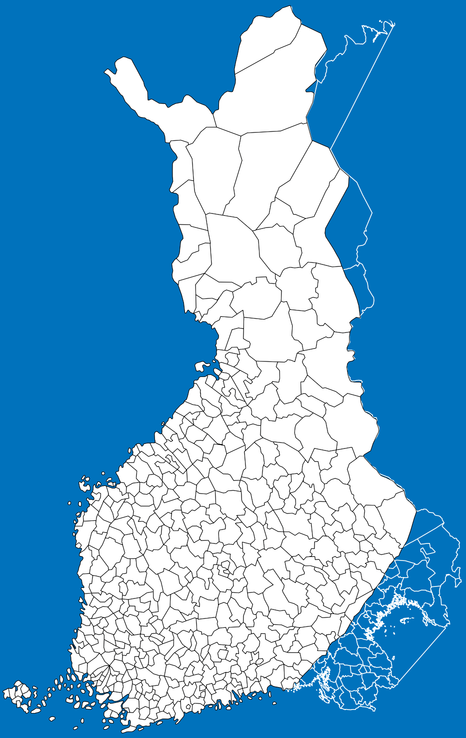
Dept. of Biomedical Engineering and Computational Science

# Bayesian Spatial and Temporal Epidemiology

## of Non-communicable Diseases and Mortality

Aki S. Havulinna



Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
**DOCTORAL DISSERTATIONS** 138/2011

# Bayesian Spatial and Temporal Epidemiology of Non-communicable Diseases and Mortality

**Aki S. Havulinna**

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium G at the Aalto University School of Science (Espoo, Finland) on the 8th of December 2011 at 12 noon.

**Aalto University**
**School of Science**
**Dept. of Biomedical Engineering and Computational Science**

**Supervisors**
Professor Jouko Lampinen
Aalto University, Espoo, Finland

**Instructors**
Docent Aki Vehtari
Aalto University, Espoo, Finland
Professor Veikko Salomaa
National Institute for Health and Welfare, Helsinki, Finland
Docent Marjatta Karvonen
National Public Health Institute, Helsinki, Finland

**Preliminary examiners**
Professor Antti Penttinen
University of Jyväskylä, Jyväskylä, Finland
Professor Seppo Koskinen
National Institute for Health and Welfare, Helsinki, Finland

**Opponents**
Professor Elja Arjas
University of Helsinki, Helsinki, Finland

Aalto University publication series
**DOCTORAL DISSERTATIONS** 138/2011

ISBN 978-952-60-4427-9 (printed)
ISBN 978-952-60-4428-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2011

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

NORDIC ECOLABEL
441 697
Printed matter

**A! Aalto University**

## Abstract

**Aalto University, P.O. Box 11000, FI-00076 Aalto** www.aalto.fi

**Author**
Aki S. Havulinna

**Name of the doctoral dissertation**
Bayesian Spatial and Temporal Epidemiology of Non-communicable Diseases and Mortality

**Publisher** School of Science

**Unit** BECS - Dept. of Biomedical Engineering and Computational Science.

**Series** Aalto University publication series DOCTORAL DISSERTATIONS 138/2011

**Field of research** Computational Science

**Manuscript submitted** 25 August 2008 **Manuscript revised** 16 September 2011

**Date of the defence** 8 December 2011 **Language** English

□ **Monograph** ☒ **Article dissertation (summary + original articles)**

### Abstract

Spatial epidemiology combines spatial statistical modelling and disease epidemiology for studying geographic variation in mortality and morbidity. The effects of putative risk factors may be examined using ecological regression models. On the other hand, age-period-cohort models can be used to study the variation of mortality and morbidity through time.

Bayesian hierarchical statistical models offer a flexible framework for these studies and enable the estimation of uncertainties in the results. The models are usually estimated using computer-intensive Markov chain Monte Carlo simulations.

In this dissertation the first four publications present practical epidemiological studies on geographic variation in non-communicable diseases in Finland. In the last publication we study the long-time variation in all-cause mortality in several European countries. New statistical models are developed for these studies.

This work provides new epidemiological information on the geographic variation of acute myocardial infarctions (AMI), ischaemic stroke and parkinsonism in Finland. An extended model for studying shared and disease specific geographic variation is presented using data on AMI and ischaemic stroke incidence. Existing results on the inverse association of water hardness and AMI are refined. New models for interpolation of geochemical data with non-detected values are presented with case studies using real data. Finally, the Bayesian age-period-cohort model is extended with versatile interactions and better prediction ability. The model is then used to study long-term variation in mortality in Europe.

**Keywords** Bayesian, epidemiology, mortality, non-communicable diseases, spatial, temporal

**ISBN (printed)** 978-952-60-4427-9 **ISBN (pdf)** 978-952-60-4428-6

**ISSN-L** 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

**Location of publisher** Espoo **Location of printing** Helsinki **Year** 2011

**Pages** 212 **The dissertation can be read at** http://lib.tkk.fi/Diss/

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

**Tekijä**
Aki S. Havulinna

**Väitöskirjan nimi**
Tarttumattomien tautien ja kuolleisuuden bayesilainen spatiaali- ja temporaaliepidemiologia

**Julkaisija** Perustieteiden korkeakoulu

**Yksikkö** Lääketieteellisen tekniikan ja laskennallisen tieteen laitos

**Sarja** Aalto University publication series DOCTORAL DISSERTATIONS 138/2011

**Tutkimusala** Laskennallinen tiede

**Käsikirjoituksen pvm** 25.08.2008 **Korjatun käsikirjoituksen pvm** 16.09.2011

**Väitöspäivä** 08.12.2011 **Kieli** Englanti

□ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

**Tiivistelmä**

Spatiaaliepidemiologiassa yhdistetään spatiaalitilastotiedettä ja epidemiologisia menetelmiä sairauksien tai kuolleisuuden esiintyvyyden alueellisten erojen tutkimiseen. Potentiaalisten riskitekijöiden yhteyksiä alueellisiin eroihin voidaan tutkia ekologisella regressiolla. Toisaalta sairauksien tai kuolleisuuden ajallista vaihtelua voidaan tutkia ikä-periodi-kohortti -mallien avulla.

Bayesilaiset hierarkkiset tilastolliset mallit soveltuvat hyvin näihin tutkimuksiin. Näillä malleilla voidaan joustavasti arvioida tulosten epävarmuuksia. Mallien estimointiin käytetään usein laskentaintensiivisiä Markovin ketju Monte Carlo -simulaatiomenetelmiä.

Tämän väitöskirjan neljässä ensimmäisessä osajulkaisussa esitetään käytännön epidemiologisia tutkimuksia tarttumattomien tautien esiintyvyyden alueellisesta vaihtelusta Suomessa. Viimeisessä osajulkaisussa tutkitaan kuolleisuuden pitkän aikavälin vaihtelua useammassa eurooppalaisessa maassa. Lisäksi tutkimuksia varten kehitetään uusia tilastollisia malleja.

Väitöstyö luo uutta epidemiologista tietoa akuuttien sydäninfarktien (AMI), iskeemisten aivohalvauksien ja parkinsonismin alueellisesta vaihtelusta Suomessa. Laajennettu malli kahden taudin yhteisen ja tautikohtaisen alueellisen vaihtelun tutkimiseen esitetään käyttäen havaintoja AMI:n ja iskeemisten aivohalvausten ilmaantuvuudesta. Aiemmin julkaistusta kovan juomaveden ja AMI:n esiintyvyyden käänteisestä assosiaatiosta saadaan tarkempaa tietoa. Alle määritysrajan jääviä havaintoja sisältävän geokemiallisen aineiston interpolointiin esitetään uusia tilastollisia malleja käytännön esimerkkien pohjalta. Lopuksi bayesilaista ikä-periodi-kohortti -mallia laajennetaan joustavilla interaktiotermeillä ja parannetaan mallin ennustuskykyä. Mallilla tutkitaan pitkän aikavälin muutoksia useamman eurooppalaisen maan kuolleisuudessa.

**Avainsanat** Bayesilainen, epidemiologia, kuolleisuus, spatiaalinen, tarttumattomat taudit, temporaalinen

**ISBN (painettu)** 978-952-60-4427-9 **ISBN (pdf)** 978-952-60-4428-6

**ISSN-L** 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

**Julkaisupaikka** Espoo **Painopaikka** Helsinki **Vuosi** 2011

**Sivumäärä** 212 **Luettavissa verkossa osoitteessa** http://lib.tkk.fi/Diss/

# Preface

This work has been conducted at the Dept. of Chronic Disease Prevention at the National Institute for Health and Welfare (THL, former KTL), Helsinki, and at the Dept. of Biomedical Engineering and Computational Science (BECS), Aalto University (former TKK), Espoo.

As I came to KTL in 2004, Docent Marjatta Karvonen introduced me to the field of medical geography. Encouraging me to begin my doctoral studies, she supervised Publications I,III–IV. Prof. Kirsi Virrantaus (Dept. of Surveying) was my first supervisor at TKK. I am grateful for the time and interest she took in my work.

Due to unsuccessful grant applications, I soon had to take a new direction. In 2006 I joined Prof. Veikko Salomaa's unit at KTL. He has encouraged me to continue this doctoral work as time allowed. I also started my postgraduate studies anew in 2006 at BECS, instructed by Docent Aki Vehtari and supervised by Prof. Jouko Lampinen. Aki's example has encouraged me in pursuing a deeper understanding of Bayesian statistics. The work for Publication V and for this thesis has been jointly supervised by Aki and Veikko. Veikko also supervised the work for Publication II with Marjatta.

Prof. Antti Penttinen (University of Jyväskylä) and Prof. Seppo Koskinen (THL) have done a superb job in reviewing this manuscript and providing insightful comments.

Looking back in time, discussions with a fellow student, Marko Peussa, guided my interest towards statistical modelling; the resulting master's thesis (in chemometrics) was supervised by Prof. Lauri Niinistö. Then a course at Palmenia Institute honed my skills in mathematical and statistical modelling. Dr. Jukka Sinisalo (at Kemira) got me involved with Bayesian networks. Dr. Jukka Ranta (Risk assessment unit/former EELA) finally convinced me to become a Bayesian.

Dr. Elena Moltchanova and the late Olli Taskinen made me immediately welcome at the SPAT team when I came to KTL, guiding my first steps in Bayesian modelling. Besides friendship, they both have made a major contribution on my work. Dr. Anne Kousa, first author in Publications III–IV, has had an important role in this work. Rauni Pääkkönen helped a lot with Publication II, and shared the office with me for several years. Dr. Mika Rytkönen helped me a great deal in starting my postgraduate studies and writing grant applications. Dr. Maria Nikkarinen, Dr. Pentti Tienari, Prof. Johan Eriksson, Prof. Reijo Marttila, and Dr. Kirsti Martikainen have all contributed in the publications. Prof. Juhani Hassi, Reija R., Ari V. and Pentti P. have also participated in the early phase.

My physical presence at BECS was minimal, but interactions with Jaakko R., Ville P., Jouni H., Eero P. and Jarno V. have been very helpful.

Thanks to all the people at KAHY and DIGY units for a great atmosphere to work in. Especially, I have learned a lot from the late Jorma Torppa and Vladislav M. besides enjoying the non-work discussions with them. My recent work with Olli S. and Juha K. have given me new ideas and improved my skills as a statistician. Arto P. and Mikko K. also shared the office with me for a while. Time spent with Paul, Nadja, Niina, Samuel and Marketta has provided relaxing diversions from work.

The early work was funded from Marjatta's research grants. Later, Pentti T. provided crucial support, which allowed me to continue this work. Foundations for Publication V were laid during the YYSSP'06 at the IIASA institute (Laxenburg, Austria, grant from The Academy of Finland). My work at BECS was funded by the TERANA project (part of Finn-Well/TEKES). The Finnish Foundation for Cardiovascular Research provided a grant for finishing this thesis (the original submission) and to do part of the research work which was later adapted to Publication V.

Above all, the life-long support and encouragement from my parents Katja and Pentti (he also proofread this manuscript), my siblings Heidi and Henri, and many more family members has given me strength in my good and bad times.

Helsinki, November 23, 2011,

Aki S. Havulinna

# Contents

# List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

**I** Havulinna AS, Tienari PJ, Marttila RJ, Martikainen KK, Eriksson JG, Taskinen O, Moltchanova E, Karvonen M. Geographical Variation of Medicated Parkinsonism in Finland During 1995 to 2000. *Movement Disorders*, 23:1024–1031, 2008.

**II** Havulinna AS, Pääkkönen R, Karvonen M, Salomaa V. Geographic Patterns of Incidence of Ischemic Stroke and Acute Myocardial Infarction in Finland During 1991–2003. *Annals of Epidemiology*, 18:206–213, 2008.

**III** Kousa A, Havulinna AS, Moltchanova E, Taskinen O, Nikkarinen M, Eriksson J, Karvonen M. Calcium:Magnesium Ratio in Local Groundwater and Incidence of Acute Myocardial Infarction Among Males in Rural Finland. *Environmental Health Perspectives*, 114:730–734, 2006.

**IV** Kousa A, Havulinna AS, Moltchanova E, Taskinen O, Nikkarinen M, Salomaa V, Karvonen M. Magnesium in Well Water and the Spatial Variation of Acute Myocardial Infarction Incidence in Rural Finland. *Applied Geochemistry*, 23:632–640, 2008.

**V** Havulinna AS. Bayesian Age-Period-Cohort Models with Versatile Interactions and Long-term Predictions: Mortality in Finland 1878–2060 and in Sweden 1751–2100. Submitted to *Statistics in Medicine*, 2011.

# Author's Contribution

### Publication I: "Geographical Variation of Medicated Parkinsonism in Finland During 1995 to 2000"

The author prepared and revised the manuscript, performed the analyses and participated in the design of analyses and data preparation. The second author contributed substantially to the manuscript writing. Other authors originally conceived and designed the study and acquired the data.

### Publication II: "Geographic Patterns of Incidence of Ischemic Stroke and Acute Myocardial Infarction in Finland During 1991–2003"

The author prepared and revised the manuscript, designed and performed the analyses and participated in data preparation. A part of the materials section was originally written by the second author. The other authors participated in the manusript preparation and the second and last authors participated in the original data collection.

### Publication III: "Calcium:Magnesium Ratio in Local Groundwater and Incidence of Acute Myocardial Infarction Among Males in Rural Finland"

The author performed the analyses, designed the interpolation model and participated in data preparation and study design. The author wrote the appendices and parts of the methods section and also participated in the manuscript revision. The first author prepared and revised the manuscript.

### Publication IV: "Magnesium in Well Water and the Spatial Variation of Acute Myocardial Infarction Incidence in Rural Finland"

The author performed the analyses, designed the interpolation model and participated in data preparation and study design. The author wrote appendices and parts of the methods section and also participated in the manuscript revision. The first author prepared and revised the manuscript.

### Publication V: "Bayesian Age-Period-Cohort Models with Versatile Interactions and Long-term Predictions: Mortality in Finland 1878–2060 and in Sweden 1751–2100"

This paper is completely the author's own work, except the original mortality and population data was collected and prepared by the Human Mortality Database and the population predictions have been made by Statistics Finland and Statistics Sweden.

Publications III–IV are also found in Dr. Anne Kousa's thesis.

# 1. Introduction

This thesis consists of five papers studying various aspects of applied Bayesian epidemiology using spatial and temporal smoothing models. Our original aim was to make methodological contributions also for spatiotemporal modelling (using lung cancer as an example), but this task turned out to be beyond the scope and schedule of this thesis. We have therefore made a different contribution (Publication V) for the revised version of this thesis.

The practical epidemiological studies in Publications I–IV are based on excellent nationwide public health and population registers in Finland. Publication I studies the incidence and prevalence of medicated parkinsonism. Publication II studies the shared and disease-specific geographic variation in ischaemic stroke and acute myocardial infarction (AMI) incidence. Publications III and IV study the geographically varying mineral composition of drinking water as a possible environmental risk factor in AMI incidence. Publications I–IV were all driven by practical needs to assess geographic variation in non-communicable diseases and to suggest putative environmental risk factors affecting these diseases. Their main emphasis has therefore been on the disease epidemiology.

Publication V presents methodological extensions for the Bayesian age-period-cohort (APC) model and demonstrates the utility of the extensions through the analysis of long series of total mortality in several European countries. This study is based on the carefully harmonized data from the Human Mortality Database [52]. In the future we will make further refinements to the model for assessing changes in cardiovascular disease incidence, prevalence, case-fatality and mortality. This was the original motivation for developing and studying these extensions for the APC models.

In applied statistics, the role of the actual statistical models is crucial even when the models are merely presented in the background as necessary tools for solving the study questions; it is important to be able to determine the appropriateness of the statistical tools at hand. This thesis will review the statistical models used in Publications I–V in more detail. A new interpolation model for partially censored geochemical data was briefly presented in the appendices of Publications III and IV; the model will also be described in more detail here.

We begin with a brief account on registry based studies and health geography. Non-communicable disease epidemiology will then be reviewed, focusing on ischaemic stroke, AMI and Parkinson's disease. The emphasis is on ecological modelling and environmental risk factors. Next, we take a look at mortality through time. We then introduce Bayesian statistical modelling, reviewing the relevant literature on spatial, spatiotemporal and APC modelling. After briefing the results of this thesis, we open the discussion by exploring the findings. The discussion will then touch the wider applicability of the presented methods and give an outlook of the future in spatial, spatiotemporal and APC modelling, concentrating on epidemiological applications. We conclude this thesis with a short summary of the empirical findings and statistical models which we have developed.

# 2. Health Geography and Non-communicable Disease Epidemiology

## 2.1 Registry Based Health Studies

These studies are based on registered information on disease events or status and auxiliary information, usually using at least the reference population at risk. Finland is extremely suitable for this kind of studies as we have virtually 100% nationwide registry coverage of:

- Hospitalisations (HILMO): National Institute for Health and Welfare; data starts from 1967 and the coverage was widened to cover all social institutions in 1994.

- National Causes of Death Statistics: Statistics Finland; data starts from 1969.

- Reimbursed medications: National Social Insurance Institute; data on persons entitled to reimbursed medication starts from 1964 and pharmacy data on prescribed medicine purchases starts from 1995.

- Population: National Population Register Centre; data starts from 1969.

- Additional covariate information, e.g., socioeconomic status (education, occupation, income): Statistics Finland.

Most of the registered information is available from 1969 onwards. By 1968 the unique National Social Security Insurance ID's had been issued to every Finnish citizen and permanent resident in Finland, thus enabling accurate record linkage between registers. Many registers were computerised at the same time. Some registers are available from earlier periods, mainly for specific diseases. The data collection at Finnish Cancer Registry, for example, started in 1953. Studies have shown that in some cases National Social Security Insurance ID's could be retrospectively matched with high accuracy to the earlier registers, but this is an expensive and time-consuming task. [86] There are also probabilistic methods for record linkage reconstruction. Although the theoretical basis is sound, this is naturally only the second best option [73, 200].

A central problem in registry based studies is case ascertainment. In some cases, the disease register data can be validated using independent sources (e.g., [211, 280, 150, 218]), but this is not always feasible. The issues in registry based studies in Finland are further discussed in [86, 123].

We finally note that collection of vital statistics in Finland was started already in 1749, by "Taulustolaitos" (Swedish *Tabellverket*), which later developed into Statistics Finland. The information includes births, deaths and marriages. The causes of death in the 1700's were crude observations given usually by relatives. [86] Nevertheless, regional studies could be done using this data—see, for example, the thorough study of the geographic distribution of malaria, gastroenteritis and smallpox in 234 Finnish municipalities, during 1749–1850 [309]. A historical perspective of regional mortality differentials in Finland is given in [220].

## 2.2 Geographies of Health

Several issues affect human health in the geographical setting. Each individual has her own "geography of health",[1] related to the geographic places of her everyday life. We must stress that it is the *place* that matters, not the geographic location given by coordinates. The location remains the same, but the place is under constant change. [76] The importance of this temporal aspect in human geography was first addressed in [109].

The role of place in human health and medicine was recognised already by Hippocrates (e.g., [111, 107]). Our "everyday" places are the places where we spend most of our time: home, school/workplace, outdoors, and travel between these places. These are the places that may have a positive or negative influence on our health. On the positive side, some landscapes

[1] Hence, the title is in plural form [76]

might have therapeutic effects, and good availability of public health services and leisure/recreational facilities, for instance, might induce well-being. Turning to the negative side, living near an industrial plant might affect our well-being, perhaps by some seemingly little things: the smell of air, or the fear that an accident could happen. Also fear in an unsafe neighbourhood reduces well-being. [76] A recent case-study[2] in the City of Järvenpää enabled the inhabitants (427 respondents) to indicate via an Internet questionnaire the places of positive and negative quality factors in their living environment. Most of the positive as well as negative places were within 1 km of home. However, negative places tended to cluster more. [235] It has also been noticed that the relative socioeconomic status of a person within her neighbourhood may affect health. Poor people living in a rich neighbourhood seem to cope less well than poor people living in a poor neighbourhood. One possible explanation is the persons' social stress from being constantly reminded of her poorness. [307]

In a more quantitative setting, place can be seen as a surrogate for the interaction between genetic factors, lifestyle and environment [237]. Genetic factors may pose an elevated disease risk, but usually the genetic disease expression is far from 100%. However, monogenetic mutations have 100% expression in rare mendelian disorders, of which there are several examples in the Finnish disease heritage [201]. In a general setting, genetic factors contribute to an elevated risk, which may (more profoundly) lead to disease when phenotypic (i.e. environmental and lifestyle) risk factors are unfavourable. In many cases, environmental and lifestyle risk factors themselves may be enough for the development of disease.

Cancers form one group of diseases in which there exist a multitude of modifiable lifestyle and environmental risk factors. Examples include smoking (mainly) in lung cancer, human papilloma virus mainly in cervical cancer and hepatitis B virus in liver cancer. In addition, food intake imbalance, obesity and the multitude of environmental carcinogens are risk factors for several cancers. Observed rapid changes in the incidence of several cancers cannot be attributed to genetic polymorphism, as the changes in allele frequencies require several generations. It is hypothesised that adaptation of modern lifestyle along with recent cumulation of environmental carcinogens caused by industry has elevated cancer risk in

[2] http://opus.tkk.fi/pehmogis/dokumentit/lyh_tutkrap_pehmoGIS_elinympariston_koetun_kartoittajana.pdf. Accessed September 5, 2011.

the genetically susceptible persons. [119, 14, 15]

## 2.3 Spatial Epidemiology

Spatial epidemiology concerns both describing and understanding geographical variations in health, especially in small area level. There are four types of studies [64]:

1. Disease mapping
2. Geographical association/correlation studies
3. The assessment of risk in relation to point or line source
4. Disease clustering and cluster detection.

Publications I–IV in this thesis are related to disease mapping and geographical associations.

## 2.4 Disease Mapping

Geographic mapping of diseases began as early as in the 1790's [10]. Geographers' broader interest in the analysis of disease and care started in the 1960's, forming the subdiscipline of medical geography [175]. A historical perspective of disease mapping is given in [299]. Until recently, small event numbers and data availability restricted disease mapping to rather coarse areal level aggregates. Studies in fine geographic resolution had to wait for the development of advanced statistical methods to control the inherent random noise. Development and use of those methods, however, required modern powerful computers.

An early review of the model building and spatial statistics in human geography is provided in [45]. A review of published disease atlases up to 1991 found that most of the studies did not use any kind of smoothing. The Finnish cancer atlas [224] was one of the first to show disease rates smoothed by a geographic centroid approach: weights are inversely proportional to the distance from the point being smoothed and directly proportional to the population counts. [300] The empirical Bayes smoothing method [44] set forth the use of conditional autoregressive (CAR) models. Besag, York and Mollié [28] presented the fully Bayesian convolution model, also known as the BYM model, which has become almost a *de facto* standard in the field. Although understanding the geographical phenomena and methodology remains an important part in the studies, disease mapping relies mostly on the use and development of spatial statistical methods. Disease mapping is usually conducted in terms of ecological studies (see, e.g., [64]).

## 2.5 Ecological Studies

In *ecological studies*, the analyses are done at *group level* instead of individual level. In spatial studies the aggregated groups are inhabitants of some (non-overlapping) geographic areas, represented by *areal level* data. The rich terminology reflects various viewpoints in the public health context, including health geography [76], medical geography [176, 175], geographical epidemiology [237], spatial epidemiology [64], small-area health statistics [65] and disease mapping [29].

Until recently, the geographic areas in the studies have been defined by some administrative bounds, e.g., counties, municipalities, hospital districts or postal areas. In Finland, grid (lattice) based exact population data has been available since 1970 [276]. Although some geographic studies have been conducted using the data, disease mapping studies have been done only recently, as the methods and available computer power developed. Also in the other Nordic countries, exact georeferenced data has been available for some time [276]. This enables us to perform geographic studies in high resolution and independent of any administrative boundaries.

### 2.5.1 Smoothing

As we increase the geographic resolution, the problem of *small numbers* increases. What we see on a crude map will be overwhelmed by random (Poisson) noise. One simple solution is to use Bayesian (non-spatial) shrinkage estimators, which force observations based on small numbers towards the global average. [79, 296] When there is reason to believe that the observations are spatially dependent—and usually they are—this information should be taken into account. The general solution is to perform some kind of smoothing over the map. However, ordinary image processing methods may not be powerful enough. Moreover, choices such as selecting the smoothing parameters are very subjective. These concerns have seeded the field of Bayesian disease mapping.

Whether we should use smoothing naturally depends on the question at hand. A decision maker in some small municipality might want to look at the actual crude number of disease cases (say) for reviewing health care resource allocation. However, if she were interested in (predicting) what will happen next year, there would be a high level of uncertainty because of the small numbers. In this case, her real target of interest would be prediction based on estimated underlying disease incidence rate. Also, when comparing the risk of a disease among different areas, or temporal changes, we are interested in differences in the underlying disease rates, not in the random noise.

### 2.5.2 Ecological Fallacy

The main limitation of an ecological study is its susceptibility to *ecological fallacy* [259]. When estimates are based on aggregated groups, we should not try to apply them at an individual level, as this would usually induce *ecological bias*. If we have found a region with a high incidence rate (e.g., as compared to the nearby areas), we may say that the disease risk is *on average* higher than in the nearby areas. However, the risk of a particular individual living in the high-risk area might well be much lower than the risk of a particular individual living in a low-risk area.

### 2.5.3 Modifiable Areal Unit Problem

*Modifiable areal unit problem* is a concept related to ecological fallacy. The choice of which way to divide an area into aggregate regions is not unique, and the analytical results of a study may depend on this choice. There are two components of the modifiable areal unit problem: the scale and zonation effects. The scale effect is attributed to variation in numerical results owing strictly to the number of areal units used in the study. For example, the choice of the resolution in a grid based map leads to scale effects. Zonation effects are attributed to the manner in which smaller areal units are grouped together to form larger units. Means and variances are resistant to these effects, whereas regression coefficients and correlation statistics exhibit dramatic changes. In disease mapping, the scale effect is apparent as the between area–variation becomes larger in smaller scales. Choosing a larger aggregation level, the variation is smaller, but important information might be lost. [3, 195]

### 2.5.4 Spatiotemporal Processes and Latency

There are two issues related to space and time in health geography. First, we note that the latency from exposure to disease occurrence might be quite long. In extreme cases, the accumulated life-time exposure to risk factors is relevant. During the latency period, people might have moved to another location—and even if not, the place itself could have changed, as noted above. Yet, almost all disease mapping studies consider the place of residence at the time of event and the place and time of exposure (almost) equivalent. [257]

The second concern is that geographic data is seldom purely spatial [45], and neither are epidemiological phenomena [257]. We may first look into this by considering the differences between disease *clusters* and *clustering* [27, 297]. Clusters refer to compact areas where there is an excess of disease cases. If specific clusters are detected, possible environmental associations could be investigated. The term clustering, on the other hand, refers to a general tendency of a disease to cluster, i.e., show geographic variation. If such variation is known to exist (or detected with some statistical test, e.g., [27]), this geographic variation could be mapped using the statistical disease mapping methods. These methods do not necessarily need to be restricted to using cluster models; e.g., the BYM model is commonly used. The clustering could show different characteristics in space and time—clusters may exist:

1. In the spatial dimension—this would indicate some permanent risk factors which are concentrated in certain areas.

2. In the temporal dimension—for example, an excess of AMI cases may be clustered in the cold seasons.

3. In the spatiotemporal continuum—for example, some infectious diseases could show temporary clustering in certain places.

As the above listing shows, usually there is no justification in considering disease events only in the spatial dimension. Indeed, the misuse of count data aggregated over time may lead to biases in the estimated area-specific risks [203]. Therefore, the natural framework for modelling in geography (also in general) is spatiotemporal processes. [45] In the point-referenced geographic studies, geostatistical modelling (e.g., kriging) is an often used technique. [58, 184] Some authors have applied geostatistical and point process methods in disease mapping using areal level data (e.g., [133, 16, 288]). There also exist spatiotemporal point process models in disease mapping, e.g., [33] However, usually the spatiotemporal models are constructed for areal level data. Spatiotemporal models and issues will be considered in the review of statistical methods (Chapter 4).

## 2.6 Environmental Risk Factors

By environmental disease risk factors we refer to any factors that are shared by people living in a common environment. Studies could include, for example, ground water [154], dietary intake [161], air pollution [137], bacterial/viral infections [303], soil [69], climate/weather [101], urban/rural environment [126], or exposure to animals at a farm [157]. However, recently it has become clear that Finns differ in genetic inheritance, being mainly separated by the so-called east/west gradient [160]. This will add a genetic flavour into the geographic epidemiology in Finland.

### 2.6.1 Spatial Ecological Correlation Studies

In spatial correlation studies (e.g., [239]) we are interested in determining whether some risk factor is associated (i.e., correlates) with the spatial (geographic) variation of a disease. Because of the ecological modelling framework, this can only be seen as an explorative semi-quantitative method. The results may provide important clues for associations that may be worth more rigorous studies, for example in a case-control setup. As is often stressed, ecological studies cannot reveal any causal relationships, even when the effects of potential confounders are controlled for. We must also bear in mind the ecological fallacy: any results apply on average, in the aggregated group level and are not generalisable to apply in the individual level. Hence, as the link between exposure and effect is not assessed directly, this ’incompleteness’ of the study generally leads to *ecological bias*. A technical introduction to the subject is given in the review of statistical methods (Chapter 4). Also, because of the modifiable areal unit problem, the choice of the areal aggregation might have a dramatic effect on the results.

Another issue is the problem with spatially misaligned data. Epidemiological data is usually available at an aggregated level, in part because of confidentiality reasons. Environmental data, however, is usually available in an accurate point level. How to realign these *spatially misaligned* data sets has been a subject of several studies [30, 31, 192, 87, 88]. This problem also concerns Publications III–IV in this thesis, where one possible solution is presented.

### 2.6.2 Point or Line Sources and Spatial Clustering

Ever since John Snow's success in identifying the source of a cholera epidemy in London[3] [34, 71], there has been interest in studying the association of a point source with an excess number of disease cases. For example, the Small-Area Health Statistics Unit at Imperial College London was originally established for studying disease risk from point sources [65]. As line sources we refer to roads with heavy traffic etc.

Examples among the studied sources are poor air quality in cities [258], asbestos mine [147, 146], polluted river [293], a nuclear power plant [273], an oil refinery plant [216], and magnetic fields from high voltage power transmission line [285, 294]. Significant associations were found with air quality and mortality [258] and with cancer and polluted river [293]. Magnetic fields showed some associations with multiple myeloma in men and colorectal cancer in women. The former could be a real association, but the study had multiple testing issues. Also, in the case of multiple myeloma, potential exposure confounders could not be controlled for [294]. In the later study of lung cancers near asbestos mine [146] distance related change points were found in the disease risk, but the study concentrated on methodology. Hence, the epidemiological results could not be further evaluated.

On the other hand, the media had raised uncritical concerns on elevated risk of childhood leukaemia around the Sellafield nuclear power plant, even trying to publish "scientific" studies. The study [273] assessed the available information and concluded that there was no evidence for the elevated risk. In cases like this, the media itself should be more critical

[3] Snow identified a pattern concentrating at a particular public water well, later mapping the dwelling locations of cholera cases—and in that map, he created (one of) the first Voronoi diagram

before raising public concerns. Also, the scientific community should be always careful in risk communication.

As we can conclude from above, the results have been negative, or inconclusive in many cases. Detecting any small excess would require a vast number of events. However, a negative finding in a study of possible risk sources should be considered a positive thing. At the same time, these studies have driven the development of advanced statistical methods. The design and methodological issues in ecological small-area studies have been recently discussed in [63, 11]. Recently, it has been suggested that ecological bias could be avoided with careful study designs, but even if this is the case, care should be taken in interpreting the results [296].

## 2.7 Epidemiology (Geographic) of Non-communicable Diseases

*Non-communicable* diseases refer to diseases which are not transmitted from one person to another. *Acute* diseases, e.g., acute myocardial infarction (AMI) and sudden stroke have an abrupt start. They may last a few days and then settle, or lead to a chronic condition or death. *Chronic* diseases, such as coronary heart disease (CHD) or diabetes may have an impact for the rest of life once they have emerged. [76] Chronic disease onset may be quite slow as with the development of Parkinson's disease (PD) [304]. One point we should consider is the fact that the population is constantly aging. Thereby the number of patients with chronic diseases is expected to grow dramatically in the next few decades. [171]

The risk factors vary from disease to disease, but some factors are common to several diseases. One example is the cluster of most dangerous heart attack risk factors, known as the *metabolic syndrome* (METS). METS drives the global twin epidemy of cardiovascular disease (CVD) and type 2 diabetes (T2DM). According to the International Diabetes Federation definition, a person is defined as having METS if she has central obesity plus any two of the following: raised triglycerides, reduced HDL cholesterol, raised blood pressure or raised fasting plasma glucose. [56] Note, however, that there are several other definitions of METS, e.g., the American National Cholesterol Education Program–Adult Treatment Panel (NCEP-ATP-III) definition and the WHO definition. In the International Diabetes Federation definition central obesity is considered a necessary trait, whereas in the other definitions it is considered as equal among the other risk factors. The different definitions may lead to slightly different diagnoses. [217]

The classification into non-communicable and *communicable* (i.e., infectious, transmissible) diseases is not always clear-cut. So-called microbial cancers (cervical, liver and stomach cancers) [256] make a notable example. Virtually all cervical cancer cases result from persistent genital infection with highly trasmissible human papilloma virus [49]. Transmissible *Helicobacter pylori* is associated with stomach cancer [139]. Chronic infection with hepatitis B virus is the most important risk factor for liver cancer [119, 256]. Note, however, that cancer itself is never infectious. The role of bacterial or viral infections is suspected also in many other non-communicable diseases (see below for some examples).

In the following, we review *geographic* studies on some of the major non-communicable diseases in Finland.

### 2.7.1 Cardiovascular Diseases

The geographic variation in CVD *mortality* in Finland has been known since the late 1940's [124, 151, 202, 286]. Until the 1900's, infectious diseases were a common cause of death, and the remoteness of rural areas was an asset in avoiding transmission, which was reflected in lower mortality rates. During the first part of the 1900's the differences in wealth determined the regional differences in mortality. [220] Thereafter, the regional east/west gradient in mortality rates has remained practically the same ever since the 1930's, and the higher mortality in eastern Finland is mainly attributable to CVD. Although mortality has received most attention, the east/west relative difference in CHD incidence and prevalence has also been noticed since the 1970's, starting from the "Seven Countries" study [128]. More specifically, the geographic variation in the *incidence* of AMI follows an east/west pattern [126] which is similar to that of CHD mortality. It has also been noted already in the 1960's that there were regional differences in the duration of pregnancy and the weight and length of the newborn, with a similar east/west gradient [278]. The east/west mortality differences cannot be explained by differences in demographic and socioeconomic composition of the regional population [151, 287].

Incidence and mortality rates of CVD have been constantly decreasing nationwide [252, 251, 213, 282, 265, 212, 168, 194]. In part, this favourable trend reflects the success of nationwide prevention programmes (e.g., [289]). However, changes in classic risk factors no longer explain time trends in CVD mortality [98]. Moreover, the east/west gradient in

CVD incidence and mortality has remained despite the decreasing trend at the nationwide level [126]. An east/west gradient in CVD mortality also prevails between the European countries [183]. Geographic variation in stroke incidence or mortality has not been previously studied in Finland.

Classic risk factors of CVD include: male sex, smoking, diabetes, hypertension and high LDL cholesterol. In the multinational cross-sectional INTER-HEART study, abnormal lipids, smoking, hypertension, diabetes, abdominal obesity, psychosocial factors, consumption of fruits, vegetables, and alcohol, and regular physical activity accounted for most of the differences in the risk of myocardial infarction worldwide in both sexes and at all ages in all regions [314]. The geographic distribution of cholesterol, obesity and some dietary habits in Finland has also been studied [262], but the association with geographic variation in CVD was not assessed. The role of drinking water constituents in CVD risk has been subject to several studies in Finland (see, e.g., [154] and references therein), and also in other countries (e.g., [189, 311, 191, 41]). The results generally suggest that low water hardness, especially low magnesium (Mg) concentration, is associated with increased CVD risk. However, there seems to be no consensus on the subject as of yet. Associations of viral and bacterial infections with CVD have been studied [303, 228]. There are gender differences in the presentation and clinical course of many cardiovascular disorders [312], and also generally in health in later life [6], which suggest that the geographic variation in disease risks should be evaluated genderwise. The geographic variation in CVD risk has not been studied specifically in women in Finland.

This thesis studies CVD with ischaemic (i.e. atherothrombotic) aetiology. CVD originating with inflammation or infection is therefore excluded. In addition, periferal artery disease (claudicatio intermittens) is excluded although it has an ischaemic origin. CVD can be further divided into heart diseases and cerebrovascular diseases. The diagnostic practices in this thesis are broadly based on the experience of the World Health Organization (WHO) MONICA (Multinational MONItoring of trends and determinants in CArdiovascular disease) project[4] [281] and the Finnish myocardial infarction and stroke registers which participated in the MONICA project [252, 282]. The more recent FINAMI [251] and FINSTROKE [265] registers have followed and updated the diagnostic classification of the MONICA registers. In the Finnish national health care registers the

[4] http://www.ktl.fi/monica/. Accessed September 13, 2011.

diagnoses were coded using *International Classification of Diseases, Ninth Revision* (ICD-9) until the beginning of 1996, when *International Classification of Diseases, Tenth Revision* (ICD-10) was adopted.

As a short description (in ICD-10), ischaemic heart disease includes the codes I20–I25. Of those, I21–I22 denote AMI. I20.0 denotes unstable angina pectoris. Clinically, AMI and unstable angina are often considered together as acute coronary syndrome (ACS). Cerebrovascular diseases include: subarachnoid haemorrhage (I60), intracerebral haemorrhage (I61) and cerebral infarction (I63). Stroke, not specified as haemorrhage or infarction (I64), is seldom used as most of the strokes can nowadays be classified. In the epidemiological research practice, I63 and I64 are together considered as ischaemic stroke. The exact CVD classifications used in this thesis are given in the Materials and Methods (Chapter 6).

### 2.7.2 Parkinson's Disease

Parkinson's disease (PD) is a chronic slowly progressing neurodegenerative disease with a multifactorial aetiology. Its prevalence increases with age and it is slightly more common in men than in women. [304] In the central European population, the estimated prevalence is 1.6% in population over 65 years of age [53]. Studies have been conducted for assessing the role of some environmental/occupational risk factors in PD (e.g., [157, 57]). Coffee drinking and smoking seem to be associated with lower risk [115]. Living in a rural area, exposure to pesticides or drinking well water are suggested as risk factors [157]. Other suggested risk factors include high body mass index (BMI) [117] and high total cholesterol [114]. T2DM was also suggested as a strong risk factor [116]. However, many results seem to be still controversial (see, e.g., [263]). Genetic factors have a role in PD, especially in early-onset PD (at <45 years). PD seems to be a heterogeneous disease with considerable genetic background and gene-environment interactions [12]. Some interactions have recently been suggested: Apolipoprotein E polymorphism with coffee drinking and $\alpha$-synuclein Rep1 polymorphism with smoking [173]. A report [136] showed a map of PD prevalence in Finland.

### 2.7.3 Cancers

Recent reviews of cancers in Finland are given in [226, 37]. There are numerous studies[5] on cancer occurrence, including spatial disease mapping. The Finnish Cancer Registry was established in 1952, the data collection started in 1953 and reporting all new cases of cancers has been mandatory since 1961. [270] In Finland, tumours are the second most common cause of death (23% of all deaths in 2009) after cardiovascular diseases.[6] This proportion is similar in both sexes, but cancer sites are sex specific. In men, prostate cancer is most common (5322 incident cases; 38% of all new cases in 2005), whereas breast cancer is most common in women (4021 incident cases; 32% of all new cases in 2005). [37] In absolute numbers, the yearly incidence of cancers has almost doubled from 1960 to 2005, both in men and women. In 2005, there were 14,046 new cancer cases in men and 12,415 cases in women. Much of this change is attributable to ageing population. Hence, the age-adjusted cancer incidence rate has not changed much until some recent increases around the 1990's. At the same time, age-adjusted cancer mortality has been decreasing. [37, 226]

First cancer maps in Finland were made in the late 1950s, presenting crude cancer rates in the municipalities. Later, in the 1970's cancer maps were published concerning larger regions, e.g., counties. Finally, in 1987 the smoothing method originally developed at the Geological Survey Finland (GTK) [94] was used for presenting an atlas of smoothed cancer rates in the municipality level. [223, 224]. Generally speaking, no associations of cancer and ground water minerals have been found [219]. Later cancer studies in Finland suggest some associations with environmental risk factors; see, e.g., [294, 293]. Later, atlases of cancers in the Nordic countries have been published [227, 225]. The same method has also been used for cancer maps in other countries (e.g., [260]). Unfortunately, the choice of colours in several published iso/choropleth maps (in Finland and elsewhere) is not appropriate for black and white copies; see, e.g., [223, 39]. Breast cancer in Finland has been studied in a fine grid. The geographic differences were associated with mutations in BRCA1 and BRCA2 genes. [221].

[5] For a list of references, see: http://stats.cancerregistry.fi/Publications/publications.html

[6] Statistics Finland; http://www.stat.fi/tup/suoluk/suoluk_terveys_en.html#death Both sites accessed September 13, 2011.

### 2.7.4 Diabetes

In Finland, the research of insulin dependent diabetes mellitus (Type 1 diabetes mellitus, T1DM) has been very active. The geographic variation of incidence has been mapped. During 1987–1996 the high risk areas of incident T1DM among children under 15 years of age seemed to form a belt over the central Finland. There was a slight male excess in the incidence. [234, 247, 249]. An earlier study [127] mapped the same data with a simpler Bayesian method (shrinkage towards the global mean). The relative risk (RR) in a high aggregation level (20 functional areas) seemed to exhibit the above mentioned belt pattern, but at lower aggregation levels the pattern was lost. The incidence was not associated with zink and nitrate in ground water or urban/rural status [187]. The annual decreasing trend in age of onset of T1DM was also studied in an incomplete birth cohort design, using a Bayesian spatial smoothing model. No decreasing trend was found within the birth cohorts and hence, it was concluded that the decreasing trend is mostly due to steady increasing trend in the cumulative birth cohort incidence [186].

T2DM has not been studied with disease mapping methods, but there is some evidence for regional variation in Finland. The results from three areas in Finland in men and women aged between 45 and 64 years indicate that in women the prevalence (adjusted for age and BMI) was lowest in eastern Finland (Kuopio and North Karelia), higher in southwestern Finland (Turku-Loimaa), and highest in the Helsinki-Vantaa region (p=0.003). The results in men were just the opposite but not statistically significant (p=0.52). On average, the prevalence was 10% in men and 7 % in women. [313] These results indicate that a disease mapping study could provide important new information in the regional differences of T2DM. However, the high prevalence indicates that it should be taken into account when estimating risk population counts for mapping the incidence of T2DM. This would make the estimation more difficult than in the usual disease mapping studies.

### 2.7.5 Other Diseases

Multiple sclerosis (MS) disease incidence is known to have geographic clustering in Finland: there is a large excess of MS disease cases concentrated at Seinäjoki in Southern Ostrobothnia. [274] A map of alcohol-related deaths was shown in [288]. Schizophrenia was mapped in municipality level (with no smoothing) using Finnish birth cohorts born from 1950 to 1969. There was some regional variation and significant spatial clustering of excess cases in eastern Finland. Incidence was higher in the rural areas in the oldest birth cohort, but in the younger birth cohorts incidence was higher in the urban areas. [100]

## 2.8 Spatial Epidemiology in Other Countries

Disease mapping studies and disease atlases until 1991 were reviewed in [300] and [299] updated the review until 2000. Many of the studies had concentrated on cancer mortality. Also, many Bayesian studies seem to reconsider the same few classic data sets: Scottish lip cancer [44] and Ohio lung cancer [298] using new models. Although the epidemiological novelty usually remains low, this is good for the modelling point of view, as model comparisons are easier to make. Other studies include, e.g., meningococcal disease [144], T1DM [40], MS-disease [190], Crohn's disease and ulcerative colitis [198].

# 3. Mortality Through Time

## 3.1 Collection and Analysis of Vital Statistics

The first effort to analyse vital statistics was by John Graunt in his book *Natural and Political Observations Made upon the Bills of Mortality* (1662). The London *Bills of mortality* were weekly records of causes of death which were first begun for monitoring the 1592/3 Bubonic plague outbreak. After some disuse, the monitoring was resumed at the 1603 plague outbreak, until superseded by other means of record keeping in the 19th century. [243] Although he was not the only advocate of Small Pox inoculation at the early 18th century [241], Cotton Mather's influence on noticing the efficiency of Small Pox inoculation (a practice borrowed from Africa and the Orient) in 1721 became another early example on the use of vital statistics [231].

However, it was only in the middle 1800's that the recording of vital statistics and work on public health became major issues, perhaps culminating in John Snow's classic work on the London cholera epidemic [34] which we have already mentioned.

In fact, vital statistics have been collected in some Italian cities already from the 14th century onwards. Nordic countries, however, were among the first to commence *nationwide* collection of vital statistics in the 1700's. A detailed history of the data collection is given in [85]. It is noted that the data quality in Sweden was much better in the early years than the respective data in Finland, although both data sets originate from the Swedish *Tabellverket*. [85]

Recently, harmonized nationwide time-series of vital statistics have been produced from many, mainly European countries by the Human mortality database (HMDB) [52]. The above stated lower quality of the early

Finnish data compared to the respective Swedish data was verified in a personal communication from Mila Andreeva (of HMDB).

## 3.2 Epidemiologic Transitions

The epidemiologic transition theory [208] with later refinements (e.g. [125]) views mortality as the fundamental factor in population dynamics. The theory describes demographic trends with a long-term shift in mortality and disease patterns as successive stages. We list the stages with reference to the Finnish time periods [125]:

1. 'Era of pestilence and famine'. In Finland, the last great famine was during 1866-1868. Before that, mortality was high and fluctuating, due to epidemics, famines and wars. During this stage, sustained population growth was not possible.

2. 'Era of bacteriology'. Soon after the last famine, mortality started decaying, partly due to improvements in hygiene. One important step was the work against tuberculosis. Around 1930, one Finn died of tuberculosis every hour. Mass screenings and availability of medications since the late 1940's have reduced the tuberculosis mortality to a tiny fraction of what it was 100 years earlier. [275]

3. 'Era of antibiotics'. In Finland, antibiotics were introduced for general use in public health during the late 1940's. There is some controversy regarding whether this had much impact on infectious disease mortality, which started declining already before the antibiotics were introduced. [108]

4. 'Era of delayed ageing'. After most of the infectious diseases became minor causes of death, chronic diseases took the major role in mortality. In Finland, the CVD epidemy begun in the 1950's, showing the (in)famous east/west gradient in mortality. [124] However, mortality of elderly people started decaying since the 1970's. This was not anticipated in general, nor in the seminal work on epidemiologic transitions [208].

## 3.3 Beyond Life Tables - Analysis of Mortality

John Graunt's book, cited above, was the first to present a modern life table in 1662. Some decades later (in 1693) Edmond Halley published his Breslau life table (or more strictly a population table) [95] which may be seen as the first real step in the art of life-measurement. Halley constructed an almost complete life table (population annuity by age) from observed births and deaths from a 5-year observations period. The undocumented exact methods which Halley used, including smoothing and removal of outliers, have been tried to reconstruct in [13]. Since Halley's days, several methods for measuring and predicting mortality have been presented. We do not try to make a complete survey of the methods, but merely scratch the surface here.

The age-aspect was first modelled with the Gomperz-Makeham law of mortality, and later, e.g. by the Heligman-Pollard model. [106] In the Heligman-Pollard model, the mortality curve is thought to be composed of three distinct and consecutive components.

1. *The 'infant' component* describes the rapid exponential decline in mortality during the early childhood, as the child adapts to the surrounding world, including development of the immune system.

2. *The 'accident' component* reflects the excess mortality from accidents and also the maternal mortality in women. It can be approximated as a lognormal distribution peaking around 20 years of age.

3. *The 'senescent' component* describes the gradual deterioration of the aging body. This can be modeled with the exponential Gomperz law of mortality.

A presentation of mortality on the calendar-time vs. age surface was done using the Lexis diagram in 1875. [132] Since the 1920's several publications [55, 5, 135, 72] considered (and some modelled) mortality in the age and time scales. 'Generation' effects in mortality were first considered in [55, 5], and [72] introduced the term 'cohort'. Later developments have led to so-called 'age-period-cohort' (APC) models in the 1980's. See, e.g., [113] We present the Bayesian version of the APC model in the next chapter. For completeness, we note that there are several other models for studying mortality, e.g., the Lee-Carter model [167] is quite popular.

### 3.4 Prediction of Mortality

Besides taking a look at the past mortality, predicting the future trends in mortality has become more and more important. Population aging has become an internationally important concern. As the number of elderly people is growing, while the number of youth is declining, the social and economic costs are increasing—as the old age dependency ratio increases. However, at the same time older people tend to have fewer disabilities than people at the same age had a few decades earlier, and the cognitive decline also seems to be postponed. It seems that forecasting of aging needs new thinking and new measures, as well. [255]

# 4. Review of Statistical Methods

## 4.1 Bayesian Statistical Modelling

### 4.1.1 Bayesian and Frequentist Paradigms

There are two major philosophical paradigms in statistics. Traditionally, statistics has been ruled by so-called classical *frequentist* paradigm. In the *frequency interpretation*, the probability of an *event* is the limit of its relative frequency of occurrence when the experiment is repeated a very large number of times. Probabilities are only assignable to events, in well defined random experiments. The set of all possible outcomes forms the sample space. An event forms a particular subspace of the total sample space. For an event, there are only two possibilities: either it happens or it does not happen. In practice, however, many events are unique or cannot be assigned an explicitly defined sample space. [79]

In the *Bayesian* paradigm, probability can be seen as a measure of the state of knowledge. Before any data is observed, a Bayesian statistician describes her *a priori* knowledge of a phenomenon—the degree of information—by a prior probability distribution. The likelihood of the observed data is measured by assigning it a *likelihood function*, which is often derived from a well-known probability distribution. [1] After observing the data, she updates her degree of information to form the *posterior distribution*, using the Bayes' formula as described below. There are two views on Bayesian probability. A *subjectivist* Bayesian describes probability as a personal degree of belief. So-called *objectivist* Bayesians subscribe to an axiomatic view of probability, in the spirit of Aristotelian logic. Their methods include so-called reference priors. The two books [79, 206] give

[1] Note, however, that a likelihood function is not a probability distribution

a comprehensive account on the Bayesian view. The latter book—aimed for a more advanced audience—often contrasts Bayesian and frequentist methods. The original Bayes' essay is reprinted in [9].

Because they are sometimes used in the disease mapping field, we must note that there exist also so-called *empirical Bayes* methods, which try to retain an objective approach with the frequentist view that model parameters are not random variables, by first learning parameters from data and then using them a second time in the actual model. In this thesis we only consider full Bayesian models. In philosophical terms: "Abandoning the classical frequentist probability, one might as well become fully Bayesian" (loosely quoted from [206]). But naturally there is much more to this question.

### 4.1.2 Aleatory and Epistemic Uncertainty

Another view on the Bayesian and frequentist interpretations of probability is that there are two kinds of uncertainty. *Aleatory*[2] uncertainty is induced by randomness. Aleatory uncertainty is present whenever we are interested in one or more instances of a random process. *Epistemic*[3] uncertainty is due to our imperfect knowledge of something that is not random, and so it is knowable, at least in principle. A statistical model can be viewed as a representation of (aleatory) probability distributions and (epistemic) parameters. As noted above, frequentist probability can only refer to aleatory uncertainty, requiring events to be repeatable in a process having intrinsic randomness. Epistemic uncertainties are typically associated with unique events. Usually this applies to parameters of a statistical model, as well. If we wish to use probabilities to express epistemic uncertainty, we must turn to subjective probability, i.e., become Bayesians. Expert elicitation and risk analysis are examples of fields where the distinction between aleatory and epistemic uncertainty has been emphasised. [205, 207, 204, 214]

### 4.1.3 Bayesian Inference

The frequentists view the unknown parameters of a statistical model as fixed values. In contrast, Bayesians have the view that the parameters are random variables to which they can assign probability distributions.

[2] Latin: *alea*=die, as in the words attributed to Julius Caesar "*Alea iacta est*" – the die is cast

[3] Greek: pertaining to knowledge

In the following, the terms probability distribution and probability density are used intermixed. The Bayesian statistical conclusions about a parameter $\theta$ (or unobserved data $\tilde{y}$) are made in terms of *probability statements*. These probability statements are conditioned on the observed data $y$ and are denoted $p(\theta|y)$ or $p(\tilde{y}|y)$, where the vertical bar is read "given".

Bayesian statistics relies on the Bayes theorem. We first derive it for two mutually dependent random events, $A$ and $B$. The chain rule states that $p(A, B) = p(B|A)p(A)$. Using this with the law of conditional probability $p(A|B) = \frac{p(A,B)}{p(B)}$ leads to $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$. This is known as the Bayes' rule. [79] The usefulness of this formula may be illustrated with the frequently used example[4] in determining the diagnostic accuracy of a clinical test, as follows.

Suppose that $p(D^+) = 0.01 = 1\%$ of women who participate in routine mammography screening actually have breast cancer. Hence, $p(D^-) = 0.99 = 99\%$. Further, suppose $p(T^+|D^+) = 0.8 = 80\%$ of women who have breast cancer get a positive test result (true positive rate; TPR, i.e., sensitivity). On the other hand, $p(T^+|D^-) = 0.096 = 9.6\%$ of women who do not have breast cancer get a positive test result (false positive rate, FPR), thus the specifity is $p(T^-|D^-) = 1 - FPR = 0.904 = 90.4\%$. We are then asked to calculate the probability $p(D^+|T^+)$ that a woman who gets apositive test result in the routine screening actually has breast cancer. Using the Bayes formula, we have $p(D^+|T^+) = \frac{p(T^+|D^+)p(D^+)}{p(T^+)}$, where $p(T^+) = p(T^+|D^+)p(D^+) + p(T^+|D^-)p(D^-) = 0.8 \times 0.01 + 0.096 \times 0.99 = 0.10304 = 10.3\%$. Therefore we get $p(D^+|T^+) = \frac{0.8\times 0.01}{0.10304} \approx 0.078$. In other words, only 7.8% of women who get a positive test result actually do have breast cancer.

Turning back to the Bayesian statistical modelling, we derive a model which describes our *joint probability distribution* for $\theta$ and $y$. We write this as a product of two independent components, the prior distribution $p(\theta)$ and the likelihood function of the data given the model parameter(s), $p(y|\theta)$. Hence, we have: $p(\theta, y) = p(\theta)p(y|\theta)$. As above, using the law of conditional probability, we get the Bayes' rule for posterior density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)},$$

where $p(y) = \int p(\theta)p(y|\theta)\mathrm{d}\theta$ is the marginal probability of the observed data $y$ over all possible values of $\theta$. In the discrete case $p(y) = \sum_\theta p(\theta)p(y|\theta)$. Note that the marginal probability distribution of $y$ is also the *prior pre-*

[4] See, e.g., http://yudkowsky.net/bayes/bayes.html

*dictive distribution* of observable, but not yet observed data $y$. As this *normalisation constant* is independent of the parameter(s) $\theta$, it is often omitted to obtain the unnormalised posterior density $p(\theta|y) \propto p(\theta)p(y|\theta)$. The symbol $\propto$ is read "is proportional to". From the above formulae, another distribution can be derived, namely the posterior predictive distribution $p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta$, which in the present form shows that $\tilde{y}$ and $y$ are conditionally independent given the parameter $\theta$. [79]

### 4.1.4 The Role of Prior Information

The role of priors in Bayesian inference is multifaceted. In a simple model without multilevel hierarchy, the prior represents our knowledge (or lack of it) in terms of a probability distribution, before the data has been observed. After the data has been observed, the prior distribution can be updated to the posterior probability density, using the above Bayes' rule. This density can then be considered as our prior knowledge before observing some new data. According to the *likelihood principle*, all information from the observed data is contained in the likelihood function. Hence, the prior may not depend on the data.[5] Together, these principles lead to the fact that the exact manner in which data has been collected must not affect the inference. This leads to the fact that a Gaussian model, for instance, produces equal information when observed data is obtained and added one point at a time (as in process control) or if all data is obtained simultaneously. Another example comes from epidemiology: a clinical trial could be optionally stopped prematurely, for example, when funding is withdrawn. It is not valid to analyse the incomplete data using frequentist methods, but there is no problem when the data is analysed using Bayesian methods. [206, 79]

Our prior ignorance can be represented by vague priors, and in extreme cases, objective inference is sought by using *improper* uniform, flat priors. An improper prior does not correspond to a probability density, as it cannot be integrated in order to form a normalising constant. In contrast, *proper* priors are integrable probability densities. Even when an improper prior is used, it may lead to a proper posterior density, because the likelihood function is integrable. Jeffreys' priors are one attempt to form objective priors which retain their properties even under variable transformations (but in general they do not obey the likelihood principle).

[5] As we note, empirical Bayes methods do not obey this principle

In hierarchical Bayesian models (HBM's), on the other hand, priors are used as (informative) constructions to describe our assumptions on the dependency structures in the data. These priors are then completed by giving them so-called hyperpriors, which themselves depend on fixed hyperparameters. [206, 79]

### 4.1.5 Hierarchical Modelling

We begin with the concept of *exchangeability*. Given a set of experiments, for example the lifetimes of similar light-bulbs, it is plausible to assume that we cannot distinguish between them. Therefore, permuting the experiments would not affect our information. We could estimate independent failure rate parameters for each light-bulb—with excellent model fit—but that hardly makes sense: we are interested in the average failure rate. In order to avoid *overfitting*[6] and take best possible use of the available information, we should assume a common failure rate for all the light-bulbs and estimate it with a single model parameter.

In more complex situations, however, all observations might not be exchangeable—exchangeability merely states that we do not have (or are ignorant of) any knowledge that would differentiate the experiments. In case we had such knowledge, we should use it in our model. HBM's, also known as multilevel models [92], offer a flexible framework for problems of this kind. In HBM's the prior distribution for the lower hierarchical level is complemented by a hyperprior in the next upper level and this structure may be extended to include many levels. In this case the priors at the upper levels usually describe assumed dependence structures between the observations.

Hierarchical models can be represented using tree-like structures, known as graphs. Usually, the model construction assumes that there are no causal loops in the graphs, which leads to directed acyclic graphs (DAG's). Conditional independence is exploited in multilevel hierarchical models, for instance in the frequently used WinBUGS software package. HBM's are estimated by sampling from the full conditional (FC) distributions (see below). The models often exploit *conjugate priors* so that the prior and posterior distribution remain in the same family of probability densities. Examples include Gamma-Poisson and Beta-Binomial models. Sampling from the standard probability distributions is straightforward. [267, 84,

[6] Overfitting refers to a situation where a model fits the data well, but makes poor in prediction—hence the model does not represent a general case

163, 36, 162, 79]

One concrete example [79] describes how to combine information on educational tests from eight schools. Using a hierarchical Gaussian conjugate model, we assume at the first level that within each school $i$, the scores $y_{ij}$ of each pupil $j$ are normally distributed[7] with common means $\theta_i$: $y_{ij} \sim \mathcal{N}(\theta_i, \sigma_j)$. The means in each school are normally distributed by hyperparameters $\mu$ and $\tau$: $\theta_i \sim \mathcal{N}(\mu, \tau)$. Out of convenience, a uniform hyperprior is assigned to $\mu$ given $\tau$, thus: $p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$. The model is completed by assigning a prior for $\tau$— again by convenience, we assign a flat hyperprior: $\tau \propto 1$. In this example, $\sigma_j$ are assumed to be known from other sources. The fact that these improper prior choices lead to a proper posterior distribution is given in [79].

### 4.1.6 Markov Chain Monte Carlo Sampling

Bayesian inference often leads to situations where there is no analytical solution for the posterior distribution, especially when constructing HBM's. The development of sampling methods based on simulation using pseudorandom numbers, along with the ever-increasing computer capacity, has led to current wide-spread use of Bayesian methods. MCMC in Bayesian statistics was popularised by its use in image analysis, namely in Markov random field (MRF) models—the same models that we now use in disease mapping. [25]

Monte Carlo (MC) and later Markov chain Monte Carlo (MCMC) methods were the brainchildren of the Los Alamos scientists who during WW2 developed the first electronic computer, ENIAC. The computer was built to replace a virtually limitless line of women who were constantly solving ballistical tables by cranking electromechanical hand calculators. One interesting problem for ENIAC was solving neutron diffusion in fissionable material. After an initial setting of several neutrons was established, the time evolution of the system was simulated using known statistical probabilities according to the physical and geometric factors of the experiment. Simulated events for individual neutrons were based on pseudo-random numbers. The analogy to events in a casino led to the name Monte Carlo after the famous casino.

[7] Throughout this thesis, we use *precision*, i.e. the inverse of variance, in parameterising normal distribution. In addition, the gamma distribution is parameterised as Gamma$(shape, rate)$.

Generally speaking, MC sampling (which is a frequentist method),[8] is used when a deterministic analytical solution of a problem is not available. In the method, several independent random realisations from the domain of possible model inputs are created and the outputs in each case are computed deterministically. The result is then obtained by averaging over the sampled outputs. MC sampling is useful for instance in numerical integration and optimisation in multidimensional spaces. MCMC samplers, on the other hand, produce a chain of dependent samples from the probability distribution of interest, $p(x)$. In the following, we describe the most often used MCMC samplers. [178, 177, 4]

*Metropolis and Metropolis-Hastings Sampling*

Metropolis sampling was a further development of the MC sampling. It was used in a multiparticle problem in statistical mechanics of calculating a quantity of interest $F$ in the equilibrium of state—in statistical terms, this is the expected value of $F$. We briefly introduce the method following the original development[9] which led to the Metropolis sampling algorithm. [178] The potential energy $E$ of a system with $N$ particles in any state is easily determined (Equation 1 in the article). Using the *canonical ensemble*, the microscopic states (i.e. the state of each particle) of the system can be described by the Boltzmann distribution: $p_i = \exp(-E_i/kT)$, where $k$ is the Boltzmann constant and $T$ is the absolute temperature. $p_i$ is the proportion of the particles which would exhibit a measurable macroscopic state $i$, i.e. have the potential (repulsion) energy $E_i$. Now the expectation of $F$ is calculated as $\bar{F} = \int F \exp(-E/kT) / \int \exp(-E/kT)$. This is a multidimensional integral in the $2N$ dimensional configuration space (for illustration, a system in a 2-D square was used in the article). A solution by MC would involve generating random configurations and weighting each configuration by $w = \exp(-E/kT)$. This is not practical in close packed systems as states with low weights $w$ would be selected with high probability.

Therefore a modified version of MC was developed, which would later be called MCMC. Instead of choosing configurations randomly and weighting them with $w$ as in MC, configurations are chosen with probability $w$ and weighted evenly. The sampling proceeds as follows. The system of

[8] However, so-called Bayesian Monte Carlo (BMC) has been recently introduced [236]

[9] This may be rather difficult to grasp from the original article, but it is very interesting when understood, hence the introduction this way

$N$ particles is first initialised to an arbitrary state (we use the above 2-D unit square system). Then each particle is moved to a new position: $X^* = X + \delta u_1$ and $Y^* = Y + \delta u_2$, where $u_1$ and $u_2$ are drawn uniformly between $[-1, 1]$ (these uniform densities are called proposal densities), and a tunable parameter $\delta$ is the maximum allowable displacement. If we get outside the square, the particle is re-entered from the other side. The change of energy, $\Delta E$, related to the move of particles is then calculated. If the new system has lower energy, the move is accepted.[10] If the new system has higher energy, the move is accepted with probability $w = \exp(-E/kT)$. In practice, a random number $u_3$ is drawn uniformly between $[0, 1]$, and the move is accepted if $u_3 < \exp(-E/kT)$. Otherwise the system is returned to the original position. A number of iterations are first run "in order to get rid of the effects of the initial configuration on the averages". Then after each iteration, the quantity of interest $F$ is calculated. After $M$ iterations we have simulated the expected value $\bar{F} = \frac{1}{M}\sum_{i=1}^{M} F_i$. [178]

Contemporary Metropolis samplers commonly use symmetric Gaussian random walk proposals, i.e. $X^* \leftarrow \mathcal{N}(X, \tau)$. Here the precision parameter $\tau$ is used for *tuning the acceptance* rate. Inference often concerns the expected value of $X$ itself, instead of the expected value of some function(al) which depends on $X$. Metropolis sampling was later generalised by Hastings [99] to cases where the proposal density is not necessarily symmetric. In Metropolis-Hastings sampling the acceptance ratio becomes $\alpha = \frac{p(x)q(x|x^*)}{p(x^*)q(x^*|x)}$, a random number $u$ is drawn uniformly between $[0, 1]$, and the move is accepted if $u < \alpha$. $q(x|x^*)/q(x^*|x)$ is the proposal ratio, which in the case of Metropolis sampling is 1 and can thus be omitted. [42] Finally we note that in almost all practical applications, log-likelihoods are used, i.e., all the above formulae are log-transformed, because computers have a limited precision and range of real numbers. Another computational asset of using the log-scale is that power calculations reduce to multiplications and correspondingly, multiplications reduce to additions.

As is evident from the Metropolis algorithm above, each of the generated samples in the chain only depends on the previous sample, i.e., the chain has the *Markov property*. Hence, the generated chain is a Markov chain. We note that a Markov chain has to fulfill certain criteria in order

[10] This leads to the maximum *entropy* principle, which joins information theory and statistical mechanics [120, 121]

to converge to the invariant distribution $\pi(x)$ we want to sample from.

The Markov chain theory states that any chain which is *irreducible* and *aperiodic* will have a unique stationary (=limiting) distribution and a $t$-step transition kernel[11] $\mathcal{P}(x, x^*)^t$ will "converge" to that distribution as $t \to \infty$. In an irreducible Markov chain it is possible to get to any state from any state, i.e. all states of the chain communicate with each other but not with any other state. A Markov chain is aperiodic, if it is possible to get to any state from any state in one step. In slightly tighter terms, a Markov chain is *positive recurrent* if the expected return time to any state is finite (which implies that the chain is also irreducible). An aperiodic and positive recurrent Markov chain is said to be *ergodic*. In the MCMC sampling we need an ergodic Markov chain with the property $\pi\mathcal{P}(x, x^*) = \pi$, i.e. given $x \sim \pi(x)$, if $x^* \sim \mathcal{P}(x, x^*)$ then $x^* \sim \pi(x^*)$ also. *Reversible* Markov chains have the necessary properties, i.e., they obey the *detailed balance* $\pi(x)\mathcal{P}(x, x^*) = \pi(x^*)\mathcal{P}(x^*, x)$. [36, 242]

*Gibbs Sampling*

The original paper on Gibbs sampling [82] proved the equivalence of MRF's and Gibbs distributions (of which the Boltzmann distribution is a special case), hence the method was named Gibbs sampling. The development followed the "Heat bath" version described in [178]. However, the method had been presented independently in other papers by other names. [25] The paper [82] made a formal link between statistical mechanics and image analysis.

Gibbs sampling is a special case of Metropolis-Hastings sampling, in which the proposed moves are accepted with probability 1. This is an attractive option when the FC is in the form of a standard probability distribution—which happens when using conjugate priors. Otherwise Metropolis-Hastings, or nowadays slice sampling is usually a better choice. [25]

As an example, a single update in the systematic scan (see below) Gibbs sampling proceeds as follows. We start with an arbitrary initial configuration, $\mathbf{x}^0 = \{x_1^0, \ldots, x_k^0\}$. Then each variable in turn is systematically updated:

[11] A transition kernel $\mathcal{P}(x, x^*)$ is the conditional distribution of the next state $x^*$ given the current state $x$

$$\begin{array}{lll} x_1^1 & \text{is sampled from} & p(x_1|x_2^0,\ldots,x_k^0) \\ \vdots & \vdots & \vdots \\ x_i^1 & \text{is sampled from} & p(x_i|x_1^1,\ldots,x_{i-1}^1,x_{i+1}^0,\ldots,x_k^0) \\ \vdots & \vdots & \vdots \\ x_k^1 & \text{is sampled from} & p(x_k|x_1^1,\ldots,x_{k-1}^1). \end{array}$$

*Slice Sampling*

There are certain methods for automatically tuning the Metropolis-Hastings acceptance rate, but especially in MCMC sampling of HBM's, there is a need for a generic sampler which would not require any tuning and would easily handle things like multimodal distributions. Slice sampling [197, 196] is one possible solution, and it has found widespread use e.g. in WinBUGS [267].

Slice sampling is an example of samplers which use auxiliary variable [24] methods. A single slice sampling update from a density $f(x)$ is performed as follows. First we assume that we are at some current point $x$. An auxiliary variable $y$ is drawn uniformly from $[0, f(x)]$. Using this height we form a horizontal slice by expanding alternatively to the left and right until both ends ($x_L$ and $x_R$) of the slice are at a higher point than the density, i.e. $f(x_L) < y$ and $f(x_R) < y$. In practice the expansion is limited to 10 (say) iterations, to avoid infinite loops. Now we repeatedly sample a point $x^*$ uniformly from $[x_L, x_R]$, until $f(x^*) < y$. Finally we set $x \leftarrow x^*$ and discard $y$. This procedure is then repeated until enough points $x$ are generated. In practice there are several alternatives for performing the initial expansion. Also, the sampling procedure for $x^*$ is inefficient, and in practice a "shrinkage" procedure is used. Sampling from truncated distributions is easy: we simply use the truncation points as hard boundaries when performing the expansion. [197, 196] A basic algorithm in C++ is given in Appendix D.

### 4.1.7 MCMC in Practice

Having introduced a set of MCMC samplers we now discuss how multivariate simulation is performed using MCMC. Often we have to simulate from multivariate distributions which are in a nonstandard form, and have dependent components. Simulation must then use *conditional distributions*. The chain rule states that:

$$p(x) = \prod_{i=1}^{n} p(x_i|x_{j<i}).$$

In a simple model, reordering the terms might allow sequential static simulation from $p(x)$, but normally this is not the case. From the above equation it follows that $p(x_S|x_{-S}) \propto p(x)$ for any subset $S$ of parameters. In particular, for a single parameter, $p(x_i|x_{-i}) \propto p(x)$. The above formulae are known as *full conditionals* (FCs). We note that the product form of joint distribution arises frequently in Bayesian posterior distributions, particularly in HBM's. Graphical models which form the backbone of HBM's often introduce conditional independence structures which can be exploited in simplifying the FCs. In HBM's the FC distribution of node $v$ is usually expressed as the product

$$p(v|\text{rest}) \propto p(v|\text{parents of } v) \prod_{u \in C_v} p(u|\text{parents of } u),$$

where $C_v$ represents the set of children of the node $v$. [25, 36]

There are several options for updating the FCs. Often a single-site Metropolis-Hastings or a Gibbs sampler is used. Another option would be block-updating some dependent parameters in a Metropolis-Hastings step. However, there is a limit in how many parameters can be block-updated simultaneously; this is because multivariate distributions are more sensitive to parameter changes than univariate distributions. When the proposed changes are small enough to produce a good acceptance rate, the sample autocorrelations are too high, and it could take an eternity to produce enough independent samples. There are alternatives for the parameter visiting schedule. Systematic scans are often used (e.g. in WinBUGS) even though they could produce unwanted drift effects. Random scans might be a preferable choice, also allowing the visit probabilities of individual parameters to be chosen. As modified from the random scan, a semi-regular scan would prohibit successive visits to the same site. [25, 36]

As mentioned in the treatment of Metropolis-Hastings sampling above, a MCMC sampler is first initialised to some state, which could be either random or an *a priori* probable state. A number of iterations are then simulated, until the sample paths have stabilised, i.e., the sample chains have converged to the invariate distributions from which we wish to simulate. This phase is known as the burn-in. Then we generate a large number of samples; the exact amount depends on the required accuracy (known as the Monte Carlo error) and on the available computer capacity. In some cases there could also be considerable autocorrelation in the sample chain (as MCMC generates dependent samples). Autocorrelation reduces the effective sample size. One remedy is to use thinning, i.e., store samples only in every 20$^{\text{th}}$ (say) iteration. [83, 79]

A bit outdated discussion of the sampling strategies of various researchers is in [129]. Among the choices is whether we use several shorter chains for controlling the possible effect of initial state, or one longer chain for reducing potential autocorrelation and reduce the risk that the chain would suddenly jump to another mode, which would remain unnoticed if only shorter chains were used. Another issue is how the convergence is assigned. Some authors prefer visual inspection of the generated chains (as we mainly do in this thesis) and others like to use diagnostic tests, e.g. [80]. One point to remember is that parameterisation may have a considerable effect in the model sampling efficiency [78]. Auxiliary variable methods are one possible remedy to improve model mixing [24, 110].

*Posterior Summaries*

After we have generated a sufficiently long chain of samples from the desired distribution, there is the question "how do we describe our distribution?" Although full Bayes models give the actual posterior densities, for reporting purposes the estimates of model parameters are usually expressed in terms of posterior summaries. The parameter estimates can be described by posterior means, medians or modes.[79]

Beyond point estimates, the variability of these estimates are usually described by so-called credible intervals (CI) or highest density regions (HDR). A $p$-% CI is the central interval supporting $\frac{p}{100}$ of the posterior mass. A $p$-% HDR is the most compact set supporting $\frac{p}{100}$ of the posterior mass. HDR is sometimes used when the posterior density is skewed to get a more accurate estimate. In simple terms, both CI and HDR might be called the Bayesian version of frequentist confidence interval, but there is an important difference. A Bayesian CI or HDR has p% probability that the parameter of interest lies within that interval. In contrast, the frequentist confidence interval has the interpretation that in a large number of repeated samples, p% of the calculated intervals would contain the true value of the parameter of interest. However, the frequentist confidence interval is often mistakenly considered as if it had the Bayesian interpretation. [79]

In a regression model we may check whether the 95% (say) CI of an effect ($\beta$) contains zero. If not, there is strong evidence for an association. As we see, Bayesian statistics considers the (subjective) degree of evidence, not some arbitrary threshold between "not statistically significant" and "statistically significant".

One versatile and simple measure to consider is the "Bayesian marginal posterior tail probability". As an example, suppose we have MCMC simulations of $N$ draws from the posterior distribution of two parameters $\theta$ and $\phi$. We wish to assess whether $\theta > \phi$. We calculate the posterior tail probability as $p(\theta > \phi) = \frac{1}{N}\sum_{i=1}^{N}[\theta_i > \phi_i]$, where Iverson bracket denotes the indicator function. This kind calculations are commonly used in, e.g., disease mapping to calculate region-wise posterior probability of excess disease risk. In the Bayesian setting, no direct hypothesis is done, but there are some experimental rules for decision making, to which we return later [240].

### 4.1.8 Bayesian Model Comparison and Averaging

More advanced Bayesian modelling is concerned with with the comparison of candidate models and validating the fit of the chosen model(s). Bayes factors [130] may be used for model comparison. In some cases we do not choose a single 'best fit' model, but instead use model averaging [112] to overcome the inherent uncertainty of choosing the correct model. In the model averaging methods the final posterior distribution is the average of the posterior distributions of each model, weighted by the posterior probabilities of choosing the corresponding models. Transdimensional MCMC is one of the advanced methods used in Bayesian model averaging.

*Transdimensional MCMC*

There exist a number of challenging statistical problems in which the dimension of the object of interest is not fixed [89]. Simultaneous inference on both model and parameter space is a fundamental issue in modern statistical practice [264]. Reversible jump Markov chain Monte Carlo (RjMCMC) [89] was a natural generalisation of Metropolis-Hastings algorithm to so-called Metropolis-Hastings-Green algorithm which is capable of sampling between model spaces of variable dimensions. This has enabled the use of partition models which approximate surfaces using a variable number of tiles each having a constant level. A smoothed surface is obtained by averaging over possible configurations. At the same time, possible jumps in the surface level can be easily retained. Another application is choosing variables in a regression model. Usually these applications lead to posterior model averaging [112], but another option is to select only the most probable model configuration in calculating posterior summaries.

Shortly described, the algorithm relies on fulfilling the usual reversibility requirements of the MCMC samplers. In the case of jumping between model spaces of variable dimension, new model parameters can be derived from the existing ones using an auxiliary random variable—and taking care of the Jacobian term resulting from the change of variables when calculating the Metropolis-Hastings acceptance ratio. This is illustrated in the "coal mining disasters" example in the original paper. [89] On the other hand, when new variables for jumping to a higher dimensional space do not necessarily depend (directly) on the current variables, jumps between variable dimensional parameter spaces can be made without using auxiliary variables (as e.g. in [143, 54]). We note that alternatives for RjMCMC do exist, see e.g. [271, 38, 264].

### 4.1.9 Sensitivity Analysis and Model Validation

We must bear in mind that our model could be sensitive to the underlying assumptions. The posterior distribution of the model parameters could either over- or underestimate various aspects of "true" posterior uncertainty. Typically the posterior distribution of model parameters overestimates the uncertainty in the sense that all of one's substantive knowledge is not included in the model. However, even a good model is just a simple representation of the true phenomenon.[12] Hence, we need to do posterior model checking against the observed data. It might also turn out that other reasonable models could have fit the data equally well. In model expansion, a larger model could be constructed with suitable parametrisation to contain the alternative models as special cases. Another possibility is doing model comparison or validation by checking which of the models has better predictive accuracy, possibly with penalising model complexity (number of parameters). [79, 206] Bayes factors [130] form one option for comparing two models.

A model could also be sensitive to the prior assumptions. Informative prior distributions could have an impact on the results. This may be checked by using various choices for the prior hyperparameters or by trying a prior from another family of distributions. Also the likelihood model

[12] "All models are wrong, but some models are useful." — G. E. P. Box

could affect the model sensitivity. One example is the Gaussian distribution, which could be replaced by a more robust long-tailed Student-$t$ distribution. [79, 206] The importance of priors in disease mapping was considered in [17].

*Deviance Information Criterion (DIC)*

DIC was introduced in the disease mapping context as an information criterion which would enable comparison of BHM's which are estimated using MCMC sampling. In a hierarchical model, informative prior structures could cause considerable shrinkage and therefore the effective number of parameters might be much lower than the actual number of parameters. Thus, penalising by the actual number would lead to a too conservative measure. At the moment, DIC provides only point estimates and significant differences in model performance must be deduced by some sort of rules of thumb. Quoting from the DIC FAQ,[13]

> It is difficult to say what would constitute an important difference in DIC. Very roughly, differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

Despite the criticism, DIC has remained the most commonly used measure of model fit in HBM's, especially in the disease mapping community. [268, 269]

We first define deviation as $D = -2\log(p(y|\theta)) + C$, i.e., it is -2 times the log-likelihood of the data $y$ given the parameters $\theta$. $C$ is an arbitrary constant which depends only on the data. When two models for the same data are compared, this constant cancels out. The expectation of deviance $\overline{D} = E^{\theta}[D(\theta)]$ measures model fit; the smaller it is the better the model fits. As more and more parameters are added in the model, it is easier to get a better model fit. $p_D = \bar{D} - D(\bar{\theta})$ is called the effective number of parameters. The larger this is the easier it is for the model to have a good fit. DIC is then defined as $DIC = p_D + \overline{D}$. This means that a poor model fit and a large effective number of parameters both indicate a poor model. [268, 269]

[13] http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml. Accessed September 13, 2011.

*Cross-validation*

In cross-validation (CV), we leave out a subset of the data when fitting the model and then use the model to predict the data which was left out. This is done sequentially for all observations. Most common strategies are leave-one-out CV and $K$-fold CV. In leave-one-out CV, one observation at a time is left out and predicted. In $K$-fold CV, $100/K\%$ of the data is left out. Leave-one-out CV is naturally more accurate but it requires estimating the model as many times as there are observations. $K$-fold CV requires estimating the model only $K$ times. Traditionally, model fit has been measured by mean squared error or root mean squared error. Especially in chemometrics, cross validated $\mathrm{R}^2$ has also been used.

In Bayesian CV, expected utilities are used for assessing the cross-validated model goodness. MSE and RMSE are possible measures of expected utility, but in some cases the expected utilities may be application specific. In the case of leaving observations out one at a time, importance sampling leave-one-out CV may be used to obtain some computational savings. One problem with these methods is that in frequentist models, algebraic solutions are usually available and fast to compute, but in Bayesian modelling, we usually must use MCMC methods. As a single model could take hours (if not days) to compute, this might make CV an impractical method. For further technical discussion, see e.g. [292, 291]. An approximate CV method has been presented for checking extreme observations in disease mapping [272]. There also exist various methods based on posterior predictive replicates—see, e.g., [81].

*Widely Applicable Information Criteria (WAIC)*

Recently the asymptotic equivalence of Bayes CV and widely applicable information criterion (WAIC) has been proved for singular learning machines [302, 301]. WAIC is a very promising measure for model goodness because of the simplicity of computation and broader applicability compared to DIC. However, the definition of WAIC is different between the two cited references, and therefore we must wait until the community or the author of WAIC decides which is the correct version.

## 4.2 Spatial Modelling and Smoothing

There are three basic types of spatial data [8]:

1. *Point pattern data*, where $D$ is a random subset of $\mathbb{R}^r$ (i.e, the $r \in (1, 2, 3)$-dimensional space of real numbers). The index set $\mathrm{s} \in D$ gives locations of random events. If $\forall\ \mathrm{s} : Y(\mathrm{s}) = 1$, this would define a simple *point process*. However, $Y(\mathrm{s})$ could give some additional covariate information, producing a *marked point process*.

2. *Point-referenced data*, where $Y(\mathrm{s})$ is a random vector at locations $\mathrm{s} \in \mathbb{R}^r$, which vary continuously over $D$, a fixed subset of $\mathbb{R}^r$, defining a $r$-dimensional rectangle of positive volume. This is often referred to as *geostatistical* or *geocoded* data.

3. *Areal data*, where $D$, a fixed subset of $\mathbb{R}^r$, is partitioned into a finite number of non-overlapping areal units with well defined boundaries.

Different data types call for different modelling approaches. Point process data is analysed using spatial point process models. Geostatistical data is analysed by kriging, i.e., using Gaussian process regression. Areal data is analysed using (ecological) areal level models. In the disease mapping context, conditional autoregressive (CAR) models are often used. However, these are merely the basic modelling rules.

In a wider perspective, spatial modelling can be seen as any modelling task where observations have some type of *spatial dependence* structure. These observations might be for example:

- Number of events in geographic locations (e.g. incident disease cases)

- Geographic observations (e.g. geochemical concentrations in soil)

- Number of events (or some continuous levels) in a time-series

- Nearby genotypic alleles in a DNA strand (within a chromosome)

- *Spatiotemporal* observations, i.e., a temporal series of events observed in some spatial region.

The last item reminds us that time may also be seen as a spatial dimension, although it has the directional causality property. This leads to the concept of *space-time continuum*.

We usually consider situations where:

> Everything is related to everything else, but near things are more related than distant things.

This is called Tobler's first law of geography [279]. Initially, these near things were probably considered as near in the sense of being geographically close to each other. However, things could be near w.r.t. some other measure, e.g., rural areas could be similar to each other and so could be urban areas. In mathematical terms, we consider a wide class of distance based correlation measures.

In the context of spatial epidemiology, there are basically four kinds of spatial smoothing models. In this thesis the first two model types below are used. We note that there exist several other models and model extensions which are not covered here.

### 4.2.1 Spatial Conditional Autoregressive Models

Conditional autoregressive models (CAR) have been widely applied in spatial epidemiology. Based on the seminal work [22], a CAR model joins the (usually) Gaussian random field and Markov random field models. The result is a conditional Gaussian Markov random field (GMRF) model. The model has the local Markov property, i.e., the conditional probability density in each region depends only on the observed values in the adjacent regions. The CAR model is usually used as a prior dependence structure for a spatially structured random effect in a hierarchical model. Recently, a multiple membership prior in the CAR framework was considered for spatially discontinuous regions in [47].

Although the conjugate gamma prior for Poisson regression model would have certain good properties (i.e., it would scale correctly under aggregation or refinement of regions, unlike GMRF models), the original work [22] has a pessimistic view of the possibility of gamma MRF models. Poisson/gamma model for spatial point processes was presented in [308], but at least in the disease mapping field it did not receive much interest. Few exceptions are the linear Poisson regression model (which is available in GeoBUGS [277]) for combining health and exposure data measured in disparate resolutions [30, 31], and the spatial partition model of [54]. Two Markov gamma random field models were finally independently presented in [199] and [43].

CAR models were first applied in disease mapping using an empirical Bayes method [44], and a full Bayes model (BYM) was introduced in [28]. Slightly earlier, the CAR model was applied in image restoration [23], the field from where the Gibbs sampler also originated [82]. Empirical and full Bayes models were compared in [19], and the later developments of the empirical Bayes models are discussed in [179]. In short, the estimates from empirical method have too narrow confidence intervals, as the uncertainty of the smoothing parameter (spatial precision) is not taken into account [19], although bootstrapping offers a partial remedy [19, 179]. The empirical Bayes smoothing method in Rapid Inquiry Facility [7] is based on simple gamma-Poisson smoothing towards global mean [44]. In this thesis we consider only full Bayes models.

*Intrinsic CAR Prior*

The widely used Gaussian intrinsic CAR prior (iCAR) is a special case of CAR priors, leading to an improper distribution. The formulation is based on Gaussian pairwise differences:

$$p(\lambda_i|\tau) \propto \tau_\lambda^{m_i} \exp\left\{-0.5\tau_\lambda m_i \sum_{i\sim j}(\lambda_i - \lambda_j)^2\right\},$$

where $i$ is an index of regions, $i \sim j$ refers to the $m_i$ regions $j$ that are neighbours to $i$ and $\tau_\lambda$ is the spatial precision[14] [28]. We may also write $\lambda_i \sim \mathcal{N}(\overline{\lambda_{j\sim i}}, m_i\tau_\lambda)$, which is equal to the above formulation. The fact that this particular set of fully conditional distributions leads to a joint probability density for $\lambda$ is not trivial,[15] but it is proved by Brook's lemma [35] (also known as the Hammersley-Clifford theorem [22]). We may write the joint density as:

$$p(\lambda|\tau_\lambda) \propto \mathcal{N}(\lambda, \tau_\lambda \mathbf{K}_\lambda),$$

where $\mathbf{K}_\lambda$ is the $N \times N$ structure matrix:

$$K_\lambda = \mathbf{diag}(\mathbf{m}) - W,$$

where $N$ is the number of regions and $\mathbf{m} = \{m_1, \ldots, m_N\}$. $W$ is the neighbourhood weight matrix: $W_{ij} = \mathbf{1}_{i\sim j}$. In MCMC simulation, $\lambda$ may be slice-sampled (see appendix D) element by element from the conditional distributions (see the Convolution Model below). For the sake of

[14] In this thesis we denote the precision of a parameter $\theta$ as $\tau_\theta$; correspondingly, the structure matrix is denoted as $\mathbf{K}_\theta$

[15] We note that a joint distribution can always be defined by its FCs, but only certain sets of FCs define a joint distribution

model identification, we must recenter it immediately after sampling: $\lambda \leftarrow \lambda - \overline{\lambda}$. The precision $\tau_\lambda$ may be Gibbs-sampled from its FC: $\tau_\lambda \sim \text{Gamma}(a + 0.5(N-1), b + \lambda^\mathsf{T}\mathbf{K}_\lambda\lambda)$, assuming we use a conjugate gamma prior, i.e. $\tau_\lambda \sim \text{Gamma}(a, b)$. Sparse matrix algebra is beneficial in calculating $\lambda^\mathsf{T}\mathbf{K}_\lambda\lambda$. For further discussion of the model formulation and computational issues, see, e.g. [28, 140]. The book [245] is a good source of the MRF theory and also for the below mentioned Gaussian approximations.

*Proper and Multivariate CAR*

The iCAR prior is actually a limiting case of a proper CAR prior (it is available in GeoBUGS [277]). This proper CAR prior has been considered for disease mapping, but the problem is that it cannot model considerable spatial autocorrelation [26]. Hence, the BYM iCAR prior [28] is most often used. Another option for a proper CAR prior is that of [96], which has gained popularity among some authors. An improper multivariate CAR model (MCAR) was described in [145] and the proper MCAR model was developed in [77]. In the disease mapping model comparison [29], the performance of MCAR was not good.

*Fast Sampling of GMRF's*

Block updating of the correlated parameters of the BYM model would improve model mixing and gain computational speed [145]. An approximate fast sampling algorithm of GMRF's was introduced in [244]. The algorithm exploits the fact that Cholesky decomposition can be used in generating multinormal random variables. When this is combined with a band matrix rearrangement of the very sparse covariance matrix of CAR models, we may have significant savings in sampling time. A Taylor-expansion may be used for the Poisson-likelihood of rare event data on disease occurrence. Although these methods seem attractive, so far they have not gained popularity in disease mapping studies. Later develoment suggests using Laplace approximations to completely avoid the need for sampling [246]. Related to block updating, we have tried simple Metropolis-Hastings block updates of the CAR model and found that updating $\lambda$ works for blocks up to (say) 100 regions—with more regions the random walk moves have to be too short in order to get a reasonable acceptance rate, resulting in a high autocorrelation of the generated samples. With the conditional independence structure of the CAR model, it is obvious that this kind of block updating allows parallel computing.

*Simultaneous Autoregression*

Simultaneous autoregression (SAR) was developed much earlier than CAR [306]. SAR models fit well in the maximum likelihood based inference, whereas CAR is the natural choice in BHM's. Another difference is that SAR models assume spatial stationarity (as the article name [306] suggests), whereas CAR adapts to the local, possibly non-stationary patterns, which is important in, e.g., disease mapping. [8] Note that any SAR model can be represented as a CAR model but the opposite is not necessarily true. [48]

*Convolution Model (BYM)*

In disease mapping, the commonly used Besag, York and Mollié (BYM) convolution[16] model using the iCAR prior is constructed as follows [28]. At the first hierarchical level we model Poisson rates $\mu_i$ for cases $y_i$ in each region $i$: $Y_i \sim$ Poisson$(\mu_i)$. The Poisson rates are modelled by a log-linear regression model: $\mu_i = \exp(\alpha + \lambda_i + \eta_i)e_i$, where $\alpha$ is the baseline level, and $e_i$ is the expected number of cases in the region in question. At the second level, the iCAR prior is assigned to $\lambda_i$, which is the spatially structured random effect: $\lambda_i \sim \mathcal{N}(\overline{\lambda_{j\sim i}}, m_i\tau_\lambda)$. The notation is as above, but $\tau_\lambda$ is the spatial precision. As an identifiability constraint, $\sum_i \lambda_i \equiv 0$. Optionally, when using the BYM convolution model, $\eta_i$ is the spatially unstructured random effect: $\eta_i \sim \mathcal{N}(0, \tau_\eta)$, where by model definition, $\sum_i \eta_i \equiv 0$. At the third level, we have the priors: vague Gamma priors are given for the precisions, $\tau_\lambda, \tau_\eta \sim$ Gamma$(0.01, 0.01)$ is a usual choice. A flat prior is always assigned for the baseline : $\alpha \propto 1$.

For the spatially structured random effect we have the log-FC (omitting constants): $\ell(\lambda_i) \propto [y_i\varpi_i - \exp(\varpi)] + 2m_i\tau_\lambda(\lambda_i - \overline{\lambda_{i\sim j}})$, where the first part is the Poisson-likelihood and the last term is the iCAR prior. The "log-prediction" is $\varpi_i = \log(e_i) + \alpha + \lambda_i + \eta_i$, and $\overline{\lambda_{i\sim j}}$ is called the *local mean* of $\lambda$ at the region $i$. Correspondingly, we have for the spatially unstructured random effect: $\ell(\eta_i) \propto [y_i\varpi_i - \exp(\varpi)] + 2m_i\tau_\eta(\eta_i^2)$, with the terms as above. As the prior for the baseline is flat ($\alpha \propto 1$), we have $\ell(\alpha) \propto \sum_i (y_i\varpi - \exp(\varpi))$. We may use slice-sampling with all these terms. The FC for $\tau_\lambda$ was given above in the description of iCAR prior. $\tau_\eta$ may be Gibbs-sampled from its FC, $\tau_\eta \sim$ Gamma$(a + 0.5N, b + \eta^2)$, again assuming the conjugate Gamma prior. In the sampling algorithm, we must take care to recenter $\lambda$ and $\eta$ immediately after sampling. $\alpha$ is

[16] In this case, convolution means the convolution of spatially structured and spatially unstructured random effects

the only parameter which we (must) allow to drift. However, we note that this model is available in Win/GeoBUGS [267, 277], which was used in Publications I–IV.

*Neighbourhood Structure*

The neighbours are usually defined as the regions which share any common boundary with region in question. In case of a regular lattice, this leads to the second-order neighbourhood,[17] where also the diagonal neighbours are counted—thus leading to a maximum of eight neighbours. We must note that this is only one possible choice, and not necessarily the best one (e.g. the diagonal distance in a regular lattice is $\sqrt{2}$ times the horizontal or vertical distance). Many other choices are compared with this one in [62]. More general lattices for MRF's were considered in [149] and references therein, including the hexagonal honeycomb lattice, which would lead to a less anisotropic correlation structure. Of course, the conventionally used assumption of spatially anisotropic correlation in disease mapping is a strong one, but in case it is used, the model should reflect the assumption as closely as possible. One problem with these more general lattices would be the difficulty in data aggregation. In case of irregular administrative areas, this is not (directly) relevant.

*Underlying Assumptions*

Moreover, there are certain underlying assumptions, which justify the use of a Poisson process model:

1. Individuals within the study population behave independently w.r.t. disease propensity, after allowance is made for observed and unobserved confounders. In other words, hypothetically conditional on fully specified factors for each individual, they would have independent probabilities of conducting the disease.

2. The underlying population at risk has a continuous spatial distribution within the study area. Modifications are required when there are uninhabited regions (see below).

3. The case events occur as single, unique, spatially separate events.

[17] In chess terms, this is the queens neighbourhood; c.f. first order (=rooks) neighbourhood

In this case the counts are Poisson-distributed with a region specific expectation, and this expectation is defined as a multiplicative function of a background intensity with a log-linear predictor term. [164]

*Posterior Probabilities*

One thing to consider is whether there is evidence for spatial clustering, i.e. whether there is evidence for an excess of cases in some regions. One initial option is to use tests for clustering: either global clustering, e.g. [27] or local clustering, e.g. scan statistics [156] (which have been criticised). However, if we opt for disease mapping models, the Bayesian methods allow us to calculate the marginal posterior tail probabilities or CI's. An early view considered 95% CI's of the estimated RR's and concluded (in a non-Bayesian way) the results to be statistically significant if the CI excluded 1.0 [185]. Later, decision rules were considered based on simulated data, with the conclusion that using 70-80% posterior probability that RR excluded 1.0 as a cut-point gives reasonable sensitivity with moderate expected counts ($\sim$20) and excess risks ($\sim$1.5-2.0) [240].

Calculating the posterior probability that the incidence/prevalence rate in a region exceeds the average rate is trickier. If the average rate is calculated as the mean estimated rate over the regions, it is certainly wrong as we have noticed in retrospect in [126]. The average over geographic regions does not correspond to the overall pooled rate, because the former gets a biased weighting: regions with sparse population have as much influence as the urban areas. We have experimented with weighting based on the running (or *local*) mean population counts. If these are used as weights when calculating the spatial mean rate, the result is well in accordance with the overall pooled average. However, we have not tried to prove this result which is only an approximation. Furthermore, we must note that the incidence/prevalence rate is not symmetric around the average rate. For example, if we have an average rate of 100 (in some arbitrary units), a RR of 2.0 leads to 200, or a difference of 100, but a RR of 1/2.0 only leads to a difference of 50 units. As we recall from above, the probability of exceeding the average RR would be calculated from the chain of $N$ samples as $p(\mathrm{RR}_i > \overline{\mathrm{RR}_i} = \frac{1}{N}\mathbf{1}_{\mathrm{RR}_i > \overline{\mathrm{RR}_i}}$. It is thus clear that it is much easier to exceed the average RR than to fall short of it, which is counterintuitive (but the good thing is that the high risk areas are found more easily). Therefore we recommend calculating these probabilities based on RR's. The median rate was used instead of mean rate in [190].

*Edge Effects*

Considering edge effects is important in stationary point pattern modelling using restricted sample regions of spatially continuous processes [118]. However, edge effects have not had much concern in areal level disease mapping. In part this is because we are usually modelling non-stationary patterns for data which exist only in a certain restricted area. Edge effects were considered in the book [164] for various models. The suggestions included downweighting the boundary areas or using guard areas. In the BYM model, border regions become naturally downweighted as they usually have less neighbours. The effect is that the estimated variance in the border regions is larger than in the central regions. When the restriction is by some administrative boundary, e.g., national border, it might be so that there are no observations in the outer regions which could be used as guard areas. Another suggestion in [164] was to use data augmentation to create "data" into the guard areas, but this seems rather artificial. As a recent example, edge effects were mentioned but not corrected for in [59].

### 4.2.2 Semi-parametric Partition and Cluster Models

*Models for General Clustering*

As discussed earlier, the partition models for general clustering aim to provide a more flexible model to account for discontinuities and regional differences in the geographic variation of a disease. Partition models based on *Voronoi tesselation* [91] and RjMCMC sampling were first discussed in [89]. Further development concerning spatially continuous marked point processes was done in e.g. [103]. The partition model in [143] was first to consider discrete areal level data in an irregular space. In the Voronoi tesselation, a set of cluster centres are first determined. These centres may be geographic points when using point referenced data, or discrete, non-overlapping regions when using areal data (as we do here). The rest of the areas are then assigned to the clusters so that each area belongs to the cluster centre to which it is closest. Ties are handled so that the cluster centre which is first in the list of cluster centres wins. In the disease mapping models, the RR is constant within each cluster. By generating various configurations with the MCMC algorithm, we usually base the model results on the model average (or median) over these configurations. It is not likely that there would exist only a single plausible model configuration.

The partition models are based on MCMC methods, and there are four main types of attempted updates [143, 54]:

1. *Birth* of a new cluster centre.

2. *Death* of a cluster centre.

3. *Shift* of an existing cluster centre. Also we may optionally *switch* two cluster centres in order to break possible ties in the cluster assignment faster.

4. *Update* other model parameters, e.g. RR's.

In the model of [143],[18] a log-Gaussian model is used for the Poisson rates, with diffuse hyperprior for the mean and a vague gamma prior for the precision. With a birth step, RR is proposed from the normal approximation of the FC and the RjMCMC change of variables is thus avoided. The distances are measured as the number of regions that have to be passed when going from region $A$ to region $B$. The number of clusters $k$ is given a truncated geometric distribution, resulting in $p(k+1)/p(k) = (1-c)^k$, where $c$ is suggested to be 0.02. This gives a constant penalty for adding one cluster in the model. On the other hand, conjugate gamma/Poisson or beta/binomial models are used in [54]. With the conjugate model, RR's can be marginalised out when calculating the probabilities of old and new tesselation configurations. Also, the RR's can be Gibbs-sampled from their gamma (or beta) FC's. The cited model uses the Euclidean distance measure.

Other partition models which have been studied for general clustering in disease mapping include [66] and [90], the latter using the Potts model. Potts model has found use in (spatial) population genetics [188, 70], which has its own rich field of Bayesian clustering models. Some more recent developments in disease cluster mapping include product partition models [230, 102]. The latter model is somewhat similar to those of [143, 54], and uses the gamma/Poisson conjugate prior structure.

[18] A programme is freely available at http://www.stat.uni-muenchen.de/sfb386/software/bdcd/index.html. Accessed September 7, 2011.

*Models for Specific Clusters*

According to the earlier discussion, another type of aim in cluster modelling is to locate specific clusters of an excess of disease cases. This kind of scenario is plausible if we may assume a stationary background risk of a disease and some regions where the disease risk factors have accumulated. One example of this kind of phenomenon is a sudden environmental hazard. The model in [74] is based on Markov connected component field priors. Their later model [75] is based on background level baseline risk and small circular clusters concentrated on cluster centres.

### 4.2.3 Gaussian Process Regression Models, a.k.a. Kriging Models

In the geographic point process field, geostatistical kriging models have a long history. Later, they were introduced in the machine learning community as Gaussian process models. Bayesian model based geostatistics was introduced in [58]. The problem with these models is the need to invert the covariance matrix at each iteration, an operation which is $O(n^3)$. Approximate Gaussian process regression models were presented in an unifying view in [229]. The partially independent training conditional (PITC) method produced rather good results, with savings in computational time. Naturally the approximation gets better when it is less sparse, so there is a tradeoff between speed and accuracy. Sparse log-Gaussian processes were recently applied in epidemiology in [288].

### 4.2.4 Adaptive Binned Kernel Estimators

Various adaptive binned kernel estimation models have been used for smoothing in disease mapping [224, 290] and geochemical interpolation [94] contexts. For example, the Alkemia [94] interpolation method developed at Geological Survey Finland uses a first order Butterworth kernel function $\frac{1}{1+(d/d_0)^2}$ as weights for *weighted recursive median* smoothing. $d$ is the Euclidean distance between two points and $d_0$ is the half-distance, i.e. distance where the function has dropped from 1.0 to 0.5. Weights at distances larger than a prespecified value $d_{max}$ are set to zero. Besides $d_0$ and $d_{max}$, the user-tunable parametres are: minimum number of points to include and maximum broadening factor of the window. This method has been adapted for disease mapping using weights based on inverse distance and direct population size, ever since [224]

The model described in [290] simply sums up as many neighbouring regions for a point that is needed for obtaining a prespecified number of expected cases. Each region receives equal weight. The breast cancer map in [221] was prepared using this model. Although being fast, the methods have the tendency to produce circular ripple artifacts. This model [290] has been later refined to include only a fraction of the cases in the last added regions as for reducing ripple [A. Vehtari: personal communication]. On both model types, the choice of parameters is subjective and the smoothing uncertainties are not available. Nevertheless, Adaptive Binned Kernel Estimation methods might be one feasible solution for a fast preview of new data.

### 4.2.5 Further Modelling Issues

The Poisson assumption may sometimes be a bad choice. Although this has been considered mainly with CAR models, it applies to other models as well. The rare event assumption may not hold, especially when dealing with case fatality. In those cases using the binomial logistic regression (e.g. [140]) is more appropriate. In a simple model, the cases $y_i$ in a region $i$ would be binomially distributed:
$y_i \sim$ Binom$(n_i, p_i)$, where $n_i$ is the risk population count and logit$(p_i) = \alpha + \lambda_i + \eta_i$. As we note, the risk population counts are used here instead of the expected number of cases ($e_i$ above). It seems that the binomial model is usually avoided, as it is in the general epidemiology. One reason for this might be the fact that logistic regression leads to *odds ratios* instead of the more easily interpreted RR's.

It may also happen that the disease counts are *overdispersed*, i.e., the variance might be larger than the mean, which violates the Poisson assumption. If this is the case, the model uncertainty would be underestimated. This is even more important in the case of ecological regression where the appropriateness of Poisson model should be checked. In the case of the BYM convolution model, the spatially unstructured term would model the overdispersion, but the assumption that the overdistribution is log-normally distributed is just an approximation to the Poisson model. Negative binomial regression would then be more appropriate [295].

CAR models were first presented with the above simple observed vs. expected number of cases, where the age standardisation has been done beforehand. If there is enough data, it might be more appropriate to consider the uncertainty in the age standardisation. The Cox proportional hazards assumption was used in mapping cancer survival [210]. A parametric age-group specific survival model was used in [126] and a similar model has been used in Publications I–IV. This model choice was done because at least in AMI it has been noticed that the assumption of log-linear age effect is not completely valid. This is also true in parkinsonism, as the tables in Publication I clearly show.

An early comparison of disease mapping models was done in [165]. Several disease mapping models were recently compared in [29].

*Uninhabited Areas*

Especially when using fine grid data, there exist a number of "empty" regions which are not inhabited, i.e. have no population at risk. For example, in a regular 1 km×1 km grid over Finland, 34.1% of the land area was inhabited in 1998 [276]. The spatial models which are usually developed using international data on a much coarser aggregation level do not have this concern (as is evident from the published articles).

In the case of the CAR model, there are at least two approaches. The first (e.g. [126]) is to exclude the empty regions, also from the neighbourhood structure. This is feasible, if the proportion of empty regions is not too large. The second approach [154] is to modify the Poisson distribution so that $p(y_i = 0|N_i = 0) = 1$, where $N_i = 0$ indicates the population number. The implementation ([154]) has the problem that logarithm of the population counts $\log(N_{ik})$ is used in the model formulation. The remedy was to use something like $\log(N_{ik}+10^{-5})$ to keep the logarithm defined. In practice this leads to the situation that the relative risks in empty areas become predicted from the CAR prior. The model is still identifiable as the sum to zero constraint and data make the CAR prior proper [26]. The use of zero inflated Poisson (ZIP) models is also a relevant option here, e.g. [233]

## 4.3 Ecological Regression

Naturally there are other options, but in this thesis we have used ecological regression based on the CAR/BYM model. Additional covariate data can be included in the BYM model by modifying the log-linear regression term to be: $\log(\mu_i) = \alpha + \mathbf{Z}_i^{\mathsf{T}}\xi + \lambda_i + \eta_i + \log(N_i)$, where all the other terms are as in the disease mapping model, but $\mathbf{Z}_i$ is the matrix (or vector) of covariate(s) and $\xi$ are the corresponding effect(s). Vague normal priors are given for the covariate effect(s) $\xi \sim \mathcal{N}(0, 10^{-5})$. Ecological regression with errors in covariates was considered in [20].

The ecological regression approach may look trivial, but there are several underlying issues, which we briefly describe here. For a deeper discussion, see e.g. [239]. First, we note that the spatial structure must be accounted for, in case we have an ecological regression model which studies the effect of a covariate (e.g. in the risk of a disease) over some geographic areas $i$. A simple regression model would be: $y_i = \mathbf{X}_i^\mathsf{T}\xi + \varepsilon_i$, where $y_i$ are the aggregated disease counts and $\mathbf{X}_i$ are the group-average values of the covariates. The residual error, $\varepsilon_i$, accounts for the aggregate effects of unmeasured confounders. It would be incorrect to assume that the residuals are independent, because the confounders in nearby regions are likely to be correlated. Omitting the residual correlation would lead to underestimation of the variance in the covariate effects.

The second important point concerns the ecological bias, which we have already discussed above. If we assume the commonly used multiplicative hazards model, we have the dose response for an individual: $f(x) = \exp(\alpha + x^\mathsf{T}\xi)$. The disease rate at *group level*, for the whole group is $\lambda_G = E_G[p(D|\mathbf{X})] = \int_{x \in G} f(x)H(x)\mathrm{d}x$. For example, assuming that the exposures of individuals ($H(x)$) are normally distributed, $H(\mathbf{X}) \sim \mathcal{N}(\mu_G, \mathbf{\Omega}_G^{-1})$, we have $\lambda_G = \exp(\alpha + \mu_G^\mathsf{T}\xi + 0.5\xi^\mathsf{T}\mathbf{\Omega}_G\xi)$. In practice, the attractive approximation $\lambda_G \approx \exp(\alpha + \mu_G^\mathsf{T}\xi)$ is often used, perhaps even without noticing that it is only an approximation.

A third point is that the Poisson assumption that the variance and mean are equal may not hold. The BYM convolution model accounts for possible spatially unstructured extra variability, but assumes the residuals to be log-normally distributed. As we already mentioned, the negative binomial regression model would be a more natural choice for accounting extra Poisson variation [295].

### 4.4 Shared Component Modelling

Shared component modelling was introduced in [142] and it may be seen as a version of spatial factor analysis. One aim in shared component is to pool strength in using data on related diseases. Another aim is to find similarities and dissimilarities in the geographical distribution of related diseases, as we do in Publication II. A disease can also be used as a surrogate for a health risk, as has been done using lung cancer as a surrogate for the prevalence of smoking [50]. The original model in [142] used a symmetric specification in which three spatial clustering components (each based on the cluster model [143]) are used to describe joint and disease specific variation in disease risk. In short, the model is formulated as:

$$\begin{aligned} y_{1i} &\sim \text{Poisson}(e_{1i}\theta_i^{\delta}\phi_{1i}) \\ y_{2i} &\sim \text{Poisson}(e_{2i}\theta_i^{1/\delta}\phi_{2i}), \end{aligned}$$

where $y_{1i}$ and $y_{2i}$ are the number of cases in diseases ("1" and "2") and $e_{1i}$ and $e_{2i}$ are the corresponding expected number of cases. $\theta_i$ is the clustering component shared by both diseases, with the weights $\delta$ and $1/\delta$ allowing the strength of the common component to be different in the two diseases. $\phi_{2i}$ and $\phi_{2i}$ are the disease specific clustering components. A suitable prior is given for $\delta$. In the model, nothing is said about the clustering components, they could as well be formed with BYM convolution priors (as in Publication II), or with any other suitable choice.

Later developments (e.g. [277, 105] propose using a non-symmetrical specification, so that the model for disease "2" becomes $y_{2i} \sim \text{Poisson}(y_{2i}\theta_i^{1/\delta})$. In this formulation, both diseases are assumed to share common variation, and the disease specific part of variation is accounted only for the first disease. The symmetric components can be extracted by simple arithmetic. The non-symmetrical formulation is assumed to produce a more stable model than the symmetric version, because of possible identifiability problems in assigning the variation between the shared and disease specific components. [29] Extension to more than two diseases are also discussed in [105]. The recent article [59] jointly modelled six cancers using three shared components. Issues in mapping two diseases were discussed in [50].

## 4.5 Spatiotemporal Modelling

So far, we have considered only models which assume that the geographic variation does not show any temporal development. As we discussed earlier (Chapter 2), usually this assumption is not plausible. As the people, their lifestyles and environment are under constant change, this change is usually reflected in disease rates. BHM's offer several alternatives to incorporate the temporal aspect in disease mapping. Although several time-space models have existed for a long time (e.g. so-called space-time autoregressive integrated moving average models), spatiotemporal modelling naturally followed in the footsteps of spatial disease mapping models. The variety of models is large, but we try to cover the models of main interest here.

Among the first spatiotemporal considerations in disease mapping was the paper [46]. Then, several approaches were presented for the BYM models. [18] uses multinormally distributed log-linear time-trends, where the trend on an area is conditioned on the time-average spatial RR in that area. [298] suggests using a nested specification, where the spatial random effects in each time period independently have BYM priors. A model which is separable in space and time was used in [141].

Earlier we stated that for model comparison, it is beneficial to provide a general model class which includes the models to be compared as special cases. This kind of approach was used in [140], where the variation in space and time could be divided into separable effects in space and time, and a spatiotemporally inseparable term. Each of these components may be individually included or excluded from the model. BYM priors were used in [140] for each component, but any other type of prior could be used. Later, the model was extended with a temporal lag in covariates, with the claim that the models were estimated in WinBUGS, including the spatially inseparable term. However, no WinBUGS code is included in the publication ([60]) and it remains unclear whether this is actually achievable.

Recently, there has been an increasing interest in spatiotemporal modelling. Here we merely list a few of the novel ideas. MCAR priors were used in [122]. The binomial CAR model with separable spatial and temporal terms was augmented with a mixture of low variation and high variation spatiotemporal residuals in [1]. Smoothing splines were used in [170] and later with the binomial model in [261]. Temporal autoregressive terms of higher order were considered in [172]. The shared component / latent factor framework was used in [238, 284]. Dirichlet process mixtures were used in [153]. The specific cluster model of [75] was extended to the spatiotemporal domain in [310].

## 4.6 Bayesian Age-Period-Cohort Modelling

Cohort or generation effects in mortality were first considered in the 1920's and 1930's, e.g., in tuberculosis mortality [5]. Slowly this observation has led to so-called age-period-cohort (APC) models in studying the time-trends of mortality and morbidity [113]. The APC models assume that three effects affect the trends simultaneously:

- The age effect captures the natural development of mortality or disease morbidity by age.
- The period effect reflects events that affect all people at a certain time point.
- The cohort effect reflects events that affect a certain generation, e.g., a birth cohort.

As an example, we formulate the logistic binomial APC model as[19]

$$Y_{pa} \sim \text{Binom}(N_{pa}, p_{pa}) \tag{4.1}$$

$$\log\left(\frac{p_{pa}}{1-p_{pa}}\right) = \theta_a + \phi_p + \psi_c, \tag{4.2}$$

where $Y_{pa}$ is the number of deaths and $N_{pa}$ is the population size of age group $a$ during the period $p$. Because of the linear dependency $C_{(pa)} = A - a + p$, the linear trend is identifiable only in 2 out of 3 effects. This is the well known unidentifiability problem [209]. Bayesian APC models [21] with conditional autoregressive (CAR) first or second order random walk smoothing priors and additional constraints in the parameter estimates have provided one elegant solution to make the model parameters identifiable.[20]

### 4.6.1 Conditional autoregressive random walk smoothing priors

The derivation of first or second order random walk smoothing priors is shown in detail in Publication V. Here we only note that the second order

[19] Another common option is using a log-linear Poisson model.

[20] A software package, BAMP, is freely available at http://volkerschmid.de/bamp/. Accessed September 7, 2011.

random walk prior can also be derived as the symmetric form [25, 169]:

$$\begin{aligned} p(\theta_1) = p(\theta_I) &\propto 1 \\ p(\theta_i | \theta_{i-1}, \theta_{i+1}) &\propto \mathrm{N}\left(\tfrac{\theta_{i-1}+\theta_{i+1}}{2}, 4\tau\right), \text{ for } i = \{2, \ldots, I-1\} \end{aligned} \tag{4.3}$$

$$\Leftrightarrow p(\boldsymbol{\theta}|\tau) \propto \tau^{(I-2)/2} \exp\left\{-\frac{\tau}{2}\sum_{i=2}^{I-1}(\theta_{i-1} - 2\theta_i + \theta_{i+1})^2\right\} = \tau^{(I-2)/2} \exp\left\{-\frac{\tau}{2}\boldsymbol{\theta}'\mathbf{K}\boldsymbol{\theta}\right\}. \tag{4.4}$$

As with the asymmetric form, we obtain the MRF structure matrix of the RW2 prior after expanding the square to the quadratic form:

$$\mathbf{K} = \begin{bmatrix} 1 & -2 & 1 & & & & & \\ -2 & 5 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{bmatrix}. \tag{4.5}$$

### 4.6.2 Autoregressive Integrated Moving Average Models

The Box-Jenkins approach to time series modelling frequently employs autoregressive integrated moving average models (ARIMA) [32]. We write the general ARIMA($p, d, q$) model as

$$Y^*_t = \theta_1 Y^*_{t-1} + \cdots + \theta_p Y^*_{t-p} + \varepsilon_t + \phi_1 \varepsilon_{t-1} + \cdots + \phi_q \varepsilon_{t-q}, \tag{4.6}$$

where $p$ and $q$ are non-negative integers, which respectively define the order of the autoregressive, integrated and moving average parts of the model. $\mathbf{Y}^*$ is the $d^{\text{th}}$ difference of the original time series $\mathbf{Y}$, and $\{\varepsilon_1, \cdots, \varepsilon_t\}$ are the error terms.

An autoregressive model assumes that the current state of a process linearly depends on the $p$ previous states. A moving average model, on the other hand, assumes that the current state of a process linearly depends on the $q$ previous error terms, so that the errors (also called random shocks) are correlated. The autoregressive model is more straightforward in interpretation and easier to fit than the moving average model. Depending on the application and data, both models can be used at the same time. If the time series is not stationary (i.e., the joint probability distribution changes when the process is shifted in time), differencing (i.e., using the "integrated" part of the ARIMA model) can be used to make the process stationary. The full ARIMA model is usually estimated using state space methods [61].

In this work we have only used the integrated autoregressive part of the full ARIMA model. This can be estimated using Bayesian linear regression, as detailed in Publication V.

# 5. Study Aims

The aims of this study were:

1. To examine the geographic variation in Parkinson's disease incidence and prevalence using an existing conditional autoregressive spatial smoothing model (Publication I)

2. To examine the joint and disease specific geographic variation in stroke and AMI incidence using a shared component model (Publication II)

3. To examine the role of geographically varying mineral composition of drinking water in AMI incidence and create a Bayesian method for interpolating geochemical data with censored observations (Publications III-IV)

4. To create versatile extensions for the Bayesian APC models in order to analyse and predict long time series of observed mortality and morbidity (Publication V)

# 6. Materials and Methods

## 6.1 Georeferenced Data in Finland

Georeferenced data is usually delivered in Finnish "Kartastokoordinaattijärjestelmä" (KKJ, map coordinate system), which consists of several detached strips.[1] As such, the KKJ data is not usable in modelling, but it must be transformed into a projection which treats Finland as a contiguous region. In all studies, the coordinates were transformed into Finnish "Yhtenäiskoordinaatisto" (YKJ, common coordinate system) as detailed in [155]. YKJ is a national projection system which produces very low distortion of scale across the country [155]. In practice, we treated all data as if the scale was not distorted at all.

One topic which has not received much attention is the accuracy and repeatability of geocoding. Quality seems to vary between different vendors of commercial geocoding [305]. In Finland, the national Population Register Centre has the geographic centre coordinates of each building available, and this information can be linked to every person. At least in the cities, the accuracy of the coordinates is about 20m [222], and it is therefore of no relevance in the ecological studies, where the grid size is typically 10km × 10km. We may also note that the typical everyday neighbourhood radius of a person is about 1km [235], which suggests that it is generally of no use to consider more accurate grid levels in ecological studies. The proportion of missing georeferenced data in the official statistics was 1% in 2000, thus only few people in Finland lack an official address. [193] However, the official address does not necessarily represent

[1] Since 2010, ETRS89 will become the new standard in Finland. http://www.maanmittauslaitos.fi/tiedotteet/2010/05/maanmittauslaitos-vaihtoi-etrs89-koordinaattijarjestelmaan. Accessed 7 September, 2011.

the actual place of residence. Also, it is plausible to assume that missing addresses are more common in the lower socioeconomic groups, i.e. those persons who usually have a higher disease risk.

## 6.2 Georeferenced Data Sets

### 6.2.1 Medicated Parkinsonism, in Men and Women, 1995–2000

This data set was constructed for the study in Publication I, which gives the details. The original aim was to form a data set on idiopathic PD, but it was not possible using the data available to us. Hence, we settled for a wider class of medicated parkinsonism. The data is based on two "semi-independent" sources of The Social Insurance Institution of Finland:

1. Registry of patients entitled to reimbursed medication of PD or parkinsonism (Reimbursement code 110).

2. Registry of prescribed medicine buys. A buy of at least one PD-specific drug was required; the drugs are listed in Publication I.

In addition, patients were restricted to those who were 30 years or over at the time of diagnosis. This restriction was chosen to decrease patients with extrapyramidal symptoms due to non-PD causes. Coordinates were available for >98% of the cases.

### 6.2.2 Stroke and AMI in Men and Women, 1991–2003

This data set was used in Publication II (ischaemic strokes and AMI) and in Publication IV (AMI).

All incident and recurrent AMI and Stroke events were collected in the Finnish National Cardiovascular Disease Register (CVDR) [158]. This register is constructed by a nationwide record linkage of HILMO, National Causes of Death Statistics and the drug reimbursement and prescribed medicine purchase registers of The Social Insurance Institution of Finland.

The incident cases were defined as those for whom there were no similar events in the preceding seven years [253]. We note that AMI in this data set is not directly comparable with the earlier data set (which is described below) because it had a broader definition of cases and a slightly biased definition of incident cases. The principles of case definitions are given in the project website.[2] The exact case definitions for this study, as given in Publication II, are listed below. ICD-9 and *International Classification of Diseases, Tenth Revision* (ICD-10) codes were used in both sources.

*AMI Events*

Non-fatal AMI events were identified from HILMO using the codes I21-I22(ICD-10) / 410(ICD-9) as the main or an additional diagnosis. Fatal AMI events were identified from the National Causes of Death Statistics using the diagnosis codes I20-I25, I46, R96, R98 (ICD-10) / 410-414, 798 except 7980A (ICD-9) as the underlying or immediate cause of death. Codes I21-I22(ICD-10) / 410(ICD-9) were also accepted as a contributing cause of death.

*Stroke Events*

Ischaemic stroke events were identified using the diagnosis codes I63-I64 except I63.6 (ICD-10) / 4330A, 4331A, 4339A, 4340A, 4341A, 4349A, 436 (ICD-9) in both HILMO and National Causes of Death Statistics.

*Coordinate Data*

For each case, the exact place of residence coordinates were obtained corresponding to the event date. Unexpectedly, the proportion of missing coordinates was much higher than in the previous AMI data, although the data should be more complete in more recent years. The proportion of missing coordinates was systematically biased towards higher age, earlier years and women, with a maximum of 10% for a single year/age/gender group.

The probable explanation is that there were some details in the data merging process that were not taken into account in the construction of this latter data set, despite multiple requests to the data provider. As the risk population data sets were provided earlier, this probably created a downward bias in the estimated disease rates. Therefore, some compensation was done as follows.

As described in Publication II, a part of the persons with missing coordinates could be assigned to their coordinates from a previous event—because of the incident case definition and two followed events, each person could have several events in the data set. The rest of the cases with

[2] http://www.ktl.fi/portal/7137. Accessed September 7, 2011.

missing coordinates were assigned into the data as fractional observations spread into the municipality of residence (this was available for everyone) weighted with the age-group specific population counts in each of the grid cell belonging to the municipality (see Publication II). This can be seen as a conservative approach, as each of the cases is spread out into a whole commune instead of assigning it randomly to a specific grid cell. Creating multiple imputed data sets [79] would probably have had a similar effect, but the implementation would have been unnecessarily difficult.

### 6.2.3 AMI in Men, in Years 1983, 1988 and 1993

This earlier data set on AMI was used in Publication III. The data construction is detailed in [126]. Both fatal and non-fatal cases of AMI were defined as any of the *International Classification of Diseases, Eighth Revision* (ICD-8) or ICD-9 codes 410–414 in HILMO or National Causes of Death Statistics in the years 1983, 1988 or 1993. As is clear from the ICD-9 codes, the data set in fact represents the broader category of ischaemic heart disease (IHD). However, we denote this as a data set on AMI, as was done in the original publications [126, 154] and in Publication III. Incident cases were confirmed by tracing back any earlier AMI events. As the HILMO and National Causes of Death Statistics data are available only after 1968, this forms a bias in the incident cases, as cases in the later years have been followed up for a longer time. Persons aged 35–74 years at the time of event were included in the data. For each case, the exact place of residence coordinates were obtained corresponding to the event date. The proportion of missing coordinates was low, about 3% on average over the cross section years.

### 6.2.4 Georeferenced Population at Risk

In Publications I–IV the population data sets were obtained from the National Population Registration Centre, covering years 1983, 1988, 1993, 1998, 2000 and 2002. The spatial resolution was at 1km × 1km regular grid. Population counts were available at the end of each year in question, for ages 0–74. Therefore, the risk populations were in part interpolated and extrapolated from available data. Although this led to some loss of information, it was considered a better option than the loss of money which the high costs of obtaining the additional data would have caused.

Interpolation and extrapolation were done *assuming* linear age-cohort trends, and negligible migration. The population counts in older age groups were estimated from the population counts of 75+ years old which were available for each grid cell. The proportion of people in the age groups 75–79, 80–84 and 85+ years were available at municipal level and these proportions were used as weights to assign the counts of 75+ years old people in each grid cell to the corresponding age groups. The estimated population counts in five-year age groups for the whole country were compared with the accurate counts available from Statistics Finland, showing very high accuracy.

Despite the high accuracy at countrywide level, the estimated counts at areas of low population density could have considerable random errors; see [234] for a related analysis and discussion. The probable effect is that some structured variation in the estimated maps could change into unstructured variation.

### 6.2.5 Urban and Rural Areas

This data set was used in Publications III–IV. The data is based on several reports primarily aimed for rural regional development policy (e.g. [134]). A detailed description of the principles behind urban and rural division was given in [248], which we outline here. The classification was available at the municipality level and it is based on the situation in 1993, at which time there were 455 municipalities in Finland. The municipalities were classified into four possible categories.

1. *Urban areas* are characterized with dense population and high share of secondary and tertiary sector activities. Population of built-up areas must exceed 15,000 inhabitants.

2. *Urban-adjacent rural areas* are mainly located in the western and southern Finland. Over 50% of the total population live in an area from which more than 20% of the work force is commuting to an urban area.

3. *Rural heartland areas* are either dominated by strong primary production or have achieved functional diversification. Most often large city centres are relatively distant to people living in the rural heartland. Over 50% of active farms are situated here.

4. *Remote/isolated areas* have surmounting problems. The share of pri-

mary activities is high, farming has low intensity and profitability, the population density is low, outmigration is high, and the population structure is skewed.

For the analyses, all areas except urban areas were jointly considered as rural areas. The division into urban and rural areas was applied at 10 km × 10 km grid level by assigning each grid cell the classification of the municipality which covered the major part of the cell area.

### 6.2.6 Geochemical Data

Point-referenced data on mineral constituents in drinking water originated from Geological Survey Finland as research co-operation. Concentrations of magnesium (Mg) and calcium (Ca) were among the available measurements (others are mentioned in Publication IV). The data originated mainly from Geological Survey Finland's "Thousand wells" survey [159], which was conducted in 1999 to obtain information on the physical-chemical quality of Finnish household well waters in the sparsely populated rural areas across the country. The data included samples from springs, shallow dug wells and wells dug into bedrock. It has been estimated that over 1,000,000 Finns use private well water in their households. [159]

Additional data samples were obtained later from natural springs etc. In an earlier study [154], the data were interpolated into a regular 10x10 km$^2$ grid by the Alkemia smooth interpolation method [94]. For present studies, however, we developed a Bayesian smooth interpolation method in Publications III–IV. One reason for this was that there were several nondetects in the data. Another reason was that we wanted the level of smoothing be dictated by data, not by any arbitrary choices of the researcher. A technical description of the interpolation method is given below.

## 6.3 Disease Mapping Using the iCAR Model

The disease mapping model used in Publication I followed the approach presented in [126]. As mentioned in the review of statistical methods (Chapter 4), we use a parametric model for age-group effects, which accounts for the uncertainty in age standardisation.

The hierarchical model starts with Poisson rates $\mu_{ik}$ for cases $Y$ in each grid cell $i$ and age group $k$: $Y_{ik} \sim \text{Poisson}(\mu_{ik})$. The Poisson rates are modelled by a log-linear regression model: $\log(\mu_{ik}) = \alpha + \beta_k + \lambda_i + \eta_i + \log(N_{ik})$, where $\alpha$ is the baseline level, $\beta_k$ are the age group effects (with $\beta_1 \equiv 0$ for identifiability) and $N_{ik}$ is the risk population.

The CAR prior is assigned to $\lambda_i$, which is the spatially structured random effect: $\lambda_i \sim \mathcal{N}(\overline{\lambda_{j\sim i}}, m_i\tau_\lambda)$, where $\overline{\lambda_{j\sim i}}$ is the mean of $\lambda$ in the local neighbourhood of grid cell $i$, $m_i$ is the number of neighbours $i$ has, and $\tau_\lambda$ is the spatial precision. As an identifiability constraint, $\sum_i \lambda_i \equiv 0$. Optionally, when using the BYM convolution model, $\eta_i$ is the spatially unstructured random effect: $\eta_i \sim \mathcal{N}(0, \tau_\eta)$. $\tau$'s are given vague Gamma priors: $\tau_\lambda, \tau_\eta \sim \text{Gamma}(0.01, 0.01)$ is a usual choice. A flat prior is assigned for the baseline : $\alpha \propto 1$. The age group effects are assigned vague Normal priors: $\beta_k \sim \mathcal{N}(0, 10^{-5}); \quad k \in \{2 \ldots K\}$.

## 6.4 Ecological Regression Using the iCAR Model

Following the approach presented in [126], additional covariate data can be included in the model by modifying the log-linear regression term to be: $\log(\mu_{ik}) = \alpha + \beta_k + \mathbf{X}_i^{\mathsf{T}}\xi + \lambda_i + \eta_i + \log(N_{ik})$, where all the other terms are as above, but $\mathbf{X}_i$ is the matrix of covariates and $\xi$ are the effects of the covariates. Vague Normal priors are given for the covariate effect(s) $\xi \sim \mathcal{N}(0, 10^{-5})$. This model was used in Publications III–IV for the associations of drinking water constituents and AMI and in Publication I for the effect between urban/rural areas.

## 6.5 A Shared Component iCAR Model

The model used in Publication II follows the symmetric specification [142, 277] of the shared component model. We extended the model to include age group specific effects as follows. The number of observed cases $Y_{dik}$ in area $i$ and age group $k = 1, \ldots, K$ in diseases $d = 1, 2$ were modelled with Poisson rates:

$$Y_{1ik} \sim \text{Poisson}(\mu_{1ik})$$
$$Y_{2ik} \sim \text{Poisson}(\mu_{2ik}),$$

and log-linear models were used for Poisson rates:

$$\mu_{1ik} = \exp(\alpha_1 + \beta_{1k} + \kappa_{0i}^{\delta_1} + \kappa_{1i} + \log(N_{1ik}))$$
$$\mu_{2ik} = \exp(\alpha_2 + \beta_{2k} + \kappa_{0i}^{\delta_2} + \kappa_{2i} + \log(N_{2ik})).$$

The baseline risks were assigned flat priors, $\alpha_d \propto 1$. Vague normal priors were assigned to age group effects, $\beta_{dk} \sim \mathcal{N}(0, 10^{-5})$, except $\beta_{d1} \equiv 0$ as identifiability constraints. $\kappa_{0i}$ was a BYM convolution prior for the shared variation and $\kappa_{di}$ were BYM convolution priors for disease specific variation. $\delta_d$ allowed the strength of shared variation to be different in each disease, and it was assigned somewhat informative prior, $\log(\delta_1) \sim \mathcal{N}(0, 5.9)$. This corresponded to assuming *a priori* a 95% probability that the proportion$\delta_1/\delta_2$ was between 1/5 and 5. $N_{dik}$ is the risk population. In the case of two diseases, $N_{1ik} \equiv N_{2ik}$ is usually used, but if the disease rates in men and women are compared, the risk populations are naturally different. The WinBUGS code of the model is given in Appendix B.1.

## 6.6 Interpolation of Geochemical Data Using the iCAR Model

The interpolation models in Publications III–IV were used for spatially aligning the geochemical point level observations with the areal level AMI data. Our initial aim was to develop an interpolation model which would account for the non-detected observations (observations where the concentration is below the detection limit) and for the measurement uncertainty. We note that our interpolation models are not exact in the sense that the interpolated surface does not necessarily go through the observed data points. With data aggregated to areal level this is not even possible: there may be several observations within a certain area.

### 6.6.1 Interpolation Model in Publication III

In this Publication a rather complex interpolation model was used. The WinBUGS code of the model is given in Appendix B.2.

### 6.6.2 Interpolation Model in Publication IV

In this Publication, a much simpler model was used for interpolation. The model algorithm in WinBUGS is given in the Appendix A of Publication IV. The model assumes a lognormal distribution of observations $x_{ik}$

within each region $i$: $x_{ik} \sim \mathcal{LN}(\mu_i, \tau_{reg})$, where $k(i)$ is the index of observation in region $i$, and $\tau_{reg}$ is the precision which is common to all regions. Observations below the detection limit ($DL$) are modelled as arising from the truncated distribution $x_{ik} \sim \mathcal{LN}(\mu_i, \tau_{region})|(0, DL)$. In this model we use $x_{ik} = DL/2$ as a "pseudo-observation", which in retrospect was noticed to be unnecessary; using "NA" would be the proper choice. Note that the corrected likelihood based model is given in Appendix B.3.

The interpolations in each area, $\mu_i$, were modelled with the CAR model: $mu_i = \alpha + \lambda_i$, where $\alpha$ is the baseline and $\lambda_i$ has the iCAR prior as described above (Section 6.3). The precision parameters $\tau_{reg}$ and $\tau_\lambda$ were given the usual Gamma$(0.01, 0.01)$ priors, and $\alpha \propto 1$. The interpolated observations were calculated as $\hat{x}_i = \exp(\mu_i)$.

## 6.7 The Bayesian Age-Period-Cohort Model (with Extensions)

The models are described in detail in Publication V.

# 7. Results

As the epidemiological results are already given on the original papers, we only wrap up the most important results here. However, the geochemical interpolation model was only briefly presented in the Publications III–IV, so we give it a further treatment here.

## 7.1 Medicated Parkinsonism

The results in Publication I suggest that there exists a belt of higher incidence and prevalence of medicated parkinsonism, passing across central Finland. There is also a sporadic area of excess number of cases in Southern Ostrobothnia. Based on different medicine buying patterns, the elevated risk in Kuopio and North Karelia regions (marked in the maps in Publication I as "2", "3") was the most consistent finding. Incidence rate grew steadily, until peaking at the age of 75–79 years. Prevalence peaked a bit later, at the age of 80-84 years. There was strong evidence for a male excess both in incidence and prevalence, the RR being around 1.5. As discussed in Publication I, microtubule associated protein tau haplotype 1 homozygosity is not a likely risk factor in the geographic variation of medicated parkinsonism.

## 7.2 Ischaemic Stroke and AMI

The main findings in Publication II were:

1. There is strong evidence for geographical variation in ischaemic stroke incidence.

2. There is strong evidence for geographical variation in AMI incidence,

and the variation pattern was rather consistent with previous findings (e.g. [126]).

3. The variation patterns of ischaemic stroke and AMI were quite similar in men and women. Figure 7.1 shows the shared and gender specific AMI risk using a shared component model for AMI incidence only. Figure 7.2 shows the shared and gender specific ischaemic stroke risk using a shared component model for ischaemic stroke incidence only. The data for these models was the same is in Publication II. Because of the similarity, men and women were pooled in consequent analyses. However, the age group effects in men and women had to be considered separately, in both diseases.

4. The variation patterns of ischaemic stroke and AMI have considerable independent components, although these diseases share a common atherosclerotic background.

5. The traditional east vs. west difference in CVD incidence rates still exist. The male excess is much higher in AMI but there is also male excess in ischaemic stroke incidence rate.

In addition, the geographic variation in haemorrhagic stroke was studied, but no clear patterns were found. In part this is due to small case numbers. Therefore these results were not published.

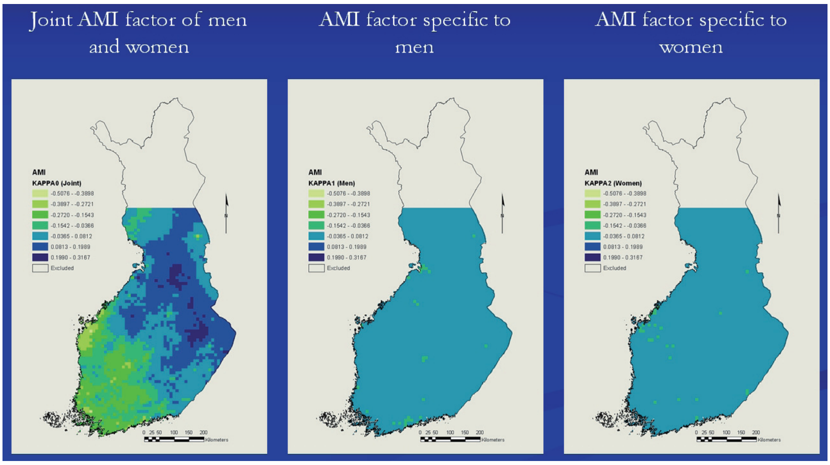


**Figure 7.1.** The shared component model for AMI incidence. Relative risk is shown in log-scale.

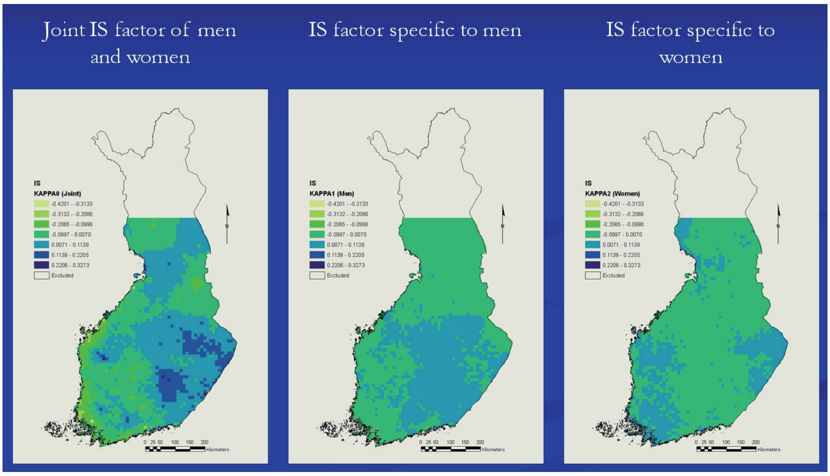


**Figure 7.2.** The shared component model for ischaemic stroke incidence. Relative risk is shown in log-scale.

## 7.3 AMI and Drinking Water

The conclusion of the earlier study (Publication III) was that hard drinking water was associated with reduced AMI risk in men. Especially water poor in Mg (as measured by the Ca:Mg ratio) was associated with increased AMI risk. Later, Publication IV refined the results by using a larger, more recent AMI incidence data set which included women. The results suggest that Mg is the beneficial constituent in hard drinking water.

## 7.4 Interpolation of Geochemical Data

The first model (Publication III) worked well for the elements with a few nondetect values. The magnitude of variance was similar within and between the grid cells. Measurement uncertainty was ignorable when compared to the regional variations. The WinBUGS implementation of the model was rather tricky, with a need to restrict some parameters to be positive. In the later work (Publication IV), there were more elements to interpolate, and a few of them had more than 40 % of the observations as nondetects (Table A1 in Publication IV). It was noticed that in those cases, the original interpolation model did not converge. As a remedy, the interpolation model (Appendix A in Publication IV) was simplified considerably. This model worked well and was very fast to execute. In both papers, the estimated posterior means were taken as the interpolations, hence omitting the posterior uncertainty in the interpolation.

## 7.5 Age-Period-Cohort Models

The results are thoroughly presented and discussed in Publication V.

# 8. Discussion

## 8.1 Epidemiological Studies

### 8.1.1 Medicated Parkinsonism

The clear clustering pattern of excess cases suggests common environmental or genetic risk factors. The genetic risk factors may be related to the settlement history in Finland [201]. On the other hand, some minerals in soil could form one environmental risk factor; manganese (Mn) is one suggested risk factor in PD. The sporadic cluster (region "4" in Publication I) seems to be located at the acid sulphate soil region [69] in the Ostrobothnia. It can also be noted that the MS cluster in Finland [274] is located in the same region. Geological Survey Finland's maps of elements in soil [148] indicate that most of the studied elements have a belt of high concentration passing through the areas that have higher risk of PD. This belt is known as the Raahe-Laatokka ore belt. However, there are high elemental concentrations also around the region of Tampere, which does not support the hypothesis. Preliminary ecological regression (using the model as in Publication IV) did not suggest any geographic association with Mn in the soil and the incidence/prevalence of parkinsonism in Finland.

Although our case ascertainment is slightly stricter than elsewhere (e.g., [117, 115, 116, 114]), there is still a moderate possibility for some regional biases, e.g. due to differences in registration practices. Even if this would be true, the bias can hardly be the sole cause of the considerable geographic variations. The fact that there was no clear geographic variation pattern in the early-onset parkinsonism patients might be in part due to small case numbers.

### 8.1.2 Ischaemic Stroke and AMI

The geographic pattern in AMI incidence still showed the traditional east-/west gradient. Interestingly, the pattern of ischaemic stroke incidence showed some independent variation, although the major part of variation (70%) was common with AMI. This is somewhat in contradiction with the fact that the two diseases with common atherosclerotic background share common risk factors. Possible explanations for the differences would include risk factors that act more strongly on the other disease, e.g. hypertension or excess use of alcohol (c.f. the map in [182, 288]) in stroke. Another explanation could be competing risks [166]: the age specific risks are somewhat different in these two diseases.

The fact that the geographic patterns of AMI and ischaemic stroke were similar in men and women further suggests that there are some common environmental or genetic risk factors underlying the diseases. In a recent study in Finland, the genetic background (as indicated by birthplace) predicted risk of prehospital sudden cardiac death independently of other risk factors in men $\leq$ 54 years of age who had migrated to Helsinki metropolitan area [283]. This gives further support to the role of genetic factors in the east/west CVD gradient, as environmental risk factors of the original birthplace have not accumulated for the whole lifetime. Associations of single nuclear polymorphisms (SNP's) have been found with known CVD risk factors (e.g [131]). Recently, SNP's associated with CVD itself have also been found [174, 254].

The different age group effects in men and women (Table III in Publication III) were in accordance with the common epidemiological knowledge. The risk accumulation in men somewhat seems to slow down with age, because those with the highest risk become selectively removed from the risk population. In contrast, the accumulation in women turns to a higher rate after the menopause. This supports the choice of independent age-group effects (instead of a log-linear age effect) in the models of Publication II.

### 8.1.3 AMI and Drinking Water

In particular, the association of Mg in drinking water (and not Ca) with lower AMI incidence is plausible on two grounds. First, the finding is consistent with physiological effects of Mg. Second, in Finland we generally have an adequate supply of Ca from dairy products.

There exist a lot of uncertainties in this kind of ecological studies. In this study the relevant questions (and some answers) include:

- Who drinks well water? We have restricted the study population to inhabitants in rural areas, who are more likely to have a well of their own. The more rural an area is, the higher percentage is expected to have own wells.

- In this study there was no way to link individuals with a particular well. Therefore it was assumed that on average people drink the kind of water that the region in question has on average.

- Is there such a thing as "average water" in a region? In part this depends on the level of aggregation. This question is also related to uncertainties in interpolating the well water data. The effects of modifiable areal unit problem must also be considered. By choosing a regular grid, the confounding effects of administrative regions can be mostly avoided.

- Is the proportion of Mg obtained from drinking water considerable when compared to other sources? As discussed in Paper IV, about 10% of the daily Mg intake is estimated to be attributable to drinking water. However, water is ingested also in the form of beverages and food, and the same drinking water will probably be used in local production.

- What about migration? It is not reasonable to assume that people live their whole life in one place. Therefore, in a strict sense, the life-long exposures should include the effects of individual migration histories. In Publications III–IV we have assumed that the migration within nearby areas is more common than migration between distant regions. In fact, migration within a municipality is roughly twice as common as migration between municipalities. The latter is directed from rural regions to more urban centres. [152] By pooling statistical power from nearby regions, local migration is in part accounted for. However, further studies should consider the whole migration history of each disease case.

In short, one strength of ecological studies is the relative easiness with which data can be collected. The uncertainties most probably dilute any existing associations.

Similar associations have been ascertained in several ecological studies in AMI and CVD in general. The consensus has been that this association is not generally ascertained in case-control studies (e.g. [191]). This may be in part related to the fact that adjusting for confounders needs to be more strict when we use smaller aggregation levels. However, a recent meta-analysis of case-control studies suggests a significant association of higher magnesium levels in drinking water and reduced CVD risk [41]. It seems that more quantitative studies with careful design are needed before final conclusions can be drawn.

Recent research suggests that Mg intake is inversely related to systemic inflammation and C-reactive protein (CRP), which is associated with increased CVD risk [138, 2]. Similar inverse association is found with Mg intake and METS [266]. In the U.S., Mg intake is much lower than the recommended daily allowance in a considerable part of the population [138].

Perhaps we may note that the most important fact is not whether Mg deficiency is mostly due to drinking water; instead it would be more important to know whether there exists a risk group of people also in Finland with inadequate magnesium supply and whether this is related to increased CVD or METS risk. On the other hand, we must note that the National FINDIET Survey 2007 (e.g. [215] suggest that Ca and Mg intake are both above the recommended levels in men and women, but the intake is slightly lower in the Helsinki/Vantaa and Turku/Loimaa region, compared to North Savo, North Karelia and Oulu. In the FINDIET studies also the urban dwellers are represented, so that the results are not directly comparable with our findings.

### 8.1.4 Mortality in Several European Countries

The study in Publication V is rather descriptive in nature. However, it shows that with proper models the observed data sets can be broken down to a highly detailed level, i.e., even obtain information at 1x1 year resolution in the age/period plane.

The complexity of the underlying phenomena still waits more research to be done for making useful predictions. It should also be noted that the population predictions depend on the predicted mortality and the population predictions obtained elsewhere could be discrepant in view of the predicted mortality. A proper model should therefore include the population predictions.

## 8.2 Statistical Models

### 8.2.1 The Shared Component Model

The shared component model in Paper II proved to be useful in studying the shared and disease specific variation in two related diseases. Our extended implementation with the age group specific effects and the symmetrical specification resulted in somewhat slow converge and moderate autocorrelation. Hence, we used 100,000 iterations with thinning in the estimation. With this complex model, the estimation took one week in WinBUGS. It is evident that as such the model is not suitable for routine use, at least not in WinBUGS.

Using the model to check whether there are any differences in the spatial variation of a disease in men and women was a novel idea, which worked well, and the models also converged fast. It remains to be tested whether there are convergence issues when there exists a difference between men and women. We may also comment that merging the incidence rates of men and women to obtain a common pooled estimate is epidemiologically somewhat meaningless—such a "genderless" population does not exist. However, this approach produced a more simple model to answer our study question.

### 8.2.2 The Geochemical Interpolation Models

The model in Publication III worked initially rather well despite the rather complex approach. There is a clear need for an interpolation model which could account for the possibility of nondetects. In a nonspatial statistical modelling, a similar likelihood based approach (based on censored observations) has been used.[1] Our model also includes the possibility of accounting for measurement errors.

However, because of the model complexity, a simple model was created for Publication IV. The asset of the model is fast estimation; with a 10x10 km$^2$ grid with 6000 observations estimation takes less than one minute in WinBUGS. In retrospect, the implementation (Winbugs Code in Appendix A) has a flaw (as does the implementation in Publication III) . For nondetects, there is no need for the pseudo-observations. Instead, the likelihoods of censored observations could be directly used, which ac-

[1] The "Nondetects and Data Analysis" package in R: http://cran.r-project.org/web/packages/NADA/. Accessed September 12, 2011.

tually was the original model idea. This is shown in the Appendix B.2 here. A spot check showed that the original model will produce a somewhat smaller range of interpolated values than a proper likelihood based approach would. Otherwise, the interpolation results were very similar. The difference in ranges would have some conservative effect in ecological regression models when covariate data contains a noticeable proportion of nondetects. In this study, mainly Ca and Mg were used as covariates in the regression models. The proportion of missing data on these elements was negligible, so the use of pseudo-observations probably had no effect.

### 8.2.3 Age-Period-Cohort Models

In Publication V we have noticed the need of versatile interactions in the age-period-cohort modelling of long time-series of mortality. Previous approaches for mortality models have usually omitted all but the most recent data when making predictions. We have also noticed that the simple conditional autoregressive random walk smoothing priors may be adequate for the observed time series, but they do not have any long-range dependence which would be needed for predictions.

Our experiences so far suggest the ARIMA family of time-series models as a good candidate for age-period-cohort models, but ARIMA models should be also used for the observed time period instead of using them only for the predictions which result in an inconsistent model. Another option would be using Gaussian process priors as the smoothers. Initial work with Jaakko Riihimäki and Dr. Aki Vehtari (Aalto University / BECS) suggests that these models work quite well, but are quite time-consuming to estimate.

## 8.3 Suggestions for Future Research

Here we mainly discuss ideas for further model development, as it is evident that there still are many more interesting epidemiological questions to be explored in Finland. A few applications to mention are model extensions for (the spread of) infectious diseases and disease mapping and other applications in spatial genetics.

As longer time series of georeferenced epidemiological information will become available, the need for spatiotemporal modelling is growing. This is already reflected in the literature. Both in spatiotemporal and purely spatial modelling, there is a need for flexible model classes as it is evident that the conventionally used BYM model has rather strong prior assumptions. Geostatistical models (e.g., [133]) provide more complex covariance structures, and partition/cluster models (e.g., [143, 54]) provide almost limitless flexibility. Another option is the use of multiscale modelling [67]. The flexibility, however, comes at a certain price, namely increased computational complexity. With the more flexible models, we must keep in mind the original aim: the rather strong prior model assumptions of e.g. the BYM models were placed because of the uncertainty of small numbers. If the numbers themselves cannot provide us with the complete picture, we must complete it with our (assumed) prior knowledge. Thus, too flexible models are not appropriate for data on rare disease events—with these models we would fall in the trap of overfitting.

With a growing variety of models at hand, the actual data may not be enough to tell us which model(s) would be appropriate for our application. Also, as we have seen, the DIC is not always useful. Some disease mapping models have been compared using simulated data (e.g., [29]), but more comparisons are clearly needed. Also, there is a need for a carefully designed simulated data sets which would be shared among the researchers.

Many interesting applications of spatial modelling await in the field of population genetics. As more and more genetic studies become available, the genetic data will also include the possibility of studying genetic components in the regional variation of a disease; see, e.g., [250].

Current approaches in spatial modelling use in part some hand made implementations, e.g., in Publications I–IV, a lot of hands-on work has been done. As more and more data and larger data sets become available, better tools are clearly needed for data management and model development. Also, it is clear a that faster, modular software code is needed for implementing and developing the models.

## 8.4 Future Perspectives in Spatial Epidemiology

Historically, ecological studies have had a large impact in epidemiology. Recently, ecological studies have received much criticism, mainly as compared to the more rigorous case/control and prospective cohort study designs, e.g. [257]. This criticism is naturally justified as a reminder of the limitations of ecological studies (e.g., ecological bias, modifiable areal unit problem), and the results are only applicable at an aggregated level, with no possibility to prove any causal pathways. However, right tools must be used in the right place. While we should keep in mind the limitations, ecological studies have many assets, and hence they have remained popular in the epidemiological field.

As spatial aggregate data is usually much more easily available than individual level data, ecological studies may be readily used as preliminary studies in generating etiological hypotheses (cf. Publications I–IV). Ecological studies are well suited for environmental epidemiology—a field of growing importance. Better use of spatiotemporal data and further development of spatiotemporal models will be an important task. Besides surveillance, properly designed spatiotemporal models, for example associated with the age-period cohort approach may be used in forecasting disease rates in the nearby future. Estimating the geographical differences in disease incidence and prevalence provides important information in the functioning of the health care system and in directing the public health resource allocation and research.

Some recent papers have considered careful ecological designs which either could avoid the ecological bias [296] or use a hybrid design with supplementary case/control data [97] to obtain more accurate estimates. However, the view in [257] that the ecological studies should try to reach the standards of case/control studies would effectively reduce if not nullify the original assets of ecological studies. However, in some cases a more thorough approach is warranted, as in [294]. Another thing to consider in disease mapping is adding the information on birthplace, which should usually be readily available, and could provide important background risk information [283]. As already mentioned, we see that spatial genetics will also become a new important field of research.

We must stress that spatial statistics is a field of its own, including much more than the applications in spatial epidemiology which we have discussed. In Publication V we have exploited spatial statistical models in the age-period-cohort model context. As we have seen, spatial statistics has been a major driving force in the development of sophisticated Bayesian computational methods and models. We expect that this favourable contribution will prevail also in future.

# 9. Summary of the Empirical Results and Statistical Models

## 9.1 Empirical Results

In Finland, regional differences in mortality and morbidity have been observed over the last six decades, since the seminal work of Väinö Kannisto [124]. Despite the continuing efforts in epidemiology and health promotion, mortality and morbidity still have regional variation in several diseases/causes. So far, most of the studies on regional health differences have been done at the level of administrative regions whereas the differences in etiological factors do not necessary follow these boundaries. Recent advances in computational technology and statistical methods have enabled us to use a high geographic resolution independent of any administrative boundaries.

In this thesis we have georeferenced morbidity and mortality data from several administrative registers to a 10x10 km$^2$ regular grid over Finland. We have used more recent observations and larger data sets to update the earlier knowledge on AMI incidence and on the incidence and prevalence of parkinsonism. The east/west relative risk (RR) is 1.23 in AMI with a large male excess risk (RR=2.5). Ischaemic stroke shows a pattern similar to AMI, but only 70% of the regional variation is shared with AMI. The east/west difference is lower (RR=1.08) and also the male excess is lower (RR=1.58) in ischaemic stroke. In parkinsonism we have mainly observed a wide belt of excess risk passing across Finland along the borderline of the historical Pähkinäsaari peace treaty. There was strong male excess in the incidence and prevalence of parkinsonism (RR=1.54), but no urban/rural difference.

One of the etiological hypotheses in the regional differences of AMI is the role of drinking water; especially hard drinking water has been suggested as a protective factor. We have further studied the role of drinking water in AMI incidence in the rural areas. Our results suggest that hard water with a low Ca/Mg ratio is associated with lower AMI incidence. However, this alone would explain only some small percentage of the regional variation. Earlier studies suggest that the geographic differences are rather similar within each socioeconomic subgroup. We see the role of genetics as one of the next focus areas in studies of regional differences.

We have also studied all-cause mortality in a long time-perspective of a few hundred years in several European countries. The results show that although there are some country-specific aspects, all the countries have followed the epidemiological transition theory and are now in the fourth era, namely the era of delayed ageing. Studying disease-wise trends and future projections of mortality, incidence and prevalence would provide important information for decision making in the public health sector.

Further studies should also inspect the spatiotemporal patterns of geographic variation in order to assess the long-term stability of the observed differences.

## 9.2 Statistical Models

We have based the research on the conditional autoregressive model of Besag, York and Mollié. From this we have created a smooth interpolation model for geochemical observations with non-detects, but both of the published model versions handle the nondetects in a complicated manner. The proper model (which we suggest in the discussion) would be directly likelihood-based. We have also made a small extension to the conditional autoregressive shared component model of Knorr-Held and Best, by including age effect covariates.

The spatial smoothing model which we have used so far is quite robust, fast and easy to implement (e.g., in WinBUGS), but it assumes that the spatial patterns are similar in each region. Therefore, some region-specific clusters or discontinuities might not be detected. There exist some cluster models which are based on the transdimensional reversible jump MCMC. This will be our next action item, also with attempt to extend these models to the spatiotemporal domain—which has proved to be a complex task.

We have used the one and two-dimensional conditional autoregressive smoothing priors also with the age-period-cohort models. However, in the study it became evident that we need a smoothing model with a longer time-dependence, such as the models in the ARIMA framework.

# A. List of Abbreviations

| | |
|---|---|
| ACS | Acute Coronary Syndrome |
| AMI | Acute Myocardial Infarction |
| APC | Age-period-cohort |
| ARIMA | Autoregressive iterative moving average |
| BMI | Body Mass Index ($weight/height^2$; $kg/m^2$) |
| BRCA1 | Breast Cancer 1, early onset gene |
| BRCA2 | Breast Cancer Type 2 susceptibility protein gene |
| BYM | Besag York and Mollié (model) |
| Ca | Calcium |
| CAR | Conditional Autoregressive |
| CHD | Coronary Heart Disease |
| CRP | C-reactive Protein |
| CV | Cross Validation |
| CVD | Cardiovascular Disease |
| CVDR | Finnish National Cardiovascular Disease Register |
| DAG | Directed Acyclic Graph |
| DIC | Deviance Information Criterion |
| DNA | Deoxyribonucleic Acid |
| ENIAC | Electronic Numerical Integrator And Computer |
| FAQ | Frequently Asked Questions |
| FC | Full Conditional (distribution) |
| FINAMI | The Finnish Myocardial Infarction Register |
| FINSTROKE | The Finnish Stroke Register |
| GMRF | Gaussian Markov Random Field |
| HBM | Hierarchical Bayesian Model |
| HDL | High Density Lipoprotein |
| HILMO | National Hospital Discharge Register |
| iCAR | Intrinsic Conditional Autoregressive |

| | |
|---|---|
| ICD-8 | International Classification of Diseases, Eighth Revision |
| ICD-9 | International Classification of Diseases, Ninth Revision |
| ICD-10 | International Classification of Diseases, Tenth Revision |
| IHD | Ischaemic Heart Disease |
| KKJ | Kartastokoordinaattijärjestelmä (a Finnish Map Coordinate System) |
| MC | Monte Carlo |
| MCMC | Markov Chain Monte Carlo |
| MCAR | Multivariate Conditional Autoregressive |
| METS | Metabolic syndrome |
| Mg | Magnesium |
| Mn | Manganese |
| MONICA | Multinational MONItoring of trends and determinants in CArdiovascular disease |
| MRF | Markov Random Field |
| MS | Multiple Sclerosis |
| PD | Parkinson’s Disease |
| PITC | Partially Independent Training Conditional |
| RIF | Rapid Inquiry Facility |
| RjMCMC | Reversible jump Markov chain Monte Carlo |
| SAR | Simultaneous Autoregression |
| SNP | Single Nuclear Polymorphism |
| T1DM | Type 1 Diabetes (Mellitus) |
| T2DM | Type 2 Diabetes (Mellitus) |
| WAIC | Widely Applicable Information Criterion |
| WinBUGS | Microsoft Windows® version of "Bayesian inference Using Gibbs Sampling" software |
| YKJ | Yhtenäiskoordinaatisto (a Finnish Common Map Coordinate System) |

# B. WinBUGS Code for Selected Models

## B.1 Shared Component Model in Publication II

Besides the basic shared component model, this programme shows: 1) one version for coping with uninhabited regions, 2) how to calculate age-adjusted incidence rate, 3) how to calculate the "properly weighted" age-adjusted incidence rate (as we suggested in the text), and 4) how to calculate the fractions of variances explained by each component.

```
# simple model for pooled diseases, joint model for two diseases
# & another use: study it there's any need to separate models for men & women.
# Aki H., 14.2.06
# using "srrun" weighting to calculate SRmean
# (those weights are disease specific)
# notation: Y1, Y2; N1, N2 etc
# lambda0 = joint
# alpha1, lambda1 = for dis. 1 etc.
# agest= age stardardizing coeffs
# using convolution priors

model;
{
for (j in 1:regions){
for (k in 1:K){
#LIKELIHOODs;
  Y1[j,k]~dpois(mu1[j,k]);
  Y2[j,k]~dpois(mu2[j,k]);
}}


for (j in 1:regions){
for (k in 1:K){
  log(mu1[j,k])<-eta1[j]+beta1[k]+log(N1[j,k]+1.0E-5)+alpha1;
  log(mu2[j,k])<-eta2[j]+beta2[k]+log(N2[j,k]+1.0E-5)+alpha2;
}}

#CAR-distributions;
lambda0[1:regions]~car.normal(map[],w[],Nneighs[],tau0);
lambda1[1:regions]~car.normal(map[],w[],Nneighs[],tau1);
lambda2[1:regions]~car.normal(map[],w[],Nneighs[],tau2);

# including unstructured components
for (i in 1:regions){
  kappa0[i]<-lambda0[i]+uns0[i]
  kappa1[i]<-lambda1[i]+uns1[i]
  kappa2[i]<-lambda2[i]+uns2[i]
  uns0[i]~dnorm(0,tau.uns0)
  uns1[i]~dnorm(0,tau.uns1)
  uns2[i]~dnorm(0,tau.uns2)
  eta1[i]<-kappa0[i]*delta+kappa1[i]
  eta2[i]<-kappa0[i]/delta+kappa2[i]
}

for(k in 1:K) {
  expb1[k]<-exp(beta1[k])
  expb2[k]<-exp(beta2[k])
}
astd1<-inprod(expb1[],agest[])
astd2<-inprod(expb2[],agest[])

for (i in 1:regions){
  # for monitoring age-standardized, dis 1;
  SR1[i]<-exp(kappa0[i]*delta+kappa1[i]+alpha1)*astd1*100000;
  SRrun1[i]<-SR1[i]*srwrun1[i];
  # for monitoring age-standardized, dis 2;
  SR2[i]<-exp(kappa0[i]/delta+kappa2[i]+alpha2)*astd2*100000;
  SRrun2[i]<-SR2[i]*srwrun2[i];
}

srrunmean1<-sum(SRrun1[1:regions])
srrunmean2<-sum(SRrun2[1:regions])

#assigning weights for CAR-distribution;
for (k in 1:neighs){
  w[k]<-1;
}

#CALCULATING P-VALUES;

for (j in 1:regions){
  P0[j]<-step(kappa0[j]);
  P1[j]<-step(kappa1[j]);
  P2[j]<-step(kappa2[j]);
  Psrrun1[j]<-step(SR1[j]-srrunmean1);
  Psrrun2[j]<-step(SR2[j]-srrunmean2);
}

#PRIORS;
beta1[1]<-0
beta2[1]<-0
Pbeta[1]<-step(alpha1-alpha2) # => actually p(alpha)
for(k in 2:K){
  beta1[k]~dnorm(0.0,1.0E-5);
  beta2[k]~dnorm(0.0,1.0E-5);
  Pbeta [k]<-step(beta1[k]-beta2[k])
}
alpha1~dflat();
alpha2~dflat();
tau0~dgamma(.01,.01);
tau1~dgamma(.01,.01);
tau2~dgamma(.01,.01);
sigma0<-1/sqrt(tau0);
sigma1<-1/sqrt(tau1);
sigma2<-1/sqrt(tau2);
tau.uns0~dgamma(.01,.01);
tau.uns1~dgamma(.01,.01);
tau.uns2~dgamma(.01,.01);
sigma.uns0<-1/sqrt(tau.uns0);
sigma.uns1<-1/sqrt(tau.uns1);
sigma.uns2<-1/sqrt(tau.uns2);
# scaling factor for relative strength of shared component for each disease
logdelta ~ dnorm(0, 5.9)
# (prior assumes 95% probability that delta^2 is between 1/5 and 5;
delta <- exp(logdelta)

#summaries
for (i in 1:regions){
  totalRR1[i]<-exp(eta1[i])
  totalRR2[i]<-exp(eta2[i])
  specRR1[i]<-exp(kappa1[i])
  specRR2[i]<-exp(kappa2[i])
  sharedRR[i]<-exp(kappa0[i])
  logsharedRR1[i]<-kappa0[i]*delta
  logsharedRR2[i]<-kappa0[i]/delta
}
var.shared1<-sd(logsharedRR1[])*sd(logsharedRR1[])
var.shared2<-sd(logsharedRR2[])*sd(logsharedRR2[])
var.spec1<-sd(kappa1[])*sd(kappa1[])
var.spec2<-sd(kappa2[])*sd(kappa2[])
frac.shared1<-var.shared1/(var.shared1+var.spec1)
frac.shared2<-var.shared2/(var.shared2+var.spec2)
frac.spec1<-1-frac.shared1
frac.spec2<-1-frac.shared2
}
```

## B.2 Interpolation Model in Publication III

This particular model was programmed for interpolating Mg observations. This model listing should be read along with the description in Publication III. Additional complexity is due to the fact that certain tricks must be used in order to limit the distributions to positive values.

```
# valid observations
for(i in 1:Mg.nvalid)
{
  # assign a common lognormal distribution
  # as with Ca, this actually not used; using dunif for Mg...
  Mg[i]~dlnorm(Mg.mu,Mg.tau.cell)
}
# observations below det.limit of 92, simulation
for(i in 1:Mg.nlo92)
{
# simulate the distributions for values below det. limit of 1992
  # use this if many of the observations belong into this class
  # Lo.Mg[i]~dlnorm(Mg.mu,Mg.tau.cell)I(,Mg.dl92)
  # use this if only a fraction of observations belong into this class
  Lo.Mg[i]~dunif(0,Mg.dl92)
}
#priors
Mg.mu~dnorm(0,1.0E-5)
Mg.tau.cell~dgamma(0.01,0.01)

####
# the spatial model for Mg
####
Mg.dlper<-Mg.dl92/2

#likelihoods
for(i in 1:Mg.nvalid)
{
  Mg.uc.tau[i]<-pow(1/uncert.mg[i],2)
  Mg.vmuu[i]~dnorm(Mg.s.mu[Mg.cell[i]],Mg.uc.tau[i])I(1.0E-6,) # limit this to >0
  L.Mg.vmuu[i]<-log(Mg.vmuu[i])
  L.Mg2[i]<-log(Mg2[i])
  L.Mg2[i]~dnorm(L.Mg.vmuu[i],Mg.sstau[Mg.cell[i]])
}
for(i in 1:Mg.nlo92)
{
  Mg.muu[i]<-Mg.s.mu[Mg.locell92[i]]+cut(Lo.Mg[i])
  L.Mg.muu[i]<-log(Mg.muu[i]*step(Mg.muu[i]-1.0E-7)+1.0E-6) # limit this to >0
  L.Mg.dlper2[i]<-log(2*Mg.dlper) # recentering: add dl/2
  L.Mg.dlper2[i]~dnorm(L.Mg.muu[i],Mg.sstau[Mg.locell92[i]])
}

for(i in 1:regions)
{
  Mg.sstaw[i]<-Mg.s.tau[i]+Mg.tau0
  Mg.sstau[i]<-Mg.sstaw[i]*step(Mg.sstaw[i])+1.0E-6
   log(Mg.s.mu[i])<-Mg.lambda0[i]+Mg.alpha0
  Mg.interp[i]<-(cut(Mg.s.mu[i]))
}

# CAR-distributions
# weights for CAR-distribution were preassigned in the data
Mg.lambda0[1:regions]~car.normal(map[],w[],Nneighs[],Mg.tau)
Mg.s.tau[1:regions]~car.normal(map[],w[],Nneighs[],Mg.s.tau.p)

# spatial priors
Mg.s.tau.p~dgamma(0.01,0.01) # for cell variance
Mg.tau~dgamma(0.01,0.01) # for CAR
Mg.alpha0~dflat()
Mg.tau0~dflat()
}
```

### B.3 Corrected Version of the Interpolation Model in Publication IV

As we can see, this model is much simpler than the above model. Again, this should be read along with the description in Publication IV. This corrected model version uses NA for observations below the detection limit, and therefore exploits the original idea of likelihood based approach for nondetects.

```
##########################################
# "arguments" == data
# Obs = valid observations
# ObsLow = low obs, give NA
# detlim = detection limit
# n.valid = number of valid observations
# grid.valid = grid cell index for each valid observation
#
# n.low = #low observations
# grid.low = grid cell index for each "low" observation
#
# map, Nneighs, regions, neighbours as usual
##########################################

# inits
# tau.car = 1; spatial CAR variance
# tau.cell = 1; spatial in-a-cell variance
# alpha0 =  log("mean baseline concentration")
# lambda0 = rep(0,regions)
# ObsLow[i] <- NA (not detlim/2)

model{
#likelihoods
for(i in 1:n.valid)
{ Obs[i]~dlnorm(spat.mu[grid.valid[i]],tau.cell) }
for(i in 1:n.low)
{ ObsLow[i]~dlnorm(spat.mu[grid.low[i]],tau.cell)I(0.00001,detlim) }
for(i in 1:regions)
{
  spat.mu[i]<-lambda0[i]+alpha0
  interp[i]<-exp(spat.mu[i])
}
# CAR-distribution
lambda0[1:regions]~car.normal(map[],w[],Nneighs[],tau.car)

# assigning weights for CAR-distribution;
for (k in 1:neighbours) { w[k]<-1 }

# spatial priors
  tau.car~dgamma(0.01,0.01) # for CAR
  sigma.car<-1/sqrt(tau.car)
  tau.cell~dgamma(0.01,0.01) # for cell
  sigma.cell<-1/sqrt(tau.cell)
  alpha0~dflat()
}
```

# Bibliography

[1] J. J. Abellan, S. Richardson, and N. Best. Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives*, 116(8):1111–1119, 2008.

[2] D. Almoznino-Sarafian, S. Berman, A. Mor, M. Shteinshnaider, O. Gorelik, I. Tzur, I. Alon, D. Modai, and N. Cohen. Magnesium and C-reactive protein in heart failure: an anti-inflammatory effect of magnesium administration. *European Journal of Nutrition*, 46:230–237, 2007.

[3] C. G. Amrhein. Searching for the elusive aggregation effect: evidence from statistical simulations. *Environment and Planning A*, 27:105–119, 1995.

[4] H. L. Anderson. Metropolis, Monte Carlo and the MANIAC. *Los Alamos Science*, 14:96–108, 1986.

[5] K. F. Andvord. Continued studies of tuberculosis considered as a generation illness. (1932). *International Journal of Epidemiology*, 37:917–922, 2008.

[6] S. Arber and H. Cooper. Gender differences in health in later life: a new paradox? *Social Science & Medicine*, 48:61–76, 1999.

[7] P. Aylin, R. Mahsewaran, J. Wakefield, S. Cockings, L. Jarup, R. Arnold, G. Wheeler, and P. Elliott. A national facility for small area disease mapping and rapid initial assessment of apparent disease clusters around a point source: the UK Small Area Health Statistics Unit. *Journal of Public Health Medicine*, 21:289–298, 1999.

[8] S. Banerjee, B. P. Carlin, and A. Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[9] G. A. Barnard and T. Bayes. Studies in the history of probability and statistics: IX. Thomas Bayes's essay towards solving a problem in the doctrine of chances. *Biometrika*, 45(3/4):293–315, 1958.

[10] F. A. Barrett. Finke's 1792 map of human diseases: the first world disease map? *Social Science & Medicine*, 50:915–920, 2000.

[11] L. Beale, J. Abellan, S. Hodgson, and L. Jarup. Methodological issues and approaches to spatial epidemiology. *Environmental Health Perspectives*, Online 25 April, 2008.

[12] A. C. Belin and M. Westerlund. Parkinson's disease: A genetic perspective. *FEBS Journal*, 275:1377–1383, 2008.

[13] D. R. Bellhouse. A new look at Halley's life table. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 173:Epub ahead of printing, 2011.

[14] D. Belpomme, P. Irigaray, L. Hardell, R. Clapp, L. Montagnier, S. Epstein, and A. J. Sasco. The multitude and diversity of environmental carcinogens. *Environmental Research*, 105:414–429, 2007.

[15] P. Belpomme. Cancer and the environment: facts, figures, methods and misinterpretations. *Biomedicine & Pharmacotherapy*, 61:611–613, 2007.

[16] V. Beneš, K. Bodlák, J. Møller, and R. Waagepetersen. A case study on point process modelling in disease mapping. *Image Analysis & Stereology*, 24:159–168, 2005.

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

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

[19] L. Bernardinelli and C. Montonoli. Empirical Bayes versus fully bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 11:983–1007, 1992.

[20] L. Bernardinelli, C. Pascutto, N. G. Best, and W. G. Gilks. Disease mapping with errors in covariates. *Statistics in Medicine*, 16:741–752, 1997.

[21] C. Berzuini, D. Clayton, and L. Bernardinelli. Disease mapping with errors in covariates. *Bulletin of the International Statistical Institute*, 50:149–164, 1993.

[22] J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36:192–236, 1974.

[23] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 48:259–302, 1986.

[24] J. Besag and P. Green. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 55(1):25–37, 1993.

[25] J. Besag, P. Green, D. Higdon, and K. Mengersen. Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–66, 1995.

[26] J. Besag and C Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 1995.

[27] J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 154:143–155, 1991.

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

[29] N. Best, S. Richardson, and A. Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.

[30] N. G. Best, K. Ickstadt, and R. L. Wolpert. Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, 95(452):1076–1087, 2000.

[31] N. G. Best, K. Ickstadt, R. L. Wolpert, and D. J. Briggs. Combining models of health and exposure data: the SAVIAH study. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: methods and applications*. Oxford University Press, NY, 2000.

[32] G. Box and G. Jenkins. *Time series analysis: Forecasting and control*. Holden-Day, San Francisco, CA, 1970.

[33] A. Brix and P. J. Diggle. Spatiotemporal prediction for log-Gaussian Cox processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4):823–841, 2001.

[34] H. Brody, M. R. Rip, P. Vinten-Johansen, N. Paneth, and S. Rachman. Map making and myth-making in Broad Street: the London cholera epidemic, 1854. *Lancet*, 356:64–68, 2000.

[35] D. Brook. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51(3/4):481–483, 1964.

[36] S. P. Brooks. Markov chain Monte Carlo method and its application. *Statistician*, 47(1):69–100, 1998.

[37] Finnish Cancer Registry. Cancer in Finland 2004 and 2005. http://www.cancerregistry.fi/tilastot/image_101.pdf, Suomen Syöpäyhdistys, Helsinki 2006.

[38] O. Cappé, C. P. Robert, and P. Rydén. Reversible-jump MCMC converging to birth-death MCMC and more general continuous time samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65:679–700, 2003.

[39] F. A. Castro, K. Haimila, K. Pasanen, M. Kaasila, T. Patama, J. Partanen, H-M. Surcel, E. Pukkala, and M. Lehtinen. Geographic distribution of cancer-associated human leucocyte antigens and cervical cancer incidence in Finland. *International Journal of STD & AIDS*, 18:672–679, 2007.

[40] A. Casu, C. Pascutto, L. Bernardinelli, and M. Songini. Type 1 diabetes among Sardinian children is increasing: the Sardinian diabetes register for children aged 0–14 years (1989–1999). *Diabetes Care*, 27:1623–1629, 2004.

[41] L. A. Catling, I. Abubakar, L. Swift, P. R. Hunter, and I. R. Lake. A systematic review of analytical observational studies investigating the association between cardiovascular disease and drinking water hardness. *Journal of Water and Health*, 6(4):433–442, 2008.

[42] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *American Statistician*, 49(4):327–335, 1995.

[43] L. Choo and S. G. Walker. A new approach to investigating spatial variations of disease. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 171(2):395–405, 2008.

[44] D. Clayton and J. Kaldor. Empirical Bayes estimates of age-standardized relative risks for disease mapping. *Biometrics*, 43:671–681, 1987.

[45] A. D. Cliff and J. K. Ord. Model building and the analysis of spatial pattern in human geography (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 37:297–348, 1975.

[46] P. Congdon. Spatiotemporal analysis of area mortality. *Statistician*, 43(4):513–528, 1994.

[47] P. Congdon. Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes. *Computational Statistics and Data Analysis*, 51:3197–3212, 2007.

[48] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, Inc, New York, 1991.

[49] F. T. Cutts, S. Franceschi, S. Goldie, X. Castellsague, S. de Sanjose, G. Garnett, W. J. Edmunds, P. Claeys, K. L. Goldenthal, D. M. Harperi, and L. Markowitz. Human papillomavirus and HPV vaccines: a review. *Bulletin of the World Health Organization*, 85(9):719–726, 2007.

[50] A. R. Dabmey and J. C. Wakefield. Issues in the mapping of two diseases. *Statistical Methods in Medical Research*, 14:83–112, 2005.

[51] S. Darby, D. Hill, A. Auvinen, J. M. Barros-Dios, H. Baysson, F. Bochicchio, H. Deo, R. Falk, F. Forastiere, M. Hakama, I. Heid, L. Kreienbrock, M. Kreutzer, F. Lagarde, I. Mäkeläinen, C. Muirhead, W. Obereigner, G. Pershagen, A. Ruano-Ravina, E. Ruosteenoja, A. Schaffrath-Rosario, M. Tirmarche, L. Tomasek, E. Whitley, H-E. Wichmann, and R. Doll. Radon in homes and lung cancer risk: collaborative analysis of individual data from 13 European case-control studies. *British Medical Journal*, 330:223–226, 2005.

[52] Human Mortality Database. University of California, Berkeley (usa), and Max Planck Institute for Demographic Research (Germany). http://www.mortality.org.

[53] M. C. de Rijk, C. Tzourio, M. M. Breteler, J. F. Dartigues, L. Amaducci, S. Lopez-Pousa, J. M. Manubens-Bertran, A. Alpérovitch, and W. A. Rocca. Prevalence of parkinsonism and Parkinson's disease in Europe: the EUROPARKINSON collaborative study. European Community concerted action on the epidemiology of Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 62(1):10–15, 1997.

[54] D. G. T Denison and C. C Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–149, 2001.

[55] V. P. A. Derrick. Observations on (1) errors of age in the population statistics of England and Wales, and (2) the changes in mortality indicated by the national records. *Journal of the Institute of Actuaries*, 58:117–159, 1927.

[56] International Diabetes Federation. The IDF consensus worldwide definition of the metabolic syndrome. http://www.idf.org/webdata/docs/IDF_Meta_def_final.pdf, Brussels, 2006.

[57] F. D. Dick, G. De Palma, A. Ahmadi, N. W. Scott, G. J. Prescott, J. Bennett, S. Semple, S. Dick, C. Counsell, P. Mozzoni, N. Haites, S. Bezzina Wettinger, A. Mutti, M. Otelea, A. Seaton, P. Söderkvist, and A. Felice. Environmental risk factors for Parkinson's disease and parkinsonism: the Geoparkinson study. *Occupational and Environmental Medicine*, 64:666–672, 2007.

[58] P. J. Diggle, R. A. Moyeed, and J. A. Tawn. Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350, 1998.

[59] A. Downing, D. Forman, M. S. Gilthorpe, K. L. Edwards, and S. O. M. Manda. Joint disease mapping using six cancers in the Yorkshire region of England. *International Journal of Health Geographics*, 7:41, 2008.

[60] E. Dreassi, A. Biggeri, and D. Catelan. Space-time models with time-dependent covariates for the analysis of the temporal lag between socioeconomic factors and lung cancer mortality. *Statistics in Medicine*, 24:1919–1932, 2005.

[61] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, New York, 2001.

[62] A. Earnest, G. Morgan, K. Mengersen, L. Ryan, R. Summerhayes, and J. Beard. Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International Journal of Health Geographics*, 6:54, 2007.

[63] P. Elliott and D. A. Savitz. Design issues in small-area studies of environment and health. *Environmental Health Perspectives*, 116:1098–1104, 2008.

[64] P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors. *Spatial Epidemiology: methods and applications*. Oxford University Press, NY, 2000.

[65] P. Elliott, A. J. Westlake, M. Hills, I. Kleinschmidt, L. Rodrigues, P. McGale, K. Marshall, and G. Rose. The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *Journal of Epidemiology & Community Health*, 46:345–349, 1992.

[66] C. Fernández and P. J. Green. Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):805–826, 2002.

[67] M. A. R Ferreira and H. K. H Le. *Multiscale modeling: a Bayesian perspective*. Springer-Verlag, NY, 2007.

[68] B. Finkelstädt, L. Held, and V. Isham, editors. *Statistical methods for spatio-temporal systems*. Chapman & Hall/CRC, Boca Raton, FL, 2007.

[69] R. M. Fältmarsch, M. E. Åström, and K-M. Vuori. Environmental risks of metals mobilised from acid sulphate soils in Finland: a literature review. *Boreal Environment Research*, 13:444–456, 2008.

[70] O. Franȩis, S. Ancelet, and G. Guillot. Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics*, 174:805–816, 2006.

[71] R. R. Frerichs. History, maps and the internet: UCLA's John Snow site. *Society of Cartographers Bulletin*, 34(2):3–7, 2001.

[72] W. H. Frost. The age selection of mortality from tuberculosis in successive decades (1939). *American Journal of Epidemiology*, 141(1):4–9, 2006.

[73] S. Gamatam, R. Carter, M. Ariet, and G. Mitchell. An empirical comparison of record linkage procedures. *Statistics in Medicine*, 21:1485–1496, 2002.

[74] R. E. Gangnon and M. K. Clayton. Bayesian detection and modeling spatial disease clustering. *Biometrics*, 56:922–935, 2000.

[75] R. E. Gangnon and M. K. Clayton. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*, 22:3213–3228, 2003.

[76] A. C. Gatrell. *Geographies of Health: An Introduction*. Blackwell Publishing, UK, 2002.

[77] A. E. Gelfand and P. Vonatsou. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–25, 2003.

[78] A. Gelman. Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99(466):537–545, 2004.

[79] A. Gelman, J. B. Carlin, H. B. Stern, and D. B. Rubin, editors. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, Second edition, 2004.

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

[81] A. Gelman, Y. Goegebeur, F. Tuerlinckx, and I. Van Mechelen. Diagnostic checks for discrete data regression using posterior predictive simulations. *Journal of Applied Statistics*, 49(2):247–268, 2000.

[82] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[83] C. J. Geyer. Practical Markov chain Monte Carlo. *Statistical Science*, 7(4):473–511, 1992.

[84] W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex Bayesian modelling. *Statistician*, 43(1):169–177, 1994.

[85] H. Gille. Demographic history of the Northern European countries in the eighteenth century. *Population Studies*, 3(1):3–65, 1949.

[86] M. Gissler and J. Haukka. Finnish health and social welfare registers in epidemiological research. *Norsk Epidemiologi*, 14(1):113–120, 2004.

[87] C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

[88] F. P. Greco, A. B. Lawson, D. Cocchi, and T. Temples. Some interpolation estimators in environmental risk assessment for spatially misaligned health data. *Environmental and Ecological Statistics*, 12:379–395, 2005.

[89] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[90] P. J. Green and S. Richardson. Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1–16, 2002.

[91] P. J. Green and R. Sibson. Computing dirichlet tessellations in the plane. *Computer Journal*, 21(2):168–173, 1978.

[92] S. Greenland. Principles of multilevel modelling. *International Journal of Epidemiology*, 29:158–167, 2000.

[93] S. Guha and L. Ryan. Spatio-temporal analysis of areal data and discovery of neighborhood in conditionally autoregressive models. http://www.bepress.com/harvardbiostat/paper61/, Harvard University Biostatistics Working Paper Series, Working Paper 61, 2006.

[94] N. Gustavsson, E. Lampio, and T. Tarvainen. Visualization of geochemical data on maps at the Geological Survey of Finland. *Journal of Geochemical Exploration*, 59:197–207(11), 1997.

[95] E. Halley. An estimate of the degrees of the mortality of the mankind, drawn from curious tables of the births and funerals at the city of Breslaw; with an attempt to ascertain the price of annuities upon lives. *Philosophical transactions of the Royal society of London*, 17:596–610, 1693.

[96] M. E. Halloran and D. Berry, editors. *The IMA volumes in mathematics and its applications: Statistical methods in epidemiology, the environment and clinical trials*, chapter Leroux, B. G., Lei, X. and Breslow, N., Estimation of spatial disease rates in small areas: a new mixed model for spatial dependence. Springer, NY, 2000.

[97] S. Haneuse and J. Wakefield. Geographic-based ecologials correlation studies using supplemental case-control data. *Statistics in Medicine*, 27:864–887, 2008.

[98] K. Harald, S. Koskinen, P. Jousilahti, J. Torppa, E. Vartiainen, and Salomaa V. Changes in traditional risk factors no longer explain time trends in cardiovascular mortality and its socioeconomic differences. *Journal of Epidemiology & Community Health*, 62(3):251–257, 2008.

[99] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[100] J. Haukka, J. Suvisaari, T. Varilo, and J. Lönnqvist. Regional variation in the incidence of schizophrenia in Finland: a study of birth cohorts born from 1950 to 1969. *Psychological Medicine*, 31:1045–1053, 2001.

[101] J. D. Healy. Excess winter mortality in Europe: a cross country analysis identifying key factors. *Journal of Epidemiology & Community Health*, 57:784–789, 2003.

[102] A. Hegarty and D. Barry. Bayesian disease mapping using product partition models. *Statistics in Medicine*, 27:3868–3893, 2008.

[103] J. Heikkinen and E. Arjas. Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25:435–450, 1998.

[104] S. Helakorpi, T. Martelin, Torppa J., E. Vartiainen, A. Uutela, and K. Patja. Impact of the 1976 tobacco control act in Finland on the proportion of ever daily smokers by socioeconomic status. *Preventive Medicine*, 46:340–345, 2007.

[105] L. Held, I. Natário, S. E. Fenton, H. Rue, and N. Becker. Towards joint disease mapping. *Statistical Methods in Medical Research*, 14:61–82, 2005.

[106] L. Heligman and J. H. Pollard. The age pattern of mortality. *Journal of the Institute of Actuaries*, pages 61–82, 1980.

[107] D. Helwig. Medical geography: MDs should pay heed to "airs, waters, places". *Canadian Medical Association Journal*, 139:790–791, 1988.

[108] E. Hemminki and A. Paakkulainen. Auxiliary variable methods for Markov chain Monte Carlo with applications. *American Journal of Public Health*, 66(1180–1184):585–595, 1976.

[109] T. Hägerstrand. What about people in regional science? *Papers in Regional Science*, 24(1):6–21, 1970.

[110] D. M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *Journal of the American Statistical Association*, 93(442):585–595, 1998.

[111] Hippocrates. On Airs, Waters, and Places, 400BC; English translation by Francis Adams. http://ebooks.adelaide.edu.au/h/hippocrates/airs/.

[112] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–417, 1999.

[113] T. R. Holford. The estimation of age, period and cohort effects for vital rates. *Biometrics*, 39:311–324, 1983.

[114] G. Hu, R. Antikainen, P. Jousilahti, M. Kivipelto, and Tuomilehto J. Total cholesterol and the risk of Parkinson disease. *Neurology*, 70:1972–1979, 2008.

[115] G. Hu, S. Bidel, P. Jousilahti, R. Antikainen, and Tuomilehto J. Coffee and tea consumption and the risk of Parkinson's disease. *Movement Disorders*, 22:2242–2248, 2007.

[116] G. Hu, P. Jousilahti, S. Bidel, R. Antikainen, and Tuomilehto J. Type 2 diabetes and the risk of Parkinson's disease. *Diabetes Care*, 30(4):842–847, 2007.

[117] G. Hu, P. Jousilahti, A. Nissinen, R. Antikainen, M. Kivipelto, and Tuomilehto J. Body mass index and the risk of Parkinson disease. *Neurology*, 67:1955–1959, 2006.

[118] J. Illian, A. Penttinen, H. Stoyan, and D. Stoyan. *Statistical analysis and modelling of spatial point patterns*. John Wiley & Sons, Ltd., UK, 2008.

[119] P. Irigaray, J. A. Newby, Clapp. R., L. Hardell, V. Howard, L. Montagnier, S. Epstein, and D. Belpomme. Lifestyle-related factors and environmental agents causing cancer: an overview. *Biomedicine & Pharmacotherapy*, 61:640–658, 2007.

[120] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[121] E. T. Jaynes. Information theory and statistical mechanics II. *Physical Review*, 108:171–190, 1957.

[122] X. Jin and B. P. Carlin. Multivariate parametric spatiotemporal models for county level breast cancer survival data. *Lifetime Data Analysis*, 11:5–27, 2005.

[123] M. Kajantie, K. Manderbacka, A. McCallum, I-L. Notkola, M. Arffman, E. Forssas, S. Karvonen, M. Kortteinen, L. Alastair, and I. Keskitalo. How to carry register-based health services research in Finland? Compiling complex study data in the REDD project. http://www.stakes.fi/verkkojulkaisut/papers/DP1-2006.pdf, Helsinki, 2006.

[124] V. Kannisto. *Kuolemansyyt väestöllisinä tekijöinä (in Finnish, English summary: The causes of death as demographical factors in Finland)*. PhD thesis, Kansantaloudellisia tutkimuksia XV, Helsinki, 1947.

[125] V. Kannisto, O. Turpeinen, and M. Nieminen. Finnish life tables since 1751. *Demographic Research*, 1:Article 1, 2005.

[126] M. Karvonen, E. Moltchanova, M. Viik-Kajander, V. Moltchanov, M. Rytkönen, A. Kousa, and J. Tuomilehto. Regional inequality in the risk of acute myocardial infarction in Finland: a case study of 35- to 74-year old men. *Heart Drug*, 2:51–60, 2002.

[127] M. Karvonen, J. Rusanen, M. Sundberg, E. Virtala, A. Colpaert, A. Naukkarinen, and J. Tuomilehto. Regional differences in the incidence of insulin dependent diabetes mellitus among children in Finland from 1987 to 1991. *Annals of Medicine*, 29:297–304, 1997.

[128] M. J. Karvonen, E. Orma, S. Punsar, V. Kallio, M. Arstila, K. Luomanmäki, and J. Takkunen. Coronary heart disease in seven countries. VI. Five-year experience in Finland. *Circulation*, 41(4S1):I52–I62, 1970.

[129] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal. Markov chain Monte Carlo in practice: a roundtable discussion. *American Statistician*, 52(2):93–100, 1998.

[130] R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1998.

[131] S. Kathiresan, O. Melander, C. Guiducci, A. Surti, N. P. Burtt, M. J. Rieder, G. M. Cooper, C. Roos, B. F. Voight, A. S. Havulinna, B. Wahlstrand, T. Hedner, D. Corella, E. S. Tai, J. M. Ordovas, G. Berglund, E. Vartiainen, P. Jousilahti, B. Hedblad, M. R. Taskinen, C. Newton-Cheh, V. Salomaa, L. Peltonen, L. Groop, D.M. Altshuler, and M. Orho-Melander. Six

new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, 40(2):189–197, 2008.

[132] N. Keiding. Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society of London A*, 332(459):487–509, 1990.

[133] J. Kelsall and J. Wakefield. Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701, 2002.

[134] H. Keranen, P. Malinen, and O. Aulaskari. Suomen maaseututyypit. research papers 20. (in Finnish). http://www.aluekehityssaatio.fi/far/?download=selvityksia20.pdf, Finnish regional research, Sonkajärvi, Finland, 2000.

[135] W. O. Kermack, A. G. McKendrick, and P. L. McKinlay. Death-rates in Great Britain and Sweden: expression of specific mortality rates as products of two factors, and some consequences thereof. *Journal of Hygiene*, 38(4):433–457, 1934.

[136] T. Keränen and R. Marttila. Kapseli 30. Parkinsonin taudin lääkehoito (in Finnish), Helsinki, 2002.

[137] J. Kettunen, T. Lanki, P. Tiittanen, P. P. Aalto, T. Koskentalo, M. Kulmala, V. Salomaa, and J. Pekkanen. Associations of fine and ultrafine particulate air pollution with stroke mortality in an area of low air pollution levels. *Stroke*, 38:918–922, 2007.

[138] D. E. King, M. E. Mainous III, A. G. Geesey, and R. F. Woolson. Dietary magnesium and C-reactive protein levels. *Journal of the American College of Nutrition*, 24(3):166–171, 2005.

[139] P. Knekt, L. Teppo, A. Aromaa, H. Rissanen, and T. U. Kosunen. *Helicobacter pylori* IgA and IgG antibodies, serum pepsinogen I and the risk of gastric cancer: changes in the risk with extended follow-up period. *International Journal of Cancer*, 119:702–705, 2006.

[140] L. Knorr-Held. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567, 2000.

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

[142] L. Knorr-Held and N. G. Best. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 164:73–85, 2001.

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

[144] L. Knorr-Held and S. Richardson. A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Applied Statistics*, 52(2):169–183, 2003.

[145] L. Knorr-Held and H. Rue. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statististics*, 29:597–614, 2002.

[146] E. Kokki and A. Penttinen. Poisson regression with change-point prior in the modelling of disease risk around a point source. *Biometrical Journal*, 45(6):689–703, 2003.

[147] E. Kokki, J. Ranta, A. Penttinen, E. Pukkala, and J. Pekkanen. Small area estimation of incidence of cancer around a known source of exposure with fine resolution data. *Occupational and Environmental Medicine*, 58:315–320, 2001.

[148] T. Koljonen (ed.). *Suomen geokemian atlas. Osa 2. Moreeni (mainly in Finnish)*. Geologian tutkimuskeskus, 1992.

[149] F. Komaki. Homogeneous Gaussian Markov processes on general lattices. *Advances in Applied Probability*, 28:189–206, 1996.

[150] P. Korhonen, N. Malila, E. Pukkala, L. Teppo, D. Albanes, and J. Virtamo. The Finnish Cancer Registry as a follow-up source of a large cancer registry. *Acta Oncologica*, 41(4):381–388, 2002.

[151] S. Koskinen. *Origins of regional differences in mortality from ischaemic heart disease in Finland*. PhD thesis, University of Helsinki, Research and Development Centre for Welfare and Health, Research Reports 41, 1994.

[152] S. Koskinen, T. Martelin, I-L. Notkola, V. Notkola, and Pitkänen K., editors. *Suomen Väestö (in Finnish)*. Gaudeamus, Helsinki, 1994.

[153] A. Kottas, J. Duan, and A. Gelfand. Modeling disease incidence data with spatial and spatio-temporal Dirichlet process mixtures. *Biometrical Journal*, 50(1):29–42, 2008.

[154] A. Kousa, E. Moltchanova, M. Viik-Kajander, M. Rytkönen, J. Tuomilehto, T. Tarvainen, and M. Karvonen. Geochemistry of ground water and the incidence of acute myocardial infarction in Finland. *Journal of Epidemiology & Community Health*, 58:136–139, 2004.

[155] R. Kuittinen, T. Sarjakoski, M. Ollikainen, M. Poutanen, R. Nuuros, P. Tätilä, J. Peltola, R. Ruotsalainen, and M. Ollikainen. JHS 154: ETRS89 –järjestelmään liittyvät karttaprojektiot, tasokoordinaatistot ja karttalehtijako (in Finnish). http://www.jhs-suositukset.fi/suomi/jhs154, September 9, 2006.

[156] M. Kulldorff. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 164(1):61–72, 2001.

[157] A. M. Kuopio, R. J. Marttila, H. Helenius, and U. K. Rinne. Environmental risk factors in Parkinson's disease. *Movement Disorders*, 14:928–939, 1999.

[158] T. Laatikainen, P. Pajunen, R. Pääkkönen, I. Keskimäki, H. Hämäläinen, H. Rintanen, M. Niemi, V. Moltchanov, and V. Salomaa. National Cardiovascular Disease Register, statistical database. http://www.ktl.fi/cvdr/.

[159] P. Lahermo, T. Tarvainen, T. Hatakka, B. Backman, R. Juntunen, N. Kortelainen, T. Lakomaa, M. Nikkarinen, P. Vesterbacka, U. Väisänen, and P. Suomela. One thousand wells—the physical-chemical quality of Finnish

well waters in 1999 (mostly in Finnish). http://www.gsf.fi/info/publications/tr155/16372TutRap155.pdf, Espoo, 2002.

[160] T. Lappalainen, S. Koivumäki, E. Salmela, K. Huoponen, P. Sistonen, M. L. Savontaus, and P. Lahermo. Regional differences among the Finns: a Y-chromosomal perspective. *Gene*, 376(2):207–215, 2006.

[161] S. C. Larsson, M. J. Virtanen, M. Mars, S. Männistö, P. Pietinen, D. Albanes, and J. Virtamo. Magnesium, calcium, potassium, and sodium intakes and risk of stroke in male smokers. *Archives of Internal Medicine*, 168(5):459–465, 2008.

[162] S. Lauritzen and T. S. Richardson. Chain graph models and their causal interpretations (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):321–361, 2002.

[163] S. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 50:154–224, 1988.

[164] A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J-F. Viel, and R. Bertollini. *Disease mapping and risk assessment for public health*. John Wiley & Sons, Ltd., UK, 1999.

[165] A. B. Lawson, A. B. Biggeri, D. Boehning, E. Lesaffre, J-F. Viel, A. Clark, P. Schlattmann, and F. Divino. Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 19:2217–2241, 2000.

[166] A. B. Lawson and F. L. R. Williams. Spatial competing risk models in disease mapping. *Statistics in Medicine*, 19:2451–2467, 2000.

[167] R. D. Lee and L. R. Carter. Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, 87:659–671, 1992.

[168] H-R. Lehto, S. Lehto, A. S. Havulinna, M. Ketonen, A. Lehtonen, Y. A. Kesäniemi, J. Airaksinen, and V. Salomaa. Are coronary event rates declining slower in women than in men—evidence from two population-based myocardial registers in Finland. *BMC Cardiovascular Disorders*, 7:35, 2007.

[169] F. Lindgren and H. Rue. On the second order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35:691–700, 2008.

[170] Y. C. MacNab and C. B. Dean. Spatio-temporal modelling of rates for the construction of disease maps. *Statistics in Medicine*, 21:347–358, 2002.

[171] A. Majeed and P. Aylin. The ageing population of the United Kingdom and cardiovascular disease. *BMJ*, 331:1362, 2005.

[172] M. A. Martínez-Benuito, A. López-Quilez, and P. Botella-Rocamora. An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27:2874–2889, 2008.

[173] C. C. McCullogh, D. M. Kay, S. A. Factor, A. Samii, J. G. Nutt, D. S. Higgins, A. Griffith, J. W. Roberts, B. C. Leis, J. S. Montimurro, C. P. Zabetian, and H. Payami. Exploring gene-environment interactions in Parkinson's disease. *Human Genetics*, 123:257–265, 2008.

[174] R. McPherson, A. Pertsemlidis, N. Kavaslar, A. Stewart, R. Roberts, D. R. Cox, D. A. Hinds, L. A. Pennacchio, A. Tybjaerg-Hansen, A. R. Folsom, E. Boerwinkle, and J. C. Hobbs, H. H. anmd Cohen. A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316:1488–1491, 2007.

[175] M. S. Meade. Geographic analysis of disease and care. *Annual Review of Public Health*, 7:313–335, 1986.

[176] M. S. Meade and R. S. Earickson. *Medical Geography*. The Guilford Press, New York, NY, Second edition, 2005.

[177] N. Metropolis. The beginning of the Monte Carlo method. *Los Alamos Science*, Special Issue:125–130, 1987.

[178] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[179] J. L. Meza. Empirical Bayes estimation smoothing of relative risks in disease mapping. *Journal of Statistical Planning and Inference*, 112:43–62, 2003.

[180] M. Mähönen. *Poistoilmoitusrekisteri sepelvaltimotaudin epidemiologisen tutkimuksen tietolähteenä*. PhD thesis, Oulu University, STAKES tutkimuksia 28, 1993.

[181] M. Mähönen, V. Salomaa, M. Brommels, A. Molarius, H. Miettinen, K. Pyörälä, J. Tuomilehto, M. Arstila, E. Kaarsalo, M. Ketonen, K. Kuulasmaa, S. Lehto, H. Mustaniemi, M. Niemelä, P. Palomäki, J. Torppa, and T. Vuorenmaa. The validity of hospital discharge register data on coronary heart disease in Finland. *European Journal of Epidemiology*, 13:403–415, 1997.

[182] P. Mäkelä, S. Ripatti, and T. Valkonen. Alue-erot miesten alkoholikuolleisuudessa (in Finnish). *Suomen lääkärilehti*, 53(23):2513–2519, 2001.

[183] J. Müller-Nordhorn, S. Binting, S. Roll, and S. N. Willich. An update on regional variation in cardiovascular mortality within Europe. *European Heart Journal*, 29:1316–1326, 2008.

[184] R. Møller and R. Waagepetersen. Modern statistics for spatial point processes. *Scandinavian Journal of Statistics*, 34:643–684, 2007.

[185] A. Mollié. Bayesian mapping of Hodgkin's disease in France. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: methods and applications*. Oxford University Press, NY, 2000.

[186] E. Moltchanova, A. Penttinen, and M. Karvonen. A hierarchical Bayesian birth cohort analysis from incomplete registry data: evaluating the trends in the age of onset of insulin-dependent diabetes mellitus (T1DM). *Statistics in Medicine*, 24(19):2989–3004, 2005.

[187] E. Moltchanova, M. Rytkönen, A. Kousa, O. Taskinen, J. Tuomilehto, and M. Karvonen. Zinc and nitrate in the ground water and the incidence of Type 1 diabetes in Finland. *Diabetic Medicine*, 21(3):256–261, 2004.

[188] E. V. Moltchanova, J. Pitkäniemi, and L. Haapala. Potts model for haplotype associations. *BMC Genetics*, 6(Suppl. 1):S64, 2005.

[189] S. Monarca, F. Donato, I. Zerbini, R. L. Calderon, and G. F. Craun. Review of epidemiological studies on drinking water hardness and cardiovascular diseases. *European Journal of Cardiovascular Prevention & Rehabilitation*, 13(4):495–506, 2006.

[190] C. Montomoli, C. Allemani, G. Solinas, G. Motta, L. Bernardinelli, S. Clemente, B.S. Murgia, A. F. Ticca, L. Musu, M. L. Piras, R. Ferrai, A. Caria, S. Sanna, and O. Porcu. An ecologic study of geographical variation in multiple sclerosis risk in central Sardinia, Italy. *Neuroepidemiology*, 21:187–193, 2002.

[191] R. W. Morris, M. Walker, L. T. Lennon, A. G. Shaper, and P.H. Whincup. Hard drinking water does not protect against cardiovascular disease: new evidence from the British Regional Heart Study. *European Journal of Cardiovascular Prevention & Rehabilitation*, 15(2):185–189, 2008.

[192] A. S. Mugglin, B. P. Carlin, and A. E. Gelfand. Fully model based approaches for spatially misaligned data. *Journal of the American Statistical Association*, 95(451):877–887, 2000.

[193] T. Muilu and J. Rusanen. Rural young people in regional development—the case of Finland in 1970–2000. *Journal of Rural Studies*, 19:295–207, 2003.

[194] H. Mustaniemi. *Äkillisten sepelvaltimotautikohtausten kehityssuunnat Pohjois-Karjalassa 1973-1990 (in Finnish)*. PhD thesis, Kuopio University, 1993.

[195] T. Nakaya. An information statistical approach to the modifiable areal unit problem in incidence rate maps. *Environment and Planning A*, 32:91–109, 2000.

[196] R. M. Neal. Markov chain Monte Carlo methods based on 'slicing' the density function. Technical report, No. 9722, Dept. of statistics, University of Toronto, 1997.

[197] R. M. Neal. Slice sampling (with discussion). *Annals of Statistics*, 31(3):705–763, 2003.

[198] V. Nerich, E. Monnet, A. Etienne, S. Louafi, C. Ramé, S. Rican, A. Weill, N. Vallier, V. Vanbockstael, G-R. Auleley, H. Allemand, and F. Carbonnel. Geographical variations of inflammatory bowel disease in France: a study based on national health insurance data. *Inflammatory Bowel Diseases*, 12:218–226, 2006.

[199] L. E. Nieto-Bajaras. A Markov gamma random field for modelling disease mapping data. *Statistical Modelling*, 8(1):97–114, 2008.

[200] D. Nitsch, S. Morton, B. DeStavola, H. Clark, and D. A. Leon. How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950s study. *BMC Medical Research Methodology*, 6:15–23, 2006.

[201] R. Norio. The Finnish disease heritage III: the individual diseases. *Human Genetics*, 112(5–6):470–526, 2003.

[202] S. Näyhä. Geographical variations in cardiovascular mortality in Finland, 1961–1985. *Scandinavian Journal of Social Medicine*, Supplementum 40:1–48, 1989.

[203] R. Ocaña-Riola. The misuse of count data aggregated over time for disease mapping. *Statistics in Medicine*, 26:4489–4504, 2007.

[204] A. O'Hagan. Dicing with the unknown. *Significance*, 1(3):132–133, 2005.

[205] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting expert probabilities (statistical practise)*. John Wiley & Sons, Ltd., UK, 2006.

[206] A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, New York, 2003.

[207] A. O'Hagan and J. E. Oakley. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety*, 85:239–248, 2004.

[208] A. R. Omran. The epidemiological transition: A theory of the epidemiology of population change (1971). *The Milbank Quarterly*, 83:731–757, 2005.

[209] C. Osmond and M. J. Gardner. Age, period and cohort models applied to cancer mortality rates. *Statistics in Medicine*, 1:245–259, 1982.

[210] K. Osnes and O. O. Aalen. Spatial smoothing of cancer survival: a Bayesian approach. *Statistics in Medicine*, 18:2087–2099, 1999.

[211] P. Pajunen, H. Koukkunen, M. Ketonen, T. Jerkkola, P. Immonen-Räihä, P. Kärjä-Koskenkari, M. Mähönen, M. Niemelä, K. Kuulasmaa, P. Palomäki, J. Mustonen, A. Lehtonen, M. Arstila, T. Vuorenmaa, S. Lehto, H. Miettinen, J. Torppa, J. Tuomilehto, Y. A. Kesäniemi, K. Pyörälä, and V. Salomaa. The validity of the Finnish Hospital Discharge Register and Causes of Death Register data on coronary heart disease. *European Journal of Cardiovascular Prevention & Rehabilitation*, 12(2):132–137, 2005.

[212] P. Pajunen, R. Pääkkönen, H. Hämäläinen, I. Keskimäki, T. Laatikainen, M. Niemi, H. Rintanen, and V. Salomaa. Trends in fatal and non-fatal strokes among persons aged 35-85+ years during 1991-2002 in Finland. *Stroke*, 36:244–248, 2005.

[213] P. Pajunen, R. Pääkkönen, A. Juolevi, H. Hämäläinen, I. Keskimäki, T. Laatikainen, V. Moltchanov, M. Niemi, H. Rintanen, and V. Salomaa. Trends in fatal and non-fatal coronary heart disease events in Finland during 1991-2001. *Scandinavian Cardiovascular Journal*, 38:340–344, 2004.

[214] M. E. Paté-Cornell. Uncertainties in risk analysis: six levels of treatment. *Reliability Engineering and System Safety*, 54:95–111, 1996.

[215] M. Paturi, H. Tapanainen, H. Reinivuo, and P. Pietinen, editors. *Finravinto 2007 -tutkimus – The National FINDIET 2007 Survey*. Kansanterveyslaitoksen julkaisuja B23, 2007.

[216] J. Pekkanen, E. Pukkala, M. Vahteristo, and T. Vartiainen. Cancer incidence around an oil refinery as an example of a small area study based on map coordinates. *Environmental Research*, 71:128–134, 1995.

[217] M-E. Piche, S. J. Weisnagel, L. Corneau, A. Nadeau, J. Bergeron, and S. Lemieux. The WHO and NCEP/ATPIII definitions of the metabolic syndrome in postmenopausal women: are they so different? *Metabolic Syndrome and Related Disorders*, 4(1):12–27, 2006.

[218] J. Pihlajamaa, J. Suvisaari, M. Henriksson, H. Heilä, E. Karjalainen, J. Koskela, M. Cannon, and J. Lönnqvist. The validity of schizophrenia diagnosis in the Finnish Hospital Discharge Register: findings from a 10-year birth cohort sample. *Nordic Journal of Psychiatry*, 62(3):198–203, 2008.

[219] R. Piispanen. Geochemical interpretation of cancer maps of Finland. *Environmental Geochemistry and Health*, 11(3–4):145–147, 1989.

[220] K. Pitkänen, S. Koskinen, and T. Martelin. Kuolleisuuden alue-erot ja niiden historia (in Finnish). *Duodecim*, 116:1697–1710, 2000.

[221] K. Pääkkönen, S. Sauramo, L. Sarantaus, P. Vahteristo, A. Hartikainen, P. Vehmanen, J. Ignatius, V. Ollikainen, H. Kääriäinen, E. Vauramo, H. Nevanlinna, R. Krahe, K. Holli, and J. Kere. Involvement of BRCA1 and BRCA2 in breast cancer in a western Finnish sub-population. *Genetic Epidemiology*, 20:239–246, 2001.

[222] Finnish Population Register Centre. Rakennusten osoite- ja koordinaattitietojen kattavuus väestötietojärjestelmässä (in Finnish). http://www.vrk.fi/vrk/home.nsf/files/selvitys\$file/selvitys.pdf, 2004.

[223] E. Pukkala. Cancer maps of Finland: an example of small area-based mapping. *Recent Results In Cancer Research*, 114:208–215, 1989.

[224] E. Pukkala, Gustafsson N., and L. Teppo. Atlas of cancer incidence in Finland, 1953–1982, Publication 37. Cancer Society of Finland, 1987.

[225] E. Pukkala, T. Patama, G. Engholm, G. H. Ólafsdóttir, F. Bray, M. Talbäck, and K. Pasanen. Small-area based map animations of cancer incidence in the Nordic countries, 1971-2003. http://astra.cancer.fi/cancermaps/Nordic/, Nordic Cancer Union, 2007.

[226] E. Pukkala, R. Sankila, and M. Rautalahti. Syöpä Suomessa 2006 (in Finnish). http://www.cancerregistry.fi/tutkimus/image_44.pdf, Suomen Syöpäyhdistys, Helsinki 2006.

[227] E. Pukkala, B. Söderman, A. Okeanov, H. Storm, M. Rahu, T. Hakulinen, N. Becker, R. Stabenow, K. Bjarnadottir, A. Stengrevics, R. Gurevicius, E. Glattre, W. Zatonski, T. Men, and L. Barlow. Cancer atlas of Northern Europe. http://www.cancerregistry.fi/atlasweb/index.htm, Cancer Society of Finland, Helsinki, 2001.

[228] P. J. Pussinen, K. Tuomisto, P. Jousilahti, A. S. Havulinna, Sundvall J., and V. Salomaa. Endotoxemia, immune response to periodontal pathogens, and systemic inflammation associate with incident cardiovascular disease events. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 27:1433–1439, 2007.

[229] J. Quiñonero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[230] F. A. Quintana and P. L. Iglesias. Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574, 2003.

[231] Tindol R. Getting the Pox off all their houses: Cotton Mather and the rhetoric of puritan science. *Early American Literature*, 46:1–23, 2011.

[232] The Finnish Radiation and Nuclear Safety Authority. Sisäilman radon (Indoor radon, in Finnish). http://www.stuk.fi/sateilytietoa/sateily_ymparistossa/radon/fi_FI/radon/_files/71375051687264509/default/sisailman_radon.pdf, 2003.

[233] R. Ramis-Prieto, J. García-Pérez, M. Pollán, N. Aragonés, B. Pérez-Gómez, and G. López-Abente. Modelling of municipal mortality due to haematological neoplasias in Spain. *Journal of Epidemiology & Community Health*, 61:165–171, 2007.

[234] J. Ranta and A. Penttinen. Probabilistic small area risk assessment using GIS-based data: a case study on Finnish childhood diabetes. *Statistics in Medicine*, 19:2345–2359, 2000.

[235] H. Rantanen and M. Kahila. The SoftGIS approach to local knowledge. *Journal of Environmental Management*, 2008.

[236] C. E. Rasmussen and Z. Ghahramani. Bayesian Monte Carlo. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*. Mit Press, Cambridge, MA, 2003.

[237] M. Rezaeian, G. Dunn, S. St Leger, and L. Appleby. Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. *Journal of Epidemiology & Community Health*, 61:98–102, 2007.

[238] S. Richardson, J. J. Abellan, and N. Best. Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical Methods in Medical Research*, 15:385–407, 2006.

[239] S. Richardson and G. Monfort. Ecological correlation studies. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: methods and applications*. Oxford University Press, NY, 2000.

[240] S. Richarson, A. Thomson, N. Best, and P. Elliott. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112(9):1016–1025, 2004.

[241] S. Riedel. Edward Jenner and the history of smallpox and vaccination. *Baylor University Medical Center Proceedings*, 18:21–25, 2005.

[242] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, NY, 1999.

[243] K. Rothman. Lessons from John Graunt. *Lancet*, 347(8993):37–39, 1996.

[244] H. Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338, 2001.

[245] H. Rue and L. Held. *Gaussian random fields: theory and applications*. Chapman & Hall/CRC, Boca Raton, FL, 2005.

[246] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319—-392, 2009.

[247] M. Rytkönen. *Geographical study on childhood type 1 diabetes mellitus (T1DM) in Finland*. PhD thesis, Oulu University, 2004.

[248] M. Rytkönen, E. Moltchanova, J. Ranta, O. Taskinen, J. Tuomilehto, and M. Karvonen. The incidence of type I diabetes among children in Finland—rural–urban difference. *Health & Place*, 9:315–325, 2003.

[249] M. Rytkönen, J. Ranta, J. Tuomilehto, and M. Karvonen. Bayesian analysis of geographical variation in the incidence of Type I diabetes in Finland. *Diabetologia*, 44(suppl 3):B37–B44, 2001.

[250] E. Salmela, O. Taskinen, J. K. Seppänen, P. Sistonen, M. J. Daly, P. Lahermo, M-L. Savontaus, and J. Kere. Subpopulation difference scanning: a strategy for exclusion mapping of susceptibility genes. *Journal of Medical Genetics*, 43:590–597, 2006.

[251] V. Salomaa, M. Ketonen, H. Koukkunen, P. Immonen-Räihä, T. Jerkkola, P. Kärjä-Koskenkari, M. Mähönen, M. Niemelä, K. Kuulasmaa, P. Palomäki, M. Arstila, T. Vuorenmaa, A. Lehtonen, S. Lehto, H. Miettinen, J. Torppa, J. Tuomilehto, Y. A. Kesäniemi, and K. Pyörälä. Trends in coronary events in Finland during 1983-1997; the FINAMI study. *European Heart Journal*, 24:311–319, 2003.

[252] V. Salomaa, H. Miettinen, P. Palomäki, M. Arstila, H. Mustaniemi, K. Kuulasmaa, and J. Tuomilehto. Diagnostic features of acute myocardial infarction—changes over time from 1983 to 1990: results from the FINMONICA AMI register study. *Journal of Internal Medicine*, 237:151–159, 1995.

[253] V. Salomaa, R. Pääkkönen, H. Hämäläinen, M. Niemi, and T. Klaukka. Use of secondary preventive medications after the first attack of acute coronary syndrome. *European Journal of Cardiovascular Prevention & Rehabilitation*, 14:386–391, 2007.

[254] N. J. Samani, J. Erdmann, A. S. Hall, C. Hengstenberg, M. Mangino, B. Mayer, and et al. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*, 357(5):443–453, 2007.

[255] W. C. Sanderson and S. Scherbov. Remeasuring aging. *Science*, 329:1287–1288, 2010.

[256] A. Savu, J. Potter, S. Li, and Y. Yasui. Breast cancer and microbial cancer incidence in female populations around the world: a surprising hyperbolic association. *International Journal of Cancer*, 123:1094–1099, 2008.

[257] A. Schaerström. Apparent and actual disease landscapes—some reflections on the geographical definition of health and disease. *Geografiska Annaler*, 81 B(4):235–242, 1999.

[258] J. Schwartz. The effects of particulate air pollution on daily deaths: a multi-city case crossover analysis. *Occupational and Environmental Medicine*, 61:956–961, 2004.

[259] H. C. Selvin. Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology*, 63:607–619, 1958.

[260] S. Siesling, J. W. W. van der Aa, M. A. amd Coebergh, and E. Pukkala. Time-space trends in cancer incidence in the Netherlands in 1989–2003. *International Journal of Cancer*, 122:2106–2114, 2008.

[261] G. L. Silva, C. B. Dean, T. Niyonsenga, and A. Vanesse. Hierarchical Bayesian spatiotemporal analysis of revascularizarion odds using smoothing splines. *Statistics in Medicine*, 27:2381–2401, 2008.

[262] M. Similä, O. Taskinen, S. Männistö, M. Lahti-Koski, M. Karvonen, T. Laatikainen, and L. Valsta. Terveyttä edistävä ruokavalio, lihavuus ja seerumin kolesteroli karttoina (in Finnish). http://www.ktl.fi/attachments/suomi/julkaisut/julkaisusarja_b/2005/2005b20.pdf, 2005.

[263] K. C. Simon, H. Chen, M. Schwarzschild, and A. Ascherio. Hypertension, hypercholesterolemia, diabetes, and risk of Parkinson disease. *Neurology*, 69(17):1688–1695, 2007.

[264] S. A. Sisson. Transdimensional Markov Chains: a decade of progress and future perspectives. *Journal of the American Statistical Association*, 100(471):1077–1089, 2005.

[265] J. Sivenius, J. Tuomilehto, P. Immonen-Räihä, M. Kaarisalo, C. Sarti, J. Torppa, K. Kuulasmaa, M. Mähönen, A. Lehtonen, and V. Salomaa. Continuous 15-year decrease in incidence and mortality of stroke in Finland. *Stroke*, 35:420–425, 2004.

[266] Y. Song, P. M. Ridker, J. E. Manson, N. R. Cook, J. E. Buring, and S. Liu. Magnesium intake, C-reactive protein, and the prevalence of metabolic syndrome in middle-aged and older U.S. women. *Diabetes Care*, 28:1438–1444, 2005.

[267] D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. WinBUGS User Manual, Version 1.4.3. http://www.mrc-bsu.cam.ac.uk/bugs, September 2007.

[268] D. J. Spiegelhalter, N. G. Best, and B. P Carlin. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. http://citeseer.ist.psu.edu/spiegelhalter98bayesian.html, Minneapolis, 1998.

[269] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[270] A. Stang, E. Pukkala, R. Sankila, B. Söderman, and T. Hakulinen. Time trend analysis of the skin melanoma incidence of Finland from 1953 through 2003 including 16,414 cases. *International Journal of Cancer*, 119:380–384, 2006.

[271] M. Stephens. Bayesian analysis of mixture model with an unknown number of components—an alternative to reversible jump methods. *Annals of Statistics*, 28(1):40–74, 2000.

[272] H. B. Stern and N. Cressie. Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19:2377–2397, 2000.

[273] J. A. Steward, C. White, and S. Reynolds. Leukaemia incidence in Welsh children linked with low level radiation—making sense of some erroneous results published in the media. *Journal of Radiological Protection*, 28:33–43, 2008.

[274] M-L. Sumelahti. *Occurrence, survival and prognostic factors of multiple sclerosis in Finland*. PhD thesis, Tampere University, 2002.

[275] M. Tala-Heikkilä. Tuberkuloosi suomessa (in Finnish). *Duodecim*, 119:1621–1628, 2003.

[276] M. Tammilehto-Luode, L. Backer, and L. Rogstad. Grid data and area delimitation by definition towards a better European territorial statistical system. *Statistical Journal of the United Nations ECE*, 17:109–117, 2000.

[277] A. Thomas, N. Best, D. Lunn, R. Arnold, and D. Spiegelhalter. GeoBUGS User Manual, Version 1.2. http://www.mrc-bsu.cam.ac.uk/bugs, September 2004.

[278] S. Timonen, U. Uotila, P. Kuusisto, and O. Lokki. Effect of certain maternal, foetal and geographical factors on the weight and length of the newborn and on the duration of the pregnancy. *Annales Chirurgiae and Gynaecologiae Fenniae*, 55:196–213, 1966.

[279] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2):234–240, 1970.

[280] H. Tolonen, V. Salomaa, J. Torppa, J. Sivenius, P. Immonen-Räihä, and A. Lehtonen. The validation of the Finnish Hospital Discharge Register and Causes of Death Register data on stroke diagnoses. *European Journal of Cardiovascular Prevention and Rehabilitation*, 14(3):380–385, 2007.

[281] K. Tunstall-Pedoe, H. amd Kuulasmaa, P. Amouyel, D. Arveiler, A-M. Rajakangas, and A. Pająk. Myocardial infarctions and coronary deaths in the World Health Organization MONICA project. *Circulation*, 90:583–612, 1994.

[282] J. Tuomilehto, D. Rastenyte, J. Sivenius, C. Sarti, P. Immonen-Räihä, E. Kaarsalo, K. Kuulasmaa, E. V. Narva, V. Salomaa, K. Salmi, and J. Torppa. Ten-year trends in stroke incidence and mortality in the FIN-MONICA stroke study. *Stroke*, 27:825–832, 1996.

[283] P. Tyynelä, S. Goebeler, E. Ilveskoski, J. Mikkelsson, M. Perola, M. Löytönen, and P. J. Karhunen. Birthplace predicts risk for prehospital sudden cardiac death in middle-aged men who migrated to metropolitan area: the Helsinki Sudden Death Study. *Annals of Medicine*, 41(1):57–65, 2009.

[284] E. Tzala and N. Best. Bayesian latent variable modelling of multivariate spatio-temporal variation in cancer mortality. *Statistical Methods in Medical Research*, 17:97–118, 2008.

[285] J. Valjus, M. Hongisto, P. Verkasalo, P. Järvinen, K. Heikkilä, and M. Koskenvuo. Residential exposure to magnetic fields generated by 110–400 kV power lines in Finland. *Bioelectromagnetics*, 16:365–376, 1995.

[286] T. Valkonen. Male mortality from ischaemic heart disease in Finland: relation to region of birth and region of residence. *European Journal of Population*, 3:61–83, 1987.

[287] T. Valkonen. Trends in regional and socio-economic mortality differentials in Finland. *International Journal of Health Sciences*, 3(3):157–166, 1992.

[288] J. Vanhatalo and A. Vehtari. Sparse log gaussian processes via MCMC for spatial epidemiology. *JMLR: Workshop and Conference Proceedings*, 1:73–89, 2007.

[289] E. Vartiainen, P. Jousilahti, G. Alfthan, J. Sundwall, P. Pietinen, and P. Puska. Cardiovascular risk factor changes in Finland, 1972–1997. *International Journal of Epidemiology*, 29:49–56, 2000.

[290] E. Vauramo, P. Mikkola, I. Sippo-Tujunen, S. Aro, S. Pelanteri, and E. Hokkanen. Coordinate-based mapping: a new method in health services research. *Medical Informatics London*, 17:1–9, 1992.

[291] A. Vehtari. *Bayesian model assessment and selection using expected utilities*. PhD thesis, Helsinki University of Technology, 2001.

[292] A. Vehtari and J. Lampinen. Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, 14(10):2339–2468, 2002.

[293] P. K. Verkasalo, E. Kokki, E. Pukkala, T. Vartiainen, H. Kiviranta, A. Penttinen, and J. Pekkanen. Cancer risk near a polluted river in Finland. *Environmental Health Perspectives*, 112(9):1026–1031, 2004.

[294] P. K. Verkasalo, E. Pukkala, J. Kaprio, K. V. Heikkilä, and M. Koskenvuo. Magnetic fields of high voltage power lines and risk of cancer in Finnish adults: nationwide cohort study. *BMJ*, 313:1047–1051, 1996.

[295] J. Wakefield. A critique of statistical aspects of ecological studies of spatial epidemiology. *Environmental and Ecological Statistics*, 11:31–54, 2004.

[296] J. Wakefield and G. Shaddick. Health-exposure modelling and the ecological fallacy. *Biostatistics*, 7(3):438–455, 2006.

[297] L. A. Waller. A civil action and statistical assessments of the spatial pattern of disease: do we have a cluster? *Regulatory Toxicology and Pharmacology*, 32:174–183, 2000.

[298] L. A. Waller, B. P. Carlin, H. Xia, and A. E. Gelfand. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438):607–617, 1997.

[299] S. D. Walter. Disease mapping: a historical perspective. In P. Elliott, J. C. Wakefield, N. G. Best, and D. J. Briggs, editors, *Spatial Epidemiology: methods and applications*. Oxford University Press, NY, 2000.

[300] S. D. Walter and S. E Birnie. Mapping mortality and morbidity patterns: an international comparison. *International Journal of Epidemiology*, 20(3):678–689, 1991.

[301] S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[302] S. Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, 2010.

[303] C. Watson and N. J. Alp. Role of *Chlamydia pneumoniae* in atherosclerosis. *Clinical Science*, 114:509–531, 2008.

[304] D. Weintraub, C. L. Comella, and S. Horn. Parkinson' disease—Part 1: Pathophysiology, symptoms, burden, diagnosis and assessment. *American Journal of Managed Care*, 14:S40–S48, 2008.

[305] E. A. Whitsel, K. M. Rose, J. L. Wood, A. C. Henley, D. Liao, and G. Heiss. Accuracy and repeatability of commercial geocoding. *American Journal of Epidemiology*, 160(10):1023–1029, 2004.

[306] P. Whittle. On stationary processes in the plane. *Biometrika*, 41(3/4):434–449, 1954.

[307] M. Winkleby, C. Cubbin, and D. Ahn. Effect of cross-level interaction between individual and neighborhood socioeconomic status on adult mortality rates. *American Journal of Public Health*, 96(12):2145–2153, 2006.

[308] R. Wolpert and K. Ickstadt. Poisson/gamma random field models for spatial statistics. *Biometrika*, 85(2):251–267, 1998.

[309] T. J. Xaviera. *Importancia de la viruela, gastroenteritis aguda y paludismo en Finlandia entre 1749 y 1850 (in Spanish)*. PhD thesis, Oulu University, 2005.

[310] P. Yan and M. K. Clayton. A cluster model for space–time disease counts. *Statistics in Medicine*, 25:867–881, 2006.

[311] C-Y. Yang, C-C. Changa, S-S. Tsaib, and H-F. Chiuc. Calcium and magnesium in drinking water and risk of death from acute myocardial infarction in Taiwan. *Environmental Research*, 101(3):407–411, 2006.

[312] M. J. Yarnoz and A. B. Curtis. More reasons why men and women are not the same (gender differences in electrophysiology and arrhythmias). *American Journal of Cardiology*, 101(9):1291–1296, 2008.

[313] H. Ylihärsilä, J. Lindström, J. G. Eriksson, P. Jousilahti, T. T. Valle, J. Sundvall, and J. Tuomilehto. Prevalence of diabetes and impaired glucose regulation in 45- to 64-year-old individuals in three areas of Finland. *Diabetic Medicine*, 22(1):88–91, 2005.

[314] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos, and L. Lisheng. Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. *Lancet*, 364:937–952, 2004.

[315] O. Zurriaga, H. Vanaclocha, M. A. Martinez-Benuito, and P. Botella-Rocamora. Spatio-temporal evolution of female lung cancer mortality in a region of Spain: is it worth taking migration into account? *BMC Cancer*, 8:35, 2008.

# Errata

## Publication II

Table 2: east/west rate ratio in ischemic stroke / men should be 1.06, not 1.09.

Aalto-DD 138/2011

TERVEYDEN JA
HYVINVOINNIN LAITOS
National Institute for Health and Welfare

9 789526 044279

ISBN 978-952-60-4427-9
ISBN 978-952-60-4428-6 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
ISSN 1799-4942 (pdf)

**Aalto University**
**School of Science**
**Dept. of Biomedical Engineering and Computational Science**
**www.aalto.fi**