Universidad Autónoma de Madrid

Facultad de Medicina

Departamento de Medicina Preventiva y Salud Pública

FROM DISEASE MAPPING TO FOCUSED CLUSTERING: Assessment and Exploration of Spatial Epidemiology Techniques for Studying Cancer in Spain.

Thesis by

REBECA RAMIS PRIETO

Supervised by

Dr. GONZALO LÓPEZ-ABENTE

Madrid 2009

Dr. Gonzalo López-Abente, Jefe del Área de Epidemiología Ambiental y Cáncer del Centro Nacional de Epidemiología. Instituto de Salud Carlos III.

INFORMA:

Que Dña. Rebeca Ramis Prieto ha realizado bajo su dirección el trabajo titulado “From disease mapping to focused clustering: Assessment and Exploration of Spatial Epidemiology Techniques for studying Cancer in Spain”. Es un trabajo original, rigurosamente realizado, y es apto para ser defendido públicamente con el fin de obtener el grado de doctor.

Para que así conste y surta los efectos oportunos, se firma este documento en Madrid, a 23 de septiembre de 2009.

FROM DISEASE MAPPING TO FOCUSED CLUSTERING: Assessment and Exploration of Spatial Epidemiology Techniques for studying Cancer in Spain.

by

REBECA RAMIS PRIETO

Department of Environmental Epidemiology and Cancer, National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain

Consortium for Biomedical Research in Epidemiology & Public Health (CIBER en Epidemiología y Salud Pública – CIBERESP), Madrid, Spain

Supervised by

Dr. G LÓPEZ-ABENTE

Head of the Department of Environmental Epidemiology and Cancer, National Centre for Epidemiology, Carlos III Institute of Health, Madrid, Spain

Consortium for Biomedical Research in Epidemiology & Public Health (CIBER en Epidemiología y Salud Pública – CIBERESP), Madrid, Spain

MADRID 2009

A mis padres, Chus y Pepe, y mis hermanas, las kekas.

Acknowledgments

I am extremely grateful for the continuing support and guidance of the head of my department and thesis supervisor Gonzalo who introduced me to epidemiology and showed me how interesting spatial epidemiology can be. He is a great researcher and I consider him as my mentor in the scientific world.

I would like to express my sincere gratitude to all the staff in the Area of Environmental and Cancer Epidemiology of the National Centre of Epidemiology for their daily support, help and friendship. Elena, Marina, Nuria, Bea, Virginia, Javi, Quique, Pablo, Anna, Feli, Cristina and other former colleagues I have learnt a great deal from you.

I am deeply grateful to the people at the National Centre of Epidemiology who make the centre a great place to work in.

I would like to give special recognition to Peter Diggle and the CHICAS (Combining Health Information, Computing and Statistics) group from the Medicine Department at Lancaster University, for welcoming me and showing me the biostatistical point of view. I also thank Virgilio for his support, and the advice to go Lancaster University for a year.

This thesis is a collection of papers so I would also like to thank all of the co-authors.

I thank the Health Department of the Basque Country Government for providing data.

Finally, I wish to thank Alex for his support and work as English proof-reader and editor.

Index

1. INTRODUCTION

This thesis seeks to depth in the spatial epidemiology methods aiming to apply them to the study of cancer distribution and its relations with environmental factors.

1.1. SPATIAL EPIDEMIOLOGY

When an epidemiological study is carried out one of the first conclusions about the distribution of the health event is: *its occurrence is not uniformly distributed either in space or time*. Variations on the appearance of health events are consequence of population structure, population density and variations in the remaining risk factors. Health determinants depend on individual characteristics, such as age, sex and genetic factors, but also on lifestyle variables, for instance smoking and diet, along with another environmental and occupational exposures. Based on that idea there are three main aims or questions in the analysis of a disease spatial distribution:

1. Knowledge about the spatial distribution: Are there areas with higher risk than others?
2. Possible relationship with environmental factors: Is the spatial distribution of the disease somehow related to the spatial distribution of a risk factor measured at the same aggregation level?
3. Location of the high risk areas: Are high risk areas geographically clustered or randomly spread?

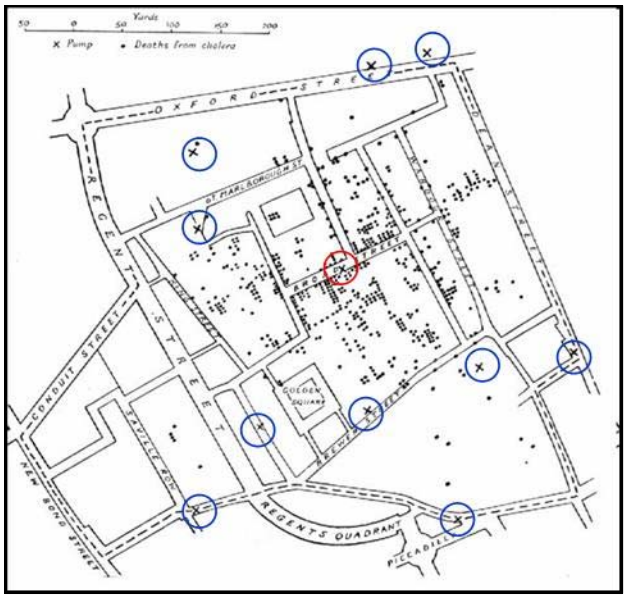
Spatial epidemiology has been developed to answer these questions. Nowadays it is the component of epidemiological science that analyses the spatial variations of health events. A formal definition can be:

> *Spatial epidemiology is the part of epidemiology that studies the variations in geographical distribution of health events, seeking the description and understanding of such variations.*

State of the art (in brief)

In 1854 John Snow presented an analysis of a cholera outbreak in London. He studied the geographical location of the cases in relation to the location of water pumps. The results were the identification of the pump responsible for the disease (Map 1.1) and a better knowledge about the cholera disease. This study is known as one of the first epidemiological analyses in history, and it can be considered the first study of spatial epidemiology [Snow, 1855]. Since then

until the 90's decade, spatial epidemiology was mainly used to create maps to describe the geographical pattern of health outcomes [Walter S D, 2000]; however during the last 25 years epidemiologists and mathematicians have worked together on the development of new techniques to solve epidemiological questions of a spatial nature [Elliott et al., 2000].



Map 1.1. Location of the cases (·) and water pumps (x) in a cholera outbreak in London in 1854.

This improvement has been possible because of the greater availability of geographically indexed health and population data and, of advances in computing and GIS (Geographical Information System). We can take as an example "The Small Area Health Statistics Unit" (SAHSU) established in the United Kingdom in 1987 (http://www.sahsu.org/index.php)[SAHSU, 2009].

Also, the increasing interest in spatial distribution of diseases is related to the increasing concern about the environment and its association with health, thus spatial epidemiology is closely linked to environmental epidemiology. Environmental exposures to harmful products can affect neighbourhoods, towns even whole regions (e.g. Chernobyl nuclear reactor meltdown in the former Soviet Union in the 1980s). The study and assessment of these environmental damages and their consequences for the exposed population need spatial characterisation and understanding; knowledge about exposures is essential and it can be approached, for instance, by maps, ecological correlation analysis or cluster analysis. Spatial epidemiology makes use of spatial statistics tools to work on these problems [Cressie N, 2000; Elliott et al., 2000].

Several books [Cressie N, 2000; Diggle, 1983; Elliott et al., 2000; Lawson A, 2001 ;Lawson, 1999; Waller and Gotway C., 2004] and an important number of papers have been published during the last 25 years presenting as many methods as their applications. Furthermore national

and regional incidence and mortality atlases are published every year around the world [Benach et al., 2001 ;Boyle and Smans M, 2008; Lopez-Abente et al., 2001; Lopez-Abente et al., 2006b; Martinez-Beneito et al., 2005].

Nowadays, under the denomination of spatial epidemiology, several statistical methods can be found corresponding to the different aims of the study [Elliott et al., 2000]:

I. Disease mapping.

II. Ecological regression (geographical correlation studies).

III. Assessment of risk in relation to a point source.

IV. Cluster detection and disease clustering.

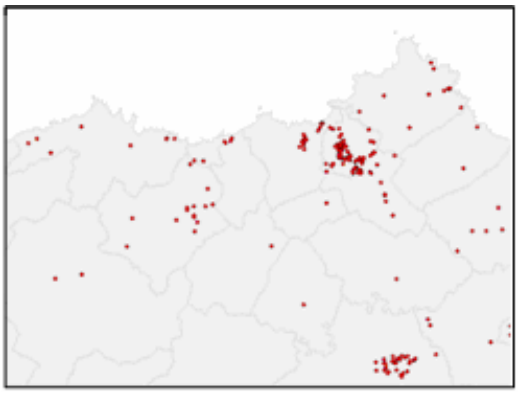
I. Disease mapping

Disease maps are representations of incidence or mortality data in their geographical context seeking to summarize the variation of the spatial distribution of diseases [Lawson, 1999]. Depending on the purpose the map can show different information: Location of individual cases within a region (Map 1.2), counts of cases in areas, rates in cities or countries or other risk estimators (Map 1.3).

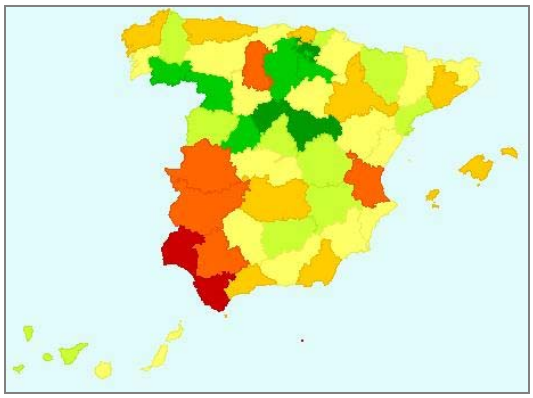
The main uses of maps:

- Descriptive studies looking for:
    - knowledge of the spatial distribution of an event.
    - Improving the knowledge about health needs of the population.
    - Better management of health resources.
- Analytical studies looking for:
    - Risk factors. For example, comparison with exposure maps
    - Assessment of health programmes and policies.

Many disease maps use simple descriptive measures such as rates or ratios to describe the distribution; however, there are several difficulties when crude ratios are used in small areas. Standard Mortality Ratio is one of the most used ratios in disease mapping. It is defined as the rate between the number of observed cases and the number of expected cases. This definition means that areas with low population have a small denominator therefore they produce SMRs with large variability and extreme values may appear when rare diseases are under study, misleading the interpretation of the map.

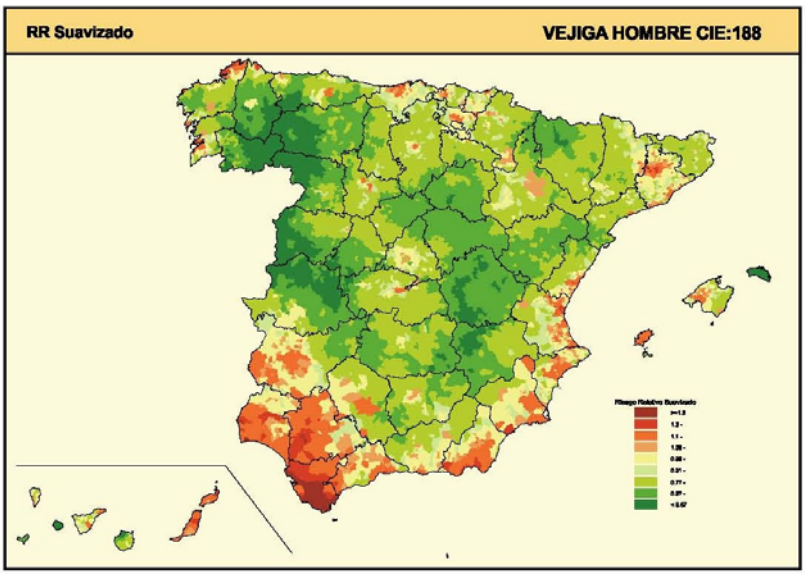


Map 1.2. Location of individual cases



Map 1.3. Aggregated cases or rates in areas

The most widely used strategy to approach these problems to estimate the spatial distribution of risk is the application models based on Poisson inference (Map 1.4) [2005] (Special Issue of Statistical Methods in Medical Research on disease mapping).



Map 1.4. Estimation of the spatial distribution of risk (Smoothed relative risk of bladder cancer)

II. Ecological regression (geographical correlation studies)

Ecological analysis examines associations between disease incidence or mortality and potential risk factors as measured on groups rather than individuals. Typically the groups are defined by geographical area such as country, region, municipality or census track. The main advantage of this kind of study is the availability of data, both health data and risk factors, at that scale. Information about lifestyle, such as diet or smoking, and exposures to environmental factors is not usually available at individual level; however, that sort of data is easy to collect at aggregated level from official (administrative) sources. On the other hand, the main disadvantage is the so-called “ecological fallacy” [Selvin et al., 1992].

Ecological fallacy is an error in the interpretation of statistical data in an ecological study that can occur when a researcher or analyst makes an inference about an individual based on aggregate data.

Accordingly, when an ecological analysis is performed over geographically referenced data, spatial autocorrelation should be also taken into account aiming to reduce the possibilities of ecological fallacy [Clayton et al., 1993].

III. Assessment of risk in relation to a point source

Point source studies are applied to assess increases in incidence or mortality of diseases in adjoining populations of potential environmental hazards. These kind of studies are sometimes carried out because of worries of the local population or media reports in reference to point sources of pollution. To deal with these analyses specific statistical methods have been developed, however, sufficient geographical resolution data are essential to produce accurate results. In addition, when the study is carried out because of a worry or a media report interpretation of the results is more difficult since the prior hypothesis could be biased.

Throughout the two last decades, concern about environmental hazard from industrial facilities has been illustrate in many studies. In the literature the majority of these studies have been focused on the detection of patterns of health events associated with air pollution and ionising radiation exposure [Kokki, 2004].

The lack of real exposure measurements in many studies has encouraged the use of estimated exposure measures. Researchers who have worked analysing air pollution from industrial facilities have mainly used the distance to the sources as surrogate of the real exposure.

During the late 80's and early 90's concern about cancer incidence in the surrounding population to nuclear plants and nuclear waste processing plants produced diverse point source studies in various countries, such as the United Kingdom [Roman et al., 1987; Gardner and Winter, 1984; Ewings et al., 1989], France [Viel and Richardson, 1990] and Spain [Lopez-Abente et al., 1999]. Furthermore in the 90s, other kind of facilities were examined too. Several studies about cancer in the vicinity of the petrochemical plant in Baglan Bay, Wales, were published [Lyons et al., 1995; Sans et al., 1995]. Also in the United Kingdom, incidence of cancer in the proximities of waste incinerators were analysed [Elliott et al., 1992; Elliott et al., 1996]. Other waste incinerators were studied in Italy [Michelozzi et al., 1998] and France [Viel et al., 2000].In 2008, Dreassi discussed the relationship between disease occurrence and distance from pollutant sources performing a sensitivity analysis over four different functional forms for the decay function of risk with increasing distance [Dreassi et al., 2008].

IV. Cluster detection and disease clustering

Finally, the last category of spatial epidemiology methods is cluster analysis. These studies seek to evaluate clusters of diseases. Normally the analysis involve the use of several statistical tests [Lawson A, 2001]. Generally cluster analysis is divided into global clustering and cluster detection. Global clustering methods test the overall global spatial correlation, and cluster detection methods identify unusual collections of events compared with others. Global clustering methods include, among others, Moran's I, Tango's and Besag–Newell's *R* statistics, being these the most widely used. On the other hand, cluster detection methods include, among others, circular and elliptic spatial scan statistics (SaTScan), flexibly shaped spatial scan statistics [Kulldorff M, 2006], Turnbull's cluster evaluation permutation procedure, local indicators of spatial association, and upper-level set scan statistics [Huang et al., 2008], where SaTScan is recognised as one of the most competitive [Duczmal et al., 2005].

Some of these methods have been widely applied in medical research, specifically for the study of disease occurrence, both communicative and chronic disease. However these methods have not been used in the present thesis.

1.2. CANCER

I. Cancer

Cancer is a heterogeneous family of diseases, consisting of over 100 different forms, that spawn from almost every cell type in the body. Each cell type gives rise to distinct forms of cancer, however despite the broad diversity, several features are common to all cancers: Cellular proliferation, circumvention of cell cycle control growth without appropriate signals, escape from programmed cell death, altered interactions between cells and the surrounding environment, evasion of immune-mediated eradication and invasiveness into normal tissue [Adami et al., 2002].

Due to the diversity of possible tumoural locations (Table 1.1) and the variety of possible risk factors the study of cancer is particularly complicated. The majority of cancers have a complex aetiology where one or more environmental risk factors interact with genetic background, age, sex, socio-demographic status and other factors [Wild, 2009].

During the last decades cancer incidence has been continuously increasing. However, scientists consider that in Western countries expansion and ageing of the population, as well as progress in cancer detection using new diagnostic and screening tests cannot fully account for the observed growing incidence of cancer. Besides, well established risks factors such as alcohol consumption and tobacco smoking in men have significantly decreased lately. On the other hand, during the same period the environment has substantially changed and many

carcinogenic factors have been accumulated in the environment [Belpomme et al., 2007b]. In this regard, many researchers have carried out studies analysing the relationship between cancer and exposure to environmental factors [Belpomme et al., 2007a].

Tumours	ICD 9	ICD 10
Buccal cavity and pharynx	140-149	C00-C14
Esophagus	150	C15
Stomach	151	C16
Colon-Rectum	153-154	C18-21
Gall-Bladder	156	C23
Pancreas	157	C25
Larynx	161	C33
Lung	162	C34
Bones	170	C40-41
Connective tissue	171	C49
Melanoma	172	C43
Breast	174	C50
Uterus	179-182	C54-55
Ovary	183	C56
Prostate	185	C61
Bladder	188	C67
Kindney	189	C64
Brain	191	C71
Non Hodgkin's limphomas	200,202	C82
Myeloma	203	C90
Leukemias	204-208	C91-95

Table 1.1. Most common tumoural locations and their codes for ICD 9 and ICD 10.

II. Cancer figures. (Burden of disease)

Cancer in Europe

The European Commission estimates that in Europe cancer affects 1 in 3 men and 1 in 4 women at some time in their lives and that 1,2 million EU citizens die from cancer each year, that is equalling about one in four deaths in Europe. Every year 3.2 million Europeans are diagnosed with cancer, which is also the second most common cause of death in Europe (29% of deaths for men, 23% for women). The most frequently tumoural locations are breast, colorectal and lung cancers [IARC, 2009a].

In 2007, Ferlay published estimated incidence and mortality rates for all European countries for 2006 [Ferlay et al., 2007]. Figures for Spain and Europe are shown in Table 1.2 and Table 1.3. From the results it can be seen that Spain presents lower incidence rates than Europe as a whole, although it has higher estimated mortality rates for colon-rectum and lung cancer among men.

	Stomach		Colon-rectum		Lung		Breast	Uterus	Prostate	All cancer	
	M	F	M	F	M	F	F			M	F
Spain	15.9	8.4	54.4	25.4	68.3	13.8	93.6	24.5	77.2	416.9	263.4
Europe	24.8	11.6	55.4	34.6	75.3	18.3	94.3	33.5	86.9	469.7	303

Table 1.2. Estimated age-standardised incidence rates (European standard population) per 100.0000 person/year by sex, 2006. Source: Ferlay 2007

	Stomach		Colon-rectum		Lung		Breast	Uterus	Prostate	All cancer	
	M	F	M	F	M	F	F			M	F
Spain	12.7	5.8	28.2	14.6	67.2	8.9	19.2	5.6	18.4	237.0	106.5
Europe	18.1	8.3	27.3	16.6	64.8	15.1	26.0	9.3	22.2	244.8	135.4

Table 1.3. Estimated age-standardised mortality rates (European standard population) per 100.0000 person/year by sex, 2006. Source: Ferlay 2007

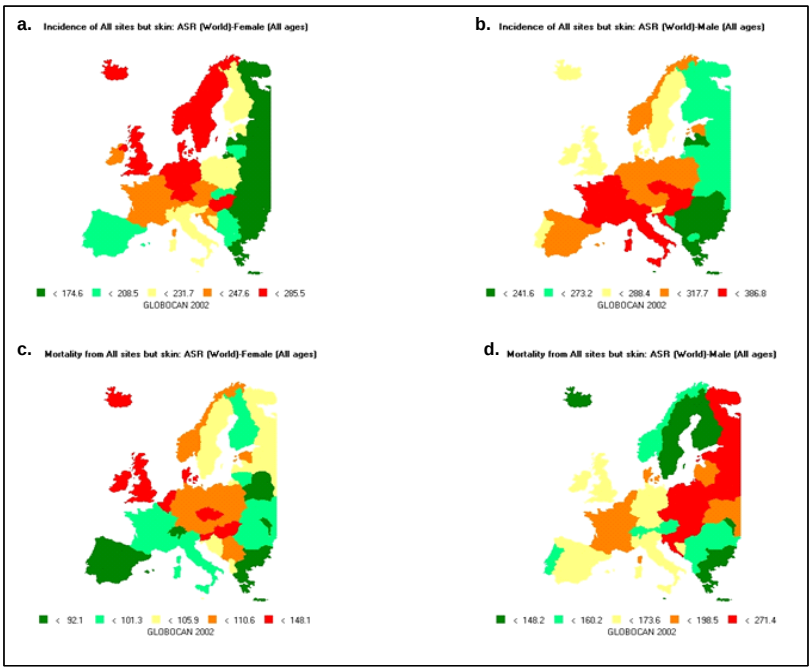


Figure 1.1. Cancer distribution in Europe: (a) estimated incidence rate from all cancer sites but skin for women; (b) estimated incidence rate from all cancer sites but skin for men; (c) estimated mortality rate from all cancer sites but skin for women; (d) estimated mortality rate from all cancer sites but skin for men. Source: Globocan 2002

On the other hand, the Globocan 2002 database from the International Agency for Research on Cancer, IARC, which has been built using data from the cancer registers of the different countries, presents estimates of incidence and mortality for 2002 [IARC, 2005] (http://www-dep.iarc.fr/). According to Globocan 2002, the estimated number of cancer cases in Europe is 2.820.774, with a sex distribution of 1.321.130 women and 1.499.664 men. As follows, we present some maps and graphs from Globocan 2002. Figure 1.1 shows four maps of Europe with the estimated incidence and mortality rates by sex. Again, the maps show that Spain generally has lower estimated rates than the European mean: only the estimated mortality rate for men reaches the mean of the interval.

The following two figures, Figures 1.2 A) ,B) and C), show the temporary trends of the incidence rates for both women and men. Data provided from the Eurpean Observatory Cancer [IARC, 2009a]. The graphs have several lines, one for each European country with available data. All the graphs confirm that the trends for the different countries present similar increase with time.

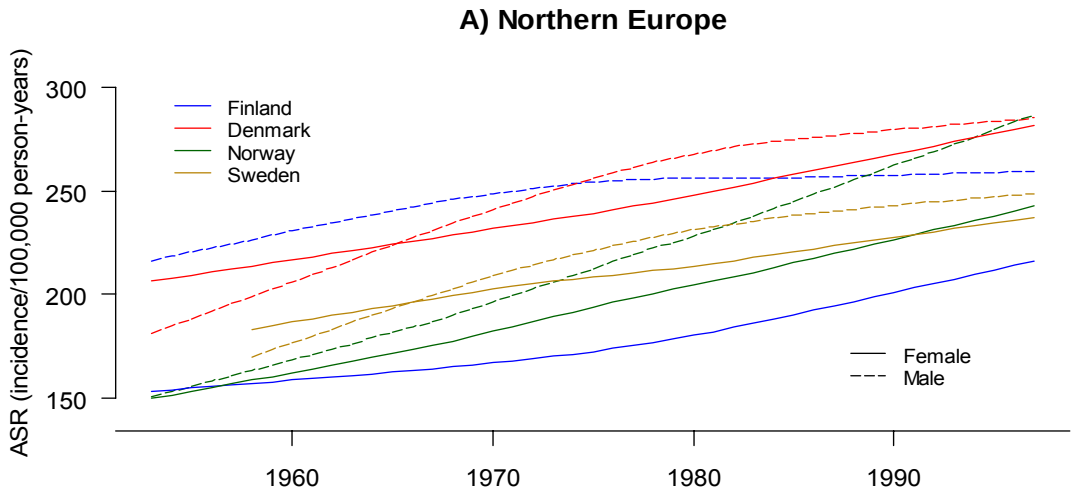


Figure 1.2. A) Time trend of incidence rate for all cancer sites but skin for women. Northern Europe. Source: European Observatory Cancer, IARC

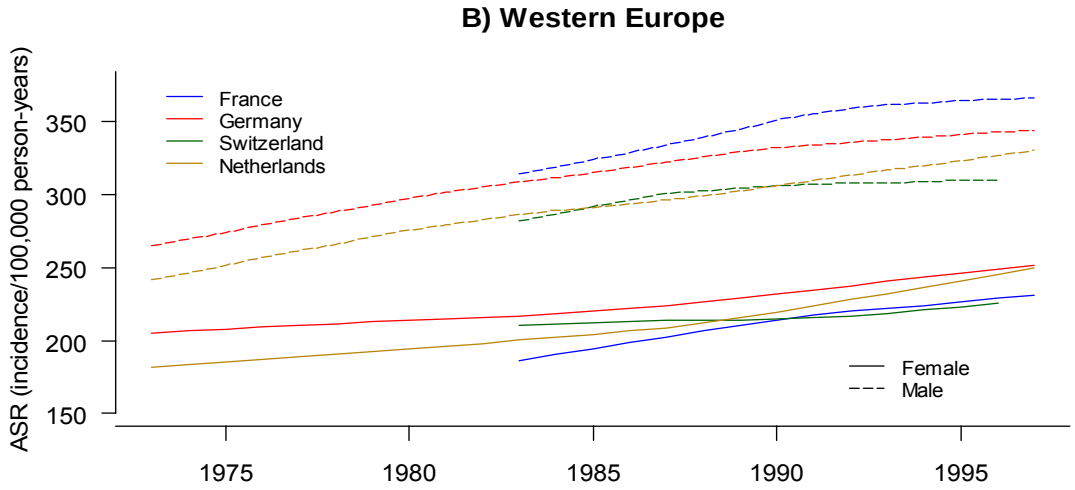


Figure 1.2. B) Time trend of incidence rate for all cancer sites but skin for men. Western Europe. Source: European Observatory Cancer, IARC

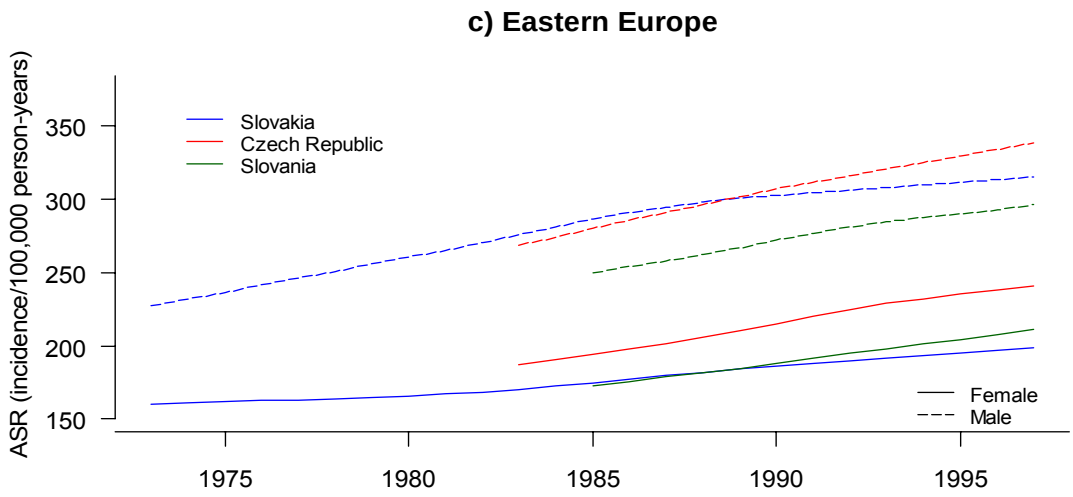


Figure 1.2. C) Time trend of incidence rate for all cancer sites but skin for men. Eastern Europe. Source: European Observatory Cancer, IARC

Cancer in Spain

Incidence rates for Spain show the same trends that those from other European Countries. (Figure 1.3). On the other hand the number of cancer cases that Globocan estimated for Spain for 2002 was 63.983 among women and 97.765 among men [IARC, 2005].

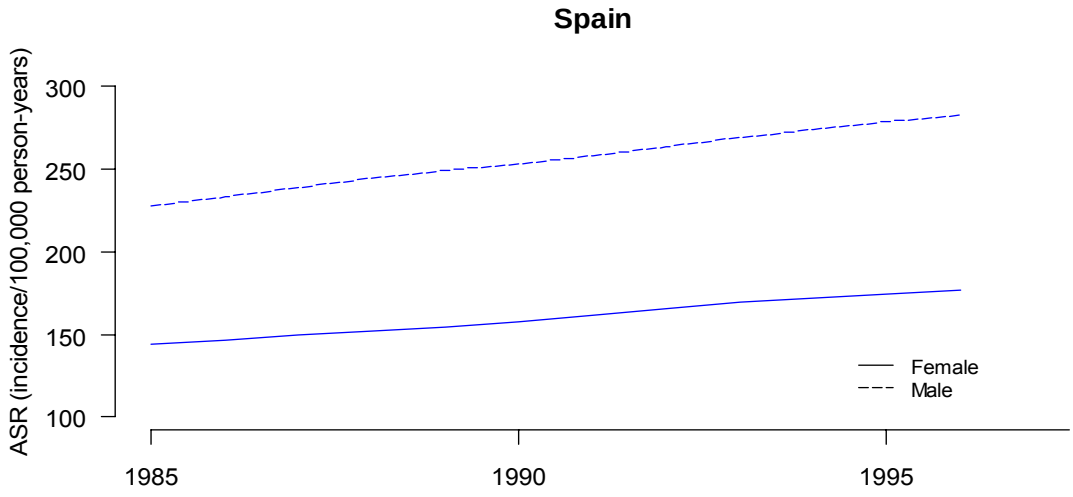


Figure 1.3. Time trend of estimated incidence rates of all cancer sites but skin. Spain. Source: European Observatory Cancer

Finally, for 2007, mortality data from the INE (National Statistic Institute) says that there were 99,763 deaths by cancer in Spain, 62,430 men and 37,333 women, up to 383,249 of total

deaths, which equates to a mean of 26%. In other words, one in four deaths was caused by cancer. Disaggregating by cancer type and sex the most important among men was lung cancer (17.162 deaths) followed by colorectal (7.857 deaths), and then prostate cancer (5.574 deaths).For women breast cancer (5.904 deaths) was the most frequent followed by colorectal (5.638 deaths) and then lung cancer (2.786 deaths).

1.3. ENVIRONMENTAL INDUSTRIAL POLLUTION

In the previous section we have talked about cancer and its aetiology. In the present thesis we aim to analyse the relation between cancer and environmental exposures from industrial pollution: however, the main difficulty in these kind of studies is the availability of suitable and accurate information.

In January 2000 the European Council approved a directive for the implementation of a European Pollutant Emission Register (EPER) (Decision 2000/ 479/CE) [EPER, 2004]. Under the terms of this Decision, all Member States were required to report industries relative to 50 pollutant emissions in excess of a given threshold. The European Pollutant Emission Register (EPER) collects information about emissions to air, soil and water from all agricultural or industrial facilities engaging in one or more activities listed in Annex I to Council Directive 96/61/EC [Commission of the European Communities, 2000].

The available information allows classification of different types of industrial activities, besides, containing abundant data on emissions of the pollutant substances and the amount released annually. In February 2004, EPER data of Spain (for 2001) were published.

Industrial activities classified in the EPER fall into the following 6 categories:

1. Energy industries;
2. Production and processing of metals;
3. Mineral industry;
4. Chemical industry and chemical installations;
5. Waste management; and
6. Other activities (which include paper and board production, manufacture of fibres or textiles, tanning of hides and skins, slaughterhouses, intensive poultry or pig rearing, installations using organic solvents, and the production of carbon or graphite).

The database also classifies the 50 declared pollutant substances into the following groups:

1. Environmental themes: methane, carbon monoxide, carbon dioxide, hydrofluorocarbons, nitrous oxide, ammonia, non-methane volatile organic compounds (NMVOC), nitrogen dioxide, perfluorocarbons, sulphur hexafluoride, sulphur dioxide, nitrogen and phosphorus.
2. Metals and metal compounds: arsenic, cadmium, chromium, copper, mercury, nickel, lead and zinc.
3. Chlorinated organic substances: dichloroethane-1,2, dichloromethane, chloroalkanes, hexachlorobenzene, hexachlorobutadiene, hexachlorocyclohexane, halogenated organic compounds, dioxins and furans, pentachlorophenol, tetrachloroethylene, tetrachloromethane, trichlorobenzenes, trichloroethane-1,1,1, trichloroethylene and trichloromethane.
4. Other organic compounds: benzene, toluene, ethylbenzene, xylenes, brominated diphenylether, organotincompounds, polycyclic aromatic hydrocarbons (PAH), phenols and total organic carbon.
5. Other compounds: chlorides, chlorine and inorganic compounds, cyanides, fluorides, fluorine and inorganic compounds, hydrogen cyanide and PM10.

The EPER register is public and all information on industrial pollution is accessible as a relational database from the European Commission server [EPER, 2004] and from the Spanish Environmental Ministry. In addition to classifying the facilities in different industrial categories and ordering the pollutant substances in groups, the register has the geographical location of each industrial facilities, which is essential information for this thesis.

Consequently, the EPER enables us to study the relationship between industrial pollution and public health consequences in Europe by analyzing the influence of spatial distribution of emissions on geographic morbidity and mortality patterns. Thus, in the years after the publication of this register a few studies of this kind have been published [Garcia-Perez et al., 2009; Monge-Corella et al., 2008].

1.4. BRIEF SUMMARY

Throughout thesis we define a methodology to study the spatial distribution of health events and its relation to environmental factors, from large disease maps for a whole country to clustering analysis focused in small areas. We have divided this work in three separate sections.

The first section is called “Modelling of municipal mortality due to haematological neoplasias in Spain”. In this chapter we assess the performance of different methods for disease mapping based on Poisson models seeking to describe spatial patters in the distribution of the disease. In particular, three Bayesian hierarchical models for relative risk smoothing are analysed: the

Besag, York and Mollié model; a model based on zero-inflated Poisson (ZIP) distribution, which allowed a large number of event-free areas; and a mixture of distributions that enabled discontinuities (jumps in the pattern) to be modelled. The major characteristic of these methods is the use of the conditional autoregressive distribution (CAR) to include the spatial autocorrelation in the model to create an interpretable risk surface.

The second part of this thesis is entitled “Study of non-Hodgkin's lymphoma mortality associated with industrial pollution in Spain, using Poisson models”. In this second step we analyse the association between spatial disease patterns and the exposure to industrial pollution. We use three models of ecological regression to estimate the relative risk associated with the proximity to pollutant factories: Poisson Regression; mixed Poisson model with random provincial effect; and spatial autoregressive modelling (BYM model). To define the exposure variable we classify the municipalities either as exposed or non-exposed relative to the distance from the industrial facilities.

Finally, the third section is called ”Risk around putative focus in a multy-source scenario. Non-lineal regression models”. In this last step we study in depth the effect on cancer distribution of industrial air pollutants released from the different facilities sited within an urban area. We have applied an unique model that included all the factories under study and aggregated health data in small areas. Due to the lack of real exposure measures we approximate them by using the distance between the focus and the areas’ centroid. As above a Poisson regression is used as a basic model and is extended with a non-linear term that estimates the variation of the risk with the variation of the distance from the focus.

Each of these sections has been published or submitted as a paper in a internatinal journal. The publication’s details are listed below. They are also included at the end of this thesis.

- Modelling of municipal mortality due to haematological neoplasias in Spain.

Rebeca Ramis, Valentín Hernández-Barrera, Marina Pollán, Nuria Aragonés, Beatriz Pérez-Gómez, Gonzalo López-Abente.

Journal of Epidemiology and Community Health. 2007, 61:2.

- Study of non-Hodgkin's lymphoma mortality associated with industrial pollution in Spain, using Poisson models.

Rebeca Ramis, Enrique Vidal, Javier García-Pérez, Virginia Lope, Nuria Aragonés, Beatriz Pérez-Gómez, Marina Pollán and Gonzalo López-Abente.

BMC Public Health. 2009, 9:26

- Risk around putative focus in a multy-source scenario. Non-lineal regression models.

Rebeca Ramis, Peter Diggle, Koldo Cambra and Gonzalo López-Abente.

Submitted in Epidemiology. 2009

2. HYPOTHESIS AND OBJECTIVES

Hypothesis:

Residential proximity of population to one or several pollutant industrial facilities directly affects incidence and mortality risk for different malignant tumours.

Objectives:

I. To assess which methodology is more efficient to estimate risk surfaces (maps), using models for smoothing standard mortality ratios (SMR), seeking to identify spatial pattern of diseases and regions at higher risk for different tumoural locations.

II. To study the relation between cancer mortality risk at small area level and ecological exposure to pollutant emissions from industrial factories using the distance as surrogate of the real exposure.

III. To develop a methodology would enable to study risk associated to exposures from more than one pollutant focus in an unique statistical model, using he distances to the focuses as surrogate of the exposures.

    a. To obtain a global risk estimation of the effect for the residential proximity to industrial pollutant focuses in a spatial framework when data are aggregated in small areas.

    b. To detect the more influential focuses in the mortality pattern.

3. MODELLING OF MUNICIPAL MORTALITY DUE TO HAEMATOLOGICAL NEOPLASIAS IN SPAIN

3.1 INTRODUCTION

Spatial analysis of health events (spatial epidemiology) is a discipline that, despite still being in the development phase, can already claim its own domain in the field of health research [Elliott et al., 2000; Lawson, 1999]. Its ability to suggest and detect possible sources of heterogeneity which may account for spatial incidence and mortality patterns in different diseases, vest this tool with great interest in the sphere of epidemiology and public health.

Moreover, its potential is being reinforced by the ever increasing availability of geographically-indexed population mortality and incidence data, as well as ongoing advances in computation techniques and Geographic Information Systems. This is a situation that tends, in turn, to favour analysis of geographical distribution of health data of ever-finer resolution [Elliott et al., 2000], a category into which the so-called “small area studies” fall.

The main advantages of small area studies are: a) better interpretability than larger-scale studies; b) lower susceptibility to ecological biases; and c) greater capacity to detect local effects linked to environmental problems, such as industrial pollution of the environment [Richardson et al., 2004]. The disadvantages, on the other hand, are well known and determine the complexity of the analytical techniques. These drawbacks are: a) the data may be very disperse, with a large number of event-free areas; b) the data tend to evidence overdispersion, c) as a general rule, there is interdependence among observations, associated with the phenomenon of correlation between adjoining areas not taken into account by classic Poisson regression models and d) another important disadvantage is measurement of errors in both numerators and denominators.

The most widely used strategy for tackling the problems posed by small area analysis is to estimate the spatial distribution of risk by means of simulation based on Bayesian hierarchical models [Gilks et al., 1996]. This approach enables relative risk maps to be estimated for an entire country embracing a great number of areas, given that there are very few constraints over the model complexity and the number of terms included in the linear predictor. It does, however, add several problems to the four difficulties enumerated above. These are: a) computation times; and b) the use of homogeneous smoothing criteria for the whole country in cases where the components of spatial structure might vary between regions.

In view of the fact that there are several methodological alternatives for generating estimates

with Bayesian hierarchical models, comparison of the results yielded by such different approaches would probably help ascertain the true surface of risk. Apart from reporting the municipal pattern of distribution for haematological tumour mortality in Spain, this study sought to compare the goodness of fit of three different models, namely: a) the Besag York and Mollié model [Besag J et al., 1991]; b) a model based on zero-inflated Poisson (ZIP) distribution, which highlighted a large number of event-free areas [Lambert, 1992]; and c) a mixtures model that enables discontinuities (jumps in the pattern) to be modelled [Lawson and Clark, 2002]. Several authors have already tried to compare the performance of different spatial models [Fernandez and Green, 2002; Lawson and Clark, 2002], but neither of them have evaluated a ZIP model.

The application of these models to the study of haematological tumours is justified because the preliminary results of the umbrella project that encompasses this study (Atlas of Municipal Cancer Mortality in Spain) show that leukaemias, non-Hodgkin's lymphomas (NHL) and multiple myeloma have a similar distribution pattern, with a number of areas of increased risk, suggesting the possible implication of environmental factors in their aetiology.

This study seeks arguments for help to decide between the different methods of modelling [Fernandez and Green, 2002; Lawson and Clark, 2002] geographical patterns in situations that include an important number of small areas (viz.: a complete country), and, as one application of the methods, to know the municipal mortality distribution of haematological tumours in Spain. Additionally, in this case, we have used a accessible software tool [Spiegelhalter et al., 2002; Spiegelhalter et al., 2003], showing that these models can be easily applied in epidemiology and Public Health.

3.2 MATERIALS AND METHODS

Cases were sourced from individual entries recording deaths due to leukaemias, non-Hodgkin's lymphomas (NHL) and multiple myeloma (ICD-9 codes 200, 202, 203 and 204-208), registered at a municipal level nation-wide for the period 1989-1998. These data were supplied by the National Statistics Institute (*Instituto Nacional de Estadística*) for the production of a municipal cancer mortality atlas, of which these results form part.

In Figure 3.1 we show a political map of Spain.

Figure 3.1. Political map of Spain.

The municipal populations, broken down by age group (18 groups) and sex, were drawn from the 1991 census and 1996 electoral roll. These years correspond to the mid-point of the two quinquennia that comprise the study period (1989-1993 and 1994-1998). Person-years for each such five-year period were then obtained by multiplying the above populations by 5.

Standard mortality ratios (SMR) were calculated as the ratio between observed and expected deaths. To calculate expected cases, the overall Spanish mortality rates for the above two 5-year periods were multiplied by each town's person-years by age group, sex and quinquennium.

To plot the maps, smoothed municipal relative risks (RR) were then calculated, using conditional autoregressive (CAR) models. These models were based on fitting spatial Poisson models with two random-effects terms that took the following into account: a) municipal contiguity (spatial term); and b) municipal heterogeneity (Figure 3.2). Three different models were fitted.

The Besag, York and Mollié (BYM) Model

This model was introduced by Clayton and Kaldor [Clayton and Kaldor, 1987] developed by Besag, York and Mollié [Besag J et al., 1991] and subsequently applied in the field of ecological studies [Clayton et al., 1993]. It is the most widely used model and is formulated as follows:

$$O_i \sim Po(E_i \lambda_i)$$
$$\log(\lambda_i) = \alpha + h_i + b_i$$

$$h_i \sim Normal(\mu, \tau_h)$$
$$b_i \sim Car.Normal(\eta_i, \tau_b)$$

$$\tau_h \sim Gamma(\alpha, \beta)$$
$$\tau_b \sim Gamma(\gamma, \delta)$$

where: λ_i is the relative risk in area i.

O_i is the number of deaths in area i.

E_i are the expected cases.

h_i is the municipal heterogeneity term from a Normal distribution.

b_i is the spatial term from a Car.Normal distribution.

τ_h is the hyperparameter of the Normal distribution.

τ_b is the hyperparameter of the Car.Normal distribution.

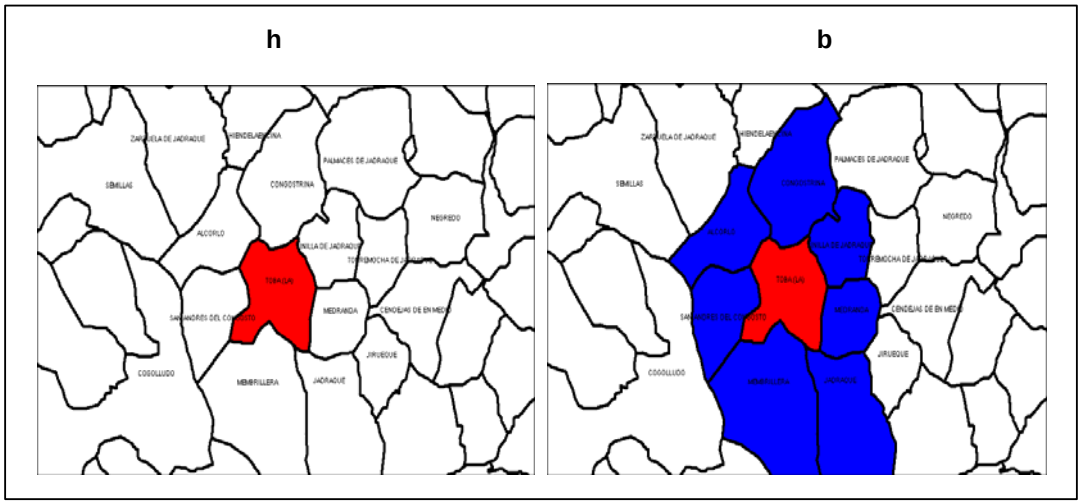


Figure 3.2. Random effects: h= municipal heterogeneity; b = municipal contiguity (spatial term).

The Lawson Mixtures Model

The second model was proposed by Lawson [Lawson, 2005]. Basing himself on the BYM model, Lawson specifies a model that assumes the breakdown of relative risk into three components, one addressing heterogeneity (h), and the other two forming a mixture which addresses the different behaviours of the spatial correlation (b and d), with b being the spatial correlation component and d the component that models the jumps in distribution. Lastly, p_i is

the proportion of random effect b_i for area I, and (1-p_i) the proportion of d_i. This mixture in the spatial term is included as treatment for any possible discontinuities displayed by distribution of the data.

$$O_i \sim Po(E_i \lambda_i)$$
$$\log(\lambda_i) = \alpha + h_i + (p_i * b_i + (1 - p_i) * d_i)$$

The zero-inflated Poisson model

Lastly, we used the zero-inflated Poisson (ZIP) model [Hall, 2000;Lambert, 1992]. This model is constructed as a mixture of Poisson distributions, one of which has the parameter λ=0 to include the high proportion of zeros possessed by these types of distributions. For study purposes, we used a ZIP model proposed by Durham *et al* [Durham et al., 2004], in which the Poisson distribution with λ>0 is taken directly from the BYM model.

$$O_i \sim (p_i * Po(E_i \lambda_i) + (1 - p_i) * Po(0))$$
$$\log(\lambda_i) = \alpha + h_i + b_i$$

The models were fitted using Markov chain Monte Carlo simulation methods with non-informative priors [Gilks et al., 1996]. Convergence of the simulations was verified using the BOA (Bayesian Output Analysis) R programme library [Smith BJ, 2001]. In view of the great number of parameters of the respective models, the convergence analysis was performed on a randomly selected sample of 10 towns and cities, taking strata defined by municipal size.

The Deviance Information Criterion (DIC) [Richardson et al., 2004] was used as the criterion for model selection. This criterion entails Bayesian generalisation of the Akaike Information Criterion (AIC) and approximately describes the expected posterior loss when a particular model is adopted, i.e., it is the description of the expected divergence of the model vis-à-vis the real data. The DIC is the parameter used in Bayesian models to assess the goodness of fit.

Posterior distributions of relative risk were obtained using WinBugs [Spiegelhalter et al., 1996; Spiegelhalter et al., 2003]. The criterion of contiguity used was adjacency of municipal boundaries [Ferrandiz et al., 2002]. Convergence of estimators was achieved before 100,000 iterations. For the maps shown, a burn-in (iterations discarded to ensure convergence) of 300,000 iterations was performed and the posterior distribution was derived with 5,000 iterations.

The results of the models were included in a Geographic Information System to plot municipal maps that depicted smoothed RR estimates and the distribution of the posterior probability that

RR>1. Insofar as this indicator is concerned, we applied Richardson's criterion [Richardson et al., 2004], which recommends that probabilities in excess of 0.8 be deemed significant.

3.3 RESULTS

A total of 55430 deaths attributable to the haematological tumours covered by this analysis were registered from 1989 to 1998 in Spain. There are 8077 towns, in 3061 towns of them, no death due to this cause was registered. Using these data, and conventional computers, we were able to compile and obtain the posterior distribution of relative risk on the basis of a single spatial model, including all of Spain's towns and the 46398 adjacencies existing between them, for the three strategies outlined above. Table 3.1 displays a number of descriptive statistics for the population and disease data. The total population was under 40 million, and leukemia mortality was two times higher than that of multiple myeloma and quite superior of NHL mortality. The mean number of cases per area, for haematological tumours, is 6.9 and the median is 1.

	Total	Mean	Median	Standard deviation	Min.	Max.	No. (%) of areas with zero counts
Population	38872268	7812.7	600	44081.1	5	3010492	0 (0)
Observed	55430	6.86	1	70.96	0	4774	3061 (37.8)
Expected	55744.65	6.9	1.3	65.17	0	4514.3	0
Observed NHL	18363	2.27	0	25.02	0	1654	5008 (62.0)
Expected NHL	18471.6	2.28	0.422	21.67	0	1501.7	0
Observed Myeloma	11634	1.44	0	14.97	0	1039	5400 (66.8)
Expected Myeloma	11707.2	1.45	0.29	13.9	0	966.6	0
Observed Leukaemias	25433	3.15	0	31.13	0	2081	4215 (52.2)
Expected Leukaemias	25565.9	3.16	0.59	29.6	0	2045.9	0

Table 3.1 Summaries of population and haematological tumours mortality in the 8077 Spanish towns.

Results of comparison of models

The DIC values listed in Table 3.2 show that Lawson's model furnished the lowest values and was thus the one that best fitted our data.

Model	Expected deviance. E(D)	Deviance evaluated with respect to expected values. D(E)	Number of effective model parameters. (Pd)	DIC
BYM	8114	7522	592	8706
Lawson	8058	7436	622	8680
ZIP	8206	7698	508	8714

Table 3.2. Deviance Information Criterion (DIC) for the three models.

The correlation coefficients between the estimated relative risks yielded by the three models were very high, with the correlation between the BYM and Lawson models being slightly higher than that for the other two (BYM-Lawson r=0.964; BYM-ZIP r=0.938; Lawson-ZIP r=0.932).

Figure 3.3 shows the combined representations of the cloud of points corresponding to the results for each pair of models. In all three cases, the data can be seen to be aligned along the main diagonal, indicating equality of results yielded by the two models for any given town. As the correlation coefficients confirmed, the BYM and Lawson models were the ones to yield the most similar results, since the data are more closely superimposed along the diagonal than in the other two cases, in which the clouds of points are wider.

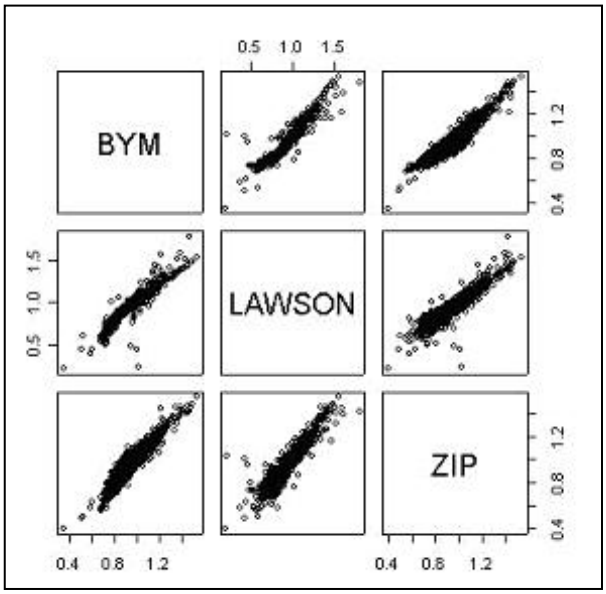


Figure 3.3. Clouds of points representing the RR distributions of the respective models, taken in pairs.

Figures 3.4 a), b) and c) plot the smoothed RR maps for the three models. Comparison of the maps shows that the spatial distribution of relative risks was practically identical in all three cases. When it came to detecting towns that registered high and medium risks, the behaviour of the three models was identical.

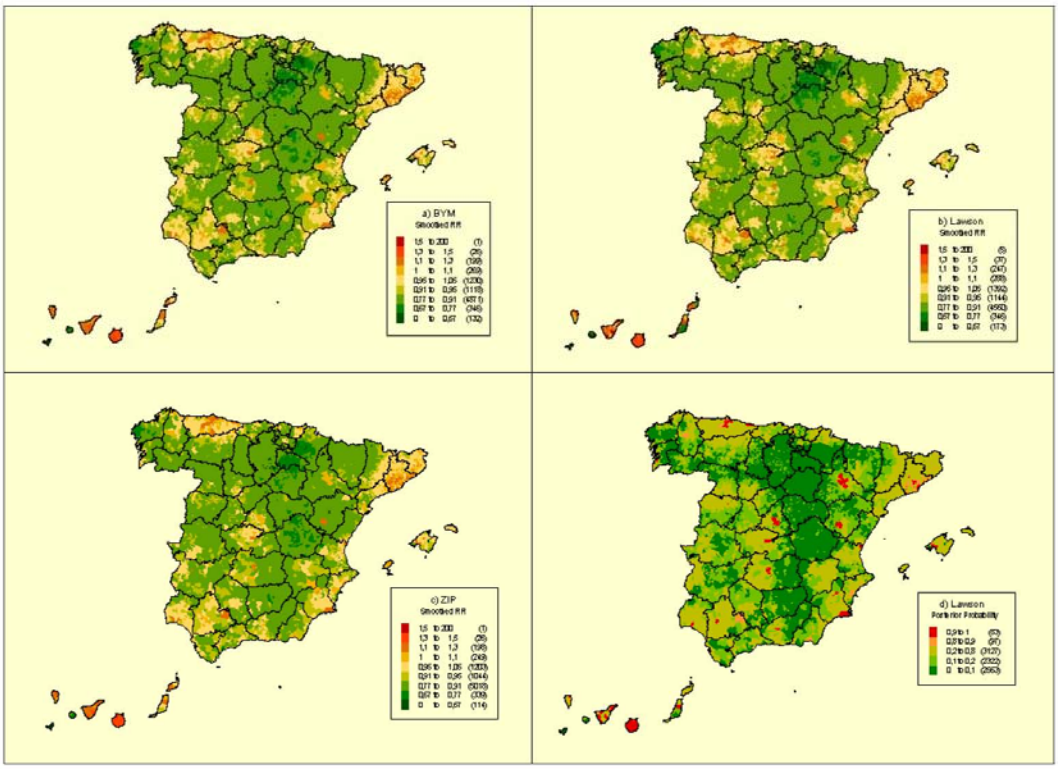


Figure 3.4. Municipal distribution of haematological tumours mortality in Spain. Distribution pattern of smoothed relative risk (RR), according to: a) BYM model; b) Lawson model; c) Zip model; d) Posterior probability of RR being greater than 1. Haematological tumours mortality, Spain 1989-1998.

The differences emerged when it came to allocating low relative risks, i.e., those below 0.77. The ZIP model registered most towns with relative risks of this type (804). In contrast, the traditional BYM model was the one that had the smallest number of towns with low relative risks (316). The Lawson model registered the greatest number of towns with extreme relative risks of both types, i.e., very low, below 0.67 (164), and very high, above 1.5 (6).

Cases in which discordances appeared between the results obtained for smoothed relative risks and those obtained for SMR were targeted for specific analysis. These discordances can go two different directions: on the one hand, there is the case where the relative risk is high versus an SMR of less than 1, viz., when the relative risk has been overestimated, and, on the other hand, there is the contrary case, where the smoothed relative risk is less than 1 versus an SMR that is greater than unity, viz., when the relative risk has been underestimated. The appearance of such cases may be attributable to the influence of the spatial component on the smoothing process.

Examination of these two events in the results yielded by the three models showed that the best model in terms of appearance of cases with RR≥1 and SMR<1 was the BYM model, with the

lowest percentage of cases. In contrast, the Lawson and ZIP models registered a lower number of locations with RR<1and SMR≥1, with both displaying the same percentage.

The differences in the number of locations with opposite association between RR and SMR yielded by the three models were not sufficiently great to allow this criterion to be used for comparative purposes. For each model percentages of small areas with RR≥1 and statistically significant SMR<1, α=0.95, were around of 3.7%, and those with RR<1 and statistically significant SMR≥1 around 5%.

Spatial distribution of haematological tumour mortality

From maps a) b) and c) in Figure 3.3, which depict the smoothed relative risks estimated by the three models, it will be clearly seen that the areas of highest risk were Barcelona Province and the Canary Islands (the islands of Gran Canaria, Tenerife and La Palma in particular), though Asturias also had a number of towns with high relative risks.

In terms of the spatial location of the areas with opposite association, the results showed that most of the cases with RR≥1 and SMR<1 were concentrated in the Canary Islands and Barcelona, whereas the areas with RR<1and SMR≥1 were mostly in the Castile-León Region.

Map d) in Figure 3.3 depicts the distribution of posterior probability under Lawson's model, in as much as this was the model that furnished the lowest DIC. According to this model, there were 199 towns with a probability of greater than 0.8 of their estimator of real risk being higher than unity. Their geographical distribution displayed two clearly differentiated patterns, namely: one with a majority of towns in Barcelona Province (97 towns) and the Canary Islands (49 towns), and the remainder divided up among a series of provinces, e.g., Madrid, Seville, Zaragoza, Salamanca, Toledo and Huelva; and the other, with major cities such as Gijón, Vigo and Cartagena.

By combining the two patterns, the following emerges: on the one hand, there are the provinces in which the towns having the highest relative risks are concentrated, namely, Barcelona and the two Canary Island Provinces; and on the other, there are important towns and cities that do not have such high relative risks but, in contrast, do have an almost certain likelihood of such relative risks being greater than 1. Table 2.3 shows the towns -excluding those in the Canary Island Autonomous Region- which registered probabilities in excess of 0.9.

Provincia	Town	Obs	Exp	SMR	RR (LAWSON)	RR (BYM)	RR (ZIP)	p(RR>1) (LAWSON)
Baleares	PALMA DE MALLORCA	463	402.9	1.149	1.130	1.129	1.112	0.994
Barcelona	BADALONA	314	253.9	1.237	1.210	1.219	1.213	0.999
	BARCELONA	3186	2690.9	1.184	1.180	1.182	1.179	1.000
	CALELLA	25	17.8	1.408	1.267	1.249	1.294	0.950
	LLAGOSTA (LA)	18	11.7	1.533	1.202	1.201	1.192	0.913
	MANRESA	128	110.2	1.162	1.112	1.124	1.127	0.940
	MATARO	151	130.7	1.155	1.136	1.143	1.144	0.960
	MOLINS DE REI	33	24.7	1.336	1.156	1.139	1.141	0.915
	MONTCADA I REIXAC	33	32.6	1.012	1.093	1.074	1.117	0.902
	PIERA	14	10.1	1.383	1.156	1.169	1.136	0.921
	PINEDA DE MAR	29	20.1	1.439	1.255	1.243	1.251	0.957
	PREMIA DE MAR	44	26.9	1.632	1.374	1.331	1.254	0.994
	RUBI	65	55.8	1.164	1.128	1.133	1.139	0.924
	SABADELL	313	254.0	1.232	1.196	1.207	1.194	0.999
	S ADRIA DE BESOS	51	39.6	1.288	1.216	1.214	1.24	0.970
	VILASSAR DE MAR	22	16.8	1.310	1.189	1.184	1.208	0.924
	S PERE DE RIUDEBITLLES	8	3.5	2.269	1.278	1.195	1.264	0.919
	S COLOMA DE GRAMENET	172	147.7	1.165	1.153	1.155	1.174	0.989
	TERRASSA	255	220.9	1.154	1.133	1.143	1.134	0.992
	VILAFRANCA DEL PENEDES	64	39.6	1.614	1.374	1.342	1.353	0.995
	VILANOVA I LA GELTRU	89	66.0	1.347	1.258	1.251	1.251	0.990
Cuidad Real	CIUDAD REAL	95	71.8	1.322	1.146	1.139	1.085	0.937
A Coruña	CORUÑA (A) (CORUNNA)	399	337.6	1.182	1.157	1.156	1.161	0.997
Granada	GRANADA	359	319.7	1.123	1.083	1.091	1.071	0.956
Guipúzcoa	SAN SEBASTIAN	316	267.2	1.182	1.141	1.149	1.143	0.995
Huelva	HUELVA	185	157.6	1.174	1.124	1.126	1.121	0.944
Madrid	MADRID	4774	4514.3	1.058	1.055	1.056	1.054	0.999
Murcia	CARTAGENA	240	210.7	1.139	1.107	1.112	1.112	0.947
	UNION (LA)	23	14.9	1.534	1.333	1.242	1.333	0.924
Navarra	ALSASUA	22	9.7	2.263	1.188	1.198	1.17	0.921
	PAMPLONA	330	256.8	1.285	1.179	1.204	1.171	0.999
Asturias	CORVERA DE ASTURIAS	30	20.6	1.458	1.229	1.211	1.211	0.951
	GIJON	478	400.6	1.193	1.177	1.179	1.186	1.000
	LLANERA	24	17.1	1.401	1.196	1.177	1.191	0.927
	OVIEDO	340	296.6	1.146	1.125	1.132	1.13	0.991
Pontevedra	VIGO	393	347.8	1.130	1.101	1.103	1.105	0.972
Salamanca	SALAMANCA	275	240.3	1.144	1.100	1.101	1.088	0.951
Seville	S JUAN DE AZNALFARACHE	35	23.8	1.470	1.232	1.187	1.208	0.912
	SEVILLE	936	850.7	1.100	1.090	1.094	1.088	0.995
Toledo	TOLEDO	117	82.9	1.410	1.266	1.228	1.158	0.993
Valencia	CANET D'EN BERENGUER	8	2.7	2.902	1.588	1.212	1.436	0.917
	SAGUNTO	106	82.0	1.292	1.150	1.156	1.112	0.925
Vizcaya	BARAKALDO	173	145.2	1.191	1.104	1.125	1.107	0.920
Zaragoza	ZARAGOZA	953	868.4	1.097	1.065	1.081	1.062	0.972

Table 3.3. *Haematological tumours mortality in Spain. Towns having a posterior probability (PP) superior to 0.9 of having an RR greater than 1 (p(RR>1)). Towns, excluding those in the Canary Island Autonomous Region, listed in order of province. Observed and expected deaths (Obs, Exp); estimated relative risk (RR); standard mortality ratio (SMR).*

3.4 DISCUSSION

The geographical patterns, performance and conclusions derived from the results of the three models discussed in this study are very similar. Even though conclusions of previous studies suggests the mixture models are more appropriate modelling relative risk pattern and do no over-smooth maps [Fernandez and Green, 2002; Lawson and Clark, 2002]. Haematological tumours register a geographical pattern that might possibly be partially explained by environmental determinants, since many of the highest-risk towns are located in heavily industrialised areas. The distribution pattern supports the hypothesis that differences in lifestyle and urban air pollution may determine the urban mortality pattern of these tumours [Lopez-Abente et al., 2001], and this conclusion can be reached from any of the models.

Previous studies on provincial lymphohaematopoietic tumour mortality patterns have reported results with negligible geographical variability and without any defined pattern, save for the excess mortality observed for the Canary Islands [Doll, 1991]. The use of towns as a level of study allows patterns to emerge that would otherwise remain hidden by virtue of provincial averaging, this advantage have been highlighted in previous studies of small areas in Spain about different causes of mortality which have not included haematological tumours [Benach et al., 2004].

Provinces that display the highest number of towns with excess mortality are Las Palmas, Tenerife, Barcelona, Asturias and Girona. Equally important are municipal areas, many in the form of isolated areas, which correspond to major towns and cities, and the results for which are shown in Table 3.3.

The case of the Canary Islands calls for special mention. This excess mortality was already visible in earlier studies, though the origin of this pattern remained unidentified. The pattern is repeated for both lymphomas and multiple myeloma, and to a lesser extent, for leukaemias, tumours that register a higher mortality in Las Palmas than in Tenerife [Doll, 1991]. At various times, the effect of the proportion of the foreign population on mortality patterns in the Canary Islands (reliability of census data and case allocation) has been discussed. One of the problems detected is the difference between the population census figures and the municipal electoral roll, a difference higher than those found in other Autonomous Regions [Godenau and Arteaga, 2004]. In our study, both sources were used as denominators. Were the problem to lie in the denominators, excess mortality, and probably excess incidence, would be observed in all the causes studied. Nevertheless, according to the information drawn from the Canary Island cancer registry, reported NHL incidence in both sexes is higher there than for the other registries in Spain, with no such excesses being in evidence for the remaining tumour sites. It would therefore be of great interest if an in-depth study were to be conducted into the

determinants of these tumours in the Canary Island Autonomous Region, since this difference in mortality would not appear to be solely attributable to census-related or demographic artefacts.

With respect to the results yielded by the different study models, the geographical pattern that emerges is very similar. Models that seek to remedy the excess of zeros, display a pattern that is almost identical to the classic BYM model, this suggest ZIP model does not distinguish between areas with no cases and areas with cases. Although goodness-of-fit criteria indicate that the model proposed by Lawson is that which best fits our data, the choice of one or another probably has scant practical consequences. In regard to the use of ZIP model, it could be unrealistic to think of null risk areas, however this distribution has been used in small areas with rare diseases in previous studies [Congdon, 2001; Ugarte et al., 2004].

In general, Bayesian models for plotting disease maps are conservative, in that they have a low sensitivity for detecting areas with moderate increases in risk, but, in contrast, have a pronounced specificity for detecting areas of high risk [Richardson et al., 2004]. That is to say, when the smoothing process yields high relative risk values, this is because the relevant SMR is high. Environmental risks are low, however, and as a result these methods have a limited power for detecting them [Richardson et al., 2004].

The results that point to large cities could simply be attributable to the greater populations to be found there. In other words, statistically significant excesses are detected in places where the comparison has sufficient power. Yet, this does not happen with other tumours or groups of tumours. The municipal pattern for haematological tumours is thus very specific and is different to that observed for other tumour sites.

With the exception of ionizing radiations and benzene in myeloid leukaemias, the aetiology of lymphohaematopoietic tumours is little known. Nevertheless, suspicions surrounding the multiple risk factors present in the study, namely, ionizing and non-ionizing radiations and exposure to different chemical substances (petroleum by-products, hydrocarbons, pesticides, solvents) [Schotenfeld and Fraumen, 1996], are shared vis-à-vis haematological tumours (leukaemias, NHL and myeloma). There has been some evidence that leukaemia and lymphomas occur in neighbourhoods that contained industrial sites [Benedetti et al., 2001; Parodi et al., 2003]. The pattern of municipal distribution linked to large cities suggests that factors associated with the process of urbanisation, such as air and/or industrial pollution, may be implicated in the aetiology of such processes.

The geographical pattern is determined by deaths in adults, who account for over 85% of cases, since lethality among children is low. Indeed, on examining municipal leukaemia mortality distribution in the under-25 age group in Spain, no geographical pattern whatsoever is in evidence (data not shown). Although infectious aetiology may be present in haematological

tumours, it seems highly unlikely that it would determine the pattern plotted for all age groups. With respect to the mechanisms implicated in the infectious aetiology of haematological tumours in childhood, the following three hypotheses have been advanced: exposure in the uterus or in the period immediately preceding birth; delayed exposure to common infections after the first year of life [Greaves, 1997]; and unusual population mixing [Kinlen, 1996; Mcnally and Eden, 2004]. The population-mixing hypothesis was initially formulated in terms of situations of immigration to isolated, sparsely populated areas [Kinlen et al., 1995]. The influence of migratory phenomena on leukaemia mortality has been studied and it has been suggested that rural-urban migration may be implicated in leukaemia mortality in Italy and Greece [Kinlen and Petridou, 1995]. In the period 1960-1970, important migratory phenomena of this type took place in Spain, with Catalonia being a net recipient of immigration from many areas, thereby rendering the population-mixing hypothesis plausible. Internal migratory flows were linked to the intensification of the industrialisation process and a decline in Spain's rural population [Capel, 1967]. As a consequence, population mixing and exposure to environmental and industrial pollution are very closely related phenomena in this country.

The different Bayesian models used in this study furnished some very similar results. The high frequency of areas without cases would not seem to pose a serious difficulty to fitting these models, at least in this group of causes. It would be advisable to ascertain whether this conclusion can be generalised and, by extension, whether the above observations are therefore applicable to other tumour sites.

4. STUDY OF NON-HODGKIN'S LYMPHOMA MORTALITY ASSOCIATED WITH INDUSTRIAL POLLUTION IN SPAIN, USING POISSON MODELS

4.1 INTRODUCTION

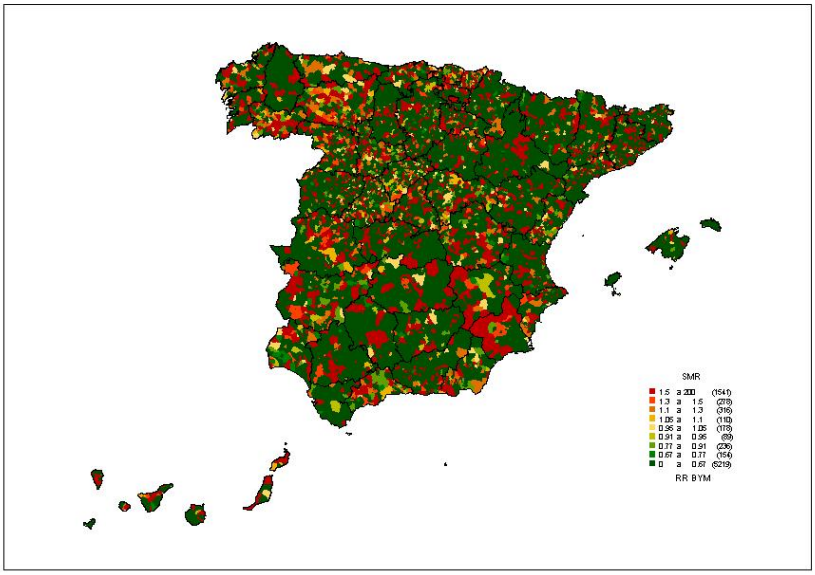
In general, industrial activities constantly release a great amount of toxic substances into the environment. At present, evidence regarding the health risk posed by residing near pollutant industries and, by extension, being exposed to their emissions, is limited. Non-Hodgkin's lymphomas (NHLs) constitute one of the tumour sites that has been linked in the literature to proximity to industrial areas[Johnson et al., 2003; Sans et al., 1995; Sharp et al., 1996]. During the second half of the 20th century, NHLs witnessed a marked increase world-wide, in terms of both incidence and mortality [Muller et al., 2005], which means that they form part of the group of so-called emerging tumours. This same increase has also been observed in Spain [Pollan et al., 1998].

Although this tumour's aetiology is rather unknown, its relationship with the immune system has generated theories about its increase being connected with the HIV epidemic [Eltom et al., 2002], though the inclusion of Highly Active Antiretroviral Treatments (HAARTs) does not appear to have affected the rising trend in NHLs [Fisher and Fisher, 2004].

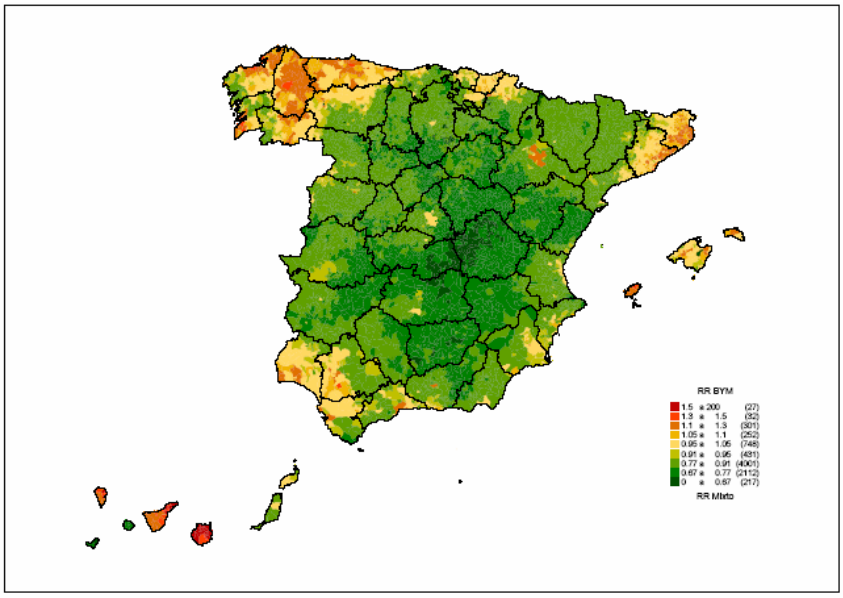
From the environmental point of view, there are some studies that link lymphomas to exposure to substances such as agricultural chemicals [Fisher and Fisher, 2004], and dioxins released by incinerators [Floret et al., 2003]. Mention should also be made of the fact that a number of occupational exposure studies have reported higher NHL incidence and mortality among workers exposed to industrial solvents [Blair et al., 1998; Burnett et al., 1999; Eltom et al., 2002]. According to Spanish mortality data, NHLs are particularly frequent in the Canary Islands [Lopez-Abente et al., 2006b], while on the mainland, higher NHL mortality is observed in Asturias, the Basque Country and Catalonia, three of Spain's most industrialised regions [Garcia-Perez et al., 2007]. As follow two maps showing risk estimations are presented, Map 4.1 shows the SMR and Map 4.2 the smoothed relative risk, both at municipal level.

At the beginning of this decade, specific legislation was passed, both in Spain and in Europe, governing the control of pollutant emissions. This initiative included the setting-up of the public European Pollutant Emission Register (EPER) [EPER, 2004], which records pollutant emissions reported by industries that admit to exceeding pre-established pollution thresholds included in the Decision (2000/479/EC) of the European Commission [EPER, 2004; Commission Of The

European Communities, 2000]. This database furnishes information on the location of industrial foci, 50 specific pollutants, and a long list of industrial processes that release emissions to air and water, thereby offering a wide range of possibilities for using such information to study possible associations between risk of incidence or mortality due to different causes and proximity to sources of industrial emissions.

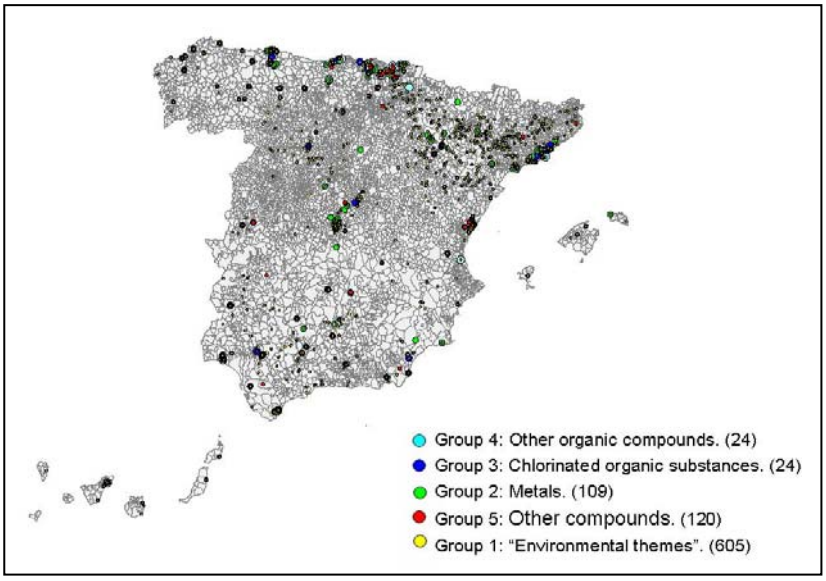


Map 4.1. Standard Mortality Ratio of LNH at municipal level.



Map 4.2. Smoothed relative risk of LNH at municipal level.

The following map (Map 4.3) shows the location of the industrial facilities registered in EPER sited in Spain.



Map 4.3. Location of the industrial facilities registered in EPER sited in Spain.

Quality information on industrial pollution, as part of the overall environmental pollution to which the population is exposed, is a critical point when it comes to evaluating its effects. Due to the dearth of such information, a recourse widely used in scientific literature is to estimate exposure based on the distance to the polluting source [Elliott et al., 2000; Johnson et al., 2003; Sans et al., 1995; Sharp et al., 1996].

This study sought to explore the relationship between municipal NHL mortality in Spain and distance to EPER-registered industries, as an indirect measure of exposure to industrial pollution, using a series of Poisson-regression-based mathematical models for the purpose.

4.2 MATERIALS AND METHODS

4.2.1 Data

Observed NHL cases, broken down by death, sex and age group (18 groups), were drawn from entries of individual deaths recorded by the National Statistics Institute (*Instituto Nacional de Estadística - INE*) with ICD9 for the period 1994-2003, in respect of the 8073 Spanish towns registered in the 2001 census.

Municipal populations, likewise broken down by sex and age group, were used to calculate expected cases These populations were obtained from the 1996 electoral roll and the 2001 census, which respectively correspond to the mid-point of the two five-year periods included in the study (1994-1998 and 1999-2003). Person-years for each quinquennium were calculated by multiplying the respective populations by 5. Expected cases resulted from multiplying the mortality rates for Spain as a whole, for each sex, age group and quinquennium, by the person-years of each town, broken down by the same strata.

Industrial pollution data were obtained from the EPER figures published in 2004, which include industries that voluntarily reported pollutant emissions exceeding a designated reporting threshold for 50 toxic substances. This database contains information identifying the industrial activity, the substances emitted, and the installation's geographical location by reference to its co-ordinates, previously validated and corrected for poor geocoding [Garcia-Perez et al., 2008]. The emission data correspond to information reported by industries for 2001. The 452 industries that reported releases to air to the EPER were grouped by industrial sector (Figure 4.1). In this study, farms were excluded from the analysis.

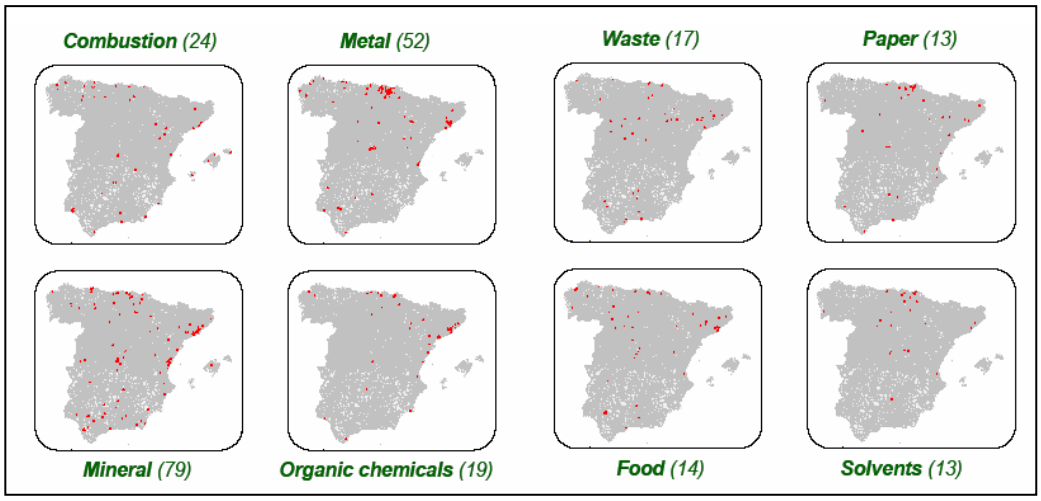


Figure 4.1. Location of the industrial facilities registered in EPER sited in Spain for the studied sectors.

For the construction of the exposure variable and calculation of RRs on the basis of spatial autocorrelation models, maps of municipal boundaries and co-ordinates of the centroids of population centres were used. This is the only available geographical information for each municipality, boundaries of the township and its centroid. We do not know the real limits of the inhabited areas; consequently, we assume that the whole population of each town lives in its centroid.

Distance to the emission source was used as an estimator of exposure to pollutant substances released by industries, [Diggle, 1990; Elliott et al., 2000; Garcia-Perez et al., 2008; Lawson A, 2001; Selvin et al., 1992]. Using this criterion, exposed populations were defined as any

population corresponding to a town that had EPER-registered industries situated within a radius of 2000, 1500 and 1000 metres, in a circle drawn with the municipal centroid as its centre. For study purposes, we only considered industrial groups that had a minimum of 10 towns within the 2-, 1.5- and 1-kilometre areas respectively. On the basis of this definition, an exposure variable was constructed for each industrial sector that showed more than 10 towns having industries within the predefined radius, for the total Spanish population. This variable was defined as a factor with three possible levels, which distinguished among: towns that had no industrial installation within the designated radius (unexposed); towns that had installations corresponding to the industrial group studied within the designated radius (exposed); and towns that had some other type of industrial installation. Finally, with the aim of controlling possible confounding effects, the following socio-demographic variables were included in the analysis: percentage of illiteracy; percentage of unemployed persons; size of household (persons per home), obtained from the 1991 census; and mean income level [Banco Español de Credito, 1993].

4.2.2 Models

Firstly, Poisson regression models were fitted, using the following formula:

$$O_i \sim Po\left(\mu_i = E_i \lambda_i\right)$$
$$\log\left(\lambda_i\right) = \sum_j \beta_j x_i \Rightarrow \log\left(\mu_i\right) = \log\left(E_i\right) + \sum_j \beta_j x_i$$

where: λ_i is the relative risk in area i; O_i is the number of deaths in area i; E_i are the expected cases; and x_i are the socio-demographic variables.

This risk estimation method takes no account of any possible spatial correlation in data drawn from contiguous areas, such as towns in a given region or country. To take such correlation into account, we therefore considered a mixed Poisson-regression-based model that included a provincial random effect:

$$O_i \sim Po\left(\mu_i = E_i \lambda_i\right)$$
$$\log\left(\lambda_i\right) = \sum_j \beta_j x_i + p_i \Rightarrow \log\left(\mu_i\right) = \log\left(E_i\right) + \sum_j \beta_j x_i + p_i$$

where p_i is the provincial random term.

Lastly, a Bayesian hierarchical model was used [Clayton et al., 1993; Ramis et al., 2007; Wakefield, 2007]. These types of models, which fall within the category of the so-called conditional autoregressive models (CAR), include two random-effects terms that take the

following into account: a) municipal contiguity (spatial term); and b) municipal heterogeneity. In our case, we used the model proposed by Besag, York and Molliè (BYM) [Besag J et al., 1991]:

$$O_i \sim Po(\mu_i = E_i \lambda_i)$$
$$\log(\lambda_i) = \sum_j \beta_j x_i + h_i + b_i \Rightarrow \log(\mu_i) = \log(E_i) + \sum_j \beta_j x_i + h_i + b_i$$

$$h_i \sim Normal(\mu, \tau_h)$$
$$b_i \sim Car.Normal(\eta_i, \tau_b)$$

$$\tau_h \sim Gamma(\alpha, \beta)$$
$$\tau_b \sim Gamma(\gamma, \delta)$$

where: h_i is the term of municipal heterogeneity; and b_i is the spatial term.

With each of the three methodologies used, a multivariate model was fitted including the distance to the locus of each type of industry, individually, and the remaining possible confounders mentioned above. In the Poisson and mixed models, the estimates were calculated using the *glm* and *glmmPQL* functions of the R software programme [R Development Core Team, 2005]. Spatial autocorrelation models were fitted with the aid of the WinBUGS Bayesian estimation programme [Spiegelhalter et al., 1996]. To obtain results from the spatial model, a burn-in period of 150,000 iterations was performed, which guaranteed convergence of the model parameters, and the posterior distribution was derived with a further 25,000 iterations. Approximately 15 hours on a conventional computer was required to complete this process.

4.3 RESULTS

From 1994 and 2003 there were 22,262 NHL-related deaths in Spain, accounting for 2.7% of all cancer deaths. In 4758 towns (59%) there was no death due to this cause.

The industrial sectors considered, together with the number of towns respectively located less than 2000, 1500 and 1000 metres away, are shown in Table 4.1. This table also includes the population belonging to towns deemed to be exposed within a radius of 2000 metres.

	Industrial sector	*Total No. factories*	No. towns with installations at x metres 2000	1500	1000	Population exposed at <2000 m
1	Combustion installations > 50 MW	*59*	24	13	6	1,034,398
2	Mineral oil and gas refineries	*10*	3	3	1	9,520
3	Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	*68*	52	35	22	113,953
4	Installations for the production of cement clinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	*55*	79	49	27	665,785
5	Basic organic chemicals	*37*	18	8	3	284,852
6	Basic inorganic chemicals or fertilisers	*25*	5	3	1	169,963
7	Pharmaceutical products	*8*	4	2	1	216,590
8	Installations for the disposal or recovery of hazardous waste (>10t/d) or municipal waste (>3t/h)	*8*	4	0	0	62,853
9	Installations for the disposal of nonhazardous waste (>50t/d) and landfills (>10t/d)	*43*	15	4	1	146,811
10	Industrial plants for pulp from timber or other fibrous materials and paper or board production (>20t/d)	*18*	13	6	3	725,225
11	Slaughterhouses (>50t/d), plants for the production of milk (>200t/d), other animal raw materials (>75t/d) or vegetable raw materials (>300t/d)	*12*	14	9	4	208,227
12	Installations for surface treatment or products using organic solvents (>200t/y)	*12*	13	6	2	418,749

Table 4.1. Industrial sectors. Number of towns with installations at distances of 2000, 1500, 1000 and 500 metres from the municipal centroid, by type of industry. Population exposed to emissions from each industrial sector at a distance of 2000 metres.

Shown in Table 4.2 and represented in Figure 4.2 are the RRs associated with each of the industrial sectors studied, for the respective radii of 2000, 1500 and 1000 metres. This table also includes the confidence (models 1 and 2) and credibility intervals (model 3) of the estimates. From these estimates, it will be seen that in towns situated within a radius of 2000 metres of paper, pulp and board installations, exposure to pollutant emissions from this industry was associated with excess NHL mortality. This excess risk was statistically significant in all 3 models, namely: 1.163 (95% CI: 1.06,1.27) for Poisson regression; 1.24 (95% CI: 1.09,1.42) for the mixed model; and 1.21 (95% CI: 1.01,1.45) for the spatial BYM model. Analysing the RRs associated with the variable of exposure to the paper, pulp and board industry in Table 3.2, it will be seen that the highest RR estimate was yielded by the spatial mixed model, followed by the BYM model and Poisson regression, in that order.

		Poisson			Mixed			BYM		
Radius	Industrial Sector	RR	Lower limit	Upper limit	RR	Lower limit	Upper limit	RR	2.50%	97.50 %
2000 m	Combustion installations > 50 MW	1.06	0.94	1.19	1.08	0.95	1.24	1.10	0.93	1.29
	Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	0.92	0.85	1.00	0.96	0.87	1.06	0.97	0.85	1.10
	Installations for the production of cement clinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	0.90	0.81	1.01	0.96	0.86	1.08	0.96	0.84	1.10
	Basic organic chemicals	1.01	0.87	1.17	1.02	0.88	1.20	1.08	0.89	1.30
	Installations for the disposal of non-hazardous waste (>50t/d) and landfills (>10t/d)	0.89	0.71	1.11	0.96	0.76	1.21	0.92	0.70	1.19
	Industrial plants for pulp from timber or other fibrous materials and paper or board production (>20t/d)	**1.16**	**1.06**	**1.27**	**1.24**	**1.09**	**1.42**	**1.21**	**1.01**	**1.45**
	Slaughterhouses (>50t/d), plants for the production of milk (>200t/d), other animal raw materials (>75t/d) or vegetable raw materials (>300t/d)	0.91	0.75	1.10	0.99	0.81	1.21	0.99	0.77	1.27
	Installations for surface treatment or products using organic solvents (>200t/y)	0.95	0.83	1.10	1.10	0.93	1.31	1.14	0.91	1.43
1500 m	Combustion installations > 50 MW	1.11	0.93	1.33	1.09	0.91	1.31	1.07	0.84	1.35
	Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	0.92	0.81	1.05	0.93	0.81	1.06	0.96	0.82	1.12
	Installations for the production of cement clinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	0.83	0.71	0.98	0.87	0.74	1.03	0.89	0.73	1.07
1000 m	Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	0.92	0.78	1.08	0.89	0.75	1.05	0.91	0.74	1.10
	Installations for the production of cement clinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	0.97	0.77	1.23	1.00	0.79	1.26	1.03	0.79	1.33

Table 4.2. *Relative risks and 95% confidence and credibility intervals for towns with installations lying within a radius of 2000, 1500 and 1000 metres from the municipal centroid. Estimates adjusted for age, sex and socio-demographic variables.*

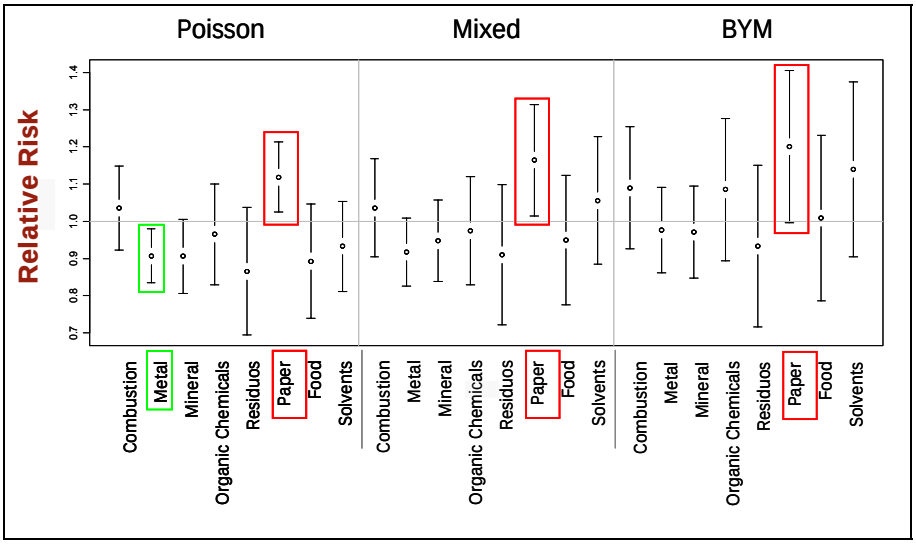


Figure 4.2. *Relative risks and 95% confidence and credibility intervals for the relative risk associated to the different industrial sectors for a radius of 2000*

4.4 DISCUSSION

The results show a possible association between exposure to air pollution from the paper, pulp and board industry and excess risk of NHL mortality, regardless of which model is used. Analysing the information contained in the EPER for 2001 shows that almost all the paper, pulp and board industries reported emissions of the following compounds, above the threshold established for their inclusion in the registry: CO; CO_2; NO_2; sulphur dioxide; organochlorinated compound mixtures; and organic carbon. Taken individually, some of these industries also reported emissions of metals (chrome, copper, nickel, lead and zinc), as well as phosphorous, nitrogen and PM10 particulate matter.

In the literature, there are few studies that link NHL to environmental exposure to chemical substances. Some occupational studies suggest a positive association with exposure to organic solvents, such as benzene [Blair et al., 1993;Hardell et al., 1998;Hayes et al., 1997], trichloroethylene (TCE), tetrachloroethylene (PCE) and styrene [Wartenberg et al., 2000]. Other occupational studies associate exposure to pesticides with an elevated risk of NHLs [Garabrant and Philbert, 2002; Lynge et al., 1997]. Lastly, different studies addressing the relationship between NHLs and exposure to dioxins furnish contradictory results [EPIYMPH, 2007; Cole et al., 2003]. In one study on a large cohort of paper industry workers, mortality from non-Hodgkin's lymphoma and leukaemia was higher among workers with elevated SO(2) exposure, and a dose-response relationship with cumulative SO(2) exposure was suggested for non-Hodgkin's lymphoma. The cohort included 57,613 workers who had been employed for a minimum of 1 year in the pulp and paper industry in 12 countries [Lee et al., 2002]. Aside from environmental exposures, there is evidence to indicate that situations associated with chronic antigenic stimulation or immunosupression favour the appearance of these tumours [Eltom et al., 2002; Fisher and Fisher, 2004].

Assessment of exposure to environmental agents that are noxious to human health is a very complex process. At present, there is a great variety of exposure-measurement strategies, depending on the timeliness and availability of resources, which include the use of remote sensors, biomarkers, or estimates of pollutant dispersion using theoretical or statistical models [Nieuwenhuijsen et al., 2006]. With respect to this last avenue of research, there are a number of studies in the literature that seek to estimate the risk associated with proximity to hazardous sites (focused clustering) [Sans et al., 1995; Wakefield and Morris, 2001]. In these and other studies, the authors have explored the idea of estimating risk according to distance [Elliott et al., 2000; Muller et al., 2005; Sans et al., 1995].

At present, the real availability of data from remote sensors or biomarkers is negligible. Hence, in the absence of such information, many studies have used distance as an exposure marker. This approach has been further refined, by endeavouring to model pollutant dispersion using

anisotropic models that take data, such as wind direction or geographical relief [Lawson A, 2001], into account. These models could not be applied to this study, however, for lack of information of this type.

With respect to our study, using the distance from the industry to the municipal centroid means that, as the study radius is reduced, the number of towns deemed to be exposed falls drastically. This situation leads to the elimination of exposure variables and the impossibility of studying variation in risk according to a more stringent definition of exposure for most of the emissions considered. Based on the results for the two industrial groups analysed at the three distances (production and processing of metals and mineral industries), no conclusion can be reached as to variation in risk with variation in distance to the emission source.

In ecological spatial correlation studies, Poisson regression is one of the basic tools applied to analysing the association between risk of mortality and the various potential risk factors [Elliott et al., 2000; Lawson A, 2001]. This type of regression forms part of so-called generalised linear models and assumes independence between observations or counts, an assumption that could be violated when working with data that have a spatial structure [Clayton et al., 1993; Elliott et al., 2000]. Nevertheless, the use of Poisson regression may help obtain an initial assessment of the presence or absence of this association. Indeed, a number of authors have used this method to evaluate the relationship between risk factors and excess incidence or mortality in the study of non-communicable diseases in a spatial context [Kokki and Penttinen, 2003; Wakefield and Morris, 2001]. The second model used -the mixed model- is included as an intermediate step between a model that assumes total independence and a model that assumes autocorrelation among observations, and has the advantage of circumventing the problems of extra-Poisson dispersion, lending robustness to the estimators and using the provincial level to approach autocorrelation, which amounts to a form of stratification in the comparisons. Lastly, the third model -the BYM model- assumes that each observation is conditionally independent of the others, i.e., that observations are spatially correlated amongst themselves, with the aim of modelling the spatial effect of the risk [Besag J et al., 1991; Congdon, 2001; Kokki and Penttinen, 2003]. In none of the models, multiple comparison adjustment was considered. The probability of one spurious test result was 0.33. Due to this low probability and the number of comparisons, we decided to asses the adjustment for multiple testing by the consistency of the associations showed by the results of the different models.

In our results for almost all the industrial sectors considered, the related risks were observed to increase as the random effects covered by the spatial structure of the data were included. The relative risks yielded by the mixed model are, in general, higher than those yielded by the Poisson regression, while those yielded by the BYM model are the highest for most of the variables. The inclusion of random spatial effects terms in risk estimation, not only improves the study of the associations between environmental exposures and mortality, but also reduces

proneness to "ecological bias" as a result of working on a larger scale and adjusting for unknown confounders which have a spatial distribution different to that of mortality [Clayton et al., 1993]. However, bearing the similarity of results in mind, the decision to apply the spatial model in exploratory studies of this magnitude must be carefully evaluated, due to the excessive time of computation. The ever increasing availability of health and exposure data calls for the definition of a fast and easy methodology of analysis that would optimise available resources within research groups when it came to embarking upon exploratory studies [Ramis et al., 2007].

None of the socio-demographic variables considered in our study appeared to act as a potential confounder, inasmuch as their elimination in the various models led to no substantial changes in the effect estimators of the distance to the industrial foci studied (data not shown). Furthermore, these possible confounding variables, defined *a priori*, displayed no important direct effect on risk of NHL mortality, registering RRs close to unity.

As stated above, little is known about the possible role of environmental exposures in NHL aetiology, which may be due to the fact most of the studies undertaken to date focused on small towns and poor-quality exposure measures. This implies a limited statistical power that hinders the estimate of modest RRs [Floret et al., 2003]. This paper presents a first approach to the exploration of the influence of exposures to industrial air pollution and risk of NHL mortality vis-à-vis the entire population of a country, something that is an advantage in terms of the sheer size of the exposed population but is a drawback in terms of possible misclassification of exposure or the uniqueness of each of the installations.

Other possible limitation is the use of ICD9, that classification has not different code for each type of lymphoma included in the LNH; as a result we can not know the spatial patterns of each individual type. Moreover, mortality data only includes the more aggressive type of lymphoma. Less aggressive lymphomas have a low mortality rate and, consequently, they are not included in this study.

It should also be pointed out that the data referring to environmental industrial exposures were drawn from the first edition of the EPER. The quality of this information may conceivably improve with the new European Pollutant Release and Transfer Register (E-PRTR), which will completely replace the EPER in 2009, thereby allowing for the validity of a study of this type to be enhanced, with the possibility of evaluating the effect of specific pollutants. Moreover, though the "near versus far" analysis conducted in this study assumes all the industries of a single sector to be equal, it must nevertheless be borne in mind that each industrial source has its own characteristics, and subsequent studies will therefore have to address these on a case-by-case basis.

Finally, we should not forget that the use of aggregated data implies important assumptions. We assume that the whole population within a municipality lives in its centroid; even more, we assume that they have always been living there. Also, we do not consider the daily movement of the people to go to work or study, for instance. Hence, we are assuming that everybody within an area is exposed to the same type and amount of pollutant substances.

The results suggest a possible increased risk of NHL mortality among populations residing in the vicinity of paper and pulp industries, an excess mortality that is observable using different models. In order to confirm or reject these results, it would be of great interest to seek to improve the exposure markers and ascertain precisely what is happening in the environs of each specific installation. In addition, the availability of incidence data would be very useful to study less aggressive lymphomas with low mortality rate, which are not included in this study. Those data would provide valuable information to analyse the spatial patterns of individual type of lymphomas integrated in modern classifications of the LNH in reference to specific locations and exposures. Unfortunately, currently there are no incidence data available at national level in Spain.

5. RISK AROUND PUTATIVE FOCUS IN A MULTY-SOURCE SCENARIO. NON-LINEAL REGRESSION MODELS

5.1 INTRODUCTION

At present, there is a constant release of toxic substances to the environment from industrial activity. However evidence regarding the health risk of living near to pollutant factories and, therefore, being exposed to their pollution is limited. One of the most studied health problems related to exposure to pollution is cancer. Some authors have described associations between lung cancer, metallurgical industry and other industrial areas [Gottlieb and Carr, 1982; Monge-Corella et al., 2008; Parodi et al., 2005]. Also, lymphomas and leukaemia are more frequent in the proximities of industrial areas [Benedetti et al., 2001; Gottlieb and Carr, 1982; Lopez-Abente et al., 1999; Sans et al., 1995; Sharp et al., 1996; Viel et al., 2000]. However, others studies have not found association between cancer and proximity to industrial facilities and incinerators [Elliott et al., 1992; Michelozzi et al., 1998; Pekkanen et al., 1995]. On the other hand, a municipal mortality atlas recently published in Spain presents heterogeneous patterns of spatial distributions for some cancer causes which suggest that environment factors may be important in their aetiology [Lopez-Abente et al., 2006b].

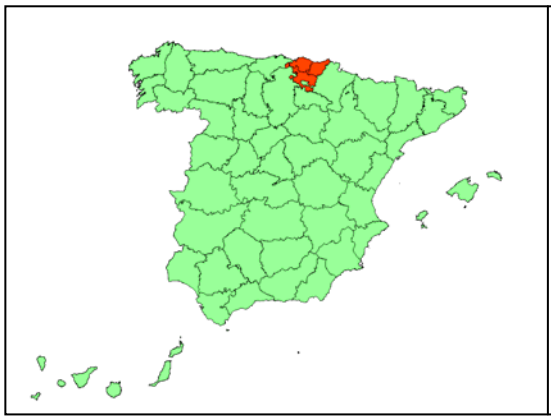
Assessment of exposure to environmental agents that are noxious to human health is a very complex process. At present, there is a great variety of exposure measurement strategies, depending on the availability of resources, which include the use of remote sensors, biomarkers, or estimates of pollutant dispersion using theoretical or statistical models [Nieuwenhuijsen et al., 2006]. With respect to this last research possibility, there are a number of studies in the literature that seek to estimate the risk associated with proximity to hazardous sites (focused clustering) [Elliott et al., 2000]. In these and other studies, the authors have explored the idea of estimating risk according to distance [Biggeri et al., 1996; Diggle and Rowlingson, 1994; Draper et al., 2005; Elliott et al., 1996; Maule et al., 2007]. At present, the availability of data from remote sensors or biomarkers in this context is very limited. Hence, in the absence of such information, many studies have used distance as an exposure marker. This approach has been further refined by endeavouring to model pollutant dispersion assuming multiplicative risk factors from separate sources [Diggle et al., 1997].

This study seeks to explore the relationship between municipal cancer mortality in Spain and distance from industrial facilities, as an indirect measure of exposure to industrial pollution in a multi-source scenario, using a Poisson-regression-based model.

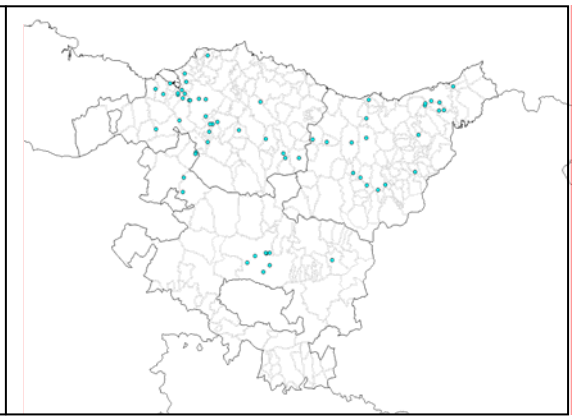
5.2 MATERIALS AND METHODS.

5.2.1 Data

The study region is the Basque Country, sited in the north of Spain (Map 4.1). Considering the data from the 2001 official census, the population of the Basque Country is 2.082.587 inhabitants, distributed between 247 municipal areas or 1645 census tracts, and the total extension of the region is 7.234 km^2 hence the population density is 289 inhabitants per km^2. This region is one of the most industrialize of Spain. Specifically there are 77 industrial facilities, registered in EPER*, sited within the region (Map 5.2). Moreover, the Basque Government facilitated census tract mortality and geographical data to carryout this study.



Map 5.1. Spain. Basque Country in red



Map 5.2. Basque Country, municipalities and factories

I. Cases: Mortality data

This study uses two different sets of cancer mortality data. Even though in both sets the source of cases is the individual death entries of the mortality register provided by the National Statistics Institute (*Instituto Nacional de Estadística – INE*).

The first set collects the number of deaths caused by cancer during the period 1994-2003, aggregated at municipal level. Table 5.1 displays the list of causes considered as well as the number of deaths per cause and the rate per 1000 inhabitants. This data was furnished by the National Statistics Institute (*Instituto Nacional de Estadística – INE*) for the production of a municipal cancer mortality atlas [Lopez-Abente et al., 2006b].

Tumours	ICD 9	Obs	rate=obs/100000
Buccal cavity and pharynx	140-149	1593	79.65
Esophagus	150	1449	72.45
Stomach	151	3745	187.25
Colon-Rectum	153-154	6288	314.4
Gall-Bladder	156	820	41
Pancreas	157	2328	116.4
Larynx	161	1166	58.3
Lung	162	9121	456.05
Bones	170	123	6.15
Connective tissue	171	212	10.6
Melanoma	172	362	18.1
Breast	174	3187	159.35
Uterus	179-182	921	46.05
Ovary	183	894	44.7
Prostate	185	2753	137.65
Bladder	188	2001	100.05
Kindney	189	1149	57.45
Brain	191	1264	63.2
Non Hodgkings limphomas	200,202	1205	60.25
Myeloma	203	760	38
Leukemias	204-208	1325	66.25

Table 5.1. Causes, ICD 9, number of cases and rate per 100.000 inhabitants for the period 1994-2003

The second data set belongs to the Health Department of the Basque Country Government; it gathers the number of deaths between the years 1996 and 2003, but in this case the data is broken down by census tract, the much finer spatial resolution. Table 5.2 shows the causes, the number of deaths and the rate per 1000 inhabitants. The populations of the census tracts vary between 1000 and 2000 inhabitants

Tumours	ICD 9	Obs	rate=obs/1000
Esophagus	150	1156	57.8
Stomach	151	2960	148
Colon-Rectum	153-154	8750	437.5
Larynx	161	909	45.45
Lung	162	7385	369.25
Breast	174	2544	127.2
Prostate	185	2268	113.4
Bladder	188	1639	81.95
Kindney	189	936	46.8
Haematilogical	200-208	2831	141.55

Table 5.2. Causes, ICD 9, number of cases and crude rate per 100.000 inhabitants for the period 1996-2003

II. Expected cases

The estimation of expected cases is done using indirect standardization as follows. For the first data set the whole period of time under study was divided in two quinquennia (1994–1998 and 1999-2003). The overall Spanish mortality rates for the above two 5-year periods are multiplied by each town's person-years, age group, sex and quinquennium. The person-years for each five-year period are obtained by multiplying the populations by 5; the municipal populations,

broken down by age group (18 groups) and sex, are obtained from the 2001 census and the 1996 municipal roll. These years correspond to the mid-points of the quinquennia.

In the second data set, the number of expected cases in each census tract is also estimated using the overall Spanish mortality rates but, in contrast to the first case, the data set is not divided into two periods. The census tracts populations are extracted from the 2001 census and processed using same strategy employed in the first data set [Barcelo et al., 2008].

III. Socio-demographic covariates

Previous to introducing the socio-demographic covariates included in this study we are going to present the concept of confounding.

> "Confounding can be defined as confusion, or mixing, of effects. The effect of the exposure variable is mixed together with the effect of another variable that is associated with the exposure and is an independent risk factor for the disease. The consequence is that the estimated association of the exposure is not the same as its true effect" [Rothman K, 2002].

Cancer incidence and mortality have many known and unknown risk factors. Some of the known factors are socio-demographic characteristics of the population. The influence of these factors should be controlled when the aim is to explore the effect of different factors in order to control the possible confounding. Age and sex are very important cancer risk factors and they should be always taken into account. In this study they are controlled by the use of indirect standardization when the number of expected cases is estimated. However, there are many more socio-demographic characteristics in a population that can determine the distribution of cancer over the population. Also, it is very important to consider the latency period of a disease such as cancer. Specialists suggest that for most of the cancer locations the latency period could be quite long, 10 years or more. For this reason the socio-demographic information used for this kind of study should be previous to the studied period, in our case the best information available comes from the 1991 census, even though the aggregation level of this data is municipal, not census tract.

The selected covariates from 1991 census to be included in the analysis are: *percentage of illiterates*, *percentage of unemployed* and *cohabitants per house*. The census does not include data about the socio-economic status. For this purpose we use an indicator of socio-economic level, *income*, provided by the Spanish Credit Bank for 1991 [Banco Español de Credito, 1993]. This index classifies towns and cities into 10 levels according to the estimated average domestic income. All these covariates are standardized at national level. Finally, we also wanted to consider prevalence of tobacco as a covariate but unfortunately such information is no available at the required aggregation level. Consequently, we decided to use the Standard

Mortality Ratio of lung cancer as an approximation of the *tobacco prevalence* [Lopez-Abente et al., 2006a].

Summarizing, the five socio-demographic covariates are:

1. Percentage of illiterates = *Education (-)*
2. Percentage of unemployed
3. Cohabitants per house = *cph*
4. Income
5. RR lung cancer = *Tobacco*

IV. Factories

As a source of information about the industrial facilities we used the European Pollutant Emission Register (EPER) [Garcia-Perez et al., 2008;Garcia-Perez et al., 2009]. This data-base collects information regarding emissions to air, soil and water from agricultural or industrial facilities and data of 50 pollutant substances. The information available allows us to identify different types of industrial activities. In February 2004, EPER data on Spain (for 2001) was published. Industrial activities classified in the EPER fall into the following 6 categories: 1) Energy industries; 2) Production and processing of metals; 3) Mineral industry; 4) Chemical industry and chemical installations; 5) Waste management; and 6) Other activities (which include paper and board production, manufacture of fibres or textiles, tanning of hides and skins, slaughterhouses, intensive poultry or pig rearing, installations using organic solvents, and the production of carbon or graphite).

In the present study, we are working with the industrial facilities that declare emissions to air only. For this specific group of industries, EPER collects information about 61 industrial facilities sited in the Basque Country. The distribution of the factories between the six main industrial categories is show in the next table.

	Industrial categories	**Nº of facilities**
1	Energy industries	4
2	Production and processing of metals	28
3	Mineral industry	8
4	Chemical industry and chemical installations	4
5	Waste management	5
6	Other activities (which include paper and board production, manufacture of fibres or textiles, tanning of hides and skins, slaughterhouses, intensive poultry or pig rearing, installations using organic solvents, and the production of carbon or graphite)	12

Table 5.3. Industrial categories and number of facilities.

An exhaustive classification (Table 5.4) can be performed based on a more specific definition of the industrial activity (Industrial sector).

Industrial categories	Industrial Sector	N° of facilities
1	Combustion installations > 50 MW	3
1	Mineral oil and gas refineries	1
2	Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	28
3	Installations for the production of cement klinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	8
4	Basic organic chemicals	2
4	Basic inorganic chemicals or fertilisers	2
5	Installations for the disposal of nonhazardous waste (>50t/d) and landfills (>10t/d)	5
6	Industrial plants for pulp from timber or other fibrous materials and paper or board production (>20t/d)	4
6	Slaughterhouses (>50t/d), plants for the production of milk (>200t/d), other animal raw materials (>75t/d) or vegetable raw materials (>300t/d)	1
6	Installations for surface treatment or products using organic solvents (>200t/y)	7

Table 5.4. Industrial categories, industrial sectors and number of facilities.

5.2.2 Methods

I. Model

In epidemiology the standard method to analyse aggregated data is ecological regression, specifically the Poisson regression is used for chronic diseases such as cancer. On this occasion, we have extended the standard Poisson model with the inclusion of a term based on the distance to the point source, to analyse the effect of the exposure to pollutant substances released by industrial facilities over the spatial distribution of cancer mortality. The log-linear formulation of the standard Poisson regression is unrealistic for this study because of the need to combine an elevated risk close to the source with a neutral long-distance effect; therefore, we extend the model by the inclusion of a non-linear distance function proposed by Diggle [Diggle and Rowlingson, 1994], $f(d_{ij})$

$$O_i \sim Po(E_i \mu_i) \tag{1}$$

$$\mu_i = \rho \exp\left[\sum_k (\vartheta_k Z_{ik})\right] \prod_j f(d_{ij}); \qquad f(d_{ij}) = 1 + \alpha_j \exp\left[-(d_{ij} / \beta_j)^2\right]$$

- ρ is the overall risk
- θ_k are the parameters of the socio-demographic covariates Z_{ik}
- αj and βj are the parameters of the distance function, and d_{ij} is the distance between the centroid of the area *i* and the focus *j*.

II. Inference

The approximate log-likelihood function for this model without constant term is [Diggle et al., 1997]:

$$L(\rho, \vartheta, \alpha, \beta) = -\sum_i \mu_i + \sum_i O_i \log(\mu_i) \tag{2}$$

III. Parameter estimation

The estimators of the parameters are obtained by direct maximisation of the likelihood function using the numerical optimization function "optim", included in R. The R code for this function is in the appendix.

We have carried out examples to assess the performance of this function, comparing its results with those from the *"nlr"* function included in the *"gnlm"* library by J. K. Lindsey [Linsey, 2001], which was developed to fit non-linear regression models.

Example: as observed data we use the stomach cancer cases aggregated at municipal level and the exposure from the industrial facilities 3689 and 3716 belonging to "basic organic chemical" sector. The socio-demographic covariates are also included in the models. (Table 5.5)

Parameter estimators	*Study function*	*Lindsey Function*
rho	1.0158	1.0219
education	-0.0477	-0.0460
unemployment	0.1378	0.1363
income	0.0927	0.0927
cohabitans	-0.0957	-0.0948
alpha1	0.1836	0.1836
alpha2	0.0735	0.0679
beta1	0.1644	0.1657
beta2	0.6299	0.5745

Table 5.5. Parameter estimators. Study function vs Lindsey function.

IV. Standard error calculations

A common way to approximate the standard errors of the parameters estimators in a non-linear regression model is through the inverse of the Hessian. Usually optimization algorithms in R, such as “optim”, provide the Hessian matrix, though we have found that for point source models like the one described above, even when numerically accurate values are returned for the maximum likelihood parameter estimates, the associated standard errors derived by inverting the estimated Hessian can be unreliable. As an alternative strategy, we obtain standard errors by combining the R function for direct maximisation of the likelihood with replicated Monte Carlo simulations of the fitted model. In Table 5.6 presents results of a simulation experiment; we have simulated a dataset of observed cases from a model with four socio-demographic covariates and two pollutant sources. In the left column the given values for the parameters are presented. The next 3 columns give standard errors; the first from the left (*Standard errors*) have the Monte Carlo standard errors calculated through the given values; the second (*Monte Carlo standard errors*) have the Monte Carlo standard errors calculated through the estimated parameters; and the last column shows the standard errors provided from Lindsey algorithm calculated with the Hessian Matrix.

	Values	Standard errors	Monte Carlo standard errors	Lindsey standard errors (Hessian)
ρ	**0.019**	**0.160** (0.146-0.177)	**0.216** (0.196-0.239)	**0.245**
θ1	**0.111**	**0.172** (0.156-0.190)	**0.254** (0.231-0.282)	**0.263**
θ2	**0.099**	**0.100** (0.091-0.110)	**0.129** (0.117-0.143)	**0.125**
θ3	**-0.020**	**0.066** (0.060-0.073)	**0.089** (0.081-0.099)	**0.101**
θ4	**-0.093**	**0.074** (0.067-0.082)	**0.098** (0.089-0.108)	**0.100**
α1	**0.100**	**0.238** (0.216-0.263)	**0.270** (0.246-0.299)	**0.192**
α2	**0.100**	**0.179** (0.162-0.198)	**0.208** (0.189-0.230)	**0.180**
β1	**0.200**	**0.232** (0.211-0.257)	**0.174** (0.159-0.193)	**0.123**
β2	**0.400**	**0.340** (0.308-0.375)	**0.395** (0.360-0.438)	**1.149**

Table 5.6. Real values, real standard errors, Monte Carlo standard errors and Lindsey standard errors (Hessian).

V. Hypotheses testing

To test hypotheses about the parameters we use the likelihood ratio test. Likelihood ratios statistic D;

$$D = 2\left\{ L\left(\hat{\varphi} \right) - L_0\left(\hat{\varphi}_0 \right) \right\}$$

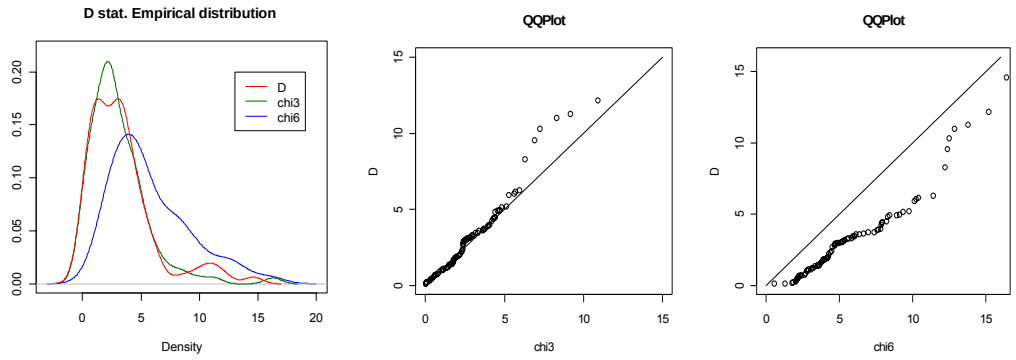
VI. Approximate null distribution of likelihood ratio statistic D

Previous studies pointed out that usual asymptotic properties of the likelihood ratio test are not clear for models with a non-linear component [Diggle and Rowlingson, 1994;Diggle et al., 1997].

To clarify this point we run a simulation experiment generating data from the following models for each area of the region:

1. Null model: $\mu_i = \rho$
2. Distance model : $\mu_i = \rho \prod_j f(d_{ij})$

For the distance model we contemplate two scenarios, one with 3 focuses and another with 4. For each of the scenarios we set of 100 simulations and calculate the corresponding likelihood ratio statistic D. Due to the form of the distance function, when α =0, β is indeterminate. We think this fact may affect to the number of effective parameter of the model and consequently may affect to the degrees of freedom. Thus we consider two reference distributions to test, χ^2_n and χ^2_{2n}, where n is the number of focuses in the empirical model. We have performed graphical and numerical tests, such as QQ-plot and Kolmogorov-Smirnov test, to contrast the form of the empirical distribution against the two theoretical distributions; we also include a graph of densities. For the first scenario Figure 5.1 shows three graphs: a density graph with the empirical distribution and the two reference distributions (a); Q-Q plot of sample D-values with χ^2_3 (b); and Q-Q plot of sample D-values with χ^2_6 (c). Moreover, as follows, we give the p-values for the Kolmogorov-Smirnov goodness-of-fit statistic. In the left graph we can see how the density of D-statistic almost overlap the density of χ^2_3; on the contrary, the shape of the χ^2_6 density is different from the empirical density. In the QQ-plot for χ^2_3 almost all points are over the main diagonal, only the last dots are away of it; alternatively dots in the second QQ-plot do not follow the main diagonal. Finally, results of Kolmogorov-Smirnov test are consistent with the graphical tests. P-value for the first contrast, empirical distribution of D vs χ^2_3, is 0.4767; hence the null hypothesis can not be rejected. The second p-value is $6.881e^{-07}$, therefore null hypothesis is rejected for χ^2_6.

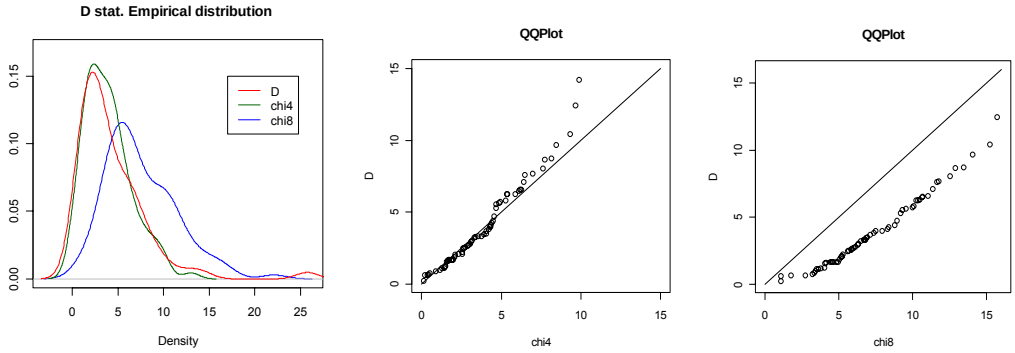


(a) (b) (c)

Figure 5.1 Comparison between the empirical distribution and the reference distributions in a three focus scenario. Density graph with the empirical distribution and the two reference distributions (a); Q-Q plot of sample D-values with χ^2_3 (b); and Q-Q plot of sample D-values with χ^2_6 (c).

```
        Two-sample Kolmogorov-Smirnov test
data:  D and chi3, D = 0.121, p-value = 0.4767
data:  D and chi6, D = 0.3973, p-value = 6.881e-07
```

Next, Figure 5.2 shows the graphs for the scenario with four focuses followed by the p-values for the Kolmogorov-Smirnov goodness-of-fit statistic. As in the previous case, results of the tests are clear in their conclusions. The empirical distribution of D-statistic can be approximated by a χ^2_4 distribution, but not for a χ^2_8 distribution.



(a) (b) (c)

Figure 5.2 Comparison between the empirical distribution and the reference distributions in a four focus scenario. Density graph with the empirical distribution and the two reference distributions (a); Q-Q plot of sample D-values with χ^2_4 (b); and Q-Q plot of sample D-values with χ^2_8 (c).

```
        Two-sample Kolmogorov-Smirnov test
data:  D and chi4, D = 0.0888, p-value = 0.8707
data:  D and chi8, D = 0.4676, p-value = 1.497e-08
```

The results of this simulation experiment suggest that a χ^2_n is a good approximation of the null sampling distribution of the likelihood ratio statistic D. On the other hand χ^2_{2n} seems to be a bad a approximation. This conclusion is opposite to that taken from previous papers [Diggle and Rowlingson, 1994;Diggle et al., 1997], where χ^2_{2n} was the distribution of the generalized likelihood ratio statistic, however some authors have discussed about the failure of asymptotic properties in non-regular likelihood when indeterminate parameters are involved .

5.2.3 Exploratory analysis

The three data sets used in this project, two mortality data sets and one pollutant emissions data set, provide a large number of possible analyses. To reduce this number and to focus only in those associations with a potentially positive result we first carry out an exploratory analysis where we fit a standard Poisson regression model that include the socio-demographic

covariates (Z_{ik}) and a variable derived from the distance between the point source and the centroid of the area (D_i).

$$O_i \sim Po(E_i \mu_i)$$

$$\mu_i = \rho \exp\left[\sum_k (\vartheta_k Z_{ik})\right] \exp[\phi D_i] \qquad (3)$$

- ρ is the overall risk
- θ_k are the parameters of the socio-demographic covariates Z_{ik}
- ϕ is the parameter associate to the distance variable D_i

For this first approach, the distance based variable takes a simple form. We create a binomial variable giving value *1* to those municipal areas, or census tracts, which have a factory within a circumference of fixed radius, and value *0* to the remaining areas. To assess the possible variation of the risk with the variation of the distance, we vary the length of the radius over the following values: 0.5km, 1km, 1.5km, 2km, 3km, 4km, 5km and 6km. Moreover, to reduce the number of regressions to fit, we aggregate the factories by industrial sector. Tables 5.7 and 5.8 show the number of areas within the fixed radius for the different distances aggregated by sector.

Industrial Sector	**Nº factories**	**Number of areas within the distance of * metres from a factory**							
		500*	**1000***	**1500***	**2000***	**3000***	**4000***	**5000***	**6000***
Combustion installations > 50 MW	*3*	0	2	3	4	5	10	14	18
Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	*28*	5	18	25	34	51	63	78	93
Installations for the production of cement klinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	*8*	1	3	6	9	17	22	39	51
Basic organic chemicals	*2*	1	1	1	2	4	6	9	10
Installations for the disposal of nonhazardous waste (>50t/d) and landfills (>10t/d)	*5*	0	1	2	5	6	10	20	22
Industrial plants for pulp from timber or other fibrous materials and paper or board production (>20t/d)	*5*	0	1	3	6	8	13	20	22
Installations for surface treatment or products using organic solvents (>200t/y)	*6*	0	2	4	7	11	17	22	32

Table 5.7. Municipal level. Number of areas aggregated by industrial sector within the fixed radius.

Industrial Sector	Nº factories	Number of areas within the distance of * metres from a factory							
		500*	1000*	1500*	2000*	3000*	4000*	5000*	6000*
Combustion installations > 50 MW	*3*	2	15	62	132	264	383	450	517
Metal industry and metal ore roasting or sintering installations, Installations for the production of ferrous and non-ferrous metals	*28*	29	104	283	484	938	1528	2370	3104
Installations for the production of cement klinker (>500t/d), lime (>50t/d), glass (>20t/d), mineral substances (>20t/d) or ceramic products (>75t/d)	*8*	17	87	196	307	606	867	1032	1221
Basic organic chemicals	*2*	6	36	74	96	165	253	314	414
Installations for the disposal of nonhazardous waste (>50t/d) and landfills (>10t/d)	*5*	1	4	18	46	133	304	556	749
Industrial plants for pulp from timber or other fibrous materials and paper or board production (>20t/d)	*5*	1	2	6	11	36	73	92	141
Installations for surface treatment or products using organic solvents (>200t/y)	*6*	4	33	80	125	247	347	450	539

Table 5.8. Census tract level. Number of areas aggregated by industrial sector within the fixed radius.

A different model is fitted for each cause, industrial sector and distance. Three sectors are not analysed for the following reasons: *alimentation* has only one facility; *combustion* has a small number of areas within the circumferences; and *Inorganic chemicals* has both facilities located in the same municipality.

I. Municipal analysis results

Covariates	Esophagus	Stomach	Colon-Rectum	Gall-Bladder	Pancreas	Larynx	Lung	Breast	Uterus	Ovary	Prostate	Bladder	Kidney	Brain	Leukemias
Percentage of illiterates		1.303				2.485						1.494			
Percentage of unemployed	1.368	1.120	1.247			1.568	1.479	1.180				1.193	0.590		
Income	0.862		0.890					0.890					1.346		
Cohabitants per house												0.884			

Table 5.9. Risk estimations for the covariates by cancer cause. Municipal level.

Using model (3); the relative risks for each cause linked to the socio-demographic covariates suggest some associations (Table 5.9). The percentage of illiterates is a risk factor for stomach, larynx and bladder cancer, specifically the relative risk is very high for larynx, 2.485. The percentage of unemployed is associated with esophagus, stomach, colon-rectum, larynx, lung, breast and bladder cancer. Income is a protective factor for esophagus, colon-rectum and breast, and a risk factor for kidney cancer. Finally, the covariate “cohabitants per house” is a protective factor for bladder cancer.

The following table shows only the distances with some statistically significant relative risks. Lung cancer does not appear because of did not show statistically significant associations.

Some sectors seem to have a trend connected to the distance, the risks increase with the proximity to the focus. However, others show isolated risks that appear to be unrelated to interpretable any distance effect. (Table 5.10)

	Dist	Nº Areas	Esophagus	Stomach	Colon-Rectum	Gall-Bladder	Pancreas	Larynx	Breast	Uterus	Ovary	Prostate	Bladder	Kidney	Brain	Leukemias
Combustion installations	*6000*	*18*		1.059	1.107	1.185							1.193			
	5000	*14*		1.107	1.085	1.25							1.146			
	4000	*10*		1.091	1.101	1.306							1.157			
	3000	*5*		1.17	1.079	1.382							1.245			
	2000	*4*		1.156	1.1	1.472	1.118						1.243			1.14
	1500	*3*				1.447	1.285								1.536	
	1000	*2*	1.661		1.256			1.54					1.496			
Metal industry	*6000*	*93*						1.029	1.017							
	5000	*78*						1.037								
	4000	*63*		1.052								1.012				
	3000	*51*		1.053								1.016				
	2000	*34*										1.033				1.108
	1500	*25*										1.072				
	1000	*18*										1.268				
	500	*5*										1.234				
Mineral	*6000*	*51*		1.074	1.077											
	5000	*39*		1.073	1.083											
	4000	*22*		1.087	1.069											
	3000	*17*		1.1												
	2000	*9*							1.116							
	500	*1*				4.115										
Organic chemicals	*6000*	*10*			1.087						1.319					
	5000	*9*			1.086											
	4000	*6*								1.33						
	3000	*4*								1.318						
	2000	*2*								1.326						1.285
	1500	*1*								1.567						1.244
Nonhazardous waste Installations	*6000*	*22*			1.071	1.133							1.115			
	5000	*20*	1.088		1.091	1.132							1.188			
	4000	*10*							1.109							
	3000	*6*	1.555													
	2000	*5*	1.564													
	1000	*1*							2.897							
Paper industry	*6000*	*22*						1.314		1.318						
	5000	*20*						1.248		1.329						
	4000	*13*						1.311		1.378						
	3000	*8*	1.394													
	1500	*3*		1.557												
Solvents	*6000*	*32*		1.061		1.169	1.039						1.126	1.084		
	5000	*22*		1.083		1.178	1.053						1.113	1.092		
	4000	*17*	1.016	1.086		1.184	1.109						1.114	1.096		
	3000	*11*		1.153		1.377	1.124						1.2	1.169		
	2000	*7*		1.557		1.392	1.129						1.219			
	1500	*4*					1.392									
	1000	*2*	1.523													

Table 5.10. Risk estimations for the distance variables by cancer cause. Municipal level.

Summarizing the result by industrial sector, four causes show a trend in risk in relation with the combustion installations: stomach, colon-rectum, gall-bladder and bladder cancer. Only prostate cancer has a statistically significant distance-trend risk relative to the metal industry. The mineral industry may be related in long distances with stomach and colon-rectum cancer risks. Chemical organic sector analysis reveals a distance-trend risk with uterus cancer, though the number of areas exposed is very low. The risk associated with the non-hazardous waste installations can be considered as punctual and not connected to changes in the distance between the point source and the centroid. Paper factories yield statistically significant trend-distance risks with larynx and uterus. Industrial facilities which use solvents have distance-trend associations with: stomach, gall-bladder, pancreas, bladder and kidney. Finally, in contrast, the following cancer types do not show interpretable distance-trends: esophagus, lung, breast, ovary, brain and leukaemia.

II. Census tracts analysis results

The socio-demographic data from the 1991 census is only available at municipal level, rather than at the disaggregated lever of census tracts. Accordingly, we use the municipal value as an

approximation of the census tract value to introduce socio-demographic information in the following models:

Covariates	Esophagus	Stomach	Colon-Rectum	Larynx	Lung	Breast	Prostate	Bladder	Haematological
Percentage of illiterates				2.710	0.974			1.749	
Percentage of unemployed	1.274	1.178	1.212	1.556	1.493	1.302			1.162
Income									
Cohabitants per house			0.870					0.785	

Table 5.11. Risk estimations for the covariates by cancer cause. Census tract.

Risks associated to the covariates have almost the same behaviour as at the municipal level. The percentage of unemployed seems to be a risk factor to nearly all causes. However we find no statistically significant risks related to income.

	Dist	Nº Areas	Esophagus	Stomach	Colon-Rectum	Larynx	Lung	Breast	Prostate	Bladder	Haematological
Combustion installations	6000	517		1.115	1.116			1.082		1.134	
	5000	450		1.129	1.122					1.091	
	4000	383		1.113	1.124					1.137	
	3000	264		1.18	1.098					1.249	
	2000	132		1.107						1.325	
	1500	62				1.424	1.108			1.247	
	1000	15				1.659	1.307			1.693	
	500	2					1.549				
Metal industry	6000	3104			1.021						
	4000	1528			1.037						
	1500	283	1.119								
	1000	104	1.302						1.332		
Mineral	6000	1221			1.089	1.107					
	5000	1032			1.078						
	4000	867			1.067						
	3000	606		1.095							
	2000	307					1.088	1.107			
	1500	196					1.067	1.131			
	1000	87			1.143						
Organic chemicals	6000	414						1.071			1.088
	5000	314						1.091			
	4000	253						1.141			
	3000	165									1.134
	2000	96									1.179
	1500	74									1.156
Nonhazardous waste Installations	6000	749		1.083	1.071			1.067			
	5000	556						1.083			
	4000	304						1.071			
	3000	133						1.114			
Paper industry	1500	6	1.961								
Solvents	6000	539								1.1	
	5000	450		1.079						1.083	
	4000	347		1.18						1.113	
	3000	247		1.116						1.174	
	2000	125		1.211		1.274					
	1500	80				1.449					
	1000	33				1.81					

Table 5.12. Risk estimations for the distance variables by cancer cause. Census tract level.

The following cancer types reproduce the associations shown in the previous analysis such as: stomach cancer, colon-rectum and bladder. The remaining cancer types have behaved differently. Lung and breast cancer show distance-trends that do not appear in the municipal level analysis.

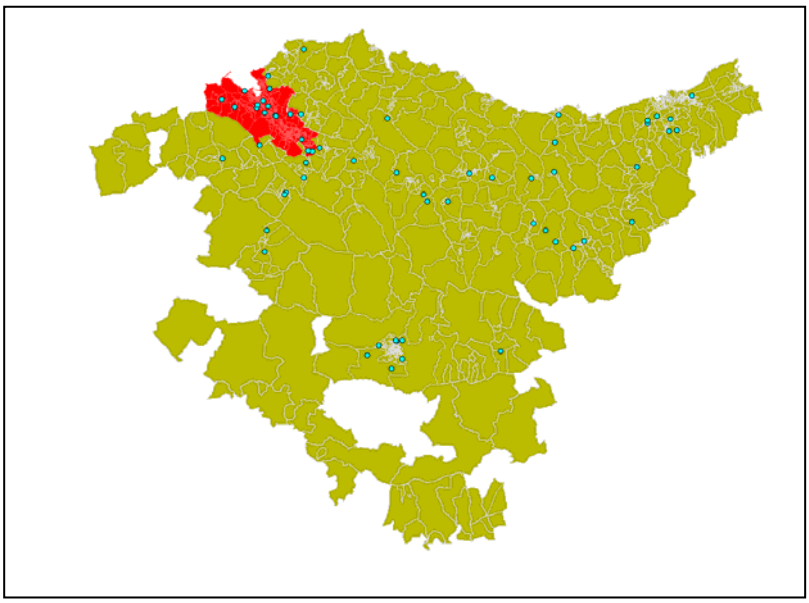
III. Conclusions of the exploratory analysis

Both previous analyses suggest that three cancer types, stomach, colon-rectum and bladder, have associations with some industrial sectors derived from the distance between the point source and the centroid of the area. Based on these results, these three causes seem suitable candidates for a deeper analysis.

The spatial distribution of bladder cancer mortality in Spain has been related to the exposure of industrial pollution in a recent study [Lopez-Abente et al., 2006a]. On the other hand, stomach and colon-rectum are part of the digestive system and tumours located in these organs seem to be more associated with diet.

We also reduce the area of study to focus on a multi-focus scenario. The new study area is the so-called “Gran Bilbao” region, which includes 15 municipalities, one of them is the city of Bilbao, and is divided into 657 census tracts. Its population is 906.222 inhabitants and the population density is 1.811,1 habitants per km^2. There are 20 industrial facilities either within the area or on its borders, half of them belong to the metal industry. (Map 5.3)

Accordingly, the next part of this report will study bladder cancer mortality. Also we are going to repeat this analysis over other two causes: haematological tumours and prostate cancer. In the literature it can be found studies where haematological tumours seem to be associated with exposure to industrial pollution [Parodi et al., 2003; Ramis et al., 2009]. Finally, in this previous analysis prostate cancer was the only tumours showing a distance-trend risk relative to the metal industry.



Map 5.3. *Basque Country by census tract. Gran Bilbao is the red area. Factories locations blue dots.*

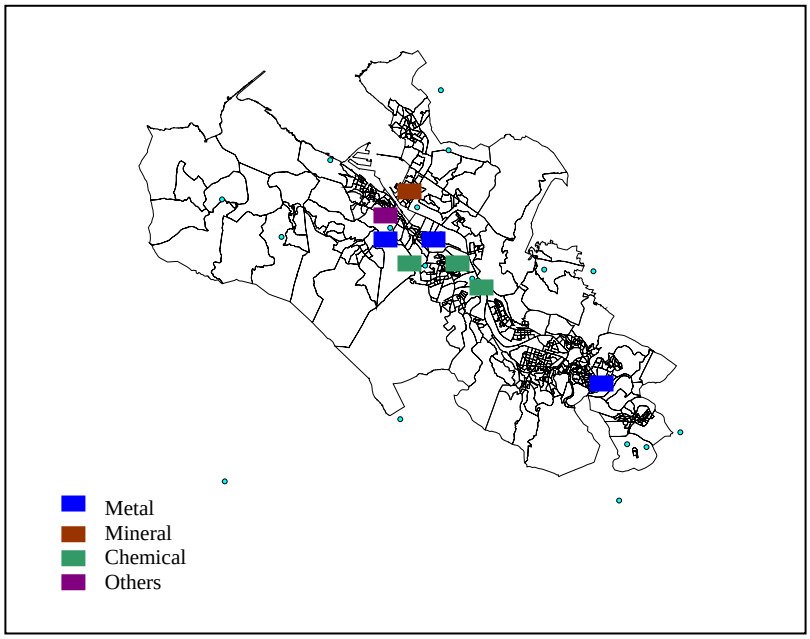
5.3 RESULTS. MORTALITY IN GRAND BILBAO

In the following models the level of aggregation of the data is census tracts; therefore, the analysis includes 657 areas and 20 industrial facilities.

Factories

As a source of information about the industrial facilities we used the European Pollutant Emission Register (EPER). The information available allows us to identify different types of industrial activities. The register presents the following 6 categories: 1) Energy industries; 2) Production and processing of metals; 3) Mineral industry; 4) Chemical industry and chemical installations; 5) Waste management; and 6) Other activities (which include paper and board production, manufacture of fibres or textiles, tanning of hides and skins, slaughterhouses, intensive poultry or pig rearing, installations using organic solvents, and the production of carbon or graphite). In the present study, we are working with eight industries located in central axis of the area; three metal factories, one mineral factory, three chemicals factories and one from group "others activities". For the following analyses, the exposure variables are defined as distance between the centroid of the census tract and the location of the factories. Moreover, we aggregate these variables in the four industrial categories: metal, mineral, chemical and other activities; in order to increasing the statistical power the data.

The following figure is a map of Gran Bilbao by census tract where locations of factories are represented by squares of different colours according to its industrial category,

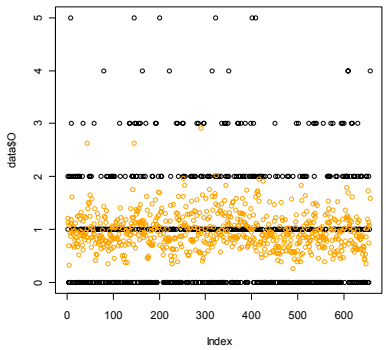


Map 5.4. Gran Bilbao by census tract. Factories by industrial categories in colour squares blue for metal industry, brown for mineral industry, green for chemical industry and purple for other activities. Remaining factories in blue dots.

5.3.1 Bladder cancer.

I. Descriptive analysis

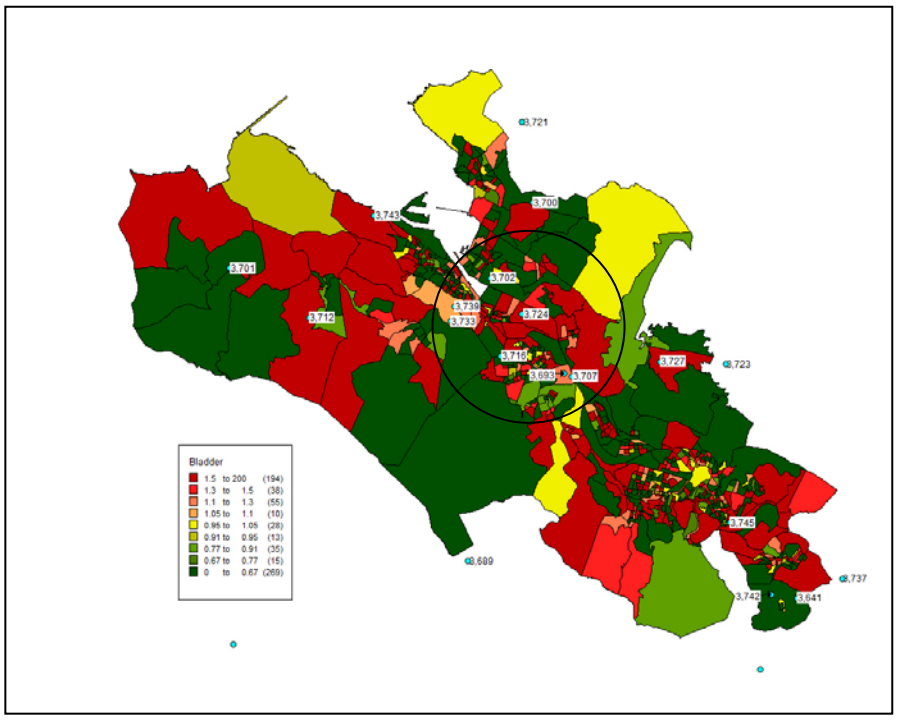
During the period under study there were 664 deaths caused by bladder cancer. The graph shows a scatter-plot of the observed cases (black dots) and the expected cases (orange dots).



On average there is one case per census tract, the median and the mean are 1 and 1.01 respectively. The value of the standard deviation is 1.05; consequently, there is no overdispersion in the data.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
0.000	0.000	1.000	1.011	2.000	5.000	1.051

The following map (Map 5.5) shows the distribution of the standard mortality ratio (observed/expected) all across the Gran Bilbao area. The industrial facilities are also shown. Finally, the circumference marks an area around the “3724” facility of 4 km of radius.



Map 5.5. *Standard Mortality Ratio of bladder cancer by census tract in Gran Bilbao*

On this occasion, the exposure to each pollutant focus is estimated by the distance between the centroid of the census tract and the location of the factory. With the aim of helping with the computation process the unit distance used in the models is 100km.

II. Regression Poisson. Covariates

Multiple regressions have been fitted to find the possible association between the socio-demographic covariates and the distribution of the bladder cancer mortality in Gran Bilbao. Three covariates yield a statistically significant relative risk: percentage of illiterates, income and standard mortality ratio of lung cancer (tobacco prevalence). In accordance with the results, the percentage of illiterates and level of tobacco prevalence are risk factors whilst income is a protective factor. The remaining socio-demographic covariates seem to be unrelated with bladder cancer mortality.

```
Model 1
glm(formula = O ~ offset(log(E)) + educ + income + lung, family = poisson)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.46410    0.20538   2.260 0.023841
educ         1.36290    0.39409   3.458 0.000544
income      -0.18414    0.10180  -1.809 0.070485
lung         0.11677    0.06809   1.715 0.086354
---
    Null deviance: 733.38  on 656  degrees of freedom
Residual deviance: 717.11  on 653  degrees of freedom
```

These three covariates are included in all the following analyses where we use the model (1) described in the methods section.

III. Model (1): individual regressions

Firstly, we have studied the 20 industrial locations one by one, adding to the spatial model the distance variable and the socio-demographic covariates. We have fitted an independent model for each point source. The results show that the deviances for these 20 models are very similar to the deviance of model 1 (717.11) and according to the likelihood test none of them is significantly better than the initial Poisson model. Results are included in the appendix.

IV. Model (1). Multiple regression inside the circumference

The second approach is done across the area inside the circle. The industrial facilities 3693, 3702, 3702, 3716, 3724, 7333 and 3739 are located within this area. However, there are two

pairs of factories very close to each other, 3693-3707 and 3722-3739; therefore, each pair is treated as just one pollutant focus. We have fitted a model that includes the five focuses. The statistic of the likelihood test has a value of 3.6726 and the 5% critical value of chi-square with 5 df is 11.07. These results suggest that the model with 5 focuses is not better than the model with just covariates. (More results are included in the appendix)

V. Model (1). Multiple regressions with 8 focus (Multiple focus scenario)

As a final step, we have studied the whole area of Gran Bilbao again, but this time the industrial facilities are grouped by industrial area (table 3). Moreover, only 8 of the industrial facilities, which are centred in the area, are introduced in the models in order to analyse the potentially most influential hazardous locations according to the distribution of the population. These eight factories belong to four different industrial areas:

- Metal: 3724, 3733, 3745
- Mineral: 3702
- Chemical: 3693, 3707, 3716
- Others: 3739

Several models are fitted. The variables introduced in each model are in Table 5.13.

	rho	Education (-)	Income	Tobacco	Metal 3f	Mineral 1f	Chemical 3f	Others 1f
Model 1	X	X	X	X				
Model 2	X	X	X	X	X			
Model 3	X	X	X	X	X	X		
Model 4	X	X	X	X	X		X	
Model 5	X	X	X	X	X	X	X	
Model 6	X	X	X	X	X	X	X	X

Table 5.13. Covariates introduced in the multiple regressions

Although, the null model is the one with no covariates, in this case we are going to consider model 1 as the reference model for the likelihood ratio tests. It can be seen in Table 5.14 that the deviances obtained from all the fitted models are just slightly smaller than the deviance of model 1 and the likelihoods are slightly bigger. As a result, none of the likelihood ratio tests are statistically significant.

As follows we give the results for different analysis performed over other two tumoural causes: haematological tumours and prostate cancer. A similar scheme has been developed.

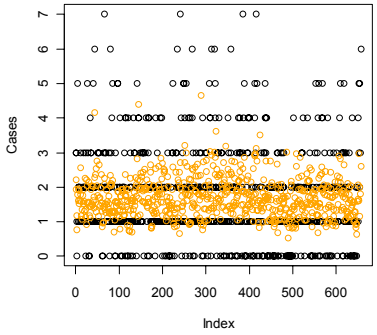
	Deviance	Likelihood	D Stat	P-value
Null	733.4			
Model 1	717.11	-612.81		
Model 2	716.02	-612.26	1.10	0.29
Model 3	715.90	-612.20	1.22	0.54
Model 4	715.65	-612.08	1.46	0.69
Model 5	715.77	-612.14	1.34	0.73
Model 6	715.81	-612.16	1.30	0.86

Table 5.14. Deviances, likelihood, D statistics and P- values for a χ^2_n for the multiple regressions

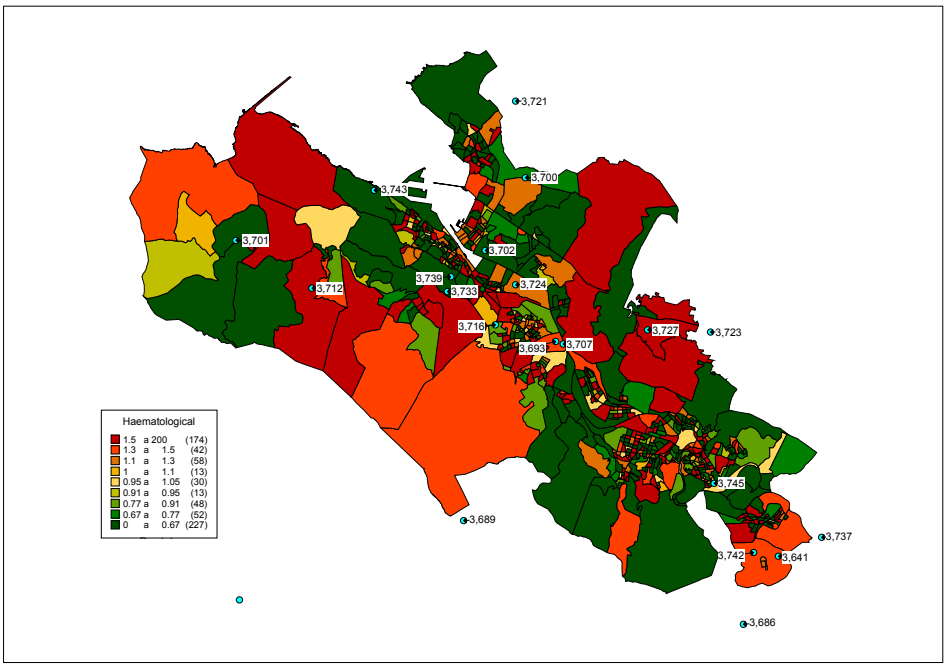
5.3.2 Haematological tumours

I. Descriptive analysis

The total number of deaths by haematological tumours during the period 1996-2003 in "Gran Bilbao" was 1175. On average, there were 1.788 per area.



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
0.000	1.000	2.000	1.788	3.000	7.000	1.4375



Map 5.6. Standard Mortality Ratio of haematological tumours by census tract in Gran Bilbao

II. Regression Poisson. Covariates

Initially, we fit multiple regressions with the socio-demographic covariates. Income, cohabitants per house (cph) and "tobacco" are risks for haematological tumours morality. (Model 1)

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.41733    0.13624   3.063  0.00219
Income       0.15131    0.08546   1.770  0.07664
cph         -0.37284    0.14641  -2.547  0.01088
lung        -0.09006    0.05250  -1.715  0.08630
---
    Null deviance: 714.43  on 656  degrees of freedom
Residual deviance: 705.44  on 653  degrees of freedom
Likelihood: -427.0215
```

III. Model (1). Multiple regressions with 8 focus

In a second analysis, we work over the *multiple focus scenario*, in other words, we work with the 8 factories sited within the region and all the census tracks of "Gran Bilbao". Two models with distance variables are built: model 2 and model 3. The following table (Table 5.15) shows the sequence of covariates introduced in each model.

	rho	Income	Cohab. house	Tobacco	Metal 3f	Mineral 1f	Chemical 3f	Others 1f
Model 1	X	X	X	X				
Model 2	X	X	X	X	X	X	X	
Model 3	X	X	X	X	X	X	X	X

Table 5.15. Covariates introduced in the multiple regressions

Table 5.16 gives the results for these models. The inclusion of the distance variables does not improve significantly the fitting of the data. Values of deviance and likelihood are very similar for the three models and p-values for the D statistic indicate non advance in models 2 and 3 with reference to the model 1.

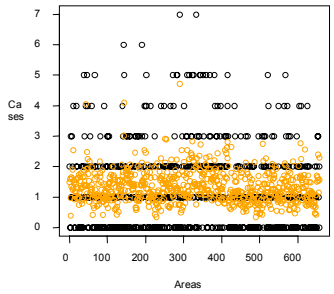
	Deviance	Likelihood	D Stat	P- value
Null	714.43			
Model 1	705.44	-427.02		
Model 2	704.38	-426.49	1.06	0.78
Model 3	704.19	-426.40	1.25	0.87

Table 5.16. Deviances, likelihood, D statistic and P-value for a χ^2_n for the multiple regressions

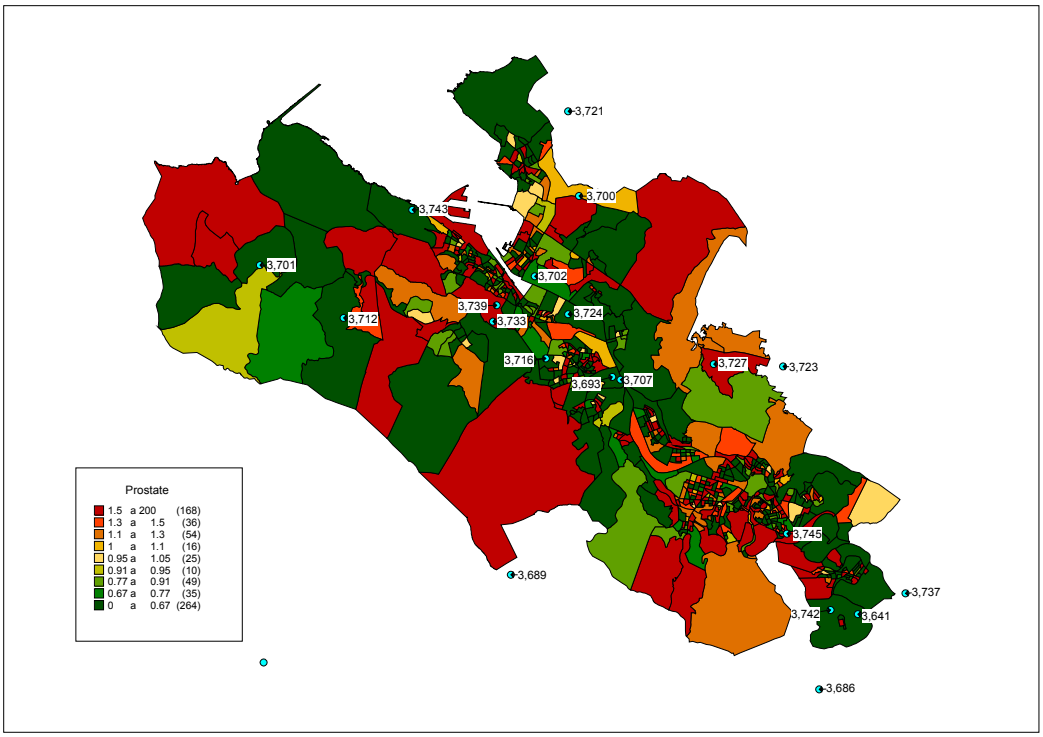
5.3.3 Prostate cancer

I. Descriptive analysis

There were a total of 883 deaths by prostate cancer in the region of "Gran Bilbao" during the period 1996-2003. On average, there were 1.344 cases per census track with a standard deviation of 1.36.



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	sd
0.000	0.000	1.000	1.344	2.000	7.000	1.36



Map 5.7. Standard Mortality Ratio of prostate cancer by census tract in Gran Bilbao

II. Regression Poisson. Covariates

As we have done with the previous cancer causes, we start the analysis by fitting a multiple Poisson regression. This regression yields the following estimators of risk for the five socio-demographic covariates:

```
glm(formula = O ~ offset(log(E)) + educ + unemploy + income + cph + lung, family
= poisson)

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.25224    0.60168  -0.419    0.675
educ        -0.04131    0.62761  -0.066    0.948
unemploy     0.15271    0.41977   0.364    0.716
income      -0.14905    0.19922  -0.748    0.454
cph         -0.16938    0.32768  -0.517    0.605
lung         0.06971    0.05965   1.169    0.243

    Null deviance: 818.04  on 656  degrees of freedom
Residual deviance: 807.51  on 651  degrees of freedom
Likelihood -537.8797
```

III. Model (1). Multiple regressions with 8 focus

We also repeat analysis over the *multiple focus scenario* defined before.

On this occasion the five socio-demographic covariates are confounders for the effect of the distance variables; consequently, we are going to use these five covariates in the following models.

Table 5.17 shows the sequence of fitted models and the different covariates included in each of them.

	rho	Education (-)	Unemployed	Income	Cohab. house	Tobacco	Metal 3f	Mineral 1f	Chemical 3f	Others 1f
Model 1	X	X	X	X	X	X				
Model 2	X	X	X	X	X	X	X			
Model 3	X	X	X	X	X	X	X	X		
Model 4	X	X	X	X	X	X	X		X	
Model 5	X	X	X	X	X	X	X	X	X	
Model 6	X	X	X	X	X	X	X	X	X	X

Table 5.17. Covariates introduced in the multiple regressions

Table 5.18 gives the values for the deviance and likelihood for the 6 models. The table shows also the P-values of the likelihood ratios test with model 1 as reference. There is evidence that the inclusion of the distance variables in model 6 is an improvement over model 1. The value of the D statistic is 10.42 with a P-value of 0.03. With regard to the remaining models, there seems to be no improvement in contrast to model 1 since their P-values are superior to 0.05.

	Deviance	Likelihood	D Stat	P-value
Null	818.04			
Model 1	807.51	-537.88		
Model 2	807.38	-537.81	0.13	0.72
Model 3	803.05	-535.65	4.46	0.10
Model 4	802.54	-535.39	4.97	0.17
Model 5	802.54	-535.37	5.03	0.17
Model 6	**797.09**	**-532.67**	**10.42**	**0.03**

Table 5.18. Deviances, likelihood, D statistic and P- value for a χ^2_n for the multiple regressions

As we just saw, only model 6 is significantly better than the model with just socio-demographic covariates, model 1.

Relative risks associated to each socio-demographic variables for models 6 are in Figure 5.3.

Covariate	rho	Education (-)	Unemployed	Income	Cohabitants house	Tobacco
Relative Risk	0.827	1.192	0.671	1.265	0.488	1.080

Industrial area	Distance functions
Metal	$1 + 0.4 \exp\left(-\left(dist\ /0.1\right)^2\right)$
Mineral	$1 + 0.42 \exp\left(-\left(dist\ /18.1\right)^2\right)$

Industrial area	Distance functions
Chemical	$1 - 0.18 \exp\left(-\left(dist\ /0.4\right)^2\right)$
Others	$1 + 13.2 \exp\left(-\left(dist\ /0.002\right)^2\right)$

Figure 5.3. Relative risk of the socio-demographic covariates and distance functions.

In concordance with these estimations, income, percentage of illiterates (education -) and RR of lung cancer (tobacco prevalence) are risk factors for prostate cancer mortality. On the other hand, percentage of unemployed and cohabitants per house have the opposite effect. Risks associated with the distance from the industrial factories are also presented in table 17, though the understanding of these mathematical functions is easier with the help of the graphs showed in Figure 5.4.

As follows, we have several graphs related to the results of this model. Firstly, we have graphs with the densities of the empirical distributions of the estimation of the parameters for the socio-demographic covariates and the intercept (Figure 5.6). These graphs show the shape of the densities and their location in reference to 1, even though, as we said before, all socio-demographic covariates are confounders for the distance variable effects. In general, all densities are rather symmetric with a sharp shape pointing the mean. As a particular case, the effect of tobacco shows a extremely sharp density in a narrow interval. For this reason, even if the relative risk associated with tobacco is not the largest, 8%, the estimation is the most

consistent.

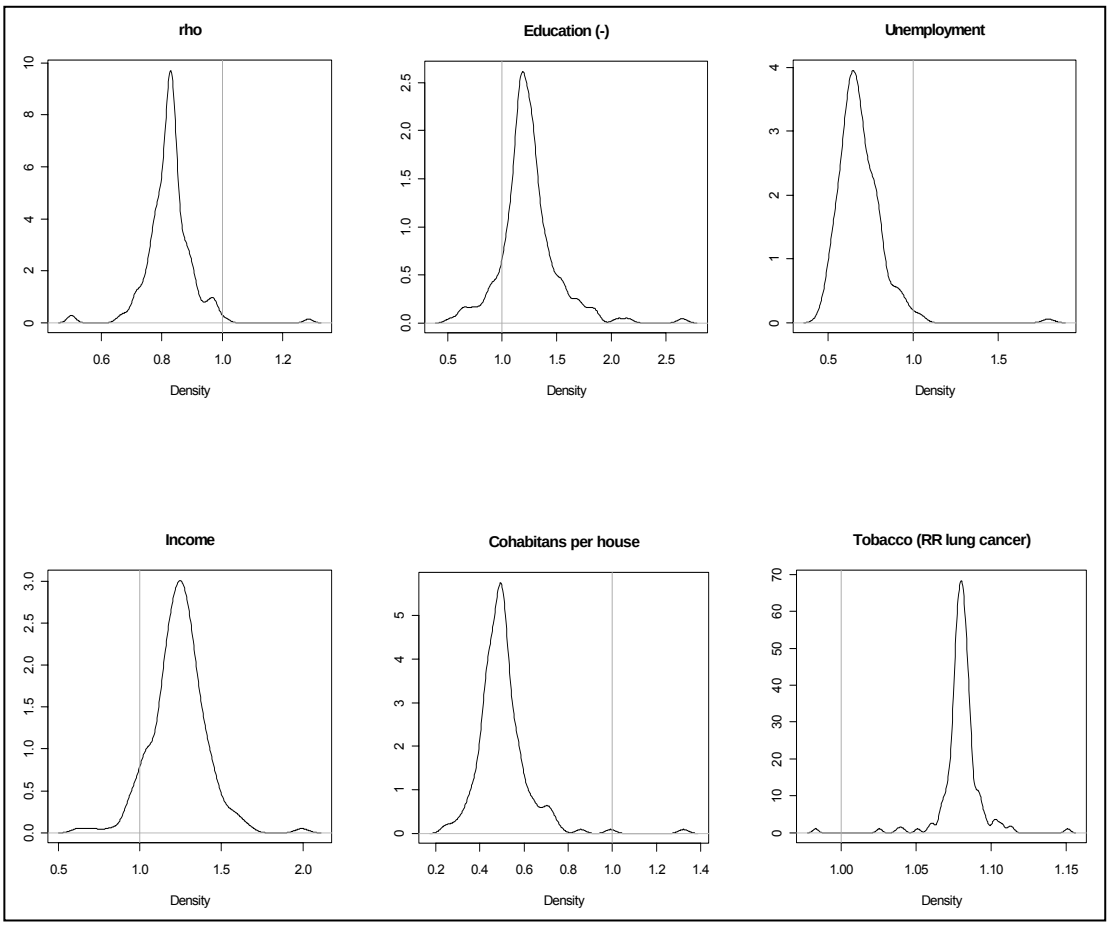


Figure 5.4. Densities of the empirical distributions of the estimation of the parameters

Secondly, the following graphs represent risk functions and confidence intervals at 95% linked to the distance from the factories of each industrial activity. As seen below, the graphs of each function have a different shape, meaning different risk effects. Risk related to metal industries has a starting value of 1.4 decaying with distance until 1.085 at 12 km, though the confident interval reaches the neutral value, 1, slightly above 10 km. The risk function associated with the mineral factory is constant and does not change with distance to the focus. Alternatively, the risk function linked to distance from chemicals factories has a positive slope with a risk of 0.83 at distance 0 increasing until 1 at 6 km. But the confident interval reaches the value 1 before 3 km. Finally, the risk function for “other kind of industries” is almost vertical dropping from 2.7 to 1 in 500m; however, the lower limit of the interval is 1 for the entire range.

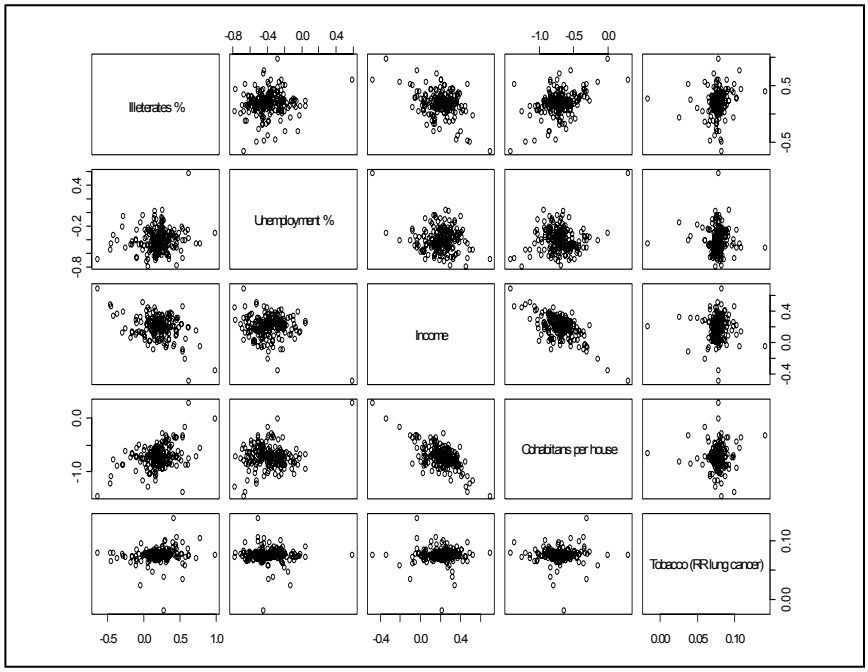


Figure 5.5. Pairwise scatterplot of the empirical distributions of the estimation of the parameters

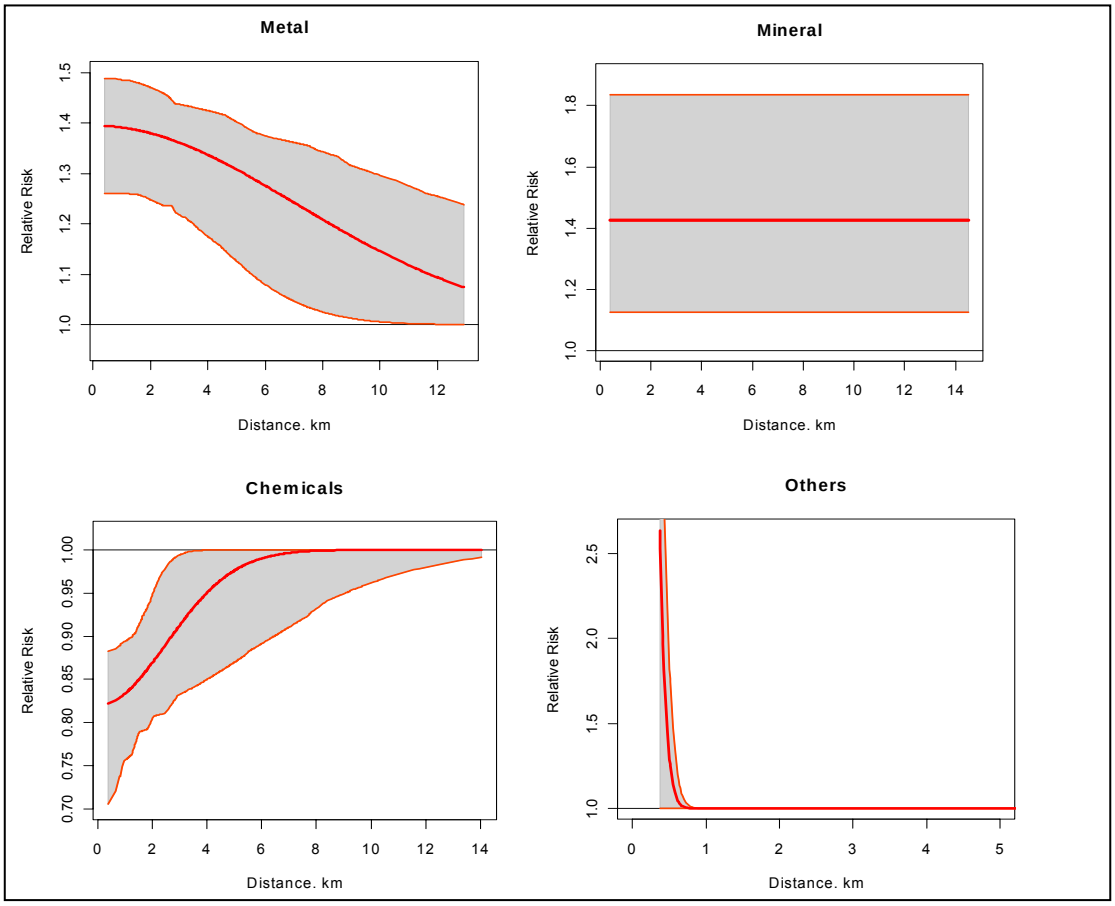


Figure 5.6. Risk functions and confident intervals (95%) for the distance to the factories by industrial area

Finally, we have graphs of the residuals in order to assess the goodness-of-fit of the model (Figure 5.7). The top graph shows deviance residuals against fitted values; the remaining graphs show deviance residuals against distances from the industries. In all the graphs a non spatial structure can be detected.

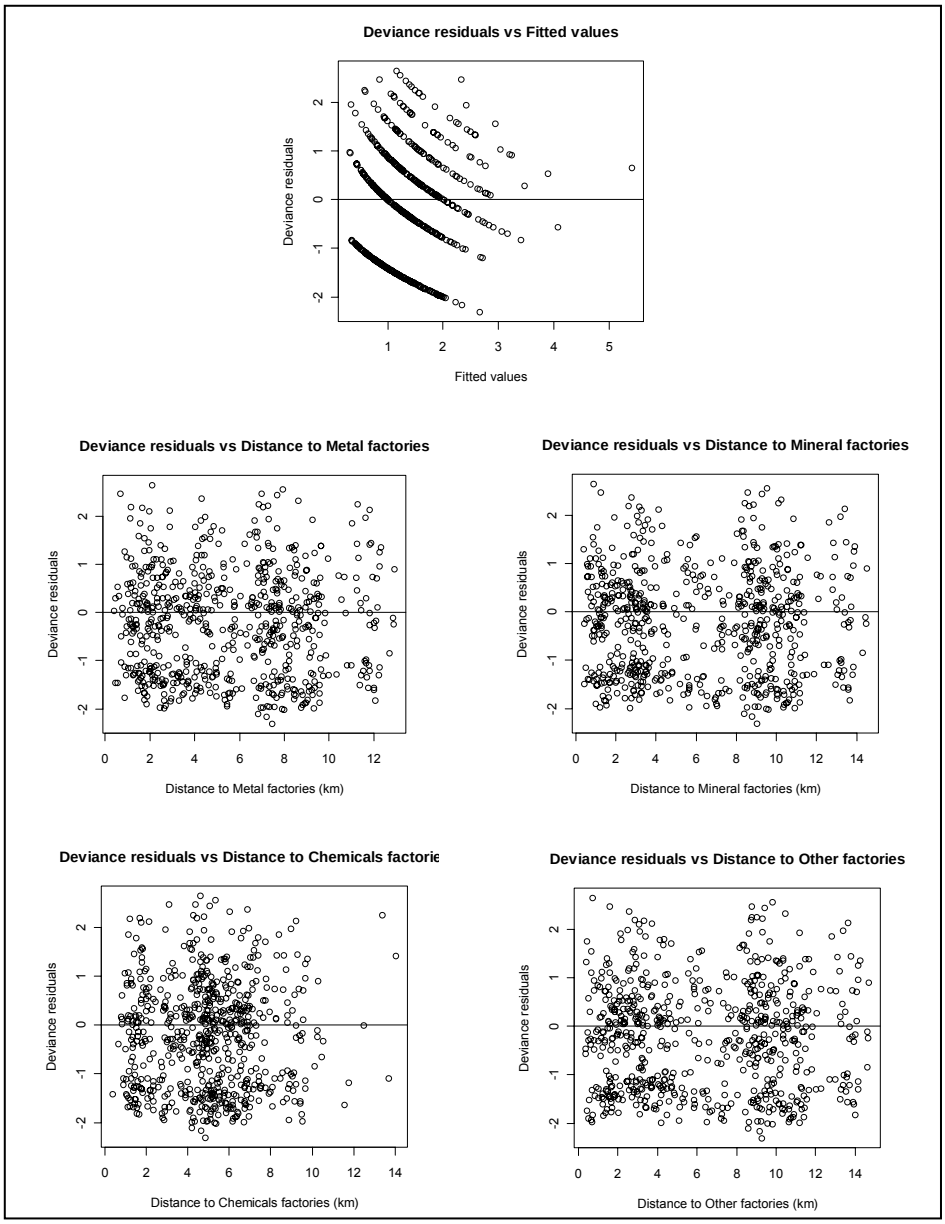


Figure 5.7. Graphs of residuals. Deviance residuals vs fitted values and distances.

5.4 DISCUSSION

In the empirical example, we have studied the distribution of bladder, haematological and prostate cancer mortality in the so-called Gran Bilbao area in relation to the exposure to pollutant substances emitted from the factories sited within the area, using the model described in the methods section. This model was initially developed by Diggle and Rowlingson to fit individual data; however, in this study we are using aggregated data.

In the analysis of the first data set, bladder cancer mortality, different approaches have been used in reference to the inclusion of factories and the extension of the area under study whilst searching for evidence. The final scenario includes the 8 factories sited in the central axis of the region, aggregated in four industrial categories, metal, mineral, chemical and other activities. The main result is that the model which better fits the available data is the one that only includes the socio-demographic covariates, income, cohabitants per house and education; in other words, relation between distance to industrial factories and bladder cancer mortality is not statistically significant. The remaining data sets, haematological tumours and prostate cancer mortality, have been studied in the last scenario only. For haematological tumours mortality no evidence of association with distance to factories has been found, moreover, as with the example of bladder cancer, some socio-demographic characteristics can be considered as risk factors, such as income, cohabitants per house and tobacco. Finally, results for prostate cancer mortality suggest an association with distance to the factories. The proposed model identifies different risk functions for the different activities. The metal industries function has decaying slope with starting risk value of 1.4 and reaches the neutral effect above 10km of distance.

This is one of the first studies analysing the relation between the spatial distribution of cancer mortality and the exposure to industrial pollution using aggregated data in Spain. Thus, it is important to discuss some conceptual and methodological issues.

Cancer is a complex disease and has many known and unknown risk factors [IARC, 2009b]. Environmental exposure could be one important factor, although there are many others involved. Lifestyle is the main factor, diet (30-35%), tobacco (25-30%) and obesity (10-20%). On the other hand infections (15-20%). And finally genetic predisposition (5-10%). However, the interaction between these factors is very important in the development of the disease [Anand et al., 2008]. In the present study only ecological data about the socio-demographic status of the population and estimation of the exposition to industrial pollution have been included, which means that important information is not being taken into account. For some tumoural locations, those with high survival rate, a weakness is the use of mortality data instead of incidence data. This implies that the data is biased because many cases of cancer are not taken into consideration for the study.

It is important to discuss the definition of distance when data are spatially aggregated whilst it can be introduced misclassification. In this work we have considered the centroid of the area as the reference point to calculate distances from sources. That decision may bias the results due to the use of centroids as co-ordinates to position an area's entire population, when, in reality, the population may be considerably dispersed. This classification error becomes much less important in smaller-sized areas.

There are others important assumptions linked to the use of aggregated data. Initial, we assume that the whole population within an area, municipality or census tract, lives in its centroid; even more, we assume that they have always been living there. Also, we do not consider the daily movement of the people to go to work or study, for instance. Hence, we are assuming that everybody within an area is exposed to the same type and amount of pollutant substances.

Finally, it should be mentioned that other sources of environmental pollution, such as traffic or indoor pollution, are not included in this study. Exposure to such pollution can contribute to the development of cancers [Belpomme et al., 2007a]. As example, substances such as polycyclic aromatic hydrocarbons produced by combustion of organic fuels are considered as mutagens [IARC (International Agency for Research on Cancer), 1989] and indoor pollutants as volatile organic compounds, benzene for instance, are rated as carcinogens [IARC (International Agency for Research on Cancer), 1995].

On the other hand, it should also be pointed out that the data referring to environmental industrial exposures was drawn from the first edition of the EPER. The quality of this information may conceivably improve with the new European Pollutant Release and Transfer Register (E-PRTR), which will completely replace the EPER in 2009, allowing enhancement of the validity of a study of this type, with the possibility of evaluating the effect of specific pollutants.

Conclusions

The proposed model is able to identify different risk functions associated with different focus when we work in a multiple focus scenario, using aggregated data in small areas.

We have found evidence of association linking the distribution of prostate cancer mortality aggregated by census tracts and exposure to pollutant substances from the metal industrial facilities located within the area; exposure estimated through the distance between the point source and the centroid of the census tract.

The socio-demographic characteristics of the population are related to many cancer causes, as the results for the previous analysis yield.

6. GENERAL DISCUSION AND FUTURE WORK

Spatial epidemiology

Before the general discussion we end this thesis with a summary of general conclusions for the three approached areas of the spatial epidemiology.

I. Disease mapping

Disease maps are the best method to represent any health event data in their geographical context when the aim is to summarize the variation of spatial distribution of diseases.

Disease mapping methods have proved to be an excellent instrument for the description of the spatial distribution of incidence or mortality rates. They are, as well, a helpful hypothesis generator, useful in the assessment of inequalities and the allocation of health care resources.

The most common summary risk measure represented in the disease maps is the standardised mortality or incidence ratio (SMR or SIR). However, the SMR and SIR are inconsistent estimations with high sampling variability when the aggregation unit is small, such as municipalities or census tract. Nevertheless, this variability can be reduced by the smoothing of the raw rates via hierarchical modelling giving the so-called smoothed relative risk. Thus, when the basic unit of aggregation is big, such as provinces or whole countries the use of SMR or SIR gives a accurate representation of the risk surface. However, when the basic aggregation unit is small the best estimated risk measure is the smoothed relative risk.

II. Ecological regression (Poisson regression)

Poisson regression is one of the basic tools applied to the analysis of the association between disease and potential risk factors. The main advantage of this analytical method is the increasing availability of information, both health and risk factors, at aggregated scale. However, we should not forget the necessity and relevance of ecological analyses when environmental factors and effects are under exploration, despite the fact that no cause-effect relationship conclusions can be accounted for. However, the main disadvantage of this method is the so-called “Ecological fallacy”, so non-individual level associations can be explained by ecological regression studies.

In general, in spite of its sometimes imputed weaknesses, the use of ecological regression is of major help to achieve an initial assessment of the presence or absence of association among the studied risk factors and the disease. Moreover, it allows to work out the relationships and interactions between different risk factors over a disease outcome when they are studied jointly.

III. Assessment of risk in relation to a point source

Point source studies can be applied to assess increases in incidence or mortality of diseases in adjoining populations of potential environmental hazards.

For this kind of studies the main difficulty is the measurement of the real exposure suffered by the population: consequently several strategies have been developed to cope with it. In many studies distance to the source is employed as a surrogate of the real exposure by defining a decay function of the risk as the distance increases. Different authors propose different methods to approach the definition of the decay function, although the accuracy of this approximation it has been widely disputed.

Lastly, ad-hoc studies regarding specific pollution sources are carried out when the media or the political authorities express concern in relation to the risk of pre-specified exposures. In some of these cases there is not a prior biological hypothesis what causes an especially complicated interpretation of the results.

6.1 DISCUSSION

Although each section has a discussion, below we are going to discuss briefly the common materials and methods.

6.1.1 Mortality data

It has already been mentioned several times throughout this thesis that mortality data have some weakness because they just include lethal cases of a disease. A better dataset would be one with all cases, mortality and incidence from tumoural registers. Unfortunately, as we have said before, nowadays there is not a nation wide cancer register in Spain. The lack of information about non-lethal cancer cases in the data set may bias the analysis for some tumoural locations, those with high survival rate, while, on the other hand, tumours with lower survival rates are well represented using death certificates. In Spain, quality of cancer death certificates was analysed by Pérez-Gómez and Aragonés in 2006 [Perez-Gomez et al., 2006]. Their main conclusions were: first, overall accuracy of cancer death certificates in Spain was comparable to that reported for other industrialised countries. Secondly, the accuracy contrasts by tumoural location, the main leading cancer sites were well certificated (i.e. lung, colon-rectum, female breast cancer, prostate, haematological..); however, less common locations are less accurate certified (i.e. larynx, bladder and ovary). According to this study our data are useful to analyse cancer mortality in Spain.

6.1.2 Population data

The municipal populations used for this study come from the 1996 Electoral Roll and 2001 Census. These two years correspond to the midway points of the two quinquennia that comprise the study period (1994-1998 and 1999-2003). To estimate the number of person-years these population have been multiplied by five. Census and electoral rolls are the most exhaustive source of information related to population; however, they are not completely reliable. Census data are collected for a snapshot in time, every decade, consequently they do not consider changes in population between census counts. Alternatively, in Spain electoral rolls are now continuous, updating every month, which can mean an advantage over census data but there are a percentage of the population who do not live in the same municipality where they are registered in; therefore, for some municipalities the electoral roll overestimates the population while in others it underestimates. In spite of these weaknesses the census and the electoral roll are the best source of information for population counts in Spain and they are the only ones available at municipal and census tract level.

6.1.3 Aggregated data

In this thesis we have used aggregated data for health events, population, socio demographical variables and industrial pollution exposure approximation. Many authors have considered the limitations of ecological studies in spatial epidemiology [Beale et al., 2008; Elliott et al., 2000; Lawson A, 2001]. First, when a study is based solely on aggregated data its results must not be interpreted at the individual level because they can suggest misleading conclusions about associations (problem known as ecological fallacy or ecological bias, [Selvin, 1958]). For instance, when we exploit socio demographical variables at area level we must be aware of the assumptions involved: we suppose the entire population in the same area has the same socio demographical characteristics. Nevertheless, using small area data reduces the ecological bias with more detailed information but by no means rules it out. Moreover, in small area studies local effects (e.g. pollution from local sources or local health experiences) can be assessed.

Alternatively, in reference to the estimation of the pollution's exposure, we assume that the whole population within an area lives in its centroid; even more, we assume that they have always been living there. Also, we do not consider the daily movement of the people to go to work or study, for instance. Hence, we are assuming that everybody within an area is exposed to the same type and amount of pollutant substances. However, exposure data at area level can be more accurate than the corresponding individual exposures [Richardson, 1992]. Furthermore, in term of risk estimates, for certain exposure measures misclassification at ecological level is less important that misclassification at individual level [Armstrong B, 2004].

In this regard, recent studies have started to consider the daily mobility of the population with the intention of reducing this misclassification. Specify, one of these studies has included the daily mobility by taking into account where people live and work [Jerrett et al., 2005].

6.1.4 Pollution data.

Distance as a proxy of exposure

Sections 4 and 5 already discuss this point; however, a more general discussion can be done. The lack of real exposure measures to harmful pollutant substances released from industrial facilities hinders the study of their potential effect on health. Scientists and researchers have developed different strategies to deal with this problem and during the last few years several methodologies have been presented (in the introduction to section 5 some of these strategies are mentioned). When exposure is estimated by the distance to the focus many assumptions ought to be considered and the results must be carefully interpreted. Furthermore, cause-effect associations can not be concluded, although the results may point to an unknown environmental health problem, supporting or rejecting a previous hypothesis. Additionally studies that use these kind of proxies should be the first approach to deeper analyse exposure to specific pollutants and health problems.

Data source. EPER

The first data published from the EPER, corresponding to 2001, included 1,437 companies. Those installations had reported pollutant emissions excess over the established thresholds for one or more of the pollutants listed in European Union Decision 2000/479/CE. This first list had several weaknesses [Garcia-Perez et al., 2008] and was unreliable in reference to the amount of substances released; however, it has enabled us to locate the most pollutant industrial facilities and to study the distribution of cancer in their vicinity.

Since 2008 the EPER has been replaced by the European Pollutant Release and Transfer Register (E-PRTR), which includes more comprehensive information on industrial pollution from 91 substances and 65 industrial activities and, besides, it is compulsory.

6.1.5 Socio demographical variables

Aims of this thesis do not comprise the study of the associations between cancer mortality and the socio-demographic characteristics of the population. However, those factors are important in the cancer aetiology. Furthermore, areas with more exposure to pollution are generally the areas with high poverty rates, thus both factors should be studied together even when we are interested primarily in one of them. In a 2005 study on mortality in small areas associated with air pollution and social-demographic variables, it was shown that some socio-demographic

variables were confounded by pollution, specifically the poverty effect on mortality was reduced by 50% by the inclusion of pollution as a covariate [Jerrett et al., 2005]. Hence, this relationship between the social-demographic characteristics and exposure to pollution should always be integrated in the models to avoid misleading conclusions in the analyses.

Data source

Information about the socio-demographic characteristics comes from the 1991 census. We have already discussed the validity of census data and its advantages and weaknesses. For the purpose of this study the 1991 census was the best data source available.

6.1.6 Bayesian inference versus classic inference.

Throughout this thesis we have applied several statistical models to estimate their parameters. For the third section three Bayesian models have been assessed. In the fourth section both Bayesian and classical inference are performed. And, finally, the proposed model of the third section is fitted by classical maximum likelihood estimation.

The use of Bayesian or classical statistical estimation has depended on the convenience of the method more than the preference for one of them over the other.

6.1.7 Summary of the methodology

The main goal of this thesis has been the setting of a methodology to study the spatial distribution of health events and its relation to environmental factors, from large disease maps for a whole country to clustering analysis focused in small areas. To achieve this main objective three steps have been taken. First, we have explored the performance of different methods for disease mapping based on Poisson models seeking to describe spatial patterns in the distribution of the disease. In particular, three Bayesian hierarchical models for relative risk smoothing have been assessed: the Besag, York and Mollié model; a model based on zero-inflated Poisson (ZIP) distribution, which allowed a large number of event-free areas; and a mixture of distributions that enabled discontinuities (jumps in the pattern) to be modelled. The major characteristic of these methods is the use of the CAR distribution to include the spatial autocorrelation in the model to create a interpretable risk surface.

In a second step we have sought to analyse the association between the spatial disease patterns and the exposure to industrial pollution. Again, we have used three models of ecological regression to estimate the relative risk associated with the proximity to pollutant emitting factories: Poisson Regression; mixed Poisson model with random provincial effect; and spatial autoregressive modelling (BYM model). We have classified as exposed populations those having an industry within a radius of 1, 1.5, or 2 kilometres from the municipal centroid

and as reference populations those outside those radii. To analyse particular harms related to different industrial activities we have aggregated the facilities by sectors.

Finally, the last step has been to study in depth the effect on public health of industrial air pollutants released from the different facilities sited within an urban area. For this purpose we have applied an unique model that included all the factories under study and aggregated health data in small areas. Due to the lack of real exposure measures we have approximated it by using the distance between the focus and the areas' centroid. The model is able to capture a risk increase around the factories and a risk decay in long distances. As above a Poisson regression is used as a basic model and is extended with a non-linear term to model that risk decay; distance's function. This distance function has two parameters, the first one is the risk at focus and the second is the decay parameter.

In summary, this thesis should provide environmental epidemiologists and other researchers who are unfamiliar with techniques of spatial analysis of environmental factors the tools for defining an appropriated methodology to approach these kind of studies.

6.2 FUTURE WORK

We conclude this thesis with a brief account of areas of future work that are related to the three proposed objectives.

Objective 1: An interesting improvement to the study of the spatial distribution of the risk would be its longitudinal analysis by adding the time effect as factor and converting the spatial model in a spatio-temporal model. This advance would contribute in the following points:

1. Establishment of the temporal pattern of the disease.
2. Study of the temporal persistence of patterns and its association with steady risk factors, such as environmental conditions, welfare services, etc.
3. Detection of unusual spatio-temporal patterns by the insertion of the interaction effect when the pattern is linked to short-term environmental harms or changes in the data collection, for instance.
4. Improvement in the epidemiological interpretation of the risk patterns.

The functional form of this spatio-temporal model could be:

$$O_i \sim Po\left(\mu_{it} = E_{ti}\lambda_{it}\right)$$

$$\log(\lambda_{it}) = \alpha + \vartheta_i + \xi_t + \nu_{it}$$

Baseline risk — Spatial effect — Temporal trend — Spatio-temporal interaction

Recently, some authors have approached the spatio-temporal analysis proposing different methodologies. Martinez-Beneito considers the inclusion of the temporal trend in the classical model of Besag, York and Mollié [Martinez-Beneito et al., 2008]. Richardson presents a Bayesian spatio-temporal analysis of joint patterns of two diseases [Richardson et al., 2006]. Consequently, the temporal extension of the model could follow any of these methodologies.

Objective 2: As we have seen throughout this thesis, in many situations the availability of data about risk factors is insufficient, so new strategies to assess these factors are needed. With this aim, several studies have used the rates of a disease with well established risk factors to analyse the influence of those risk factors over a second disease [Best and Hansell, 2009; Lopez-Abente et al., 2006a; Dabney and Wakefield, 2005; Dabney and Wakefield, 2005; Held et al., 2005].

Therefore, another possible extension to this thesis is the joint study of several cancer locations seeking for environmental and socio-demographic common risk factors, in the spatial framework. We would use the recently published works of cluster analysis that use generalized linear models [Jung, 2009; Zhang and Lin, 2009] together with the joint diseases analysis, searching for a methodology able to identify the aggregation of areas with high risk for those different diseases and similar values for the socio-demographic covariates.

With this model we would be able to locate the high risk areas shared for the different diseases and to identify common risk factors.

Objective 3: Finally, a motivating extension for Section 3 would be the expansion of the proposed model with spatial autocorrelation effects. This extension could be done by including the spatial contiguity effects in the multi-focus model. In other words, the distance function would be introduced in the BYM model as a new term.

$$O_i \sim Po(E_i\mu_i)$$

$$\mu_i = \rho \exp\left[\sum_k (\vartheta_k Z_{ik}) + h_i + b_i\right] \prod_j f\left(d_{ij}\right); \qquad f\left(d_{ij}\right) = 1 + \alpha_j \exp\left[-\left(d_{ij} / \beta_j\right)^2\right]$$

However, this extension would transform the model into a hierarchical Bayesian model.

$$h_i \sim Normal(\mu, \tau_h)$$
$$b_i \sim Car.Normal(\eta_i, \tau_b)$$

$$\tau_h \sim Gamma(\alpha, \beta)$$
$$\tau_b \sim Gamma(\gamma, \delta)$$

These spatial autocorrelation terms would work in the model as a surrogate of unmeasured confounders with spatial behaviour, which would contribute to improving the interpretation of the relative risk associated to the environmental exposure from the sources.

Finally, we would like to mention that spatial epidemiology is a growing interdisciplinary area where new methods are developed by methodologists and statisticians in collaboration with epidemiologists and other specialists. Health politicians and administrators should, then, make good use of them.

7. CONCLUSIONS

Conclusions objective 1

1. The three assessed models generate a very similar geographical pattern for the distribution of haematological tumours.

2. The model that seeks to remedy the excess of zeros (ZIP), display a pattern that is almost identical to the classic BYM model, suggesting that ZIP model does not improve substantially the performance of BYM model when it tries to differentiate between areas with no cases and areas with cases.

3. The goodness-of-fit criteria points as the best model the one proposed by Lawson; however, the choice of one or another probably has scant practical consequences.

4. The different Bayesian models used furnished some very similar results. The high frequency of areas without cases would not seem to pose a serious difficulty to fitting these models, at least in the studied causes, haematological tumours.

Conclusions objective 2

1. The results suggest a possible increased risk of NHL mortality among populations residing in the vicinity of paper and pulp industries, an excess of mortality that is observable using different models.

2. The three different approaches produce similar results. Therefore, the decision to apply the spatial model in exploratory studies of this magnitude must be carefully evaluated due to the excessive time of computation.

3. Distance as a surrogate of the real exposure helps researchers to identify possible harmful industrial sectors when there are not direct measures; however, it has many weaknesses.

Conclusions objective 3

1. The proposed model is able to identify different risk functions associated with different focus when we work in a multiple focus scenario, using aggregated data in small areas.

2. The distance function is a useful estimation of the risk because it allows to know the risk in the focus and the decay with the distance. Therefore, the most influential focuses can be identified.

3. We have found evidence of statistical association linking the distribution of prostate cancer mortality aggregated by census tracts and exposure to pollutant substances from metal industrial facilities located within the studied area; exposure estimated through the distance between the point source and the centroid of the census tract.

4. Our exploratory analysis suggests that the socio-demographic characteristics of the population are related to many cancer causes.

8. BIBLIOGRAPHY

1. Special Issue of Statistical Methods in Medical Research on disease mapping. Statistical Methods in Medical Research 14. 2005. Ref Type: Journal (Full)

2. Adami H, Hunter D, Trichopoulos D (2002) *Text book of cancer epidemiology*. Oxford University Press, New York

3. Anand P, Kunnumakkara AB, Sundaram C, Harikumar KB, Tharakan ST, Lai OS, Sung B, Aggarwal BB (2008) Cancer is a preventable disease that requires major lifestyle changes. *Pharm Res* 25: 2097-2116

4. Armstrong B (2004) Exposure measurement error: consequences and design issues. In *Exposure Assessment in Occupational and Environmental Epidemiology*, pp 181-200. Oxford University Press: Oxford
5. Banco Español de Credito. Anuario del Mercado Español. 1993. Madrid, Banco Español de Credito. Ref Type: Report
6. Barcelo MA, Saez M, Cano-Serral G, Martinez-Beneito MA, Martinez JM, Borrell C, Ocana-Riola R, Montoya I, Calvo M, Lopez-Abente G, Rodriguez-Sanz M, Toro S, Alcala JT, Saurina C, Sanchez-Villegas P, Figueiras A (2008) [Methods to smooth mortality indicators: application to analysis of inequalities in mortality in Spanish cities [the MEDEA Project]]. *Gac Sanit* 22: 596-608

7. Beale L, Abellan JJ, Hodgson S, Jarup L (2008) Methodologic issues and approaches to spatial epidemiology. *Environ Health Perspect* 116: 1105-1110

8. Belpomme D, Irigaray P, Hardell L, Clapp R, Montagnier L, Epstein S, Sasco AJ (2007a) The multitude and diversity of environmental carcinogens. *Environ Res* 105: 414-429

9. Belpomme D, Irigaray P, Sasco AJ, Newby JA, Howard V, Clapp R, Hardell L (2007b) The growing incidence of cancer: role of lifestyle and screening detection (Review). *Int J Oncol* 30: 1037-1049

10. Benach J, Yasui Y, Borrell C, Rosa E, Pasarin MI, Benach N, Español E, Martinez JM, Daponte A (2001) *Atlas de mortalidad en areas pequeñas en España. 1985-1995*. Pompeu Fabra University: Tarrasa

11. Benach J, Yasui Y, Martinez JM, Borrell C, Pasarin MI, Daponte A (2004) The geography of the highest mortality areas in Spain: a striking cluster in the southwestern region of the country. *Occupational and Environmental Medicine* 61: 280-281

12. Benedetti M, Lavarone I, Comba P (2001) Cancer risk associated with residential proximity to industrial sites: A review. *Archives of Environmental Health* 56: 342-349

13. Besag J, York J, Mollià A (1991) Bayesian image restoration, with applications in spatial statistics. *AISM* 43: 1-59

14. Best N, Hansell AL (2009) Geographic variations in risk: adjusting for unmeasured confounders through joint modeling of multiple diseases. *Epidemiology* 20: 400-410

15. Biggeri A, Barbone F, Lagazio C, Bovenzi M, Stanta G (1996) Air pollution and lung cancer in Trieste, Italy: spatial analysis of risk as a function of distance from sources. *Environ Health Perspect* 104: 750-754

16. Blair A, Hartge P, Stewart PA, McAdams M, Lubin J (1998) Mortality and cancer incidence of aircraft maintenance workers exposed to trichloroethylene and other organic solvents and chemicals: extended follow up. *Occup Environ Med* 55: 161-171

17. Blair A, Linos A, Stewart PA, Burmeister LF, Gibson R, Everett G, Schuman L, Cantor KP (1993) Evaluation of risks for non-Hodgkin's lymphoma by occupation and industry exposures from a case-control study. *Am J Ind Med* 23: 301-312

18. Boyle P, Smans M (2008) *Atlas fo cancer mortality in the European Union and the Europan Economic Area. 1993-1997*. IARC: Lyon

19. Burnett C, Robinson C, Walker J (1999) Cancer mortality in health and science technicians. *Am J Ind Med* 36: 155-158

20. Capel H (1967) Los estudios acerca de las migraciones interiores en España. *Revista de Geografía, Universidad de Barcelona* 1: 77-101

21. Clayton D, Bernardinelli L, Montomoli C (1993) Spatial Correlation in Ecological Analysis. *International Journal of Epidemiology* 22: 1193-1202

22. Clayton D, Kaldor J (1987) Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics* 43: 671-681

23. Cole P, Trichopoulos D, Pastides H, Starr T, Mandel JS (2003) Dioxin and cancer: a critical review. *Regul Toxicol Pharmacol* 38: 378-388

24. COMMISSION OF THE EUROPEAN COMMUNITIES. COMMISSION DECISION of 17 July 2000 on the implementation of a European pollutant emission register (EPER) according to Article 15 of Council Directive 96/61/EC concerning integrated pollution

prevention and control (IPPC). 36-46. 2000. Official Journal of the European Communities. 192. Ref Type: Bill/Resolution

25. Congdon P (2001) *Bayesian statistical modeling*. Wiley: West Sussex

26. Cressie N (2000) Geostatistical methods for mapping environmental exposures. Elliott P WJCBNGBDJ (ed) Oxford Medical Publications: Oxford

27. Dabney AR, Wakefield JC (2005) Issues in the mapping of two diseases. *Statistical Methods in Medical Research* 14: 83-112

28. Diggle PJ (1983) *Statistical Analysis of Spatial Point Patterns*. Academic Press: London

29. Diggle PJ, Elliott P, Morris S, Shaddick G (1997) Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A* 160: 491-505

30. Diggle PJ, Rowlingson B (1994) A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A* 157: 433-440

31. Diggle PJ (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *J R Statist Soc A* 153: 349-362

32. Doll R (1991) Urban and rural factors in the aetiology of cancer. *Int J Cancer* 47: 803-810

33. Draper G, Vincent T, Kroll ME, Swanson J (2005) Childhood cancer in relation to distance from high voltage power lines in England and Wales: a case-control study. *BMJ* 330: 1290

34. Dreassi E, Lagazio C, Maule MM, Magnani C, Biggeri A (2008) Sensitivity analysis of the relationship between disease occurrence and distance from a putative source of pollution. *Geospat Health* 2: 263-271

35. Duczmal L, Kulddorff M, Huang L (2005) Evaluation of spatial scan statistics for irregular shaped clusters. *Journal ofComputational and Graphical Statistics* 15: 1-15

36. Durham CA, Pardoe L, Vega H (2004) A methodology for evaluating how product characteristics impact choice in retail settings with many zero observations: An application to restaurant wine purchase. *Journal of Agricultural and Resource Economics* 29: 112-131

37. Elliott P, Hills M, Beresford J, Kleinschmidt I, Jolley D, Pattenden S, Rodrigues L, Westlake A, Rose G (1992) Incidence of cancers of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet* 339: 854-858

38. Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J, Grundy C (1996) Cancer incidence near municipal solid waste incinerators in Great Britain. *Br J Cancer* 73: 702-710

39. Elliott P, Wakefield JC, Best N, Briggs D (2000) *Spatial epidemiology: Methods and Applications*. Oxford Medical Publications: Oxford

40. Eltom MA, Jemal A, Mbulaiteye SM, Devesa SS, Biggar RJ (2002) Trends in Kaposi's sarcoma and non-Hodgkin's lymphoma incidence in the United States from 1973 through 1998. *J Natl Cancer Inst* 94: 1204-1210

41. EPER. European Pollutant Emission Register (EPER). European Union . 2004. Ref Type: Electronic Citation

42. EPIYMPH. EPILYMPH Project. Environmental exposures and lymphoid neoplasms. EPILYMPH Project . 2007. Ref Type: Electronic Citation

43. Ewings PD, Bowie C, Phillips MJ, Johnson SA (1989) Incidence of leukaemia in young people in the vicinity of Hinkley Point nuclear power station, 1959-86. *BMJ* 299: 289-293

44. Ferlay J, Autier P, Boniol M, Heanue M, Colombet M, Boyle P (2007) Estimates of the cancer incidence and mortality in Europe in 2006. *Ann Oncol* 18: 581-592

45. Fernandez C, Green PJ (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 64: 805-826

46. Ferrandiz J, Abellan JJ, Lopez-Quilez A (2002) Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain). In *GIS for Emergency Preparedness and Health Risk Reduction*, Briggs D, Forer P, Jarup L, Stewart PA (eds) pp 267-282. Dordrecht

47. Fisher S, Fisher R (2004) The epidemiology of non-Hodgkin lymphoma. *Oncogene* 23: 6524-6534

48. Floret N, Mauny F, Challier B, Arveux P, Cahn JY, Viel JF (2003) Dioxin emissions from a solid waste incinerator and risk of non-Hodgkin lymphoma. *Epidemiology* 14: 392-398

49. Garabrant DH, Philbert MA (2002) Review of 2,4-dichlorophenoxyacetic acid (2,4-D) epidemiology and toxicology. *Crit Rev Toxicol* 32: 233-257

50. Garcia-Perez J, Boldo E, Ramis R, Pollan M, Perez-Gomez B, Aragones N, Lopez-Abente G (2007) Description of industrial pollution in Spain. *BMC Public Health* 7: 40

51. Garcia-Perez J, Boldo E, Ramis R, Vidal E, Aragones N, Perez-Gomez B, Pollan M, Lopez-Abente G (2008) Validation of the geographic position of EPER-Spain industries. *Int J Health Geogr* 7: 1

52. Garcia-Perez J, Pollan M, Boldo E, Perez-Gomez B, Aragones N, Lope V, Ramis R, Vidal E, Lopez-Abente G (2009) Mortality due to lung, laryngeal and bladder cancer in towns lying in the vicinity of combustion installations. *Sci Total Environ* 407: 2593-2602

53. Gardner MJ, Winter PD (1984) Mortality in Cumberland during 1959-78 with reference to cancer in young people around Windscale. *Lancet* 1: 216-217

54. Gilks W, Richardson S, Spiegelhalter DJ (1996) *Markov Chain Monte Carlo in Practice*. Chapman Hall: London

55. Godenau, D and Arteaga, S. Fiabilidad de las cifras censales y padronales en Canarias. IX Congreso de Población Española. Granada. 2004. Granada.

56. Gottlieb MS, Carr JK (1982) Case-control cancer mortality study and chlorination of drinking water in Louisiana. *Environ Health Perspect* 46: 169-177

57. Greaves MF (1997) Aetiology of acute leukaemia. *Lancet* 349: 344-349

58. Hall DB (2000) Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* 56: 1030-1039

59. Hardell L, Lindstrom G, van Bavel B, Fredrikson M, Liljegren G (1998) Some aspects of the etiology of non-Hodgkin's lymphoma. *Environ Health Perspect* 106 Suppl 2: 679-681

60. Hayes RB, Yin SN, Dosemeci M, Li GL, Wacholder S, Travis LB, Li CY, Rothman N, Hoover RN, Linet MS (1997) Benzene and the dose-related incidence of hematologic neoplasms in China. Chinese Academy of Preventive Medicine--National Cancer Institute Benzene Study Group. *J Natl Cancer Inst* 89: 1065-1071

61. Held L, Natario I, Fenton SE, Rue H, Becker N (2005) Towards joint disease mapping. *Statistical Methods in Medical Research* 14: 61-82

62. Huang L, Pickle LW, Das B (2008) Evaluating spatial methods for investigating global clustering and cluster detection of cancer cases. *Stat Med* 27: 5111-5142

63. IARC. Globocan 2002. http://www-dep.iarc.fr/ . 2005. Ref Type: Electronic Citation

64. IARC. European Cancer Observatory. IARC. http://ec.europa.eu/health-eu/health_problems/cancer/index_en.htm . 2009a. Ref Type: Electronic Citation

65. IARC. Monographs on the Evaluation of Carcinogenic Risks to Humans. IARC . 2009b. Ref Type: Electronic Citation

66. IARC (International Agency for Research on Cancer) (1989) *IARC Monographs on the Evaluation of Carcinogenic Risk to Humans. Occupational Exposures in Petroleum Refining; Crude Oil and Major Petroleum Fuels*. IARC Press: Lyon

67. IARC (International Agency for Research on Cancer) (1995) *IARC Monographs on the Evaluation of Carcinogenic Risk to Humans. Wood dust and Formaldehyde*. IARC Press: Lyon

68. Jerrett M, Buzzelli M, Burnett RT, DeLuca PF (2005) Particulate air pollution, social confounders, and mortality in small areas of an industrial city. *Soc Sci Med* 60: 2845-2863

69. Johnson KC, Pan S, Fry R, Mao Y (2003) Residential proximity to industrial plants and non-Hodgkin lymphoma. *Epidemiology* 14: 687-693

70. Jung I (2009) A generalized linear models approach to spatial scan statistics for covariate adjustment. *Stat Med* 28: 1131-1143

71. Kinlen LJ (1996) Epidemiological evidence for an infective basis in childhood leukaemia. *Journal of the Royal Society of Health* 116: 393-399

72. Kinlen LJ, Dickson M, Stiller CA (1995) Childhood leukaemia and non-Hodgkin's lymphoma near large rural construction sites, with a comparison with Sellafield nuclear site. *BMJ* 310: 763-768

73. Kinlen LJ, Petridou E (1995) Childhood leukemia and rural population movements: Greece, Italy, and other countries. *Cancer Causes Control* 6: 445-450

74. Kokki, E. Spatial small area analyses of disease risk around sources of environmental pollution: modelling tools for a system using high resolution register data. 2004. University of Jyväskalä. Ref Type: Thesis/Dissertation

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

76. Kulldorff M (2006) Tests of spatial randomness adjusted for an inhomogeneity: a general framework. *Journal of theAmerican Statistical Association* 101(475): 1289-1305

77. Lambert D (1992) Zero-Inflated Poisson Regression, with An Application to Defects in Manufacturing. *Technometrics* 34: 1-14

78. Lawson A (2001) *Statistical methods in spatial epidemiology*. Wiley: Chichester

79. Lawson A (1999) *Disease mapping and risk assessment for public health*. Wiley: Chichester

80. Lawson A (2005) Editorial: SMMR special issue on disease mapping. *Statistical Methods in Medical Research* 14: 1-2

81. Lawson A, Clark A (2002) Spatial mixture relative risk models applied to disease mapping. *Stat Med* 21: 359-370

82. Lee WJ, Teschke K, Kauppinen T, Andersen A, Jappinen P, Szadkowska-Stanczyk I, Pearce N, Persson B, Bergeret A, Facchini LA, Kishi R, Kielkowski D, Rix BA, Henneberger P, Sunyer J, Colin D, Kogevinas M, Boffetta P (2002) Mortality from lung cancer in workers exposed to sulfur dioxide in the pulp and paper industry. *Environ Health Perspect* 110: 991-995

83. Linsey J (2001) *Nonlinear Models in Medical Statistics*. Oxford University Press: New York

84. Lopez-Abente G, Aragones N, Pollan M, Ruiz M, Gandarillas A (1999) Leukemia, lymphomas, and myeloma mortality in the vicinity of nuclear power plants and nuclear fuel facilities in Spain. *Cancer Epidemiol Biomarkers Prev* 8: 925-934

85. Lopez-Abente G, Aragones N, Ramis R, Hernandez-Barrera V, Perez-Gomez B, Escolar-Pujolar A, Pollan M (2006a) Municipal distribution of bladder cancer mortality in Spain: possible role of mining and industry. *BMC Public Health* 6: 17

86. Lopez-Abente G, Pollan M, Escolar-Pujolar A, Errezola M, Abraira V (2001) *Atlas de mortalidad por cáncer y otras causas en España. 1978-1992*. Instituto de Salud Carlos III: Madrid

87. Lopez-Abente G, Ramis R, Pollan M, Aragones N, Perez-Gomez B, Gomez-Barroso D, Carrasco J, Lope V, Garcia-Perez J, Boldo E, Garcia-Perez J, Garcia-Mendizabal M (2006b) *Atlas municipal de mortalidad por cáncer en España, 1989-1998 (Atlas of municipal cancer mortality in Spain 1989-1998)*. Instituto de Salud Carlos III: Madrid

88. Lynge E, Anttila A, Hemminki K (1997) Organic solvents and cancer. *Cancer Causes Control* 8: 406-419

89. Lyons RA, Monaghan SP, Heaven M, Littlepage BN, Vincent TJ, Draper GJ (1995) Incidence of leukaemia and lymphoma in young people in the vicinity of the petrochemical plant at Baglan Bay, South Wales, 1974 to 1991. *Occup Environ Med* 52: 225-228

90. Martinez-Beneito M, Lopez-Quilez A, Amador A, Melchor I, Botella-Rocamora P, Abellan J, Verdejo F, Zuriaga O, Banaclocha H, Escolano M (2005) *Atlas de mortalidad de la Comunidad Valenciana 1991-2000*. Applied Spatial Statistics for Public Health Data: Valencia

91. Martinez-Beneito M, Lopez-Quilez A, Botella-Rocamora P (2008) An autoregressive approach to spatio-temporal disease mapping. *Stat Med* 27: 2874-2889

92. Maule MM, Magnani C, Dalmasso P, Mirabelli D, Merletti F, Biggeri A (2007) Modeling mesothelioma risk associated with environmental asbestos exposure. *Environ Health Perspect* 115: 1066-1071

93. Mcnally RJQ, Eden TOB (2004) An infectious aetiology for childhood acute leukaemia: a review of the evidence. *British Journal of Haematology* 127: 243-263

94. Michelozzi P, Fusco D, Forastiere F, Ancona C, Dell'Orco V, Perucci CA (1998) Small area study of mortality among people living near multiple sources of air pollution. *Occup Environ Med* 55: 611-615

95. Monge-Corella S, Garcia-Perez J, Aragones N, Pollan M, Perez-Gomez B, Lopez-Abente G (2008) Lung cancer mortality in towns near paper, pulp and board industries in Spain: a point source pollution study. *BMC Public Health* 8: 288

96. Muller AM, Ihorst G, Mertelsmann R, Engelhardt M (2005) Epidemiology of non-Hodgkin's lymphoma (NHL): trends, geographic distribution, and etiology. *Ann Hematol* 84: 1-12

97. Nieuwenhuijsen M, Paustenbach D, Duarte-Davidson R (2006) New developments in exposure assessment: the impact on the practice of health risk assessment and epidemiological studies. *Environ Int* 32: 996-1009

98. Parodi S, Stagnaro E, Casella C, Puppo A, Daminelli E, Fontana V, Valerio F, Vercelli M (2005) Lung cancer in an urban area in Northern Italy near a coke oven plant. *Lung Cancer* 47: 155-164

99. Parodi S, Vercelli M, Stella A, Stagnaro E, Valerio F (2003) Lymphohaematopoietic system cancer incidence in an urban area near a coke oven plant: an ecological investigation. *Occupational and Environmental Medicine* 60: 187-194

100. Pekkanen J, Pukkala E, Vahteristo M, Vartiainen T (1995) Cancer incidence around an oil refinery as an example of a small area study based on map coordinates. *Environ Res* 71: 128-134

101. Perez-Gomez B, Aragones N, Pollan M, Suarez B, Lope V, Llacer A, Lopez-Abente G (2006) Accuracy of cancer death certificates in Spain: a summary of available information. *Gac Sanit* 20 Suppl 3: 42-51

102. Pollan M, Lopez-Abente G, Moreno C, Vergara A, Aragones N, Ruiz M, Ardanaz E, Moreo P (1998) Rising incidence of non-Hodgkin's lymphoma in Spain: analysis of period of diagnosis and cohort effects. *Cancer Epidemiol Biomarkers Prev* 7: 621-625

103. R Development Core Team (2005) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Viena

104. Ramis R, Garcia-Perez J, Pollan M, Aragones N, Perez-Gomez B, Lopez-Abente G (2007) Modelling of municipal mortality due to haematological neoplasias in Spain. *J Epidemiol Community Health* 61: 165-171

105. Ramis R, Vidal E, Garcia-Perez J, Lope V, Aragones N, Perez-Gomez B, Pollan M, Lopez-Abente G (2009) Study of non-Hodgkin's lymphoma mortality associated with industrial pollution in Spain, using Poisson models. *BMC Public Health* 9: 26

106. Richardson S (1992) Statistical modeling of spatial variations in epidemiology. *Epidemiol Sante Publique* 40: 33-45

107. Richardson S, Abellan JJ, Best N (2006) Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Stat Methods Med Res* 15: 385-407

108. Richardson S, Thomson A, Best N, Elliott P (2004) Interpreting posterior relative risk estimates in disease-mapping studies. *Environ Health Perspect* 112: 1016-1025

109. Roman E, Beral V, Carpenter L, Watson A, Barton C, Ryder H, Aston DL (1987) Childhood leukaemia in the West Berkshire and Basingstoke and North Hampshire District Health Authorities in relation to nuclear establishments in the vicinity. *Br Med J (Clin Res Ed)* 294: 597-602

110. Rothman K (2002) *Epidemiology. An introduction*. Oxford University Press: New York

111. SAHSU. Small Area Health Statistics Unit. SAHSU. Imperial Collegue London . 2009. Ref Type: Electronic Citation

112. Sans S, Elliott P, Kleinschmidt I, Shaddick G, Pattenden S, Walls P, Grundy C, Dolk H (1995) Cancer incidence and mortality near the Baglan Bay petrochemical works, South Wales. *Occup Environ Med* 52: 217-224

113. Schotenfeld D, Fraumen J (1996) *Cancer Epidemiology and Prevention*. Oxford University Press: New York

114. Selvin HC (1958) Durkheim's 'suicide' and problems of empirical research. *American Journal of Sociology* 63: 607-619

115. Selvin S, Schulman J, Merrill DW (1992) Distance and risk measures for the analysis of spatial data: a study of childhood cancers. *Soc Sci Med* 34: 769-777

116. Sharp L, Black RJ, Harkness EF, McKinney PA (1996) Incidence of childhood leukaemia and non-Hodgkin's lymphoma in the vicinity of nuclear sites in Scotland, 1968-93. *Occup Environ Med* 53: 823-831

117. Smith BJ. Bayesian Output Analysis Program (BOA), Version 0.99.1 for S-PLUS and R [On-line]. 2001. Ref Type: Computer Program

118. Snow J (1855) *On the mode of communication of cholera (2nd edn)*. Churchill: London

119. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 64: 583-616

120. Spiegelhalter DJ, Thomas D, Best N (2003) *WinBUGS Manual user. Version1.4*. MRC: Cambridge

121. Spiegelhalter, D. J., Thomas D, Best, N., and Gilks, WR. BUGS: Bayesian inference using Gibbs sampling. [0.50]. 1996. Cambridge. Ref Type: Computer Program

122. Ugarte MD, Ibanez B, Militino AR (2004) Testing for Poisson zero inflation in disease mapping. *Biometrical Journal* 46: 526-539

123. Viel JF, Arveux P, Baverel J, Cahn JY (2000) Soft-tissue sarcoma and non-Hodgkin's lymphoma clusters around a municipal solid waste incinerator with high dioxin emission levels. *Am J Epidemiol* 152: 13-19

124. Viel JF, Richardson ST (1990) Childhood leukaemia around the La Hague nuclear waste reprocessing plant. *BMJ* 300: 580-581

125. Wakefield JC (2007) Disease mapping and spatial regression with count data. *Biostatistics* 8: 158-183

126. Wakefield JC, Morris S (2001) The Bayesian modelling of disease risk in relation to a point source. *Journal of the American Statistical Society* 96: 77-91

127. Waller A, Gotway C. (2004) *Applied Spatial Statistics for Public Health Data*. Wiley: Chischeter

128. Walter S D (2000) Disease mapping: a historical perspective. In *Spatial epidemiology: Methods and Applications*, Elliott P WJCBNGBDJ (ed) Oxford Medical Publications: Oxford

129. Wartenberg D, Reyner D, Scott CS (2000) Trichloroethylene and cancer: epidemiologic evidence. *Environ Health Perspect* 108 Suppl 2: 161-176

130. Wild CP (2009) Environmental exposure measurement in cancer epidemiology. *Mutagenesis* 24: 117-125

131. Zhang J, Lin G (2009) Spatial scan statistics in loglinear models. *Computational Statistics and Data Analysis* 53: 2851-2858

9. ABSTRACTS

9.1 MODELLING OF MUNICIPAL MORTALITY DUE TO HAEMATOLOGICAL NEOPLASIAS IN SPAIN

Background..

Spatial analysis of health events (spatial epidemiology) has the ability to suggest and detect possible sources of heterogeneity which may account for spatial incidence and mortality patterns in different diseases.

This study seeks to explore the geographical pattern of mortality by haematological tumours in Spain at municipal level using three models and to compare their the goodness of fit.

Methods.

The fitted Bayesian hierarchical models were: a) the Besag, York and Mollié model; b) a model based on zero-inflated Poisson (ZIP) distribution, which allowed a large number of event-free areas; and c) a mixture of distributions that enabled discontinuities (jumps in the pattern) to be modelled. The tree models allow to obtain smoothed relative risk maps for the all country. The goodness of fit was evaluated using the deviance information criteria.

Results.

The three models yielded very similar results. The ZIP model plotted a pattern almost identical to the BYM model. The goodness-of-fit criteria indicate that the mixture model is the one that best fits our data. Haematological tumours display a geographical pattern that could possibly be in part explained by environmental determinants, since many of the highest-risk towns belong to heavily industrialised areas.

Conclusions.

The choice of one or another model has scant practical consequences. The pattern of distribution supports the hypothesis that differences in lifestyles, air/industrial pollution and migratory phenomena may determine the pattern of urban mortality due to these tumours.

9.2 STUDY OF NON-HODGKIN'S LYMPHOMA MORTALITY ASSOCIATED WITH INDUSTRIAL POLLUTION IN SPAIN, USING POISSON MODELS

Background.

Non-Hodgkin's lymphomas (NHLs) have been linked to proximity to industrial areas, but evidence regarding the health risk posed by residence near pollutant industries is very limited. The European Pollutant Emission Register (EPER) is a public register that furnishes valuable information on industries that release pollutants to air and water, along with their geographical location. This study sought to explore the relationship between NHL mortality in small areas in Spain and environmental exposure to pollutant emissions from EPER-registered industries, using three Poisson-regression-based mathematical models.

Methods.

Observed cases were drawn from mortality registries in Spain for the period 1994-2003. Industries were grouped into the following sectors: energy; metal; mineral; organic chemicals; waste; paper; food; and use of solvents. Populations having an industry within a radius of 1, 1.5, or 2 kilometres from the municipal centroid were deemed to be exposed. Municipalities outside those radii were considered as reference populations.

The relative risks (RRs) associated with proximity to pollutant industries were estimated using the following methods: Poisson Regression; mixed Poisson model with random provincial effect; and spatial autoregressive modelling (BYM model).

Results.

Only proximity of paper industries to population centres (>2 km) could be associated with a greater risk of NHL mortality (mixed model: RR:1.24, 95% CI:1.09-1.42; BYM model: RR:1.21, 95% CI:1.01-1.45; Poisson model: RR:1.16, 95% CI:1.06-1.27). Spatial models yielded higher estimates.

Conclusions.

The reported association between exposure to air pollution from the paper, pulp and board industry and NHL mortality is independent of the model used. Inclusion of spatial random effects terms in the risk estimate improves the study of associations between environmental exposures and mortality.

The EPER could be of great utility when studying the effects of industrial pollution on the health of the population.

9.3 RISK AROUND PUTATIVE FOCUS IN A MULTY-SOURCE SCENARIO. NON-LINEAL REGRESSION MODELS

Backgrounds.

We consider the problem of investigating the risk of non-infectious diseases in populations exposed to pollution from different point sources.

The data most commonly available to study this question consist of counts of cases of disease over given areas (O_i) and distances between the focus and a central point within the areas (d_{ij}). Also covariates related to the socio-economic status are considered (Z_k).

This study seeks to explore the relationship between small area (municipalities or census tracts) cancer mortality in Spain and distance from industrial facilities, as an indirect measure of exposure to industrial pollution in a multi-source scenario, using a Poisson-regression-based model.

Methods.

The classic approach to the study of non-infectious disease with data counts is the ecological regression. Although, for our specific problem this Poisson regression is extended with the inclusion of a distance's function ($f(d_{ij})$) and the result is non-lineal model. This function models an elevated risk close to the source (alpha) with a neutral long-distance effect (beta).

$$O_i \sim Po(\mu_i)$$

$$\mu_i = \rho * \sum_k (\vartheta_k * Z_{ik}) * \prod_j f(d_{ij}); \qquad f(d_{ij}) = 1 + \alpha_j * \exp\left(-\left(\frac{d_{ij}}{\beta_j}\right)^2\right)$$

- ρ is the overall risk
- θ_k are the parameters of the socio-demographic covariates Z_{ik}
- αj and βj are the parameters of the distance function, and d_{ij} is the distance between the centroid of the area *i* and the focus *j*.

This model is applied to study the spatial variation of the cancer mortality risk in Gran Bilbao region related to exposure to pollutant substances released from the industrial facilities located within the region. Data is aggregated in census tracts and socio-demographic information has been included in the models as covariates.

Results.

We have studied the distribution of bladder, haematological and prostate cancer mortality.

The used model has given different risk functions associated with different focus. However, only for prostate cancer mortality the model with the distance's function was statistically significantly better than the model with the socio-demographic covariates only. For the remaining models the maximum likelihood tests were not statistically significant.

Conclusions.

The proposed model is able to identify different risk functions associated with different focus when we work in a multiple focus scenario, using aggregated data in small areas.

We have found evidence of association linking the distribution of prostate cancer mortality aggregated by census tracts and exposure to pollutant substances from the metal industrial facilities located within the area; exposure estimated through the distance between the point source and the centroid of the census tract.

The socio-demographic characteristics of the population are related to many cancer causes, as the results for the previous analysis yield.

10. APPENDIX

10.1 APPENDIX SECTION 3

WinBUGS code for the BYM Model

```
model
{
  for (i in 1 : N) {
      O[i]  ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]+.000001) + alpha + b[i] + h[i]
      theta[i]<- log(E[i]+0.00000001) + alpha + b[i] + h[i]
      RR[i] <- exp(alpha +  b[i] + h[i])   # Area-specific relative risk (for
maps)
      h[i] ~ dnorm(0, tau.h)        # Unstructured random effects
     PP[i] <- step(RR[i]-1)
       dev.i[i] <- O[i]*log((O[i]+step(-O[i]))/mu[i])-O[i]+mu[i]
  }

  # CAR prior distribution for spatial random effects:
  b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
  for(k in 1:sumNumNeigh) {
      weights[k] <- 1
  }

  # Other priors:
  alpha  ~ dflat()
  tau.b  ~ dgamma(0.5, 0.0005)
  sigma.b <- sqrt(1 / tau.b)
  tau.h  ~ dgamma(0.5, 0.0005)
  sigma.h <- sqrt(1 / tau.h)
  dev <- 2*sum(dev.i[])
}

DATOS
INITS
```

WinBUGS code for the Lawson Model

```
model
{
d[1:regions]~car.l1(adj[],weights[],num[],tau1)
b[1:regions] ~ car.normal(adj[], weights[], num[], tau1)
b.mean <- mean(b[])
d.mean<-mean(d[])
dev <- 2*sum(dev.i[])
for (i in 1 : regions) {
O[i] ~ dpois(mu[i])
log(mu[i]) <- log(E[i] + .000001) + alpha0 + a[i]+p[i]* b[i]+(1-p[i])*d[i]
theta[i]<- log(E[i] + .000001) + alpha0 + a[i]+p[i]* b[i]+(1-p[i])*d[i]
dev.i[i] <- O[i]*log((O[i]+step(-O[i]))/mu[i])-O[i]+mu[i]

RR[i] <- mu[i] / (E[i] + .000001)

PP[i] <- step(RR[i]-1)
}
 for(k in 1:sumNumNeigh) {
      weights[k] <- 1
  }
alpha0 ~ dflat()
tau1 ~ dgamma(rstar, dstar) sigma1<- 1 / sqrt(tau1)
tau2~dgamma(0.5,0.0005) sigma2<-1/sqrt(tau2)
for(j in 1:regions){
p[j]~dbeta(0.7,0.7)
a[j]~dnorm(0.0,tau2)}
}


DATOS

INITS
```

WinBUGS code for the ZIP Model

```
model
{
  for (i in 1 : N) {
      z[i]<-0
      z[i]  ~ dpois(phi[i])
      phi[i] <--ll[i]
      p[i]~ dbeta(1,1)
       ll[i]<-zero[i]*(log(p[i])-mu[i]+log(1-p[i]))+(1-zero[i])*(log(1-p[i])-
       mu[i]+O[i]*log(mu[i])-logfact(O[i]))
      zero[i] <- equals(O[i],0)
      log(mu[i]) <- log(E[i]+.0000001) + alpha + b[i] + h[i]
      theta[i]<- log(E[i]+.00000001) + alpha + b[i] + h[i]
      RR[i] <- exp(alpha +  b[i] + h[i])  # Area-specific relative risk
      h[i] ~ dnorm(0, tau.h)        # Unstructured random effects
       dev.i[i] <- O[i]*log((O[i]+step(-O[i]))/mu[i])-O[i]+mu[i]
      PP[i] <- step(RR[i]-1)
      }
  # CAR prior distribution for spatial random effects:
  b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
  for(k in 1:sumNumNeigh) {
      weights[k] <- 1
  }
  # Other priors:
  alpha  ~ dflat()
  tau.b  ~ dgamma(0.5, 0.0005)
  #sigma.b <- sqrt(1 / tau.b)
  tau.h  ~ dgamma(0.5, 0.0005)
  #sigma.h <- sqrt(1 / tau.h)
  dev <- 2*sum(dev.i[])
}
```

10.2 APPENDIX SECTION 4

R code for Poisson Regression and Mixed Model

```
##libraries
library(MASS)
library(nlme)


## fuctions loading for confidence intervals CI
source("C:/Eper/R/Poisson/ic-rr.R")
source("C:/Eper/R/Poisson/ic.rr.glmm.R")


##data files loading eper y lnh
source("C:/Eper/lnh/datos lnh ambos.dmp")


##tables
tt3<-tt.3
tt3$O<-datos$O
tt3$E<-datos$E


######################## regressions ################################
##gp1#####

rg1_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp1)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)

rg1<-glm(O~ offset(log(E+.000001))+factor(gp1)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)

##graphs
plot(rg1_m)
plot(rg1) ##la regresión sin efecto aleatorio (glm) tiene 4 gráficos

##ic
rg1_ic<-ic.rr(summary(glm(O~ offset(log(E+.000001))+factor(gp1)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)))
```

```
rg1_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp1)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))



###gp5#########
rg5_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp5)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)


rg5<-glm(O~ offset(log(E+.000001))+factor(gp5)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic
rg5_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp5)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg5_ic<-ic.rr(summary(glm(O~ offset(log(E+.000001))+factor(gp5)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)))


###gp6#######
rg6_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp6)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)
rg6<-glm(O~ offset(log(E+.000001))+factor(gp6)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic
rg6_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp6)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg6_ic<-ic.rr(summary(glm(O~ offset(log(E+.000001))+factor(gp6)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)))



###gp8#######
rg8_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp8)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)
```

```
rg8<-glm(O~ offset(log(E+.000001))+factor(gp8)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic
rg8_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp8)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg8_ic<-ic.rr(summary(glm(O~ offset(log(E+.000001))+factor(gp8)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)))


###gp13##############
rg13_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp13)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)


rg13<-glm(O~ offset(log(E+.000001))+factor(gp13)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic
rg13_m<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp13)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg13<-ic.rr(summary(glm(O~ offset(log(E+.000001))+factor(gp13)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)))


###gp14#############
rg14_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp14)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)


rg14<-glm(O~ offset(log(E+.000001))+factor(gp14)+factor(gpob)+
analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic
rg14_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp14)+factor(gpob)+
analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg14_ic<-ic.rr(summary(glm(O~
offset(log(E+.000001))+factor(gp14)+factor(gpob)+
```

```
analk+parok+rentk+phogk, family=poisson, data=tt3)))



###gp17###############

rg17_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp17)+factor(gpob)+

analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)


rg17<-glm(O~ offset(log(E+.000001))+factor(gp17)+factor(gpob)+

analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic

rg17_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp17)+factor(gpob)+

analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg17_ic<-ic.rr(summary(glm(glm(O~
offset(log(E+.000001))+factor(gp17)+factor(gpob)+

analk+parok+rentk+phogk, family=poisson, data=tt3)))



###gp20#############

rg20_m<-glmmPQL(O~ offset(log(E+.000001))+factor(gp20)+factor(gpob)+

analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)


rg20<-glm(O~ offset(log(E+.000001))+factor(gp20)+factor(gpob)+

analk+parok+rentk+phogk, family=poisson, data=tt3)


##ic

rg20_m_ic<-ic.rr.glmm(summary(glmmPQL(O~
offset(log(E+.000001))+factor(gp20)+factor(gpob)+

analk+parok+rentk+phogk, random= ~ 1|factor(PROV), family=poisson, verbose=F,
data=tt3)))


rg20_ic<-ic.rr(summary(glm(O~
offset(log(E+.000001))+factor(gp20)+factor(gpob)+

analk+parok+rentk+phogk, family=poisson, data=tt3)))
```

WinBUGS code for the BYM Model

```
model
{
  for (i in 1 : N) {
      O[i]  ~ dpois(mu[i])
      log(mu[i]) <- log(E[i]+.000001) + alpha + b[i] + h[i]
+Id*Paro[i]+Ir*Rent[i]+Ih*Phog[i]+ bcont*cont[i]+Ia*Analf[i]
      theta[i]<- log(E[i]) + alpha + b[i] + h[i]
+Id*Paro[i]+Ir*Rent[i]+Ih*Phog[i]+ bcont*cont[i]+ Ia*Analf[i]
      RR[i] <- exp(alpha +  b[i] + h[i] +Id*Paro[i]+Ir*Rent[i]+Ih*Phog[i]+
bcont*cont[i]+ Ia*Analf[i])
  # Area-specific relative risk (for maps)
      h[i] ~ dnorm(0, tau.h)        # Unstructured random effects
      PP[i] <- step(RR[i]-1)
      #dev.i[i] <- O[i]*log((O[i]+step(-O[i]))/mu[i])-O[i]+mu[i]
  }

  # CAR prior distribution for spatial random effects:
  b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
  for(k in 1:sumNumNeigh) {
      weights[k] <- 1
  }

  # Other priors:
  Ia ~ dnorm(0.0, 1.0E-5)
  Id ~ dnorm(0.0, 1.0E-5)
  Ir ~ dnorm(0.0, 1.0E-5)
  Ih ~ dnorm(0.0, 1.0E-5)
  bcont ~ dnorm(0.0, 1.0E-5)

  alpha  ~ dflat()
  tau.b  ~ dgamma(0.5, 0.0005)
  sigma.b <- sqrt(1 / tau.b)
  tau.h  ~ dgamma(0.5, 0.0005)
  sigma.h <- sqrt(1 / tau.h)
  #dev <- 2*sum(dev.i[])
}
```

10.3 APPENDIX SECTION 5

R code.

```
#### Function to estimate parameters in a Poisson regression
#### with a non-linear term for the distance effect to multiple pollutant focuses.
#### we include lung cancer smr as Proxy of tobacco consumption

## data: INE;O;E;Analf;Rent;pulmon;distancias
## (distance unit 100.000 metres)

## the function gives the estimated parameters and the Monte Carlos simulations
## to calculate the variances

reg.dist2<-function(data)
{

##inits
  n<-ncol(data)-6
  rg<-glm(O~ offset(log(E))+Analf+Rent+pulmon, family=poisson, data=data)
  alpha1<-rep(0.1,n)
  beta1<-rep(0.3,n)
  rg.coe<-as.numeric(rg$coefficients)
  inic<-c(rg.coe,alpha1,beta1)

## Likelihood function
  Log.lik<-function(data, Theta, n)
  {
    sc<-as.matrix(data[,4:6])
    rho<-Theta[1]
    theta<-Theta[2:4]
    alpha<-Theta[5:(5+n-1)]
    beta<-Theta[(5+n):(5+2*n-1)]
    Q<-sc%*%theta
    dist<-data[,7:(6+n)]
    mu=0
    k=0
    for(i in 1:nrow(data))
    {
      t<-0
      for(j in 1:n)
      {
      t[j]<-1+alpha[j]*exp(-(dist[i,j]/beta[j])^2)
      }
    k[i]<-prod(t)
```

```
    mu[i]<-data$E[i]*exp(rho+Q[i])*k[i]
    }
    L<--sum(mu)+sum(data$O*log(mu))
    L1<--L
    list(L1)
  }

## mid function
Log.Lik<-function(Theta)
  {
    data=data
    n<-ncol(data)-6
    res<-Log.lik(data,Theta,n)
    res
  }

## optimization function
  theta.hat<-optim(par=inic,fn=Log.Lik)

## convergence
  conver<-theta.hat[[4]]
  inic1<-theta.hat[[1]]
  con<-1
  while(conver>0)
  {
    theta.hat<-optim(par=inic1, fn=Log.Lik)
    conver<-theta.hat[[4]]
    inic1<-theta.hat[[1]]
    con<-con+1
  }
print(theta.hat)

## MLE
  theta<-theta.hat[[1]] ## estimadores

######## Standard errors ############ Simulatión
  sc<-as.matrix(data[,4:6])
  rho<-theta[1]
  t1<-theta[2:4]
  alpha<-theta[5:(4+n)]
  beta<-theta[(5+n):length(theta)]
  Q<-sc%*%t1
  dist1<-data[,7:(7+n-1)]
  mu=0
```

```
  k=0
  for(i in 1:nrow(data))
  {
    t<-0
    for(j in 1:n)
    {
    t[j]<-1+alpha[j]*exp(-(dist1[i,j]/beta[j])^2)
    }
  k[i]<-prod(t)
  mu[i]<-data$E[i]*exp(rho+Q[i])*k[i]
  }

## Monte Carlo simulations
  data1<-data
  theta.hat.S<-0
  for(k in 1:100)
   {
   print(k)
   ###simulaciones de Oi -> Yi
     Y=0
     for(i in 1:nrow(data))
     {
       Y[i]<-rpois(1,mu[i])
     }
     data1$O<-Y
     Log.Lik<-function(Theta)
     {
       data=data1
       res<-Log.lik(data,Theta,n)
       res
     }
     theta.hat.S[k]<-optim(par=theta, fn=Log.Lik)
    }
  Theta.hat.S<-t(as.data.frame(theta.hat.S))
  sd<-apply(Theta.hat.S,2,sd)
list(theta,sd)
}
```

More results of section 5.

(5.2.2) Standard errors

Simulations experiment to assess the performance of Hessian standard errors against Montecarlo standard errors.

We generate a sample of observed cases from a model with four socio-demographic covariates and two factories (the model has *real values* as parameters). We calculate the *real standard errors* via Monte Carlo. We generate 100 different samples from the model and we estimate the parameters; with these 100 estimators we get the standard errors. We use the first sample we calculate the Monte Carlo standard errors and the Hessian standard errors. Results of this experiment are included in the next table.

	Real Values	Real standard errors	Monte Carlo standard errors	Hessian standard errors
ρ	0.01877	**0.16027**	**0.1824**	**0.2446**
$\theta 1$	0.11056	**0.17157**	**0.2255**	**0.2630**
$\theta 2$	0.09949	**0.09970**	**0.1149**	**0.1247**
$\theta 3$	-0.01997	**0.06585**	**0.1000**	**0.1012**
$\theta 4$	-0.09325	**0.07401**	**0.0936**	**0.1000**
$\alpha 1$	0.10000	**0.23807**	**0.2654**	**0.1915**
$\alpha 2$	0.10000	**0.17869**	**0.2164**	**0.1801**
$\beta 1$	0.20000	**0.23244**	**0.1725**	**0.1232**
$\beta 2$	0.40000	**0.33953**	**0.4110**	**1.1490**

***Table**. Real values, real standard errors, Monte Carlo standard errors and Hessian standard errors.*

(5.3.1) More results of the analysis of bladder cancer. Tables.

Spatial model: individual regressions

Firstly, we study the 20 industrial locations one by one, adding to the spatial model the distance variable and the socio-demographic covariates. We fit an independent model for each point source. The results show that the deviances for these 20 models are very similar to the deviance of model 1 (717.11) and according to the likelihood test none of them is significantly better than the initial Poisson model.

Industrial facility	Deviance
d3641	716.68
d3686	716.44
d3689	716.11
d3693	716.13
d3700	716.55
d3701	716.38
d3702	715.78
d3707	715.91
d3712	716.12
d3716	715.71

Industrial facility	Deviance
d3721	716.24
d3723	716.17
d3724	715.92
d3727	716.22
d3733	716.06
d3737	716.17
d3739	716.04
d3742	716.33
d3743	716.14
d3745	716.28

***Table**. Deviances for the individual regressions.*

Industrial facility	Deviance	Alpha	Lower limit	Upper limit	Beta	Lower limit	Upper limit
d3641	716.68	-0.184	-0.572	0.204	37.255	19.715	54.795
d3686	716.44	0.535	-0.262	1.332	0.104	-0.049	0.258
d3689	716.11	1.311	0.041	2.581	0.049	0.006	0.092
d3693	716.13	0.294	-0.276	0.865	0.04	-0.049	0.128
d3700	716.55	0.237	-0.503	0.978	0.085	-0.234	0.404
d3701	716.38	0.171	-0.563	0.905	0.178	-0.483	0.838
d3702	715.78	0.674	0.014	1.333	0.11	-0.085	0.305
d3707	715.91	0.311	-0.543	1.164	0.039	-0.084	0.162
d3712	716.12	0.168	-0.529	0.865	0.182	-0.646	1.011
d3716	715.71	0.85	0.127	1.574	0.091	-0.069	0.252
d3721	716.24	0.447	-0.549	1.443	0.061	-0.033	0.155
d3723	716.17	-0.029	-0.489	0.431	6.445	1.664	11.227
d3724	715.92	0.325	-0.377	1.026	0.07	-0.145	0.285
d3727	716.22	0.185	-0.52	0.889	0.066	-0.318	0.45
d3733	716.06	1.246	0.482	2.01	0.112	-0.014	0.238
d3737	716.17	0.757	-0.637	2.151	0.025	-0.022	0.073
d3739	716.04	0.408	-0.266	1.082	0.134	-1.284	1.553
d3742	716.33	0.26	-0.548	1.068	0.386	-1.383	2.154
d3743	716.14	0.517	-0.022	1.057	0.23	-7.928	8.387
d3745	716.28	0.268	-0.276	0.812	0.099	-0.356	0.554

***Table**. Deviances and estimators of the parameters of the distance function for the 20 focuses of Gran Bilbao.*

Only one of the focuses (3689) has both parameters of the distance function, alpha and beta, are inside the limits, even though this factory is located outside of the area of study.

Spatial model. Multiple regression inside the circumference

The second approach is done across the area inside the circle. The industrial facilities 3693, 3702, 3702, 3716, 3724, 7333 and 3739 are located within this area. However, there are two pairs of factories very close to each other, 3693-3707 and 3722-3739; therefore, each pair is treated as just one pollutant focus. We fit a model that includes the five focuses. Moreover, when we reduce the area of study two of the three covariates are not statistically significant anymore, income and tobacco prevalence, thus we remove these two covariates from the model.

```
glm(formula = O ~ offset(log(E)) + educ + income + lung, family = poisson)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7514  -1.1476  -0.1687   0.5012   2.8590
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.5143     0.3524   1.459   0.1445
educ          1.7809     0.8170   2.180   0.0293 *
income       -0.4288     0.4728  -0.907   0.3644
lung          0.1485     0.1137   1.307   0.1914
---
    Null deviance: 260.68  on 247  degrees of freedom
```

```
Residual deviance: 253.05  on 244  degrees of freedom
AIC: 594.32

glm(formula = O ~ offset(log(E)) + educ, family = poisson)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7363  -1.1573  -0.1909   0.5170   2.7334
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6642     0.3117   2.131   0.0331 *
educ          1.3672     0.6249   2.188   0.0287 *
---
    Null deviance: 260.68  on 247  degrees of freedom
Residual deviance: 255.51  on 246  degrees of freedom
AIC: 592.77
```

The likelihood for the model with one covariate, education level, and 5 focuses is -225.777 and the likelihood for the Poisson model with one covariate is -227.6133. The statistic of the likelihood test has a value of 3.6726 and the 5% critical value of chi-square with 5 df is 11.07. These results suggest that the model with 5 focuses is not better than the model with just one covariate.

11 PAPERS

- Modelling of municipal mortality due to haematological neoplasias in Spain.

Rebeca Ramis, Valentín Hernández-Barrera, Marina Pollán, Nuria Aragonés, Beatriz Pérez-Gómez, Gonzalo Lopez-Abente.

Journal of Epidemiology and Community Health. 2007, 61:2.

- Study of non-Hodgkin's lymphoma mortality associated with industrial pollution in Spain, using Poisson models.

Rebeca Ramis, Enrique Vidal, Javier García-Pérez, Virginia Lope, Nuria Aragonés, Beatriz Pérez-Gómez, Marina Pollán and Gonzalo López-Abente.

BMC Public Health. 2009, 9:26

- Risk around putative focus in a multy-source scenario. Non-lineal regression models.

Rebeca Ramis, Peter Diggle, Koldo Cambra and Gonzalo López-Abente.

Submitted in Epidemiology. 2009