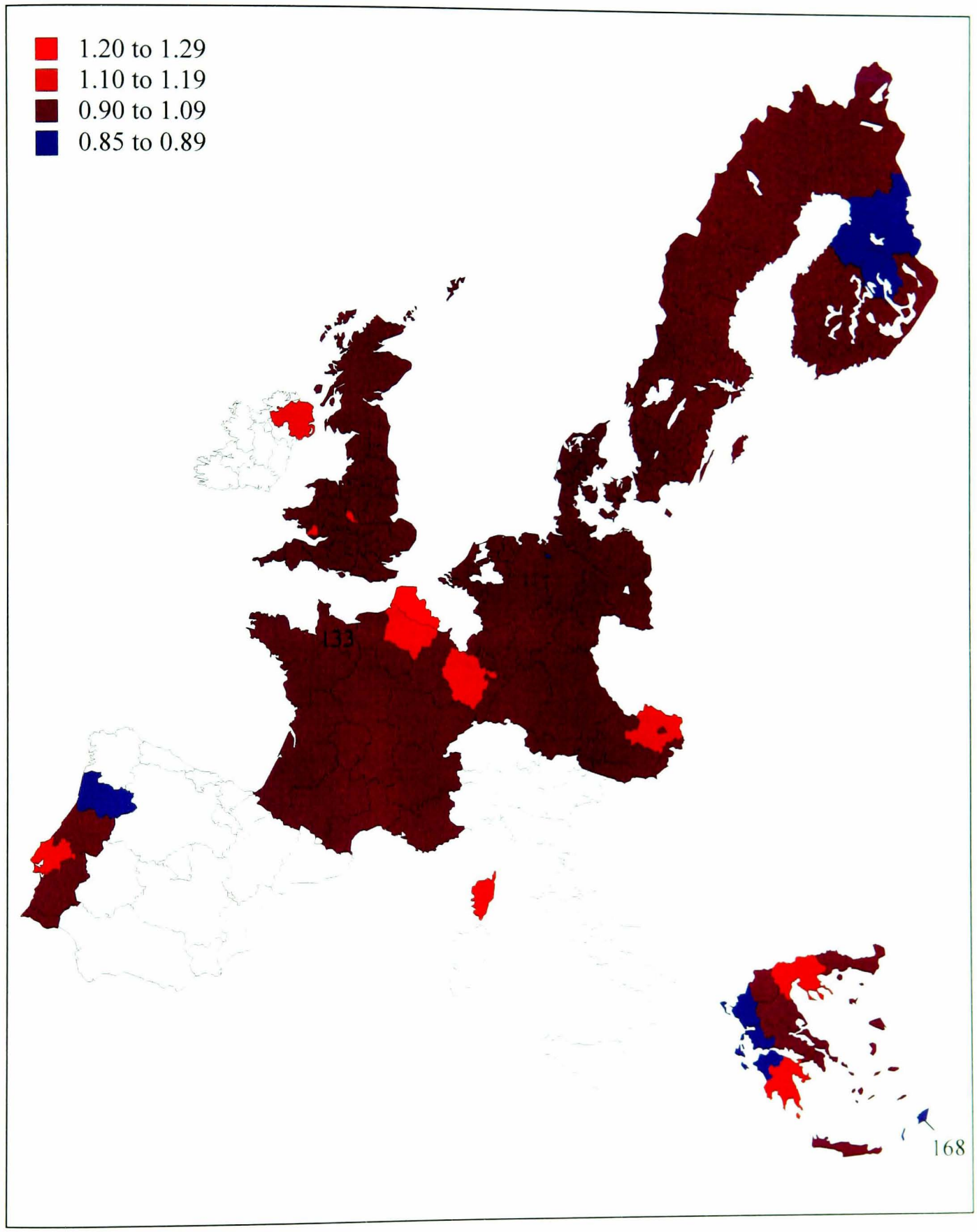


Figure 6.7 Plot of Latitude against Standardised Mortality Ratios

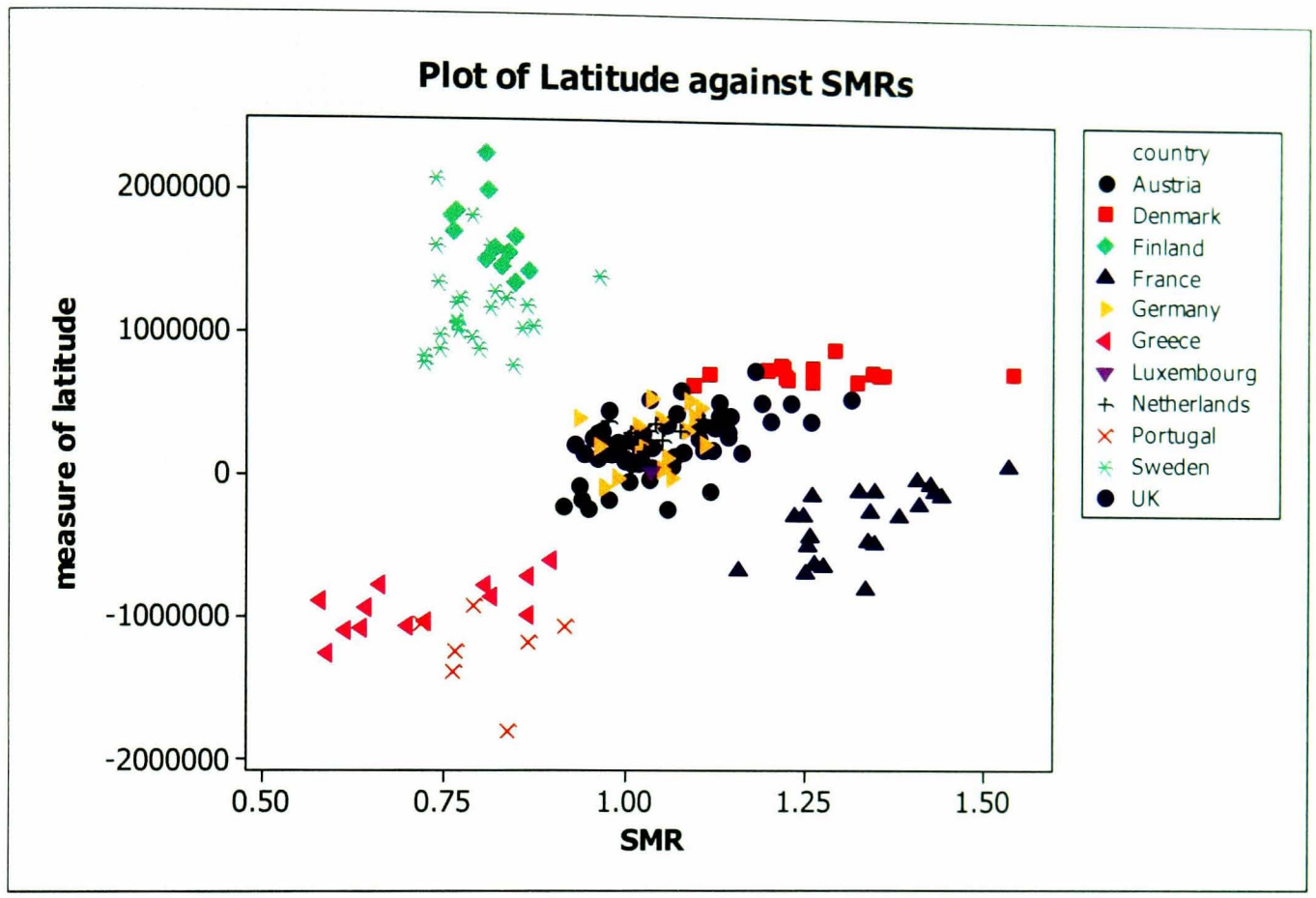


Figure 6.8 Plot of Latitude against Relative Risks from fully Bayesian null model

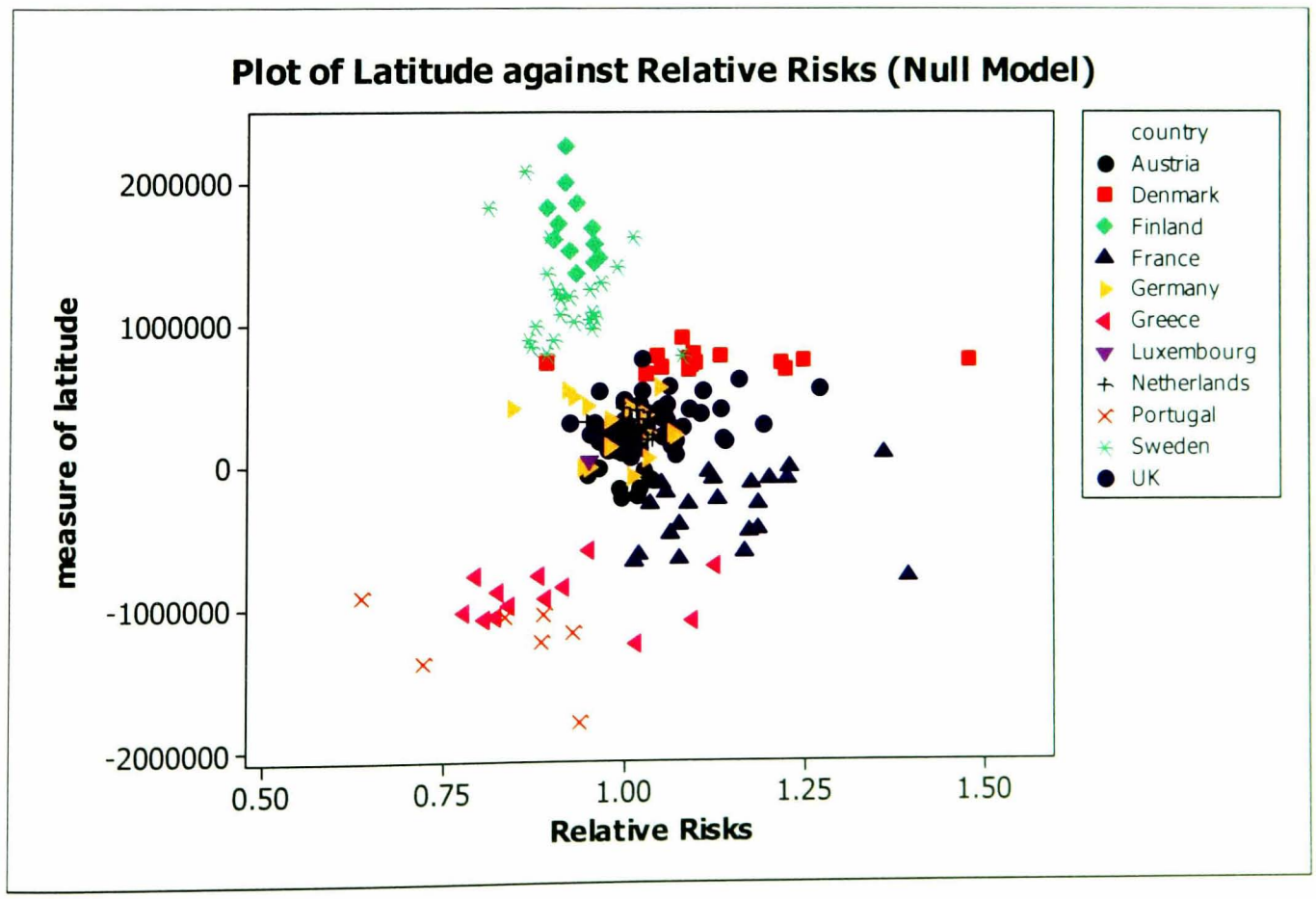


Figure 6.9 Plot of Latitude against Relative Risks from fully Bayesian full model

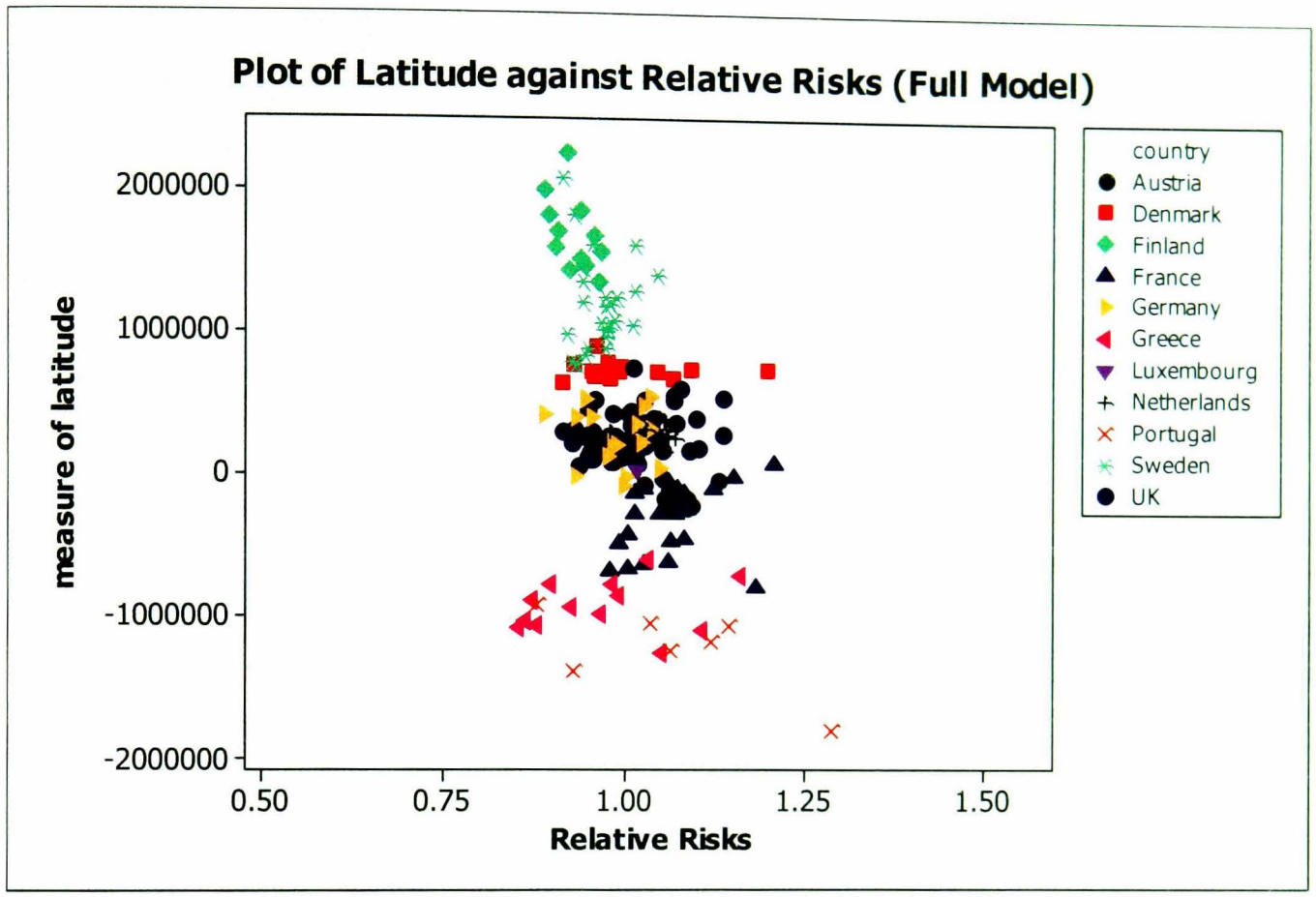
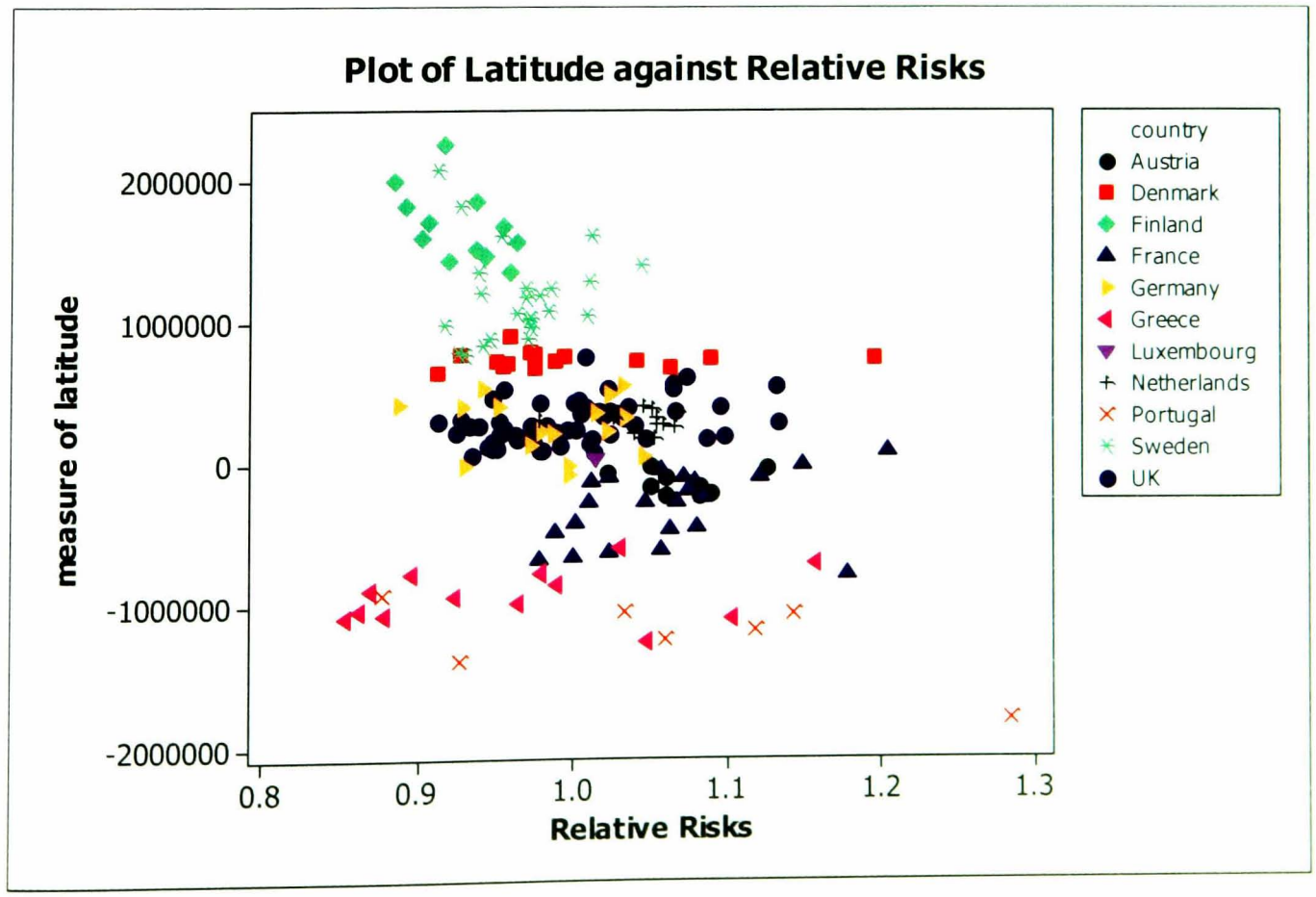


Figure 6.10 Plot of Latitude against Relative Risks from fully Bayesian full model (different scale)



6.8 Adding a Further Level

Spatial clustering effects can be incorporated into a multilevel model indirectly through the use of higher levels of geography as additional levels in the model. The European cancer mortality data set has a higher level, country, available and can be added as a third level to account for the fact that regions within a country are more likely to be homogeneous than regions from different countries.

Equations (6.5) and (6.6) can be extended to include a further level by adding another variance parameter, y_k say, into the random part of the model, where k = 1,.., 11 are the country identifiers:

$$O_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \log(\mathrm{E}_i) + \alpha + \sum_m \beta_m x_{mi} + u_{a[i]} + \sum_{j \in \partial[i]} w_{i,j} v_j + y_k .$$

The random effect y_k is assigned a normal distribution with mean zero and variance σ_y^2 :

$$\mathrm{y_k} \sim \mathrm{N}(0, \sigma_y^2) .$$

The hyper-prior for the inverse variance term, Σ_y^{-1} , is defined to be Gamma distributed:

$$\Sigma_y^{-1} \sim Gamma(0.001, 0.001) .$$

The burn-in period was determined by examining trace plots from parallel chains and the Gelman-Rubin diagnostic test. Figure 6.11 shows the Gelman-Rubin test results for the full 3 level model from iterations 450,000 to 500,000. It can be seen that the $\hat{\mathrm{R}}$ (red line) value is stabilising close to 1 and the estimates of the variance of ω (blue line) and the average of the within-sequence variances, W (green line) are consistently overlapping. Before this, some of the parameters had not yet stabilised. This indicates that all of the parameters have converged to stability and 500,000 iterations is therefore a suitable burn-in period for this model.

After a burn-in period of 500,000, the number of further iterations to run was decided by examining when the Monte Carlo error was small enough in relation to the posterior standard deviation. Table 6.5 shows that by a further 100,000 iterations the Monte Carlo error as a percentage of the sample standard deviation is less than 5% for all the parameters. Therefore running 100,000 iterations after convergence gives a suitable set of samples from which accurate posterior inference can be made.

The posterior estimates from the full model are given in Table 6.6. The results from the null model are also given. Smoking and fruit consumption are the only factors affecting the risk of cancer mortality after taking account of the non-independence between regions within the same country. This is due to the width of the posterior credible intervals being wider than they are when fitting the two-level model (see Table 6.2). The introduction of the country level random effects has resulted in the covariate being measured with more uncertainty. Comparing the null model with the full model shows that adding the covariates reduces the total variance by around 40%, which is less of a reduction that we saw in the two-level model. In the two-level model, the covariates were effectively explaining the differences between countries, which has now been taken into account, hence, reducing their usefulness slightly.

As can be seen from examining the random terms, there has been further partitioning of the variance; with σ_y^2 representing the amount of variation that is due to differences between countries. A high amount of the total variance (81%) is attributable to country level difference. This is due to the high amount of country level clustering and differences between countries that were evident from the initial disease maps. Comparing the total variance (= 0.03) to that of the two-level model (= 0.01 (see Table 6.2)) it can be seen that there has been an increase in overall variation. This is somewhat surprising as, in practice, adding a further level to multilevel models does not tend to affect the total variance very much. The effect of adding a further geographical level, whilst already accounting for the non-independence of region level risks through the spatial factor, will be discussed further in Chapter 8, after looking at further examples. It becomes evident that it is

likely that the three-level model is mainly describing the differences between countries, hence the increase in total variance, and that the two-level model is more useful for describing the differences between regions and the affects covariates have on these.

Table 6.7 gives selected region's relative risk of cancer mortality. Comparing the two-level null model with the three-level null model shows some changes in the estimates; adding a higher country level has pulled the relative risks towards the country's overall relative risk. Some of the estimated relative risks change again after adjusting for the covariates; this is expected as the risk of cancer mortality is likely to change in some manner if the risk factors are having a significant effect. The width of the posterior credible intervals increases somewhat after the addition of covariates. The introduction of country-level random effects has resulted in the covariates being measured with more error which in turn has resulted in introducing uncertainty into the residuals and hence, the relative risks. This will be explored further in Chapter 8.

Figure 6.12, the map of relative risks from the three-level model with no covariates, and Figure 6.14, the corresponding plot against latitude, shows a very high amount country level clustering. It is likely to reflect the true distribution of the country level risk of cancer mortality (before adjusting for risk factors). Figure 6.13 shows the map of relative risks after adjusting for risk factors, and the corresponding plot is given in Figure 6.15, and it can be seen that country level clustering is still evident but to a lesser degree. Finland clearly have the lowest relative risk of cancer mortality after adjusting for all covariates and the spatial patterning of the disease. Some within country variation is evident across the map and regions in France and Austria are standing out as disease 'hotspots'. Other factors, which have not been taken into account in these models, are likely to be affecting these areas; such as health care provision (cancer treatments or availability of screening) or differing levels of the modelled risk factors within the countries.

Table 6.5 Monte Carlo error as percentage of posterior standard deviation

Parameters	100,000 iterations		
	MC error	**Posterior SD**	***MC error as % of SD***
β_0	0.3099	0.01395	4.5%
β_1	0.00016	0.00001	4.5%
β_2	0.00213	0.0001	4.5%
β_3	0.00109	0.00005	4.4%
β_4	0.01361	0.00061	4.5%
β_5	0.00209	0.00009	4.5%
β_6	1.60E-06	4.50E-08	2.9%
σ_u^2	0.00059	0.00001	1.2%
σ_{uv}	0.00087	0.00001	1.0%
σ_v^2	0.00281	0.00004	1.4%
σ_y^2	0.02018	0.00056	2.8%

Table 6.6 Results from fully Bayesian spatial multilevel models (three levels)

Parameters		Null Model		Full Model	
		Estimate	**Credible Interval**	**Estimate**	**Credible Interval**
β_0		7.54	(7.42 , 7.63)	6.93	(6.41 , 7.51)
β_1	**(smoke)**			**0.0004**	**(0.0002 , 0.0007)**
β_2	**(fruit)**			**-0.0065**	**(-0.0108 , -0.0020)**
β_3	(veg)			-0.0011	(-0.0029 , 0.0014)
β_4	(animal)			0.0216	(-0.0136 , 0.0462)
β_5	(alcohol)			0.0008	(-0.0041 , 0.0058)
β_6	(gdp)			3.1E-7	(-2.7E-6 , 3.4E-6)
σ_u^2		0.0028	(0.0018 , 0.0042)	0.0029	(0.0019 , 0.0042)
σ_{uv}		0.0009	(-0.0007 , 0.0027)	0.0011	(-0.0005 , 0.0029)
σ_v^2		0.0110	(0.0067 , 0.0173)	0.0108	(0.0061 , 0.0172)
σ_y^2		0.0446	(0.0174 , 0.1094)	0.0246	(0.0061 , 0.0775)

Figure 6.11 Gelman-Rubin plots at 500,000 iterations (note that only every tenth iteration is stored)

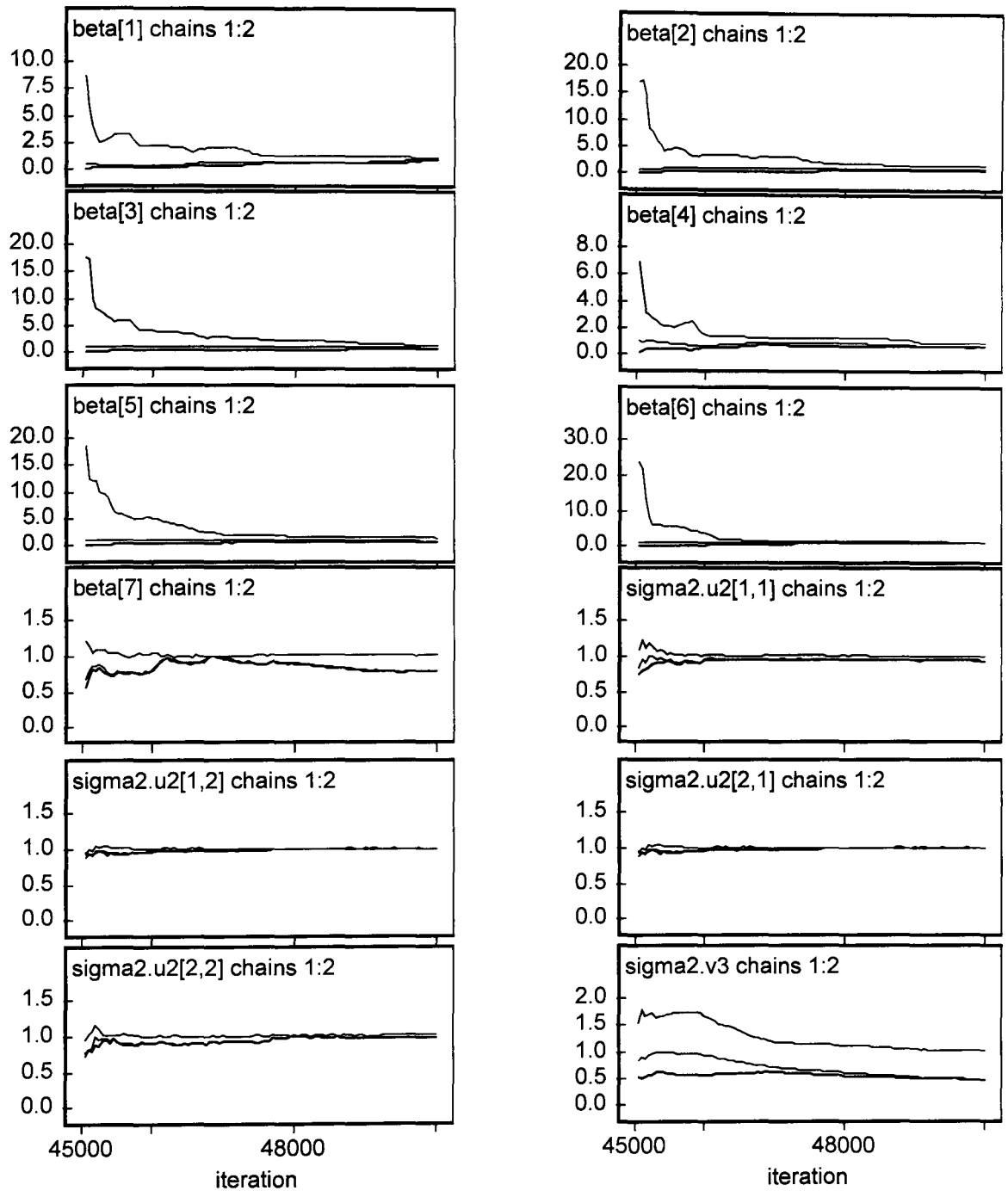


Table 6.7 Estimates of relative risks of mortality from cancer (selected regions shown, ordered by decreasing RR from three-level full model)

			Two-level null model		Three-level null model		Three-level full model	
i	**Region**	**Country**	$\mathbf{RR_{mean}}$	$\mathbf{PI_{95\%}}$	$\mathbf{RR_{mean}}$	$\mathbf{PI_{95\%}}$	$\mathbf{RR_{mean}}$	$\mathbf{PI_{95\%}}$
2	Burgenland	Austria	1.03	(0.76 , 1.40)	1.02	(0.75 , 1.40)	1.22	(0.90 , 1.73)
138	Nord-Pas-de-Calais	France	1.36	(0.82 , 2.32)	1.44	(0.98 , 2.17)	1.21	(0.80 , 1.74)
133	Picardie	France	**1.23**	**(1.07 , 1.41)**	**1.40**	**(1.14 , 1.76)**	1.21	(0.96 , 1.45)
3	Carinthia	Austria	0.99	(0.81 , 1.21)	1.01	(0.79 , 1.30)	1.21	(0.95 , 1.59)
4	Lower Austria	Austria	1.02	(0.88 , 1.18)	1.00	(0.79 , 1.26)	1.19	(0.95 , 1.54)
139	Lorraine	France	**1.22**	**(1.06 , 1.44)**	**1.39**	**(1.13 , 1.72)**	1.19	(0.93 , 1.43)
158	Macedonia Central	Greece	1.12	(0.90 , 1.39)	0.88	(0.68 , 1.17)	1.18	(0.92 , 1.52)
244	Overijssel	Netherlands	1.03	(0.90 , 1.19)	1.04	(0.85 , 1.31)	1.18	(0.99 , 1.38)
241	Groningen	Netherlands	1.00	(0.79 , 1.25)	1.04	(0.81 , 1.38)	1.18	(0.95 , 1.45)
327	Madeira	Portugal	0.94	(0.72 , 1.22)	0.88	(0.61 , 1.29)	1.17	(0.82 , 1.70)
...	...	...	...	...	...	...	...	...
123	Kymi	Finland	0.95	(0.75 , 1.20)	0.83	(0.64 , 1.10)	0.82	(0.61 , 1.07)
119	Ahvenanmaa	Finland	0.93	(0.76 , 1.14)	0.83	(0.63 , 1.09)	0.81	(0.61 , 1.06)
129	Uusimaa	Finland	0.96	(0.80 , 1.14)	0.83	(0.66 , 1.06)	0.81	(0.62 , 1.02)
120	Hame	Finland	0.92	(0.77 , 1.10)	0.81	(0.65 , 1.04)	0.79	(0.61 , 1.00)
126	Oulu	Finland	0.91	(0.76 , 1.09)	0.81	(0.64 , 1.06)	0.79	(0.60 , 1.02)
124	Lappi	Finland	0.91	(0.70 , 1.20)	0.81	(0.60 , 1.11)	0.79	(0.57 , 1.07)
125	Mikkeli	Finland	0.90	(0.76 , 1.06)	0.81	(0.64 , 1.04)	0.79	(0.61 , 1.01)
127	Pohjois-Karjala	Finland	0.93	(0.76 , 1.13)	0.81	(0.63 , 1.06)	0.79	(0.59 , 1.02)
121	Keski-Suomi	Finland	0.90	(0.76 , 1.07)	0.79	(0.63 , 1.02)	**0.78**	**(0.59 , 0.99)**
122	Kuopio	Finland	0.89	(0.73 , 1.09)	0.79	(0.61 , 1.03)	**0.77**	**(0.59 , 0.99)**

Figure 6.12 Map of RRs from fully Bayesian model (three-level, no covariates)

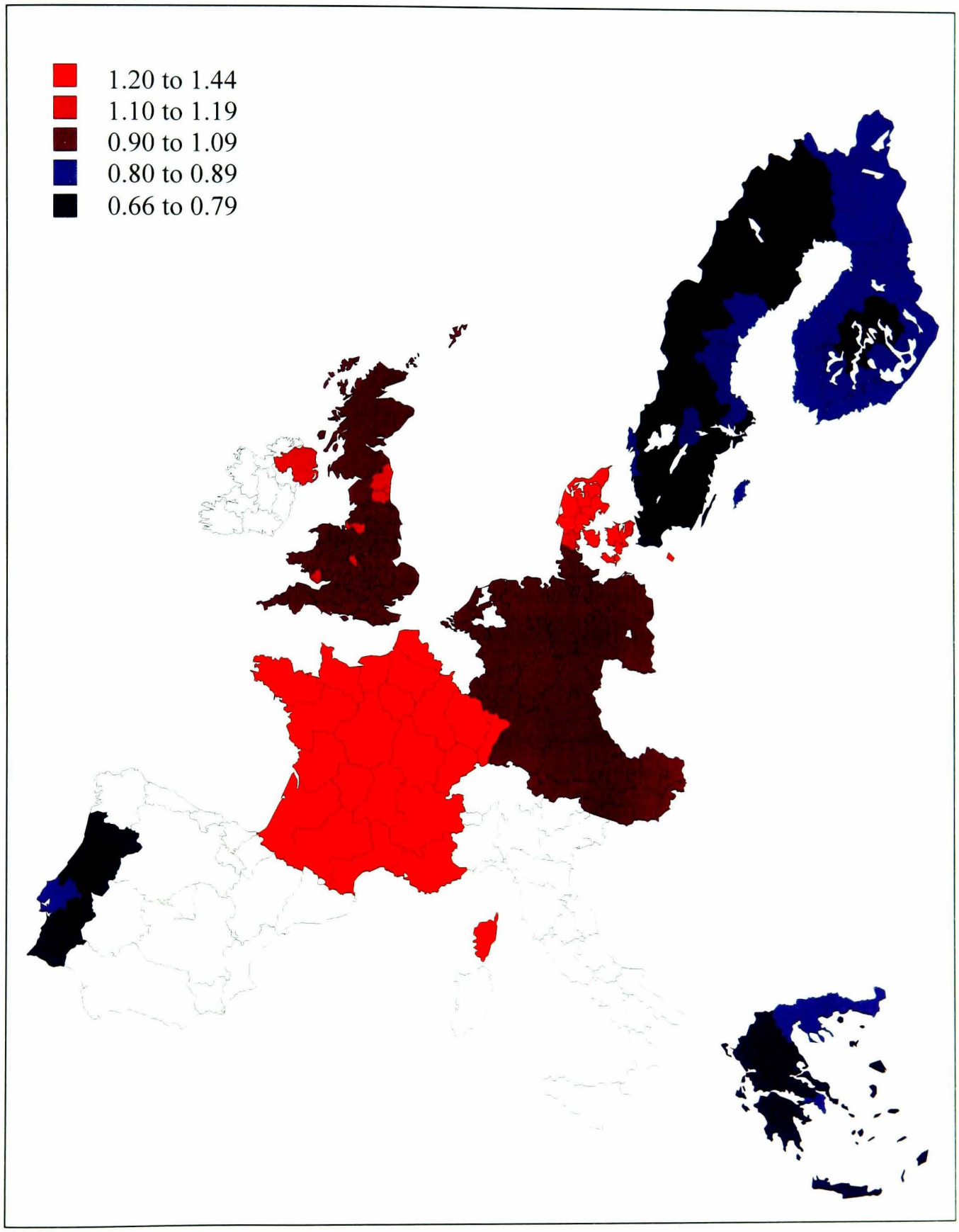


Figure 6.13 Map of RRs from fully Bayesian model (three-level, all covariates)

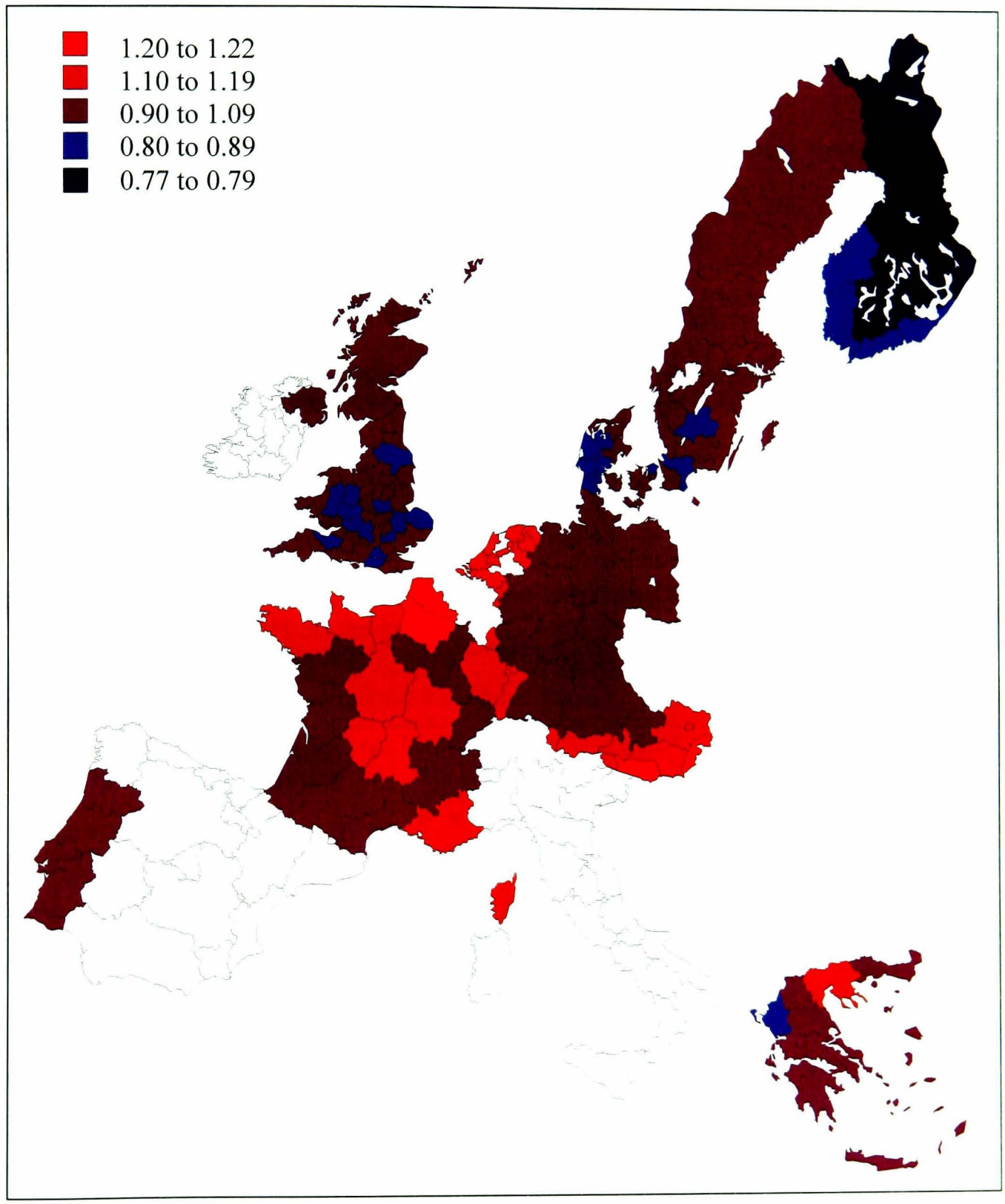


Figure 6.14 Plot of Latitude against Relative Risks from fully Bayesian model (three-level, no covariates)

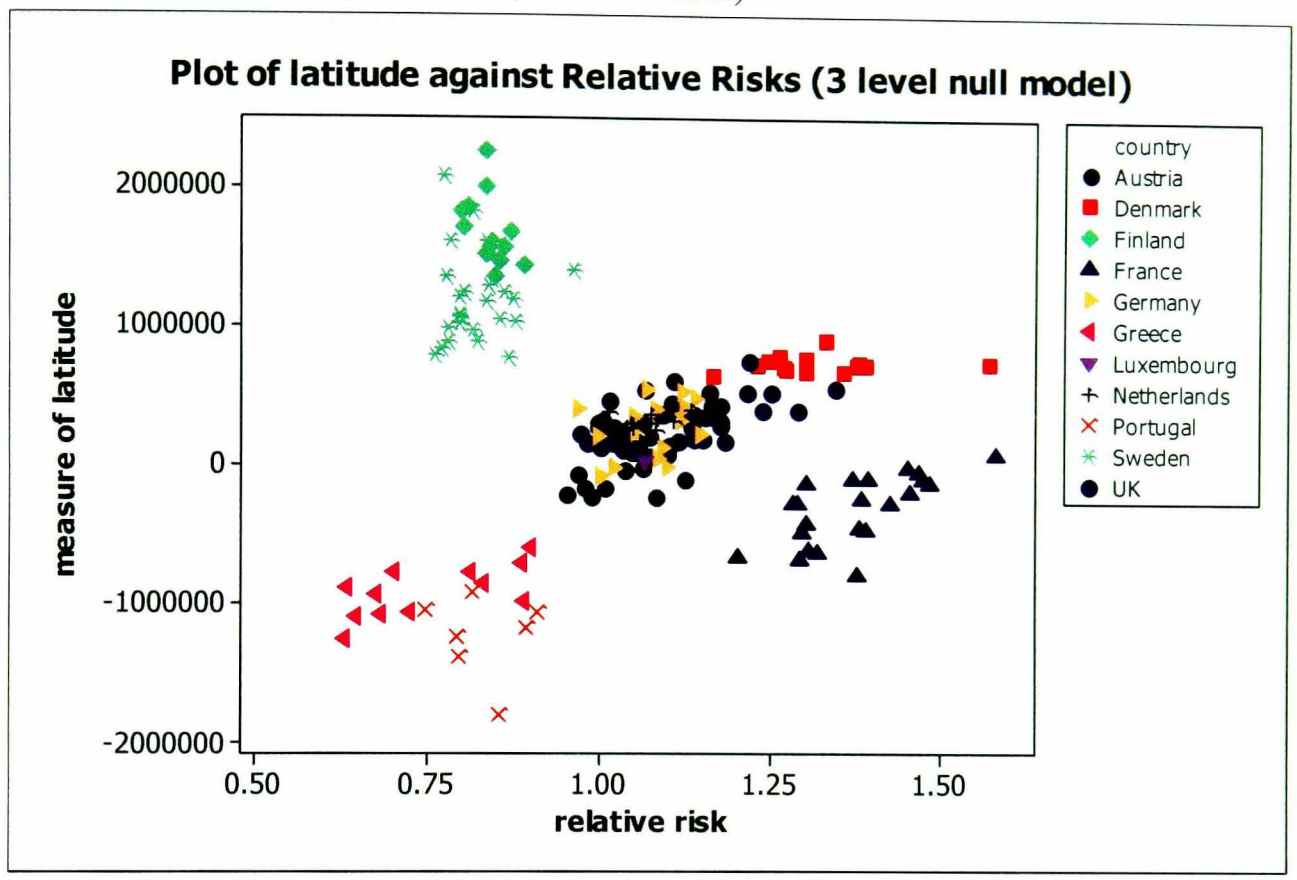
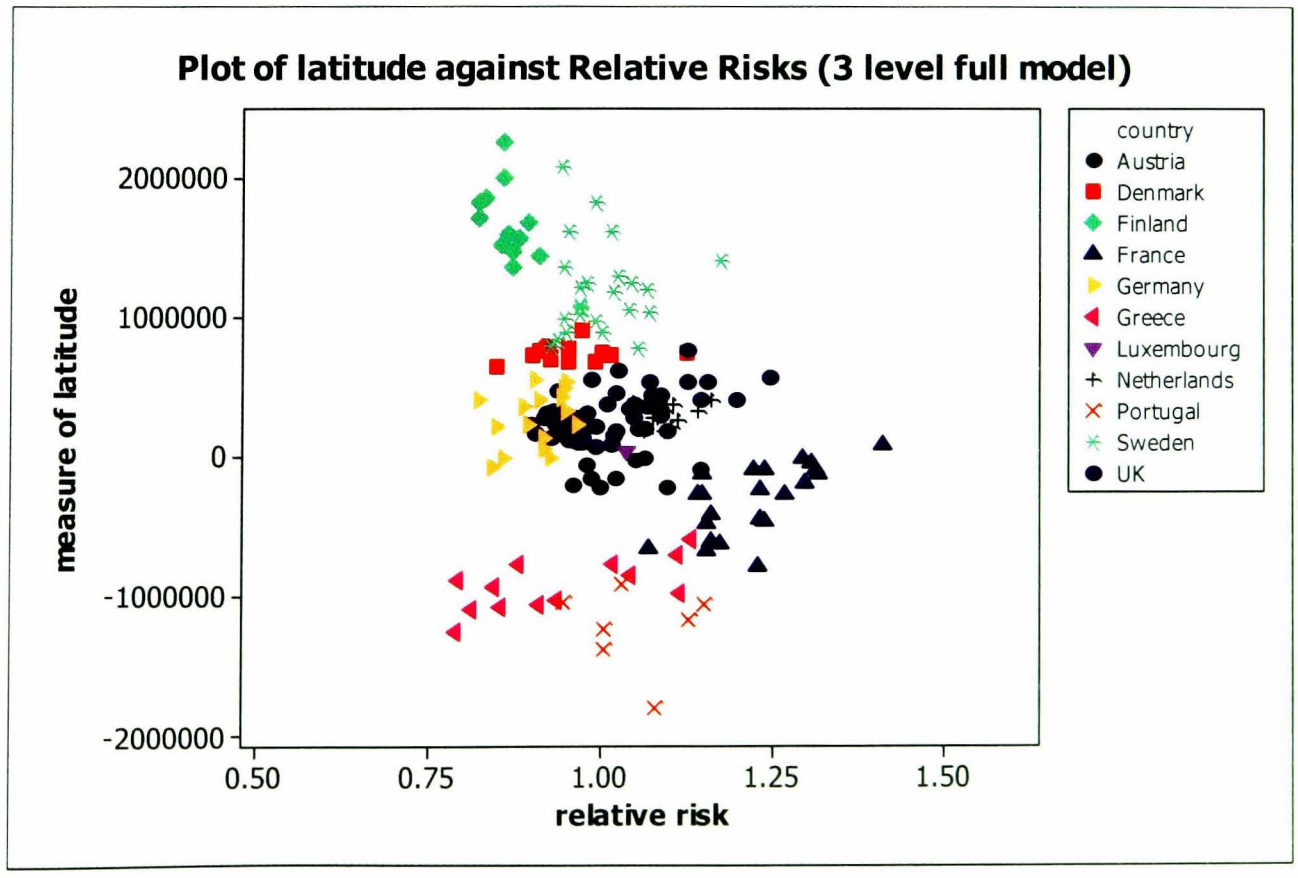


Figure 6.15 Plot of Latitude against Relative Risks from fully Bayesian model (three-level, all covariates)



Chapter 7

7 Spatial Model Comparisons

7.1 Spatial Models

It has been shown that modelling the underlying spatial structure to produce smoothed disease maps has its advantages. An effective way to do so is through the use of random effects modelling and so far a spatial multilevel model allowing for correlation between regional and spatial random effects has been explored. To fit these models it is necessary to use either iterative procedures, such as iterative generalised least squares (IGLS (124)), or simulation based methods, such as Gibbs sampling (137). MCMC simulation based methods have been becoming more common in recent years due to the increasing ability and speed of computers. These methods are popular for fitting complex models and have been used to develop multilevel models with a complex underlying structure.

A set of multilevel models whose complex structure does not fit into the standard multilevel framework are multiple membership models. These models do not strictly follow the nested structure of the data in multilevel models but can be fitted using IGLS methods with constraints (138). However, Browne at al (139) discuss the problems associated with fitting such models and how using MCMC methods proves very useful. Fitting such models in a spatial disease context will be compared with the models fitted in Chapters 5 and 6.

Also, a spatial model commonly used for disease outcomes is the conditional autoregressive (CAR) model. This will also be fitted to the EU cancer mortality dataset and will be compared with the various spatial multilevel model structures.

7.2 Multiple Membership Model

A multiple membership model is an extension to the standard multilevel framework and considers the case when the lowest level unit is a member of more than one higher classification unit. This family of models was first used by Hill and Goldstein (140) and further developed by Browne et al (139). They are actually part of the larger family of generalised linear mixed models (GLMM); these are a combination of the linear mixed model (141) and the generalised linear model framework (142). An extension to these models (139) is to use them in a spatial context (5) by considering the areas as one classification and the neighbours as another multiple-membership classification; this is known as a multiple-membership multiple-classification (MMMC) model. This is very similar to the spatial multilevel model described previously but does not allow correlation between the area and neighbour residuals. Using the notation developed by Browne et al (139), modelling the observed counts of cancer deaths in 1991 for 187 regions with known neighbourhood structure can be written as

$$\begin{aligned} y_i &\sim Poisson(\lambda_i), \\ \log(\lambda_i) &= \log(E_i) + \beta_o + X_i\beta_1 + \ldots + X_i\beta_k \\ &\quad + u^{(2)}_{a[i]} + \sum_{j\in\partial[i]} w^{(3)}_{i,j} u^{(3)}_j, \end{aligned} \tag{7.1}$$

$$u^{(2)}_j \sim N(0, \sigma^2_{u(2)}), \;\; u^{(3)}_j \sim N(0, \sigma^2_{u(3)}).$$

In (7.1), y_i is the count of cancer deaths for the ith region in the dataset; a[i] is the area from which the observed count was taken, and $\partial[i]$ is the set of neighbouring areas to the area from which the count was taken. β is a vector of fixed effect parameters, and $u_i^{(2)}$, $u_i^{(3)}$ are the vectors of residuals for random effects for classifications 2 (area) and 3 (neighbours) respectively.

Similar to the spatial multilevel model described in Chapters 5 and 6, the observed counts of cancer deaths are affected by k covariates, the area where the

counts came from and the neighbouring areas. The weights used in this model are such that

$$\sum_{j\in\partial[i]} w_{i,j}^{(3)} = 1 \ \forall\, i,$$

and all neighbours are given equal weights so that

$$w_{i,j}^{(3)} = \frac{1}{n_i},$$

where n_i is the number of neighbours to $a[i]$.

The multiple-membership multiple-classification model was fitted using MCMC methods; the prior distributions for the fixed effect parameters are flat,

$$\beta_k \sim U(),$$

for the inverse of the classification 3 variance term ($\sum_{u(3)}^{-1}$), a Gamma prior was allocated:

$$\sum_{u(3)}^{-1} \sim Gamma(0.001,0.001),$$

and similarly for the classification 2 variance matrix:

$$\sum_{u(2)}^{-1} \sim Gamma(0.001,0.001).$$

The MCMC algorithm that is used to fit the model is based on a combination of univariate Metropolis Hastings (MH) steps and Gibbs steps and has been implemented in MLwiN (143). The model can also be run in WinBUGS (7) and the code to do so is given in Appendix A2.1.

7.3 Conditional Autoregressive Model

A set of Bayesian spatial models that have been frequently used to model disease counts are based on the conditional autoregressive (CAR) prior (110); these were

described in more detail in Chapter 4 (see equations (4.18) to (4.22)). The version of this model appropriate for the EU cancer dataset can be written as follows:

$$\begin{gathered} y_i \sim Poisson\,(\lambda_i), \\ \log(\lambda_i) = \log(E_i) + X_i\beta_1 + \ldots + X_i\beta_k + u_i + v_i, \\ u_i \sim N(0,\sigma_u^2),\ \ v_i \sim N(\bar{v}_i, \sigma_v^2/n_i) \\ \text{where } \bar{v}_i = \sum_{j\in\partial(i)} v_j/n_i\,. \end{gathered} \tag{7.2}$$

Again, n_i is the number of neighbouring regions for region i. The same priors are used to fit the CAR model using MCMC methods as were used in the MMMC model. As can be seen the model is similar to the MMMC model as it has two sets of random effects; the difference is that spatial correlation is achieved through the variance structure instead of through the multiple membership relationship resulting in the neighbourhood random effects not being independent. Whilst the MMMC model has $r_{a[i]}$ random effects for each observation, where $r_{a[i]}$ is the number of neighbours for region i, the CAR model has one random effect for each observation. These random effects have the average of the surrounding random effects as their expected value. The CAR distribution is improper, and in order to produce a model that has a unique solution, a constraint has to be made on the model. This model was fitted in MLwiN and the common procedure when fitting the model in the package is to remove the intercept term to make the model identifiable (5). The code for running this model in WinBUGS is given in Appendix A2.2.

7.4 Model Comparisons

It is of interest to compare the spatial multilevel model with correlation between random effects with the MMMC model and the CAR model. The results from fitting these models using MCMC methods will be presented along with results of fitting a variance components and spatial multilevel model using quasi-likelihood methods in a frequentist setting, as described in Chapter 5.

7.4.1 Parameter estimates

For an initial comparison, the parameter estimates for each model are given in Table 7.1. The estimates along with the respective confidence or posterior credible intervals are presented for the five different models:

Model A: the variance components model fitted using empirical Bayes methods

Model B: the spatial multilevel model with correlated residuals fitted using empirical Bayes

Model C: the spatial multilevel model with correlated residuals fitted using fully Bayesian techniques

Model D: the multiple-membership multiple-classification model fitted using fully Bayesian methods

Model E: the conditional autoregressive model fitted by fully Bayesian methods.

The significant variables in each model are presented in bold and it can be seen that the significant fixed parameter estimates are similar in models A to D, with the only discrepancy being that β_3 is just non-significant in model B. Model E, however, shows differences in the magnitudes and signs of some of the fixed parameter estimates; β_2 has a lesser effect on cancer mortality than in the other models and β_3 has the opposite effect. As expected, the confidence intervals for the fixed parameters in model E cover different values than in the case of the other models. They are also generally wider, covering a larger range of values than all the other models. The confidence intervals for models A to D cover much of the same values. When comparing the intervals for models A and B, the intervals are about double the width in model B; since model A has ignored the positive autocorrelation that has been shown to exist, it is likely that incorrect inference has occurred; in particular standard errors will have been underestimated. Models C and D have attached similar errors for the fixed parameters; one would expect these models to produce the most similar results since these models are very similar in principle. The errors or ranges in the confidence intervals in these two models are also smaller than in the empirical Bayes estimated model B. Since models B and C are exactly the same except for the method of model fitting, it

suggests that the fully Bayesian method is more precise in estimating the parameters.

If we look at the random parameters, models C and D again are in close agreement both in estimates and confidence intervals. However, the random parameter estimates of models B and C are very different. Model B has a larger total variance and also attributes a higher percentage of this to the spatial effects, σ_v^2. Overall there appears to be more variability when fitting model B, as, along with a higher total variance, the confidence intervals attached to these estimates are also wider. In addition to this, a negative estimate is suggested as a plausible value for σ_u^2. This is obviously not plausible and possibly shows a fault in this model-fitting procedure which allows such estimates to be produced. Finally, it would appear that model E has less total variation and has less variation attributable to spatial heterogeneity; however, it should be noted that the variance terms are not directly comparable due to the different use of weights. The variance terms are not actually very different from those in models C and D, after taking account of the average number of neighbours possessed by a region.

It should be noted that, for models A and B, the confidence intervals are obtained from the estimate ± 1.96 × standard error. The model estimation that is being used for these models is based on an iterative generalised least squares (IGLS) framework which uses penalised quasilikelihood (PQL) estimation to approximate the Poisson distribution (6). This may lead to inaccuracy in the estimation of standard errors resulting in the confidence intervals not being estimated well which may account for some of the discrepancies between the width of the model intervals.

7.4.2 Residuals

The set of models that have been fitted can also be compared by examining residuals. This method is used in disease mapping as local goodness of fit

measures to help assess how well the model fits the data (5). The composite residuals can be written as

$$r_i = u_i + \sum_j w_{ij} u_j \tag{7.3}$$

for models B, C and D and

$$r_i = u_i + v_i \tag{7.4}$$

for model E. To make these residuals comparable to the other models, $\bar{v}_i$ can be subtracted from (7.4):

$$r_i = u_i + (v_i - \bar{v}_i)\,. \tag{7.5}$$

Model A, the variance components model, has one set of residuals representing heterogeneity effects. Composite residuals are calculated for model B incorporating the heterogeneity residuals and the spatial residuals taking account of the number of neighbours each region has. In the Bayesian settings the residuals were simply formed at each iteration of a posterior sampler and averaged over the converged sample (144). Composite residuals, similar to those calculated for model B, were examined for models C and D and for E, the composite residuals that were examined are shown in equation (7.5).

Table 7.2 presents the residuals for each of the five models; these are given in decreasing order of the residuals from model C, the fully Bayesian spatial multilevel model. It can be seen that the residuals vary somewhat across the models; however, the trends in the estimates for models A to D are fairly similar with no particular pattern emerging, such as one model having obviously higher or lower residuals. Model E appears to show more differences in the residuals values and this can be seen more clearly from the matrix plot in Figure 7.1. This plots all of the residuals from models A to E. In each of the plots, except for the bottom row, the points tend to lie along the diagonal indicating agreement in the residuals for the models A to D. However, there is evidently variability between the models and the best agreement is between models C and D, the two multilevel models fitted using fully Bayesian methods, with most points lying close to the

line of equality. The plots comparing the residuals from models A to D with model E show much different pictures. The points do not appear to follow a similar pattern as the other models despite the initial adjustment of the residuals.

It is more informative to compare the confidence or posterior credible intervals for residuals from the models. These are presented for the four spatial models in Table 7.3. Looking across models B to D, it can now be seen that the intervals for each of the regions do overlap somewhat, covering the same range of values. Model B, the empirical Bayes spatial multilevel model, appears to have the widest intervals in most cases. The intervals for Model C (fully Bayes equivalent of C) and D (fully Bayes MMMC), as well as covering similar values, have similar ranges of the residuals. As expected, more discrepancies occur when comparing the intervals with those from model E. There is some overlap but also many of the intervals are very different in the values covered and in the width of the intervals. More of the intervals from model E, than from the other models, do not include zero.

Figures 7.2 to 7.6 display maps of the residuals from models A to E respectively. If the spatial model fits well and all relevant covariates are included, the result should be a spatially smooth map. Examining these "smooth" maps as a diagnostic technique should display unusual features that are inevitably not accounted for by the model; these will be highlighted by clusters of high positive or negative residuals. Figure 7.2 shows a map with a lot of variability and little smoothing, which is expected because model A does not incorporate the spatial structuring of the data into the modelling. Looking at Figure 7.3 it is clear that the map of residuals is much more smooth and an area with a particularly high residual can be clearly identified from this map. The range of the residuals has also reduced by 28% (model A: -0.253 to 0.253, model B: -0.150 to 0.215) and definite clustering of the residuals is evident within countries. Overall this map suggests model B to be more useful at modelling the disease than model A.

Table 7.1 Parameter estimates and confidence intervals: models A-E

Parameters	Empirical Bayes (RIGLS)		Full Bayes (MCMC)		
	Variance Components (A)	**Spatial Multilevel (B)**	**Spatial Multilevel (C)**	**Multiple Membership (D)**	**Conditional Autoregressive (E)**
β_0	**6.83 (6.69 , 6.98)**	**6.83 (6.53 , 7.13)**	**6.88 (6.67 , 7.08)**	**6.92 (6.71 , 7.13)**	
β_1 Smoking	**0.0005 (0.0004 , 0.0006)**	**0.0005 (0.0003 , 0.0007)**	**0.0006 (0.0004 , 0.007)**	**0.0005 (0.0004 , 0.0006)**	0.0007 (-0.0002 , 0.0016)
β_2 Fruit	**-0.0076 (-0.0086 , -0.0065)**	**-0.0075 (-0.0096 , -0.0054)**	**-0.0086 (-0.0098 , -0.0071)**	**-0.0079 (-0.0095 , -0.0064)**	**-0.0030 (-0.0057 , -0.0003)**
β_3 Vegetable	**-0.0011 (-0.0017 , -0.0006)**	-0.0011 (-0.0022, 0.00002)	**-0.0014 (-0.0022 , -0.0005)**	**-0.0011 (-0.0018 , -0.0004)**	**0.0082 (0.0022 , 0.0142)**
β_4 Animal Fat	**0.0264 (0.0194 , 0.0333)**	**0.0260 (0.0120 , 0.0400)**	**0.0305 (0.0227 , 0.0369)**	**0.0280 (0.0199 , 0.0360)**	**0.0258** (-0.0184 , 0.0699)
β_5 Alcohol	0.0006 (-0.0002 , 0.0014)	0.0007 (-0.0009 , 0.0023)	-0.0009 (-0.0019 , 0.0002)	-0.0003 (-0.0013 , 0.0007)	-0.0007 (-0.0033 , 0.0018)
β_6 GDP	2.00e-7 (-4.51e-6 , 4.91e-6)	3.27e-7 (-9.2e-6 , 9.9e-6)	1.47e-7 (-3.76e-6 , 3.13e-6)	-8.86e-7 (-4.13e-6 , 2.35e-6)	-6.30e-8 (-3.41e-6 , 3.29e-6)
σ_u^2	**0.0135 (0.0106 , 0.0164)**	0.0018 (-0.0136 , 0.0173)	**0.0038 (0.0026 , 0.0057)**	**0.0027 (0.0015 , 0.0039)**	**0.0028 (0.0013 , 0.0044)**
σ_v^2		**0.0436 (0.0113 , 0.0759)**	**0.0270 (0.0178 , 0.0390)**	**0.0283 (0.0185 , 0.0381)**	0.0060 (-0.0012 , 0.0130)
σ_{uv}		0.0114 (-0.0003 , 0.0231)	**0.0052 (0.0023 , 0.0086)**		

Table 7.2 Model residuals (ordered by decreasing residuals from model C)

Country	Region	R_A	R_B	R_C	R_D	R_E
Portugal	Madeira	0.1468	0.1689	0.2513	0.1293	-0.0475
France	Nord-Pas-de-Calais	0.2528	0.3984	0.1860	0.1611	0.0547
Denmark	Copenhagen *	0.1596	0.0562	0.1792	0.1170	0.2889
France	Corsica	0.1840	0.2010	0.1642	0.0599	-0.0049
Greece	Macedonia Central	0.0750	0.0081	0.1472	0.0941	0.0958
France	Picardie	0.1924	0.2467	0.1386	0.1254	-0.2879
Portugal	Azores	0.2243	0.2032	0.1338	0.1039	0.0532
UK	West Midlands	0.0272	-0.0130	0.1243	0.0460	0.0768
UK	Northern Ireland	-0.0535	0.0115	0.1240	0.1324	-0.0957
Austria	Lower Austria	0.0735	0.1015	0.1180	0.0659	0.0098
...	...		...	...	...	...
UK	Bedfordshire	-0.0233	-0.0758	0.0015	-0.0279	0.0038
France	Poitou-Charentes	0.0766	0.1884	0.0015	0.0554	-0.0430
UK	Humberside	0.0662	0.0277	0.00004	0.0150	0.0587
France	Midi-Pyrénées	-0.0171	0.1420	-0.0004	0.0468	-0.1235
Germany	Saarland	-0.0433	0.0642	-0.0032	-0.0481	0.0330
Germany	Baden-Wurttemberg	-0.2105	-0.0259	-0.0040	-0.0860	-0.0643
UK	Dyfed	-0.0495	-0.0136	-0.0056	-0.0062	-0.0676
Denmark	Ringkobing	-0.0792	-0.1451	-0.0070	-0.0149	-0.1904
UK	Hampshire	-0.0628	-0.0782	-0.0098	-0.0195	0.0156
Greece	Thessaly	0.0400	-0.0893	-0.0110	-0.0179	0.0965
...	...		...	...	...	...
Finland	Mikkeli	-0.1093	-0.2191	-0.1042	-0.0554	0.2209
Greece	Aegean South	-0.1100	-0.1164	-0.1100	-0.1021	-0.0697
Finland	Kuopio	-0.1849	-0.2466	-0.1159	-0.0700	0.3683
Germany	Bremen	0.0096	-0.0563	-0.1210	-0.0224	0.0522
Finland	Oulu	-0.1512	-0.2148	-0.1236	-0.0733	-0.0184
Greece	Greece West	-0.1072	-0.1355	-0.1310	-0.1103	-0.0052
Portugal	Norte	0.0220	0.1026	-0.1315	-0.0334	0.0382
Greece	Epirus	-0.2461	-0.1194	-0.1399	-0.1463	-0.1678
Greece	Ionian Islands	-0.0346	-0.1786	-0.1492	-0.1635	0.0247
Greece	Aegean North	-0.1393	-0.1693	-0.1599	-0.1263	-0.0586

*Copenhagen and Frederiksberg (city)

Looking at the residuals from model C (Figure 7.4), it can again be seen that a smoothed map is displayed with two regions of high positive residuals and one with high negative residuals; thus this model is also useful for highlighting extreme residuals. Despite the majority of regions having residuals close to zero ie orange, yellow or green, there is slightly more variability within countries perhaps reflecting a more similar picture of the true distribution of the residuals. The overall range in the residuals from model C (-0.160 to 0.251) is actually slightly wider than those from model B.

Examining Figure 7.5, the residuals from model D, it can be seen that the range of residuals (-0.160 to 0.161) is slightly narrower than any of the other models. However, comparing the map with Figure 7.4, a very similar picture can be seen; a spatially smoothed map is evident with two areas standing out as having extreme residuals. Both maps display a picture whereby the models have clearly smoothed the data well, reduced the variability and hopefully provided a picture that is close to the true residual surface.

Finally, looking at the map from model E (Figure 7.6) we can see that the overall range in residuals is much wider than from the other models. For the purpose of comparison, the residuals have been split into similar categories as for Figures 7.2 to 7.5, but this is perhaps not the most ideal choice of ranges due to the wider spread of the residuals. However, as can be seen from the map, the majority of regions do have residuals which still lie close to zero. There are much more 'extreme' residuals (blue and red areas) than were evident on the other map, but this could be explained by the residuals not being directly comparable and being grouped wrongly.

As previously mentioned, few regions have residual confidence intervals that do not include zero (see Table 7.3). This indicates that few regions actually have relative risks significantly different from unity and from examining the maps for models B, C and D (Figures 7.3, 7.4 and 7.5) it can be seen that they also only suggest one or two areas with particularly high residuals. This suggests these disease maps are useful for detecting disease 'hotspots' where the populations are at significant risk of mortality.

Table 7.3 Residual confidence intervals (ordered as in Table 7.2)

Region	$CI_{95\%}$ (B)	$PI_{95\%}$ (C)	$CI_{95\%}$ (D)	$CI_{95\%}$ (E)
Madeira	(-0.046 ,0.476)	(-0.008 ,0.497)	(-0.082 ,0.341)	(-0.144 ,0.049)
Nord-Pas-de-Calais	(-0.304 ,0.564)	(-0.112 ,0.567)	(-0.158 ,0.480)	(-0.031 .0.140)
Copenhagen *	(-0.409 ,0.453)	(-0.133 ,0.545)	(-0.248 ,0.482)	(0.115 .0.463)
Corsica	(-0.349 ,0.502)	(-0.114 ,0.453)	(-0.264 ,0.384)	(-0.106 .0.096)
Macedonia Central	(-0.132 ,0.232)	(-0.035 ,0.317)	(-0.062 ,0.251)	(0.017 ,0.175)
Picardie	(0.135 ,0.243)	(0.022 ,0.243)	(0.035 ,0.216)	(-0.363 ,-0.213)
Azores	(-0.106 ,0.416)	(-0.117 ,0.408)	(-0.114 ,0.321)	(-0.052 ,0.158)
West Midlands	(-0.198 ,0.208)	(-0.064 ,0.329)	(-0.177 ,0.269)	(-0.011 ,0.165)
Northern Ireland	(-0.401 ,0.448)	(-0.206 ,0.507)	(-0.255 ,0.520)	(-0.190 ,-0.002)
Lower Austria	(-0.059 ,0.191)	(-0.021 ,0.248)	(-0.041 ,0.172)	(-0.061 ,0.080)
...	... ...	... ...	... ...	... ...
Bedfordshire	(-0.191 ,0.151)	(-0.165 ,0.163)	(-0.163 ,0.107)	(-0.079 ,0.086)
Poitou-Charentes	(-0.103 ,0.221)	(-0.112 ,0.122)	(-0.084 ,0.195)	(-0.088 ,0.002)
Humberside	(-0.163 ,0.175)	(-0.132 ,0.149)	(-0.126 ,0.156)	(-0.015 ,0.132)
Midi-Pyrénées	(-0.107 ,0.215)	(-0.118 ,0.153)	(-0.058 ,0.152)	(-0.142 ,-0.105)
Saarland	(-0.181 ,0.211)	(-0.146 ,0.157)	(-0.193 ,0.097)	(-0.064 ,0.130)
Baden-Wurttemberg	(-0.173 ,0.166)	(-0.117 ,0.132)	(-0.220 ,0.048)	(-0.129 ,0.001)
Dyfed	(-0.192 ,0.184)	(-0.183 ,0.182)	(-0.158 ,0.146)	(-0.154 ,0.018)
Ringkobing	(-0.218 ,0.098)	(-0.152 ,0.135)	(-0.172 ,0.142)	(-0.320 ,-0.061)
Hampshire	(-0.160 ,0.104)	(-0.130 ,0.111)	(-0.124 ,0.085)	(-0.061 ,0.092)
Thessaly	(-0.168 ,0.128)	(-0.136 ,0.123)	(-0.162 ,0.126)	(0.012 ,0.181)
...	... ...	... ...	... ...	... ...
Mikkeli	(-0.225 ,0.030)	(-0.230 ,0.021)	(-0.178 ,0.067)	(0.192 ,0.250)
Aegean South	(-0.211 ,0.160)	(-0.264 ,0.042)	(-0.268 ,0.064)	(-0.152 ,0.012)
Kuopio	(-0.251 ,0.079)	(-0.274 ,0.044)	(-0.220 ,0.080)	(0.419 ,0.318)
Bremen	(-0.459 ,0.428)	(-0.565 ,0.272)	(-0.376 ,0.331)	(-0.025 ,0.129)
Oulu	(-0.257 ,0.040)	(-0.275 ,0.019)	(-0.212 ,0.065)	(-0.089 ,0.053)
Greece West	(-0.227 ,0.040)	(-0.263 ,0.013)	(-0.241 ,0.021)	(-0.083 ,0.072)
Norte	(-0.426 ,0.435)	(-0.467 ,0.211)	(-0.358 ,0.291)	(-0.072 ,0.148)
Epirus	(-0.192 ,0.091)	(-0.283 ,-0.011)	(-0.279 ,-0.014)	(-0.260 ,-0.075)
Ionian Islands	(-0.318 ,0.205)	(-0.391 ,0.097)	(-0.400 ,0.073)	(-0.079 ,0.128)
Aegean North	(-0.249 ,0.062)	(-0.306 ,-0.024)	(-0.263 .0.011)	(-0.133 ,0.016)

Figure 7.1

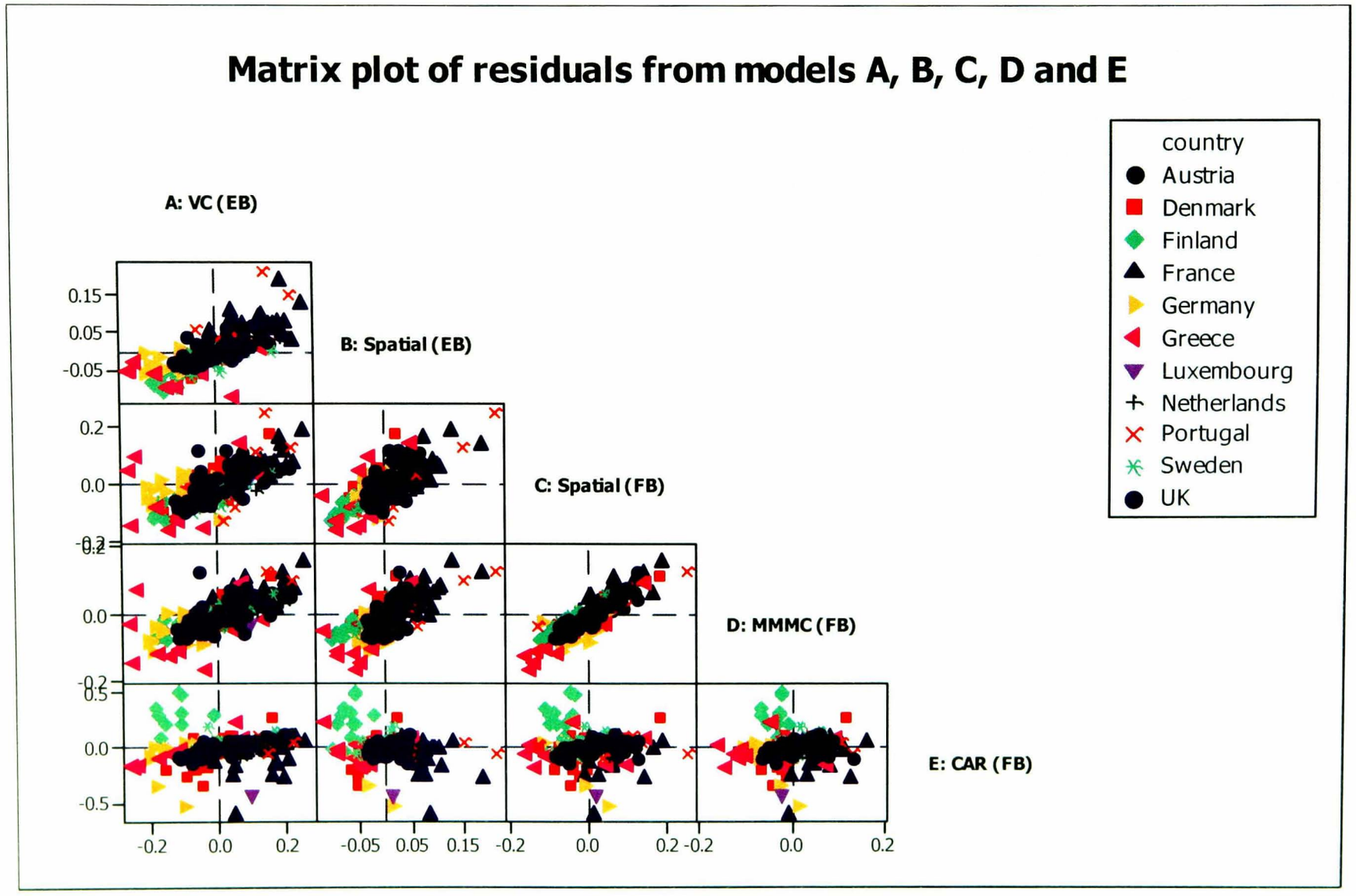

Matrix plot of residuals from models A, B, C, D and E
A: VC (EB)
B: Spatial (EB)
C: Spatial (FB)
D: MMMC (FB)
E: CAR (FB)
country
Austria
Denmark
Finland
France
Germany
Greece
Luxembourg
Netherlands
Portugal
Sweden
UK

Figure 7.2 Map of residuals from model A (VC fitted using EB)

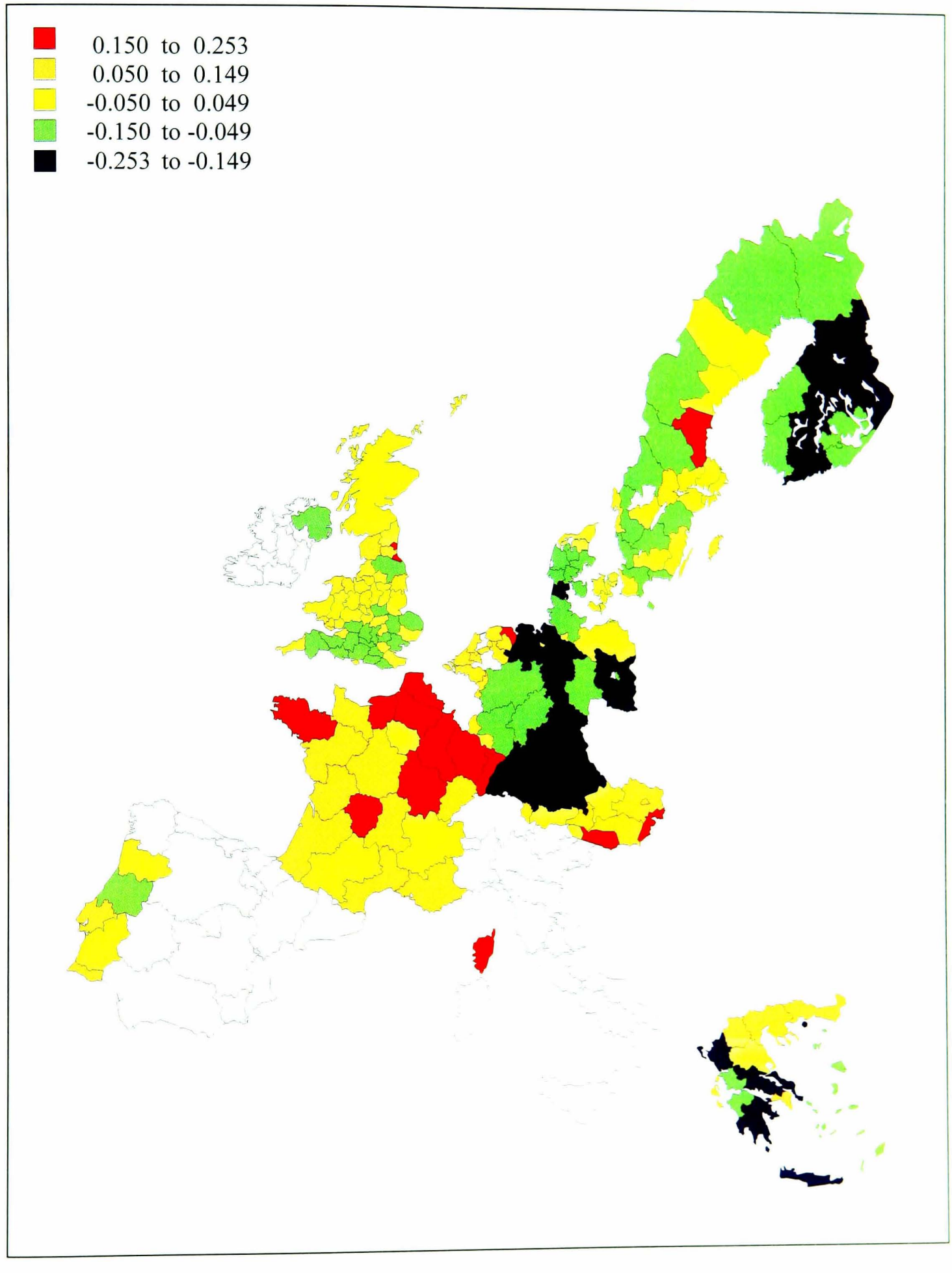


Figure 7.3 Map of residuals from model B (Spatial fitted using EB)

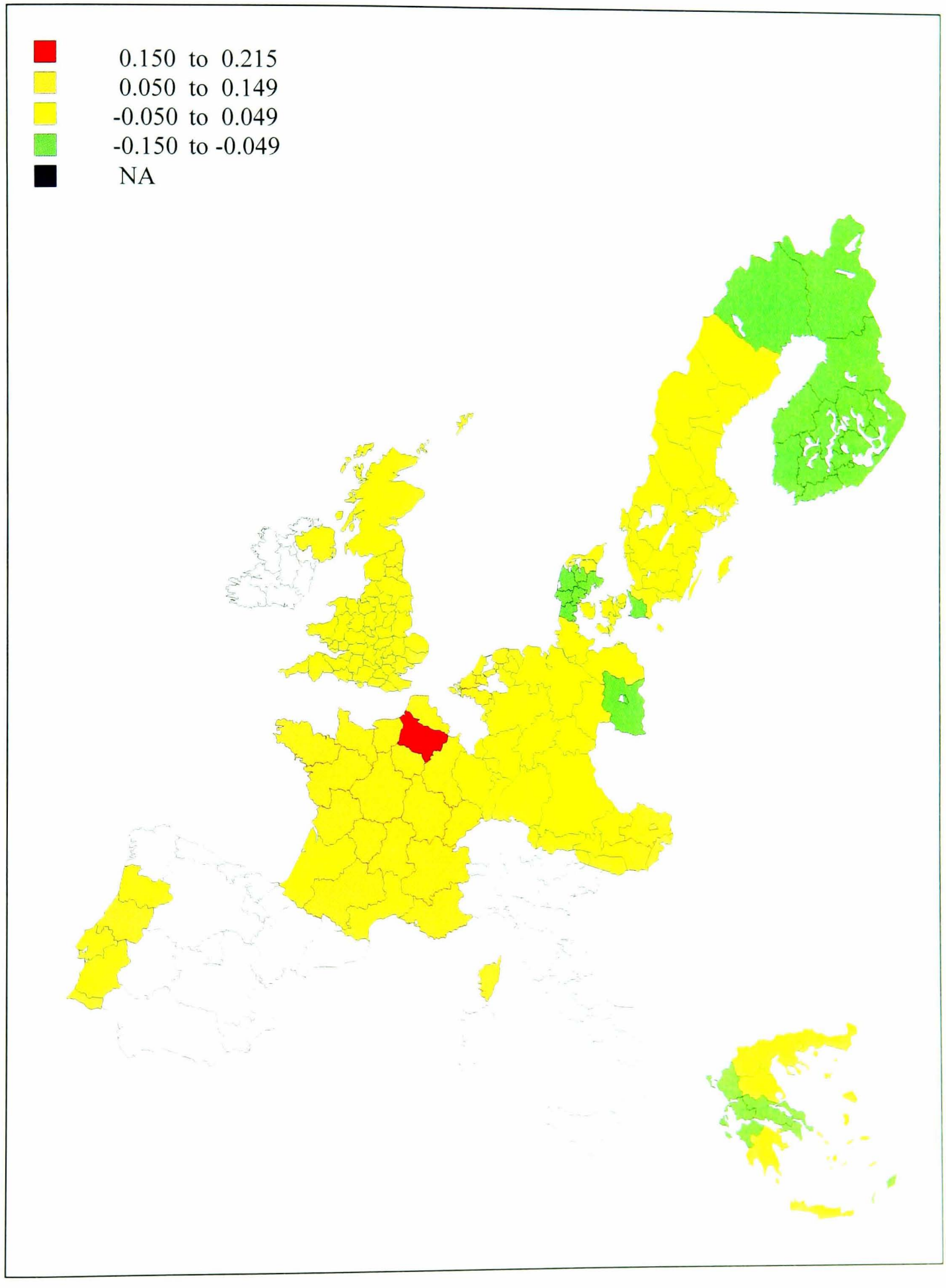


Figure 7.4 Map of residuals from model C (Spatial fitted using FB)

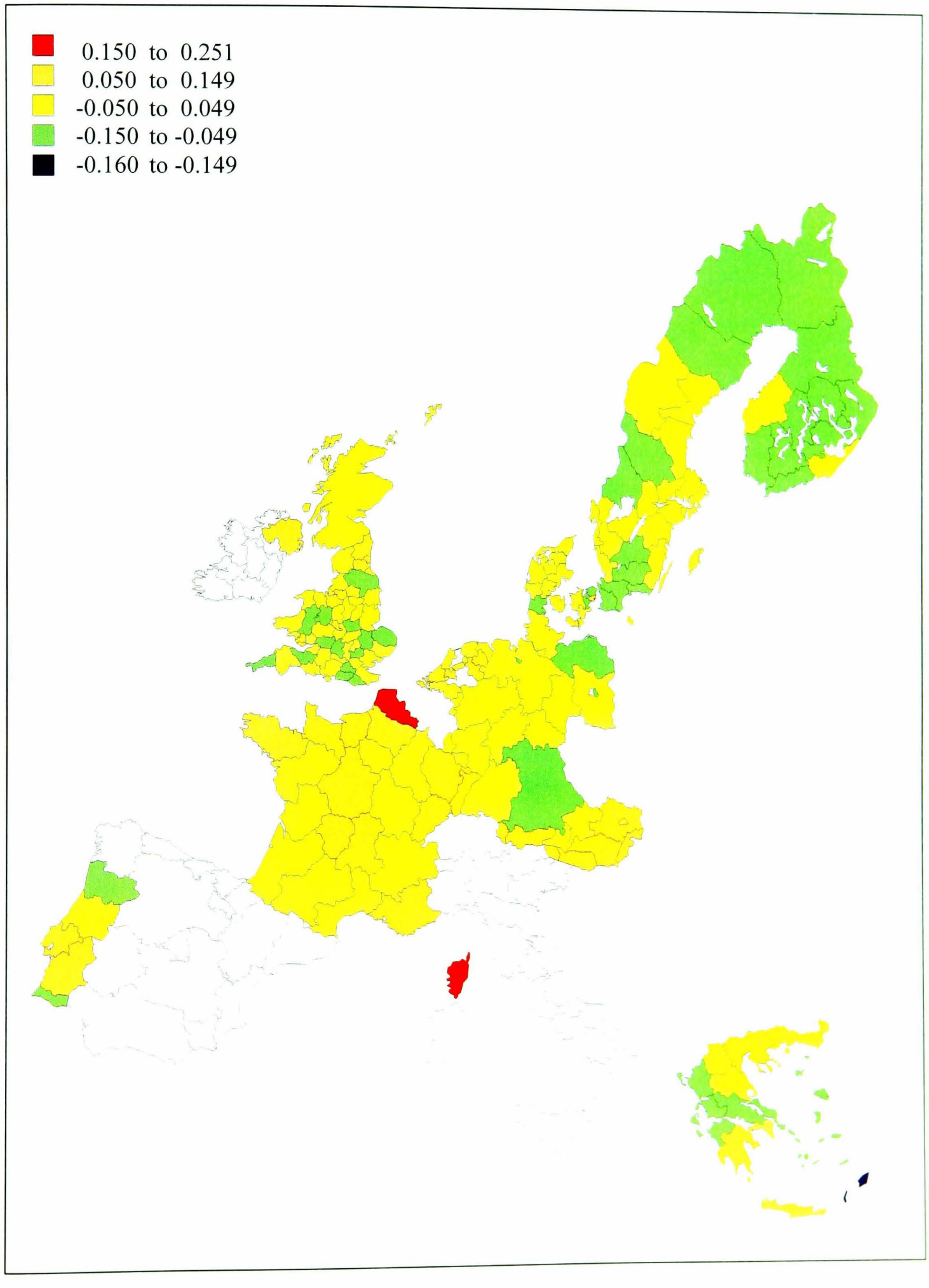


Figure 7.5 Map of residuals from model D (MMMC fitted using FB)

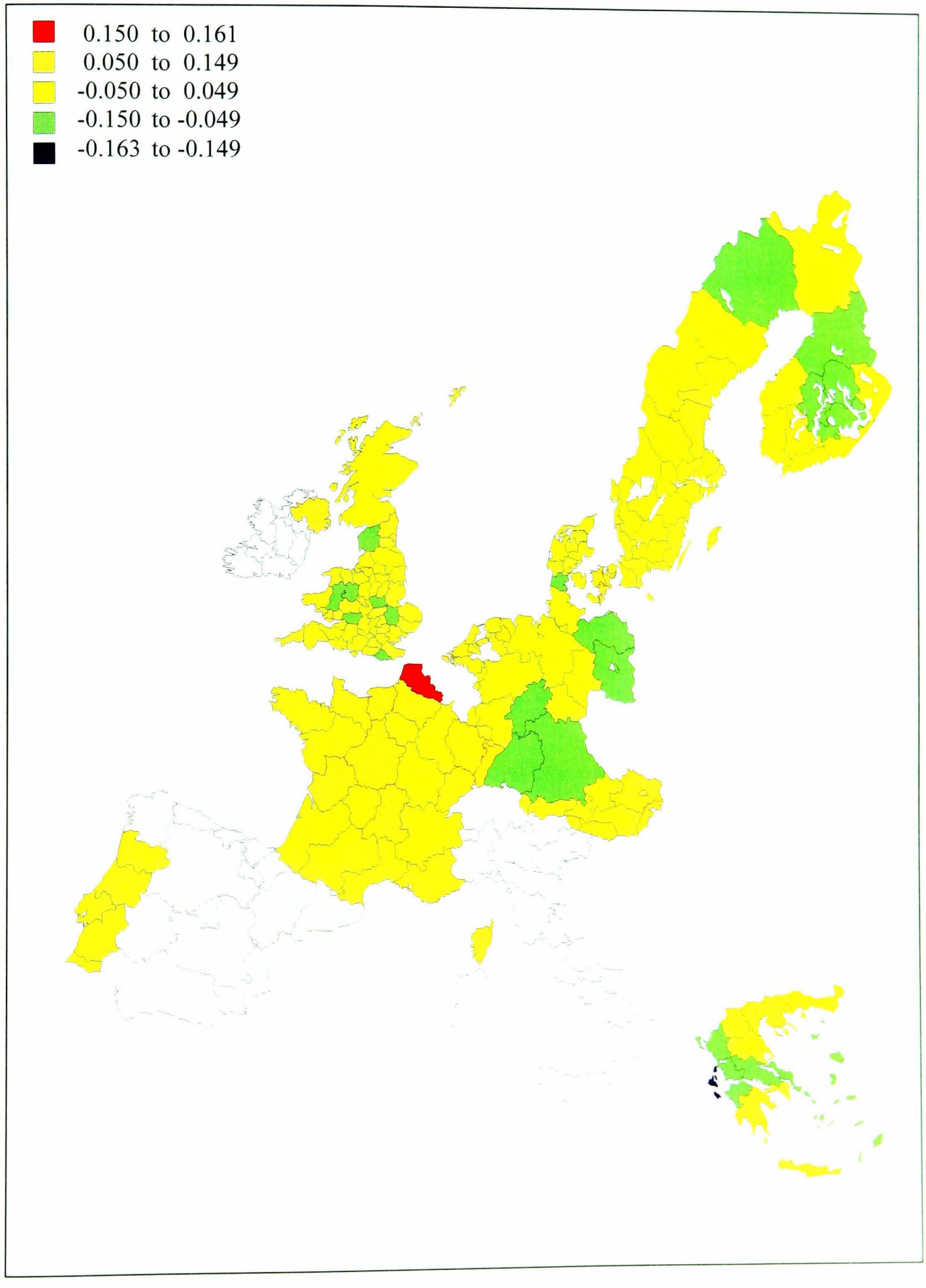
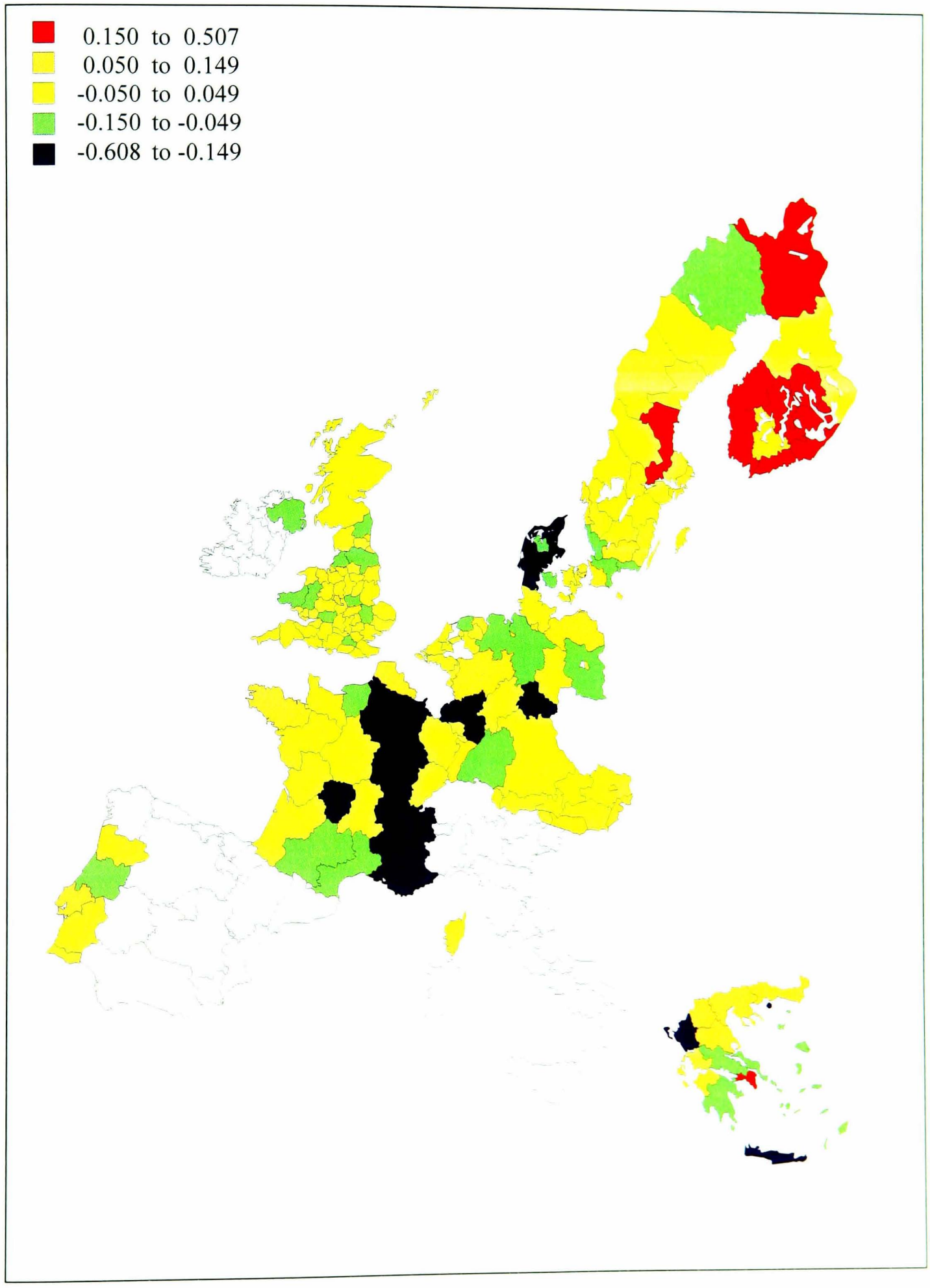


Figure 7.6 Map of residuals from model E (CAR fitted using FB)



7.4.3 Deviance Information Criterion (DIC)

The complexity and fit of the fully Bayesian models were compared using the deviance information criterion (DIC) (145). This is based on the posterior distribution of $D(C)$ and consists of two components, a term that measures goodness-of-fit and a penalty term for increasing model complexity; it is written as

$$DIC = \overline{D} + p_D . \tag{7.6}$$

The first term is a Bayesian measure of model fit and is defined as the posterior expectation of the deviance

$$\overline{D} = \mathrm{E}_{\theta|y}\left[D(\theta)\right] = \mathrm{E}_{\theta|y}\left[-2\ln f(y|\theta)\right] \tag{7.7}$$

where $f(y|\theta)$ is the likelihood function ie the joint conditional probability function of the observations given the unknown parameters. The better the model fits the data, the larger the values for the likelihood. Since D is defined using minus twice the log-likelihood, smaller values represent 'better' models.

The second term, p_D, is an estimate of the 'effective' number of parameters and can be thought of as a penalty term reflecting the model complexity or degrees of freedom. p_D is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean $\overline{\theta}$ of the parameters

$$\begin{aligned} p_D &= \overline{D} - D(\theta) = \mathrm{E}_{\theta|y}\left[D(\theta)\right] - D\left(\mathrm{E}_{\theta|y}\left[\theta\right]\right) \\ &= \mathrm{E}_{\theta|y}\left[-2\ln f\left(y|\theta\right)\right] + 2\ln f\left(y|\overline{\theta}\right). \end{aligned} \tag{7.8}$$

p_D should capture the amount of shrinkage obtained by the hierarchical prior. A p_D value that is small relative to the number of data points indicates that the prior provides a lot of information, and therefore there has been little borrowing of strength across the regions. It makes sense to compare the estimated p_D for each model to gain some knowledge of the amount of important structural information

provided by the second stage prior. On the other hand, models with negligible prior information will have a p_D approximately equal to the number of parameters. In this case the DIC will be close to Akaike's Information Criterion (AIC); later in this section the AIC will be used to compare the empirical Bayes models.

The DIC will be used to compare the performance of the given disease mapping models; it is effectively a method of selecting the model that leads to the best prediction of the risk surface in the areas of interest. For each of the models fitted using fully Bayesian methods, the DIC is presented in Table 7.4. The table also presents the separate contributions of fit ($\overline{D}$), complexity (P_D) and deviance ($D(\overline{\theta})$).

Table 7.4 DIC for models fitted using fully Bayesian

Models		$\overline{D}$	$D(\overline{\theta})$	P_D	**DIC**
Spatial Multilevel	**(C)**	1939.0	1770.0	168.4	**2107.4**
Multiple Membership	**(D)**	1938.6	1771.0	167.6	**2106.2**
Conditional Autoregressive	**(E)**	1945.2	1779.3	166.0	**2111.2**

Firstly it should be seen that all three models are highly complex leading to a considerable borrowing of strength, with between 166 and 168 'effective' parameters being needed to fit the 187 data points. The DIC values indicate that model D (MMMC model) is the 'best' model out of the three considered here as it has the lowest DIC value. However, Spiegelhalter et al (145) suggested that models with DIC values within 1 or 2 of the 'best' model should be strongly supported, those within 3 and 7 should be weakly supported and those with a DIC more than 7 higher than the 'best' are very much inferior. This suggests model E is weaker than models C and D which are effectively equal in terms of their fit

and performance and have priors providing appropriate information about the spatial structure.

To compare the empirical Bayes models, where the prior is completely specificied with no hyperparameters, the Akaike Information Criterion can be used. The AIC can be written as

$$\text{AIC} = D(\hat{\theta}) + 2p, \tag{7.9}$$

where p is the number of parameters and $\hat{\theta}$ denotes the maximum likelihood estimate, or in this case the quasi-likelihood equivalent. In fact, equation (7.8) can be rearranged and written as $\overline{D} = D(\theta) + p_D$, then the DIC defined in equation (7.6) can then be re-written as

$$\text{DIC} = D(\bar{\theta}) + 2p_D. \tag{7.10}$$

which is the same as equation (7.9) but with the posterior mean $\bar{\theta}$ substituted by the ML, or QL, estimate of θ. Thus, the DIC can be seen as a generalisation of the AIC and, in the special case where the prior is flat, such as with the empirical Bayes analysis, AIC equals DIC since the ML estimate coincides with the posterior mean. This is also approximately true for QL estimates since the estimates may not maximise the likelihood but should approximate it. It follows that a model with a smaller AIC is favoured and comparing the empirical Bayes estimates would suggest that the variance components model, with a lower AIC, is the 'best' model. However, in this case, model B is favoured since A ignores important information about the spatial structure. Also, if the QL estimates are a good approximation of ML estimates, the -2ln($f(y|\theta)$) value for model B should be lower than that of model A. This is not true, as can be seen from Table 7.5, and is a disadvantage of using model fitting using quasilikelihood methods. Looking at the two sets of models using the DIC and AIC values gives a crude comparison of all the models and it is clear the fully Bayesian model should be supported.

Table 7.5 AIC for models fitted using empirical Bayes

Models		-2ln($f(y\|\theta)$)	$2p$	AIC
Variance Components	**(A)**	2547.5	16	**2563.5**
Spatial Multilevel	**(B)**	2669.2	20	**2689.2**

7.4.4 Iterations

A final factor that should be considered when making model choices, especially when MCMC methods are being used to fit complex models, is the time it takes to run the required number of iterations. Actual times have not been given here but Table 7.6 displays the number of iterations required for each of the models to reach convergence (see Appendix A3.1–A3.2 for convergence diagnostics). For models C to E, that were fitted using MCMC techniques, a burn-in period plus further iterations, as described in Chapter 6, was required. As can be seen, model E required the most iterations in total, with a burn-in period of 2,000,000 being needed followed by 100,000 further iterations. As a result this model took the longest time to run. Model D required fewer iterations in total followed by model C which required the least overall; which took much less time to run. However, it should be noted that all models required many hours of simulation time before suitable posterior distributions were obtained.

Finally, due to the computational intensity of the MCMC simulations, iterative quasi-likelihood procedures have an advantage. From Table 7.6 it can be seen that both models A and B required very few iterations before convergence of parameters. These took minutes to run and therefore have the advantage of much less time being needed to fit the models.

Table 7.6 Number of iterations required for convergence

Models		Burn-in	Further Iterations
Variance Components *	**(A)**	-	6
Spatial Multilevel *	**(B)**	-	8
Spatial Multilevel	**(C)**	200,000	50,000
Multiple Membership	**(D)**	1,000,000	100,000
Conditional Autoregressive	**(E)**	2,000,000	100,000

* models fitted using empirical Bayes methods so only the number of iterations required until convergence presented

7.5 Model Choice

The main goal of disease mapping is to remove random noise and any natural variation in the human population allowing identification of areas with high or low rates. To accurately produce such a map a model can be used that allows the borrowing of strength across the whole of the study region and reduces the variance through the use of shrinkage estimators. In this chapter five such models were examined in an attempt to find the "best" model for mapping regional level cancer mortality.

Examining the parameter estimates initially showed the importance of taking account of spatial autocorrelation. Comparing the fixed parameters from the spatial models suggested that, with the exception of the CAR model, the fully Bayesian methods gave more precise estimates. The random parameter estimates for the fully Bayesian spatial multilevel model (C) and MMMC model (D) were very similar but discrepancies were evident when comparing them to the empirical Bayes spatial multilevel model (B); different estimates, wider intervals

and a negative variance being included in the confidence interval all suggested C and D were superior. The differences in the fixed parameter estimates in the CAR model (E) lead to concerns about its uses for modelling this type of data.

The best agreement in the residuals was between models C and D. With the exception of the CAR model, which again was not directly comparable, the variance components model (A) produced the most extreme residuals. Comparing the residuals' confidence intervals for models B, C and D again showed model B to cover the widest range of values; models C and D were narrower and very similar. The maps of the residuals for models C and D appeared to produce the most smoothed maps and were the most useful diagnostic tools. The map of model D showed that the residuals covered the narrowest range overall.

Examining the Deviance Information Criterions suggests models C and D were equally superior to model E and both provided appropriate information about the spatial structure in the data through the use of adequate priors. Comparing the Akaike Information Criterion for the models fitted using Empirical Bayes methods suggested that the variance components model should be favoured, but, this ignores the spatial structure and therefore has its disadvantages. Using the AIC and the DIC values to give a crude comparison of model fit across the empirical and fully Bayesian methods showed that models C and D had much lower information criterion values, again suggesting 'better' models.

Finally, time taken to fit the various models was considered when choosing a model. As discussed, the empirical Bayes models required much less time than the fully Bayes models. Among the MCMC models, that all required numerous iterations, model C required the least and could be fitted in less time than models D and E. In fact, this was one of the main deciding factors in model choice between the two favoured disease mapping models. As described in this section, models C and D, the fully Bayesian spatial multilevel model and the multiple-membership model, are superior in different manners to the others. Both provide similar estimates and are very useful for disease mapping, but, model C could be

fitted in about a fifth of the time of that for model D making it a more favourable choice.

It should also be noted that model C provides an additional piece of information, namely a correlation between the two sets of random effects, u and v. From the parameter estimates (Table 7.1) it can be seen that this term is significant showing a positive correlation between the two terms and suggesting it is useful to incorporate it in the modelling. Also, because they were fitted using MCMC methods, model C, and effectively model D, have another advantage of allowing more complex multilevel modelling structures to be fitted, such as the addition of higher geographical levels; these modelling extensions proved difficult using empirical Bayes methods.

For the above reasons the natural decision was to model and disease map further similar data using model C, the fully Bayesian spatial multilevel model. This model will be used to explore the spatial distributions of various specific cancers in Chapter 8.

Chapter 8

8 Specific Cancers

8.1 Examining Specific Cancer Rates

Examining all cancers grouped together has proved informative and is commonly used to measure variation in health across regions and countries. However, it is a fairly crude measure of analysis and so far has ignored the fact that specific cancers may follow different spatial patterns in the EU. Also, literature suggests that relationships with various risk factors vary for different cancers, and therefore it makes sense to consider different cancers individually.

The mortality data used in previous Chapters are all malignant neoplasms (ICD-9 140-208) for 187 regions in 11 EU countries (Table 3.3). This dataset can be broken down further into specific cancers by regions. The malignant neoplasms that will be examined further in this chapter are i) malignant neoplasm of trachea, bronchus and lung (which will be referred to as lung cancer from now on), ICD-9 162; ii) malignant neoplasm of oesophagus (oesophageal cancer), ICD-9 150; and iii) malignant neoplasm of the colon, rectum rectosigmoid junction and anus (colorectal cancer) ICD-9 153,154.

8.2 Lung Cancer Mortality

In 1998, 32% of the world's lung cancer deaths occurred in Europe (146) despite Europe comprising approximately an eighth of the world's population (147). Lung cancer is the most common cause of death from cancer in European men; in 1995 29% of the total male cancer burden was due to lung cancer (2). Recent studies

have shown an increase in both incidence of and mortality from lung cancer among woman in Europe (148-151). There has been a "rapid increase" in female lung cancer mortality in some countries in Europe (150) and rates in EU countries appear to have doubled in the period 1955-1994 (148).

8.2.1 Modelling Lung Cancer Mortality

Previous studies such as these have examined lung cancer mortality patterns and trends in Europe at a country level (2, 152, 153). It is of interest to evaluate whether these patterns remain after taking regional variations into account. Also, using modelling methods that allow adjustment for spatial patterning and potential risk factors may produce a different picture of European lung cancer mortality rates and more accurately quantify the burden of the disease.

Again, 187 regions in 11 EU countries in 1991 are examined. The data are modelled using the fully Bayesian spatial multilevel model (equations (6.5) and (6.6)). The full model fitted the same explanatory variables as were used when examining all cancers together. Convergence was monitored using time series plots and Gelman-Rubin plots for each parameter from simultaneous runs of the model resulting in a burn-in period of 50,000 being needed for the two-level null model. The number of further iterations required was determined by examining the Monte Carlo error in relation to the posterior standard deviation for each parameter, and the two-level null model required 10,000 further iterations until a suitable posterior distribution was available from which to sample. The two-level full model required a burn-in of 200,000 iterations and 300,000 further iterations. The full three-level model took 500,000 iterations until convergence and a further 300,000 iterations were then required.

8.2.2 Model Results: Lung Cancer

Tables 8.1 and 8.2 give estimates of relative risks of lung cancer mortality for selected regions; extreme (two maximum and two minimum) rates are given for

each country (except Luxembourg where only country level data are available), ordered by declining posterior relative risk estimates within each country. Table 8.1 gives the standardised mortality ratios and 95% confidence intervals based on the Poisson distribution, and the relative risks (posterior means) obtained from the two-level model with no covariates and corresponding 95% credible intervals (posterior 0.025 and 0.975 quantiles). Table 8.2 gives the relative risks and 95% posterior credible intervals from the two-level and three-level models with all six covariates included. Table 8.3 gives the country level SMRs and relative risks from the three-level full model. Table 8.4 gives the parameter estimates (posterior means and 95% posterior credible intervals) from the three previously mentioned models. Relative risks were calculated from the fixed effect parameter estimates to compare regions with high and low levels of exposure to the given risk factors and these are given in Table 8.5. Figures 8.1-8.4 are maps of the SMRs and the predicted relative risks from the three models.

8.2.2.1 Two-Level Null Model: Lung Cancer

Table 8.1 gives the two highest and the two lowest relative risks within each country, predicted from the two-level spatial model with no covariates. The SMR for each of these regions is also given. The predicted relative risks are taking into account the spatial patterning of lung cancer mortality and any extra-Poisson variation and the estimates show that the area with the highest relative risk of lung cancer mortality is Northern Ireland in the United Kingdom with a RR of 1.59, followed by Copenhagen and Frederiksberg (city) (RR=1.53) and Northumberland in the UK (RR=1.46). Therefore, lung cancer mortality in Northern Ireland is 59% higher than expected. Before taking account of risk factors, areas with the lowest relative risk of lung cancer mortality in the EU countries under investigation are Norte in Portugal (RR=0.60) and Vasterbotten in Sweden (RR=0.67); this means that lung cancer mortality is 40% lower than expected in Norte and 33% lower than expected in Vasterbotten. Regions in Portugal display the widest range in relative risks (0.60 – 1.33) closely followed by the UK (0.88 – 1.59). The

Netherlands have the smallest range in relative risks (0.97 – 1.16) followed by Austria (0.84 – 1.08).

8.2.2.2 Lung Cancer SMRs

Also, for each region given in Table 8.1, the corresponding SMRs are presented. The disadvantages of examining these have already been discussed but they are shown here for comparison purposes. Examining SMRs alone ignores population sizes at risk, spatial patterning etc and gives a different picture of the risk of lung cancer mortality in the EU. For example, Northern Ireland has the highest predicted relative risk from the two-level null spatial model. However, looking at the SMR alone suggests that the risk of lung cancer mortality is only 21% higher than expected (compared to 59%). Looking at Norte, the region with the lowest predicted relative risk, it can be seen the SMR has again dropped. Also, the confidence interval for the SMR (based on the Poisson distribution) suggests that Norte has a significantly lower than expected risk of lung cancer mortality; 42% - 49% lower, whereas the model predicts that this region's relative risk does not differ significantly from unity. This pattern emerges for many regions.

8.2.2.3 Lung Cancer Disease Maps I (SMRs and Two-Level Null Model RRs)

The map of SMRs (Figure 8.1) shows that there is a very high amount of variation within the EU, ranging from 0.33 to 2.16. It can also be seen that within some countries there is a high amount of clustering, with the UK and the Netherlands generally having high SMRs and Sweden and Portugal having low SMRs.

Estimating the relative risks from the two-level null spatial model reduces the number of regions with extreme rates (see Figure 8.2). Here, the within-area effects are modelled with a Poisson distribution, and relative risks between areas are considered as having a log-normal distribution with the mean for each area being centred on the mean of its neighbours. This, in effect, is providing a smoother map and is getting closer to the true picture of lung cancer mortality in

the EU. Areas that stood out as having extreme rates on the disease map of mortality previous to modelling may have high variability in their estimates due to having smaller population sizes. In such a case, one cannot confidently identify a region as, for example, a disease 'hotspot'. Modelling the data to remove this variability produces estimates of the 'true' risk of the disease, therefore extreme rates that dominate this map can be more accurately interpreted as lung cancer mortality 'hotspots'. From the map it can be seen that areas with a particularly high risk of lung cancer mortality are some UK regions, Lorraine in France and two regions in east Greece. Those areas now standing out as having particularly low risk of lung cancer mortality are regions in Portugal, Bavaria in Germany and some parts of Sweden.

8.2.2.4 Two-Level Full Model: Lung Cancer

As discussed in Chapter 2, there are other factors affecting the risk of lung cancer mortality that so far have not been taken into account. It is of interest to examine the risk of mortality from this disease after taking into account measures of risk and protective factors and also after taking into account the within-country clustering that was evident from the map of SMRs (Figure 8.1). Table 8.2 gives the estimated relative risks of lung cancer mortality from the two-level spatial model including the six covariates used in previous model fitting: fruit, vegetable, animal fat and alcohol consumption, cigarette smoking and gross domestic product. The two highest and the two lowest relative risks within each country are given. The estimates are also given for the three-level model, incorporating country as a higher level to account for the fact that regions within a country are more likely to be homogeneous than regions from different countries.

Table 8.1 Relative risks (SMR and RR from two-level null model) of mortality from lung cancer (for each country extreme rates given, ordered by decreasing RR)

i	Country	Region	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
3		Carinthia	284	253.6	1.12	(0.99 ,1.26)	1.08	(0.78 , 1.51)
7		Styria	497	560.9	0.89	(0.81 ,0.97)	1.06	(0.83 , 1.35)
...	Austria	...	...	...	...	... ...	...	... ...
2		Burgenland	136	129.6	1.05	(0.88 ,1.24)	0.91	(0.59 , 1.40)
5		Upper Austria	476	579.8	0.82	(0.75 ,0.90)	0.84	(0.64 , 1.11)
85		Copenhagen*	549	312.9	1.75	(1.61 ,1.91)	1.53	(0.82 , 2.91)
86		Copenhagen	400	298.0	1.34	(1.21 ,1.48)	1.28	(1.01 , 1.65)
...	Denmark	...	...	...	...	... ...	...	... ...
94		Ribe	136	107.9	1.26	(1.06 ,1.49)	0.96	(0.68 , 1.37)
98		Viborg	133	127.7	1.04	(0.87 ,1.23)	0.94	(0.66 , 1.32)
126		Oulu	166	174.8	0.95	(0.81 ,1.11)	1.17	(0.87 , 1.57)
124		Lappi	81	82.4	0.98	(0.78 ,1.22)	1.03	(0.69 , 1.55)
...	Finland	...	...	...	...	... ...	...	... ...
120		Hame	230	312.2	0.74	(0.64 ,0.84)	0.90	(0.67 , 1.21)
130		Vassa	166	214.9	0.77	(0.66 ,0.90)	0.87	(0.61 , 1.24)
139		Lorraine	1156	843.1	1.37	(1.29 ,1.45)	1.22	(0.96 , 1.56)
134		Haute-Normandie	714	629.7	1.13	(1.05 ,1.22)	1.13	(0.86 , 1.48)
...	France	...	...	...	...	... ...	...	... ...
142		Pays de la Loire	851	1207.0	0.71	(0.66 ,0.75)	0.84	(0.62 , 1.11)
143		Bretagne	964	1159.8	0.83	(0.78 ,0.89)	0.84	(0.54 , 1.31)
77		Saarland	639	520.5	1.23	(1.13 ,1.33)	1.14	(0.78 , 1.68)
74		Rheinland-Palatinate	1905	1869.1	1.02	(0.97 ,1.07)	1.08	(0.85 , 1.37)
...	Germany	...	...	...	...	... ...	...	... ...
76		Bavaria	3900	5466.4	0.71	(0.69 ,0.74)	0.79	(0.65 , 0.96)
71		Bremen	367	361.0	1.02	(0.92 ,1.13)	0.76	(0.36 , 1.53)
157		Macedonia East+	323	271.6	1.19	(1.06 ,1.33)	1.24	(0.85 , 1.83)
158		Macedonia Central	874	765.4	1.14	(1.07 ,1.22)	1.20	(0.88 , 1.67)
...	Greece	...	...	...	...	... ...	...	... ...
163		Greece West	335	359.7	0.93	(0.83 ,1.04)	0.88	(0.69 , 1.13)
166		Attica	1619	1545.0	1.05	(1.00 ,1.10)	0.83	(0.62 , 1.10)
238	Luxembourg	...	184	175.7	1.05	(0.90 ,1.21)	1.10	(0.75 , 1.61)
242		Friesland	326	280.8	1.16	(1.04 ,1.29)	1.16	(0.86 , 1.59)
248		Limburg	651	489.4	1.33	(1.23 ,1.44)	1.15	(0.83 , 1.60)
...	Netherlands	...	...	...	...	... ...	...	... ...
246		Flevoland	108	76.9	1.40	(1.15 ,1.70)	0.98	(0.69 , 1.39)
241		Groningen	337	262.4	1.28	(1.15 ,1.43)	0.97	(0.68 , 1.38)
327		Madeira	74	93.8	0.79	(0.62 ,0.99)	1.33	(0.82 , 2.14)
326		Azores	78	95.9	0.81	(0.64 ,1.01)	1.28	(0.80 , 2.06)
...	Portugal	...	...	...	...	... ...	...	... ...
323		Lisboa e Vale do Tejo	787	1490.6	0.53	(0.49 ,0.57)	0.72	(0.50 , 1.01)
321		Norte	733	1347.4	0.54	(0.51 ,0.58)	0.60	(0.32 , 1.10)
462		Orebro	112	169.1	0.66	(0.55 ,0.80)	1.01	(0.77 , 1.34)
467		Uppsala	85	138.4	0.61	(0.49 ,0.76)	0.99	(0.74 , 1.34)
...	Sweden	...	...	...	...	... ...	...	... ...
454		Jamtland	32	88.5	0.36	(0.25 ,0.51)	0.71	(0.51 , 0.99)
469		Vasterbotten	50	141.7	0.35	(0.26 ,0.47)	0.67	(0.47 , 0.94)
565		Northern Ireland	788	650.9	1.21	(1.13 ,1.30)	1.59	(0.83 , 3.13)
513		Northumberland	252	164.2	1.53	(1.35 ,1.74)	1.46	(1.10 , 1.98)
...	UK	...	...	...	...	... ...	...	... ...
532		East Sussex	515	464.7	1.11	(1.01 ,1.21)	0.90	(0.64 , 1.27)
521		Leicestershire	428	428.4	1.00	(0.91 ,1.10)	0.88	(0.69 , 1.10)

*Copenhagen and Frederiksberg (city), +Macedonia East and Thrace

Figure 8.1 Map of lung cancer SMRs

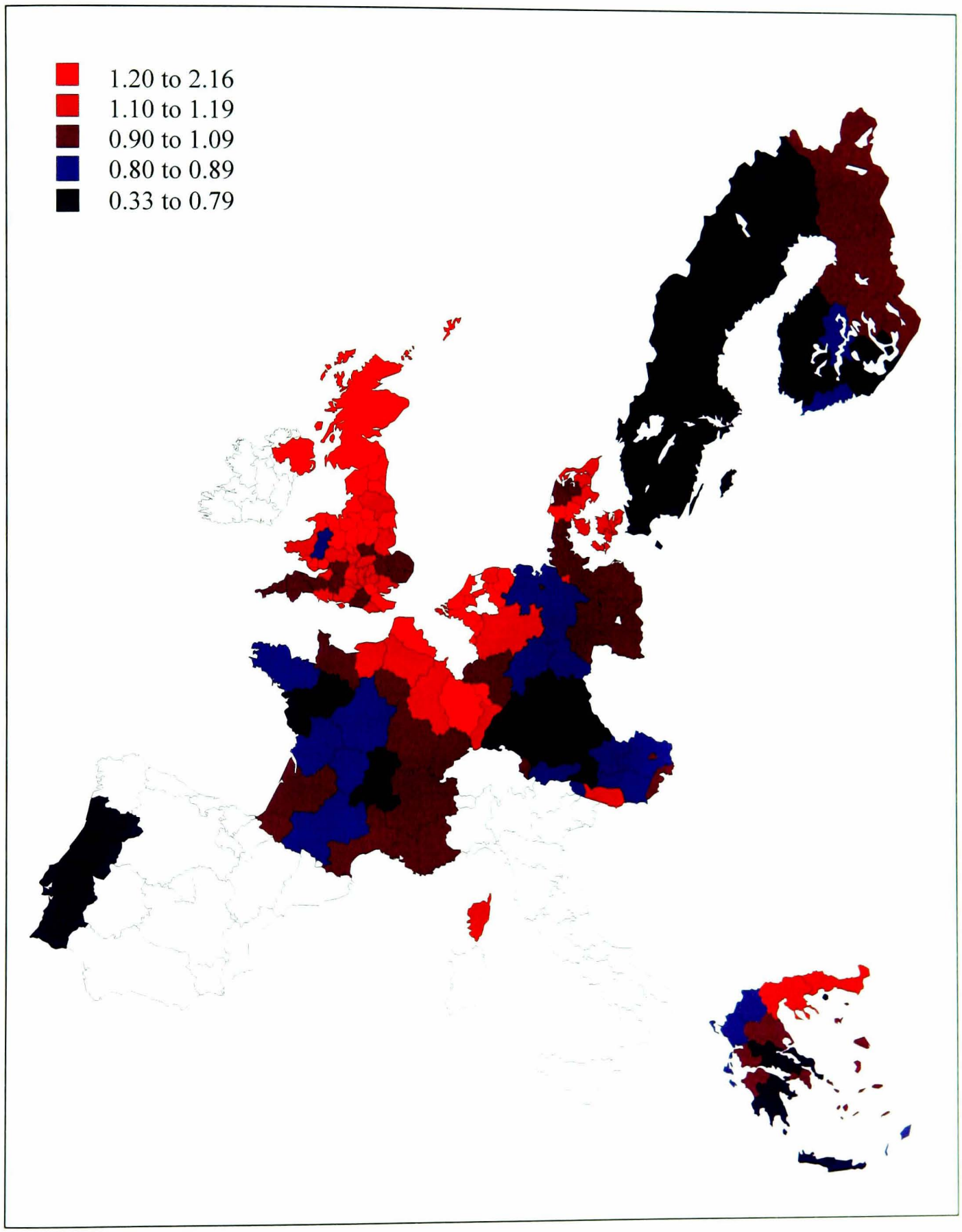
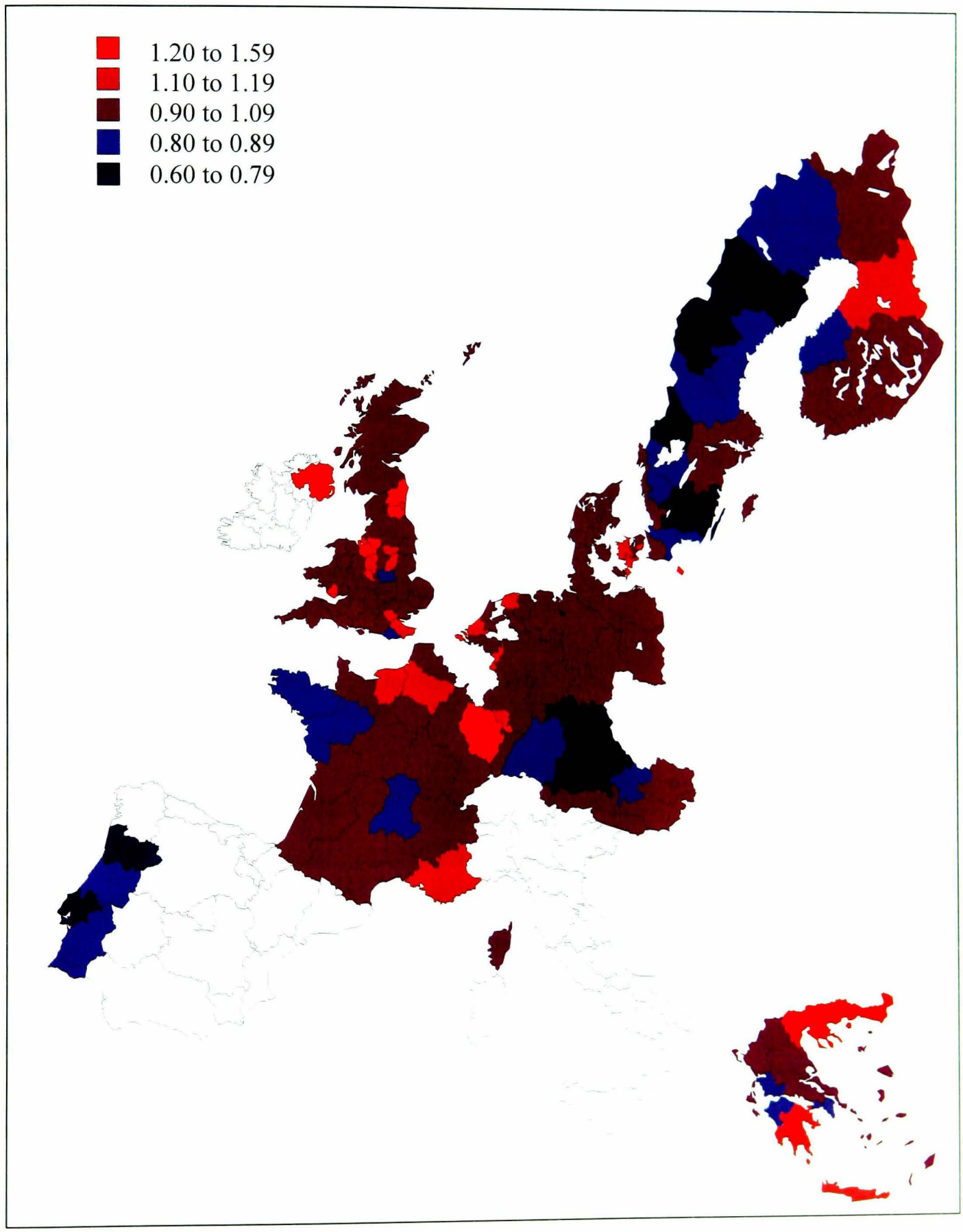


Figure 8.2 Map of lung cancer RRs from null two-level model



The estimates from the two-level model, taking into account potential risk factors, show that the region with the highest relative risk of cancer mortality in the EU countries under investigation is Madeira in Portugal (RR=1.48) closely followed by Copenhagen and Frederiksberg (city) (RR=1.45) and Northern Ireland in the UK (RR=1.44). The lowest relative risks were found in Norte in Portugal (RR=0.70), Bavaria in Germany (RR=0.72) and Vasterbotten in Sweden (RR=0.78). The regions with extreme risks are similar to those identified from the model with no covariates, but the actual estimates have changed somewhat. They are now closer to unity, which is due to the risk and protective factors being measured at the level of country and are therefore helping explain some of the differences between countries.

When comparing the 95% Bayesian credible intervals from the null and full model, it can be seen that after adding covariates many more of the intervals are significantly different from unity. Therefore, these regions have a significantly greater or lesser risk of lung cancer mortality after adjusting for known risk factors. This suggests that there are other factors, not taken account of here, which are affecting lung cancer mortality in areas of the EU or perhaps the distribution of the risk factors within these counties is not constant and it reflects particularly high (or low) regional levels of exposure to the covariates.

8.2.2.5 Lung Cancer Disease Maps II (Two and Three-Level Full Model RRs)

Looking at the maps of relative risks from the null model (Figure 8.2), most of the regions that stand out as 'hotspots' are not actually significant using this modelling method, and since they dominate the map visually they may be misinterpreted as being significant. However, looking at Figure 8.3, in which the significant regions have been named, it can be seen that these tend to be the areas that stand out visually on this smoother map, showing the usefulness of modelling and mapping the lung cancer mortality data in this manner.

Figure 8.4 maps the relative risks from the three-level model. The estimates in the regions within countries that had high overall risks are now closer to country level risk. This creates a different picture of the distribution of lung cancer, with relative risks within some countries now appearing more similar. For example, all of the regions in the UK have risks close to or greater than unity, whereas using the two-level model there was more variation with some regions having relative risks less than 0.9. Similarly, all relative risks in Germany are lower than 1 (Figure 8.4) but before accounting for the higher level they varied much more around unity.

8.2.2.6 Three-level Full Model: Lung Cancer

Looking at the final columns of Table 8.2, the results from adding a further geographical level to the model, as expected it can be seen here that the estimates change. The same picture that emerged on the map is evident, but what is also noticeable is that the 95% credible intervals get much wider. Here we are incorporating spatial clustering effects into the model indirectly by adding a third level that accounts for the fact that regions within a country are more likely to be homogeneous than regions from different countries. The introduction of country-level random effects appears to result in the covariates being measured with more error; this causes uncertainty in the estimation of the parameters associated with the covariates which in turn introduces uncertainty into the residual estimation.

Table 8.2 Relative risks (two and three level full models) of mortality from lung cancer (for each country extreme rates given, ordered by decreasing RR from three-level model)

i	Country	Region	O_i	E_i	$RR_{mean(2)}$	$PI_{95\%(2)}$	$RR_{mean(3)}$	$PI_{95\%(3)}$
3		Carinthia	284	253.6	1.23	(1.03 , 1.65)	1.30	(0.75 , 2.27)
7		Styria	497	560.9	1.14	(0.98 , 1.43)	1.21	(0.75 , 1.98)
...	Austria	...	...	...	...	... ...	...	... ...
2		Burgenland	136	129.6	0.99	(0.83 1.51)	1.12	(0.57 , 2.19)
5		Upper Austria	476	579.8	0.94	(0.81 , 1.22)	1.01	(0.60 , 1.71)
85		Copenhagen*	549	312.9	1.45	(1.20 , 2.75)	1.38	(0.57 , 3.59)
86		Copenhagen	400	298.0	1.25	(1.06 , 1.61)	1.19	(0.71 , 2.09)
	Denmark	...	...	...	...	... ...	...	... ...
87		Frederiksborg	174	156.3	0.96	(0.81 , 1.40)	0.91	(0.47 , 1.82)
98		Viborg	133	127.7	0.94	(0.79 , 1.30)	0.90	(0.50 , 1.70)
126		Oulu	166	174.8	1.14	(0.96 , 1.49)	1.19	(0.74 , 2.01)
124		Lappi	81	82.4	1.08	(0.90 , 1.60)	1.13	(0.63 , 2.14)
	Finland	...	...	...	...	... ...	...	... ...
120		Hame	230	312.2	0.92	(0.78 , 1.19)	0.97	(0.62 , 1.60)
130		Vassa	166	214.9	0.92	(0.77 , 1.28)	0.97	(0.57 , 1.71)
139		Lorraine	1156	843.1	1.23	(1.06 , 1.52)	1.25	(0.82 , 2.08)
134		Haute-Normandie	714	629.7	1.14	(0.98 , 1.46)	1.18	(0.75 , 2.02)
	France	...	...	...	...	... ...	...	... ...
143		Bretagne	964	1159.8	0.85	(0.71 , 1.28)	0.89	(0.48 , 1.77)
142		Pays de la Loire	851	1207.0	0.80	(0.68 , 1.04)	0.84	(0.52 , 1.45)
72		Nth Rhine_Westphalia	9518	8234.9	1.10	(0.97 , 1.33)	0.94	(0.62 , 1.40)
77		Saarland	639	520.5	1.10	(0.92 , 1.53)	0.95	(0.55 , 1.61)
	Germany	...	...	...	...	... ...	...	... ...
75		Baden-Wurttemberg	3181	4548.3	0.78	(0.66 , 1.02)	0.67	(0.41 , 1.07)
76		Bavaria	3900	5466.4	0.72	(0.62 , 0.86)	0.64	(0.43 , 0.94)
157		Macedonia East+	323	271.6	1.23	(1.02 , 1.81)	1.19	(0.57 , 2.42)
158		Macedonia Central	874	765.4	1.18	(0.99 , 1.60)	1.14	(0.58 , 2.14)
	Greece	...	...	...	...	... ...	...	... ...
166		Attica	1619	1545.0	0.90	(0.76 , 1.18)	0.88	(0.46 , 1.59)
163		Greece West	335	359.7	0.90	(0.76 , 1.13)	0.88	(0.48 , 1.52)
238	Luxembourg	...	184	175.7	1.25	(1.03 , 1.77)	1.22	(0.64 , 2.26)
248		Limburg	651	489.4	1.10	(0.94 , 1.49)	1.00	(0.47 , 1.81)
252		Zeeland	209	187.0	1.06	(0.87 , 2.01)	0.98	(0.33 , 2.47)
	Netherlands	...	...	...	...	... ...	...	... ...
241		Groningen	337	262.4	0.96	(0.81 , 1.32)	0.88	(0.40 , 1.60)
249		Utrecht	531	423.1	0.95	(0.80 , 1.29)	0.88	(0.41 , 1.59)
327		Madeira	74	93.8	1.48	(1.23 , 2.43)	1.57	(0.73 , 3.59)
326		Azores	78	95.9	1.42	(1.18 , 2.35)	1.51	(0.70 , 3.47)
	Portugal	...	...	...	...	... ...	...	... ...
323		Lisboa e Vale do Tejo	787	1490.6	0.82	(0.71 , 1.15)	0.87	(0.47 , 1.70)
321		Norte	733	1347.4	0.70	(0.58 , 1.28)	0.74	(0.31 , 1.88)
462		Orebro	112	169.1	1.10	(0.93 , 1.42)	1.09	(0.66 , 1.77)
467		Uppsala	85	138.4	1.09	(0.91 , 1.43)	1.09	(0.65 , 1.80)
	Sweden	...	...	...	...	... ...	...	... ...
455		Jonkoping	73	187.1	0.80	(0.67 1.03)	0.82	(0.49 1.31
469		Vasterbotten	50	141.7	0.78	(0.64 , 1.07)	0.79	(0.44 , 1.36)
565		Northern Ireland	788	650.9	1.44	(1.17 , 2.75)	1.59	(0.68 , 4.13)
513		Northumberland	252	164.2	1.32	(1.12 , 1.71)	1.45	(0.90 , 2.57)
	UK	...	...	...	...	... ...	...	... ...
534		West Sussex	432	437.5	0.86	(0.73 , 1.11)	0.95	(0.59 , 1.67)
521		Leicestershire	428	428.4	0.83	(0.72 , 1.02)	0.93	(0.61 , 1.56)

*Copenhagen and Frederiksberg (city), +Macedonia East and Thrace

Figure 8.3 Map of lung cancer RRs from full two-level model

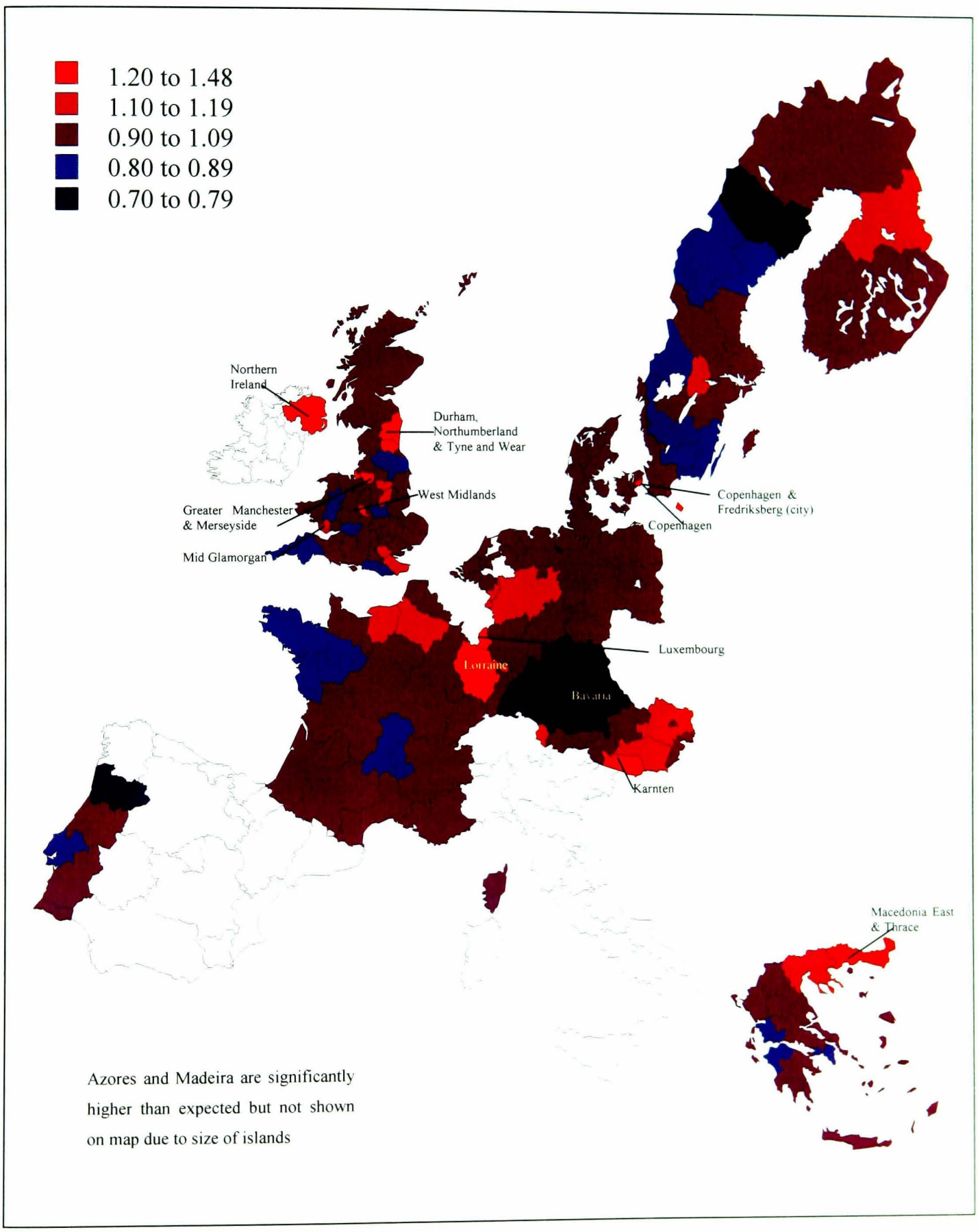
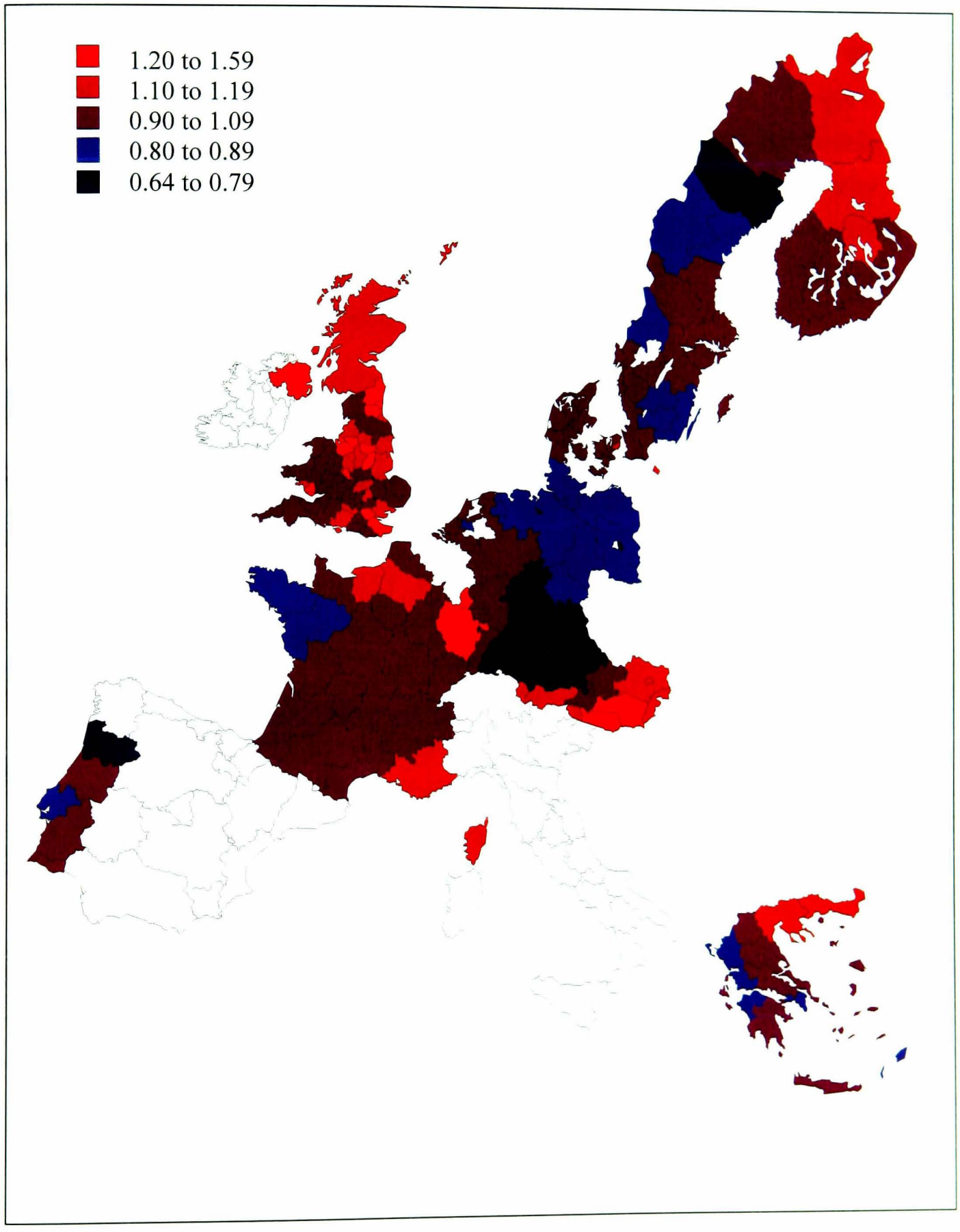


Figure 8.4 Map of lung cancer RRs from full three-level model



8.2.2.7 Lung Cancer Country Level Results

Table 8.3 gives the SMRs and corresponding confidence intervals for each country along with the relative risks and posterior intervals from fitting the 3-level full spatial model. From the modelling results it can be seen that Germany has the lowest lung cancer mortality relative risk (RR=0.83) and Austria has the highest (RR=1.14), closely followed by the UK (RR=1.12). None of the posterior credible intervals exclude unity indicating no country has a risk of lung cancer mortality significantly higher or lower than expected. This is because the within country variation is large for most countries and looking at regional level relative risks within these countries does prove more informative.

Table 8.3 Relative risks of mortality from lung cancer at country level

Country	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
Austria	3278	3633.2	0.90	(0.87 , 0.93)	1.14	(0.87 , 1.49)
Germany	35037	37653	0.93	(0.92 , 0.94)	0.83	(0.66 , 1.01)
Denmark	3206	2649.6	1.21	(1.17 , 1.25)	0.96	(0.72 , 1.32)
Finland	1820	2223.1	0.82	(0.78 , 0.86)	1.04	(0.85 , 1.34)
France	22232	22458	0.99	(0.98 , 1.00)	1.04	(0.84 , 1.40)
Greece	4877	5091.1	0.96	(0.93 , 0.99)	0.97	(0.66 , 1.35)
Luxembourg	184	175.7	1.05	(0.90 , 1.21)	1.05	(0.77 , 1.39)
Netherlands	8416	6632.7	1.27	(1.24 , 1.30)	0.92	(0.58 , 1.24)
Portugal	2270	4459.4	0.51	(0.49 , 0.53)	1.04	(0.78 , 1.46)
Sweden	2812	4944.9	0.57	(0.55 , 0.59)	0.98	(0.75 , 1.25)
UK	39121	28663	1.36	(1.35 , 1.38)	1.12	(0.89 , 1.54)

8.2.2.8 Lung Cancer Parameter Estimates

The parameter estimates from fitting the three models are given in Table 8.4. The estimates and 95% credible intervals are given for the two-level spatial model with no explanatory variables, then after including all six explanatory variables and finally including the higher country level. The overall intercept β_0 represents a logarithm of the average number of lung cancer deaths in all regions included in the study in addition to the centred logarithm of the expected deaths, when all other fixed coefficients are zero. The estimates of the intercept are not particularly informative as they reflect an unlikely situation where-by an area has zero exposure to any of the risk or protective factors.

Looking at the full two-level model first, the estimate of β_1 is the mean, or fixed slope for the explanatory variable smoke and it can be seen the estimate is positive and, judging significance from the 95% posterior credible intervals, significant. This implies that when taking the other variables into account an increase in cigarette consumption increases lung cancer mortality on average in the EU. The parameter estimate of 0.001 is a log relative risk of lung cancer mortality for each one unit/cigarette smoked increase per person per year. This suggests that every 100 increase in cigarettes smoked per person per year is associated with an increase in the risk of lung cancer mortality of about 11% (RR=exp{0.1}=1.105). It can be seen that other variables which significantly affect cancer mortality are fruit and vegetable consumptions, both showing an inverse association with lung cancer mortality. The actual effect size of these variables will be discussed later on.

Looking at the random parts of the models it can be seen that, for the null and full two-level models, the variance has been partitioned into that which is due to differences between regions, σ_u^2, and that which arises due to the spatial structure in lung cancer mortality, σ_v^2. To interpret the spatial effects, the average number of neighbours a region has should be taken into account, and after doing so the adjusted variance from the spatial part of the model is 0.046 ($\sigma_v^2 / \bar{n}$ =

0.191/4.128) and the total variance under the null model is then 0.065 ($=\sigma_u^2 + \sigma_v^2 / \bar{n}$ = 0.019 + 0.046). Seventy-one percent (0.046/0.065 *100) of this variance arises from spatial effects. Looking at σ_u^2 and σ_v^2 from the full two-level model, the total variance decreases to 0.033 (0.013 + 0.081/4.128). Therefore, taking into account the measures of exposure to risk and protective factors has reduced the variation by around 50%. Of this remaining variance (0.033), 61% is attributable to spatial patterning in the data. Finally looking at the results from the three-level model, it can be seen that the significant fixed parameter estimates do not change very much from the two-level model. However, the random part has now been partitioned further and σ_y^2 now represents the variance that is due to differences between countries. The total variance is now around 0.058, which has nearly doubled after adding a third higher level. Now, 18% of the total variance is due to heterogeneity between regions, 30% is attributable to spatial effects and 52% is due to differences between countries.

8.2.2.9 Lung Cancer Risk Factor Effect Size

It is also of interest to quantify the effect size of the risk or protective factors when examining lung cancer mortality rates. Using both the two-level and three-level models, relative risks were calculated for each variable comparing countries with high and low levels of exposure to risk factors and are given in Table 8.5. Firstly, looking at the two-level model and the smoking variable, the table is showing that Greece has the highest level of cigarette consumption in these eleven countries and Sweden has the lowest. The relative risk of 7.69 suggests that consuming, on average, the same amount of cigarettes as Greece leads to a relative risk of lung cancer mortality that is 7.7 times as high as if smoking was on the same level as Sweden. Also, looking at the fruit variable suggests that a population consuming, on average, the same amount of fruit as Greece has a risk of lung cancer mortality that is 51% lower than if consumption was on the same level as the UK. The final significant variable is vegetable consumption and the parameter estimates from the two-level model suggest that consuming the same level of vegetables as Greece

leads to a risk of lung cancer mortality that is 57% less than if consumption was on the same level as Finland.

The results are also given for the three-level model and as can be seen the estimates change slightly. Also, as noted for the parameter estimates, the 95% posterior credible intervals are wider. Adding a third higher level to the model results in increasing the variation, hence reducing the precision of the relative risks.

There may be implications for ignoring (or even including) country in the multilevel structure when fitting such spatial models. The non-independence of regional mortality within countries has been taken into account, firstly through allowing for spatial autocorrelation between neighbouring areas and secondly by using country level covariates. It has been shown that both aspects of the modelling are important; we have seen that there is significant spatial variation existing in the data suggesting there is evidence that spatial patterning of the disease does exist and the importance of the country effects shows that there are discontinuities between countries. The results seen so far in this chapter, along with those from Chapter 6, suggest that different models may be useful for the estimation of different effects. Adding a third level appears to be a useful model for explaining differences between countries; and, at least for lung cancer mortality, there is a higher level of heterogeneity between countries than between regions within countries. The spatial model, without the additional level, appears more useful at detecting differences between regions, such as disease 'hotspots' before and after adjusting for covariates, and gives more precise estimates of the fixed effects.

Table 8.4 Parameter estimates from modelling lung cancer mortality rates

Parameters		Null two-level Model		Full two-level Model		Full three-level Model	
		Estimate	**Credible Interval**	**Estimate**	**Credible Interval**	**Estimate**	**Credible Interval**
β_0		5.85	(5.76 , 5.93)	4.93	(4.43 , 5.39)	4.49	(3.47 , 5.42)
β_1	(smoke)			**0.0010**	**(0.0007 , 0.0013)**	**0.0011**	**(0.0006 , 0.0014)**
β_2	(fruit)			**-0.0104**	**(-0.0135 , -0.0076)**	**-0.0100**	**(-0.0156 , -0.0024)**
β_3	(veg)			**-0.0035**	**(-0.0051 , -0.0021)**	**-0.0040**	**(-0.0069 , -0.0013)**
β_4	(animal)			0.0085	(-0.0100 , 0.0272)	0.0120	(-0.0090 , 0.0464)
β_5	(alcohol)			-0.0003	(-0.0026 , 0.0020)	-0.0003	(-0.0046 , 0.0038)
β_6	(gdp)			3.44e-6	(-3.16e-6 , 1.02e-5)	2.55e-6	(-3.42e-6 , 8.62e-6)
σ_u^2		0.0185	(0.0102 , 0.0294)	0.0125	(0.0075 , 0.0190)	0.0105	(0.0062 , 0.0162)
σ_{uv}		0.0496	(0.0328 , 0.0708)	0.0198	(0.0108 , 0.0315)	0.0164	(0.0081 , 0.0268)
σ_v^2		0.1909	(0.1392 , 0.2585)	0.0814	(0.0484 , 0.1236)	0.0728	(0.0432 , 0.1109)
σ_y^2						0.0298	(0.0034 , 0.1134)

Table 8.5 Effect size of covariates from two- and three-level full lung cancer models

Covariate	Min	(country/ region)	Max	(country/ region)	Two-level model		Three-level model	
					RR	95% CI	RR	95% CI
Smoking	**1550**	**(Sweden)**	**3590**	**(Greece)**	**7.69**	**(4.15 , 12.7)**	**9.43**	**(3.61 , 19.3)**
Fruit	**74.5**	**(UK)**	**142.6**	**(Greece)**	**0.49**	**(0.40 , 0.60)**	**0.51**	**(0.34 , 0.85)**
Vegetable	**58.8**	**(Finland)**	**300.4**	**(Greece)**	**0.43**	**(0.29 , 0.60)**	**0.38**	**(0.19 , 0.73)**
Animal fat	2.3	(Greece)	26.8	(Luxembourg)	1.23	(0.78 , 1.95)	1.61	(0.80 , 3.12)
Alcohol	60.0	(Greece)	173.9	(Germany)	0.97	(0.75 , 1.25)	0.97	(0.60 , 1.55)
GDP	5611	(Epirus - Greeece)	44711	(Copenhagen and Frederikberg (city) - Denmark)	1.14	(0.88 , 1.49)	1.10	(0.87 , 1.40)

8.2.3 Lung Cancer Discussion

It has been shown that spatial variation in lung cancer mortality exists both within and between countries throughout the EU. Whilst a large part of the variation can be accounted for by adjusting for known risk and protective factors for lung cancer, variation still clearly exists throughout the European countries under examination. Some regions in Portugal, Germany and Sweden display particularly low lung cancer mortality rates, so it may be of interest to investigate why these areas avoid lung cancer mortality. Probably of more interest are the areas that stand out as lung cancer mortality 'hotspots' after taking account of factors that are known to affect the disease rates. There are twelve regions in the Austria, Denmark, France, Germany, Greece and the UK that show a particularly high risk of lung cancer mortality and it may be of a public health interest to investigate these areas further.

It is of obvious public health advantage to promote smoking cessation in the EU countries for the prevention of many diseases, particularly lung cancer, and similarly with healthy eating promotions. However, it is of interest to determine any other factors that are affecting EU lung cancer mortality rates. A possible cause of the lung cancer 'hotspots' remaining after adjusting for country level smoking and dietary habits is that exposure to the factors may be particularly high or low within these countries. This could be determined by carrying out smaller scale studies within these countries and specific smoking cessation promotion focus may then have to be made in these 'hotspot' areas.

Current information does not support screening for lung cancer (154). A recent lung cancer screening recommendation statement (155) suggested that some types of screening would be more likely to detect lung cancer at an early stage than would be detected in an unscreened population; however, they also found poor evidence that any screening strategy for lung cancer decreases mortality. Therefore, geographical differences in the provision of screening programmes is not a factor that is likely to be affecting the EU pattern of lung cancer mortality. Because of the very high fatality from the disease other health

care provisions, such as variations in treatment, are also unlikely to be affecting the mortality rates.

The association between lung cancer and specific occupations is well established in reports dating back to the 1950s (44). The IARC Monographs Programme (156) has reviewed many of the associations between occupational agents and different types of cancers and it can be seen that risk is increased among workers employed in a number of industries and occupations. The types of occupations are too specific to be taken into account of in a population based study; however, this risk factor may explain some of the lung cancer 'hotpots'.

There is abundant evidence in existing literature of the strong associations lung cancer has with smoking, fruit and vegetable consumption. These are also the factors that have been shown in this study to have a significant association with lung cancer mortality at the population level. It was seen that around half of the variation that exists between regions in the EU can be explained by taking into account fairly crude measures of exposure to the above significant risk and protective factors.

Further modelling showed that around 60% of the remaining variance is spatially patterned and this suggests that there are other factors not included in this study which are spatially patterned that also influence lung cancer mortality. Such factors which may be affecting these spatial patterns of mortality could be factors previously discussed or genetic predisposition to lung cancer. It has been established that a gene that predisposes lung cancer does exist and recent research in the area identified the location on the chromosomes that carry this gene (157). The precise gene has yet to be pinpointed but this research complements other work that suggests that genetic predisposition is a risk factor for lung cancer. This risk factor is likely to be affecting mortality rates in the EU in some form. Many gene frequencies are spatially patterned (158) and this leads to the possibility that this is a risk factor that may be causing some of the unexplained spatially patterned variance. Further research is necessary in this area.

8.3 Colorectal Cancer Mortality

Cancer of the colon and rectum is the second most common cancer in both men and woman in Europe (146). In men, colorectal cancer comprises 22% of all cancer cases and 14% in woman (2). There are major between-country differences in colorectal cancer survival rates in Europe (159). The countries with highest survival have a five-year survival rate which is less than 60% (160) and the overall European five-year survival is very similar in colon cancer (51%) and rectal cancer (48%). European deaths from colorectal cancer are ranked the second most common cause of cancer deaths, with around 11% of all cancer mortality being due to neoplasms in these sites (2). Second to cancer of the lung, colorectal cancer demonstrates one of the most serious cancer burdens in Europe.

8.3.1 Modelling Colorectal Cancer Mortality

It is of interest to examine the true distribution of colorectal cancer both within and between countries in Europe. Again, we are able to do so by modelling colorectal cancer mortality relative risks after adjusting for spatial patterning and potential risk factors.

As discussed in chapter 2, alcohol intake and different dietary components have been shown to be the main risk factors for the disease and data reflecting levels of these have been modelled. There is conflicting evidence on the effect smoking has on colorectal cancer mortality; this has been included as a covariate in the latter models to determine its effect on European colorectal cancer mortality rates. There is no evidence of socio-economic status affecting colorectal cancer, and GDP has not been included in the model results shown in this section as adding this covariate slowed convergence time substantially. The variables were fitted as covariates when fitting the fully Bayesian spatial multilevel models. The same sets of models are fitted as were examined for lung cancer mortality. Running simultaneous models and monitoring convergence resulted in a burn-in period of 200,000 being needed for the two-level null model. These samples were

discarded and 100,000 further iterations were run; at this stage the Monte Carlo error as a percentage of the posterior standard deviation was much less than 5% so fewer further iterations were actually required to obtain a suitable sample of reliable posterior estimates. The full two-level model required a burn-in of 200,000 iterations and 100,000 further iterations. For the full three-level model 900,000 iterations were discarded as a burn-in period and it took a further 300,000 iterations until a suitable posterior distribution was available.

8.3.2 Model Results: Colorectal Cancer

Relative risks for colorectal cancer mortality are presented in Tables 8.6 and 8.8. As with the lung cancer results, estimates are given for selected regions within each country. Table 8.6 shows the standardised mortality ratios and respective 95% confidence intervals and the posterior mean relative risks obtained from the null two-level model and corresponding 95% posterior credible intervals. Similarly Table 8.8 gives the relative risks and credible intervals from the two-level and three-level models including the five covariates discussed previously. Table 8.7 gives the eleven country level SMRs and relative risks from the posterior estimates from the full three-level model. Table 8.9 presents the posterior means and 95% credible intervals for each of the parameters from these three models. Table 8.10 shows effect sizes of the parameter estimates by giving relative risks comparing areas with high and low levels of exposure to the risk factors. Figures 8.5-8.8 are maps of the SMRs and posterior relative risks from the three models.

8.3.2.1 Two Level Null Model: Colorectal cancer Model

Table 8.6 give the two highest and two lowest relative risks within each country and corresponding intervals in brackets from the two-level spatial model with no covariates. It can be seen that the area with the highest relative risk of colorectal cancer mortality is Nord-Pas-de-Calais in France, with a relative risk of 1.61,

followed by Burgenland in Austria (RR=1.43), Cornwall in the UK (RR=1.43) and Storstrøm in Denmark (RR=1.42). Before taking account of risk factors, the areas with the lowest risk of colorectal cancer mortality is Norte in Portugal (RR=0.28) followed by the Algarve in Portugal (RR=0.48) and Dytiki Ellada and Ionian Islands in Greece (RR=0.50). Regions in Portugal show the most variability in relative risks ranging from 0.28 to 1.17. Finland have the smallest range in relative risks (0.73 – 0.98) closely followed by Sweden (0.81 – 1.07).

8.3.2.2 Colorectal Cancer SMRs

Standardised mortality ratios are given for each region and presented in Table 8.6. As previously discussed, the relative risks can be much more reliably interpreted as close to the true risk of cancer mortality; however, they are presented here for comparison purposes. It is of interest to show the discrepancies between the 'traditional' measure of disease risk and the modern spatial modelling approaches to estimate the risk. Most of the extreme predicted posterior risks differ substantially from the SMRs, often giving risk estimates on the opposite side of unity for the same region. This emphasises the importance of using accurate modelling methods to examine disease distribution.

8.3.2.3 Colorectal Cancer Disease Maps I (SMRs and Two-Level Null Model RRs)

Figure 8.5 presents the map of the SMRs and it appears to indicate that colorectal cancer mortality is generally low in the EU. The SMRs range from 0.13 to 1.27 and it can be seen that the blue regions (SMR < 0.89) tend to dominate the map. The few regions with high rates are confined to France and Denmark. Again, clustering within countries is evident from this map. However, it is more informative to examine the modelled rates.

Table 8.6 Relative Risks (SMR and RR from two-level null model) of mortality from colorectal cancer (for each country extreme rates given, ordered by decreasing RR)

i	Country	Region	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
2		Burgenland	115	112.8	1.02	(0.84 , 1.22)	1.43	(0.91 , 2.26)
4		Lower Austria	615	620.6	0.99	(0.91 , 1.07)	1.22	(0.96 , 1.56)
	Austria		...	...	...	... ...	...	... ...
7		Styria	389	486.1	0.80	(0.72 , 0.88)	0.98	(0.78 , 1.23)
3		Carinthia	164	216.6	0.76	(0.65 , 0.88)	0.85	(0.61 , 1.18)
90		Storstrøm	112	125.5	0.89	(0.73 , 1.07)	1.42	(1.03 , 1.97)
85		Copenhagen	318	308.1	1.03	(0.92 , 1.15)	1.38	(0.73 , 2.69)
	Denmark		...	...	...	... ...	...	... ...
99		Nordjylland	229	212.0	1.08	(0.94 , 1.23)	0.91	(0.54 , 1.52)
96		Ringkobing	94	105.6	0.89	(0.72 , 1.09)	0.88	(0.66 , 1.16)
129		Uusimaa	225	409.1	0.55	(0.48 , 0.63)	0.98	(0.74 , 1.32)
121		Keski-Suomi	54	92.9	0.58	(0.44 , 0.76)	0.95	(0.71 , 1.28)
	Finland		...	...	...	... ...	...	... ...
126		Oulu	61	135.9	0.45	(0.34 , 0.58)	0.78	(0.57 , 1.08)
122		Pohjois-Savo	35	96.3	0.36	(0.25 , 0.51)	0.73	(0.50 , 1.04)
138		Nord - Pas-de-Calais	1114	927.3	1.20	(1.13 , 1.27)	1.61	(0.73 , 3.58)
141		Franche-Comté	279	285.7	0.98	(0.87 , 1.10)	1.31	(0.99 , 1.70)
	France		...	...	...	... ...	...	... ...
134		Haute-Normandie	427	422.6	1.01	(0.92 , 1.11)	1.04	(0.80 , 1.35)
143		Bretagne	864	801.2	1.08	(1.01 , 1.15)	1.02	(0.65 , 1.66)
84		Thuringen	878	988.8	0.89	(0.83 , 0.95)	1.26	(0.98 , 1.63)
75		Baden-Wurttemberg	3429	3961.7	0.87	(0.84 , 0.89)	1.19	(0.88 , 1.60)
	Germany		...	...	...	... ...	...	... ...
69		Hamburg	745	813.7	0.92	(0.85 , 0.98)	0.91	(0.60 , 1.38)
78		Berlin	1211	1439.4	0.84	(0.79 , 0.89)	0.91	(0.42 , 1.93)
158		Macedonia Central	241	582.7	0.41	(0.36 , 0.47)	0.96	(0.68 , 1.38)
165		Peloponnese	77	360.7	0.21	(0.17 , 0.27)	0.93	(0.58 , 1.48)
	Greece		...	...	...	... ...	...	... ...
162		Ionian Islands	21	109.1	0.19	(0.12 , 0.29)	0.50	(0.28 , 0.86)
163		Peloponnese	59	299.5	0.20	(0.15 , 0.25)	0.50	(0.37 , 0.67)
238	Luxembourg		110	146.0	0.75	(0.62 , 0.91)	1.01	(0.65 , 1.52)
252		Zeeland	104	157.3	0.66	(0.54 , 0.80)	1.23	(0.57 , 2.69)
248		Limburg	320	378.0	0.85	(0.76 , 0.94)	1.18	(0.85 , 1.63)
	Netherlands		...	...	...	... ...	...	... ...
242		Friesland	159	234.5	0.68	(0.58 , 0.79)	0.93	(0.67 , 1.28)
250		Noord-Holland	648	899.3	0.72	(0.67 , 0.78)	0.87	(0.54 , 1.39)
323		Lisboa e Vale do Tejo	894	1170.3	0.76	(0.71 , 0.82)	1.17	(0.86 , 1.59)
322		Centro	364	742.1	0.49	(0.44 , 0.54)	1.01	(0.85 , 1.20)
	Portugal		...	...	...	... ...	...	... ...
325		Algarve	73	153.6	0.48	(0.37 , 0.60)	0.48	(0.26 , 0.89)
321		Norte	594	1066.3	0.56	(0.51 , 0.60)	0.28	(0.16 , 0.50)
460		Malmöhus	258	386.4	0.67	(0.59 , 0.75)	1.07	(0.62 , 1.85)
470		Västernorrland	85	137.9	0.62	(0.49 , 0.76)	1.06	(0.70 , 1.60)
	Sweden		...	...	...	... ...	...	... ...
448		Älvsborg	130	218.1	0.60	(0.50 0.71)	0.82	(0.61 1.09)
469		Västerbotten	78	114.5	0.68	(0.54 , 0.85)	0.81	(0.57 , 1.14)
543		Cornwall	189	246.2	0.77	(0.66 , 0.89)	1.43	(0.69 , 3.06)
565		Northern Ireland	427	536.3	0.80	(0.72 , 0.88)	1.35	(0.67 , 2.73)
	UK		...	...	...	... ...	...	... ...
546		Somerset	163	236.4	0.69	(0.59 , 0.80)	0.95	(0.71 , 1.25)
528		Hertfordshire	266	389.0	0.68	(0.60 , 0.77)	0.90	(0.66 , 1.19)

Figure 8.6 maps the relative risks estimated from the two-level null spatial model and a very different picture of the distribution of colorectal cancer is evident. The relative risks range from 0.28 to 1.61 and the pattern on the overall map is displaying a high amount of variation both within and between counties. The map is now showing fewer areas with extreme low risk of cancer mortality and more areas with high risks of this disease (RR > 1.1). Countries that appear to have areas with very high risk (RR > 1.2) of colorectal cancer mortality before taking into account potential risk factors for the disease are France, Germany, Austria, Denmark and the UK. These regions could now be viewed as colorectal cancer mortality 'hotspots'. Some regions in Finland, Greece and Portugal display very low risk of colorectal cancer mortality (RR < 0.8). Taking into account risk and protective factors of colorectal cancer will hopefully help explain some of the variation that is clearly evident on this map.

8.3.2.4 Two-Level Full Model: Colorectal Cancer

As previously discussed, there are risk factors that have been shown to affect the risk of colorectal cancer mortality but so far these have been ignored in the modelling. Table 8.8 gives the estimated relative risks of colorectal cancer from the two-level spatial model including five of the covariates used in previous model fitting; fruit, vegetable, animal fat and alcohol consumption and cigarette smoking. Again, the two highest and two lowest relative risks within each country are given. The table also shows the estimates from the full three-level model; these estimates take account of the within country clustering that was evident from the initial map of SMRs.

Figure 8.5 Map of colorectal cancer SMRs

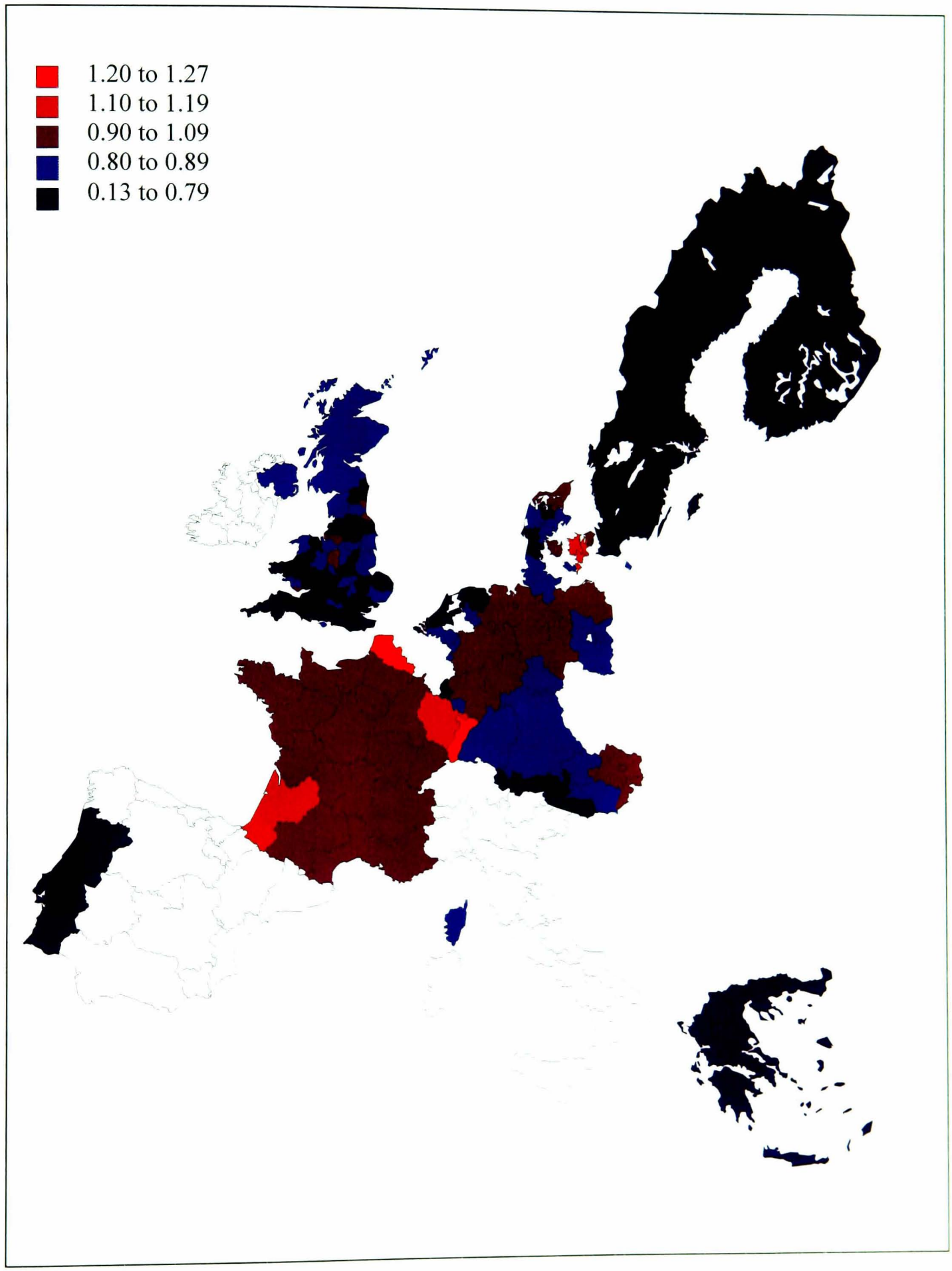
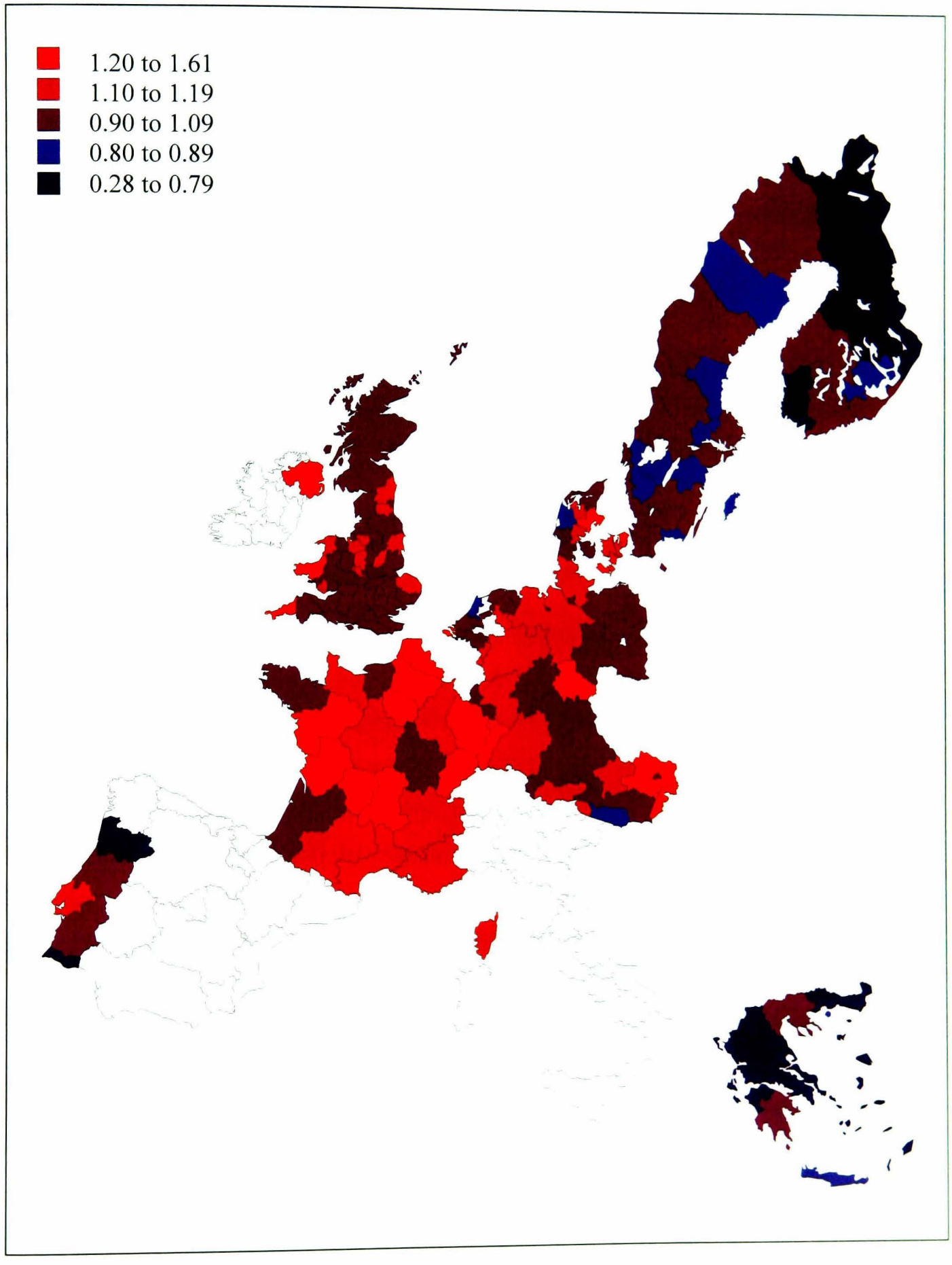


Figure 8.6 Map of colorectal cancer RRs from null two-level model



The estimates from the two-level model taking the covariates into account show that the region with the highest relative risk of mortality in the EU countries under investigation is Lisbon and Vale do Tejo in Portugal (RR=1.49), followed by Nord-Pas-de-Calais in France (RR=1.39) and Lower Austria (RR=1.33). The lowest relative risks were found in Norte in Portugal (RR=0.54) followed by Greece West (RR=0.69) and Ionian Islands (RR=0.72) in Greece. Many of the regions identified as extreme before adding the covariates to the model remain as extreme within their country. However, they now tend to be less extreme as the covariates are helping to explain some of the country level variation.

Regions within Portugal still show the most variability, with the relative risk of cancer mortality ranging from 0.54 to 1.49. Regions within Finland display the least variability, with all relative risks ranging from 0.91 to 1.06.

8.3.2.5 Three-level Full Model: Colorectal Cancer

Looking at the final columns of Table 8.8, which are the relative risks from adding a third level to the model (country), it can be seen both the estimates and the credible intervals have changed somewhat when compared with the two-level full results. The relative risks now take into account the fact that regions within a country are more homogenous than regions from different countries. The estimates are shrunk towards the overall relative risk of the country. All regions in France, which displayed the most areas with very high risks (see Figure 8.6), now have predicted relative risks which are fairly high. On the other hand, Finland's relative risks have shrunk somewhat based on the fact that they had many regions with very low risk of colorectal cancer mortality; the variability between regions has reduced through the addition of country as a level resulting in there is now being less variability within countries.

The 95% credible intervals for these relative risks again appear to have widened so there is now much more uncertainty attached to the risks, which is likely to be due to the covariance that exists between the country level random

effects and fixed effects; this will be causing uncertainty in the estimation of the parameters associated with the covariates which leads to uncertainty in the relative risks.

8.3.2.6 Colorectal Cancer Disease Maps II (Two and Three-Level Full Model RRs)

Figure 8.7 maps the relative risks from the full two-level model. The relative risks that stand out on the map are the bright red regions (RR > 1.2) and these are areas that possibly need further investigation into why they remain as colorectal cancer mortality 'hotspots' after adjusting for the potential risk and protective factors. There are ten such regions in Portugal, France, Austria and Greece, and areas of low risk also stand out on the map; there are such regions in Finland, Portugal and Greece.

The 'hotspot' areas can now be clearly identified on the map because the map that has been produced has been spatially smoothed and variation has been reduced through adjusting for colorectal cancer risk factors. The smooth map is dominated by purple regions in which the relative risks are close to unity; none of these risks are significantly different from unity based on 95% posterior credible intervals. In fact, the only relative risks significantly different from unity are the extreme risks, ie those greater than 1.2 or less than 0.80.

Finally examining the map of relative risks from the three-level model (Figure 8.8) it can be seen that most countries now display a high amount of clustering within the country. So the regional relative risks have been drawn towards the overall country level risk. Countries that stand out as still having high variability within them are Portugal and Greece. Looking at the results from the two-level model (Table 8.8), the regions within these countries did have the highest amount of variability in relative risks, 0.54-1.49 and 0.69-1.27 respectively.

Although this map is smooth in that it appears to be free of random noise and not dominated by natural variation it is of limited use for the purpose of identifying regional 'hotspots' of the disease.

8.3.2.7 Colorectal Cancer Country Level Results

The table below (Table 8.7) gives the SMRs and corresponding confidence intervals for each country and the relative risks and posterior intervals from fitting the three-level full model. The modelling results coincide with the map obtained from fitting the three-level model (Figure 8.8) in that France has the highest relative risk (RR=1.51) of colorectal cancer mortality by far. None of the other countries display RRs significantly different from unity based on the 95% posterior credible intervals.

Table 8.7 Relative risks of mortality from colorectal cancer at country level

Country	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
Austria	2785	3224.2	0.86	(0.83 , 0.90)	1.09	(0.77 , 1.67)
Germany	29867	33190.8	0.90	(0.89 , 0.91)	0.95	(0.71 , 1.40)
Denmark	2077	2207.8	0.94	(0.90 , 0.98)	0.96	(0.65 , 1.64)
Finland	966	1843.6	0.52	(0.49 , 0.56)	0.71	(0.50 , 0.94)
France	15778	15184.4	1.04	(1.02 , 1.06)	1.51	(1.15 , 2.08)
Greece	1234	4087.3	0.30	(0.29 , 0.32)	0.87	(0.47 , 1.65)
Luxembourg	110	146.0	0.75	(0.62 , 0.91)	1.09	(0.71 , 2.13)
Netherlands	4090	5401.1	0.76	(0.73 , 0.78)	1.01	(0.47 , 1.59)
Portugal	2151	3554.6	0.61	(0.58 , 0.63)	1.03	(0.64 , 1.64)
Sweden	2499	4185.2	0.60	(0.57 , 0.62)	1.01	(0.68 , 1.46)
UK	19220	24308.9	0.79	(0.78 , 0.80)	0.98	(0.61 , 1.41)

8.3.2.8 Colorectal Cancer Parameter Estimates

The parameter estimates from fitting the three models are given in Table 8.9 along with their posterior credible intervals. β_o represents the overall intercept, but however these values are not very informative as they reflect a situation where an area has zero exposure to all the risk factors. The other fixed parameter estimates are of more interest. Looking at the full two-level model first it can be seen that the estimate of the fixed slope for the smoke variable is positive and, judging from the 95% posterior interval, significant. Therefore, taking the other variables into account, an increase in cigarette smoking increases colorectal cancer on average in the EU. The parameter estimate of 0.0006 is a log relative risk of lung cancer mortality for each one cigarette smoked per person per year. This suggests that every increase of 100 cigarettes smoked per person per year is associated with an increase in the risk of colorectal cancer mortality of about 6% (RR=exp{0.06}=1.06). It can be seen that other variables which significantly affect colorectal cancer mortality are fruit consumption (RR=0.91), vegetable consumption (RR=0.97) and animal fat consumption (RR=1.39); all are based on a 10kg increase in the food or drink per person per year.

Looking at the random part of the null and two-level model, the variance has been partitioned into that which is due to regional differences, σ_u^2, and that which is due to the spatial structure of colorectal cancer mortality, σ_v^2. The spatial variance under the null model, after taking the regional average number of neighbours into account, is 0.040, resulting in the total variance being 0.056. Therefore, 71% of the total variance in colorectal cancer mortality, before taking risk factors into account, arises from spatial effects. The total variance from the full two-level model is 0.028. Therefore, taking into account the measures of exposure to the various risk factors has reduced the overall variation in colorectal cancer mortality by 50%. Of this remaining variation 68% is attributable to spatial patterning in the data.

Table 8.8 Relative risks (RR from two- and three-level full models) of mortality from colorectal cancer (for each country extreme rates given, ordered by decreasing RR from three-level model)

i	Country	Region	O_i	E_i	$RR_{mean(2)}$	$PI_{95\%(2)}$	$RR_{mean(3)}$	$PI_{95\%(3)}$
4		Lower Austria	615	620.6	1.33	(1.07 , 1.67)	1.25	(0.76 , 2.28)
2		Burgenland	115	112.8	1.30	(0.90 , 1.91)	1.22	(0.67 , 2.47)
	Austria							
6		Salzburg	119	172.3	1.06	(0.84 , 1.34)	1.02	(0.61 , 1.86)
3		Carinthea	164	216.6	1.01	(0.78 , 1.33)	1.01	(0.59 , 1.87)
90		Storstrom	112	125.5	1.09	(0.83 , 1.43)	1.04	(0.57 , 2.19)
85		Copenhagen*	318	308.1	1.00	(0.57 , 1.73)	1.03	(0.47 , 2.65)
	Denmark							
93		Sonderjylland	76	106.7	0.84	(0.66 , 1.07)	0.89	(0.50 , 1.82)
96		Ringkobing	94	105.6	0.80	(0.62 , 1.02)	0.88	(0.49 , 1.81)
129		Uusimaa	225	409.1	0.91	(0.72 , 1.14)	0.74	(0.44 , 1.16)
120		Hame	154	264.6	0.88	(0.69 , 1.13)	0.74	(0.44 , 1.17)
	Finland							
124		Lappi	31	62.6	0.81	(0.56 , 1.16)	0.66	(0.35 , 1.13)
122		Kuopio	35	96.3	0.77	(0.57 , 1.01)	0.65	(0.37 , 1.04)
138		Nord-Pas-de-Calais	1114	927.3	1.39	(0.77 , 2.54)	1.68	(0.86 , 3.50)
140		Alsace	495	390.6	1.22	(0.97 , 1.57)	1.63	(1.04 , 2.67)
	France							
151		Provence-Alpes-Cote d'Azur	1179	1313.4	1.04	(0.85 , 1.27)	1.35	(0.86 , 2.21)
152		Corsica	64	75.9	1.03	(0.60 , 1.78)	1.30	(0.67 , 2.64)
72		Nth-Rhine Westphalia	6841	7209.2	1.09	(0.91 , 1.29)	1.00	(0.66 , 1.67)
83		Saxony-Anhalt	1078	1114.9	1.07	(0.85 , 1.35)	1.00	(0.63 , 1.73)
	Germany							
76		Bavaria	4085	4757.4	0.94	(0.81 , 1.10)	0.92	(0.61 , 1.52)
78		Berlin	1211	1439.4	0.85	(0.47 , 1.54)	0.85	(0.42 , 1.86)
158		Central Macedonia	241	582.7	1.27	(0.94 , 1.72)	1.10	(0.48 , 2.66)
157		Macedonia East+	83	208.1	1.04	(0.71 , 1.53)	1.01	(0.42 , 2.53)
	Greece							
162		Ionian Islands	21	109.1	0.72	(0.46 , 1.10)	0.75	(0.30 , 1.89)
163		Greece West	59	299.5	0.69	(0.53 , 0.87)	0.72	(0.32 , 1.62)
238	Luxembourg		110	146.0	0.99	(0.72 , 1.35)	1.09	(0.57 , 2.68)
248		Limburg	320	378.0	1.17	(0.90 , 1.55)	1.11	(0.43 , 2.13)
244		Overijssel	311	362.3	1.08	(0.90 , 1.32)	1.06	(0.43 , 1.93)
	Netherlands							
249		Utrecht	238	350.2	0.99	(0.74 , 1.31)	0.94	(0.36 , 1.81)
250		Noord-Holland	648	899.3	0.90	(0.62 , 1.31)	0.93	(0.34 , 1.88)
323		Lisboa e Vale do Tejo	894	1170.3	1.49	(1.10 , 2.04)	1.32	(0.64 , 2.75)
324		Alentejo	157	268.6	1.19	(0.95 , 1.49)	1.11	(0.57 , 2.15)
	Portugal							
326		Azores	40	75.6	0.86	(0.53 , 1.40)	0.87	(0.37 , 2.01)
321		Norte	594	1066.3	0.54	(0.31 , 0.92)	0.73	(0.29 , 1.77)
466		Stockholm	473	706.3	1.04	(0.75 , 1.43)	1.06	(0.58 , 1.90)
470		Vasternorrland	85	137.9	1.06	(0.77 , 1.46)	1.05	(0.57 , 1.89)
	Sweden							
471		Vastmanland	63	122.1	0.92	(0.72 1.17)	0.97	(0.55 1.67)
464		Skaraborg	69	140.7	0.94	(0.73 , 1.20)	0.97	(0.55 , 1.67)
552		Cheshire	343	384.1	1.12	(0.92 , 1.37)	1.08	(0.58 , 1.79)
511		Durham	222	251.2	1.10	(0.90 , 1.36)	1.07	(0.57 , 1.79)
	UK							
533		Surrey	322	463.5	0.94	(0.75 , 1.16)	0.93	(0.49 , 1.54)
528		Hertfordshire	266	389.0	0.89	(0.71 , 1.12)	0.91	(0.48 , 1.52)

*Copenhagen and Frederiksberg (city), +Macedonia East and Thrace

Figure 8.7 Map of colorectal RRs cancer from full two-level model

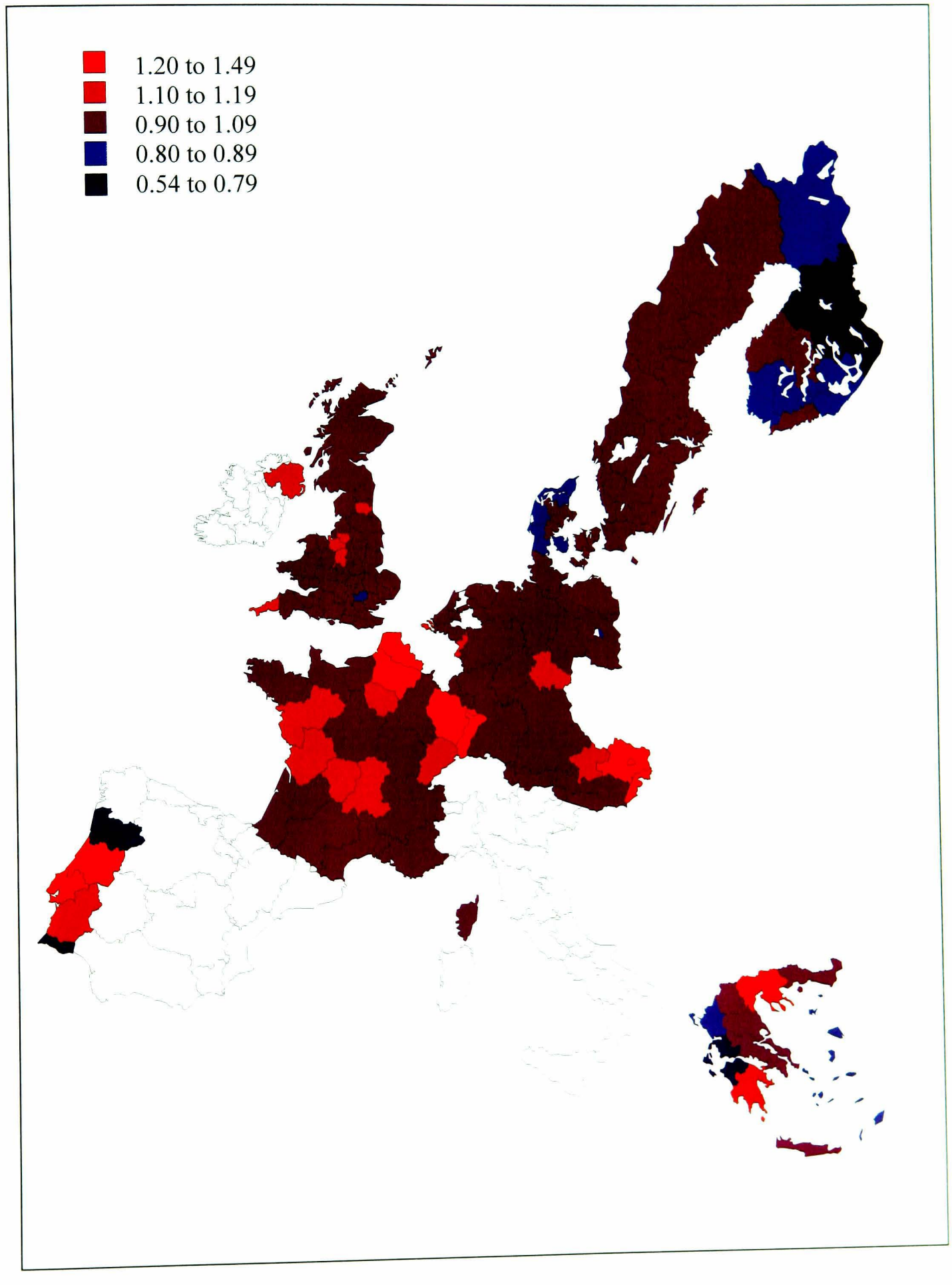
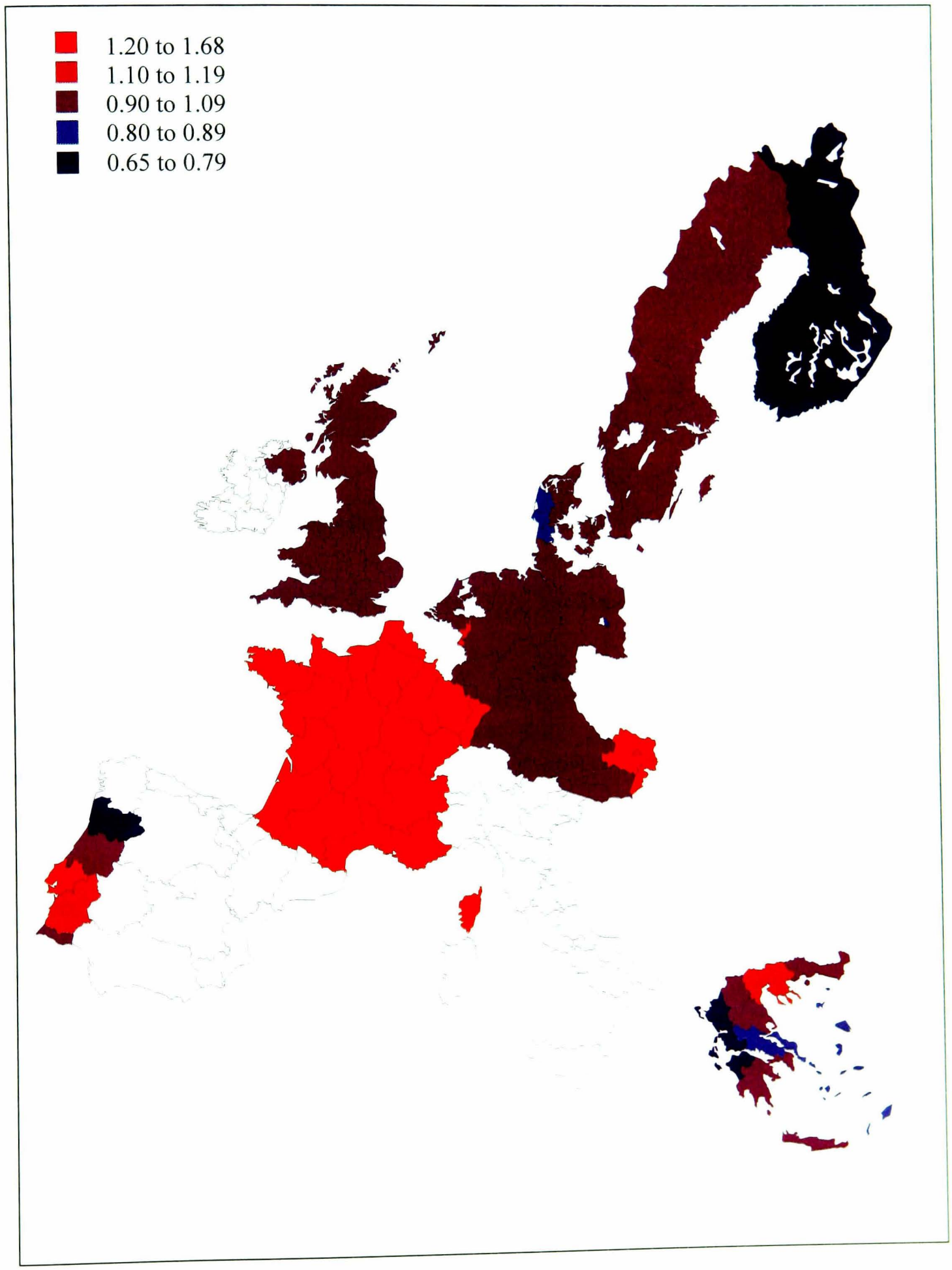


Figure 8.8 Map of colorectal cancer RRs from full three-level model



Finally, looking at the three-level model it can be seen that the parameter estimates are now all non-significant based on the 95% posterior intervals. More uncertainty appears to be attached to the estimates, which was also evident in the relative risks (Table 8.8). In contrast to lung cancer, the differences in colorectal cancer between countries cannot be explained, to a great extent, by the various risk factors, hence, the non-significant parameter estimates. In fact, the large differences between countries remain after taking account of the risk factors; see Figure 8.8.

8.3.2.9 Colorectal Cancer Risk Factor Effect Size

The parameter estimates for the fixed effects were explained in the previous section. However, to give a clearer picture of the actual effect size of these estimates, relative risks were calculated comparing countries with high and low levels of exposure to each of the risk factors. These are presented for the two-level and three-level models in Table 8.10. Concentrating on the two-level model results and looking at the smoking variable first it can be seen that smoking, on average, the same amount of cigarettes as in Greece leads to a relative risk of colorectal cancer mortality that is 3.4 times as high as if smoking was on the same level as in Sweden. Consideration of the dietary variables leads to the following conclusions: consuming, on average, the same amount of fruit as in Greece has a risk of colorectal cancer mortality that is 49% lower than if consumption was on the same level as in the UK; vegetable consumption on the same level as in Greece gives a risk of mortality 55% lower than if consuming as much as Finland; and animal fat consumption on the same level as Luxembourg results in a colorectal cancer mortality risk 2.3 times as high than if consumption was similar to Greece.

8.3.3 Colorectal Cancer Discussion

As was evident from studying all cancers and lung cancer, spatial variation in colorectal cancer mortality exists both within and between countries in the EU. About half of this variation can be accounted for by taking into account the known risk and protective factors for the disease; however, variation still exists throughout the eleven countries. Portugal, Finland and Greece have regions in which very low risks of mortality exist; further investigations in these areas may help determine ways to prevent colorectal cancer mortality. However, areas of high risk tend to be of more interest from the viewpoint of public health interventions and those which exist for colorectal cancer mortality after accounting for the risk factors are ten regions across France, Austria, Portugal and Greece.

It would be interest to investigate further why these 'hotspot' areas exist after accounting for the various risk and protective factors. As with lung cancer, differing within-country exposure to one or more of the risk factors could account for the high risks. For example, people in the 'hotspot' regions in France may smoke more than in the rest of the country but this has not been adjusted for so instead it shows as a very high risk area on the map. There is abundant evidence supporting the benefits of early screening on colorectal cancer survival (161-166) and it has been estimated that more than one third of deaths from colorectal cancer could be prevented if those over fifty were screened regularly for the disease (167). The availability of screening for colorectal cancer in these 'hotspot' areas may be affecting rates. Accurate diagnosis and effective surgery for early-stage cancer has also been shown to improve colorectal cancer mortality (168), and variations in the provisions of these within countries may also be affecting the rates.

Table 8.9 Parameter estimates from modelling colorectal cancer mortality rates

Parameters		Null two-level Model		Full two-level Model		Full three-level Model	
		Estimate	**Credible Interval**	**Estimate**	**Credible Interval**	**Estimate**	**Credible Interval**
β_0		5.36	(5.28 , 5.44)	4.92	(4.50 , 5.39)	4.84	(3.68 , 6.08)
β_1	**(smoke)**			**0.0006**	**(0.0003 , 0.0009)**	0.0004	(-0.0002 , 0.0009)
β_2	**(fruit)**			**-0.0099**	**(-0.0131 , -0.0069)**	-0.0061	(-0.0128 , 0.0012)
β_3	**(veg)**			**-0.0033**	**(-0.0048 , -0.0019)**	-0.0033	(-0.0089 , 0.0005)
β_4	**(animal)**			**0.0331**	**(0.0167 , 0.0498)**	0.0155	(-0.0312 , 0.0561)
β_5	(alcohol)			-0.0003	(-0.0027 , 0.0019)	0.0031	(-0.0034 , 0.0092)
β_6	(gdp)					3.9e-07	(-4.1e-06 , 4.9e-06)
σ_u^2		0.0163	(0.0090 , 0.0255)	0.0090	(0.0049 , 0.0145)	0.0047	(0.0023 , 0.0080)
σ_{uv}		0.0479	(0.0326 , 0.0673)	0.0226	(0.0138 , 0.0339)	0.0065	(0.0023 , 0.0121)
σ_v^2		0.1630	(0.1187 , 0.2178)	0.0795	(0.0521 , 0.1146)	0.0260	(0.0122 , 0.0454)
σ_y^2						0.0867	(0.0204 , 0.2968)

Table 8.10 Effect size of covariates from two-and three-level full colorectal cancer models

Covariate	Min	(country)	Max	(country)	2 level model		3 level model	
					RR	95% CI	RR	95% CI
Smoking	**1550**	**(Sweden)**	**3590**	**(Greece)**	**3.40**	**(1.84 , 6.27)**	2.26	(0.66 , 6.27)
Fruit	**74.5**	**(UK)**	**142.6**	**(Greece)**	**0.51**	**(0.41 , 0.63)**	0.66	(0.42 , 1.09)
Vegetable	**58.8**	**(Finland)**	**300.4**	**(Greece)**	**0.45**	**(0.31 , 0.63)**	0.45	(0.12 , 1.13)
Animal fat	**2.3**	**(Greece)**	**26.8**	**(Luxembourg)**	**2.25**	**(1.51 , 3.39)**	1.46	(0.47 , 3.95)
Alcohol	60.0	(Greece)	173.9	(Germany)	0.97	(0.74 , 1.24)	1.42	(0.68 , 2.85)

In this study, the factors that were shown to have a significant association with European colorectal cancer mortality rates complemented other research on the disease (as discussed in Chapter 2). Smoking and animal fat consumption were shown to be strong risk factors for the disease, and fruit and vegetable consumption exert a protective effect on colorectal cancer mortality. Around 50% of the variation that was seen to exist between regions in the EU can be explained by taking account of these fairly crude measures of exposure to the given risk and protective factors.

There was still a fairly high amount of variance evident even after adjusting for covariates, and modelling showed that around 68% of this is spatially patterned. This suggests that there are other factors not included in the study that are also spatially patterned that also influence colorectal cancer mortality. About 75% of patients with colorectal cancer have sporadic disease in that there is no apparent evidence of having inherited the disorder (169), and the remaining patients have a family history suggesting a genetic contribution, common exposures among family members, or a combination of both. It is well established that gene frequency varies considerably from place to place, but usually there is little difference between neighbouring populations (158). Therefore, genetic susceptibility to the disease may be one of the underlying spatial factors showing a strong relationship with colorectal cancer. Some genetic mutations have been identified as the cause of inherited cancer risk; these mutations are estimated to account for only 5% to 6% of colorectal cancer cases overall and therefore it is likely that other undiscovered major genes and background genetic factors contribute to the development of colorectal cancer, along with nongenetic risk factors such as those previously discussed (169). It was noted that different types of health care provision play an important role in colorectal cancer incidence and mortality; this also may be a factor that is spatially patterned and could account for some of the unexplained spatially patterned variation. Similarly, the importance of country may relate to countrywide differences in screening practices.

8.4 Oesophageal Cancer Mortality

Oesophageal cancer is one of the most deadly malignancies (58). In Europe, five-year relative survival rates are around 10% with significant between country survival differences existing (170). Worldwide, cancer of the oesophagus is the sixth highest malignancy amongst men and ninth highest amongst woman (171). Cases of the disease are much more common in economically less developed regions and, although around 80% occur in such areas, it should still be approached as a public health concern in Europe. Substantial variation in incidence, and therefore mortality due to poor prognosis, has been observed in Europe; in Sweden, a low-risk country, the age adjusted incidence rates are 3.1 and 1.0, in England and Wales 7.6 and 3.2, in Scotland 9.4 and 5.0 and in Calvados in France 22.3 and 1.1 per 100,000 person-years for men and woman respectively (172). The high between country variation within Europe and the very high fatality of oesophageal cancer patients emphasises the need for further investigation into its true distribution and the reasons for such patterns existing.

8.4.1 Modelling Oesophageal Cancer Mortality

As with previous cancers examined, spatial multilevel modelling techniques are used to examine the oesophageal cancer mortality patterns throughout eleven EU countries. As previously discussed, cancer of the oesophagus has been strongly associated with smoking and alcohol intake, and dietary factors (mainly the protective effects of fruit and vegetable consumption) have been shown to influence the patterns of the disease. Again, data that reflects different levels of these within the EU countries has been included in the models. The same group of models have been fitted as in previous sections; burn-in periods of 20,000, 300,000 and 800,000 were needed for the null, two-level and three-level spatial models respectively. Further iterations of 100,000, 100,000, and 400,000 were run for the three models, respectively, to gain a suitable posterior distribution from which to sample from. It should be noted that, as with modelling colorectal cancer, many fewer further iterations would have sufficed for the null model.

8.4.2 Model Results: Oesophageal Cancer

Tables 8.11 and 8.12 present the standardised mortality ratios and relative risks from fitting the 3 models. As in the previous sections the most extreme rates within each country are given. Table 8.13 shows the country-level SMRs and relative risks from fitting the three-level model. Table 8.14 gives the fixed and random parameter estimates from each of the models and Table 8.15 shows relative risks comparing high and low levels of exposure to each of the covariates. The SMRs and relative risks from each model have been mapped and are displayed in Figures 8.9-8.12.

8.4.2.1 Two-Level Null Model: Oesophageal Cancer

Looking at Table 8.11 it can be seen that, based on the relative risks from the null model, the region with the highest risk of oesophageal cancer in the EU countries being examined is Copenhagen in Denmark with a very high relative risk of 4.40. This is followed by Nord-Pas-de-Calasis in France (RR=2.94) and Northern Ireland in the UK (RR=2.79) who, despite having a risk of oesophageal cancer much lower than Copenhagen, before taking account of any risk factors, still have a very high relative risk of the disease. The region with the lowest relative risk emerges to be the Algarve in Portugal with a risk of oesophageal cancer mortality 70% lower than what is expected for that area. This is closely followed by Bornholm in Denmark (RR=0.35) and Vienna in Austria (RR=0.37). These figures alone display the very high variability of the disease both within and between countries in the EU. It is clear that regions within Denmark show the most variability in risks of oesophageal cancer mortality, and the country displaying the least amount of variability is Greece with very low relative risks ranging from 0.46 to 0.74.

8.4.2.2 Oesophageal Cancer SMRs

Comparing the relative risks to the SMRs displayed in table 8.11 it can again be seen that there are many discrepancies between the two. Much research on variability of disease within countries and across numerous countries often looks solely at age and sex standardised rates. This table, along with the similar tables for other cancers, clearly shows that spatially modelled risks, which should be closer to the true distribution of the disease, often differ very much from the 'raw' rates emphasising that such rates should be taken with caution.

8.4.2.3 Oesophageal Cancer Disease Maps I (SMRs and Two-Level Null Model RRs)

Figure 8.9 clearly shows that there is much clustering of SMRs within countries in the EU and also displays much variation between countries. The SMRs range from 0 to 5.33; this is an example of another problem when using these ratios. Since oesophageal cancer is not as common a form of malignancy as cancers such as lung cancer, sometimes an area, probably with a small population, will observe no deaths from the disease. In this case it is a fairly small region in Finland. The population in this area is then presented as having zero risk of mortality from oesophageal cancer. This is however false and the modelling approaches subsequently used take this into account through modelling extra-Poisson variation.

Figure 8.10 shows a much different picture of the distribution of oesophageal cancer mortality. The relative risks before taking account of risk and protective factors now show more variation within most of the countries. The relative risks range from 0.29 to 4.40. However, there are now fewer areas of extreme risk. Most of the high relative risks can be found in Denmark, France and the UK. Most of the regions with lower risk of oesophageal cancer than expected before taking account of risk factors are in Greece, West Germany, Sweden and Finland.

8.4.2.4 Two-Level Full Model: Oesophageal Cancer

After adjusting for risk and protective factors (Table 8.12) shows that the relative risks for oesophageal cancer mortality predicted from the two-level model change somewhat. The most extreme risk of the disease can now be found in Azores, Portugal; taking into account Portugal's smoking and drinking habits and average fruit and vegetable consumption the risk of oesophageal cancer mortality is 3.9 times higher than expected. Centro in Portugal has the second highest risk of oesophageal cancer mortality (RR=1.95), followed by Zeeland in the Netherlands (RR=1.70). Provence-Alpes-Cote d'Azur in France has the lowest risk of mortality after adjusting for covariates with a relative risk of 0.63, closely followed by Bremen in Germany (RR=0.66; not evident from table) and Uusimaa in Finland (RR=0.69). Regions within Finland display the least amount of variability in relative risks with a range of 0.69 to 0.98.

Some extreme regions tend to be less extreme after adding the covariates; the relative risks in Copenhagen and Frederiksberg (city) and Strørstrom reduce considerably after adjusting for the covariates. Therefore, after taking into account the fact that people living in Denmark have low fruit and vegetable consumption and high animal fat and alcohol intake, they actually have a much lower risk of oesophageal cancer mortality. Greece's relative risks, which were very low previously, grew closer to unity after adding the covariates. Greece have a "healthy" lifestyle in general, except for their smoking habits, in that they have very high fruit and vegetable consumption and very low animal fat and alcohol consumption; after taking this into account the relative risks of oesophageal cancer mortality are much higher.

8.4.2.5 Three-level Full Model: Oesophageal Cancer

Again it can be seen that the uncertainty attached to the relative risks is much greater after adding a third hierarchical level to the model. Many of the confidence intervals more than double in size.

As expected, the estimates of the relative risks in countries which generally had high risks previously increase somewhat; this can be seen for example in Denmark and the Netherlands. Similarly, those countries with clustering of low relative risks prior to adding the higher level such as Finland and Germany now have lower estimates.

8.4.2.6 Oesophageal Cancer Disease Maps II (Two and Three-Level Full Model RRs)

Again, adding the covariates gives a different picture of oesophageal cancer mortality (Figure 8.11). There is still a high amount of variation within and between countries. However, fewer regions now have very high (RR > 1.2) risk of mortality, due to the risk and protective factors explaining the reasons for some areas having such high risks. There are twenty nine regions that remain oesophageal cancer 'hotspots' and possibly need further investigation into why they are so. Apart from those in the UK and Sweden, these tend to be small clusters within France, the Netherlands, Austria, Denmark and Portugal. The areas of low risk that stand out on the map are regions in Finland, Portugal and Greece and clusters within France, Germany and Sweden. Again, it may be of interest to look at these areas on a smaller scale to try and determine why they have such low oesophageal cancer mortality after taking into account country level exposures to the significant risk factors.

Finally, looking at the map of the relative risks, after adding country as a higher level to the model (Figure 8.12), there is more clustering evident within countries. All the regions within some countries are very homogeneous, eg Germany, Finland, Netherlands and Denmark, and some countries show two or three distinct clusters within the country, eg France, Austria and Greece. The last two maps give a smoother picture of the distribution of oesophageal cancer mortality in the EU.

Table 8.11 Relative risks (SMR and RR from two-level null model) of mortality from oesophageal cancer (for each country extreme rates given, ordered by decreasing RR)

i	Country	Region	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
8		Tirol	18	36.4	0.49	(0.29 , 0.78)	0.98	(0.51 , 1.86)
2		Burgenland	17	18.2	0.94	(0.54 , 1.50)	0.98	(0.37 , 2.65)
	Austria							
5		Upper Austria	38	81.1	0.47	(0.33 , 0.64)	0.65	(0.36 , 1.15)
10		Vienna	51	113.8	0.45	(0.33 , 0.59)	0.37	(0.11 , 1.08)
85		Copenhagen*	231	43.3	5.33	(4.67 , 6.07)	4.40	(1.49 , 13.8)
90		Storstrom	77	20.7	3.73	(2.94 , 4.66)	2.67	(1.58 , 4.87)
	Denmark							
99		Nordjylland	159	35.2	4.51	(3.84 , 5.27)	1.25	(0.52 , 3.35)
91		Bornholm	16	3.7	4.34	(2.48 , 7.04)	0.35	(0.14 , 1.01)
123		Kymi	11	23.1	0.48	(0.24 , 0.85)	1.14	(0.50 , 2.50)
130		Vassa	20	29.9	0.67	(0.41 , 1.03)	0.97	(0.44 , 2.08)
	Finland							
121		Keski-Suomi	6	16.1	0.37	(0.14 , 0.81)	0.70	(0.37 , 1.22)
124		Lappi	10	11.7	0.85	(0.41 , 1.57)	0.59	(0.22 , 1.50)
138		Nord-Pas-de-Calais	526	194.5	2.70	(2.48 , 2.95)	2.94	(0.63 , 13.8)
141		Franche-Comte	83	61.2	1.36	(1.08 , 1.68)	1.87	(0.99 , 3.10)
	France							
151		Provence-Alpes-Cote d'Azur	271	271.5	1.00	(0.88 , 1.12)	0.77	(0.49 , 1.20)
145		Aquitaine	215	178.9	1.20	(1.05 , 1.37)	0.74	(0.37 , 1.59)
75		Baden-Wurttemberg	431	641.1	0.67	(0.61 , 0.74)	1.15	(0.62 , 2.03)
78		Berlin	136	219.2	0.62	(0.52 , 0.73)	1.12	(0.22 , 5.36)
	Germany							
83		Saxony-Anhalt	106	179.8	0.59	(0.48 , 0.71)	0.67	(0.37 , 1.20)
69		Hamburg	96	123.0	0.78	(0.63 , 0.95)	0.51	(0.21 , 1.23)
159		Macedonia West	8	18.2	0.44	(0.19 , 0.87)	0.74	(0.33 , 1.63)
167		Aegean South	7	17.0	0.41	(0.16 , 0.85)	0.72	(0.36 , 1.43)
	Greece							
168		Segean North	3	18.7	0.16	(0.03 , 0.47)	0.50	(0.25 , 0.94)
169		Crete	5	41.5	0.12	(0.04 , 0.28)	0.46	(0.14 , 1.34)
238	Luxembourg		20	24.7	0.81	(0.49 , 1.25)	0.90	(0.41 , 1.85)
252		Zeeland	27	25.6	1.05	(0.69 , 1.53)	1.46	(0.28 , 8.02)
243		Drenthe	22	29.5	0.75	(0.47 , 1.13)	1.26	(0.64 , 2.41)
	Netherlands							
241		Groningen	36	36.4	0.99	(0.69 , 1.37)	0.75	(0.35 , 1.68)
246		Flevoland	5	10.9	0.46	(0.15 , 1.07)	0.73	(0.35 , 1.45)
326		Azores	7	13.1	0.53	(0.21 , 1.10)	1.83	(0.84 , 4.03)
322		Centro	81	121.8	0.67	(0.53 , 0.83)	1.03	(0.75 , 1.41)
	Portugal							
327		Madeira	19	12.8	1.49	(0.89 , 2.32)	0.48	(0.17 , 1.20)
325		Algarve	9	26.0	0.35	(0.16 , 0.66)	0.30	(0.07 , 1.14)
460		Malmohus	36	61.1	0.59	(0.41 , 0.82)	2.63	(0.87 , 7.50)
458		Kristianstad	12	24.4	0.49	(0.25 , 0.86)	1.14	(0.64 , 1.92)
	Sweden							
469		Vasterbotten	9	19.1	0.47	(0.21 , 0.89)	0.52	(0.24 , 1.04)
455		Jonkoping	8	25.1	0.32	(0.14 , 0.63)	0.51	(0.28 , 0.87)
565		Northern Ireland	126	90.3	1.40	(1.16 , 1.66)	2.79	(0.63 , 11.8)
555		Merseyside	185	97.1	1.91	(1.64 , 2.20)	1.64	(0.79 , 3.32)
	UK							
529		Berkshire	48	45.0	1.07	(0.79 , 1.41)	0.90	(0.55 , 1.41)
563		West Glamorgan	33	27.1	1.22	(0.84 , 1.71)	0.82	(0.41 , 1.57)

*Copenhagen and Frederiksberg (city)

Figure 8.9 Map of oesophageal cancer SMRs

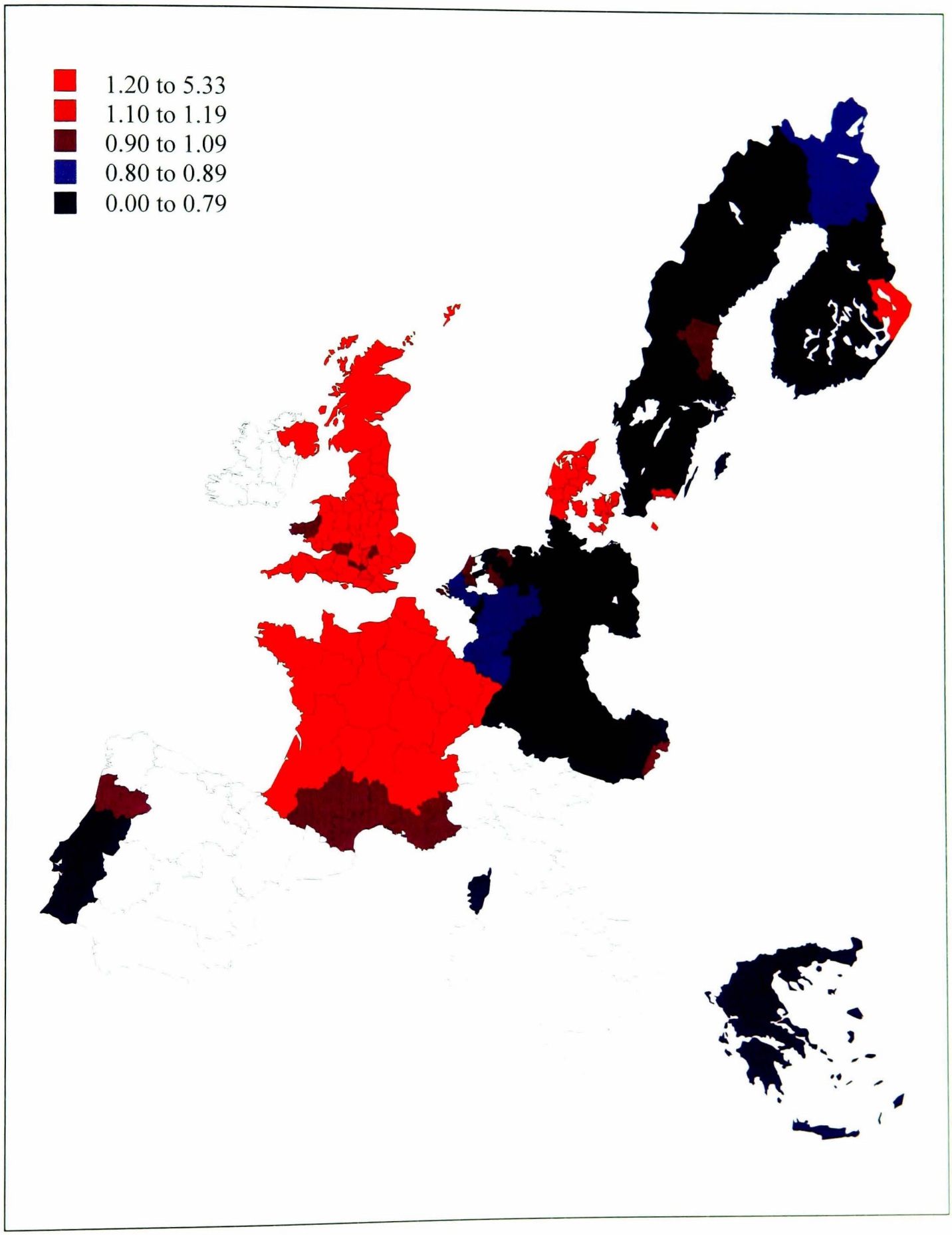


Figure 8.10 Map of oesophageal cancer RRs from null two-level model

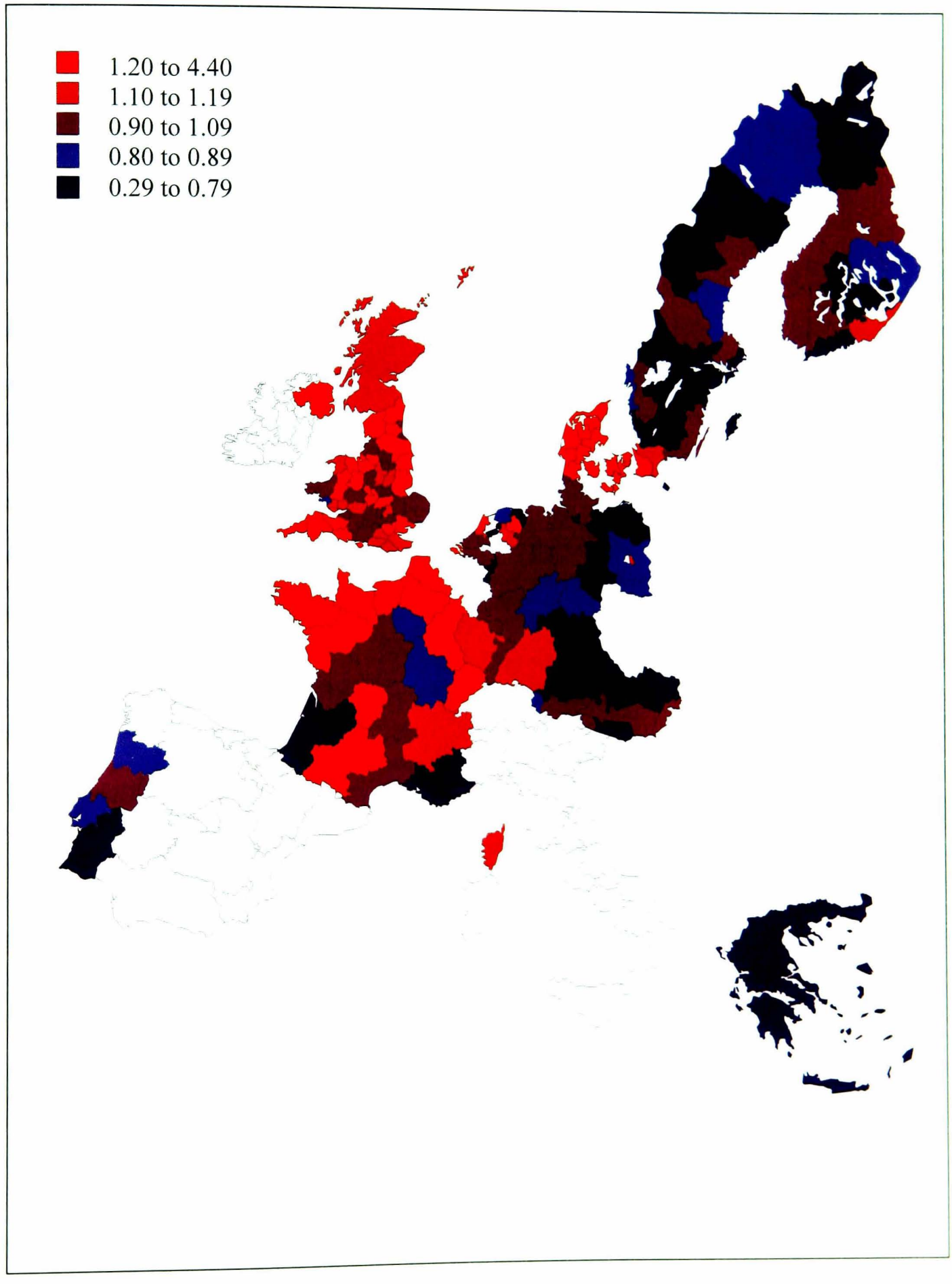


Table 8.12 Relative risks (from two-and three-level full models) of mortality from oesophageal cancer (for each country extreme rates given, ordered by decreasing RR from three-level model)

i	Country	Region	O_i	E_i	$RR_{mean(2)}$	$PI_{95\%(2)}$	$RR_{mean(3)}$	$PI_{95\%(3)}$
7		Styria	53	77.4	1.41	(0.98 , 2.09)	1.06	(0.40 , 3.07)
2		Burgenland	17	18.2	1.17	(0.60 , 2.32)	1.00	(0.31 , 3.54)
	Austria							
5		Upper Austria	38	81.1	0.90	(0.58 , 1.38)	0.82	(0.30 , 2.42)
10		Upper Austria	51	113.8	0.76	(0.38 , 1.51)	0.73	(0.22 , 2.49)
85		Copenhagen*	231	43.3	1.63	(0.71 , 3.82)	1.85	(0.38 , 9.96)
90		Storstrom	77	20.7	1.59	(1.02 , 2.50)	1.82	(0.51 , 7.09)
	Denmark							
87		Frederiksborg	78	22.0	1.08	(0.61 , 1.87)	1.36	(0.35 , 5.64)
93		Sonderjylland	42	17.8	0.96	(0.65 , 1.41)	1.30	(0.37 , 4.81)
119		Ahvenanmaa	0	1.8	0.98	(0.62 , 1.53)	0.66	(0.22 , 1.86)
127		Pohjois-Karjala	14	11.7	0.89	(0.55 , 1.42)	0.64	(0.21 , 1.84)
	Finland							
129		Uusimaa	41	70.8	0.69	(0.45 , 1.03)	0.57	(0.19 , 1.53)
121		Keski-Suomi	6	16.1	0.72	(0.45 , 1.09)	0.56	(0.19 , 1.52)
138		Nord - Pas-de-Calais	526	194.5	1.51	(0.56 , 4.17)	1.51	(0.39 , 6.42)
136		Basse-Normandie	180	78.3	1.41	(0.97 , 2.11)	1.45	(0.57 , 3.88)
	France							
152		Corsica	9	16.1	0.71	(0.29 , 1.73)	0.74	(0.19 , 2.90)
151		Provence-Alpes-Cote d'Azur	271	271.5	0.63	(0.42 , 0.92)	0.71	(0.27 , 1.90)
74		Rheinland-Palatinate	215	262.5	0.99	(0.71 , 1.37)	0.75	(0.29 , 2.18)
70		Lower Saxony	378	514.7	0.97	(0.79 , 1.19)	0.73	(0.31 , 1.96)
	Germany							
84		Thuringen	76	160.5	0.76	(0.52 , 1.08)	0.54	(0.20 , 1.61)
82		Saxony	140	318.4	0.71	(0.46 , 1.06)	0.53	(0.19 , 1.65)
157		Macedonia East and Thrace	16	37.3	1.06	(0.55 , 2.07)	1.05	(0.18 , 5.64)
158		Macedonia Central	33	105.0	1.07	(0.63 , 1.81)	1.04	(0.20 , 5.02)
	Greece							
168		Aegean North	3	18.7	0.77	(0.47 , 1.24)	0.80	(0.15 , 3.71)
169		Crete	5	41.5	0.74	(0.36 , 1.50)	0.77	(0.13 , 4.10)
238	Luxembourg		20	24.7	1.05	(0.61 , 1.82)	1.20	(0.35 , 4.26)
252		Zeeland	27	25.6	1.70	(0.62 , 4.87)	2.04	(0.42 , 12.3)
249		Utrecht	59	59.9	1.30	(0.80 , 2.17)	1.76	(0.52 , 7.10)
	Netherlands							
246		Flevoland	5	10.9	1.04	(0.60 , 1.77)	1.44	(0.41 , 5.97)
248		Limburg	35	69.2	0.84	(0.52 , 1.33)	1.30	(0.38 , 5.18)
326		Azores	7	13.1	3.90	(1.69 , 9.33)	1.95	(0.34 , 9.34)
322		Centro	81	121.8	1.85	(1.29 , 2.78)	1.39	(0.34 , 4.52)
	Portugal							
324		Alentejo	21	44.2	1.14	(0.75 , 1.73)	0.93	(0.22 , 3.04)
325		Algarve	9	26.0	0.76	(0.27 , 2.03)	0.82	(0.13 , 3.97)
449		Blekinge	15	12.6	1.21	(0.69 , 2.18)	1.12	(0.26 , 4.29)
466		Stockholm	80	115.8	1.13	(0.65 , 1.98)	1.10	(0.26 , 4.08)
	Sweden							
463		Ostergotland	8	31.8	0.78	(0.50 , 1.17)	0.82	(0.21 , 2.73)
455		Jonkoping	8	25.1	0.73	(0.47 , 1.09)	0.81	(0.21 , 2.70)
565		Northern Ireland	126	90.3	1.51	(0.58 , 3.95)	1.32	(0.29 , 5.77)
553		Greater Manchester	326	168.2	1.28	(0.87 , 1.89)	1.25	(0.43 , 3.43)
	UK							
529		Berkshire	48	45.0	0.81	(0.56 , 1.14)	0.86	(0.30 , 2.33)
563		West Glamorgan	33	27.1	0.77	(0.46 , 1.24)	0.85	(0.26 , 2.56)

Figure 8.11 Map of oesophageal cancer RRs from full two-level model

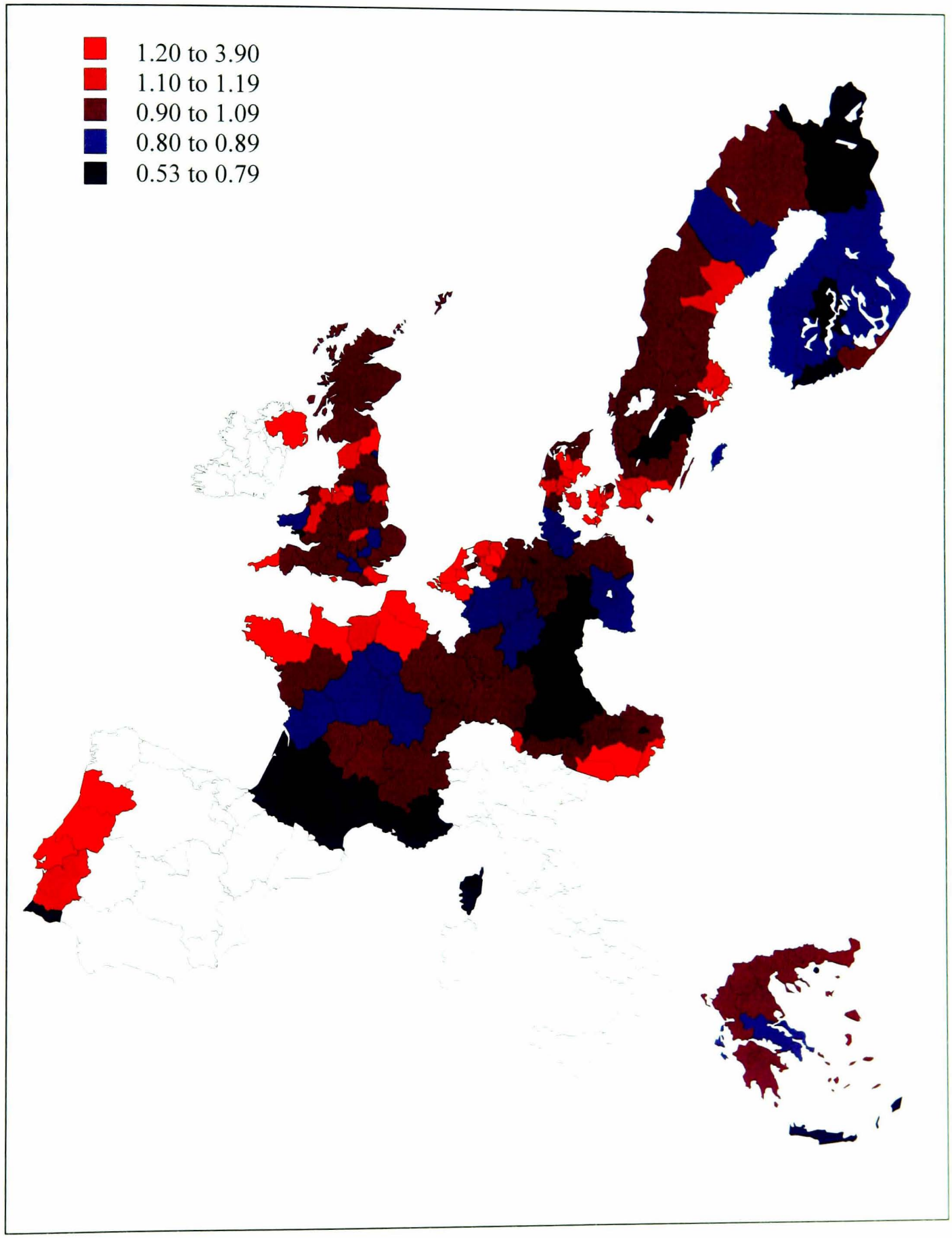
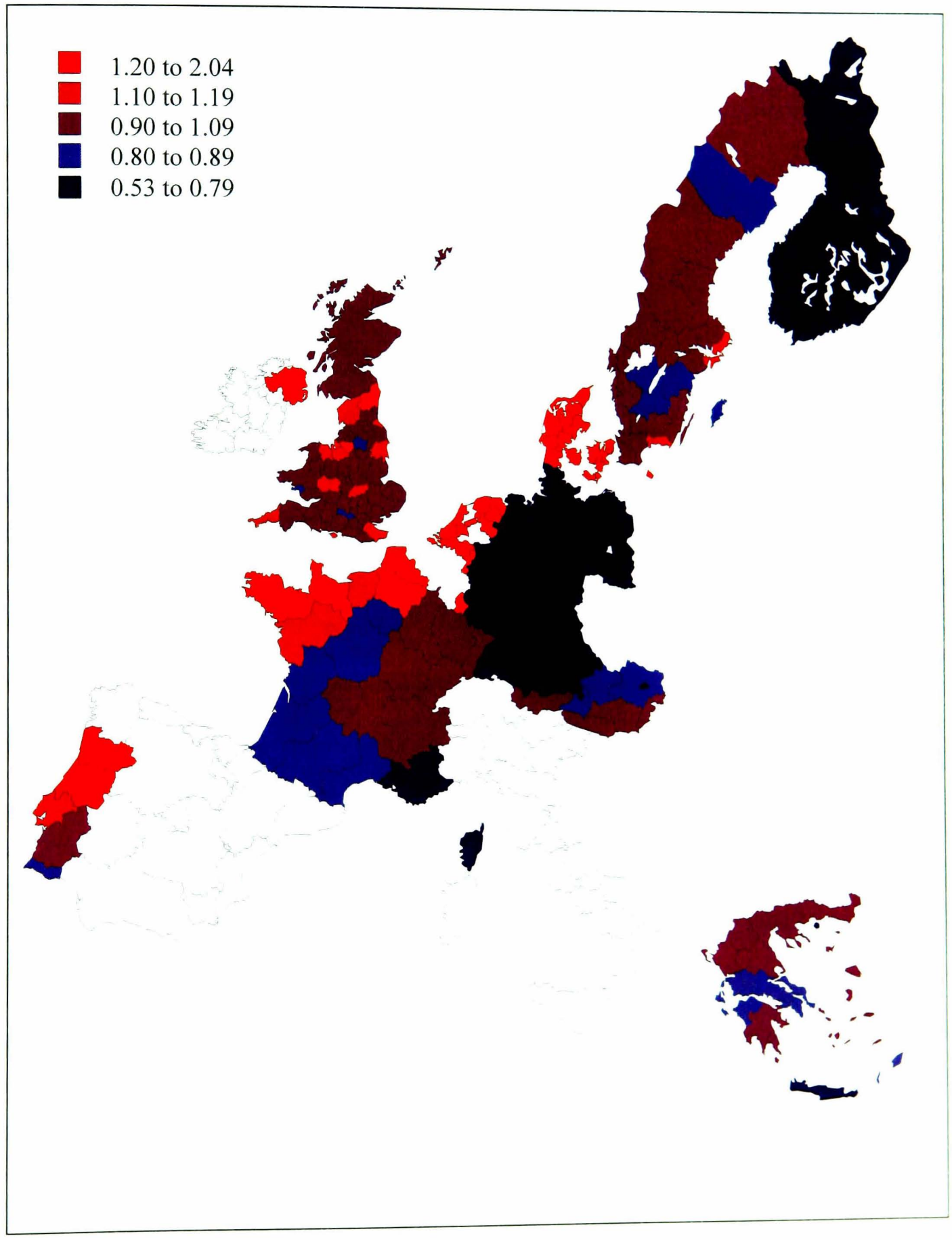


Figure 8.12 Map of oesophageal cancer RRs from full three-level model



8.4.2.7 Country Level Oesophageal Cancer Results

Comparing the country level relative risks obtained from fitting the full three-level multilevel model (Table 8.13) it can be seen that the Netherlands and Denmark have by far the highest risks of oesophageal cancer mortality (RR of 1.55 and 1.50 respectively). The country with the overall lowest relative risk is, as expected, Finland (RR=0.62), followed by Germany (RR= 0.67).

Table 8.13 Relative risks of mortality from oesophageal cancer at country level

Country	O_i	E_i	SMR	$CI_{95\%}$(SMR)	RR_{mean}	$PI_{95\%}$
Austria	255	506.6	0.50	(0.44 , 0.57)	0.88	(0.45 , 1.84)
Germany	3690	5316.7	0.69	(0.67 , 0.72)	0.67	(0.34 , 1.52)
Denmark	1423	365.2	3.90	(3.70 , 4.10)	1.50	(0.60 , 4.05)
Finland	192	314.7	0.61	(0.53 , 0.70)	0.62	(0.29 , 1.20)
France	4835	3203.9	1.51	(1.47 , 1.55)	1.01	(0.54 , 1.96)
Greece	181	690.3	0.26	(0.23 , 0.30)	0.91	(0.26 , 2.91)
Luxembourg	20	24.7	0.81	(0.49 , 1.25)	1.14	(0.51 , 2.68)
Netherlands	773	931.4	0.83	(0.77 , 0.89)	1.55	(0.67 , 4.27)
Portugal	460	611.1	0.75	(0.69 , 0.82)	1.16	(0.40 , 2.67)
Sweden	346	671.8	0.52	(0.46 , 0.57)	0.96	(0.35 , 2.35)
UK	6071	3957.1	1.53	(1.50 , 1.57)	1.02	(0.47 , 2.12)

8.4.2.8 Oesophageal Cancer Parameter Estimates

The fixed parameter estimates from the full two-level model (Table 8.14) show that the country level covariates that are significantly affecting oesophageal cancer mortality whilst taking the other variables into account are smoking and fruit, vegetable and animal fat consumption. The estimate for smoking suggests

that every increase of 100 cigarettes smoked per person per year is associated with an increase in the risk of oesophageal cancer mortality of about 17% (RR=exp{0.16}=1.17}. An increase in fruit consumption of 10 kg per person per year reduces the risk of oesophageal cancer by 28% (RR=0.72). A similar increase in vegetable consumption results in a relative risk of 0.63 and in animal fat consumption gives a relative risk of 2.16.

The random part of the null model shows that the spatial variance, σ_v^2, after taking the average number of nearest neighbours ($\bar{n}$ = 4.128) into account is 0.180, resulting in a total variance of 0.228. Therefore, 79% of the total variance in oesophageal cancer mortality, before taking risk factors into account, arises from spatial effects. The total variance from the full two-level model is 0.078. Therefore, taking into account the measures of country level exposure to cigarette smoking and fruit, vegetable and animal fat consumption has reduced the overall variation in oesophageal cancer mortality by 66%. Sixty four percent of this remaining variation can then be attributed to spatial patterning in the data.

Looking at the three-level model, it can be seen that the parameter estimate for fruit consumption is the only variable that remains significant. After adding country as a level to the hierarchical model it can again be seen that there is more error attached to all the fixed parameter estimates with the posterior credible intervals being around three times wider in most cases. The intervals for the random parameters actually reduce in size; however, a large amount of the variation is now attributable to the third level, country. Partitioning the variance so that it includes σ_y^2 results in the total variance increasing to 0.353. The majority of the variance from this model is attributable to the differences between countries.

8.4.2.8 Oesophageal Cancer Risk Factor Effect Size

Concentrating on the significant parameter estimate relative risks from the two-level model, Table 8.15, shows that smoking on average the same amount of

cigarettes as in Greece leads to a relative risk that is 26 times as high than if smoking was on the same level as in Sweden. Consuming an equivalent average amount of fruit as Greece reduces the risk of oesophageal cancer mortality by 89% compared to the UK. Vegetable consumption on the same level as in Greece has a risk of mortality that is 67% lower than if consumption was on the same level as in Finland. Finally, animal fat consumption on the same level as in Luxembourg leads to a relative risk of oesophageal cancer 6.6 times what it would be if consumption was similar to Greece.

8.4.3 Oesophageal Cancer Discussion

Mortality patterns of oesophageal cancer are similar to other cancers in that it is clear from disease maps and estimates of relative risks that much variation exists both within and between countries in the EU. Some of this variation can be accounted for by taking into account country level risk and protective factors for the disease. This changes the pattern of mortality: it smoothes the distribution of the disease more and makes small clusters of extreme risks more evident. After such modelling, the most prominent areas with very low risk of the disease are South France, West Germany and some Scandinavian regions. Clusters of high risk that are clearly visible from this smoothed map are in North Portugal, North France, West Denmark, South Austria, the Netherlands and a few UK regions.

It may be of public health interest to investigate why some regions in the EU remain as oesophageal 'hotspots' after accounting for exposure to known risk and protective factors. It could simply be that there are variations in levels of these factors within countries; this has not been taken into account, due to data availability, and may reflect the within country variations. Levels of exposure to the various factors would have to be examined at the regional level, at least, within these countries, and, if particularly high or low levels of exposure exist, specific countries may wish to aim public health promotions, specific to this disease, in these areas.

Table 8.14 Parameter estimates from modelling oesophageal cancer mortality rates

Parameters		Null 2 level Model		Full 2 level Model		Full 3 level Model	
		Estimate	**Credible Interval**	**Estimate**	**Credible Interval**	**Estimate**	**Credible Interval**
β_0		3.87	(3.71 , 4.03)	3.04	(2.31 , 3.78)	3.34	(0.28 , 6.98)
β_1	**(smoke)**			**0.0016**	**(0.0011 , 0.0021)**	0.0010	(-0.0008 , 0.0020)
β_2	**(fruit)**			**-0.0326**	**(-0.0378 , -0.0276)**	**-0.0267**	**(-0.0444 , -0.0054)**
β_3	**(veg)**			**-0.0046**	**(-0.0072 , -0.0021)**	-0.0022	(-0.0095 , 0.0070)
β_4	**(animal)**			**0.0768**	**(0.0495 , 0.1058)**	0.0481	(-0.0613 , 0.1342)
β_5	(alcohol)			-0.0015	(-0.0052 , 0.0022)	0.0045	(-0.0090 , 0.0154)
β_6	(gdp)						
σ_u^2		0.0475	(0.0171 , 0.1007)	0.0278	(0.0124 , 0.0478)	0.0166	(0.0072 , 0.0290)
σ_{uv}		0.1659	(0.0970 , 0.2530)	0.0675	(0.0413 , 0.1012)	0.0298	(0.0150 , 0.0491)
σ_v^2		0.7436	(0.5046 , 1.0130)	0.2052	(0.1292 , 0.3065)	0.0890	(0.0465 , 0.1483)
σ_y^2						0.3145	(0.0618 , 1.0790)

Table 8.15 Effect size of covariates from 2 and 3 level full oesophageal cancer models

Covariate	Min	(country)	Max	(country)	2 level model		3 level model	
					RR	95% CI	RR	95% CI
Smoking	**1550**	**(Sweden)**	**3590**	**(Greece)**	**26.10**	**(10.19 , 69.49)**	7.06	(0.20 , 62.62)
Fruit	**74.5**	**(UK)**	**142.6**	**(Greece)**	**0.11**	**(0.08 , 0.15)**	**0.16**	**(0.05 , 0.69)**
Vegetable	**58.8**	**(Finland)**	**300.4**	**(Greece)**	**0.33**	**(0.17 , 0.61)**	0.59	(0.10 , 5.38)
Animal fat	**2.3**	**(Greece)**	**26.8**	**(Luxembourg)**	**6.57**	**(3.36 , 13.36)**	3.25	(0.22 , 26.79)
Alcohol	60.0	(Greece)	173.9	(Germany)	0.84	(0.55 , 1.28)	1.66	(0.36 , 5.80)

Some areas stood out visually on the disease map before adjusting for the various covariates, but after doing so the relative risks shrunk closer to unity. Therefore the risk of oesophageal cancer in these areas is affected by the significant risk factors in the modelling. These regions or clusters are also areas of public health concern; they may benefit further from health promotion in the areas of smoking and healthy eating.

There is reasonable evidence that screening would result in no (or minimal) decrease in mortality from oesophageal cancer in the US population (173). After examining European survival rates, Faivre et al (170) concluded that stage of diagnosis and different types of surgery are likely to improve survival. However, there is no evidence of such health provisions affecting mortality rates, and therefore will not be affecting variations in European rates. This suggests that, if attempting to reduce oesophageal cancer mortality in 'hotspot' areas, prevention should be the focus.

Obesity has emerged as a major risk factor for this disease with a positive association being shown to exist between BMI or relative weight and oesophageal cancer (59-63). Due to the unavailability of consistent obesity or BMI data across the various populations, this risk factor was not included in modelling. Hopefully the dietary factors will reflect levels of obesity in some manner. However, the unaccounted for BMI levels in the populations may explain some of the oesophageal cancer mortality 'hotspots'.

It is somewhat surprising that alcohol intake is not a significant risk factor, at the population level, for oesophageal cancer mortality in the EU. As previously discussed (Chapter 2), the positive association between alcohol use and the risk of oesophageal cancer is well established, and many studies have shown it to be a strong risk factor. This suggests that alcohol levels within countries would be associated with the disease rates and this again may account for the clusters of very high (or low) rates.

It was observed that the factors that have a significant positive association with European oesophageal cancer mortality rates were smoking and animal fat

consumption. The risk of dying from the disease was much higher in areas with high smoking level compared to low; the risk factor, 26.1, demonstrates the huge burden smoking has on this disease. Animal fat consumption also increases the risk of oesophageal cancer by a great amount and should be considered as an area on which to focus if aiming for prevention of the disease. As with all other cancers examined at the population level, fruit and vegetable consumption has a strong protective effect against this malignancy. Taking account of the country level measures of exposure to these diseases explained about 66% of the variation that existed between regions in the EU.

The disease maps and model results showed that there was still a high amount of variation after adjusting for covariates and that around 64% of this is spatially patterned. The unexplained spatially-patterned variation again suggests there are other spatially-patterned risk factors for oesophageal cancer. The factors that may be causing the disease 'hotspots' could also be spatially patterned and account for some of this variation. Since gene frequency is often spatially patterned (158) it is not unreasonable to consider this as an unaccounted for factor that is affecting the disease. It has been reported that there is apparent familial clustering of oesophageal cancer patients (174, 175) but it is unclear whether this represents a common exposure to environmental factors or a genetic predisposition. There is only one, rare, recognised genetic abnormality that predisposes patients to a type of cancer of the oesophagus (176). However, further research is needed in this area and it cannot be ruled out as a spatially patterned factor affecting the disease.

8.5 Comparing Cancer Patterns

To compare cancer mortality patterns across the EU, disease maps were examined and have proved a very useful tool for analysing the spatial patterns of cancer mortality throughout Europe. The maps provide a clear picture of the estimated risk of the specific cancers across the regions of interest. However, mapping estimates from different models can provide different pictures and interpretations

of the distribution of the disease. Maps were produced of the risk of the disease without adjusting for any covariates. These provide a true reflection of cancer rates across the EU and one can be fairly confident that areas which stand out as having extremely high rates are actually disease 'hotspots'. Such maps do not give explanations as to why areas have extremely high (or low) rates of the disease but do indicate where further public health intervention may be required. In an attempt to explain such patterning of the disease, relative risks were modelled after adjusting for various covariates that were thought to be influencing the rates of the disease. A comparison across both maps would show if the risk in areas that were previously standing out as 'hotspots' has reduced. If so, this would suggest the covariate(s) have explained these high rates, eg if smoking is a strong significant covariate in the model this would suggest the high rates of, say, lung cancer mortality can be explained by the country's smoking levels. This would then give scope to introduce public health policy on reducing smoking in that area. Areas that remain as hotspots after adjusting for risk factors gives an added piece of information in that some other factor, not accounted for in the modelling, is causing these areas of high risk. The region could perhaps have higher (or lower) levels of exposure to the covariates than the country on average or perhaps some other unaccounted for factor is causing the hotspot such as a social or lifestyle factor that has not been taken into account or an area of point-source pollution. Disease maps after adjusting for covariates may be useful to public health specialists who may be able to identify or recognise the patterns as those relating to other covariates. Overall, when modelling we wish to compare rates across regions allowing for varying age and sex structure of the populations as it is known these are affecting cancer mortality. However, there is obviously no wish to change the age and sex structure of a population to influence these rates. In contrast, it is often of public health interest to change the levels of other factors which vary across regions and influence the mortality rates. For this reason, it is appropriate to always standardise for age and sex but fit both unadjusted and adjusted models with regards to other factors such as diet and smoking. This

allows comparison and interpretation of various disease maps as discussed above and hopefully determines where and what public health interventions are needed.

The three cancers examined in this chapter are different in many respects. The number of deaths within the eleven EU countries varies substantially between the malignancies; the total number of lung cancer deaths was 123,253 in 1991, colorectal cancer deaths totalled 80,777 and oesophageal cancer deaths totalled 18,246.

The risk/protective factors which were identified in this study as significantly affecting each of the cancers were smoking, fruit consumption and vegetable consumption, with smoking, as expected, being a risk factor and fruit and vegetable consumption exerting protective effects against the diseases. Animal fat consumption was shown to be a significant risk factor for colorectal and oesophageal cancer mortality. The effect sizes of these risk and protective factors differed between the three cancer types. Smoking, as expected, was a strong risk factor for each group of mortalities. However, the highest risk from smoking was observed with oesophageal cancer, which had a relative risk three times as high as the risk associated with lung cancer. The associations fruit consumption had with lung and colorectal cancer mortality rates were similar; this was also the case for vegetable consumption. Again, oesophageal cancer had the strongest relationship with both of these risk factors. Consumption of animal fats increased the risk of both colorectal and oesophageal cancer mortality, and again the association with the latter was the strongest.

The covariates that were found to be associated with the specific cancers do, in general, tie in with existing literature on cancer mortality risk factors. However, as discussed in Chapter 2, several epidemiological studies have shown a link between oesophageal cancer and alcohol intake, but this was not found in this study. This could be due to the variable being used to account for alcohol intake not being the most appropriate, or the fact that the variable is at country level. A more specific type of alcohol, instead of all alcohols, or perhaps alcohol data which is of a lower form of aggregation would have to be examined to show

evidence of a significant effect. Other examples of strong associations that have been shown to exist in literature but are not apparent in this study are the link between socio-economic status and lung cancer and animal fat consumption and colorectal cancer. Again, more appropriate types of data may be needed to show such effects. On the other hand, we found an association between colorectal cancer mortality and smoking levels, but there is inconsistent evidence of an association in existing literature. It may also be useful to include other risk factors that were mentioned in Chapter 2 such as consumption levels of fish or fish oil which has been shown to have a protective effect on colorectal cancer.

The variability across relative risks of cancer mortality also varied from site to site; the relative risks after taking into account the relevant risk and protective factors ranged from 0.70 to 1.48 for lung cancer, with an inter-quartile range of 0.94 to 1.07, colorectal cancer was similar, ranging from 0.50 to 1.49 (inter-quartile range 0.95 – 1.07), and oesophageal cancer ranged from 0.63 to 3.9 (inter-quartile range 0.88 – 1.14). Similar results were observed across cancer sites when examining the random part of the models, with the total variation in the specific cancer mortality relative risks being reduced by a half or more through adjusting for risk and protective factors. Of the remaining variation, around 60% is attributable to the spatial patterning of each of the diseases.

Examining cancer mortality from all sites together (Chapter 5) gave relative risks, after adjusting for risk factors, which range from 0.85 to 1.29 (inter-quartile range 0.95 – 1.05). The risk factors shown to significantly affect all cancer mortality when comparing areas of high and low exposure were smoking (RR=2.99) and fruit (RR=0.58), vegetable (RR=0.72) and animal fat (RR=2.03) consumption. Modelling all cancer mortality together showed that there was a 65% reduction in the total variation of relative risks after adjusting for the risk and protective factors. It also showed that 85% of the remaining variation was due to spatial effects. Despite the similarities these results have with those predicted from modelling the three specific cancers, grouping all cancers together is

concealing different patterns in mortality from individual cancer sites and different relationships these have with risk factors.

Although similarities were observed, there is evidence of differences in European patterns of cancer mortality at the different cancer sites, suggesting that it makes sense to examine mortality patterns separately for specific malignancies. Examining the spatial mortality patterns for all cancers together is informative and is used as a way of determining cancer burden (177). However, most recent research in European cancer mortality also examines specific cancer mortality (2, 10, 148, 149, 152, 160, 178-181). These studies complement the work carried out in this chapter by suggesting that examining specific cancer mortality allows the burden of cancer to be delineated in more detail.

A final similarity that was observed across cancers was that Finland consistently had low risks of the disease. Effective measures were adopted on tobacco as well as on diet in Finland, and it has shown that total cancer mortality declined by over 40% in males aged 55 to 64 over a forty-year period up until 1994 (182). This country is considered to have the most effective overall programmes to reduce cancer mortality (179) and indicates the importance and scope for intervention on cancer control on population level in the EU. Much of the European cancer mortality observed here is, at least in theory, highly preventable with some regions, previously mentioned, requiring urgent preventative intervention on tobacco and diet modification.

Although the main purpose of this chapter was to explore the spatial distribution of cancer mortality in the EU, examining the various datasets provided an opportunity to explore further the effects of introducing a third higher level to the spatial model. Adding country as a higher level (random effect) takes into account the fact that areas within a country are more likely to be homogeneous than regions from different countries. The fact that regions are not independent has been taken into account through the spatial part of the model but including a higher country level takes account of the fact that there may be added differences at the borders where neighbours may be more heterogeneous. This

would probably be due to political and cultural differences and differences between health care systems across the nations. However, the disease maps from such models tended to over-smooth within countries and over-emphasize differences between countries and therefore appear to be useful for examining country level differences but are of limited use for determining the effect risk factors have on cancer rates in these countries and for providing smooth disease maps that enable cancer 'hotspots' to be identified.

Chapter 9

9 Discussion

9.1 Conclusions

The general aim of this thesis is to develop existing spatial modelling methods to provide an accurate account of the spatial patterning of cancer mortality across Europe. This task was effectively split into two parts; an ecological analysis on the burden of cancer mortality in Europe and the development of spatial multilevel models to analyse regional mortality data across various countries.

The ecological analysis initially involved identifying risk and protective factors for specific cancers and obtaining data that reflect the different levels of these factors. Examining the fixed parameter estimates from modelling these data allowed the relationship between the geographical variation of cancer mortality and the various explanatory covariates to be described. The random parameter estimates allowed the assessment of the contribution spatial factors have upon the disease. Relative risks of mortality were predicted from the models and used to provide smoothed disease maps of the risk of all cancer mortality and three specific cancer mortalities. Various maps were examined for each group of cancers enabling an overall assessment of the true underlying distribution of the disease. Variability within and between countries was evident across Europe for each of the cancer groups examined. Much of this variation could be accounted for by risk factors such as smoking and diet, which had strong yet differing effects on each of the cancer groups. Accounting for spatial effects was also shown to reduce variation in cancer mortality substantially across the EU. Cancer 'hotspots'

were identified and we provided evidence that urgent preventative public health intervention is required in many European regions.

Developing the spatial multilevel model involved extending work by Langford et al (6) in which they proposed a model with correlated random effects. This model had been fitted using Empirical Bayes procedures and was further explored in this thesis by fitting it to cancer mortality data across various countries. Using iterative generalised least squares procedures to fit this model had its disadvantages, one being the restrictions on adding further hierarchical levels. To overcome this problem the model was further developed using Markov Chain Monte Carlo (MCMC) procedures and overall resulted in a more flexible set of disease mapping models. Using this fully Bayesian approach gave the means to extend the model by adding a higher geographical level and the effects of doing so were explored; there was the suggestion that, depending on the specific reason for producing the disease map, adding a higher level to the multilevel model, such as country, often proved not useful. The models proposed in this thesis were compared to existing spatial models that can be implemented in the software MlwiN; the multiple-membership multiple-classification (MMMC) model and the conditional autoregressive (CAR) model. Overall the fully Bayesian spatial multilevel model proved to be more accurate and efficient at disease-risk estimation.

9.2 Limitations and Further Work

There are limitations to this study in the form of data available to carry out the ecological analysis, methods used to describe cancer mortality risk and further modelling issues that were beyond the scope of this PhD. These are discussed below with indications of areas where further work would be of interest.

9.2.1 Limitations with Data and Methods

There are some disadvantages to this type of ecological study even before any analysis has been carried out. All of the data are in the form of aggregate information, but an ideal situation would be to have data available at the individual level where, obviously, much more information would be available. However, this study was carried out on such a large scale geographically that obtaining this type of information is virtually impossible. The conclusions that have been drawn throughout the thesis on the burden of cancer mortality come under the risk of being affected by ecological fallacy whereby observations based on aggregate data are improperly inferred to an individual level. For these reasons it is important to make it clear that conclusions referring to the risk and distribution of cancer are being made about populations, mainly at a regional level.

Due to data availability, most of the risk factor data were only available at country level. This high level of aggregation is again not ideal due to the loss of information. However, it was shown that these fairly crude measures of exposure to the various risk factors are very useful, as they were shown to be significant predictors of the disease and they helped to explain a high amount of variation in all of the models explored. Once again care should be taken when interpreting the results from modelling the data; the risk factors tend to explain the differences between the countries and variation that remains between regions within countries could reflect the regional differences in these risk factors.

Another limitation was choosing which time period most closely reflects population's accumulated lifetime exposures to the risk factors. The time period chosen, approximately the same period as for the mortality data, was perhaps not ideal, and it would be of interest to explore the effects of using different time periods or perhaps more than one time period.

There are issues concerning the quality of the mortality, population and risk factor data. When collecting health data over such a large geographical area, recording and handling of the data varies between countries and inevitably this

results in variations in data availability and accuracy. Organisations like WHO carry out vigorous data checking procedures to minimise poor data quality but, since we are relying on many sources of data collection, these problems will exist. In this study one should remain aware that conclusions being drawn are fairly crude and tend to be used as an exploratory tool and for the generation of hypotheses about individual risks from the disease.

As discussed above, missing data are also inevitable in this type of study. From the initial mortality and population data discussed in Chapter 3 it was seen that data availability is very poor for eastern European countries. This was the main motivation for concentrating on western European countries, where data are much more abundant and more countries have data available at the lowest level of aggregation. However, missing data were still a problem when examining cancer mortality in the EU as three countries have missing data for the time point of interest. For the EU, population and mortality data tends to be available for at least two of the four time points, and further work incorporating time in the modelling would help to overcome some problems. Extending the models to examine spatiotemporal effects would use more of the information that is available and provide disease maps of the whole of the EU. Perhaps more importantly, considering an analysis with an added temporal dimension would allow the examination of the change in disease distribution and covariate effects over time. In theory, this could be easily implemented: the spatial multilevel model could be extended by adding a further lower level and modelling the temporal trend as a random effect. A further time/year level would enable the inclusion of up to four years of data per area.

Examining only the EU data means much of the European mortality data is being discarded. It would be of interest to explore cancer patterns for the countries farther east in Europe. An initial analysis looking at Europe as a whole was carried out and briefly discussed in Chapter 5, and it emerged that the countries in Eastern and Western Europe behaved very differently in terms of distribution of the disease risk. This along with the fact that eastern Europeans have very

different lifestyle habits to those in the EU suggests that it does not make sense to model all of the countries together. It would make sense to carry out separate analyses and compare the results for the different groups of countries. To do so one would have to consider where to split Europe. Choosing to examine the EU was simply based on geopolitical borders. However, this does not necessarily mean that comparing the EU with eastern and central Europe would be the best statistical comparison. An initial analysis would need to be carried out to determine which countries are the most similar in terms of cancer risk and lifestyle habits and comparison groups could be drawn from these.

To provide a good overview of the spatial distribution of cancer mortality, all cancers grouped together were modelled and also two of the most common cancers, lung and colorectal, were examined along with oesophageal cancer which is one of those most deadly malignancies. However, there are other cancers that add to the burden of cancer mortality in Europe. Four other groups of malignant neoplasms are available from the WHO dataset, and examining the spatial patterns of these would provide a more comprehensive overview of the European cancer risk. There are also other causes of death available from the WHO mortality dataset which pose a burden on health in Europe. Applying the same models to different causes of mortality would describe the distribution of risk from these other diseases and their relationships with certain risk factors, and would provide information on where further public health interventions may be required. A further extension to the spatial multilevel model is to predict more than one outcome simultaneously (102) and potentially, with the availability of various mortality data, multiple causes of death could be examined.

As discussed in Chapter 8 there is abundant research suggesting that certain genes or gene mutations predispose people to cancers and with deadly cancers, such as lung or oesophageal, genetic predisposition is therefore related to mortality rates of the disease. There is much ongoing research into pinpointing specific genes that are related to specific cancers, and there is some existing research that suggests that gene frequencies have spatial patterning. However,

more work is needed with regards to the geographical patterning of the gene frequencies related to cancers and their interactions with the environment and lifestyles of populations.

9.2.2 Modelling Issues

The addition of further hierarchical levels to the initial spatial multilevel model was explored using an MCMC framework. Adding a higher geographical level takes account of the fact that regions within the same country are more homogeneous than regions in different countries. Multilevel modelling aims to produce more accurate estimates by taking account of the non-independence at the lower levels. However, in this case regional non-independence has already been taken account of by fitting the spatial model, and when country level is added it appears to cause further uncertainty in the residuals and parameter estimates. This was evident in each model that was fitted, but further exploration into why this is happening would be useful to determine when adding the higher geographical level is appropriate. It appears from this analysis that it may only be useful for determining country effects and in fact not useful for mapping the distribution of the disease.

The deviance information criterion (DIC) was used to aid model selection but gives little help in assessing how well the models fitted the data. Residuals were examined to informally assess overall goodness-of-fit but there is a lack of literature on measuring model fit for complex multilevel models. There is scope to carry out further work on this set of models to determine a method of assessing model fit.

It would be of interest to extend this study and carry out work that examines the effect different prior distributions have on model fit. Several different prior distributions for the variance parameters would have to be identified and fitted to the spatial multilevel model that was chosen as the 'best' disease mapping model. Various sets of priors could be fitted to the datasets already examined and parameter estimates and coverage of confidence intervals could then be compared.

Alternatively, a method that may be more informative is a simulation study. This would involve simulating many datasets from the same distributions and using each set of chosen priors, fit the model to these datasets. For each set of priors, parameter estimates are obtained by finding the mean values over the numerous simulations; these can then be compared on the basis of how biased the methods are and how well the confidence intervals they produce cover the data. This method has the advantage of the true 'answers' being known.

A final limitation to the modelling methods is the length of time it takes to fit the spatial multilevel model with correlated random effects using MCMC methods. The number of iterations and time required to run a suitable number of simulations were very high. This situation is obviously not ideal, especially as there is scope to expand the data to include more countries and timepoints which would result in slowing the process down further. The model is complex and inevitably will take numerous MCMC runs to fit. However, it would be of interest to examine the effect different prior distributions have on modelling time required. This could be monitored whilst carrying out the simulation study discussed above.

Appendix 1: WHO European Region

A1.1 EU members

EU Member States	
Previous to October 2004	After October 2004
Austria	Cyprus
Belguim	Czech Republic
Denmark	Estonia
Finland	Hungary
France	Latvia
Germany	Lithuania
Greece	Malta
Ireland	Poland
Italy	Slovakia
Luxembourg	Slovenia
Netherlands	
Portugal	
Spain	
Sweden	
UK	

A1.2 Region names for codes given in Tables 3.2 – 3.5

Country	Region name (WHO code)					
Austria	Burgenland	(AT01)	Vorarlberg	(AT08)	Vienna	(AT09)
Azerbaijan	Other regions	(AZ01)	Nakhichevan	(AZ03)	Baku	(AT04)
Belarus	Grodno	(BY02)	Minsk city	(BY0401)	Vitebsk	(BY06)
	Gomel	(BY03)				
Belgium	Flemish Region	(BE1)	Walloon Region	(BE2)	Brussels	(BE3)
Bulgaria		(BG02)	Plovdiv	(BG05)	Sofia city	(BG07)
	Mikhaylovgrad	(BG04)	Razgrad	(BG06)		
Czech Republic	Prague	(CZ01)	Jihomoravsky	(CZ03)	Severomora-vsky	(CZ05)
	Jihoeesky	(CZ02)			Stredoeesky	(CZ06)
Denmark	Copenhagen and Frederiksberg (city)	(DK011)	Frederiksborg	(DK013)	Bornholm	(DK023)
			Roskilde	(DK014)	Sonderjylland	(DK032)
	Copenhagen	(DK012)				
Finland	Ahvenanmaa	(FL01)	Mikkeli	(FL07)	Oulu	(FL11)
	Lappi	(FL06)	Uusimaa	(FL08)		
France	Ile de France	(FR01)	Limousin	(FR63)	Corsica	(FR83)
	Pays de la Loire	(FR51)				
Germany	Hamburg	(DE2)	Berden Wurttem-berg	(DE8)	Saxony	(DEE)
	Bremen	(DE4)	Brandenburg	(DEC)	Berlin (West)	(DEBW)
	North Rhine-Westphalia	(DE5)				
Greece	Attica	(GR3)	Ionian Islands	(GR22)	Aegean North	(GR42)
	Macedonia Central	(GR12)	Aegean South	(GR41)		
Hungary	Budapest	(HU05)	Komarom-Esztergom	(HU12)	Nograd	(HU15)
	Fejer	(HU07)	Somogy	(HU13)		(HU16)
Italy	Lombary	(IT2)	Liguria	(IT13)	Puglia	(IT91)
	Valle d'Aosta	(IT12)		(IT33)		(IT93)
Kazakstan	South Kazakstan	(KZ03)	Atyrau	(KZ06)	Turgay	(KZ16)
Kyrgystan	Bishkek	(KG01)	Issyk-Kul	(KG04)	Talas	(KG07)
	Dzhalal-Abad	(KG03)	Osh	(KG06)		
Netherlands	Groningen	(NL11)	Noord-Brabant	(NL51)	Zeeland	(NL74)
	Flevoland	(NL25)	Zuid-Holland	(NL73)		
Norway	Finnmark	(NO03)	Oslo og Akershus	(NO10)	Rogalnd	(NO12)
	Hedmark	(NO04)	Østfold	(NO11)		
Poland	Chelm	(PL05)	Legnica	(PL19)	Nowy Sacz	(PL24)
	Katowice	(PL13)	Lodz	(PL21)	Sieradz	(PL36)
		(PL16)				
Portugal	Azores	(PT2)	Norte	(PT11)	Alentejo	(PT14)
Romania	Arad	(RO02)	Bucharest	(RO10)	Iasi	(RO25)
	Bacau	(RO04)	Covasna	(RO16)	Vaslui	(RO39)
	Brasov	(RO08)				
Russian Federation	Tura	(RU1005)	Pskov	(RU0203)	Moskow city	(RU0314)
	Chukotka	(RU1109)	St Petersbug	(RU0204)		
Slovakia	Bratislava	(SK01)	Zapadoslovensky	(SK04)		

(Continued over page)

Spain	Asturias	(ES12)	Extremadura	(ES43)	Ceutay Melilla	(ES63)
	Madrid	(ES3)	Andalusia	(ES61)	Canary Island	(ES7)
Sweden	Gävleborg	(SE04)	Jämtland	(SE07)	Stockholm	(SE19)
	Gotland	(SE05)	Norrbotten	(SE14)		
Switzerland	Appenzell-Inner Rhoden	(CH03)	Basle city	(CH05)	Zug	(CH25)
	Basle	(CH04)	Uri	(CH22)	Zurich	(CH26)
Tajikistan	Dushanke	(TJ01)	Kuliabsk	(TJ04)	Other regions	(TJ07)
	Gorno-Budakhshan	(TJ02)	Khudzand	(TJ05)		
Turkmenistan	Ashgabat	(TM01)	Tahauz	(TM04)	Other regions	(TM05)
Ukraine	Chernihiv	(UA0102)	Kiev city	(UA0114)	Kirovohrad	(UA0304)
	Chermihiv	(UA0103)	Donetsk	(UA0302)	Zaporizhya	(UA0308)
United Kingdom	Bedfordshire	(UK511)	East Sussex	(UK531)	Powys	(UK914)
	Berkshire	(UK521)	Greater London	(UK55)	Northen Ireland	(UKB)
	Buckinghamshire	(UK522)	Isle of Wight	(UK562)		
Uzbekistan	Dzhizak	(UZ03)	Syr-Darya	(UZ11)	Tashkent city	(UZ13)
	Samarkand	(UZ09)				
Yugoslavia	Montenegro	(YU01)	Serbia	(YU03)		

Appendix 2: WinBugs code

A2.1 Code for Multiple-Membership Multiple Classification (MMMC) model

```
#----MODEL Definition----------

model
{
# Level 1 definition
for(i in 1:N) {
deaths[i] ~ dpois(mu[i])
log(mu[i]) <- offs[i] + beta[1]
+ beta[2] * smoke[i]
+ beta[3] * fruit[i]
+ beta[4] * veg[i]
+ beta[5] * animal[i]
+ beta[6] * alcohol[i]
+ beta[7] * GDP[i]
+ u2[region[i]]
+ weight1[i] * u3[neigh1[i]]
+ weight2[i] * u3[neigh2[i]]
+ weight3[i] * u3[neigh3[i]]
+ weight4[i] * u3[neigh4[i]]
+ weight5[i] * u3[neigh5[i]]
+ weight6[i] * u3[neigh6[i]]
+ weight7[i] * u3[neigh7[i]]
+ weight8[i] * u3[neigh8[i]]
+ weight9[i] * u3[neigh9[i]]
+ weight10[i] * u3[neigh10[i]]
+ weight11[i] * u3[neigh11[i]]
+ weight12[i] * u3[neigh12[i]]
}
# Higher level definitions
for (j in 1:n2) {
u2[j] ~ dnorm(0,tau.u2)
}
for (j in 1:n3) {
u3[j] ~ dnorm(0,tau.u3)
}
# Priors for fixed effects
for (k in 1:7) { beta[k] ~ dflat() }
# Priors for random terms
tau.u2 ~ dgamma(0.001000,0.001000)
sigma2.u2 <- 1/tau.u2
tau.u3 ~ dgamma(0.001000,0.001000)
sigma2.u3 <- 1/tau.u3
}
```

A2.2 Code for Conditional Autoregressive (CAR) model

```
#----MODEL Definition----------

model
{
# Level 1 definition
for(i in 1:N) {
deaths[i] ~ dpois(mu[i])
log(mu[i]) <- offs[i] + beta[1] * smoke[i]
+ beta[2] * fruit[i]
+ beta[3] * veg[i]
+ beta[4] * animal[i]
+ beta[5] * alcohol[i]
+ beta[6] * GDP[i]
+ carmean + u2[region[i]]
+ u3[region[i]]
}
# Higher level definitions
for (j in 1:n2) {
u2[j] ~ dnorm(0,tau.u2)
}
u3[1:n3] ~ car.normal(adj[],weights[],num[],tau.u3)
# Priors for fixed effects
for (k in 1:6) { beta[k] ~ dflat() }
carmean ~ dflat()
# Priors for random terms
tau.u2 ~ dgamma(0.001000,0.001000)
sigma2.u2 <- 1/tau.u2
tau.u3 ~ dgamma(0.001000,0.001000)
sigma2.u3 <- 1/tau.u3
}
```

A2.3 Data file for spatial multilevel model fitted in equations (6.5) and (6.6)

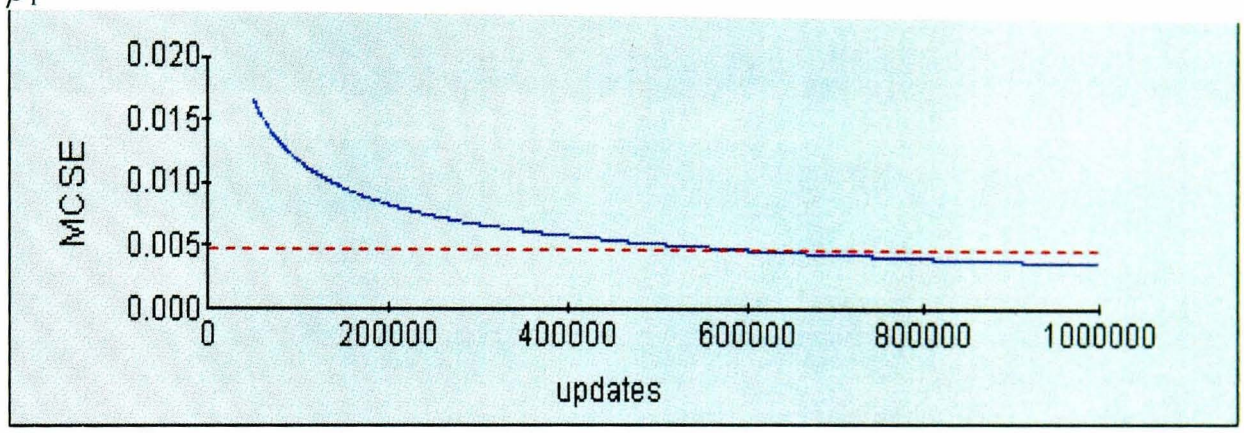
```
#----Data File---------------------------------

list(N= 187,
R2 = structure(
 .Data = c(0.0055, 0.002,
           0.002,   0.055),
 .Dim = c(2,2)),
region = c(1,2,3,4,5,6,7,8,9,10,11,12,...,187),
neigh1 = c(6,8,9,18,18,8,18,18,8,61,12,92,...,186),
neigh2 = c(3,7,8,6,7,5,8,9,3,34,10,91,...,1),
neigh3 = c(1,6,6,5,6,4,5,7,1,22,1,89,...,1),
neigh4 = c(1,5,4,3,4,3,2,6,1,12,1,25,...,1),
neigh5 = c(1,1,1,1,2,2,1,3,1,11,1,24,...,1),
neigh6 = c(1,1,1,1,1,1,1,2,1,1,1,22,...,1),
neigh7 = c(1,1,1,1,1,1,1,1,1,1,1,21,...,1),
neigh8 = c(1,1,1,1,1,1,1,1,1,1,1,15,...,1),
neigh9 = c(1,1,1,1,1,1,1,1,1,1,1,14,...,1),
neigh10 = c(1,1,1,1,1,1,1,1,1,1,1,13,...,1),
neigh11 = c(1,1,1,1,1,1,1,1,1,1,1,11,...,1),
neigh12 = c(1,1,1,1,1,1,1,1,1,1,1,10,..., 1),
weight1 = c(0.500,0.250,0.200,0.250,0.200,0.167,0.250,0.167,0.500,0.200,0.500,0.083,...,1.000),
weight2 = c(0.500,0.250,0.200,0.250,0.200,0.167,0.250,0.167,0.500,0.200,0.500,0.083,...,0.000),
weight3 = c(0.000,0.250,0.200,0.250,0.200,0.167,0.250,0.167,0.000,0.200,0.000,0.083,...,0.000),
weight4 = c(0.000,0.250,0.200,0.250,0.200,0.167,0.250,0.167,0.000,0.200,0.000,0.083,...,0.000),
weight5 = c(0.000,0.000,0.200,0.000,0.200,0.167,0.000,0.167,0.000,0.200,0.000,0.083,...,0.000),
weight6 = c(0.000,0.000,0.000,0.000,0.000,0.167,0.000,0.167,0.000,0.000,0.000,0.083,...,0.000),
weight7 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
weight8 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
weight9 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
weight10 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
weight11 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
weight12 = c(0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.000,0.083,...,0.000),
smoke = c(2210.0,2210.0,2210.0,2210.0,2210.0,2210.0,2210.0,2210.0,2210.0,2360.0,2360.0,
2360.0,2360.0,...,2210.0),
fruit = c(139.2,139.2,139.2,139.2,139.2,139.2,139.2,139.2,139.2,118.8,118.8,118.8,...,74.5),
veg = c(80.6,80.6,80.6,80.6,80.6,80.6,80.6,80.6,80.6,78.6,78.1,78.1,...,88.2),
animal = c(21.1,21.1,21.1,21.1,21.1,21.1,21.1,21.1,21.1,19.9,19.9,19.9,..., 9.7),
alcohol = c(166.0,166.0,166.0,166.0,166.0,166.0,166.0,166.0,166.0,173.9,173.9,173.9,...,123.2),
gdp = c(13756.0,18239.0,17338.0,20710.0,25118.0,17903.0,22431.0,22519.0,33902.0,22100.0,
38689.0,21355.0,...,11461.0),
offs = c(-0.998,-0.339,0.700,0.498,-0.551,0.461,-0.296,-1.015,0.886,1.337,0.943,2.355,...,0.593),
deaths = c(769.0,1407.0,3779.0,2875.0,1012.0,2890.0,1267.0,642.0,4676.0,7379.0,5310.0,
20009.0,...,3486.0))
```

The data file is shown for regions 1-12 and 187. The weight structure is given through the variables neigh1,...,neigh12 and weight1,...weight12. For example region 1 has 2 neighbours (regions 6 and 3) and those neighbours are given equal weights of $1/n_i = 0.5$.

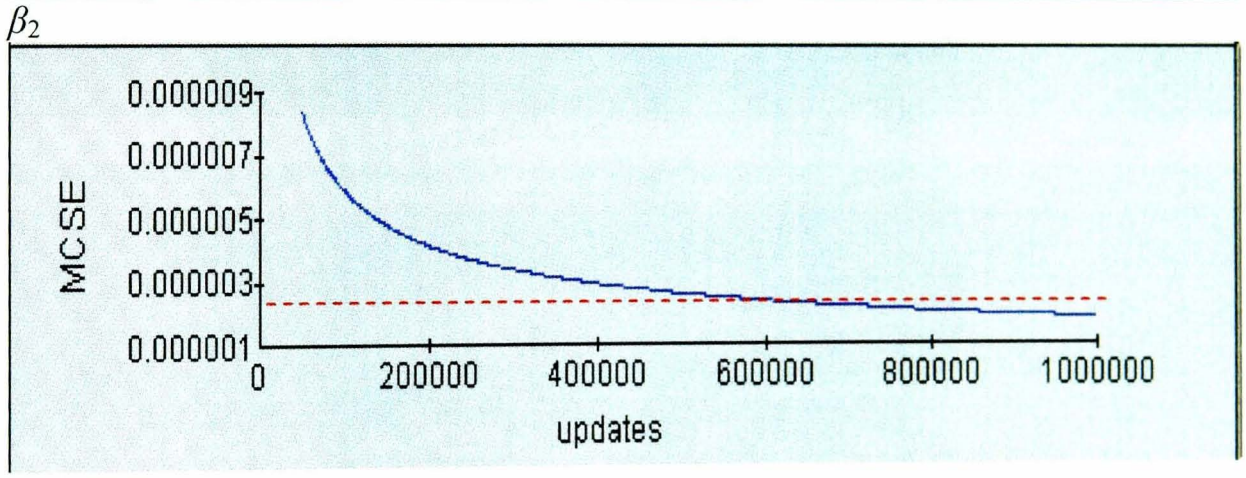
Appendix 3: Convergence Diagnostic Plots

A3.1 Convergence plots for fixed parameters from MMMC model

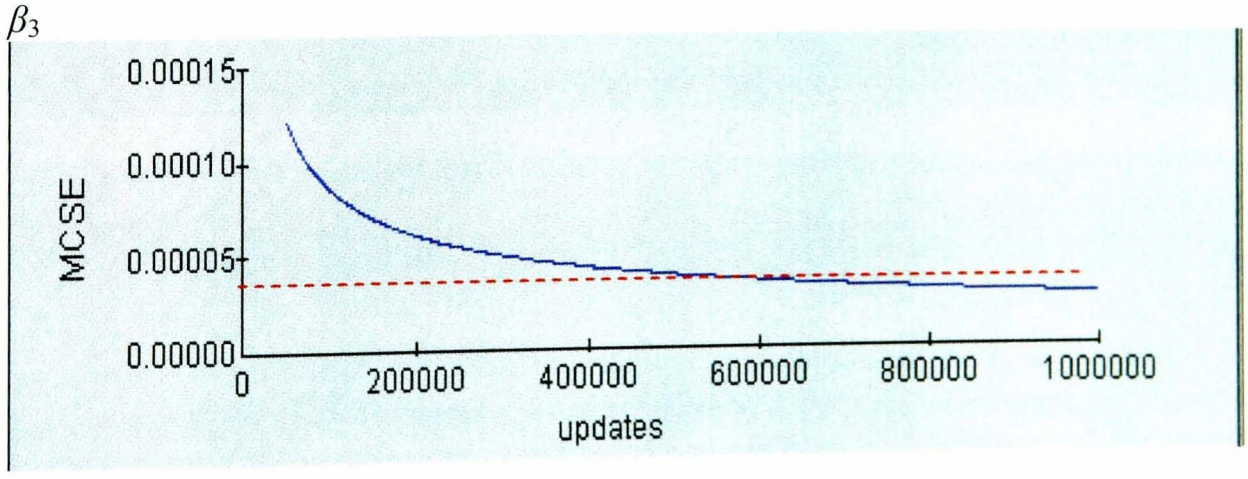
β_1



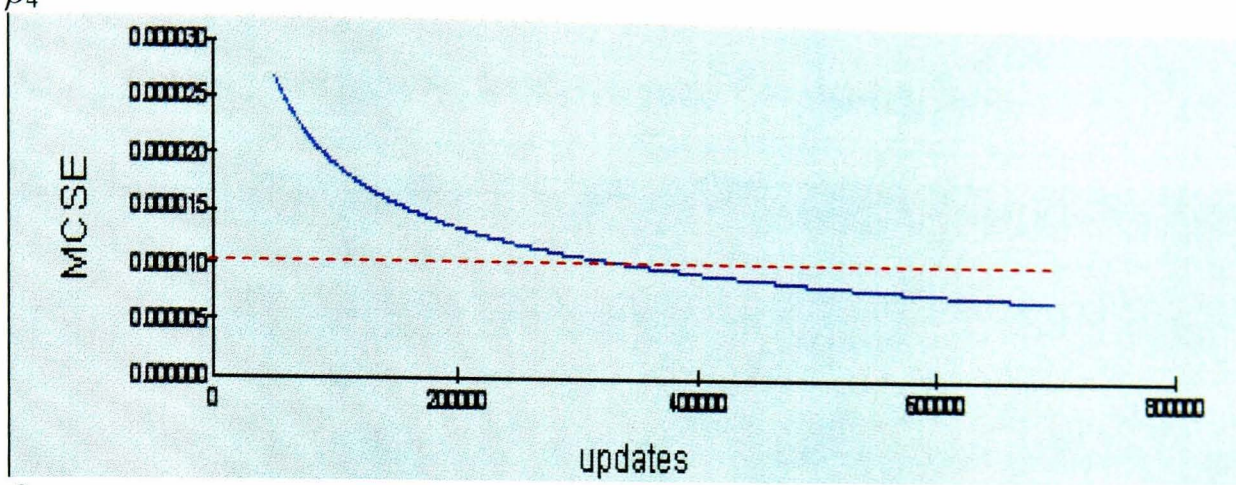
β_2



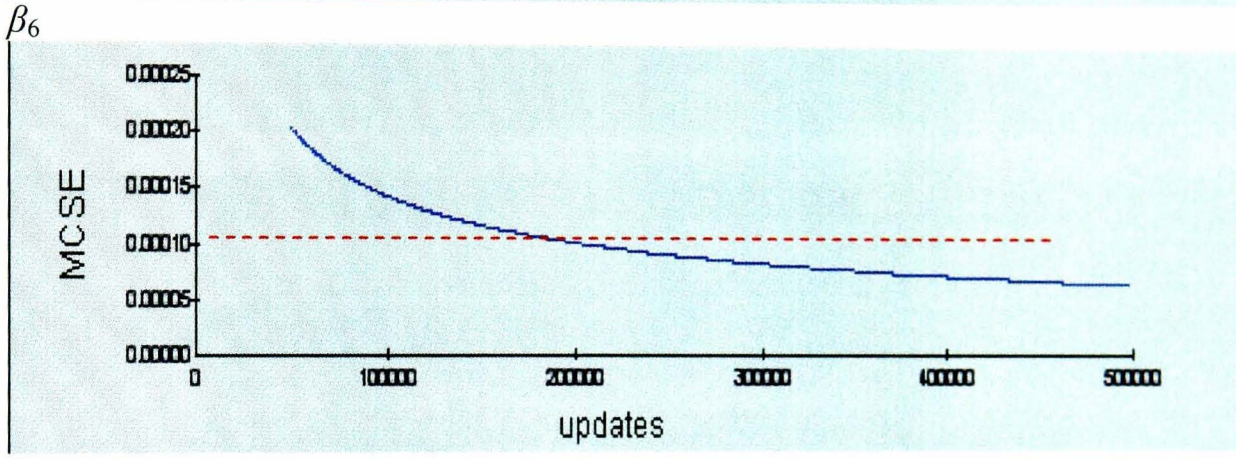
β_3



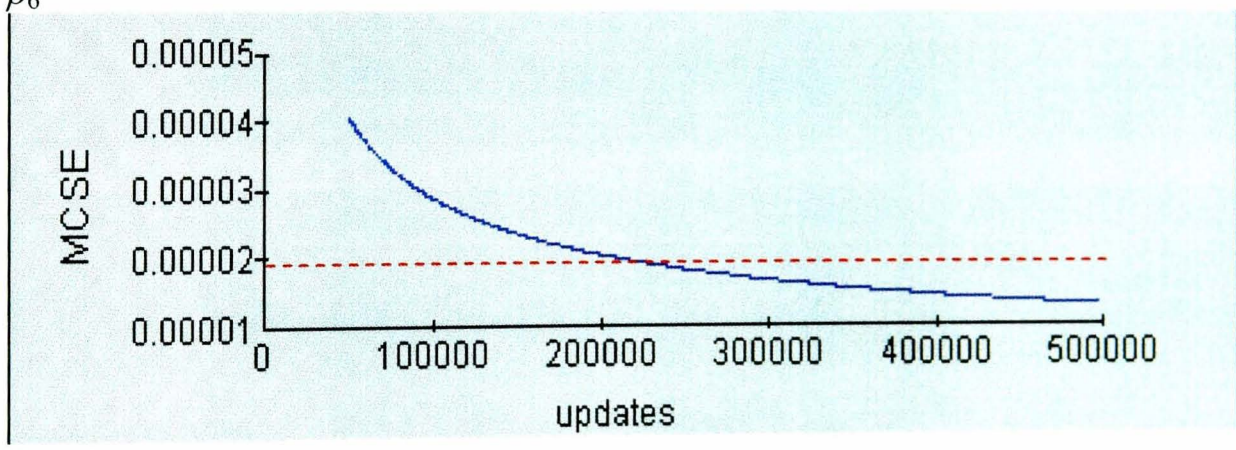
β_4



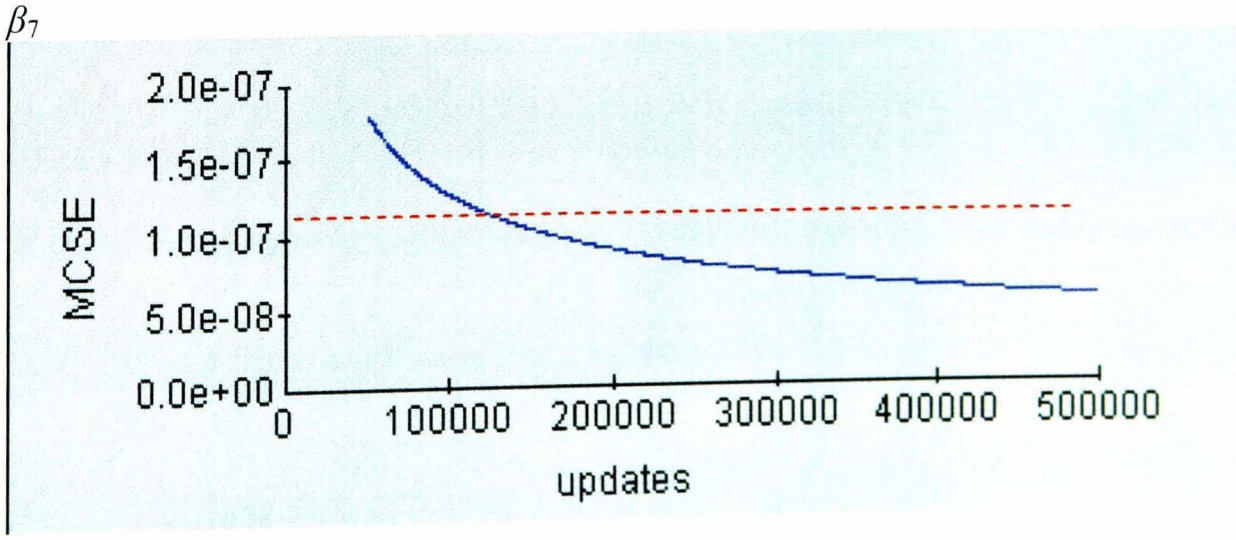
β_6



β_6

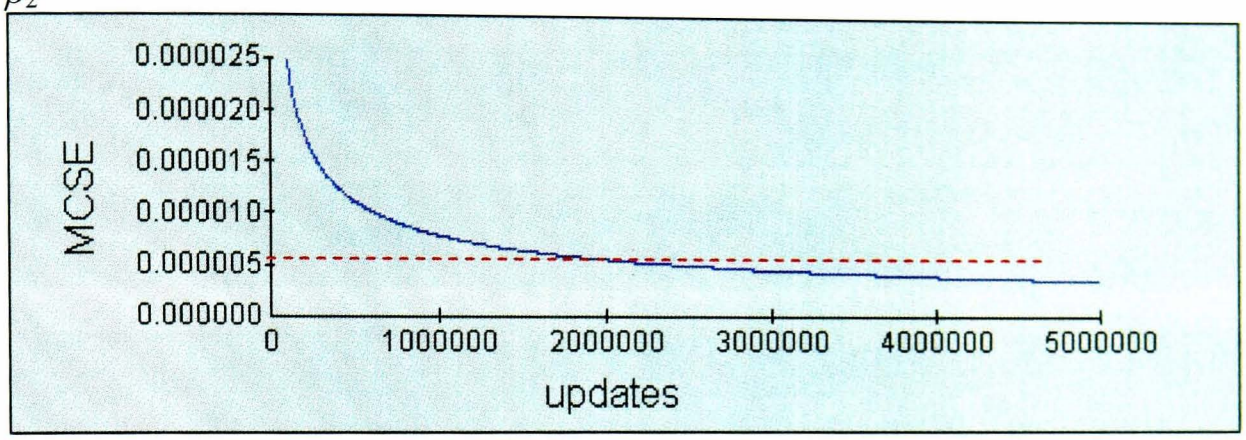


β_7

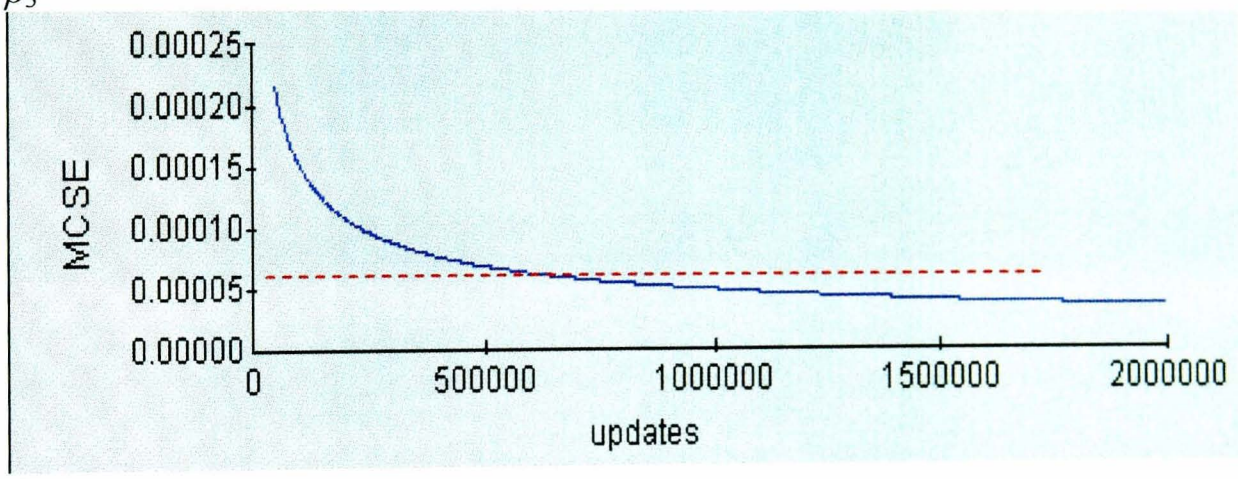


A3.2 Convergence plots for fixed parameters from MMMC model

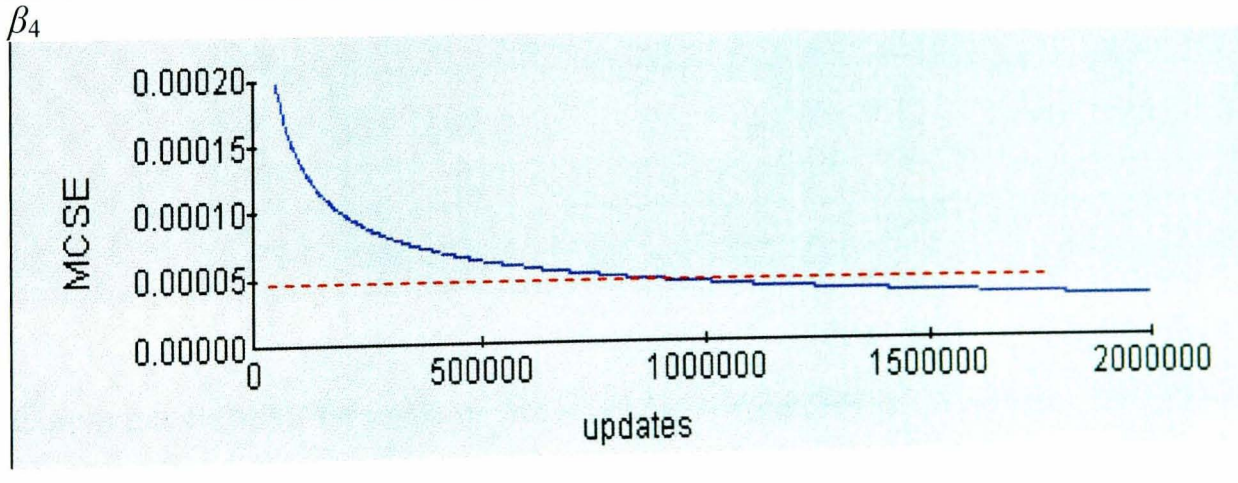
β_2



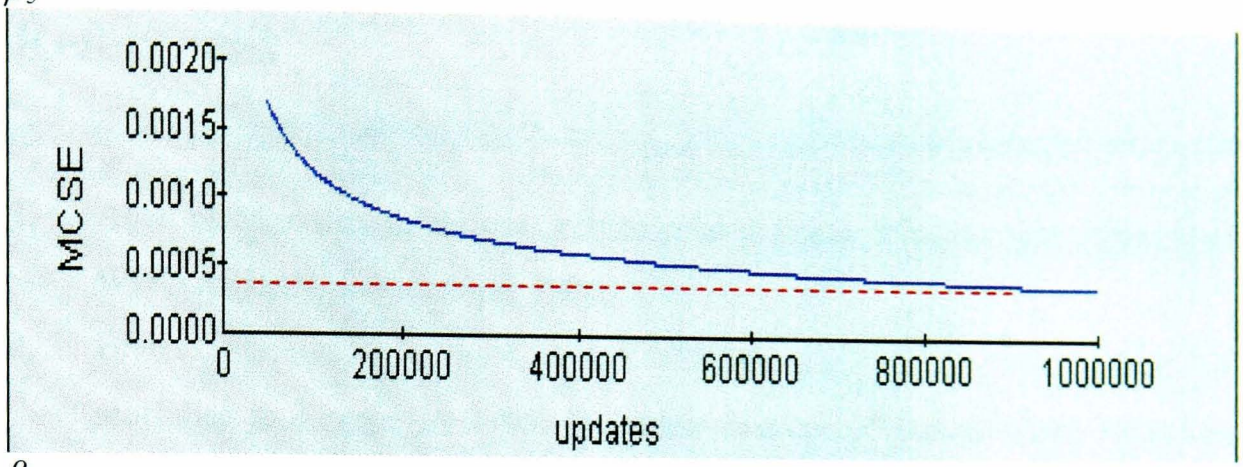
β_3



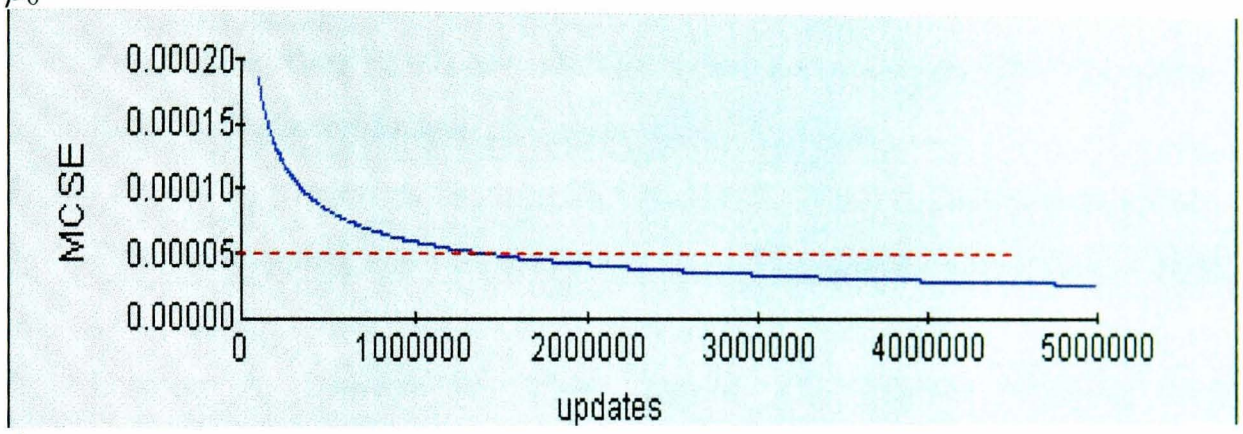
β_4



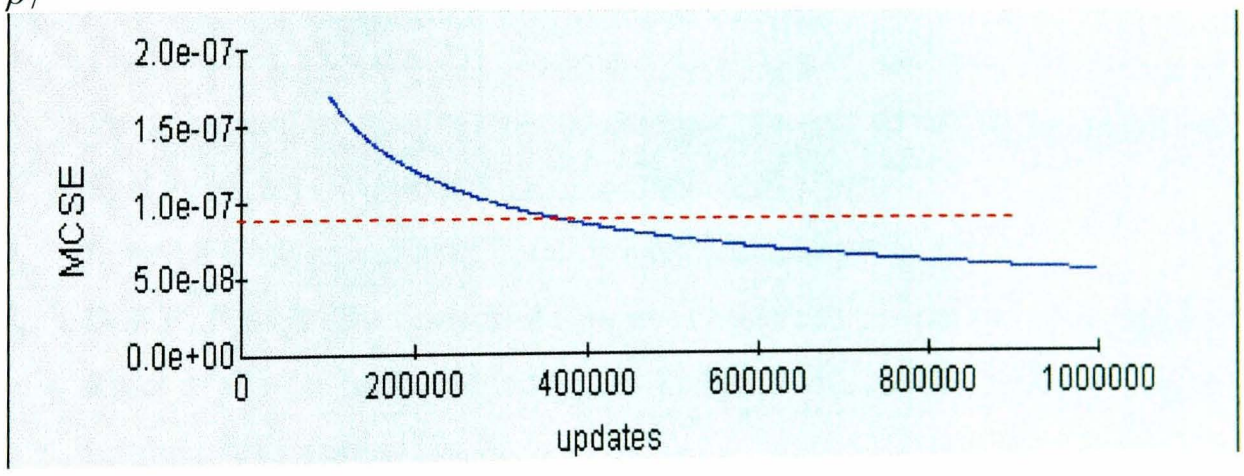
β_5



β_6



β_7



Convergence plots for each of the fixed parameters are given above. The Monte Carlo error is given on the Y-axis and the number of iterations is given on the X-axis. The blue line plots how many iterations are needed to achieve a posterior mean parameter estimate with a desired Monte Carlo error value. The red line plots the MC error value that is 5% of the posterior standard deviation; hence a suitable burn-in with an adequate number of further iterations has been achieved. Only the fixed parameters are given as previous modelling showed that these require the most number of iterations.

References

1. Atlas of Mortality in Europe: subnational patterns, 1980/81 and 1990/1991: WHO regional publications; 1997.
2. Bray F, Sankila R, Ferlay J, Parkin DM. Estimates of cancer incidence and mortality in Europe in 1995. European Journal of Cancer 2002;38(1):99-166.
3. Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000. The global picture. European Journal of Cancer 2001;37:S4-S66.
4. Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel J-F, Bertollini R, editors. Disease mapping and risk assesment for public health: John Wiley & Sons Ltd.; 1999.
5. Lawson A, Browne W, Vidal Rodeiro CL. Disease Mapping with WinBUGS and MLwiN. West Sussex: John Wiley & Sons Ltd; 2003.
6. Langford IH, Leyland AH, Rasbash J, Goldstein H. Multilevel modelling of the geographical distributions of diseases. Journal of the Royal Statistical Society Series C-Applied Statistics 1999;48:253-268.
7. Imperial College and MRC, UK. WinBUGS. In; 1996-2004.
8. Doll R, Peto R. The Causes of Cancer - Quantitative Estimates of Avoidable Risks of Cancer in the United-States Today. Journal of the National Cancer Institute 1981;66(6):1191-&.
9. Survival of Cancer Patients in Northern Ireland 1993-1996. Belfast: N. Ireland Cancer Registry; 2001.
10. Facts and Figures of Cancer in the European Community: Internaltional Agency for Research on Cancer, Commission of the European Communities, WHO "Europe Against Cancer"; 1993.
11. Berrino F, Capocaccia R, Esteve J, Gatta G, Hakulinen T, Micheli A, et al. Survival of Cancer Patients in Europe: The EUROCARE-2 Study: IARC Scientific Publications; 1999.

12. Lee MM, Lin SS. Dietary fat and breast cancer. Annual Review of Nutrition 2000;20:221-248.
13. Grant WB. An ecologic study of dietary and solar ultraviolet-B links to breast carcinoma mortality rates. Cancer 2002;94(1):272-281.
14. Boyle P, Maisonneuve P, Autier P. Update on cancer control in women. International Journal of Gynecology & Obstetrics 2000;70(2):263-303.
15. Ahn J, Gammon MD, Santella RM, Gaudet MM, Britton JA, Teitelbaum SL, et al. Myeloperoxidase genotype, fruit and vegetable consumption, and breast cancer risk. Cancer Research 2004;64(20):7634-7639.
16. Gandini. Meta-analysis of studies on breast cancer risk and diet: the role of fruit and vegetable consumption and the intake of associated micronutrients (vol 36, pg 636, 2000). European Journal of Cancer 2000;36(12):1588-1588.
17. Jain M. Dairy foods, dairy fats, and cancer: A review of epidemiological evidence. Nutrition Research 1998;18(5):905-937.
18. Mattisson I, Wirfalt E, Wallstrom P, Gullberg B, Olsson H, Berglund G. High fat and alcohol intakes are risk factors of postmenopausal breast cancer: a prospective study from the Malmo diet and cancer cohort. International Journal of Cancer 2004;110(4):589-97.
19. Tjonneland A, Thomsen BL, Stripp C, Christensen J, Overvad K, Mellemkaer L, et al. Alcohol intake, drinking patterns and risk of postmenopausal breast cancer in Denmark: a prospective cohort study. Cancer Causes & Control 2003;14(3):277-284.
20. Rehm J, Room R, Graham K, Monteiro M, Gmel G, Sempos CT. The relationship of average volume of alcohol consumption and patterns of drinking to burden of disease: an overview. Addiction 2003;98(9):1209-1228.
21. Bowlin SJ, Leske MC, Varma A, Nasca P, Weinstein A, Caplan L. Breast cancer risk and alcohol consumption: Results from a large case-control study. International Journal of Epidemiology 1997;26(5):915-923.
22. Howe GR, Hirohata T, Hislop TG, Iscovich JM, Yuan JM, Katsouyanni K, et al. Dietary factors and risk of breast cancer: combined analysis of 12 case-

control studies. Journal of the National Cancer Institute 1990;83(20):1501-1507.

23. Willett WC, Rickhill B, Hankinson SE, Hunter DJ, Colditz GA. Non-genetic factors in the causation of breast cancer. In: Harris J, Lippman ME, Morrow M, Osborne CK, editors. Diseases of the Breast. Philadelphia: Lippincott-Raven Press; 2000. p. 175-220.
24. Dove-Edwin I, Thomas HJW. Review article: The prevention of colorectal cancer. Alimentary Pharmacology & Therapeutics 2001;15(3):323-336.
25. Detels R, Holland WW, McEwen J, Omenn GS, editors. Oxford Textbook of Public Health: Oxford Medical Publications; 1997.
26. Slattery ML, Levin TR, Ma K, Goldgar D, Holubkov R, Edwards S. Family history and colorectal cancer: predictors of risk. Cancer Causes & Control 2003;14(9):879-887.
27. Willett W. The Search for the Causes of Breast and Colon Cancer. Nature 1989;338(6214):389-394.
28. Willett WC, Stampfer MJ, Colditz GA, Rosner BA, Speizer FE. Relation of Meat, Fat, and Fiber Intake to the Risk of Colon Cancer in a Prospective-Study among Women. New England Journal of Medicine 1990;323(24):1664-1672.
29. Neagoe A, Molnar AM, Acalovschi M, Seicean A, Serban A. Risk factors for colorectal cancer: an epidemiologic descriptive study of a series of 333 patients. Rom J Gastroenterol 2004;13(3):187-193.
30. Fernandez E, Chatenoud L, La Vecchia C, Negri E, Franceschi S. Fish consumption and cancer risk. American Journal of Clinical Nutrition 1999;70(1):85-90.
31. Tavani A, Pregnolato A, LaVecchia C, Negri E, Talamini R, Franceschi S. Coffee and tea intake and risk of cancers of the colon and rectum: A study of 3,530 cases and 7,057 controls. International Journal of Cancer 1997;73(2):193-197.

32. Klatsky AL, Armstrong MA, Friedman GD, Hiatt RA. The Relations of Alcoholic Beverage Use to Colon and Rectal- Cancer. American Journal of Epidemiology 1988;128(5):1007-1015.
33. Schouten LJ. Oesophageal and stomach cancer. In: Damhuis RAM, Schouten LJ, Visser O, editors. Gastrointestinal Cancer in the Netherlands 1989-1992. Netherlands: Netherlands Cancer Registry; 2001. p. 1-8.
34. Shetty PS, W.P.T J. Nutrition. In: Detels R, Holland WW, McEwen J, Omenn GS, editors. Oxford Textbook of Public Health. 3rd ed: Oxford University Press; 1997. p. 157-174.
35. Cayuela A, Vioque J, Bolumar F. Esophageal Cancer Mortality - Relationship with Alcohol Intake and Cigarette-Smoking in Spain. Journal of Epidemiology and Community Health 1991;45(4):273-276.
36. World Cancer Researxh Fund. Lung. In: Food, Nutrition and the Prevention of Cancer: A Global Perspective (Part II, Cancers, Nutrition and Food). Washingoton DC: American Institute for Cancer Research; 1997. p. 130-147.
37. Hill MJ. Changes and developments in cancer prevention. Journal of the Royal Society for the Promotion of Health 2001;121(2):94-97.
38. Mulder I, Jansen M, Smit HA, Jacobs DR, Menotti A, Nissinen A, et al. Role of smoking and diet in the cross-cultural variation in lung-cancer mortality: the Seven Countries Study (vol 88, pg 665, 2000). International Journal of Cancer 2001;91(6):901-901.
39. Breslow RA, Graubard BI, Sinha R, Subar AF. Diet and lung cancer mortality: a 1987 National Health Interview Survey cohort study. Cancer Causes & Control 2000;11(5):419-431.
40. Kubik AK, Zatloukal P, Tomasek L, Petruzelka L. Lung cancer risk among Czech women: A case-control study. Preventive Medicine 2002;34(4):436-444.
41. Chow WH, Schuman LM, McLaughlin JK, Bjelke E, Gridley G, Wacholder S, et al. A Cohort Study of Tobacco Use, Diet, Occupation, and Lung-Cancer Mortality. Cancer Causes & Control 1992;3(3):247-254.

42. Skuladottir H, Tjoenneland A, Overvad K, Stripp C, Christensen J, Raaschou-Nielsen O, et al. Does insufficient adjustment for smoking explain the preventive effects of fruit and vegetables on lung cancer? Lung Cancer 2004;45(1):1-10.
43. Fabricius P, Lange P. Diet and lung cancer. Monaldi Arch Chest Dis 2003;59(3):207-211.
44. Boffetta P, Trichopoulos D. Cancer of the Lung, Larynx, and Pleura. In: Adami H, Hunter D, Trichopoulos D, editors. Textbook of cancer epidemiology. Oxford: Oxford University Press, Inc.; 2002.
45. Manjer J, Andersson I, Berglund G, Bondesson L, Garne JP, Janzon L, et al. Survival of women with breast cancer in relation to smoking. European Journal of Surgery 2000;166(11):852-858.
46. Collaborative Group on Hormonal Factors in Breast Cancer. Alcohol, tobacco and breast cancer - collaborative reanalysis of individual data from 53 epidemiological studies, including 58 515 women with breast cancer and 95 067 women without the disease. British Journal of Cancer 2002;87(11):1234-1245.
47. Rookus MA, Verloop J, de Vries F, van der Kooy K, van Leeuwen FE. Passive and active smoking and the risk of breast cancer. American Journal of Epidemiology 2000;151(11):109.
48. Wartenberg D, Calle EE, Thun MJ, Heath CW, Lally C, Woodruff T. Does passive smoking exposure increase female breast cancer mortality? American Journal of Epidemiology 2000;151(11):110.
49. Robert SA, Strombom I, Trentham-Dietz A, Hampton JM, McElroy JA, Newcomb PA, et al. Socioeconomic risk factors for breast cancer: distinguishing individual- and community-level effects. Epidemiology 2004;15(4):442-450.
50. Tavani A, Gallus S, La Vecchia C, Negri E, Montella M, Dal Maso L, et al. Risk factors for breast cancer in women under 40 years. European Journal of Cancer 1999;35(9):1361-1367.

51. Schrijvers CTM, Mackenbach JP, Lutz JM, Quinn MJ, Coleman MP. Deprivation and Survival from Breast-Cancer. British Journal of Cancer 1995;72(3):738-743.
52. Consedine NS, Magai C, Conway F, Neugut AI. Obesity and awareness of obesity as risk factors for breast cancer in six ethnic groups. Obes Res 2004;12(10):1680-1689.
53. Sweeney C, Blair CK, Anderson KE, Lazovich D, Folsom AR. Risk factors for breast cancer in elderly women. American Journal of Epidemiology 2004;160(9):868-875.
54. Enger SM, Greif JM, Polikoff J, Press M. Body weight correlates with mortality in early-stage breast cancer. Arch Surg. 2004;139(9):954-958.
55. Giovannucci E, Rimm EB, Stampfer MJ, Colditz GA, Ascherio A, Kearney J, et al. A prospective study of cigarette smoking and risk of colorectal adenoma and colorectal cancer in U.S. men. Journal of the National Cancer Institute 1994;86(3):162-164.
56. Tavani A, Gallus S, Negri E, Franceschi S, Talamini R, La Vecchia C. Cigarette smoking and risk of cancers of the colon and rectum: a case-control study from Italy. European Journal of Epidemiology 1998;14(7):675-691.
57. Nyren O, Bergstrom R, Nystrom L, Engholm G, Ekbom A, Adami HO, et al. Smoking and colorectal cancer: a 20-year follow-up study of Swedish construction workers. Journal of the National Cancer Institute 1997;89(1):95-96.
58. Nyren O, Adami H. Esophageal Cancer. In: Adami H, Hunter D, Trichopoulos D, editors. Textbook of cancer epidemiology. Oxford: Oxford University Press, Inc.; 2002.
59. Brown LM, Swanson CA, Gridley G, Swanson GM, Schoenberg JB, Greenberg RS, et al. Adenocarcinoma of the Esophagus - Role of Obesity and Diet. Journal of the National Cancer Institute 1995;87(2):104-109.
60. Vaughan TL, Davis S, Kristal A, Thomas DB. Obesity, Alcohol, and Tobacco as Risk-Factors for Cancers of the Esophagus and Gastric Cardia -

Adenocarcinoma Versus Squamous-Cell Carcinoma. Cancer Epidemiology Biomarkers & Prevention 1995;4(2):85-92.

61. Chow WH, Blot WJ, Vaughn TL, Risch HA, Gammon MD, Stanford JL, et al. Body mass index and risk of adenocarcinomas of the esophagus and gastric cardia. Journal of the National Cancer Institute 1998;90(2):150-155.
62. Devesa SS, Blot WJ, Fraumeni JF. Changing patterns in the incidence of esophageal and gastric carcinoma in the United States. Cancer 1998;83(10):2049-2053.
63. Lagergren J, Bergstrom R, Nyren O. Association between body mass and adenocarcinoma of the esophagus and gastric cardia. Annals of Internal Medicine 1999;130(11):883-+.
64. Baynard SP, Jinot J, Leaderer BP, Brown GB, Matrtinex FD, Simonsen NR, et al. Respiratory Health Effects of Passive Smoking: Cancer and Other Disorders. Washington DC: Office of Health and Environmental Assessment; 1992.
65. Janssen-Heijnen MLG, Gatta G, Forman D, Capocaccia R, Coebergh JWW. Variation in survival of patients with lung cancer in Europe, 1985-1989. European Journal of Cancer 1998;34(14):2191-2196.
66. Schrijvers CTM, Coebergh JWW, Vanderheijden LH, Mackenbach JP. Socioeconomic Variation in Cancer Survival in the Southeastern Netherlands, 1980-1989. Cancer 1995;75(12):2946-2953.
67. Adami H, Hunter D, Trichopoulos D, editors. Textbook of cancer epidemiology. Oxford: Oxford University Press, Inc.; 2002.
68. Murray CJL, Lopez AD, editors. The global burden of disease : a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020. Cambridge, MA: Harvard University Press; 1996.
69. Food and Agriculture Organisation of the United Nations; Online Database. In: http://apps.fao.org/.
70. World Health Organisation. Tobacco or Health: A Global Status Report. http://www.cdc.gov/tobacco/WHO/; 1997.

71. Office on Smoking and Health of the National Centre for Chronic Disease Prevention and Health Promotion. In: http://www.cdc.gov/tobacco/.
72. WHO Statistical Information System (WHOSIS). In: http://www3.who.int/whosis/menu.cfm.
73. EUROSTAT. In: http://europa.eu.int/comm/eurostat.
74. Cromley EK, McLafferty SL. GIS and Public Health. 1st ed. New York: The Guilford Press; 2002.
75. Muehrcke PC, Muehrcke JO. Map use : reading, analysis, and interpretation. 3rd ed ed. Madison, Wis.: JP Publications,; 1992.
76. Clayton D, Kaldor J. Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. Biometrics 1987;43(3):671-681.
77. Breslow NE, Day NE. Indirect standardisation and multiplicative models of rates with reference to the age adjustment of cancer incidence and relative frequency data. Journal of Chronic Diseases 1975;28:289-303.
78. Lawson A, Bohning D, Biggeri A, Lesaffre E, Viel J-F. Disease Mapping and Its Uses. In: Disease Mapping and Risk Assessment for Public Health: John Wiley & Sons Ltd.; 1999. p. 3-13.
79. Mollie A. Bayesian and Empirical Bayes Approaches to Disease Mapping. In: Lawson A, editor. Disease Mapping and Risk Assessment for Public Health: John Wiley & Sons Ltd; 1999. p. 15-29.
80. Bohning D, Schlattmann P. Disease Mapping with Hidden Structures Using Mixture Models. In: Lawson A, editor. Disease Mapping and Risk Assessment for Public Health: John Wiley & Sons Ltd; 1999. p. 49-60.
81. Tsutakawa RK. Estimation of Cancer Mortality-Rates - a Bayesian-Analysis of Small Frequencies. Biometrics 1985;41(1):69-79.
82. McCullagh P, Nelder J. Generalized Linear Models. London: Chapman & Hall; 1989.
83. Mollie A. Bayesian Mapping of Disease. In: Gilks W, Richardson S, Spiegelhalter DJ, editors. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996. p. 259-379.

84. Clayton D, Bernardinelli L. Bayesian Methods for mapping disease risk. In: Elliott P, Cuzick J, English D, Stern R, editors. Geographical and environmental epidemiology : methods for small-area studies: Oxford University Press; 1992. p. 205-220.

85. Bernardinelli L, Montomoli C. Empirical Bayes Versus Fully Bayesian-Analysis of Geographical Variation in Disease Risk. Statistics in Medicine 1992;11(8):983-1007.

86. Marshall RJ. A Review of Methods for the Statistical-Analysis of Spatial Patterns of Disease. Journal of the Royal Statistical Society Series a-Statistics in Society 1991;154:421-441.

87. Efron B, Morris C. Data analysis using Stein's estimation and its generalisation. Journal of the American Statistical Association 1975;70:311-319.

88. Tsutakawa RK, Shoop GL, Marienfeld CJ. Empirical Bayes Estimation of Cancer Mortality-Rates. Statistics in Medicine 1985;4(2):201-212.

89. Leonard T. Bayesian methods for binomial data. Biometrika 1972;59:581-589.

90. Lawson AB. Disease map reconstruction. Statistics in Medicine 2001;20(14):2183-2204.

91. Tsutakawa RK. Mixed Model for Analyzing Geographic Variability in Mortality- Rates. Journal of the American Statistical Association 1988;83(401):37-42.

92. Manton KG, Woodbury MA, Stallard E, Riggan WB, Creason JP, Pellom AC. Empirical Bayes Procedures for Stabilizing Maps of United- States Cancer Mortality-Rates. Journal of the American Statistical Association 1989;84(407):637-650.

93. Heisterkamp SH, Doornbos G, Gankema M. Disease Mapping Using Empirical Bayes and Bayes Methods on Mortality Statistics in the Netherlands. Statistics in Medicine 1993;12(19-20):1895-1913.

94. Langford IH. Using Empirical Bayes Estimates in the Geographical Analysis of Disease Risk. Area 1994;26(2):142-149.

95. Cislaghi C, Biggeri A, Braga M, Lagazio C, Marchi M. Exploratory Tools for Disease Mapping in Geographical Epidemiology. Statistics in Medicine 1995;14(21-22):2363-2381.
96. Yasui Y, Liu H, Benach J, Winget M. An empirical evaluation of various priors in the empirical Bayes estimation of small area disease risks. Statistics in Medicine 2000;19(17-18):2409-2420.
97. Manton KG, Stallard E, Woodbury MA, Riggan WB, Creason JP, Mason TJ. Statistically Adjusted Estimates of Geographic Mortality Profiles. Journal of the National Cancer Institute 1987;78(5):805-815.
98. Marshall RJ. Mapping Disease and Mortality-Rates Using Empirical Bayes Estimators. Applied Statistics-Journal of the Royal Statistical Society Series C 1991;40(2):283-294.
99. Maiti T. Hierarchical Bayes estimation of mortality rates for disease mapping. Journal of Statistical Planning and Inference 1998;69(2):339-348.
100. Laird N. Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association 1978;73:805-811.
101. Schlattmann P, Bohning D. Mixture-Models and Disease Mapping. Statistics in Medicine 1993;12(19-20):1943-1950.
102. Leyland AH, Langford IH, Rasbash J, Goldstein H. Multivariate spatial models for event data. Statistics in Medicine 2000;19(17-18):2469-2478.
103. Leyland AH. Spatial Analysis. In: Leyland AH, Goldstein H, editors. Multilevel Modelling of Health Statistics: John Wiley & Sons, Ltd; 2001.
104. Goldstein H. Multilevel Statistical Models. 3rd ed. London: Arnold; 2003.
105. Langford IH, Day RD. Poisson Regression. In: Leyland AH, Goldstein H, editors. Multilevel Modelling of Health Statistics: John Wiley & Sons, Ltd; 2001. p. 45-57.
106. Langford IH, Bentham G, McDonald A-L. Multi-level modelling of geographically aggregated health data: a case study on malignant melanoma mortality and UV exposure in the European community. Statistics in Medicine 1998;17:41-57.

107. Meza JL. Empirical Bayes estimation smoothing of relative risks in disease mapping. Journal of Statistical Planning and Inference 2003;112(1-2):43-62.
108. Besag J, Mollie A. Bayesian mapping of mortality rates. Bulletin of the International Statistical Institute, 47th session 1989;1:127-128.
109. Mollie A, Richardson S. Empirical Bayes Estimates of Cancer Mortality-Rates Using Spatial Models. Statistics in Medicine 1991;10(1):95-112.
110. Besag J, York J, Mollie A. Bayesian Image-Restoration, with 2 Applications in Spatial Statistics. Annals of the Institute of Statistical Mathematics 1991;43(1):1-20.
111. Besag J. Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B 1974;36:192-236.
112. Lawson AB, Biggeri AB, Boehning D, Lesaffre E, Viel JF, Clark A, et al. Disease mapping models: an empirical evaluation. Statistics in Medicine 2000;19(17-18):2217-2241.
113. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Statistical Methods in Medical Research 2005;14(1):35-59.
114. Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, et al. Statistical issues in the analysis of disease mapping data. Statistics in Medicine 2000;19(17-18):2493-2519.
115. Wakefield JC, Best NG, Waller LA. Bayesian approaches to disease mapping. In: Elliott P, Wakefield JC, Best NG, Briggs DJ, editors. Spatial Epidemiology: Methods and Applications. Oxford: Oxford University Press; 2000. p. 104-127.
116. Ripley BD. Spatial Statistics. New York: John Wiley & Sons; 1981.
117. Kelsall J, Wakefield J. Modeling spatial variation in disease risk: A geostatistical approach. Journal of the American Statistical Association 2002;97(459):692-701.
118. MacNab YC. Hierarchical Bayesian modeling of spatially correlated health service outcome and utilization rates. Biometrics 2003;59(2):305-316.

119. Lawson A, Clark A. Spatial mixture relative risk models applied to disease mapping - Authors' reply. Statistics in Medicine 2003;22(7):1203-1203.
120. Green PJ, Richardson S. Hidden Markov models and disease mapping. Journal of the American Statistical Association 2002;97(460):1055-1070.
121. Fernandez C, Green PJ. Modelling spatially correlated data via mixtures: a Bayesian approach. Journal of the Royal Statistical Society Series B-Statistical Methodology 2002;64:805-826.
122. Goldstein H. Multilevel Statistical Models. London: Edward Arnold; 1995.
123. Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death. Ninth revision: WHO: Geneva; 1997.
124. Goldstein H. Multilevel Mixed Linear-Model Analysis Using Iterative Generalized Least-Squares. Biometrika 1986;73(1):43-56.
125. Goldstein H, Rasbash J. Efficient computational procedures for the estimation of parameters in multilevel models based on iterative generalised least squares. Computational Statistics and Data Analysis 1992;13:63-71.
126. Browne WJ, Draper D. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Nottingham Statistics Research Reports; 2004 January.
127. Goldstein H. Restricted Unbiased Iterative Generalised Least Squares Estimation. Biometrika 1989;76:622-623.
128. Besag J, Newell J. The Detection of Clusters in Rare Diseases. Journal of the Royal Statistical Society Series a-Statistics in Society 1991;154:143-155.
129. Langford IH, Leyland AH, Rasbash J, Goldstein H, Day RD, McDonald A-L. Multilevel Modelling of Area-Based Health Data. In: Lawson A, Biggeri A, Bohning D, Lesaffre E, Viel J-F, Bertollini R, editors. Disease Mapping and Risk Assesment for Public Health: John Wiley & Sons Ltd; 1999. p. 218-228.
130. Gilks W, Richardson S, Spiegelhalter DJ. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996.
131. Ripley BD. Stochastic simulation. New York, USA: Wiley; 1987.

132. Gilks WR, Wild P. Adaptive Rejection Sampling for Gibbs Sampling. Applied Statistics-Journal of the Royal Statistical Society Series C 1992;41(2):337-348.
133. Gelman A, Rubin DB. A single sequence from the Gibb sampler gives a false sense of security. In: Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors. Bayesian Statistics. Oxford: Oxford University Press; 1992.
134. Gelman A. Inference and monitoring convergence. In: Gilks W, Richardson S, Spiegelhalter DJ, editors. Markov Chain Monte Carlo in Practive. London: Chapman & Hall; 1996. p. 131-140.
135. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science 1992;7:457-511.
136. Spiegelhalter DJ, Thomas A, Best N, Lunn D. WinBUGS User Manual. Cambridge; 2003.
137. Gelfand AE, Smith AFM. Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 1990;85(410):398-409.
138. Rasbash J, Goldstein H. Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. Journal of Behavioural Statistics 1994;19:337-350.
139. Browne W, Goldstein H, Rasbash J. Multiple membership multiple classification (MMMC) models. Statistical Modelling 2001;1(2).
140. Hill PW, Goldstein H. Multilevel modeling of educational data with cross-classification and missing identification for units. Journal of Educational and Behavioral Statistics 1998;23(2):117-128.
141. Harville D. Maximum Likelihood appraoches to variance components estimation and to related problems. Journal of the American Statistical Association 1977;72:320-340.
142. Nelder J, Wedderburn R. Generalized linear models. Journal of the Royal Statistical Society Series A 1972;135:370-384.

143. Rasbash J, Browne WJ, Goldstein H, Yang M, Plewis I, Healy M, et al. A User's Guide to MLwiN, Version 2.1,. London: Institute of Education, University of London; 2000.
144. Spiegelhalter DJ, Best N, Gilks W, Inskip H. Hepatitis B: a case study in MCMC methods. In: Gilks W, Richardson S, Spiegelhalter DJ, editors. Markov Chain Monte Carlo in Practice. London: Chapman & Hall; 1996.
145. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society Series B-Statistical Methodology 2002;64:583-616.
146. Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN 2000: Cancer Incidence, Mortality and Prevalence Worldwide, Version 1.0. In. Lyon: IARC CancerBase No. 5; 2001.
147. World Population Prospects. The 1998 Revision Volume I: Comprehensive Tables: UN DEPT OF ECON & SOCIAL AFF; 1999.
148. Levi F, Lucchini F, La Vecchia C, Negri E. Trends in mortality from cancer in the European Union, 1955-94. Lancet 1999;354(9180):742-743.
149. Levi F, Lucchini F, Negri E, Boyle P, La Vecchia C. Cancer mortality in Europe, 1995-1999, and an overview of trends since 1960. International Journal of Cancer 2004;110(2):155-169.
150. Brennan P, Bray I. Recent trends and future directions for lung cancer mortality in Europe. British Journal of Cancer 2002;87(1):43-48.
151. Bray I. Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. Journal of the Royal Statistical Society Series C-Applied Statistics 2002;51:151-164.
152. Levi F, Lucchini F, Negri E, Boyle P, La Vecchia C. Cancer mortality in Europe, 1990-1994, and an overview of trends from 1955 to 1994. European Journal of Cancer 1999;35(10):1477-1516.
153. Levi F, Lavecchia C, Lucchini F, Negri E. Cancer Mortality in Europe, 1990-92. European Journal of Cancer Prevention 1995;4(5):389-417.
154. Humphrey LL, Teutsch S, Johnson M. Lung Cancer Screening with Sputum Cytologic Examination, Chest Radiography, and Computed Tomography:

An Update for the U.S. Preventive Services Task Force. Ann Intern Med 2004;140(9):740-753.

155. U.S. Preventive Services Task Force*. Lung Cancer Screening: Recommendation Statement. Ann Intern Med 2004;140(9):738-739.
156. IARC. IARC Monographs Programme; 1972-2001.
157. Bailey-Wilson JE, Amos CI, Pinney SM, Petersen GM, de Andrade M, Wiest JS, et al. A major lung cancer susceptibility locus maps to chromosome 6q23-25. American Journal of Human Genetics 2004;75(3):460-474.
158. Cavalli-Sforza LL, Menozzi P, Piazza A. The History and Geography of Human Genes. Chichester: Princeton University Press; 1994.
159. Berrino F, Capocaccia R, Coleman MP, Esteve J, Gatta G, Hakulinen T, et al. Survival of cancer patients in Europe: the EUROCARE-3 study. Annals of Oncology 2003;14(Suppl 5).
160. Coleman MP, Gatta G, Verdecchia A, Esteve J, Sant M, Storm H, et al. EUROCARE-3 summary: cancer survival in Europe at the end of the 20th century. Annals of Oncology 2003;14(Supp 5):v128-v129.
161. Coebergh JWW. Colorectal cancer screening in Europe: first things first. European Journal of Cancer 2004;40(5):638-642.
162. Winawer SJ, Zauber AG. Colorectal cancer screening: Now is the time. Canadian Medical Association Journal 2000;163(5):543-544.
163. Levin B. Colorectal Cancer Prevention and Early Detection. In: C G, Willett., editor. Cancer of the Lower Gastrointestinal Tract: American Cancer Society and BC Decker, Inc; 2001. p. 45-52.
164. Inadomi JM, Sonnenberg A. The impact of colorectal cancer screening on life expectancy. Gastrointestinal Endoscopy 2000;51(5):517-523.
165. Gazelle GS, McMahon PM, Scholz FJ. Screening for colorectal cancer. Radiology 2000;215(2):327-335.
166. U of M study confirms importance of biennial colorectal cancer screening. Minnesota; 2000.

167. Screening for Colorectal Cancer Saves Lives: Centers for Disease Control and Prevention; 2004.
168. Centers for Disease Control. Screening for colorectal cancer - United States, 1992-1003, and new guidelines. MMWR 1996;45:106-10.
169. Genetics of Colorectal Cancer. In: National Cancer Institute, U.S. National Institutes of Health; 2004.
170. Faivre J, Forman D, Esteve J, Gatta G. Survival of patients with oesophageal and gastric cancers in Europe. European Journal of Cancer 1998;34(14):2167-2175.
171. Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. International Journal of Cancer 1999;80(6):827-841.
172. Parkin DM, Whelan SL, Ferlay J, Raymond L, Young J. Cancer Incidence in Five Continents. Lyon: International Agency for Research on Cancer; 1997.
173. ACOR. Screening for Esophageal Cancer. In. New York: http://www.acor.org/cnet/62877.html; 2004.
174. McGilchrist A, Park KGM. Risk Factors and Delays in Presentation in Scottish Audit of Gastric and Oesophageal Cancer: NHS; 2000.
175. Enzinger PC, Mayer RJ. Medical progress - Esophageal cancer. New England Journal of Medicine 2003;349(23):2241-2252.
176. Risk JM, Mills HS, Garde J, Dunn JR, Evans KE, Hollstein M, et al. The tylosis esophageal cancer (TOC) locus: more than just a familial cancer gene. Diseases of the Esophagus 1999;12(3):173-176.
177. Antunes JLF, Toporcov TN, de Andrade FR. Trends and patterns of cancer mortality in European countries. European Journal of Cancer Prevention 2003;12(5):367-372.
178. Black RJ, Bray F, Ferlay J, Parkin DM. Cancer incidence and mortality in the European Union: Cancer registry data and estimates of national incidence for 1990. European Journal of Cancer 1997;33(7):1075-1107.
179. Boyle P, d'Onofrio A, Maisonneuve P, Severi G, Robertson C, Tubiana M, et al. Measuring progress against cancer in Europe: has the 15% decline targeted for 2000 come about? Annals of Oncology 2003;14(8):1312-1325.

180. La Vecchia C, Franceschi S, Levi F. Epidemiological research on cancer with a focus on Europe. European Journal of Cancer Prevention 2003;12(1):5-14.

181. Rosenberg MS, Sokal RR, Oden NL, DiGiovanni D. Spatial autocorrelation of cancer in western Europe. European Journal of Epidemiology 1999;15(1):15-22.

182. Vartiainen E, Puska P, Jousilahti P, Korhonen HJ, Tuomilehto J, Nissinen A. 20-Year Trends in Coronary Risk-Factors in North Karelia and in Other Areas of Finland. International Journal of Epidemiology 1994;23(3):495-504.