Statistical Applications in Geographical Health Studies

José Miguel Martínez Martínez
PhD Student

Joan Benach de Rovira
Universitat Pompeu Fabra
PhD Advisor

Yutaka Yasui
University of Alberta
PhD Advisor

Josep Ginebra i Molins
Universitat Politècnica de Catalunya
PhD Tutor

Doctorate in Technical and Computer Applications of Statistics,
Operational Research and Optimization.
Universitat Politècnica de Catalunya
Barcelona, 8/05/2006

INDEX

List of tables

List of figures

ACKNOWLEDGEMENTS

In the paragraphs that follow, I would like to express my gratitude to all those people who in some way or another have helped me during the long journey which began with my Diploma in Statistics, continued with my Bachelors Degree in Statistical Science and Techniques, to finally reach the present doctoral thesis. During this period of over ten years I believe I have learned something every day, and I am conscious of how much more I have to learn, or rather aware that I will never be able to learn all I would like to know. There is no doubt that the opportunity to learn a little more and be able to apply statistical methods which may be of use to other workers in a variety of fields, and in particular those dedicated to improving the population's health, was the main motivation which led me to undertake this thesis.

I am especially grateful to *Joan Benach*, *Yutaka Yasui* and *Josep Ginebra* for their enormous help in the process of elaborating this thesis. They are not only great researchers and teachers, but also great people. I am also grateful to *Joan* for teaching me to see the utility and the limitations of science, for trusting me and letting me develop my ideas and suggestions in all aspects where we have collaborated over the last five years or more, for helping me with and being concerned about my future. Thanks to *Josep* for sharing different ideas with me and making me recognize the strengths and limitations of statistical methods. I would like to thank *Yutaka* for being always so ready to help me with and teach me about a plethora of statistical techniques applied with the object of improving the health of populations. I am also grateful to him and all his family: *Marcy Winget*, *Niko*, *Osamu* and *Lena* (Today is the day!, Yes..No, word, Do you understand my English?...Yes, I understand your English...so....we can talk, we can talk, we can talk and sing♫ ♫..), for putting me up, and treating my like just one more member of the family, during my stay in Edmonton (Canada), as well as for the great times we spent together.

Thanks also to the members of the Department of Public Health Sciences, University of Alberta. I am particularly grateful to *Nicola Cherry* and *Duncan Saunders* for their all their assistance during my stay in Edmonton, and to *Irina*, *Xiufang* and *Doug* for help with the simulations carried out.

Thanks to the members of the Inner City Health Research Unit of St. Michael's Hospital (Toronto, Canada). In particular I would like to thank *Pat O'Campo*, *Jim Dunn* and *Piotr Gozdyra*. Thanks to *Carles Muntaner*, *Samuel Noh*, *Violet Kaspar* and *Ivonne Hinds* for their multiple attentions during our stay in the Center for Addiction and Mental Health, University of Toronto (Canada). I would like to extend special thanks to *Carles* and *Pat* for their always warm and constant support, and to *Gabriel* and *Daniel* for the pleasant moments together.

Thanks also to *Pepe Tapia* and *Ana Diex-Roux* for looking after us during our stay in Ann Arbor (Michigan, USA).

I am grateful to *Pedro Delicado* for his lecture course focussing on Bayesian methods. Thanks to that course I was able to start my doctorate, and have also as a result met some great companions. I am likewise indebted, for their considerable and disinterested assistance, to *Lourdes Rodero*, *Rosa Lamarca*, *Alfonso Buil* and *Ramón Clèries*. Particularly to my great friend *Ramón* for sharing with me the passion to understand little by little more about statistics from an applied perspective, which serves to improve peoples health status (thanks buddy!). I am also grateful to *Celia Martín Marta Llabres,* and *Lupe Gómez* for their help on various aspects related with the doctorate.

Thanks to the biostatisticians of IMIM, IMAS and UPF, who form a small group, but with a large dose of companionship and friendship. My thanks, then, to *Joan Vila*, *Josep Maria Manresa, Helena Martí, Ruben Román*, *Eduard Molins*, *Josué Almansa*, *Albert Sánchez*, *Montse Martínez*, *Gemma Vilagut*, *Angels Pont*, *Ignasi Serra*, *Paco Fernández*, *Alex Amoròs*, *Raquel García*, *Estel Plana*, *Marta Benet*, *Emma Arcos* and *Mireia Vilardell* and particularly to *Laura Muñoz* (now at the Catalan Institute of Oncology), *Xavi Basagaña* (currently completing his doctorate at Harvard), *Mercè Comas* and *Isaac Subirana*. I would also like to mention my colleagues *Joan Valls*, *Valentín Navarro, Maite Encuentra, Rebeca Font, Yolanda Benavente, Gina Albero, Cristian Bocanegra, David Santos, Jaume Escudero, Ruben Romera, Juan Ramón González, Gemma Castaño, Diego Pangusion*, *Jordi Real, Aureli Tobías, Anna*

Schiaffino and my Bachelors degree companions from Granada, *Emilio López*, *Antonio Arquero* and *Miguel Rodríguez-Barranco*.

Thanks to colleagues of the Occupational Health Unit. Particularly to *Fernando G. Benavides*, *Sergi Jarque*, *George Delclòs*, *David Gimeno*, *Consol Serra*, *Emily Ahonen*, *Maite Sampere*, *Marcelo Amable*, *Jordi Castejón*, not forgetting my office companions, *Cristina Portolés*, *Montse Vergara*, *Maria Buxó*, *Victoria Porthé*, *Ruth Domínguez* and office ex-companions *Marisa Martínez*, *Núria Catot, La Noe, Merche Iglesias* and *Sira Gónzalez*.

Thanks to my teachers from the Universidad Autónoma de Barcelona Diploma course: *Maria del Mar García* and *Miguel Martín*, for their courses focussed on application of statistical methods in health sciences. Particular thanks to *Victor Moreno* for constantly being there to resolve my methodological doubts.

Thanks to my teachers from the Granada Bachelors course in Statistical Sciences and Techniques, and especially to *José Miguel Angulo*, *Maria Dolores Ruiz Medina*, *Ana Aguilera*, *Mariano José Valderrama* and *Rafael Pérez-Ocón*.

Thanks to colleagues interested in geographical disease mapping, with whom I have shared endless thoughts and ideas over these years. Particular thanks to *Miguel Ángel Martínez-Beneito*, *Oscar Zurriaga*, *Antonio López*, *Paloma Botella*, *Gemma Cano, Enric Azlor*, *Mayka Rodríguez*, *Ricardo Ocaña* and *Carmen Sánchez-Cantalejo*.

Thanks to my students of the Occupational Health II course (part of the diploma in Labor Relations, Universitat Pompeu Fabra) for showing me how to learn through teaching.

Thanks to *Carme Borrell*, *Maribel Pasarín* and *Santiago Esnaola* for teaching me, along with *Joan* and *Carles*, the importance of studying inequalities in health. I am particularly grateful to *Carme*, and to *Joan Guix*, and to the Barcelona Public Health Agency in general for their constant willingness to facilitate a variety of sources of data, necessary to be able to apply the statistical methods developed in this thesis, and in this way being able to contribute to improvement of people's health.

My thanks also to my companions from the period when I worked in the Catalan Regional Government Health Department Information and Studies Service. Particularly to *Glòria Ribas*, *Roser Bosser, Mara Barés, Anna Puigdefàbregas, JJ Coll, Josep Maria Giné*, *Sergi Cruz*, *Carme Navalón*, *Rosa Gispert*, *Xavi Puig*, *Cristina Rius* and *Glòria Pérez*.

To the staff of the Pompeu Fabra University library (Mar Campus) and above all to *Josep Gibert*, *Marina Losada* and *Rosa Feixas* for always being ready to actively collaborate in searching for information.

To the members of the Cartographic, Mortality, Census and Statistical Services of the Catalan Institute of Statistics (IDESCAT), and very specially for the help received from *Gelasio Nogueira*, *Jordi Oliveres* and *Dolors Olivares*; also to the Spanish National Institute of Statistics (INE), specially *Montserrat García*, *David Goizueta* and *Guillermo Olmo*.

To *Dave Macfarlane* for translation into English of a considerable part of this thesis, and his valuable comments. Many thanks for your help.

To *Esther Español* for help in the graphical design.

To my parents, *Juan* and *Laura*, my brother *Juan Pedro* and sister *Marisa*, and other family and friends. They understand the effort which research and study require, and they are always the ones to make sacrifices. Thanks for your comprehension, and for being there when I needed you. In particular I am grateful to my parents for teaching me to appreciate the simple things in life, to respect others as one would oneself, and for accepting me just as I am. To my big sister for financial assistance in my studies, complementing that provided by scholarships and by my parents. To my brother for sharing the family business workload with me, Bar Kalamot (in Gavà), for always helping my parents there, and for infecting me with his enthusiasm for cinema. Some of our clients also deserve special mention, particularly *Alberto Caparrós*, *Alonso*, *Aurelio*, *Eduard*, *Martín Atenza*, *Manuel Martos*, *Marcelino* and *Armando*, for the great times we have enjoyed together.

To my childhood friends: *Pedro Peña*, *Jordi Sáez*, *David Martos*, *Javi Moreno*, *Manolo Pascual*, *Jordi Sierra* and *Diego González*. To my companions at the Shyto-Ryu Cervantes Karate center of Gavà, but particularly my instructor *José Alonso* for sharing his passion for Karate with me. To my grandmother *Luisa Puertas* and to the memory of my grandparents *Juan Pedro*, *Maria* and *Miguel* and in general to my extended family members in Gavà (Barcelona), Murcia, Madrid and particularly to those in the village of El Margen (Granada) for their warm reception during the summer months and where a good part of this thesis was elaborated. To my cats *Cuqui* and *Mimi*.

To *Mar Torné*, *Carme Farré*, *Esteve Torné* and *Concepció Vila*. I particularly want to thank *Mar* for reviewing this text and her constant support and company ("piedra de la suerte").

My sincere thanks to the three external reviewers for reading this thesis and for their excellent comments.

I often think how fortunate I am to have met certain people in my life. Certainly, we only live once, and for me, sharing life's moments with them fills me with satisfaction, and provides that extra energy which at times is needed.

Finally, I want to dedicate this thesis to two people who are no longer with us, my cousin *Antonio Escudero* and my great friend *Armando Ramírez* who will live on in my memory. It is also dedicated to everyone who gets up every morning ready to continue the fight to make the world we live in a better place, ready to help others without expecting anything in return, those who know that health, family and friendship are some of the few really precious things. My thanks to all of you.

José Miguel

Barcelona, 2006

AGRADECIMIENTOS

En los siguientes párrafos quiero expresar mi agradecimiento a todos aquellos que de una forma u otra me han ayudado y han confiado en mí durante un largo camino que comenzó con la realización de la Diplomatura en Estadística, siguió con la Licenciatura en Ciencias y Técnicas Estadística y ahora continua con esta tesis doctoral. Durante este periodo de más de diez años creo haber aprendido un poco más día a día y al mismo tiempo saber que todavía me queda mucho por aprender, o al menos ser consciente de que nunca podré conocer todo lo que quisiera. Sin duda alguna aprender un poco más y poder aplicar métodos estadísticos que puedan ser útiles para otros profesionales de diferentes campos y en especial aquellos investigadores que buscan mejorar la salud de la población, fue la motivación principal que me llevó a realizar esta tesis doctoral.

Estoy especialmente agradecido a *Joan Benach*, *Yutaka Yasui* y *Josep Ginebra* por su gran ayuda en el proceso de realización de esta tesis. Ellos son tanto grandes profesores e investigadores como grandes personas. También estoy agradecido a *Joan* por enseñarme a ver la utilidad y las limitaciones de la ciencia, confiar en mí dejandome desarrollar mis ideas y propuestas en todo aquello en lo que hemos colaborado desde hace ya más de cinco años, ayudarme y preocuparse siempre por mi futuro. También a *Josep* por compartir conmigo diferentes ideas y hacerme ver las fortalezas y limitaciones de los métodos estadísticos. También a *Yutaka*, por estar siempre dispuesto a ayudarme y enseñarme multitud de métodos estadísticos aplicados para mejorar la salud de las poblaciones. También estoy agradecido a él y a toda su familia: *Marcy Winget*, *Niko*, *Osamu* y *Lena* (Today is the day!, Yes..No, word, Do you understand my English?...Yes, I understand your English...so....we can talk, we can talk, we can talk and sing♫ ♫..), por acogerme en su casa como uno más de la familia y cuidarme durante mi estancia en Edmonton (Canada), así como por los grandes momentos que vivimos juntos.

A los miembros del Department of Public Health Sciences de la Universidad de Alberta. En especial, estoy agredecido a *Nicola Cherry* y *Duncan Saunders* por sus

atenciones durante mi estancia en Edmonton y a *Irina*, *Xiufang* y *Doug* por su ayuda en las simulaciones realizadas.

A los miembros del Inner City Health Research Unit del St. Michael's Hospital (Toronto, Canada). Estoy especialmente agradecido a *Pat O'Campo*, *Jim Dunn* y *Piotr Gozdyra*. A *Carles Muntaner*, *Samuel Noh*, *Violet Kaspar* and *Ivonne Hinds* por sus múltiples atenciones durante nuestra estancia en el Centre for Addiction and Mental Health de la Universida de Toronto (Canada). En especial estoy agradecido a *Carles* y *Pat* por su siempre cálido y constante apoyo, y a *Gabriel* y *Daniel* por los agradables momentos vividos juntos.

A Pepe Tapia and Ana Diex-Roux con gratitud por sus atenciones durante nuestra estancia en Ann Arbor (Michigan, USA).

A *Pedro Delicado* por impartir una asignatura enfocada al método Bayesiano. Gracias a ésta me inicié en el doctorado y he podido conocer a grandes compañeros. Por ello también agradezco toda su gran ayuda desinteresada a *Lourdes Rodero*, *Rosa Lamarca*, *Alfonso Buil* y *Ramón Clèries*. En especial a mi gran amigo *Ramón* por compartir conmigo la pasión de aprender día a día un poco más de estadística desde un plano aplicado que sirva para poder mejorar el estado de salud de las personas (Gracias chavalín!). También estoy agradecido a *Celia Martín, Marta Llabres* y *Lupe Gómez* por su ayuda en diferentes cuestiones relacionadas con el doctorado.

A los bioestadísticos del IMIM, IMAS y UPF, juntos formamos un pequeño grupo pero grande en compañerismo y amistad. Estoy agradecido a *Joan Vila*, *Josep Maria Manresa, Helena Martí, Ruben Román*, *Eduard Molins*, *Josué Almansa*, *Albert Sánchez*, *Montse Martínez*, *Gemma Vilagut*, *Angels Pont*, *Ignasi Serra*, *Paco Fernández*, *Alex Amoròs*, *Raquel García*, *Estel Plana*, *Marta Benet*, *Emma Arcos, Mireia Vilardell* y en especial estoy agradecido a *Laura Muñoz* (actualmente en el Institut Català d'Oncologia), *Xavi Basagaña* (actualmente realizando su doctorado en Harvard), *Mercè Comas* e *Isaac Subirana*. También quiero mencionar a mis colegas *Joan Valls*, *Valentín Navarro, Maite Encuentra, Rebeca Font, Yolanda Benavente, Gina Albero, Cristian Bocanegra, David Santos, Jaume Escudero, Ruben Romera, Juan Ramón González, Gemma Castaño, Diego Pangusion, Jordi Real, Aureli Tobías, Anna*

Schiaffino y a mis compañeros de la Licenciatura de Granada *Emilio López*, *Antonio Arquero* y *Miguel Rodríguez-Barranco*.

A los compañeros de la Unidad de Investigación en Salud Laboral. En especial a *Fernando G. Benavides*, *Sergi Jarque, George Delclòs*, *David Gimeno*, *Consol Serra*, *Emily Ahonen*, *Maite Sampere*, *Marcelo Amable*, *Jordi Castejón*, mis compañeras de despacho *Cristina Portolés*, *Montse Vergara*, *Maria Buxó*, *Victoria Porthé*, *Ruth Domínguez* y mis excompañeras *Marisa Martínez*, *Núria Catot, La Noe, Merche Iglesias* and *Sira Gónzalez*.

A mis profesores de la Diplomatura en Estadística de la Universidad Autónoma de Barcelona, *Maria del Mar García* y *Miguel Martín*, por sus asignaturas enfocadas a la aplicación de métodos estadísticos en ciencias de la salud. En especial a *Victor Moreno* por ayudarme a resolver siempre mis preguntas metodológicas.

A mis profesores de la Licenciatura en Ciencias y Técnicas Estadística de Granada. En especial a *José Miguel Angulo*, *Maria Dolores Ruiz Medina*, *Ana Aguilera*, *Mariano José Valderrama* y *Rafael Pérez-Ocón*.

A los colegas interesados en el análisis geográfico de la salud con los que he compartido múltiples pensamientos e ideas durante estos años. En especial a *Miguel Ángel Martínez-Beneito*, *Oscar Zurriaga*, *Antonio López*, *Paloma Botella*, *Gemma Cano, Enric Azlor*, *Mayka Rodríguez*, *Ricardo Ocaña* and *Carmen Sánchez-Cantalejo*.

A los alumnos de la asignatura de Salud laboral II de la Diplomatura en Relaciones Laborales de la Universitat Pompeu Fabra por mostrarme como aprender enseñando.

A *Carme Borrell*, *Maribel Pasarín* y *Santiago Esnaola* por mostrarme, junto a *Joan* y *Carles*, la importancia de estudiar las desigualdades en salud. En especial, estoy agradecido a *Carme*, *Joan Guix* y en general a la Agència de Salud Pública de Barcelona, por estar siempre dispuestos a facilitarnos diferentes fuentes de datos, sobre las que poder aplicar los métodos estadísticos desarrollados en esta tesis y de esta forma poder contribuir a mejorar la salud de las personas.

A los compañeros del Servei d'Informació i Estudis de la Generalitat de Catalunya que conocí durante mi etapa en esta institución. En especial a *Glòria Ribas*, *Roser Bosser, Mara Barés, Anna Puigdefàbregas, JJ Coll, Josep Maria Giné*, *Sergi Cruz*, *Carme Navalón*, *Rosa Gispert*, *Xavi Puig*, *Cristina Rius* y *Glòria Pérez*.

A los responsables de la biblioteca de la Universitat Pompeu Fabra (área del Mar), y sobre todo a *Josep Gibert*, *Marina Losada* and *Rosa Feixas* por su siempre activa colaboración en la búsqueda de información.

A los miembros de los Servicios de Cartografía, Mortalidad, Servicios Censales e Información Estadística del Institut d'Estadístiques de Catalunya (IDESCAT), y en especial la ayuda prestada por *Gelasio Nogueira*, *Jordi Oliveres* y *Dolors Olivares*, y al Instituto Nacional de Estadística (INE), y muy especialmente a *Montserrat García*, *David Goizueta* y *Guillermo Olmo*.

A *Dave McFarlane* por la traducción al inglés de la mayor parte de esta tesis, sus buenas apreciaciones y comentarios. Gracias por tu gran ayuda.

A *Esther Español* por su gran ayuda en el diseño gráfico.

A mis padres, *Juan* y *Laura*, hermanos, *Juan Pedro* y *Marisa*, familiares y amigos. Ellos saben bien lo que supone el esfuerzo de la investigación y del estudio, ellos son siempre los grandes sacrificados. Gracias por comprenderme y estar a mi lado cuando os necesito. En especial quiero estar agradecido a mis padres por enseñarme a apreciar las cosas sencillas de la vida, respetar a los demás como a uno mismo y quererme siempre tal y como soy. A mi hermana mayor por ayudarme económicamente en mis estudios complementando la ayuda que siempre me han dispensado mis padres o las becas de estudios. A mi hermano por compartir conmigo el trabajo en el negocio familiar del Bar Kalamot de Gavà, ayudar siempre a mis padres y transmitirme su admiración por el cine. En especial quiero agradecer a nuestros clientes *Alberto Caparrós*, *Alonso*, *Aurelio*, *Eduard*, *Martín Atenza*, *Manuel Martos*, *Marcelino* y *Armando*, por los buenos momentos que hemos vivido juntos.

A mis amigos de la infancia: *Pedro Peña*, *Jordi Sáez*, *David Martos*, *Javi Moreno*, *Manolo Pascual*, *Jordi Serra* y *Diego González*. A mis compañeros del gimnasio de Karate Shyto-Ryu Cervantes de Gavà, en especial a mi profesor *José Alonso* por compartir conmigo la pasión por el Karate. A mi abuela *Luisa Puertas* y a mis abuelos *Juan Pedro*, *Maria* y *Miguel* en el recuerdo, y en general a mis familiares de Gavà (Barcelona), Murcia, Madrid y del pueblo de El Margen (Granada) por su cálida acogida durante los meses de verano donde fue realizada parte de esta tesis. A mis gatos *Cuqui* y *Mimi*.

A *Mar Torné*, *Carme Farré*, *Esteve Torné* i *Concepció Vila*. En especial quiero agradecer a *Mar* la revisión de este texto y su constante apoyo y compañía ("piedra de la suerte").

Mi sincero agradecimiento a los tres revisores externos por leer esta tesis y por sus buenos comentarios.

Muchas veces pienso en lo afortunado que soy de haber conocido a algunas de estas personas durante mi vida. Ciertamente, vivimos solo una vez y para mí compartir momentos de ésta con muchos de ellos me llena de satisfacción y me da esa energía que a veces nos falta.

Finalmente, quiero dedicar esta tesis a aquellas personas que ya no están con nosotros en especial a mi primo *Antonio Escudero* y a mi gran amigo *Armando Ramírez* que estarán siempre vivos en mi recuerdo. Y va dedicado a todos aquellos que se levantan cada mañana luchando porque este mundo en el que vivimos sea un lugar mejor, que ayudan a los demás sin pedir nada a cambio, a quienes saben que la salud, la familia y la amistad es aquello más preciado. Gracias a todos.

José Miguel

Barcelona, 2006

INTRODUCTION

This thesis consists of two related parts based on the study of health in a geographical region divided in a set of zones (small areas). The first part considers studies based on health information aggregated for each area into which the region under study has been divided. Specifically, it is a disease mapping application, based on generation of an Atlas of mortality in small areas of Catalonia over the period 1984-1998, using empirical Bayes methods. The second part considers an innovative approach, based on an integration of aggregated and individual health data in each of the zones of the region under study, using an estimating equation approach. Specifically, we consider this new approach as an extension of geographical regression.

The elaboration of the first part of this thesis is justified for different reasons. First, health atlases and the mapping of health indicators in general, has demonstrated its great utility in identifying geographical localizations of health problems, in formulation of hypotheses about disease causes, and in monitoring public health interventions. For example, the first atlases of cancer in the United States identified a strong clustering of areas with high rates of mouth cancer in the south east of the country. A subsequent epidemiological study found the clustering to be associated with the habit of chewing tobacco. Furthermore, some authors, by identifying similarities between non-adjacent areas, have managed to find a risk factor common to these locations. For example, the observation of high lung cancer mortality in some coastal areas of the USA has been attributed to ship building activities involving asbestos exposures during the Second World War. Second, most atlases of mortality at the small area level present patterns of relative mortality risk for the most important causes of death using maps with a high level of geographical resolution. These atlases combine many maps of small areas providing information about specific diseases. It is important to choose and combine key information that may have some relevance in the description and atieological study of diseases. This can lead to improvements in the study of health indicators in small area atlases with maps showing time trends in the study region and the geographical patterns in zones within small areas having a large population. The few small area atlases that have included time trend information, have assessed the evolution

of mortality indicators by comparing several maps for different time periods. Despite its value, this strategy does not show important information such as the relative mortality risk evolution of each area compared to the overall time trend of all areas combined. Additionally, this strategy may not be parsimonious in a small area health Atlas where many causes of death are considered, because it is important to combine the key geographical mortality information into a display sufficiently comprehensive to permit the different maps presented in the Atlas to be close enough to facilitate their visual comparison. The first goal of this thesis was to construct a mortality Atlas involving a decomposition of the Autonomous Community of Catalonia into 289 small areas (municipalities or aggregates thereof) and 66 primary health areas of Barcelona city, being a small area but with a large population, for the period 1984-1998. In this Atlas we combine important geographical mortality information into a comprehensive display with specific statistical methods, to obtain the relevant information displayed in the maps. For Catalonia as a whole, these maps presented, using a double-page format, the age adjusted relative risk, significantly high and low relative risk areas, relative risk in Barcelona City with respect to Catalonia and internally with respect to Barcelona, relative risk by age group (0-64 and 65+) and additionally the relative risk evolution over time in each area summarized in an single map, using spatial and temporal information modeled through Bayesian methods. Specifically, the atlas uses a strategy to include both: 1) relative risk evolution throughout the study period of each area compared to the average trend for all Catalonia and 2) the absolute relative risk evolution of each area. To our knowledge, this is the first time that both types of information have been combined in a single map. In addition, this is the first Atlas that presents information about geographical patterns in zones within small areas having a large population such as the cities of a country and includes life expectancy obtained with an empirical Bayes approach.

The second part of this thesis can be useful in epidemiological studies where we include exposure and confounding variables that may have different sources of within and between-population variability. For example, in the study of the aetiology of bladder cancer we can jointly include variables where the within-population variability is higher than the between-population variation, such as smoking status, and variables where the between-population variation can be higher than the within-population, such as chlorinated drinking water. Specifically, analyses of individual disease-exposure data

within a population are useful when exposure of interest varies sufficiently within the population. When the within-population variance of exposure is limited, however, power of the individual-data analysis within a population is reduced. In such situations, aggregated-data analyses of disease data across populations, with a sample of individual exposure data from populations, can be powerful in estimating the exposure effect if between-population variation of exposure is large. Both approaches are useful depending on where the exposure variation exists. However, although we may have knowledge of which variations dominate in each variable, exposure and/or confounding variables with different types of variation can be considered jointly. The second goal of this thesis was to consider a new analytical framework that is a combination of the individual- and aggregated-data analyses, based on an estimating equation approach ("population-based estimating equation" (PBEE) approach). The proposed analysis utilizes strengths from individual data and aggregated data in the estimation of the exposure effect of interest, depending on which of the exposure variations (within- vs. between-population) dominates.

The two parts of this thesis have been structured into eight chapters. The first part, dealing with an application of disease mapping studies occupies chapters 1 through 6. Chapters 1 to 5 tackle the antecedents of geographical studies of health, mainly centering on disease mapping, while chapter 6 considers the application of these techniques in generating the Atlas of mortality in small areas of Catalonia (1984-1998). Specifically, chapter 1 deals with the definition and utility of geographical epidemiology, chapter 2 considers the indicators most commonly used to measure health in the areas of a study region, chapter 3 details certain problems deriving from these indicators in small area studies. Subsequently, chapters 4 and 5 describe the statistical methods, and currently available software, used to control for these problems. Finally, chapter 6 explains the methods used and results obtained in the Atlas of mortality in small areas of Catalonia (1984-1998), and finishes with a discussion on several aspects touched on earlier in the chapter.

The second part relates to geographical regression extensions and covers chapters 7 and 8. Chapter 7 deals with the antecedents and utility of studies based on individual-level data and how such studies may suffer limitations which in turn may be overcome using particular techniques for aggregated data. Finally chapter 8 describes a proposed

new approach that is a combination of the individual- and aggregated-data analyses, and simulations are conducted under different scenarios to show the strengths of the proposed approach in the estimation of the exposure effects of interest. The chapter ends with a brief discussion.

Furthermore, a series of appendices present a complementary set of maps (maps of life expectancy in small areas), the construction of the geographic units, a general SAS program for carrying out some of the analyses of the first part, R programs for carrying out the analyses of the second part and a list of publications derived from the present thesis including communications to congresses, books, book chapters and articles sent to a variety of scientific journals.

The first part of this thesis was partially funded by the Jaume Bofill Foundation (Barcelona), Agency for Administration of University and Research Grants (Government of Catalonia), Municipal Agency of Public Health (Barcelona). The second part was partially funded by a grant from the Fondo de Investigaciones Sanitarias (FIS 03/0586) and the BBVA Foundation. In addition, the Occupational Health Research Unit was recognized as a standing research group by the Government of Catalonia (2002-2004: SGR/0005; 2005-2008: SGR/00699) and Yutaka Yasui is supported by the Canada Research Chair Program.

Finally, we hope that some of the methods and topics employed may be of use to researchers who want to improve the study of health in space and time.

CHAPTER 1

Geographical epidemiology

"The most extensive data maps, place millions of bits of information on a single page before our eyes. No other method for the display of statistical information is so powerful"

Edward E. Tufte

1.1 Introduction.

Science is a process of continual knowledge gathering which allows us to better understand social realities and natural phenomena which are often unobservable such as human societies, or genes[1]. The scientific approach is present in very diverse disciplines such as sociology, statistics, geography or medicine. Related to these and other disciplines, public health uses scientific knowledge to study, prevent and act on specific problems harmful to health of one or more human populations[2]. When interest centers on describing, quantifying and explaining the variation or distribution of health (for example, illness or death) and factors potentially harmful to it (exposure to the so called risk factors) in a specific geographical area, the term geographical, or spatial, epidemiology is used[3,4].

Among the most quoted examples of geographical epidemiology, we may note the research conducted by John Snow in 1854 which shows how a simple spatial study can be of enormous relevance for public health. The most important accepted version of Snow's research about cholera explained that in an initial stage Snow represented, on a map of London, the deaths due to cholera during the years 1848-49 and 1853-54. From this geographical representation, he was able to ascertain that the majority of deaths

were concentrated near one of the two companies distributing water in the city (Southwark and Lambeth). In a second phase, he showed by comparison that the number of deaths per thousand inhabitants among people whose water was supplied by the Southwark company was much higher than among those whose water was supplied by the Lambeth company. In this way he determined that very probably cholera was spread by contaminated water[2,3,5].

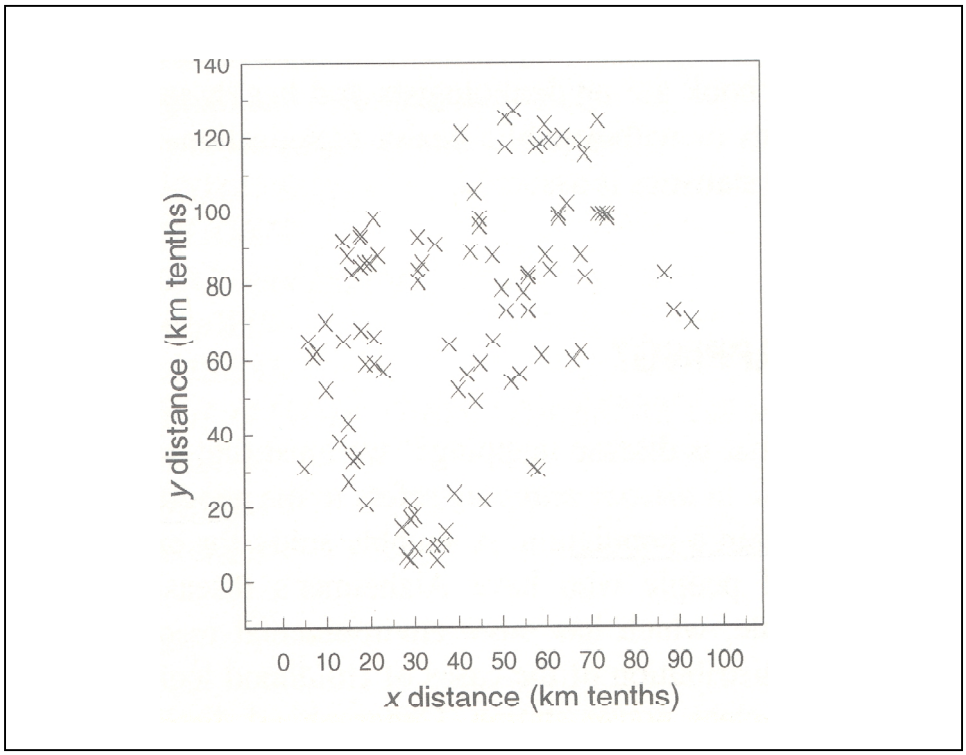
In general, geographical epidemiology studies yield results which generate hypotheses, or study the etiology or the causes and risk factors which may be associated with different states of health. Subsequently the results obtained may be studied more precisely through more detailed epidemiological analyses. Thus, public administrations can use this knowledge to establish priorities for specific geographical areas and contribute to a more adequate territorial distribution of social and health-related resources[6].

1.2 Geographical analysis levels in health studies.

In order to conduct any epidemiological study we must quantify health and the risk factors in a group of individuals. In carrying out a geographical study we begin with some geographical area of study in which we can obtain information at two geographical levels of analysis[4,7].

In the first geographical level we may obtain the exact spatial localisation of the information of each individual in points or coordinates of the study area[7,8]. For example, imagine that we are interested in studying mortality in a set of individuals in a particular geographical area over a fixed period of time. Those individuals who die will be denominated cases and those who do not, non-cases. In this example we can determine exactly the geographical coordinates or spatial localisation at which each case and each non-case occurs. Figure 1.1 is a diagram of the exact geographical location of the deaths due to bronchitis in an area in eastern Scotland for the period 1966-76[8].

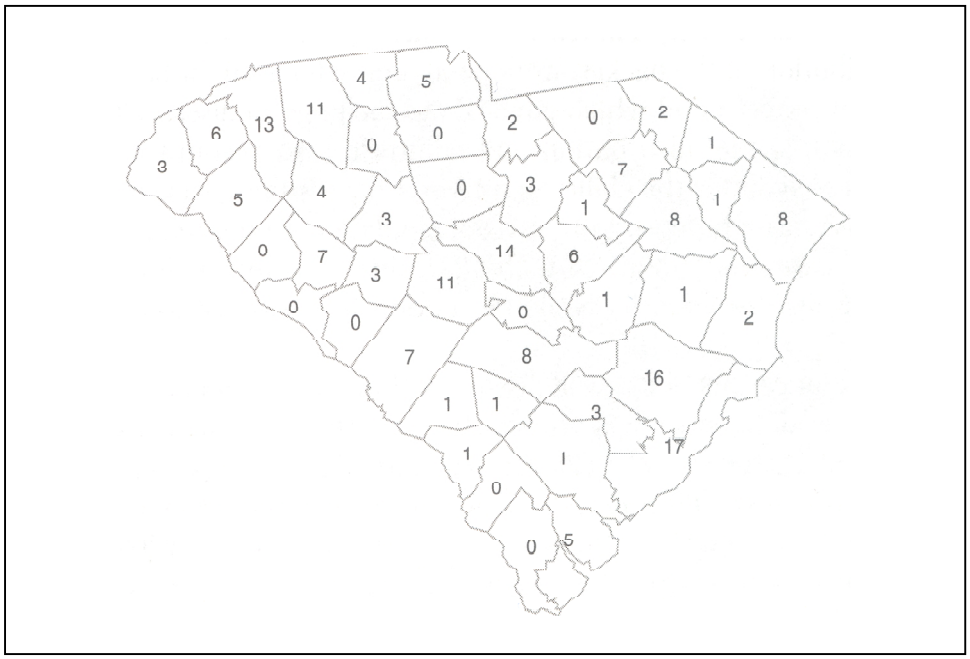
Figure 1.1 *Bronchitis deaths in an area of eastern Scotland, 1966-76.*



In the above example, we can study theoretical mechanisms that could explain the occurrence of the disease cases using statistical techniques that involve point process models[7,8,9,10]. Also, there are statistical techniques that use data with the exact location of a covariate of interest in a geographical region. These are geostatistical techniques that are applied to predict the covariate under study in points of the region under study[10,11,12].

In the second geographical level, we consider that the study area is divided into a set of mutually exclusive zones in which geographical information on health and exposure to risk factors may be aggregated over each of these zones[7,8]. Lets consider again the above example in which we want to study health through mortality of a set of individuals. In this situation we now do not have the exact coordinates for each case and non-case in the study area, but instead we will have aggregated information about the numbers of cases and non-cases in each of the geographical zones which divide up the study region. Figure 1.2 is a diagram showing deaths due to congenital malformations in areas of South Carolina, United States, for the year 1990[8].

Figure 1.2 *Congenital malformation deaths in South Carolina, 1990.*



The first part of this thesis and the rest of this chapter will be based mainly on this second geographical level, in which our study region is divided into a set of mutually exclusive geographical zones for which we have aggregated information on health and exposure to risk factors. For example, our study region could be Spain, and the areas into which it is divided could be the autonomous communities. For each autonomous community we may have as the health indicator the number of deaths with respect to the population of inhabitants, and the percentage of unemployment as the risk factor.

It should be noted that in this second case, we may also have individual information about health outcomes (disease or death) and individual covariates within the set of mutually exclusive zones forming the area under study (see section 1.6 in this chapter). Although we have individual information, we don't know the exact spatial localisation of each individual in the study area, we only know to which of the mutually exclusive areas that form the study region each of them belongs. This particular case will be the focus of the second part of the thesis presented in chapters 7 and 8.

1.3 Why use data aggregated over geographical areas?

The fact of using aggregated data may be conditioned by two fundamental reasons: on one hand, the availability of information at the aggregated level, and on the other the particular interest of studies which consider as the analysis unit the geographical areas into which the study areas is divided[13,14]. We will analyse both of these in greater detail.

The first reason arises from the impossibility of obtaining individual level data. At least three reasons could account for this absence of information: 1) collection of the data is not feasible because the information sources available do not contain it, 2) the data may be available but somewhat, or highly, incomplete, and 3) the data may be provided in aggregated form for reasons of statistical confidentiality. As an example of point 2), in Spain it is generally not possible to conduct studies on social inequalities in health using social class since the available sources of data are very incomplete. Thus using the declaration of occupation in the death register it is not possible to study the distribution of health by social class at individual level. There are certain exceptions such as Barcelona and Madrid where record linkage between the registries of death and municipal censuses have permitted the study of social inequalities in mortality at individual level using educational level[15]. As an example of point 3) we may cite a geographical study conducted in Spain on mortality due to specific causes in the period 1987-1995. The National Statistical Institute only provided the total number of deaths in geographical areas constituted by municipalities in which there were at least 3500 inhabitants in 1991. Those municipalities not achieving this figure had to be aggregated[6,16].

With regard to the second reason, the study of a set of geographical areas, such as will be described when we define the types of studies in spatial epidemiology, is of considerable relevance in public health. For example, we know that those areas having higher rates of unemployment, or having poorer quality housing (more overcrowding), also have higher mortality rates[17]. Detection of areas with high mortality can assist public administration to establish social and health policy appropriate to each need and which help to reduce mortality in the most problematic areas.

1.4 Types of studies in geographical epidemiology.

In the context of geographical epidemiology we may be considered at least three types of studies[3,4,7,18], namely, disease mapping, ecological geographical regression and disease clustering studies. Other authors also include the so-called migration studies[3]. In what follows we describe the objective of disease mapping and ecological geographical regression studies, and we will note how they may share, as we will see shortly, both similar proposals, and similar, specifically developed, statistical methods. The present document will center mainly on studies using disease mapping, readers interested in disease clustering studies can review the chapter about clustering, cluster detection and spatial variation in risk in Wakefield, et al[19]. It should be noted that results presented in the examples have been obtained using Bayes procedures that will be explained in chapter 3.

1.4.1 Disease mapping.

The general aim of these studies is to obtain an estimate of particular summary measures of health for each of the geographical areas into which the study region is divided. In this way we may compare areas in terms of health indicators. They are mainly descriptive studies, and are based on using maps to represent health indicators. As disease mapping studies use groups or geographical areas as the unit of analysis they are considered a type of ecological study. Obtaining these summary measures for each geographical area, and by representing them on the corresponding maps, allows us to tackle at least six objectives[3,4,18]:

- Reveal spatial patterns in the health indicators, principally risk of disease or death, in a set of geographical areas.

- Compare maps of the health indicators with maps of potential risk factors.

- Determine areas of high and low risk of disease or mortality.

- Permit detection of aggregated areas with excess risk of disease or mortality.

- Study time trends in health indicators for a set of geographical areas.

- Compare maps of health indicators for different causes of death.

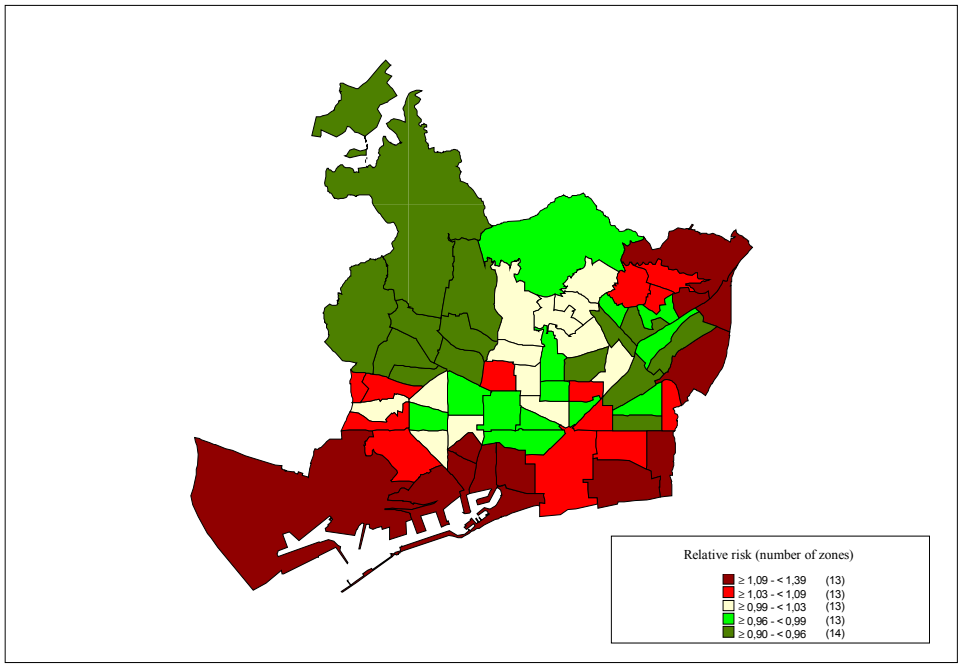
Each of these aspects is explained in more detail below.

1.4.1.1 Spatial patterns in health indicators.

In order to illustrate the utility of disease or mortality maps for revealing spatial patterns let's consider the following example: we can obtain the risk of death, for each of the 66 primary health care areas of Barcelona city, defined as the number of deaths with respect to the number of inhabitants taking into account that the distribution by age in each health area is different. We may then classify the health areas into 5 groups, on a scale of low to high risk. We now assign a colour to each group and represent the health areas on a map of Barcelona where each receives the corresponding colour. In this way we may visualise how the health areas of highest mortality are concentrated along the coastal fringe[13] (Figure 1.3). The method described, in which indicators of exposure and health are categorised into groups which are assigned a colour is known as the choropleth method[20].

The most notable application of this technique is the creation of so-called mortality atlases in which a large collection of maps for the different causes of death are presented. Notable examples include those produced in the United States[21,22] and in Spain using provinces[23] and municipalities or aggregates thereof[6] and the atlas developed in the autonomous community of Valencia[24,10].

Figure 1.3 *Mortality inequalities among men in Barcelona, 1987-1995.*



1.4.1.2 Comparison between maps showing health indicators and risk factors.

Obtaining maps with health indicators for subsequent comparison with maps of potential risk factors allows us to descriptively evaluate a possible relationship between the two, which may give us clues about the etiology of a particular disease[3,4]. For example, figures 1.4 and 1.5 present maps showing the municipalities of Catalonia (or aggregates thereof). We may observe that there is a certain similarity, particularly in the coastal area, between the map of mortality risk and the map of percentage unemployment[13]. Unemployment is considered a risk factor for health, related with social exclusion and relative poverty, and falls in the group of so-called indicators of material deprivation[25].

Figure 1.4 *Percentage of men unemployed in Catalonia, 1991.*

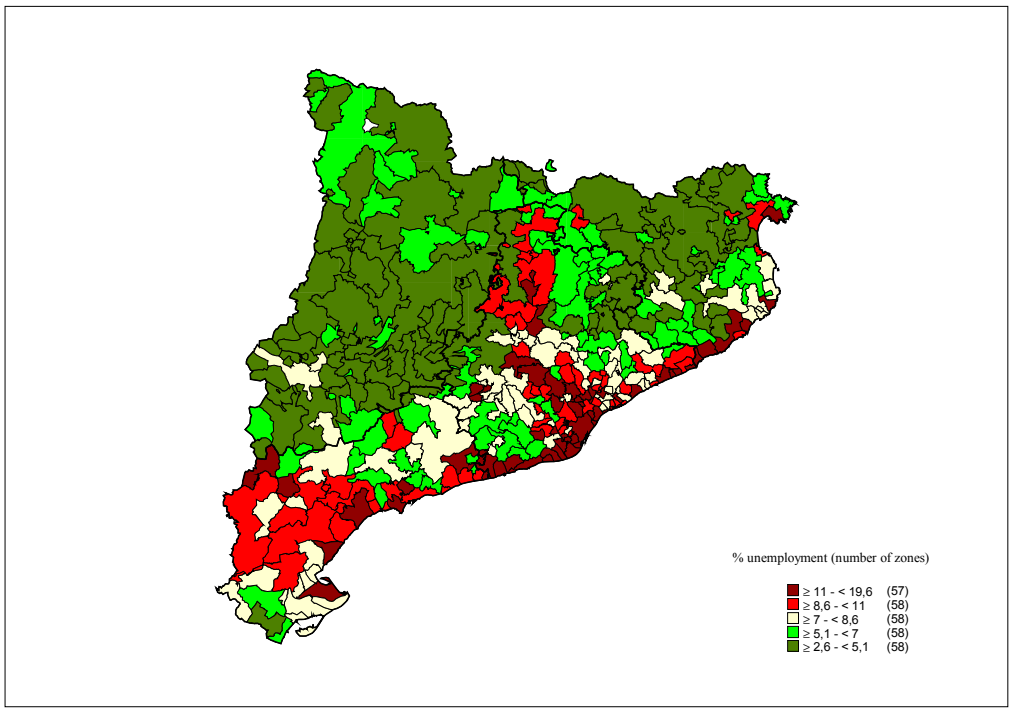
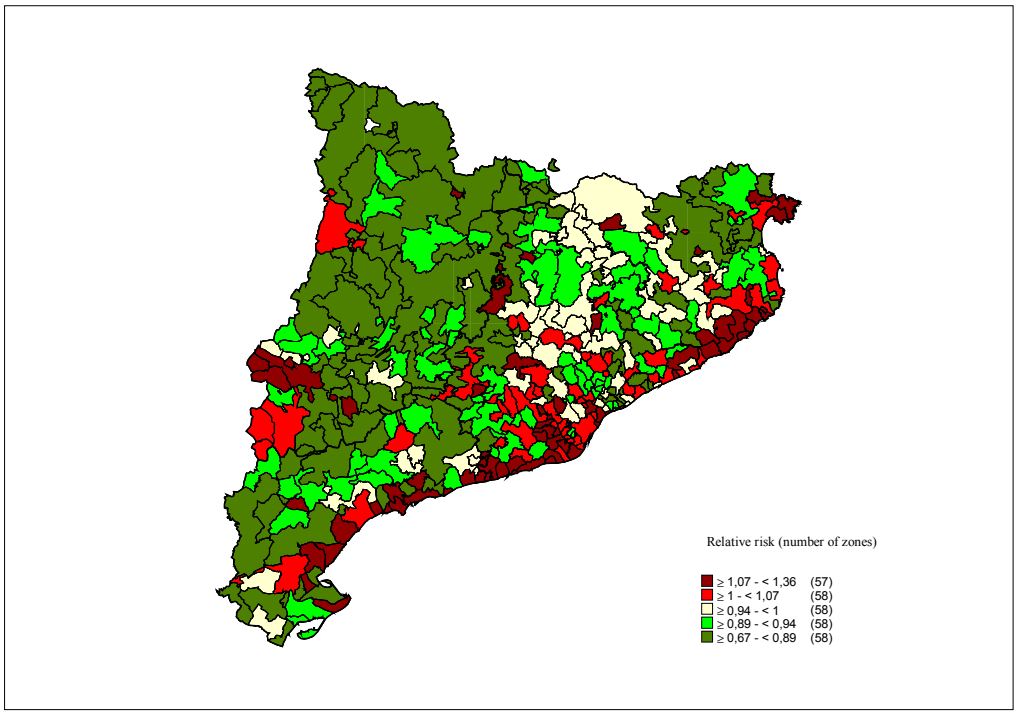


Figure 1.5 *Mortality among men in Catalonia, 1987-1995.*

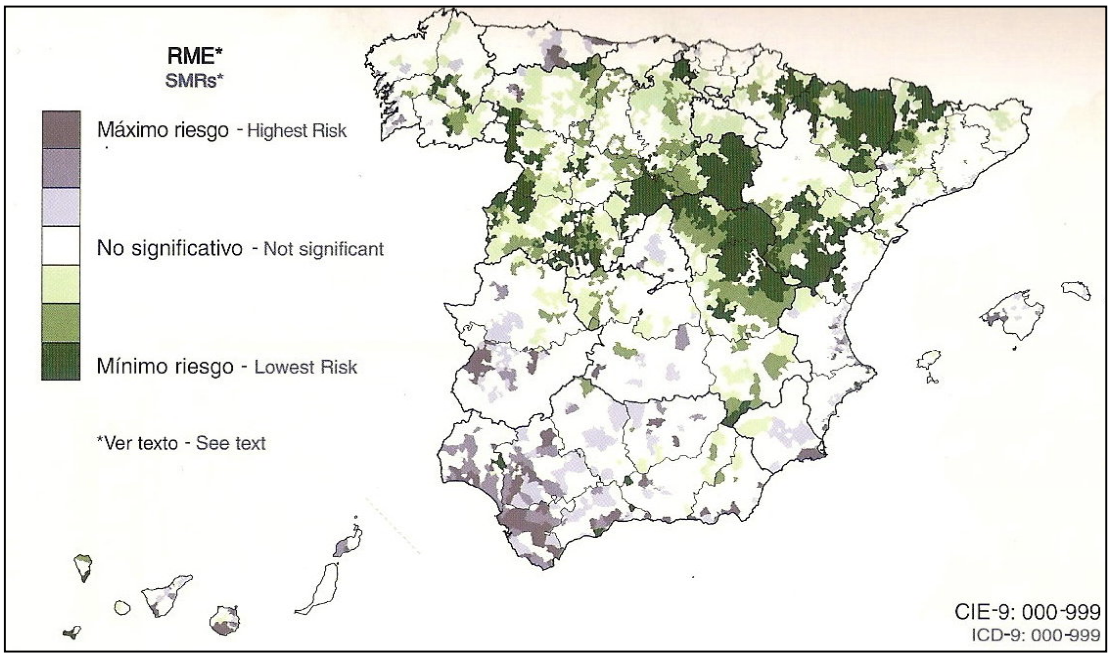


In a subsequent section, describing geographical correlation studies, it will be shown that this hypothesis may be evaluated from a statistical perspective through the use of statistical regression models.

1.4.1.3 Visualising areas of high and low risk of disease or death on a map.

This application of geographical studies allows the areas with the highest and the lowest health indicators to be located on a map. We can also carry out certain types of statistical tests which tell us whether the observed difference between the risk on one area and that of some reference population with which we compare it is statistically significant, or not just due to chance. It should be noted that in this latter case the descriptive information shown on the map is combined with a series of statistical hypothesis tests. Figure 1.6 shows the areas with a mortality risk statistically higher compared with the reference of all-causes mortality for Spain as a whole[6].

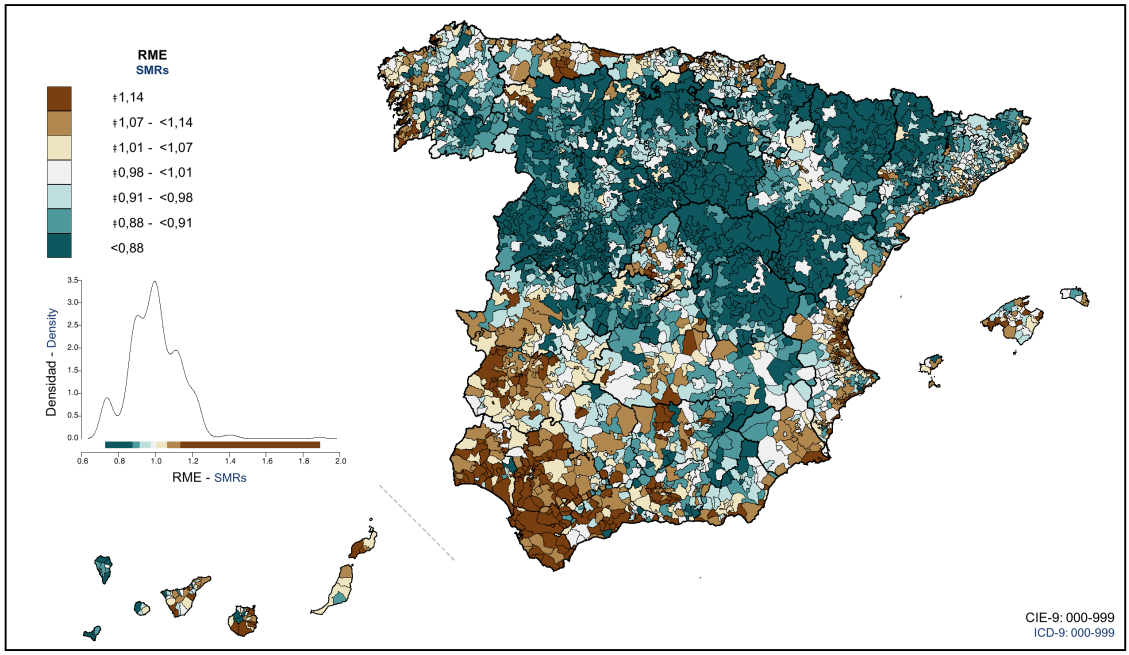
Figure 1.6 *Areas of high (low) risk among men, all causes of death, 1987-1995.*



1.4.1.4 Detecting aggregated areas with an excess risk of disease on a map.

Through the visualisation of health indicators on a map we can detect whether the areas of high risk form groups in particular geographical locations or areas. Thus we can detect clusters of areas with elevated risk of disease or death. Figures 1.6 and 1.7 show a clear clustering of areas with high mortality risk in the south-west of Spain[6].

Figure 1.7 *Mortality risk among men for all causes of death, Spain, 1987-1995.*



1.4.1.5 Time trends in health indicators for a set of geographical areas.

In order to study time trends of health indicators in small areas we may opt to construct several maps for different periods of time. For example, figures 1.8, 1.9 and 1.10 present the distribution of life expectancy for men in small areas for the periods 1990-1992 (figure 1.8) and 1996-1998 (figures 1.9 and 1.10)[26]. Figures 1.8 and 1.9 show that the geographical distribution of life expectancy has not varied greatly between the two periods. On the other hand, by applying the scale for life expectancy of

figure 1.8 (period 1990-1992) to figure 1.10 (period 1996-1998) we may observe that although the distribution of mortality is similar in the two periods, there has been a general rise in life expectancy.

Figure 1.8 *Geographical distribution of life expectancy in men in small areas. Spain, 1990-1992.*

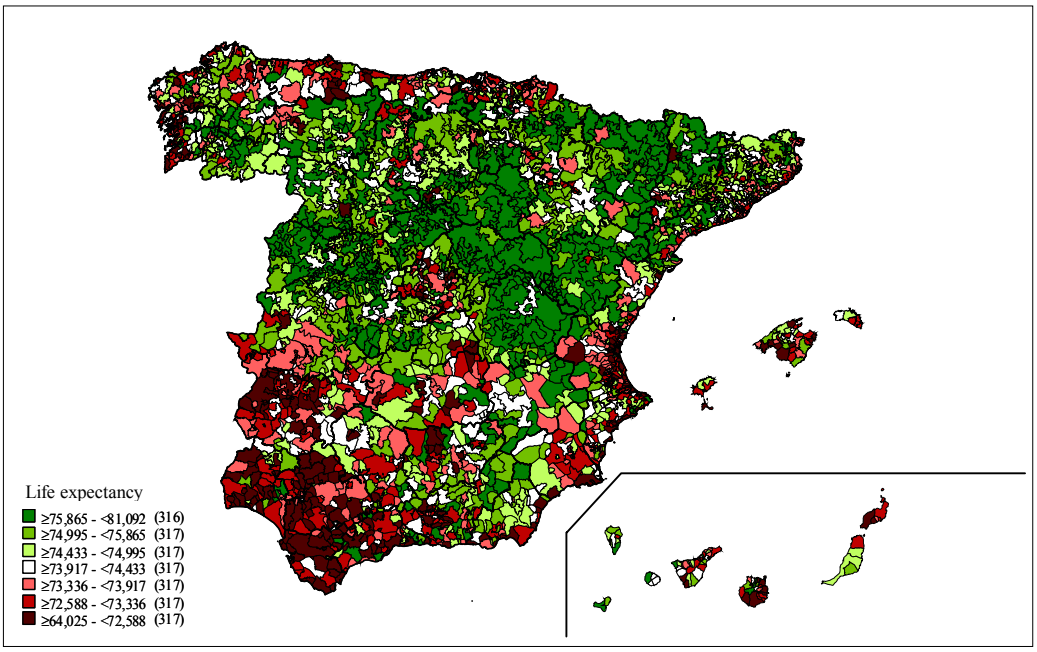


Figure 1.9 *Geographical distribution of life expectancy in men in small areas. Spain, 1996-1998.*

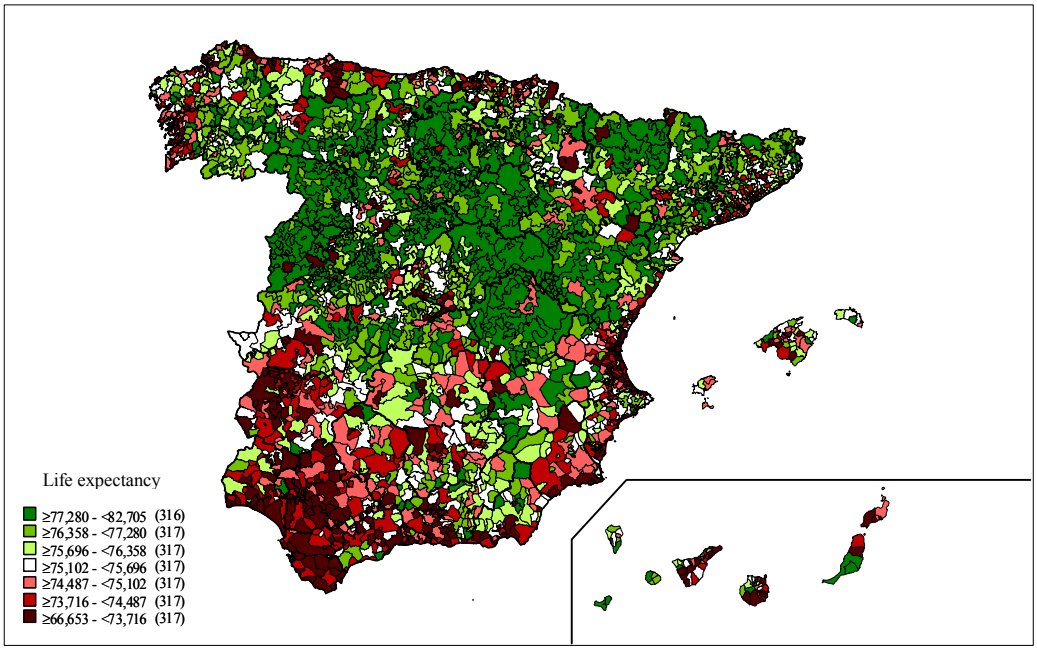
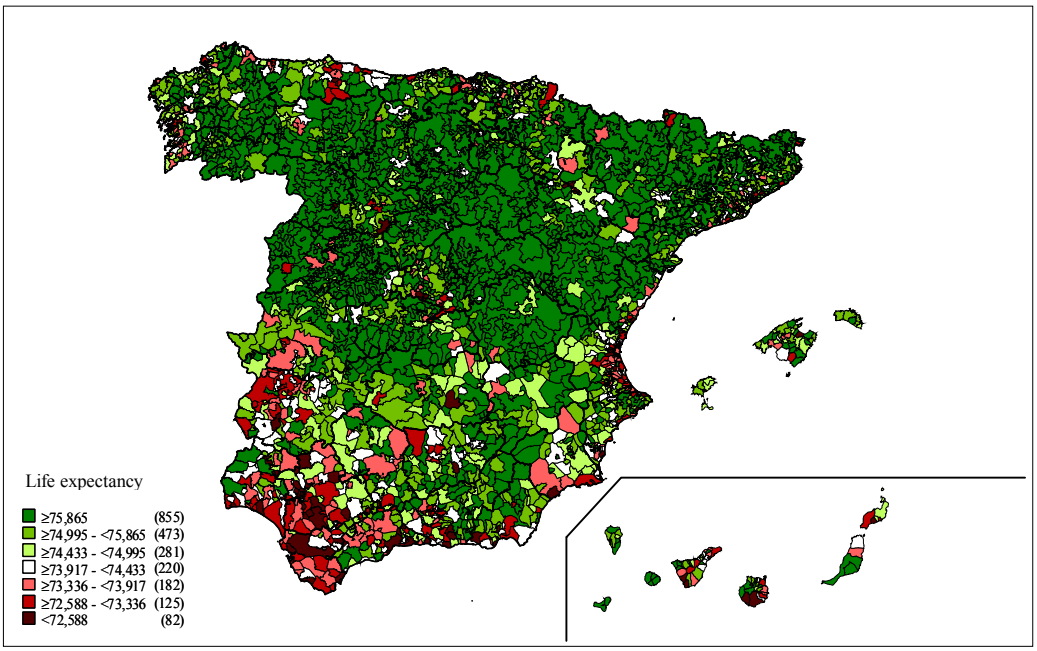


Figure 1.10 *Geographical distribution of life expectancy in men in small areas (with life expectancy scale from 1990-1992 period). Spain, 1996-1998.*



Another method of assessing time trends in health indicators for small areas will be shown in chapter 6, using a single summarizing map.

1.4.1.6 Compare maps of health indicators for different causes of death.

The comparison of maps of health indicators for different causes of death may lead to generation of hypotheses regarding the aetiology of the diseases. According to Linda W. Pickle [6]: "*more clues about the etiology of a disease and its burden on the population can be gleaned from a comparison of many different causes of death. For example, similarities of patterns on maps of lung cancer and other respiratory diseases might suggest the presence of an airborne pollutant*".

1.4.2 Ecological regression studies.

Ecological regression studies, a particular type of ecological study, in the context of geographical epidemiology, may also be referred to as geographical correlation studies. When we speak of regression or correlation the idea is one of studying the relationships between two or more variables. Thus, these studies attempt to determine how health indicators vary in relation to the variations in particular indicators of exposure taking some set of groups as the unit of analysis.

According to Morgenstern the aim of ecological studies is two-fold[27]: 1) to generate or test etiological hypotheses (explain why the disease occurs) and 2) assess the efficacy of interventions in the population (test whether prevention policies to reduce the numbers of cases of the disease and promote health have had any effect).

In the context of geographical epidemiology the groups will be geographical areas and we will examine the relationship between geographical variations in the disease and variations in the degree or levels of exposure to a particular risk factor[3]. Risk factors can be some environmental agent, particular factors of lifestyle such as diet or the above-mentioned indicators of "material deprivation". Figure 1.11 shows an ecological regression in areas of Catalonia corresponding to municipalities (or aggregates thereof) which relates the risk of dying with the percentage of unemployment grouped in quintiles where the lowest quintile represent the lowest level of unemployment[13]. From the figure it may be appreciated that there is a upward trend in mortality excess for those areas with the highest rates of unemployment.

Figure 1.11 *Excess mortality (percentage) and 95% confidence interval by quintiles of unemployment in rural areas (under 20,000 inhabitants). Men, Catalonia 1987-1995.*

1.5 Advantages and limitations of ecological studies.

As mentioned above, disease mapping and ecological regression are types of ecological studies. Ecological studies present a series of advantages and disadvantages, briefly covered below, with respect to studies which I will call analytics, based on data about individuals.

1.5.1 Advantages of ecological studies.

The use of aggregated data, as mentioned above, could be conditioned by the impossibility of obtaining them at individual level since the available data sources may not have them, although the information could be available for groups. Therefore this information will not be analysable at individual level, but only at ecological level, something which must be seen as an advantage. This lack of data may arise through two situations: 1) although it might be possible to obtain such data at individual level, the information sources have not collected it, except at group level, 2) it is impossible to obtain an individual measure of the data, but a group measure is possible. In this second case the variables are known as integral variables and as examples we may cite certain types of laws, political systems or environmental variables[28,29].

Ecological studies tend to be low cost[14] since the aggregated information is available in many information sources and is easy to access and manipulate.

Furthermore they are more suitable for evaluating the efficacy of population interventions where the aim is to evaluate the impact of collective (and not individual) actions upon health such as a new program for the prevention of some particular disease[27]. For example, we can draw a simple graph evaluating the time trend in the annual mortality rate for a particular disease over a period of 10 years. Suppose that in the middle year (year 5 of 10) a specific intervention aimed at reducing mortality was implemented. If this intervention has had an effect on the population the graph will show a declining trend after the middle year. If we want a more precise assessment of whether there is a point of inflection we can apply statistical techniques developed for this purpose such as the so-called joinpoint models[30].

Analogously, as mentioned, geographical studies allow us to detect in a simple, quick and practical way, areas with high indices of mortality and risk factors. This can aid public administrations in establishing social and health policies suited to each case, helping to reduce mortality in the most needy areas and contributing to a more adequate territorial distribution of social and health resources[6].

On the other hand, in some cases the measurement of particular variables in individuals can lead to measurement errors. When such measures are aggregated and analysed at group level we achieve a reduction in the measurement error when the number of individuals per group is stable[27]. Aggregation of individual level variables eliminates correlations within the observations in each group since we obtain a single summary measure for each group. It should be noted however that there is a danger of falling into the trap of the "ecological fallacy", which will be defined in the next section.

In other cases, studies at individual level cannot detect effects due to exposure to particular risk factors when these present very little variation within groups, yet they may present variations at group level. For example, if we study the relationship between breast cancer and diet in different geographical regions we find that diet varies little within the geographical areas, but between different areas the diet patterns may show greater variation. In this latter case so-called aggregated data studies are also used, a type of study not considered ecological even though the units of analysis are groups. Aggregated data studies, unlike ecological studies, are centred on the associations at individual level and can control for factors which may modify such associations[31]. This type of study will be reviewed in chapter 7 and will be an important element in the second part of this thesis.

1.5.2 Limitations of ecological studies.

The main limitation of ecological studies is that it is impossible to make inferences at individual level when that is our aim. This inconvenience is due to not knowing the joint distribution of health and exposure in the individuals in each

group[14,28]. In other words in these studies the starting point is data on health and exposure obtained for each geographical area through summary measures, unlike individual-level studies in which disease and exposure is measured in each individual. For example, imagine that we want to relate some disease with the fact of not working as a risk factor. Although we could determine whether each individual was ill or not, we will only be able to have the percentage of unemployment in the geographical area of residence, so we cannot have for the whole area simultaneously whether each individual in the area was ill or not and whether he was unemployed or not.

If we consider a summary measure of exposure at area level it is a reflection of the individuals who live there, and if we assign the same value for exposure to all the individuals in the area we could be committing an error since the individuals within each area may present different levels of exposure. This fact means that we can only draw conclusions at group or area level to avoid falling in the so-called ecological fallacy. The ecological fallacy consists of considering that the relationship obtained between disease and the risk factor at group level also occurs in each individual, in other words making inferences at individual level when the data analysed belong to groups[27,28,29]. The error or bias arising from assuming that the effects of the variables obtained at group level are the same at individual level is known as ecological bias.

Following the example expounded by Diez-Roux[28,29], imagine we want to evaluate the relationship between traffic deaths and income level taking a set of countries as the analysis unit. Upon doing an ecological regression we may find that the higher the income the higher the mortality from traffic accidents, however, if we perform this comparison for individuals we will find that the situation is reversed, since as individual income rises, traffic accident mortality falls.

In the above example the relation between traffic injuries deaths and exposure was inverted, however even if the relationship was in the same direction, the bias remains in the sense that the magnitude of the effect is usually different, greater in the ecological study than in the individual level study[27].

It should be noted that at individual level there is the possibility of committing the so-called atomist error which consists in making inferences about groups when the data analysed are on individuals[29].

In a subsequent section dealing with geographical area size, it will be explained how certain geographical studies using very small areas, the probability that group or area data reflect individuals can be raised. By taking areas or other units which are small we can reduce the inference problems inherent in ecological studies[27].

1.6 Ecological studies and multilevel studies.

As mentioned above, the fact of using studies with aggregated data can be a result of the available information since mortality and exposure data are available in aggregated form for areas but not at individual level. If information was available at individual and at geographical area levels we could consider multilevel studies in which there is a hierarchy of levels in the data. For example in our case individuals would represent the lowest level and each geographical area into which the individuals are grouped would represent a higher level in the hierarchy.

Using these models we can consider simultaneously group and individual variables in order to make inferences a individual level controlling for the correlation structure in the data due to the organization of the individual within groups[28,29,30,31,32,33]. In studies of geographical correlation we can conduct multilevel studies using the group or geographical area as the lowest unit and with a hierarchy in which these areas are grouped together into larger geographical units[7].

Multilevel models are also known as random effects models since they control for the correlation structure through random effects. These models are useful for both data organised in groups and for repeated measures or longitudinal data. There is also another type of approach suitable for this type of data that permits controlling for intra-group correlation by establishing a structure of covariance or of correlation between the observations of the same group instead of using random effects. These approach use an estimation method based on generalised estimating equations (GEE)[34,35,36,37]. In the

second part of this thesis (chapter 7 and 8) we will return to random effects models and the estimating equation approach.

1.7 Why use maps to represent statistical information?

As we have seen in several examples, the use of maps instead of presenting the information by means of descriptive tables is very useful above all when we are dealing with a large number of geographical areas[3]. I present below some quotes from well known authors revealing their opinions about the use and utility of maps:

- Edward R.Tuffte: "*Examining the scatter over the surface of the map, Snow observed that cholera occurred almost entirely among those who lived near (and drank from) the Broad Street water pump (...). Of course, the link between the pump and the disease might have been revealed by computation and analysis without graphics, with some good luck and hard work. But, here at least, graphical analysis testifies about the data far more efficiently than calculation*"[38].

- David English: "*Statistical tables, while able to present more data than maps, cannot easily convey these spatial patterns, and so are less comprehensible or accessible means of presenting geographical data*"[3].

- Linda W. Pickle: "*many epidemiologists questioned the utility of mapping small area rates. After all, the data had been published in tabular form years before! However, geographic patterns in the data could not be discerned from alphabetized listings of small area rates, so many surprising findings occurred after the publication of the maps*"[39].

1.8 Geographical unit of analysis: why use small geographical areas?

In order to be able to study the spatial distribution of health and particular risk factors in a set of geographical areas we can consider different levels of precision. For

example, health in the various regions of Spain may be studied through a summary measure or indicator which quantifies the risk of death. Thus we may obtain mortality indicators for each autonomous community or by provinces (Figure 1.12) and compare them to ascertain which present high mortality. Although the information provided by this perspective is useful and of interest for social and health policy planning, we may consider that both provinces and autonomous communities are large areas composed of smaller ones. These smaller geographical units may also present different patterns of the mortality indicator which go undetected when studying them aggregated into the larger units of provinces or autonomous communities. Thus to obtain mortality indicators at a more detailed level of precision we could consider smaller areas formed from the municipalities of Spain (Figure 1.13).

According to Elliott and Cuzik there is no exact definition of what constitutes a *small area* since this depends on the context and the number of cases of disease observed[40]. For example, in the context of the above example, municipalities represent small areas with respect to the autonomous communities or provinces, while city neighbourhoods or the census tracts of any town represent areas even smaller than the municipality. This subdivision into neighbourhoods or census tracts may be of considerable interest in the study of large capitals such as Barcelona or Madrid in Spain. Note that in this example an area is considered smaller when the geographical territory it occupies is less.

Regarding the number of cases, Elliott and Cuzik define a small area as follows: "as a rough guide, any region containing fewer than about 20 cases of disease can be considered a small area". However these 20 cases are in relative terms since they also say that the area they define as small will depend on the rate of disease or death, the period of time studies (number of years) and the population density[40]. Furthermore they also consider that except in certain specific cases a minimum population size is also required, and suggest a figure of not less than 10000 inhabitants.

From the statements made above it can be deduced, without do a formal definition, that a small area is a geographical area in which we may have a small number of cases of disease and small population, that it may be a geographical territory of reduced size, which allows us to study the spatial variation in health and particular

associated risk factors, with a more detailed level of precision than in the case for other larger geographical regions.

When ecological studies were defined, we described that it was a mistake to consider for a geographical area that its summary exposure measure reflected that of its inhabitants. This was due to the fact that the individuals within each area may present different exposure levels. Another advantage of working with small areas with respect to other larger geographical regions is that the smaller the areas analysed, the closer will be the measures of exposure and health to the real values of individuals[3].

In recent years thanks to advances in the production of statistical information, in computing and in the availability of geographical information systems, it is now possible to in many countries to conduct studies in small areas with a high level of resolution.

Figure 1.12 *Maps of the autonomous communities and provinces of Spain.*

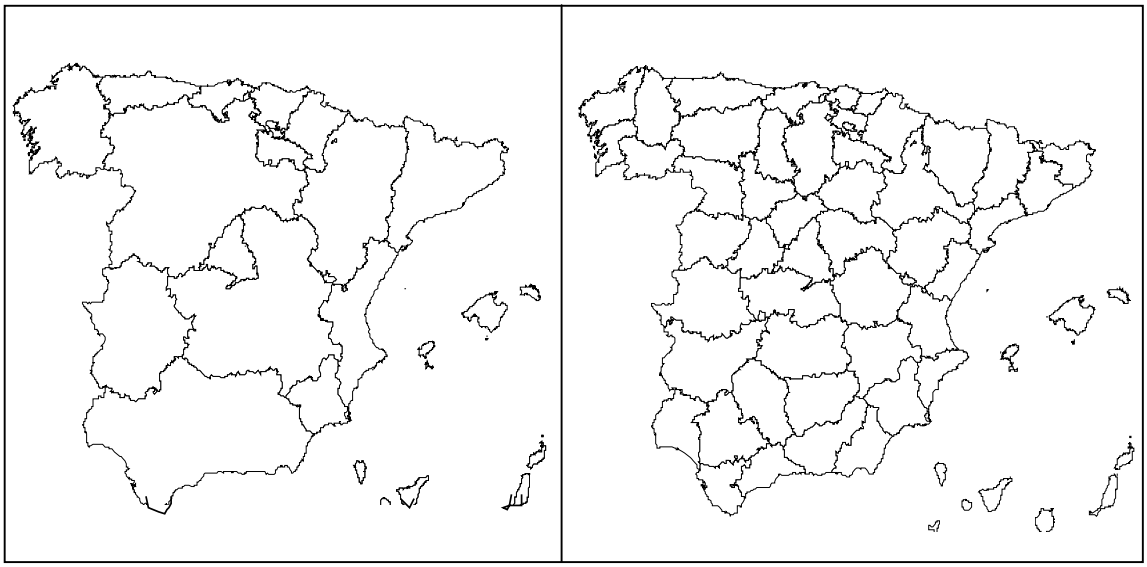
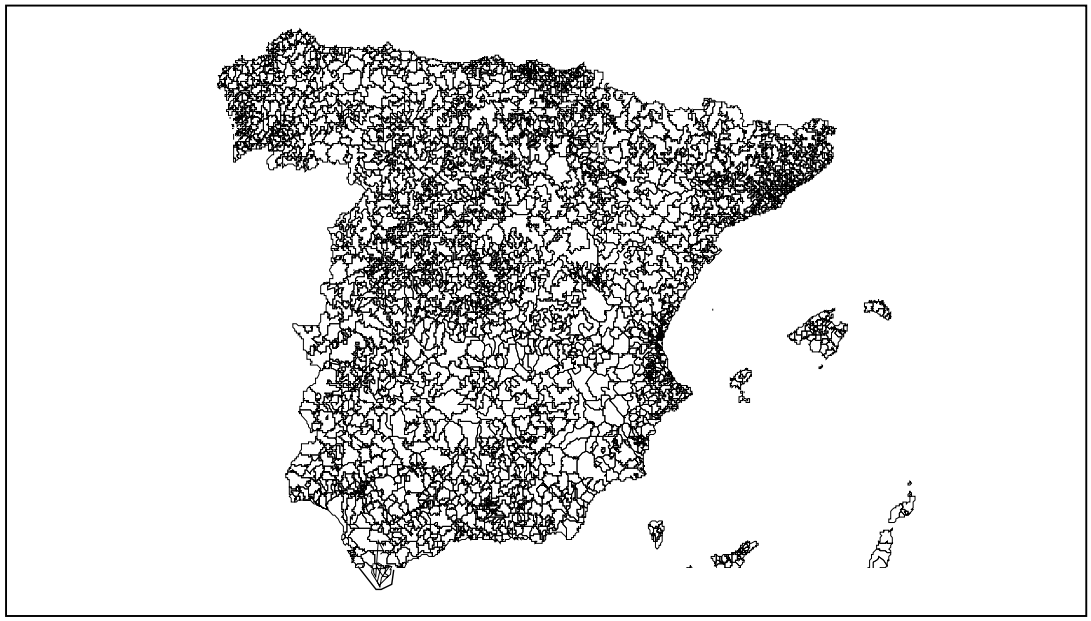


Figure 1.13 *Maps of the municipalities of Spain or aggregates of them.*



1.9 Quality of data and forms of graphical representation.

As in any study we must take data quality into account before performing any statistical analysis. Because of their greater accessibility and good general quality, mortality data are usually used in disease mapping. In order to calculate health indicators in each geographical area, we must know the number of cases and the population of inhabitants stratified with respect to the different variables of interest. Knowing the quality or limitations of the data is of interest in geographical studies or indeed in any other type of study. In Spain studies exist which have found that the quality of mortality data is good[41,42]. On the other hand, population data mainly come from the national or local censuses. Census data are obtained by the National Institute of Statistics, and with a good general level of quality. However, local census data may overestimate the population figures since they are self-reported by the town councils of each municipality and there is a relationship between the number of inhabitants and the subsidies conceded, subsidies being higher when the number of inhabitants is larger.

In relation to representation on maps, other authors have dealt with the limitation of the use of maps and how they may even lead to visual lies, for example when the

geographical areas are not all of a homogeneous size[43,44]. Also, the colours used may influence the visual perception, for example stronger tones for the higher risk groups and softer tones for lower risk may highlight higher risk areas more than lower risk ones. On the other hand, specific studies have been conducted which allow to discern which colours are the most suitable, the main objective being to make the maps easily visible to people with vision defects[45]. Thus the representation of data through maps can also condition the results finally obtained.

CHAPTER 2

Health Indicators

"*Research questions are always formulated at conceptual level and require such notions in order to be expressed intelligibly. The answers obtained, partial or otherwise, must of necessity be summarized*"

Luis Carlos Silva

2.1 Introduction.

In order to determine the health status of a population we must quantify it using a set of summary measures or indicators. These may be absolute numbers or relative measures such as proportions, rates or ratios[46]. Among the most used indicators are those reflecting mortality in the population. The remainder of this document will refer to mortality indicators.

In our case we will consider that we have information on the number of cases of death and of the population in each geographical area in a particular period of time as functions of different characteristics or variables such as sex or age. From the quotient between the observed cases and the population of the geographical area we obtain the relative measure which I will call mortality incidence rate, or simply mortality rate. Incidence rates capture phenomena related with time and measure the speed of appearance of new cases of disease or death in relation to the size of the population[47]. For more on the specific terminology of each of the various measures of disease in the context of epidemiology and public health, one may consult a variety of texts[47,48]. In what follows I will describe the concept of crude and specific rates.

2.2 Crude mortality rate.

Suppose that our study region is composed of n mutually exclusive geographical areas. The crude rate for the i-th area (i=1, ..., n) is denoted by λ_i and its estimate is defined as:

$$\hat{\lambda}_i = \frac{o_i}{p_i}$$

where

o_i = Observed number of deaths in the i-th area.

p_i = Population of the i-th area.

p_i may also be known as the number of person-time units at risk. If the period of time is not too long and the population is stable (immigrations compensate emmigrations), p_i may be estimated by the population corresponding to the mid-point of the study period multiply by the time period[47]. Periods of one year are generally used and therefore the estimate of p_i would correspond to the population on 1 July.

2.3 Specific mortality rates.

As pointed out, a rate allows us to quantify the number of cases of death with respect to a set of individuals. This measure will be useful to compare the health of several populations among themselves (in our case geographical areas) and in this way identify populations in which deaths occur faster.

The populations are represented by different characteristics or variables which may influence the crude rate and whose effect we must eliminate in order to compare them appropriately. For example, age is a determining factor in mortality, in other words, at more advanced ages more deaths generally occur (although there may also be specific causes for which mortality is higher at other ages, such as for example traffic injury deaths). Therefore, if we compare the crude rates of two populations in which the age distributions are unequal, for example one with a large majority of elderly and another with few, we could find that the two rates are different. However this difference would be largely due to the effect of age rather than to other possible risk factors

susceptible of being treated and/or controlled by public administrations. So, if we consider the crude rate when comparing mortality between two populations we do not take into account the effect or different distribution of age, and we would obtain erroneous results[49]. These factors whose effects we wish to control are known as confounding factors, since as their name indicates, they "confuse" the results obtained[50].

One possible solution to the situation described is to stratify the population by levels of the confounding factor and to compare the crude incidence rates between the same groups of the factor. Thus if we have J levels of a confounding factor and two populations, A and B, we would compare the crude rate of A for level 1 with the crude rate of B for level 1, the crude rate of A for level 2 with the crude rate of B for level 2, and so on for all J comparisons. The crude rates obtained for the subgroups of a variable are known as specific rates. The definition of specific rates for a particular variable (in our case a confounding factor) is given below.

The specific mortality rate in the i-th area (i=1, ..., n) and the j-th group of a variable (j=1, ..., J) is denoted by λ_{ij} and its estimate is defined as:

$$\hat{\lambda}_{ij} = \frac{o_{ij}}{p_{ij}}$$

where

o_{ij} = Number of cases of death in the j-th group of the variable in the i-th area.

p_{ij} = Population in the j-th group of the variable in the i-th area.

As mentioned earlier, the crude rate is calculated for each of the J groups of the confounding variable. Hence these groups ought to be chosen in such a way that the distribution of mortality over them is as homogeneous as possible, in order for them to be comparable. This can mean that within each group we take very few effectives and in consequence we have a large number of groups. For example, age is usually stratified in groups of 5 years so for an age range of 0 to 90 years we must form 18 groups. The comparison of 18 specific rates between two or more populations in order to ascertain which ones present higher rates of mortality is not a very convenient, practical or parsimonious method. To resolve this we will consider a summary measure for each population or geographical area which controls or eliminates the effect of confounding

variables (such as age) which can lead us to draw erroneous conclusions in the comparisons between populations[51,52,53]. These summary measures will receive the name of standardised mortality indicators. Figures 2.1 and 2.2 present an example where we compute a mortality indicator in each one of 2,218 small areas of Spain. For figure 2.1 a crude mortality indicator has been used, while for figure 2.2 a standardized mortality indicator was used that controls for the confounding effect of age. It may be observed how different results may be obtained if we don't control for the confounding effect of age. The comparison of the small areas with respect to the standardized mortality indicator show that the high mortality is in the South West of Spain for all causes of death in men.

Figure 2.1 *Geographical distribution of relative risk (not standardized) in men in small areas. Spain, 1990-1998.*

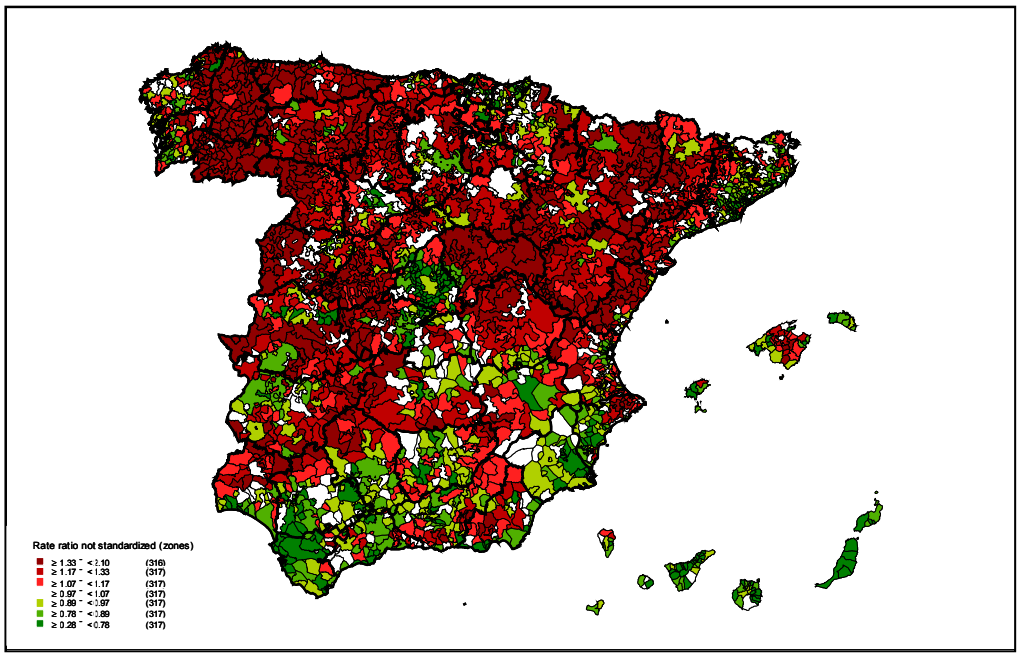
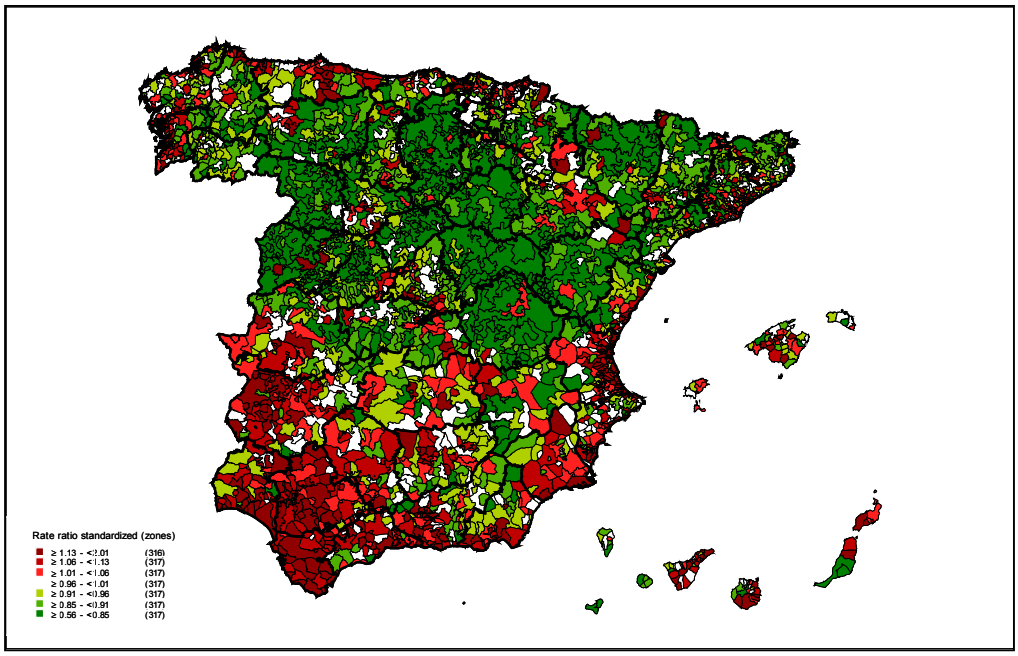


Figure 2.2 *Geographical distribution of standardized relative risk in men in small areas. Spain, 1990-1998.*



2.4 Standardised mortality indicators.

Standardised mortality indicators receive this name because their calculation is usually done based on a reference or standard population[51]. The reference population may be "external" such as the world population or that of Europe, or it may alternatively be obtained "internally" from our data. Generally the ideal standard population is that which is most similar to the populations being compared[49]. The use of an external population may serve to permit international comparisons between studies in different countries which have used the same standard population. In studies of small areas in which comparisons are made between geographical areas of a larger region the standard population is usually obtained internally.

To simplify we will consider the adjustment for a single confounding factor consisting of J categories. It should be noted that obtaining these standardised measures always implies a certain loss of information. We describe separately below the calculation of adjusted rates and the ratio of adjusted rates.

2.4.1 Adjusted mortality rates.

Adjusted rates can in general be expressed and estimated as follows[51,52]:

$$\sum_{j=1}^{J} w_j \hat{\lambda}_{ij}$$

where

w_j = Weight assigned to the j-th group of the confounding factor depending on the standardisation method.

It may be observed that adjusted incidence rates are the specific incidence rates weighted by w_j. Depending on this weight different methods of standardisation arise, of which the direct and indirect methods are the most utilised. Table 2.1, extracted and adapted from the article by Inskip et al.[51] shows the weight used and the formula for standardised rate for each of these two methods. In these formulas the specific mortality rate of the reference population in the j-th group of the confounding factor appears denoted by φ_j and the definition of its estimate is:

$$\hat{\varphi}_j = \frac{D_j}{N_j}$$

where

D_j = Number of cases of death in the j-th group of the confounding factor in the reference population.

N_j = Population in the j-th group of the confounding factor in the reference population.

On the other hand the crude incidence rate of the reference population is denoted by φ and its estimate will be equal to:

$$\hat{\varphi} = \frac{D}{N}$$

where $D = \sum_{j=1}^{J} D_j$ and $N = \sum_{j=1}^{J} N_j$

Table 2.1 *Crude rate, rate standardised by the direct method, and rate standardised by the indirect method.*

Incidence rate	**Weight** w_j	**Formula** $\sum_{j=1}^{J} w_j \hat{\lambda}_{ij}$
Crude	$\frac{p_{ij}}{p_i}$	$\sum_{j=1}^{J} \frac{p_{ij}}{p_i} \frac{o_{ij}}{p_{ij}} = \frac{\sum_{j=1}^{J} o_{ij}}{p_i} = \frac{o_i}{p_i}$
Adjusted (indirect method)	$\frac{\hat{\varphi} p_{ij}}{\sum_{j=1}^{J} \hat{\varphi}_j p_{ij}}$	$\sum_{j=1}^{J} \frac{\hat{\varphi} p_{ij}}{\sum_{j=1}^{J} \hat{\varphi}_j p_{ij}} \frac{o_{ij}}{p_{ij}} = \frac{\hat{\varphi} o_i}{\sum_{j=1}^{J} \hat{\varphi}_j p_{ij}}$
Adjusted (direct method)	$\frac{N_j}{N}$	$\sum_{j=1}^{J} \frac{N_j}{N} \frac{o_{ij}}{p_{ij}} = \frac{1}{N} \sum_{j=1}^{J} N_j \frac{o_{ij}}{p_{ij}}$

2.4.2 Standardised mortality rate ratios.

Adjusted rates are used in order to be compared between different populations. Therefore, once the rate for a population has been standardised the next step and objective is the comparison with the standardised rate of another population. In order to compare the standardised rates of two populations, their ratio can be used. Thus we have the Standardised Mortality Ratio (SMR), obtained from the ratio of rates standardised by the indirect method, and the Comparative Mortality Figure (CMF), obtained from the ratio of two rates standardised by the direct method[52].

The ratios of rates can also be expressed based on the specific rates[51]. Furthermore it should be noted that the ratio of two CMF is also a CMF, whereas the ratio of two SMRs is not itself an SMR. For more details consult the work of Inskip H et al., and of Breslow and Day[51,52,53].

2.4.2.1 Standardised mortality ratio.

The SMR is defined as the ratio of standardised rates obtained by the indirect method which compare a particular population with a reference population[52,53]. We see therefore, consulting table 2.1 that the standardised rate estimated by the indirect method for our population (i-th area) taking as reference the population defined by the specific rates φ_j is defined as:

$$\frac{\hat{\varphi} o_i}{\sum_{j=1}^{J} \hat{\varphi}_j p_{ij}}$$

On the other hand, we obtain the estimate of the standardised rate for the reference population with respect to the population defined by the specific rates φ_j, i.e. with respect to itself. We find that this is equal to its crude rate:

$$\frac{\hat{\varphi} D}{\sum_{j=1}^{J} \hat{\varphi}_j N_j} = \frac{\hat{\varphi} D}{\sum_{j=1}^{J} \frac{D_j}{N_j} N_j} = \frac{\hat{\varphi} D}{D} = \hat{\varphi}$$

Thus the SMR for the i-th area is defined as the ratio between the two standardised rates:

$$SMR_i = \frac{\hat{\varphi} o_i \Big/ \sum_{j=1}^{J} \hat{\varphi}_j p_{ij}}{\hat{\varphi}}$$

from which it may be deduced that

$$SMR_i = \frac{o_i}{\sum_{j=1}^{J} \hat{\varphi}_j p_{ij}}$$

2.4.2.2 Comparative mortality figure.

Analogously, the CMF is a ratio of two standardised rates obtained with the direct method which compares a particular population with a reference population. From table 2.1 it may be demonstrated that the standardised rate for the reference

population is equal to its crude rate. Therefore, the CMF of the i-th area is also calculated dividing the standardised rate for the i-th, estimated with the direct method, by the crude rate of the reference population. Thus:

$$CMF_i = \frac{\frac{1}{N}\sum_{j=1}^{J} N_j \frac{o_{ij}}{p_{ij}}}{\frac{D}{N}}$$

from which it may be deduced that:

$$CMF_i = \frac{\sum_{j=1}^{J} N_j \frac{o_{ij}}{p_{ij}}}{D}$$

2.4.2.3 Alternative expressions for the standardised mortality ratio and the comparative mortality figure.

The formulas for SMR and CMF may be expressed as ratios between observed and expected values. From table 2.1 we may obtain:

$$SMR_i = \frac{o_i}{e_i}$$

where e_i represents the expected cases of death in the i-th area if our population were to follow the pattern of mortality of the standard or reference population. In other words,

$$e_i = e_{i1} + \ldots + e_{iJ} = \hat{\varphi}_1 p_{i1} + \ldots + \hat{\varphi}_J p_{iJ} = \sum_{j=1}^{J} \hat{\varphi}_j p_{ij}$$

Similarly:

$$CMF_i = \frac{e_i^*}{D}$$

where e_i^* are the deaths expected in the standard population if this were to follow the mortality pattern of the population of the i-th area. That is,

$$e_i^* = \sum_{j=1}^{J} N_j \frac{o_{ij}}{p_{ij}}$$

2.4.3 Choice of the standardisation method: comparative mortality figure or standardised mortality ratio?

2.4.3.1 Advantages of the direct method with respect to the indirect method: Property of consistency in the standardised indicators.

Our objective is centred on comparing populations using standardised rates. For this we could consider different measures such as the differences or the ratios described above. Usually the ratios are used since in the comparison of specific rates of two populations they are usually more constant or stable, not presenting the variability of other measures[52].

It is not practical to compare the various populations between themselves by using all the combinations of ratios. As an alternative we may take some population R from the set of populations as the reference and calculate the ratio of each of the others with respect to R. In this way, if a population A presents an adjusted rate 1.5 times higher than the population R and another population B presents an adjusted rate 3 times greater than that of R, we can affirm that population B presents greater mortality than population A ($3 > 1.5$). In order to be able to compare ratios of two populations in this manner we must require that such ratios reflect the true relationship between the ratios of the specific rates of the two populations. This fact gives rise to the concept of conservation of consistency in the standardised indicators as we will see below[51,52,53,54]. The reference population R with respect to which the ratio is calculated may be the same population with respect to which we standardised, and had designated also as the "reference" or standard population. If we use the standard population as the reference population R we will find that the ratios of the standardised rates of each population with respect to R will be the SMR and CMF defined above. Thus in order to be able to compare the populations under study we may use the indicators SMR and CMF defined above.

2.4.3.1.1 Consistency in the standardised indicators.

In order that a method of standardisation allow us to compare a set of populations with respect to their incidence rates it must conserve the consistency between the different populations being compared[51,53]. Inskip H et al., refer to the property of consistency as follows: if in each one of the J groups of the confounding variable we compare the specific rates of two populations, in other words we carry out J comparisons where in each one we compare the j-th group of one population with the same group of the other population, and we find that in one of these populations all the specific rates are higher than those of the other population, then the standardised rate of the first population should be higher than the standardised rate of the second one independently of the reference population utilised[49].

Note that in this property we start from the fact that the J comparisons of the specific rates must present the same magnitude in the sense that they must all be larger or equal (or all smaller or equal) in one of the populations in which we compare them. In situations where the non-compliance with this premise is very pronounced, stratification into age groups that meet the premise is necessary, and then we calculate the adjusted rates in each subgroup. Given that the direct method complies with the property of consistency it is generally preferred to the indirect method which may not comply.

Breslow and Day talk about the property of consistency from the point of view of comparisons of SMRs and CMFs between populations[52,54]. In the case in which we want to compare two populations, consistency demands that the ratios of two SMRs or two CMFs ought to reflect the set of J ratios of their corresponding specific rates. From this it may be deduced that if the ratio between specific rates is higher in the J comparisons for some population A with respect to another population B then it must be the case that the SMR or CMF be higher for population A. This property once again is complied with by the direct method of standardisation[52].

For example, in the extreme idealised situation in which the ratios of specific rates are all equal to a constant κ then the ratio of CMFs will also be equal to κ and therefore represent correctly the ratio between specific rates[52]. In contrast, the indirect method need not necessarily comply with the premise in the above extreme situation. Therefore the ratios of two SMRs may not reflect the set of ratios of the corresponding

specific rates of the two populations. In this way the comparison by means of the ratio of two SMRs could lead us to erroneous conclusions.

However, if in addition to considering that the ratios of specific rates of the two populations all be equal to κ, we require that the ratios between the specific rates of the two populations we want to compare and the specific rates of the "reference" population also be constant, then the ratios of the SMRs of the two populations will be equal to κ[52,54]. Therefore we can compare SMRs in two or more populations if the variation between the ratios of specific rates is not too high and we additionally require that the specific rates for all the populations we want to compare (n geographical areas in our case) be constant with respect to the specific rates of the reference population. Breslow and Day propose a simple model for the specific rates which meet the property of consistency defined as[52,54]:

$$\lambda_{ij} = \theta_i \varphi_j \qquad (2.1)$$

where θ_i is the ratio of the j-th specific rate of the i-th area with respect to (relative to) the j-th specific rate of the reference population. We will also call the ratio of rates denoted by θ_i the relative risk.

This assumption can be validated using a graph of the specific rates of each area with respect to the specific rates of the reference population, and we find that we do indeed obtain a line with a slope of unity[55]. Validation may also be carried out comparing the statistical model expressed by (2.1) with model (2.2) where proportionality will not be complied with if the effect γ_{ij} corresponding to an interaction between the i-th area and the confounding factor is significant[55]. We will see later that observed death data may be modelled based on a Poisson distribution, and hence allows us to define a statistical model for the rates.

$$\lambda_{ij} = \theta_i \varphi_j \gamma_{ij} \qquad (2.2)$$

However, in the context of studies of small areas where in each geographical area we will have a reduced number of observations this assumption may be impossible

to validate due to having insufficient data[56]. It should be noted that other authors have not found large differences between the results obtained with the direct and indirect standardization methods[57].

2.4.3.2 Advantages of the indirect method over the direct method: a question of variability in the standardised indicators.

We saw earlier that the main advantage of the direct method was the conservation of consistency. In what follows we deal with certain circumstances where the indirect method is the more appropriate.

As several authors have stated, when adjusted rates are calculated in populations with a reduced number of deaths the indirect method is more appropriate since the standardised indicators are more stable, i.e. their standard error is smaller[51,52,53]. Another way of seeing that the SMR is more appropriate for situations with small numbers of deaths, is to estimate what the risk of death would be, if we assume a distribution for the observed deaths suitable for when we have few cases of death in relation to the population. The distribution usually used is the Poisson distribution, appropriate for independent populations with infrequent events. Under the Poisson assumption the maximum likelihood estimator of the relative risk of death is the SMR as Breslow and Day briefly demonstrate[51,52,53,54]. In other words, if O_{ij} and p_{ij} are respectively the observed deaths and the population of the j-th group of the confounding factor in the i-th area and we consider that $O_{ij}|\lambda_{ij}$ are independent random variables following a Poisson distribution with expected value $\lambda_{ij}p_{ij}$ where λ_{ij} is defined according to model (2.1) we find that the maximum likelihood estimator of θ_i, denoted by $\hat{\theta}_i^{MV}$ is equal to:

$$\hat{\theta}_i^{MV} = SMR_i = \frac{O_i}{E_i}$$

where E_i indicates a new notation for the expected values e_i defined above for the SMR.

This property suggests that the SMR could be used to compare populations or geographical areas in which we have small numbers of deaths in relation to the

population at risk. Analogously to the previous case it can be demonstrated that for a rare non-infectious disease (or for mortality data) we may consider that $O_i|\theta_i,E_i$ are independent and follow a Poisson distribution with expected value $\theta_i E_i$. Thus we have that the maximum likelihood estimator of the relative risk θ_i is once again the SMR. In order to simplify the notation for the remainder of this text the conditioning by E_i will be suppressed. For example, we will write $O_i|\theta_i$ instead of $O_i|\theta_i,E_i$.

It should be noted that in the first case the effect of the confounding factor is taken into account by including the reference rates stratified by levels of the confounding factor through the term φ_j in model (2.1). This way we obtained an adjusted relative risk which as shown corresponded to the standardised mortality ratio for this confounding factor. In the second case the effect of the confounding factor is represented by the expected numbers of cases, denoted by E_i.

For these reasons SMR is a priori the best choice for the mortality risk in studies of small areas in which generally there are few deaths in each area with respect to the population at risk. In cases dealing with more common diseases we can consider the binomial distribution as the more appropriate for modelling the observed numbers of deaths[58]. In reality the use of the Poisson distribution is due to its being an approximation for the binomial distribution when we are dealing with rare diseases. In the remainder of this document we will use the Poisson distribution for the observed numbers of deaths.

The SMR also presents other advantages such as the possibility of considering a Poisson distribution on the observed cases of death thus allowing them to be modelled obtaining parameters interpretable from the epidemiological point of view which represent ratios of rates or relative risks and may easily be extended to add random effects as we will see in the next chapter. In contrast the CMF could be modelled carrying out some type of transformation which would approximate it to the Normal distribution. However, its interpretation would not be easy and it would not represent any measure of health of interest epidemiologically speaking.

Another inconvenience of the direct method is that it requires knowledge about the observed cases in the study population and of the population at risk for each of the J strata of the confounding factor. The indirect method on the other hand only requires knowledge of the population at risk in each of these J strata and the total observed cases of death[47,48,49,50].

2.4.4 Obtaining reference rates internally.

There are several methods for estimating the standard population internally. The simplest method consists of directly aggregating the data according to the groups of the confounding factor. That is, for j=1, ..., J,

$$\hat{\varphi}_j = \frac{\sum_{i=1}^{n} o_{ij}}{\sum_{i=1}^{n} p_{ij}}$$

However, as some authors point out this method may be incorrect since it may eliminate some effects of the geographical area[55]. In order to take this fact into account one may consider an estimation of the specific rates internally based on a model which involves the structure of the geographical areas by carrying out a joint estimation of the J specific rates and the relative risk in the n geographical areas[52,54,55,59].

Clayton and Kaldor[59] also perform a joint estimation of the relative risk in each one of the areas and reference rates on the basis of the algorithm proposed by Mantel and Stark[60]. Other authors carry out the internal estimation of the reference rates taking into account the structure of the geographical areas based on Generalized Estimating Equations (GEE) approach[6].

2.4.5 Interpretation of the standardised mortality ratio.

In our case we will consider that the rates have been estimated internally based on our own population. The SMR have a simple interpretation in terms of the observed and expected cases of death. An SMR value greater than (less than) 1 indicates more (less) deaths than those expected if mortality in the area were the same as that occurring

in our entire population or study region (divided into n geographical areas) taking into account the effect of the confounding factor. As we have seen earlier it can also be interpreted as a ratio of rates relative to the complete set of reference rates, in this case if the SMR value is higher (lower) than 1 it would indicate that the mortality rate of the i-th area is higher (lower) than the rate for the study region (divided into n geographical areas) taking into account the effect of the confounding factor.

CHAPTER 3

Bayesian models in disease mapping

"We note the essential duality between a sample and the density (distribution) from which it is generated. Clearly, the density generates the sample; conversely, given a sample we can approximately recreate the density"

Smith and Gelfand

3.1 Introduction.

In disease mapping our main objective is to compare each one of the geographical areas on the basis of some particular health characteristic. In the previous chapter two summary measures were mentioned which could express such characteristics taking into account the effect of variables which could confound the results. Of these measures the SMR was chosen for its superior properties when we have rare diseases and because it allows modeling based on the Poisson distribution. In the present chapter we will see how the SMR can be unstable in small areas, what the effects of such instability may be and some alternatives that different authors have used to deal with it through the incorporation of additional information.

3.2 Disadvantages of the standardized mortality ratio for small areas.

In order to obtain an estimate of the relative mortality risk we described the case in which $O_i|\theta_i$ is a random variable with known observed values which follow a Poisson distribution with expected value $\theta_i E_i$ where E_i is a known fixed value which allows us to

take into account the effect of confounding factors. In this case θ_i is an unknown fixed quantity which represents the relative risk for the i-th area. Based on this assumption the maximum likelihood estimator of θ_i was the SMR[52]:

$$\hat{\theta}_i^{MV} = SMR_i = \frac{O_i}{E_i}$$

We may observe that:

$$E[\hat{\theta}_i^{MV}] = E[\frac{O_i}{E_i}] = \frac{1}{E_i} E[O_i] = \frac{1}{E_i} \theta_i E_i = \theta_i$$

$$V[\hat{\theta}_i^{MV}] = Var[\frac{O_i}{E_i}] = \frac{1}{E_i^2} Var[O_i] = \frac{1}{E_i^2} \theta_i E_i = \frac{\theta_i}{E_i}$$

From this it may be deduced that the main disadvantage of using the SMR as an estimator of the relative risk is its statistical instability in small areas[59,61,62]. In spite of being an unbiased estimator of the relative risk its variance or variability in small areas tends to be large since such areas may represent a reduced population and in consequence the expected value E_i is small. In other words, the estimator of the variability of the SMR is equal to:

$$\hat{V}[\hat{\theta}_i^{MV}] = \frac{O_i}{E_i^2}$$

Thus when E_i is small, the variability in SMR will be large.

In order to obtain a correct estimation of the relative risk in each area we must control the statistical instability of the SMRs since otherwise we may be led to draw erroneous conclusions.

Often the most extreme SMR (high and low) will occur in the areas with least population. Therefore, if we do not control for the statistical instability of the SMRs we could consider as having the greatest or the least risk geographical areas which in reality do not, the extreme values merely resulting from the reduced population. In the face of uncertainty about the exactness of the true reason for such an extreme SMR, i.e. whether it really corresponds to a good approximation of the true risk in that area or whether it is due to the small population, we assume the latter.

Another alternative would be to use maps in which the statistical significance for each geographical areas is considered. However, the highly significant areas will be those having the greatest population even when the value of the relative risk is not notable (either high or low)[61,62].

Another disadvantage of the SMR is that it does not incorporate spatial configuration in its estimation[62,63]. Furthermore, although the SMR was the best estimation of relative risk under the Poisson assumption, it can be shown that alternative methods, known as Bayesian, to be described below, offer estimates of relative risk which have smaller mean squared error when the number of areas is higher than 3[62].

The approach described in this section which obtained the SMR as a maximum likelihood estimate for the relative mortality risk under the assumption that $O_i|\theta_i$ follows a Poisson distribution with mean $\theta_i E_i$ will be called the classical approach.

3.3 Alternative approaches to the standardized mortality ratio.

Several authors have proposed alternative approaches for estimating the relative mortality risk in such a way as to minimize the problem of statistical instability of the SMRs in low population areas. Broadly speaking we may distinguish four alternative approaches[63]:

- Smoothing models based on non-parametric methods.
- Linear Bayes methods, based on linear functions of the SMRs.
- Fully Bayes methods.
- Empirical Bayes methods.

The remainder of this document will deal only with the empirical Bayes (EB) and fully Bayes (FB) approaches since these present certain advantages over the other two approaches, for example the use of the likelihood function and the possibility of

estimating measures and even a complete distribution, known as the posterior distribution, which allows us to obtain a great deal of information about the behavior of the relative mortality risks. From the posterior distribution we can obtain various measures to estimate the relative risk such as the mean, mode, and median of the posterior distribution and intervals associated with their variations, known as credibility intervals in the FB approach. Perhaps due to these and other reasons, these methods are the ones usually used in the context of estimation of relative mortality risk in small areas.

3.4. How do Bayesian methods control instability in the standardized mortality ratio?

3.4.1. Utility of the Bayesian methods in estimating the relative risk in small areas. Definition of the prior distribution.

We saw earlier that the information which the classical approach provides does not allow us to control statistical instability in the SMRs. In order to control this fact, we must incorporate additional information such that when we want to estimate the relative risk of an area, we control for whether the area has a large or a small population.

To achieve this aim, the Bayesian methods establish a weighting or a compromise between the information about the area for which we want to estimate the relative risk and the information provided by the other areas[61]. If a particular area of interest has a reduced population, then the estimation of the relative risk must borrow strengths from the information provided by all the areas assigning less weight to the unstable information provided by this area. On the other hand, if the area of interest has a large population there is no need for the estimate to borrow strengths and a greater weight is assigned to the stable information provided by this area[61].

This means that the estimation for relative risk of each area is its SMR smoothed in such a way that when an area has a small population its relative risk will tend towards

the mean value of all the relative risks, which is usually 1 (however if the reference population is obtained externally they do not necessarily have to be centered around one[61]). In contrast, if the area has a large population the estimate of the relative risk will be close to the SMR for that area.

Such methods can be extended when there is evidence of a spatial pattern in the mortality[61]. In this case the information provided by the remainder of areas will be of a local nature instead of global and the relative risk of a particular area of interest will not tend towards the global mean but rather towards a value similar to the relative risk of adjacent areas.

The Bayesian approaches collect information about the remainder of areas or about adjacent areas into a probability distribution considered on the relative risks. This is known as the prior distribution and incorporates the variation of the relative risks of the geographical areas of the region under study[61,62].

We will see below, from the point of view of the prior distribution, the concept presented earlier relating to the control of instability of the SMRs by Bayesian methods. For example, consider the EB gamma-poisson method which will be dealt with shortly. This method considers the relative risks as independent and identically distributed, following Gamma prior distribution with scale parameter α and form parameter ν:

$$[\theta_i|\alpha,\nu] \sim \text{Gamma}(\alpha,\nu)$$

Under this assumption we may obtain, as the estimator of the relative risk for the i-th area, the following expression[7]:

$$\hat{\theta}_i^G = w_i E[\theta_i \mid \alpha,\beta] + (1 - w_i)SMR_i \qquad (3.1)$$

where

- $E[\theta_i|\alpha,\beta]$ is the mean of the prior distribution on the relative risks, which in the case of a Gamma is defined as the ratio between its parameters of scale and of form (α/ν).

- SMR_i is the value of the standardized mortality ratio of the i-th area.
- w_i is a weight ($0 \leq w_i \leq 1$) with the following value:

$$w_i = \frac{\nu}{(\nu + E_i)} \qquad (3.2)$$

From (3.1) and (3.2) it can be deduced that when the area has a small population (small E_i) w_i will tend towards 1 and more weight will be given to the information from the rest of the areas reflected in the mean of the prior distribution $E[\theta_i|\alpha,\beta]$. On the other hand when the area has a large population (E_i is large) w_i will tend towards 0 and more weight will be given to the information of the area for which we want to estimate the relative risk and which is represented by SMR_i, i.e. by the observed and expected cases of death in this area in the form of the ratio, O_i/E_i.

From expressions (3.1) and (3.2) we can also see the process of smoothing of the SMRs mentioned above occurs. The smaller the population of the area in which we estimate the relative risk, the more this estimate will be smoothed tending more towards the global mean of the relative risks[55,62].

Table 3.1 shows an example taken from Mollié[64] which shows the relative risk estimated using the classical method and using a fully Bayesian approach in certain selected areas (departments) of France. The notation used for the i-th area is: Y_i for observed deaths, E_i for the expected deaths, SMR_i for the classical estimate of relative risk and θ_i* for the relative risk estimated using the Bayesian approach. It may be observed how the Bayesian method smoothes the relative risks and how the degree of smoothing is more pronounced in those areas with lower expected values.

Table 3.1 *Hodgkin's disease in men in France by selected departments 1986-1993.*

i	Département	Y_i	E_i	SMR	95% CI_{SMR}	θ_i^*	95% CI_{Bayes}
80	Somme	30	15.92	1.88	(1.27–2.70)	1.40	(1.10–1.80)
59	Nord	108	66.42	1.63	(1.33–1.98)	1.49	(1.25–1.78)
72	Sarthe	24	15.82	1.52	(0.97–2.25)	1.08	(0.87–1.38)
55	Meuse	9	5.97	1.51	(0.69–2.86)	1.08	(0.83–1.42)
76	Seine-Maritime	51	34.20	1.49	(1.11–1.97)	1.30	(1.04–1.62)
62	Pas-de-Calais	50	38.72	1.29	(0.96–1.70)	1.32	(1.05–1.63)
02	Aisne	17	15.60	1.09	(0.64–1.74)	1.16	(0.92–1.47)
90	Territoire-de-Belfort	4	3.85	1.04	(0.28–2.66)	0.99	(0.74–1.33)
60	Oise	20	19.67	1.02	(0.62–1.57)	1.10	(0.87–1.36)
42	Loire	21	22.47	0.93	(0.58–1.43)	0.89	(0.71–1.11)
44	Loire-Atlantique	25	29.25	0.85	(0.55–1.27)	0.89	(0.70–1.12)
57	Moselle	24	28.21	0.85	(0.55–1.26)	0.89	(0.68–1.17)
29	Finistère	15	25.93	0.58	(0.32–0.95)	0.74	(0.54–0.99)
69	Rhône	21	42.33	0.50	(0.31–0.76)	0.75	(0.58–0.94)
87	Haute-Vienne	6	12.28	0.49	(0.18–1.07)	0.89	(0.69–1.12)
48	Lozère	1	2.64	0.38	(0.01–2.11)	0.86	(0.64–1.13)
05	Hautes-Alpes	1	3.73	0.27	(0.01–1.49)	0.86	(0.63–1.14)

Analogously, figures 3.1 and 3.2 show the SMR and the relative risk obtained using an empirical Bayes method for lung cancer mortality in men and breast cancer for women for municipalities (or aggregates thereof) in Spain over the period 1987-1995[6]. In this case the information provided by the rest of the areas is of a global nature and the relative risks will tend towards their mean value (approximately 1). Of the 2218 areas analyzed these graphs show the 100 areas with lower expected value E_i (2a and 3a) and the 100 areas with a high expected value (2b and 3b). The SMR value for each area and the corresponding smoothed relative risk, denoted as $\hat{\theta}_i$, are joined by a line. If the line is a constant the effect of smoothing has been minimal, otherwise the slope of the line indicates whether the smoothing was more or less pronounced. It may be appreciated that smoothing of the SMR is greater for those areas with lower expected values.

Figure 3.1 *SMR in each area paired with the relative risk obtained through an empirical Bayesian method* ($\hat{\theta}_i$), *for the 100 areas with the lowest expected values (5.19 to 9.60)(2a) and the 100 areas with the highest expected values (84.40 to 6997.25)(2b). Lung cancer deaths for men by municipalities (or aggregates thereof) in Spain (1987-1995).*

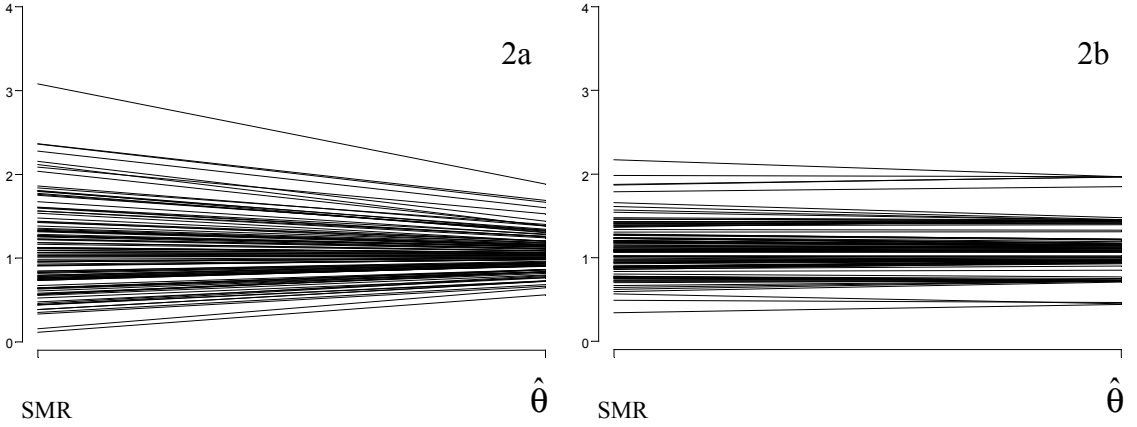
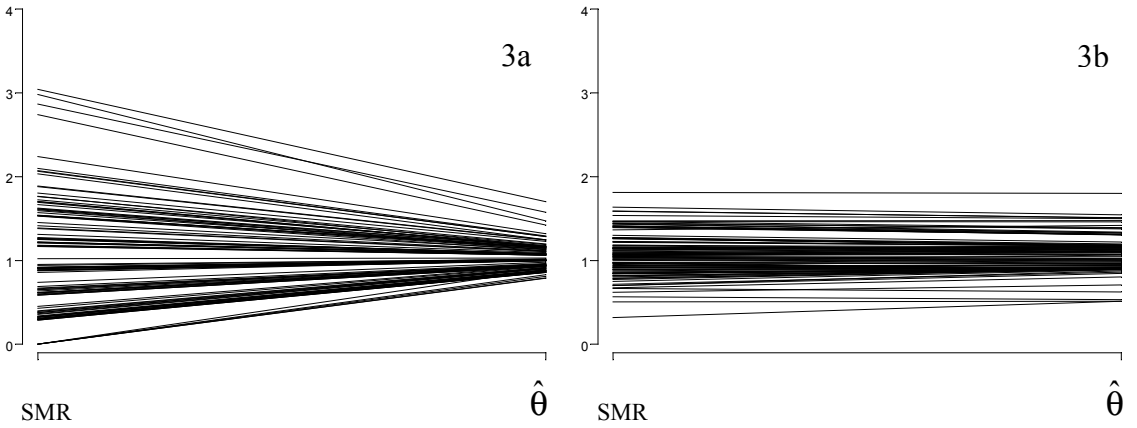


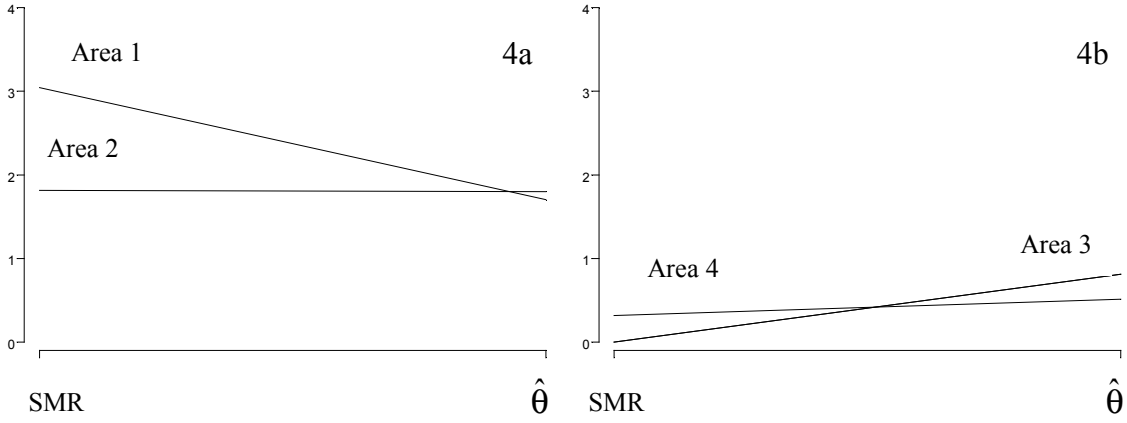
Figure 3.2 *SMR in each area paired with the relative risk obtained through an empirical Bayesian method* ($\hat{\theta}_i$), *for the 100 areas with the lowest expected value (1.71 to 3.45)(3a) and the 100 areas with the highest expected values (34.13 to 3291.34)(3b). Breast cancer deaths for women by municipalities (or aggregates thereof) in Spain (1987-1995).*



We will now compare the areas with extreme SMR of figure 3.2, i.e. the 2 areas in graphs 3a and 3b with the highest SMRs and the 2 areas in graphs 3a and 3b having the lowest SMRs. Figure 3.3 shows the 2 areas with the highest SMRs along with the

relative risk obtained by the Bayesian method (4a) and the 2 areas with the lowest SMRs along with the relative risk obtained by the Bayesian method (4b). As described earlier, if we consider the SMR as an estimate of the relative risk we could erroneously conclude that area 1 presents greater risk than area 2, however the Bayesian estimate reveals that area 2 presents greater or approximately equal risk as areas 1. Analogously, if we use the SMR we could erroneously conclude that area 3 presents lower risk than area 4, whereas the Bayesian estimate indicates that area 4 presents lower risk than area 3. It should once again be stressed that in the face of uncertainty of knowing exactly the real reason for this SMR we assume it to be due to the reduced population size.

Figure 3.3 *SMR paired with the relative risk obtained by an empirical Bayesian method* ($\hat{\theta}_i$) *for areas with the lowest expected values (areas 1 and 3) (expected values 1.71 to 3.45) and for areas with the highest expected values (areas 2 and 4)(expected values 34.13 and 3291.34). Breast cancer deaths in women for municipalities (or aggregates thereof) in Spain (1987-1995).*



3.4.2 Posterior distribution: compromise between the data information and prior distribution.

In the above section we have described conceptually the control of statistical instability in the SMRs for areas with low population using the Bayesian method. From the point of view of a statistical model the Bayesian approach controls separately two types of variations[61,65]:

1) Variation due to the observed deaths which is controlled by considering a Poisson distribution over these deaths conditioned by the unknown values of the relative risks.

2) Variation due to the relative risks not covered by the Poisson distribution, also known as extra-Poisson variability or over-dispersion, which is controlled by considering the prior distribution over these relative risks.

In the Poisson model over-dispersion occurs when the variability in the deaths observed in the i-th area is higher than that expected[58], i.e.,

$$Var(O_i) > E[O_i]$$

As we will see later over-dispersion can have spatial or non-spatial effects, and can be originated by, for example, variables not included in the analysis, or inaccuracies in the data[58].

Thus the two approaches EB and FB may be seen as a two-level hierarchical model where each level controls for a different type of variation.

- Level 1 (Poisson variability):

$$[O_i|\theta_i] \sim Poisson(\theta_i E_i)$$

- Level 2 (Extra-Poisson variability):

$$\theta|\gamma \sim \pi(\theta|\gamma)$$

Where $\theta = (\theta_1,..., \theta_n)^T$ and $\pi(\theta|\gamma)$ represents a prior distribution. As described below the FB approach will also consider a third level in which variability of the γ parameters of the prior distribution are controlled for by considering a probability distribution over them. The FB model also receives the name of Bayesian hierarchical model.

Thus, in order to estimate the relative risk in each area we will have two types of information:

1) Information about the area of interest obtained from the likelihood function of the observed deaths obtained through the Poisson assumption.

2) Information from the rest of the areas (or adjacent areas), provided by the prior probability distribution, on the relative risks.

The combination of these two types of information in order to permit obtaining the estimation of the relative risks will be carried out based on Bayes' Theorem.

We will denote the set of observed deaths by $O = (O_1,..., O_n)^T$, expected deaths by $E = (E_1,...,E_n)^T$ and the set of relative risks of the rest of the n geographical areas by $\theta = (\theta_1,..., \theta_n)^T$. Under the Poisson assumption on the observed deaths**,** we can obtain the likelihood function for the sample (when the function is with respect to θ) or joint distribution of $O|\theta$ which we will denote by $L(O \mid \theta)$.

Furthermore, we will consider a prior distribution over θ. Earlier we used $\pi(\theta|\gamma)$ to denote the joint prior distribution over the relative risks where γ are a set of parameters which define it and receive the name of hyperparameters[64,65]. We will subsequently see that the prior distribution may be over each one of the relative risks of each area if we do not consider there is a spatial structure in the relative risks. On the other hand, when a spatial structure is present the prior distribution will be fitted over a vector composed of the relative risks or for each relative risk conditioned to the relative risks of the adjacent areas.

Therefore, applying Bayes' Theorem, from $L(O|\theta)$ and $\pi(\theta|\gamma)$ we may obtain:

$$P(\theta \mid O,\gamma) = \frac{P(O,\theta \mid \gamma)}{P(O \mid \gamma)} = \frac{L(O \mid \theta)\pi(\theta \mid \gamma)}{\int_{\Theta} L(O \mid \theta)\pi(\theta \mid \gamma)\partial\theta} \qquad (3.3)$$

If the distribution on the relative risks were discrete the integral would be replace by a summation.

We may observe that the conjunction of both types of information provides us with a distribution of the relative risks which take into account the observations of the sample since it is conditioned to the vector of observations O. The distribution of θ conditioned to the data O and denoted by P(θ|O) is known as the posterior distribution and is the basis of Bayesian inference[61,62]. This distribution presents excellent properties, for example the compliance with the likehood principle [66] and the obtaining of a set of measures and confidence intervals (credibility intervals in the fully Bayesian context) which will permit us to make inferences about the relative risks θ_i. The distribution expressed in (3.3) does not yet correspond to P(θ|O) since it also depends on the hyperparameters γ. We will see below how the way of dealing with the hyperparameters γ, in order to arrive at the posterior distribution P(θ|O), represents the greatest difference between EB and FB approaches.

3.4.3 Differences between the empirical Bayes and the fully Bayes approach.

3.4.3.1 The empirical Bayes approach.

The EB approach tries to approximate P(θ|O) using P(θ|O,γ)[61,65]. To do so it estimates the hyperparameters γ of the prior distribution over the relative risks based on observed data, and for this reason is called "empirical"[67]. Specifically, the estimation of γ is based on the marginal distribution of the vector of observed data O. Thus, once γ has been estimated and substituted into (3.3) we consider that the two distributions are approximately equal, in other words, and denoting by $\hat{\gamma}$ the estimation of γ:

$$P(\theta \mid O) \approx P(\theta \mid O, \hat{\gamma})$$

where for (3.3) and with K a constant:

$$P(\theta \mid O, \hat{\gamma}) = \frac{1}{K} L(O \mid \theta)\pi(\theta \mid \hat{\gamma})$$

$$\alpha \quad L(O \mid \theta)\pi(\theta \mid \hat{\gamma})$$

In this way once γ has been estimated we can obtain the posterior distribution of relative risk for the i-th area by solving[64]:

$$P(\theta_i \mid O, \hat{\gamma}) = \int_{\theta_1} \ldots \int_{\theta_{i-1}} \int_{\theta_{i+1}} \ldots \int_{\theta_n} P(\theta \mid O, \hat{\gamma}) \partial\theta_1 \ldots \partial\theta_{i-1} \partial\theta_{i+1} \ldots \partial\theta_n \qquad (3.4)$$

We can obtain an estimation of the relative risk with, for example, the mean of the posterior distribution, $E[\theta_i \mid O, \hat{\gamma}]$, obtained throught another integration process.

In the case in which the relative risks are independent following the prior distribution the approximation of the posterior distribution of the relative risk of the i-th area can be simplified to yield:

$$P(\theta_i \mid O_i, \hat{\gamma}) = \frac{P(O_i \mid \theta_i)\pi(\theta_i \mid \hat{\gamma})}{P(O_i \mid \hat{\gamma})}$$

$$\alpha \quad P(O_i \mid \theta_i)\pi(\theta_i \mid \hat{\gamma}) \qquad (3.5)$$

As already mentioned the estimation γ is carried out based on the marginal distribution of $O|\gamma$ defined for γ a vector of fixed parameters as:

$$P(O \mid \gamma) = \int_{\Theta} L(O \mid \theta)\pi(\theta \mid \gamma)\partial\theta \qquad (3.6)$$

This may be seen as a likelihood function to estimate γ, and in consequence receives the name of marginal likelihood.

When the integral given in (3.6) can be solved analytically the marginal distribution, denoted by $P(O|\gamma)$, may be obtained. Thus, based on $P(O|\gamma)$ we will be able to estimate the γ values using the method of maximum likelihood.

On the other hand if the integral of (3.6) is too difficult to solve it will be necessary to use other methods to obtain the estimation of γ through maximization of the marginal likelihood. Methods which approximate integrals over the random effects (in this case parameters of relative risks) may also be used, such as that known as adaptive Gaussian quadrature, and as a specific case the Laplace approximation[68]. This approximation can be maximized using optimization algorithms such as the dual quasi-Newton[68]. The EM algorithm may also be used both to obtain a point estimate as well as for confidence intervals[62,65,69].

Once the estimation of γ has been obtained we must solve the integrals given in (3.4 – 3.6). When the prior distribution is a conjugate prior, the posterior distribution is simple to obtain, for example in the gamma-poisson model. If not we must resort again to the approximate methods described earlier.

The EB approach is also seen as a Generalized linear mixed model (GLMM)[58,70] because it can combine fixed effects (due to covariates) with the random effects (relative risks). For this reason it is also possible to obtain an estimate of the relative risks through methods developed for obtaining estimates of the fixed and random effects in GLMM. Among these methods, attention is drawn to the method of penalized quasi-likelihood (PQL)[70,71] based on linearization of the model by Taylor series expansion and described by Breslow and Clayton[70], among others. We can also consider the procedure of pseudo-likelihood which includes PQL as a special case, or restricted pseudo-likelihood developed by Wolfinger and O'Connell[71]. For a recent revision of methods of inference in GLMM consult the text of Basagaña et al[32].

The EB approach based on the approximation $P(\theta_i \mid O, \hat{\gamma})$ provides good point estimates of the relative risk, however it underestimates the variability of this estimate[61,62]. This happens because there is no control over the variability of γ when estimating the relative risks. In addition to the EM algorithm, methods have been

proposed for calculating the confidence intervals of the estimates based on bootstrap techniques, although their validity is questionable[61,65]. Other alternatives are based on using the delta method and a conditional mean square error[61,68,118].

3.4.3.2 The fully Bayes approach.

The handling of the vector γ of hyperparameters in the FB approach is different to that of EB. The FB approach does not estimate the γ parameters of the data but rather uses a third level in which a distribution is considered on γ in addition to the two levels in which we control Poisson and extra-Poisson variability. The distribution over γ is termed hyperprior distribution. This distribution consists of certain parameters here called hyperparameters which are established by the researcher based on a series of criteria.

If we denote the distribution on the γ parameters by h(γ), we may obtain $P(\theta,\gamma \mid O)$ defined as:

$$P(\theta,\gamma \mid O) = \frac{L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)}{\int_{\Gamma}\int_{\Theta} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\theta\partial\gamma}$$

Integrating over **γ** we obtain the posterior distribution $P(\theta \mid O)$

$$P(\theta \mid O) = \int_{\Gamma} P(\theta,\gamma \mid O)\partial\gamma$$

After successive substitutions we arrive at

$$P(\theta \mid O) = \frac{\int_{\Gamma} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\gamma}{\int_{\Gamma}\int_{\Theta} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\theta\partial\gamma}$$

where

$$P(O) = \int_{\Gamma}\int_{\Theta} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\theta\partial\gamma$$

Therefore

$$P(\theta \mid O)\alpha \int_{\Gamma} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\gamma$$

The posterior distribution of relative risk in the i-th area will be equal to:

$$P(\theta_i \mid O) \ \alpha \ \int_{\Gamma}\int_{\theta_1}\ldots\int_{\theta_{i-1}}\int_{\theta_{i+1}}\ldots\int_{\theta_n} L(O \mid \theta)\pi(\theta \mid \gamma)h(\gamma)\partial\theta_1\ldots\partial\theta_{i-1}\partial\theta_{i+1}\ldots\partial\theta_n\partial\gamma$$

Calculation of the integrals is unmanageable and for this reason the FB approach resorts to Montecarlo methods based on Markov chains (MCMC)[72] which will use a series of algorithms to simulate samples of the joint posterior distribution $P(\theta,\gamma \mid O)$ and in particular of $P(\theta_i \mid O)$[62,65]. As Wakefield et al.[58] point out in relation to an observation made by Smith and Gelfand[73]:" *We note the essential duality between a sample and the density (distribution) from which it is generated. Clearly, the density generates the sample; conversely, given a sample we can approximately recreate the density*". Therefore, based on samples generated from the posterior distribution we can reconstruct it. Among the algorithms used to generate samples of the posterior distribution are the Metropolis algorithm, their generalization known as Metropolis-Hastings and a modification of the latter known as Gibbs Sampling[7,72].

Just as in the EB approach the FB method provides correct point estimates of summary measures of the posterior distribution (for example, mean and median) as estimators of relative risk, but unlike the EB approach it also provides appropriate estimates for confidence intervals of the measures, which in the FB context are known as credibility intervals[62,65].

CHAPTER 4

Prior distributions on relative risks

"He who adds nothing to his knowledge, reduces it"

Talmud

4.1 Introduction.

In the previous chapter we described how the prior distribution can reflect information about the geographical variation of the relative risks in the areas into which the study region is divided. Geographical variation in relative risks may be of two types. The first type assumes that the relative risks do not present any kind of spatial structure and is known as heterogeneity. The second considers that the relative risks may present a spatial structure in the sense that a particular area may present similar risks to other geographical areas in the study region (usually neighboring) and is referred to as clustering. If it is not known with certainty which of these two types of variation is present in the relative risks, both may be taken into consideration jointly (heterogeneity plus clustering). Each prior distribution, depending on the geographical variation, will determine the level of smoothing. In the case of heterogeneity this will be of a global nature, i.e. based on all the areas, while in the case of clustering it will be local, in other words based on adjacent areas[58]. Below we define the prior distributions for each of the mentioned types of geographical variation, and for the combination of the two.

4.2 Prior distributions for representing heterogeneity variation.

For this type of variation we assume that the relative risks of each area $\theta_i|\gamma$ (i=1,...,n) are independent and identically distributed following the prior distribution. The prior distributions usually used in this case are the Gamma distribution, a discrete distribution consisting of K probability distributions evaluated in K observed values of relative risk, and the Normal or Gaussian distribution for the logarithm of relative risk. We will give the expression for the posterior mean in the EB approach for the Gamma distribution and the discrete distribution. It should be noted that the prior distributions for the case of heterogeneity are also valid for the FB approach provided an appropriate hyperprior is taken for the values of the vector γ[7].

4.2.1 Gamma prior distribution.

Consider that the relative risks, $\theta_i|\gamma$ (i=1,...,n), with $\gamma = (\alpha,\nu)^T$ are independent and identically distributed following a Gamma prior distribution with scale parameter α and form parameter ν[59]:

$$[\theta_i|\alpha,\nu] \sim \text{Gamma}(\alpha,\nu) \qquad i=1,...,n$$

As $O_i|\theta_i$ (i=1,...,n) are independent and follow a Poisson distribution with mean $\theta_i E_i$ and $\theta_i|\alpha,\nu$ (i=1,...,n) are independent and identically distributed following a Gamma(α,ν) distribution, this model also receives the name of Gamma-Poisson model. As the Gamma distribution is a conjugate distribution with the Poisson we obtain that the posterior distribution is also Gamma, with scale parameter $\alpha+O_i$ and form parameter $\nu+E_i$.

The parameters α and ν will be estimated by maximum likelihood based on the joint marginal distribution denoted $P(O|\alpha,\nu)$. In this case, $P(O_i|\alpha,\nu)$ is a Negative Binomial distribution with mean and variance[59].

$$E[O_i \mid \nu,\alpha] = E_i + \frac{\nu}{\alpha}$$

$$Var[O_i \mid \nu,\alpha] = E_i \frac{\nu}{\alpha} + E_i^2 \frac{\nu}{\alpha^2}$$

Therefore, from the maximum likelihood estimates of α and ν denoted by $\hat{\alpha}$ and $\hat{\nu}$ we may obtain an estimate for the relative risk θ_i of each geographical area based on the mean of the posterior distribution. In this case the posterior distribution is a Gamma distribution and hence the estimate of the mean is defined as the ration between its scale and form parameters, i.e. the relative risk of the i-th area is equal to:

$$\hat{\theta}_i^G = E[\theta_i \mid O_i, \hat{\alpha}, \hat{\nu}] = \frac{O_i + \hat{\alpha}}{E_i + \hat{\nu}}$$

Clayton and Kaldor also describe the extension of this model when we want to include covariates in order to obtain a more precise prediction of the relative risk. In this case the relative risk of the i-th area will not tend towards the global mean of all relative risks, but instead towards an estimate of the relative risk representative of the covariates of its zone[59].

4.2.2 Normal prior distribution.

Consider that the vector consisting of the logarithm of the relative risks defined by $V = (\log \theta_1 ,..., \log \theta_n)^T = (V_1,...,V_n)^T$ follows a multivariate Normal distribution with an nx1 vector of means denoted by μ and an nxn matrix of variances-covariances denoted by Σ respectively equal to:

$$E[V] = \mu = (0,...,0)T \qquad (4.1)$$

$$Cov[V] = \Sigma = \sigma_v^2 I \qquad (4.2)$$

where I is an nxn identity matrix and σ_v^2 is the variance of V_i, i.e. $\sigma_v^2 = \mathrm{Var}[V_i]$, (i=1,...,n), that controls the variability of the effects V_i between geographical areas[58,59,64].

Therefore, from (4.1) and (4.2) we obtain that $V_i \mid \sigma_v^2$ (i=1,...,n) are independent and identically distributed following a prior Normal distribution with mean 0 and variance σ_v^2:

$$[V_i \mid \sigma_v^2] \sim \mathrm{Normal}(0, \sigma_v^2) \qquad (4.3)$$

It is usual to consider that V_i represents covariates or risk factors not observed which are common to the individuals of the i-th area and which do not present a spatial pattern of any kind[58].

Instead of considering the distribution for V_i given in (4.3) we may also use a conditional prior distribution for $[V_i|V_j, j\neq i]$, (i=1,...,n)[65,74]. This formulation and (4.3) are equivalent under the linear restriction:

$$\sum_{i=1}^{n} V_i = 0 \qquad (4.4)$$

This prior distribution is defined as[65,74]:

$$[V_i \mid V_j, j \neq i, \sigma_v^2] \sim \mathrm{Normal}\,(\overline{V}_{-i}, \sigma_v^2) \qquad i=1,...,n$$

where $\overline{V}_{-i} = \frac{1}{(n-1)} \sum_{j\neq i}^{n} V_j$

In this way we find that the relative risk in the i-th area is displaced towards the global mean $\overline{V}_{-i}$ [65].

This conditional prior defines an improper distribution for the vector V consisting of the logarithms of the relative risks[58,65]. An improper distribution is one which is not defined by a probability measure or is not a probability density function[66]. Even though the joint distribution V is improper the posterior parameters V_i will be

identifiable, i.e. the data allow us to obtain information about such parameters and thus we can correctly carry out their estimation[75,76]. This prior distribution will also be described for models where the relative risks have spatial structure.

Other authors consider a decomposition of the model in which they add a constant representing the global mean (α_0) of the relative risks. For example, Clayton and Bernardinelli add a constant term when the rates of the reference population for the calculation of the expected cases E_i have been obtained externally since the SMR will not necessarily be centered around approximately 1[61]. When there is no decomposition involving a constant term we can use the conditional prior distribution directly. However if we do use the decomposition with a constant term we must impose a restriction. To illustrate this fact, we start from the following situation:

$$\log \theta_i = \alpha_0 + b_i$$

$$[b_i \mid b_j, j \neq i, \sigma_v^2] \sim \text{Normal}\,(\bar{b}_{-i}, \sigma_v^2)$$

where $\bar{b}_{-i} = \frac{1}{(n-1)} \sum_{j \neq i}^{n} b_j$

The restriction is imposed to ensure that the parameters of the model are identifiable. To do so we may consider two restrictions: 1) eliminate the constant from the model (α_0=0) thus returning to the initial situation or 2) consider that the mean of the random effects b_i (1,...,n) is equal to 0 or equivalently[58,65,77,78,79]:

$$\sum_{i=1}^{n} b_i = 0 \qquad (4.5)$$

In the FB approach, furthermore, it will be obligatory to consider a non-informative Uniform(-∞,∞) prior distribution for the constant term α_0[77,78,80]. It should be noted that these restrictions have no consequences on the posterior distributions of the parameters b_i[80].

The independent prior distribution given in (4.3) already incorporates restriction (4.5) expressed in (4.4) and so we can incorporate the constant α_0 with no additional form of restriction:

$$\log \theta_i = \alpha_0 + b_i$$

$$[b_i \,|\, \sigma_v^2] \sim \text{Normal}(0,\ \sigma_v^2)$$

On the other hand, the Normal prior distribution is preferable to the Gamma distribution described in the preceding section since it is easy to add covariates specific to the area[62]. Furthermore, it can easily be extended to include the spatial structure in the relative risks although progress has been made to do so using the Gamma distribution[58,63]. The extension to incorporate covariates does not modify the descriptions given above for the prior distribution and in the general case with a constant term this is done as follows:

$$\log \theta_i = \alpha_0 + \mathbf{X}_i^T \beta + b_i$$

Where X_i is a kx1 vector of variables of the i-th area and β is a kx1 vector of regression coefficients.

The posterior distribution of relative risk, given the sample, will not be easily obtained and as mentioned earlier we must resort to approximations[59] or appropriate algorithms which simulate the posterior distribution depending on the Bayesian approach utilized.

4.2.3 Discrete prior distribution.

Let the relative risks, $\theta_i|\gamma$, (i=1,...,n), with $\gamma = (p_1,..., p_K, \phi_1,..., \phi_K)^T$ be independent and identically distributed following a discrete prior distribution consisting of K probabilities $\{p_j; j=1,\ldots,K\}$ evaluated in $\{\phi_j; j=1,\ldots,K\}$[81,82,83,84], i.e.,

$$P(\theta_i=\phi_j) = p_j \qquad\qquad j=1,\ldots,K$$

where K may be either known or unknown. The case for K known is called fixed support, and when unknown as a flexible support. The components $\{p_j\}$ and $\{\phi_K\}$ are also usually denoted by:

$$F = \begin{bmatrix} \phi_1 \ldots \phi_K \\ p_1 \ldots p_K \end{bmatrix}$$

The parameters K, $\{p_j\}$, $\{\phi_K\}$ will be estimated by maximum likelihood based on the joint marginal or marginal likelihood denoted $P(O \mid F)$. In this case $P(O_i \mid F)$ is a finite mixture of Poisson distributions:

$$P(O_i \mid F) = \sum_{j=1}^{K} P(O_i \mid \theta_i = \phi_j) p_j$$

where

$$[O_i \mid \theta_i = \phi_j] \sim \text{Poisson}(\phi_j E_i) \qquad j=1,...,K;\ i=1,...,n$$

$$\sum_{j=1}^{K} p_j = 1,\ p_j \geq 0 \qquad j=1,...,K$$

When K is unknown the estimator of F receives the name of nonparametric maximum likelihood estimator[85].

Once the estimates of K, $\{p_j\}$ and $\{\phi_j\}$ have been obtained by maximum likelihood, denoted by $\hat{K}$, $\{\hat{p}_j\}$ and $\{\hat{\phi}_j\}$ respectively, we can calculate the non-parametric empirical Bayesian estimate of relative risk through the posterior mean:

$$\hat{\theta}_i^{NP} = \frac{\sum_{j=1}^{\hat{K}} \hat{\phi}_j P(O_i \mid \theta_i = \hat{\phi}_j) \hat{p}_j}{\sum_{j=1}^{\hat{K}} P(O_i \mid \theta_i = \hat{\phi}_j) \hat{p}_j}$$

It should be note that some authors had applied this approach under the fully Bayesian framework[86].

4.3 Prior distributions to represent clustering variation.

The prior distributions most widely used to represent spatial structure in the relative risks are the so-called Markov random fields (MRF). In MRF models the relative risk of the i-th area depends on the relative risks of the remaining areas based only on the relative risks of neighboring areas[61]. Therefore the relative risks θ_i, (i=1,...,n), are conditionally independent given the relative risks of neighboring areas.

An area "j" is considered a "neighbor" of an area "i" if knowing the relative risk of the "j" area (θ_j) provides information about the relative risk of the "i" area (θ_i) due to its similarity with respect to some "local attribute", such as for example, the level of unemployment or historical or cultural characteristics[87]. Usually the "neighbor" areas are taken as those which are contiguous with the i-th area.

MRF models include the so-called Gaussian MRF or conditional autoregressive normal (CARN) models[58,61,62,64]. These models were originally applied by a number of authors to the process of reconstruction of images[79] and they consider that the distribution of relative risk of each area conditioned to the neighboring areas is Normal.

In general each area will not have the same number of "neighbors". When the areas do not all have the same constant number of "neighbors" the CARN prior is known as intrinsic, i.e. intrinsic conditional autoregressive Gaussian distribution[61,79,82].

4.3.1 Intrinsic CARN prior distribution.

We start with the general situation in which we have a vector $U = (\log \theta_1 ,..., \log \theta_n)^T = (U_1,...,U_n)^T$. Our aim will be to consider a prior distribution over the vector U which takes account of dependencies between U_i and U_j, $j \neq i$[58]. In the choice of a prior distribution over U we have two options[58]:

1) Specify a joint multivariate probability distribution on U.

2) Consider conditional univariate distributions $[U_i|U_j=u_j,j\neq i]$ (i=1,...,n; j=1,...,n).

In order to arrive at the expression of the intrinsic CARN prior distribution we will choose the second strategy.

The strategy based on conditional distributions is the one usually used in spatial contexts. This option is easier to handle than the joint distribution strategy and it reduces to specifying a smaller number of parameters instead of having to determine all the elements of the matrix of variances-covariances[58]. Among other advantages it also permits simpler estimation of the parameters since we can avoid the need to invert the variances-covariances matrix[81]. For more information on strategy 1) consult the description by Wakefield et al[58] or Pascutto et al[55].

To consider strategy 2) we must obtain the conditional univariate distributions $[U_i|U_j=u_j,j\neq i]$. Specifically, the intrinsic CARN prior distribution on U may consider that the conditional functions $[U_i|U_j=u_j,j\neq i]$ follow univariate Normal distributions with conditional moments equal to[58,62,64]:

$$E[U_i \mid U_j = u_j, j \neq i, w_U^2] = E[U_i \mid U_j = u_j, j \in \delta_i, w_U^2] = \frac{1}{m_i}\sum_{j\in\delta_i} u_j \qquad (4.6)$$

$$Var[U_i \mid U_j = u_j, j \neq i, w_U^2] = Var[U_i \mid U_j = u_j, j \in \delta_i, w_U^2] = \frac{w_u^2}{m_i} \qquad (4.7)$$

where

δ_i = set of areas neighboring the i-th area.

m_i = number of areas neighboring the i-th area.

In this case, w_u^2 controls the variability of the random effects conditioned by the effects of the neighboring areas.

Just as in the case of heterogeneity, the intrinsic prior CARN distribution leads us to an improper multivariate distribution on U and the restrictions mentioned above must be considered in order to ensure that the model be identifiable when a constant term is incorporated. Analogously if no constant term is included the intrinsic prior

CARN distribution may be applied without any restrictions. It may also be extended to include covariables just as in the case of the normal prior distribution in 4.2.2.

4.3.2 Other prior distributions for spatial structure.

There are at least two other prior distributions apart from the Gaussian for modeling the conditional spatial structure in relative risks[58]. The double-exponential prior or Laplace distribution is particularly robust, based on the median of the relative risks of adjacent areas instead of the mean[58,77,88]. In adittion, we can considered a proper Gaussian prior distribution that has been criticed by several authors[7,77,78].

4.4 Prior distributions for representing joint variation (heterogeneity and clustering).

As mentioned earlier when it is not known with certainty which form of geographical variation is present in the relative risks the two (heterogeneity and clustering) may be considered jointly. This model is known as the Besag, York and Mollié (BYM) model after the authors who proposed it. We will consider in this case[62,64,82]:

$$\log \theta_i = V_i + U_i$$

where V_i and U_i are independent and defined as in (4.3), (4.6) and (4.7), i.e.:

$$[V_i \mid \sigma_v^2] \sim \text{Normal}(0,\ \sigma_v^2)$$

$$[U_i \mid U_j = u_j, j \in \delta_i, w_u^2] = \text{Normal}\left(\frac{1}{m_i}\sum_{j\in\delta_i} u_j, \frac{w_u^2}{m_i}\right)$$

In this case[64]:

$$E[\log \theta_i \mid U_j = u_j, V_j = v_j, j \neq i, \sigma_v^2, w_u^2] = \frac{1}{m_i}\sum_{j\in\delta_i} u_j$$

$$\text{Var}[\log \theta_i \mid U_j = u_j, V_j = v_j, j \neq i, \sigma_v^2, w_u^2] = \sigma_v^2 + \frac{w_u^2}{m_i}$$

From this it may be deduced that when:

$$\frac{w_u^2}{\sigma_v^2} > 1$$

the structured or spatial variation will dominate while in the opposite case variation without structure will dominate. As described above in section 4.2, w_u^2 controls the variability of the random effects conditioned by the effects of the neighboring areas. For this reason it would be more appropriate to compare the unstructured marginal variance σ_v^2 with the structured marginal variance denoted as σ_u^2. However, the only way of obtaining σ_u^2 is from the marginal covariance of $(\log \theta_1,...,\log \theta_n)^T$ but this does not exist[64]. To determine which of the types of variation dominates the structure of the relative risks Molliè[64] proposes comparing the ratio of the empirical marginal variances taken as approximations of σ_v^2 and σ_u^2. These empirical variances are defined, respectively, as[64]:

$$s_V^2 = \frac{1}{n-1}\sum_{i=1}^{n}(V_i - \overline{V})^2 \qquad s_U^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \overline{U})^2$$

4.5 Choice of the hyperprior distribution.

As described in the previous chapter the EB approach estimates the hyperparameters vector γ of the data. If we consider that the effects β are fixed we may use the same strategy to estimate the constant α_0 and the vector of β coefficients[58,62]. In the EB approach rather than fixed effects we could consider a prior distribution over β and consider them as random effects. In this way the parameters of the distributions fixed over β will be those estimated based on the data.

In contrast, the FB approach does not estimate the γ parameters of the data but rather uses a third level in which a hyperprior distribution is considered for γ. Analogously, we consider a distribution for the constant term α_0 and the vector of β coefficients.

The hyperprior distributions used in the FB approach are "non-informative"[62]. A non-informative distribution for some particular parameter is one which contains no information about that parameter in the sense of not favoring any of the values it may possibly take[67].

The non-informative priors fixed over β are usually improper such as the Uniform(-∞,∞) distribution or a proper prior such as the Normal distribution with a very large variance. Earlier we noted that it was obligatory to consider a Uniform(-∞,∞) non-informative prior distribution for the constant term α_0 when considering the restriction which ensured that the parameters would be identifiable[77,78,80].

The present section will only describe the hyperpriors for the heterogeneity and clustering variation cases and in consequence when we combine the two types of variation.

We could consider an improper non-informative Uniform(-∞,∞) prior distribution for $\log \sigma_u^2$ and $\log w_v^2$ as well as for the constant α_0. In the case of α_0 the posterior distribution is not improper, however for $\log \sigma_u^2$ and $\log w_v^2$ an improper posterior distribution results, the consequence being that all the relative risks estimated a posteriori are equal[62].

In this case hyperprior distributions are not fixed over $\log \sigma_u^2$ and $\log w_v^2$, but usually rather over the precision or inverse of the variances, i.e. over:

$$\frac{1}{\sigma_u^2} \qquad \text{and} \qquad \frac{1}{w_v^2}$$

The hyperprior fixed is usually a proper Gamma(a,b) distribution with scale parameter a and form parameter b, both strictly positive. As Thomas et al[77,78] and Wakefield et al[58] cite, care must be taken with the parameters which form the hyperprior distribution over the inverses of the variances and it is advisable to perform sensitivity analyses taking various distributions or different values of the hyperparameters of the

hyperprior distribution. In this way we may assess whether there are discrepancies or similarities between the results obtained a posteriori, as done by for example Molliè[64] and Bernardinelli[89]. In this latter article Bernardinelli et al. use the chi-squared distribution as the hyperprior for precision.

Molliè proposes various alternatives for obtaining the parameters of the Gamma distribution[62,64]. For example, in the case where there is no a priori information Molliè generally considers taking a Gamma(a,b) prior where the mean a/b is based on the observed log SMRs, and the variance a/b^2 is very large.

Other authors take very low values for the parameters a and b of the Gamma distribution, for example of the order of a=0.01 and b=0.001, however as Kelsall and Wakefield point out the Gamma distribution with such priors is usually highly informative[58,77,78,90]. As an alternative they propose a Gamma(0.5,0.0005) distribution for the inverse of the variances which in many cases provides a wide range of possible values for the relative risk.

Other proposals have recently be made based on a non-informative Uniform(a,b) distribution on the standard deviation of the random effects [91,92].

4.6 Which prior distribution should be chosen?

Earlier we described three main types of prior distributions which may be considered, depending on the expected variation in the relative risks. However often it is not clear what variation will be dominant and usually a model is used which combines both types of variation although such a model may not be necessary or appropriate. For example some authors believe the BYM model could oversmooth the relative risk surface and consequently have developed semi-parametric spatial models that allow discontinuities in the risk[93,94].

According to Bernardinelli et al[95]: "*The choice between the clustering and heterogeneity model depends upon our prior belief about the size of high/low risk clusters. A cluster size bigger than the area size leads to a clustering model, while a*

cluster size smaller than the area size leads to a heterogeneity model. Althought it is possible to include both these terms in the model, this may not be necessary". From this it may be deduced that for large geographical areas it seems reasonable to apply a prior for heterogeneity while for small it is better to use a clustering prior.

Other studies have gone further and evaluated empirically the different prior distributions fixed over the relative risks:

Lawson et al[63], using simulation, found that the Gamma-Poisson model, linear Bayes methods and the model combining variation of heterogeneity and clustering were generally the most robust, followed by discrete non-parametric prior based on mixtures which were less robust, and finally the non-parametric smoothing methods which in general did not behave well.

Yasui et al.[96] also evaluated various prior distributions for the estimation of relative risk in small areas based on real data corresponding to municipalities (or aggregates thereof) in Spain. In their assessment they found that the discrete non-parametric prior based on mixture models behaved well in the estimation of relative risks in areas of low risk, while the spatial priors based on the Normal distribution behaved well and provided good estimates for the areas with high risk. They also obtained an *ad hoc* estimate of the relative risks by averaging estimates of relative risks obtained using two priors (discrete non-parametric non-structured, and normal structured) which in general was well-behaved.

Militino et al[97] focus their research on investigating the performance of certain models in identifying high risk areas. In their research discrete mixture models perform well in locating regions which experience high risk. Normal models also work well in identifying high risk areas and perform better when there is spatial autocorrelation.

Recently Best et al.[93] have compared a variety of models established to date in obtaining the relative risks in disease mapping studies, including the BYM model and semi-parametric models. The results suggest that the BYM model and semi-parametric models perform well for modelling a single disease. In consequence the authors suggest that the BYM model remains an appropriate tool for small area disease mapping.

The behavior of the different models fitted can be assessed based on the residuals[7,63]. As Lawson et al.[7] report, the *Bayesian Information Criterion* is commonly used as the overall measure of goodness of fit, and more recently the *Deviance Information Criterion* has come into use[98]. From the Bayesian point of view a series of residual and specific diagnoses have also been defined. For more detail the reader may consult Lawson et al[7], Carlin and Louis[67], and Stern and Cressie[99].

4.7 Other ways of incorporating the spatial configuration: multiple-membership models.

In chapter 1 we described the multilevel models as those in which there is a hierarchy of levels in the data. In these models each unit belongs to only one unit of the level above it. For example, in Spain we may consider a geographical structure formed of municipalities which group into 53 provinces and these into 17 autonomous communities. However it can happen that an inferior unit may belong to, or be influenced by, more than one of the higher level groups of the hierarchy. In this case an extension of the multilevel models must be used, known as multiple-membership models (MM)[7,100]. For example, consider the classical application of multilevel models in which we have students grouped into schools. It may happen that the exact school which certain individuals attend may not be known, only that they attend one of a set of possible schools, or there may be students who change school during the course of their education and consequently "belong" to more than one school.[7]

If we translate this idea to the geographical context, we may consider that each area is influenced by more than one group of a superior level where each one of these superior groups will be areas defined as neighbors. The spatial configuration is added into the model directly, by incorporating a weighted sum of the random effects corresponding to each area's neighbors into the linear predictor.

As Lawson et al.[7] describe, the MM model and extensions were used in the geographical context by Langford et al.[101]. For more details, the reader may also consult Browne et al.[102].