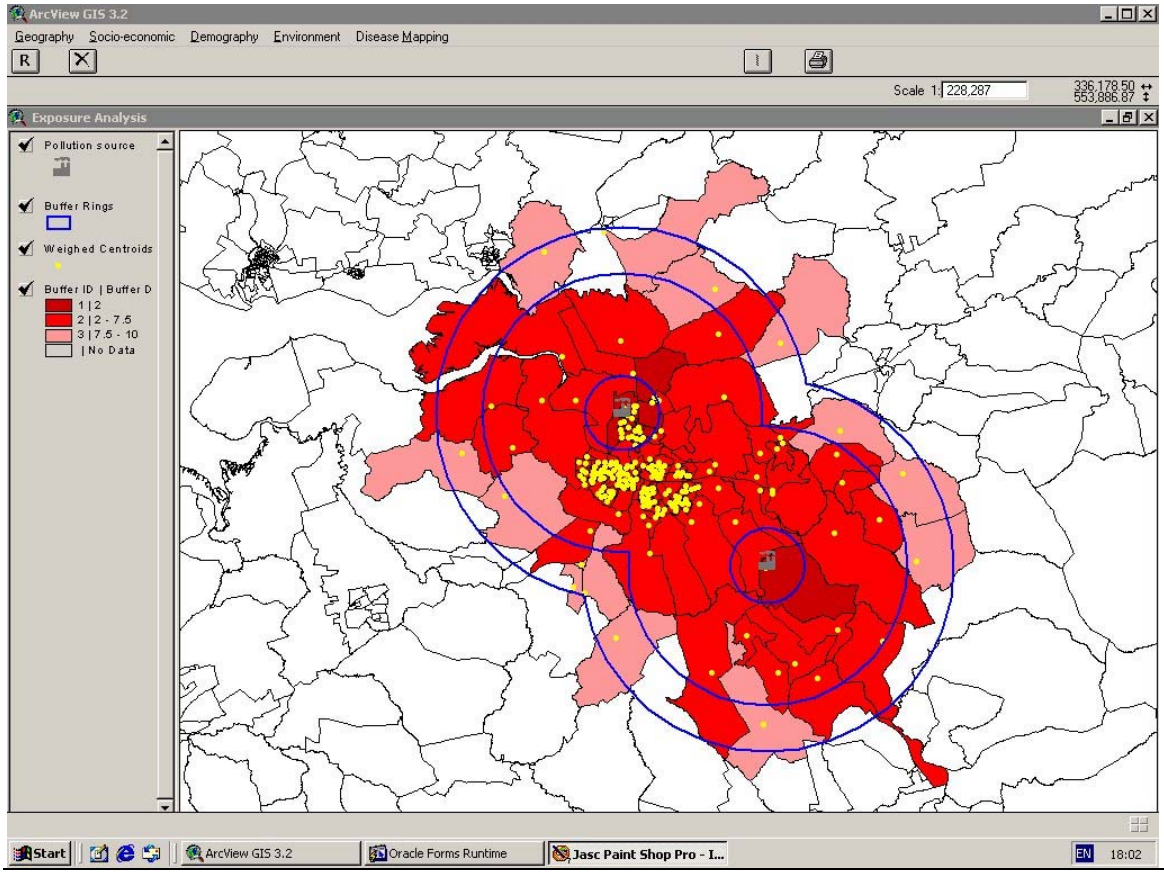
EUROHEIS

A European Health and Environment Information System for Exposure and Disease Mapping and Risk Assessment



FINAL REPORT 2003

Project Code SI2.291820 (2000 CVG2-605)

CONTENTS

ABBREVIATIONS

ARCVIEW	GIS software
AMI	Acute Myocardial Infarction
CDC	Centers for Disease Control and Prevention, Atlanta, USA
ECG	Electrocardiogram
GIS	Geographic Information Systems
HIA	Health Impact Assessment
IARC	International Agency for Research on Cancer
ICD	International Classification of Disease
ISEE	International Society for Environmental Epidemiology
ISEA	International Society of Exposure Analysis
MAUP	Modifiable Areal Unit Problem
MCMC	Markov Chain Monte Carlo method
ORACLE	Database software
PM10	Particulate Matter (10 micrometers in diameter and smaller)
PCDD/F	Polychlorinated Dibenzodioxin DibenzoFuran
RR	Relative RIsk
SAHSU	The Small Area Health Statistics Unit (UK)
SES	Socio-Economic Status
SIR	Standardised Incidence Ratio
SMASH	Small Area Statistics on Health (Finland)
SMR	Standardised Mortality Ratio
WHO	World Health Organisation
YYL	Years of Life Lost

LIST OF TABLES

LIST OF FIGURES

Executive summary

The EUROHEIS (European Health and Environment Information System) project aims to improve understanding of the links between environmental exposures, health outcome and risk through the development of integrated information systems for rapid assessment of relationships between the environment and health at a geographical level. EUROHEIS is a three-year project with each of three phases funded separately. The aims of the EUROHEIS project are to improve the analysis of health and environmental data in order to respond rapidly to putative health threats. The project enables a speedy assessment of the relationships between environmental exposures and disease within the partner countries, and improves the knowledge and understanding of health risk management.

The first phase of the project was designed to examine the feasibility of installing an analysis tool in each of the partner countries, based on the Rapid Inquiry Facility (RIF) developed at SAHSU (Imperial College London). The RIF combines geographical information software and health and environmental databases in an easy to use exploratory tool, which allows the assessment of risk and disease mapping at a small area level. Further development of the RIF was carried out to facilitate the transfer to the other partners.

The second phase of the project allowed the development and installation of a RIF to varying degrees in partner countries, taking into account the feasibility study. An analysis of socio-economic indices and their components across the partner countries as well as an investigation of the RIF system as a tool for Health Impact Assessment (HIA) were explored.

The objective for the third phase of the project was to demonstrate the usefulness of the RIF in answering questions concerning environmental health risks, utilising the system within the context of improving public health, preventing human illness and diseases, and obviating sources of danger to health. This was piloted through a series of case studies carried out within each of the partner countries to include not only RIF analyses, but also an evaluation of the role the RIF may play in providing rapid and accurate information to local policy makers.

The Danish partner tested the possibility of using routine health and population data for a study on the occurrence of cancer in a population
exposed to airborne dioxins. Exposure in the surroundings of the plant were modelled, and concentric circles around the plant at various distances were defined. Cancer risk in the defined areas were computed. The results showed no increase in the incidence of cancers in the exposed population, which was an important finding from a public health point of view.

The Finnish partner investigated the possible increase in cancer risk following exposure to chlorophenols and PCDD/F emanating from a heavily contaminated river. Total cancer as well as several specific cancer sites were studied. The exposure assessment was based on the distance of the place of residence from the river and from the Gulf of Finland. The results indicated increased excess cancer risk in farmers living close to the River Kymijoki. These important findings give first

approximations of risk - the results need to be confirmed in individual-level studies.

The Spanish partner performed three case studies to evaluate the usefulness of the implemented system. In Valencia, the relationship between nitrate concentration in drinking water and cancer mortality was assessed, as well as the possible relationship between calcium and magnesium concentrations in drinking water and cardiovascular and cerebrovascular mortality, using data on water quality and cancer mortality at municipality level. Smoothed maps of the relevant variables were produced using the RIF. Relationships between risk factors and mortality rates were explored by comparing populations at different risk levels within the RIF methodology. Finally, the health impacts of air pollution by tile industry in Castellón were investigated. Results from the three case studies have provided guidance to health policies for disease prevention and medical care facilities.

The Swedish partner illustrated the capability of the implemented RIF as a tool for disease mapping and description of disease occurrence. The occurrence of acute myocardial infarction (AMI) was described in relation to socio-economic deprivation in the Stockholm region. The RIF also produced maps of road traffic generated air pollution in the Stockholm region, which will be used for further studies of possible associations between AMI and air pollution.

The UK partner investigated cancer incidence in areas exposed to high levels of bromate. High levels of bromate were discovered in local water supplies in the Hatfield area (North of London), as well as private boreholes, likely to have come from a chemical plant manufacturing sodium and potassium bromates. Incidence and relative risk of selected cancers in the affected area were computed. The population under study consisted of two areas: enumeration districts (EDs) falling within contaminated water supply boundaries, and an area containing EDs whose population centroids are included within a 7.5km radius of the old plant. There was great public concern within the locality.

A key component of any spatial analysis of the health effects of environmental exposure is control for possible confounding by differences in deprivation between areas. The Irish partner has worked with the other EUROHEIS partners to examine proposed measures of person and area level deprivation in their chosen application areas. The quality and availability of indicators of deprivation at the individual level and the small-area level, and their usefulness in the RIF, has been evaluated in this work.

The Italian partner has applied model-based methods for Health Impact Assessment (HIA) to data on environment and health in selected Italian cities. The work analyses data of the kind that is normally used for conducting HIA exercises (i.e., population-based routinely collected data), at the small area level. The methods for HIA developed in EUROHEIS are based on models incorporating components for exposure-response function, exposure profiles and variability, existence of susceptible subgroups as well as latency time and its variability. In addition, the models make allowance for the possible sources of uncertainty using confidence limits of the exposure response and errors in the exposure measures. Mortality associated with ambient air pollution has been used to illustrate the HIA procedure.

We also gained an additional partner in EUROHEIS : the National Institute of Public Health and the Environment in the Netherlands, who successfully has started to to implement the RIF in the Netherlands during phase III.

A very successful end of project conference was held in Sweden in March 2003, attracting circa 100 delegates from the EU as well as from other countries outside the EU, such as Peru, India, Israel, Canada and the USA. The conference focused on the use of the EUROHEIS RIF for public health. The system was demonstrated and EUROHEIS partner countries presented results from the case study evaluations. The conference programme as well as proceedings form the conference are annexed to this report.

Three independent external evaluators (representing the public health, GIS and environment areas of expertise) were invited to give their opinion of how the system can be used in future applications. Their reports are included as a separate annex to this final report for the EUROHEIS project.

We have submitted a proposal for future development of EUROHEIS for the Programme of Community Action in the field of Public Health (2003-2008), aiming to further develop EUROHEIS, making the system available to additional EU and candidate countries (Hungary and Poland are partners in the new proposal). The current EUROHEIS software will be translated into a more user-friendly system to facilitate dissemination via the Internet. Further methodological development will make it possible to input data from dispersion (and similar) models to enhance exposure assessment, as well as data export facilities to make further, more advanced statistical analysis possible when needed. A technical support centre (help desk) will be set up to facilitate implementation in member countries. Training material will be developed, as well as training courses to assist installation and operation of the software. Guidelines for interpretation of results will be developed (including issues of data quality, zone design (scale and aggregation) and migration). Furthermore, we intend to set up a network of environmental health experts to provide technical, methodological and scientific support within the framework of EUROHEIS, and assist in training of network members.

The use of the system for Health Impact Assessment (HIA) (including Environmental Burden of Disease (EBD)) will be further explored, building on the experience of the current EUROHEIS project. In addition to environmental data, the system will incorporate measures of socio-economic status and other health determinants (such as smoking, obesity etc) in order to make possible analysis of environmental equity, which will be a major project task.

In light of the successful first EUROHEIS conference (March 2003), we are planning a second conference in 2006.

Finally, it should be noted that EUROHEIS has been widely recognised as a valuable tool for exploring environmental health risks, evidenced by the international participation in the EUROHEIS conference. In particular, CDC (Centers for Disease Control and Prevention, Atlanta, USA) has recently started an environmental health tracking system, with similar aims as the EUROHEIS project. Collaboration discussions have started and CDC is included as a (non-paid) partner in the new EUROHEIS proposal.

Introduction

The EUROHEIS project aims to improve understanding of the links between environmental exposures, health outcome and risk through the development of integrated information systems for rapid assessment of relationships between the environment and health at a geographical level. The partners involved in this project have been among the leaders in Europe in developing and applying epidemiological and statistical methods for the appraisal and analyses of disease occurrence in small geographical areas related to point source exposures and to disease mapping.

The EUROHEIS project reported here was initially designed as a three-year project; subsequently each of three phases were funded as separate one-year projects. The initial project period was from the 15 December 2000 until the 15 December 2001, extended to end on the 15 May 2002. The 2002 report should be seen as an interim report for the entire three year project. Phase III covers the period between 16 May 2002 and 15 May 2003. This is the report for the final year of the project (Phase III).

In the *first phase* of the project, a one-year feasibility study was undertaken to assess the possibilities of implementing systems for point source investigations and disease and exposure mapping within the participating countries, modelled on a system being developed within the UK. The UK system was also further developed to include more generalised modules for disease and exposure mapping. It suggested that data are available and suitable for the development of a system for point-source analysis and disease mapping in several of the partner countries. Prototype systems for point-source investigations, including some disease mapping facilities, had already been developed in the UK and Finland.

The *second phase* built on the feasibility study. Its aim was to implement systems for point-source investigations and disease and exposure mapping (incorporating state-of-the-art Bayesian statistical models and Geographical Information Systems (GIS)) in the participating countries, modelled on the system being developed by the UK partner. In Finland the existing SMASH system had been modified in collaboration with the EUROHEIS group. The RIF system was also being implemented. Denmark had decided to adapt the RIF to their own requirements, taking advantage of their rich data sources incorporating individual level population data. Ireland undertook an extensive review of meta-data protocols for the exchange of meta-data between partners in relation to the health and socio-economic databases available to the partners in order to develop and test appropriate and sensitive measures of deprivation at a suitable spatial level. Italy explored the possibility of using the RIF for health impact assessment. The Italian partner has also been responsible for liasing with APHEIS (an Air Pollution and Health Information System), which is another project funded by the same EU programme as EUROHEIS (DG SANCO, Action on Pollution Related Research).

In this, the *third phase* of the project, the UK partner has further developed the Rapid Inquiry Facility (RIF) to aid its implementation in several of the partner countries. The scripts have been rewritten in light of feedback from our partners. The system has been tested and finalised ready for distribution. Documentation, including the data and hardware specifications has been completed and we have provided assistance (including site visits) to partner countries where required. The UK partner has taken the lead on the development of the EUROHEIS website which includes details of the

project, project partners, interim reports, and a demonstration of the RIF as well as information on the end of project conference, 30-31 March 2003. A leaflet summarising the EUROHEIS project has also been produced. The results of the project have been disseminated through international meetings, publications and through the project website.

This final report for the EUROHEIS project focuses on the work performed in the *third phase* of the project, detailing the work of each of the partners in separate chapters, focussing on the case studies performed to the test the country specific systems.

The last chapter describes the wide dissemination of the EUROHEIS project findings, including meetings and conferences as well as published papers.

Annex 1 consists of documentation for the enhanced Rapid Inquiry Facility (RIF);
Annex 2 contains the EUROHEIS conference programme;
Annex 3 contains the proceedings from the EUROHEIS conference; and
Annex 4 contains the reports by the independent assessors

Note also that further details regarding the project can be obtained from the EUROHEIS web site (http://www.med.ic.ac.uk/divisions/60/euroheis/homepage.htm) Soon to be relocated to www.euroheis.org.

Denmark

Arne Poulstrup
Henrik L. Hansen

National Board of Health
Vejle
DENMARK

Introduction

Utilising the concepts of the English Rapid Inquiry Facility (RIF), for detecting and investigating clusters of diseases related to environmental hazards, Denmark, through the National Board of Health has decided to develop its own version utilising the Danish high-resolution databases on health issues: the Danish Cancer Registry, the National Registry of Patients (in- and out-patients), the Congenital Malformation Registry and the National Registry of Death. As denominator data are used the Danish Civil Registration Register (CPR) and included in the model will also be essential socio-economic variables on individual or aggregated levels.

The envisaged national, geographically based health database is under construction. Two pilot studies have been initiated to provide experiences to be used for the construction of the national model: one project dealing with allergy using a patient database and one dealing with an environmental database of water quality.

Furthermore a new development with a new interested and interesting partner in the national environmental health scenario has emerged:

From the government a clear wish has been expressed to enhance the utilisation of the registers for environment *and* health, in order to examine possible links between environmental exposures and negative health outcomes, and to generate more knowledge on the issues through enhanced environmental health research. Considerable financial resources have been set-aside for this activity on the national budget for 2002 – 2004.

As a consequence the Danish Environmental Protection Agency (DEPA) has just released a report, describing the possibilities of linking the comprehensive Danish *environmental* databases with the above mentioned national GIS-linked health database, www.mst.dk

Pilot

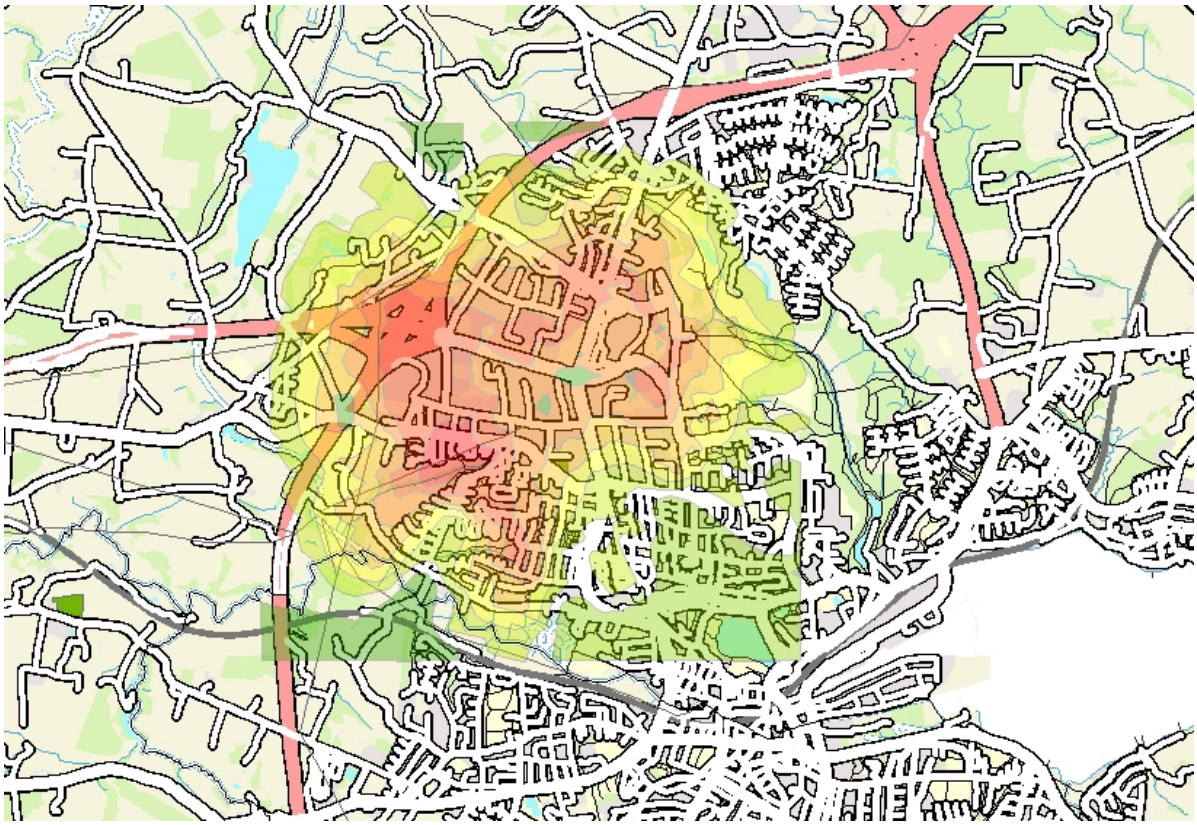
To demonstrate the power and potential of the Danish model for linkage of health data with environmental exposure data, utilising the RIF principles of data linkage through via GIS, a known case of exposure of approximately 25,000 individuals to airborne dioxin, was chosen.

In this case study it has been possible to apply an exposure model of airborne dioxin on the said population to determine the amount of dioxin to which it has been exposed over the years, i.e. since 1980. The exposure model used has been developed through a simulation procedure utilising known emission concentrations from the industries in the area combined with longitudinal meteorological data.

The model, however, does not allow for more than an estimation of average concentrations of dioxin per hour per m^3, as only very recent data of the dioxin emissions were available for the model.

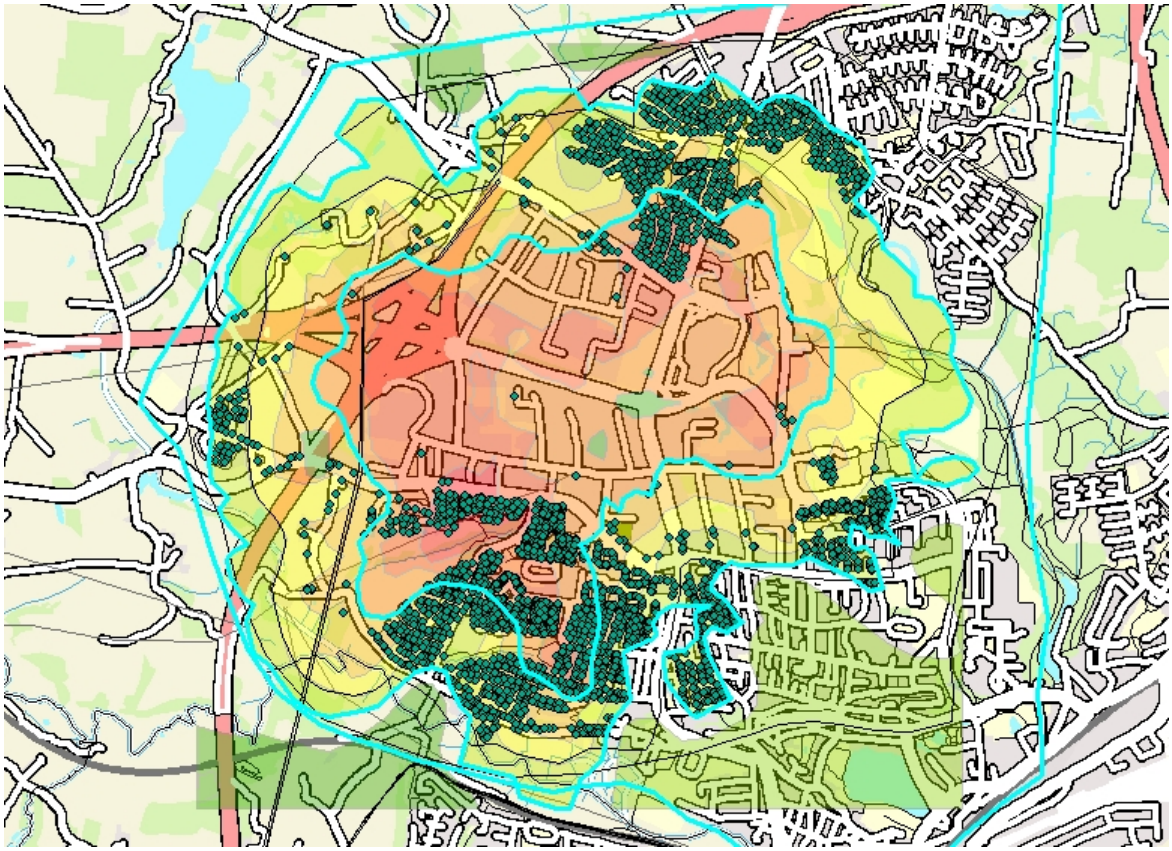
The model used is a Danish Gaussian plume model, simulating plume dispersions from one or more stacks. The model is widely used within the environmental protection authorities, and has been internationally validated.

Figure 1. The colours indicate the various concentrations of emitted dioxin: pale green is 1 picogram dioxin per hour while deep red represents a concentration of 6 picogram dioxin per hour



Thus the area and the population affected have been demarcated quite accurately as well as it has been possible to demarcate different concentration bands/polygons of dioxin.

Figure 2. The blue lines follow the concentration bands of hourly emitted dioxin and are used to demarcate the populations exposed to the various concentrations

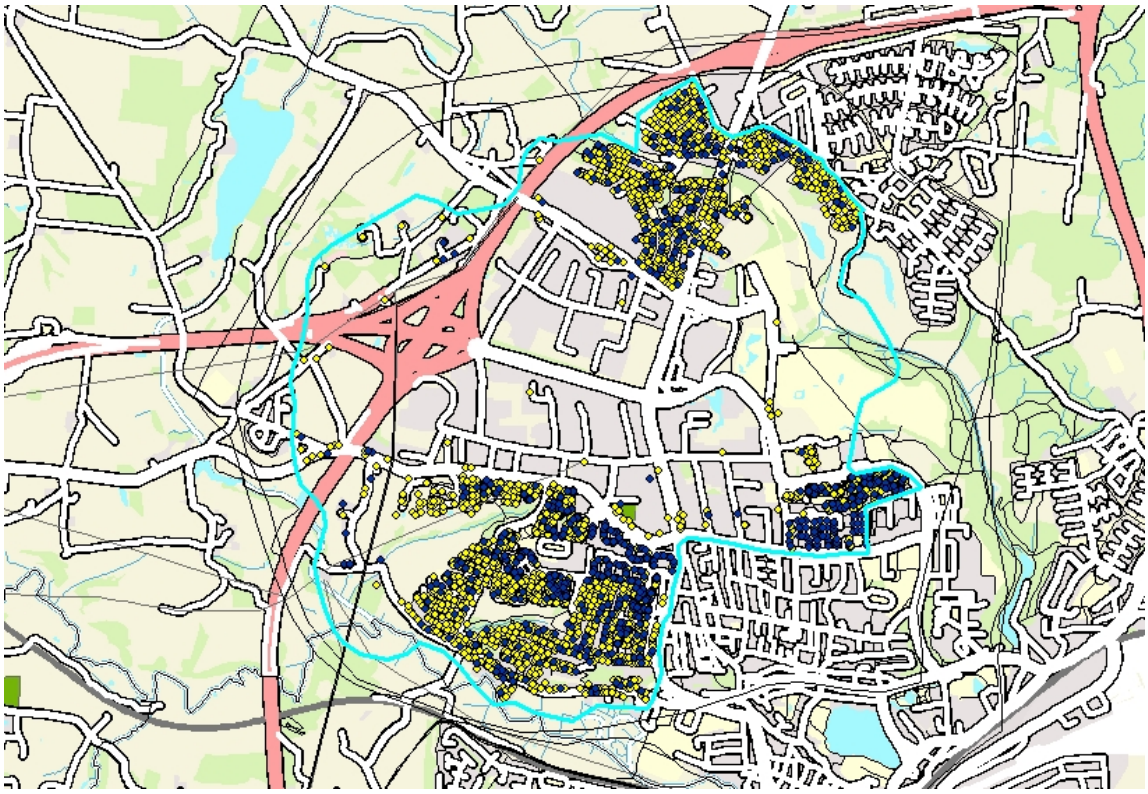


Through the application of the geo-coded addresses of all individuals in the exposure area from 1986 to 1998, it has been possible to keep record of all individuals who have been born, moved into or out of or died in the area. Hence it has to some extent been possible to keep track of the migration and the exposure time of the population.

By linkage of the exposed individuals´ personal identification numbers (CPR no´s; CPR=Central Population Register [Civil Registration System]) to the National Cancer Register, the number of cancers detected in the area in the period from 1986 – 1998 has been established.

The time span was chosen to permit for sufficient exposure time, and 1998 was the last year for which the cancer registry was up- and validated.

Figure 3. In this picture yellow dots represent addresses of the area while the blue dots represent the identified cancers of the population demarcated.



Different analyses of the cancer incidence in the exposed area compared to national figures have been made. The analyses showed no increase in the incidence of cancers in the exposed population, neither in any of the years covered nor aggregated over the years examined: 1986 till 1998.

The cancer figures analysed were cumulated incidence of all types of cancers, excluding cancer of the skin. The data were controlled for age, sex and migration.
The follow up time, however, has been short, and the effect of the recent high figures of measured dioxin in the air, will have to await further investigation.

Finland

Esa Kokki, Phil.Lic.
Pia Verkasalo, D.Med.Sc.
Juha Pekkanen, D.Med.Sc.

National Public Health Institute
Department of Environmental Health
Kuopio
FINLAND

Introduction

Department of Environmental Health, National Public Health Institute, Kuopio, Finland, has participated in EUROHEIS I, II and III – projects over the last three years. At the same time, we have collaborated intensely with researchers at the Finnish Cancer Registry and the University of Jyväskylä. The overall aim of the three projects has been in improving the understanding of the links between environmental exposures, health outcomes and risk through the development of an integrated information system for the rapid assessment of relationships between the environment and health at a geospatial level.

The focus has shifted from a feasibility study, to implementation of systems for point source investigations and to conduction of a case study on risk of cancer and proximity of residence to a river with high sediment levels of dioxins. More specifically, the aim of year one was in undertaking a feasibility study to assess the possibilities of implementing systems for point source investigations and disease and exposure mapping within the participating countries, modelled on the system developed within the UK. The aim of year two was in implementation of systems for point source investigations and disease and exposure mapping, modelled on the system being developed by the UK partner. The aim of year three was in evaluation of implemented systems for point-source analyses and disease and exposure mapping using country-specific case studies. An additional aim of the Finnish partner was in the extension of statistical methods to solve the problems of high resolution and sparse data in estimation of cancer risk around a point source. In this report, we summarize the results of the three projects and suggest needs and ways for future developments.

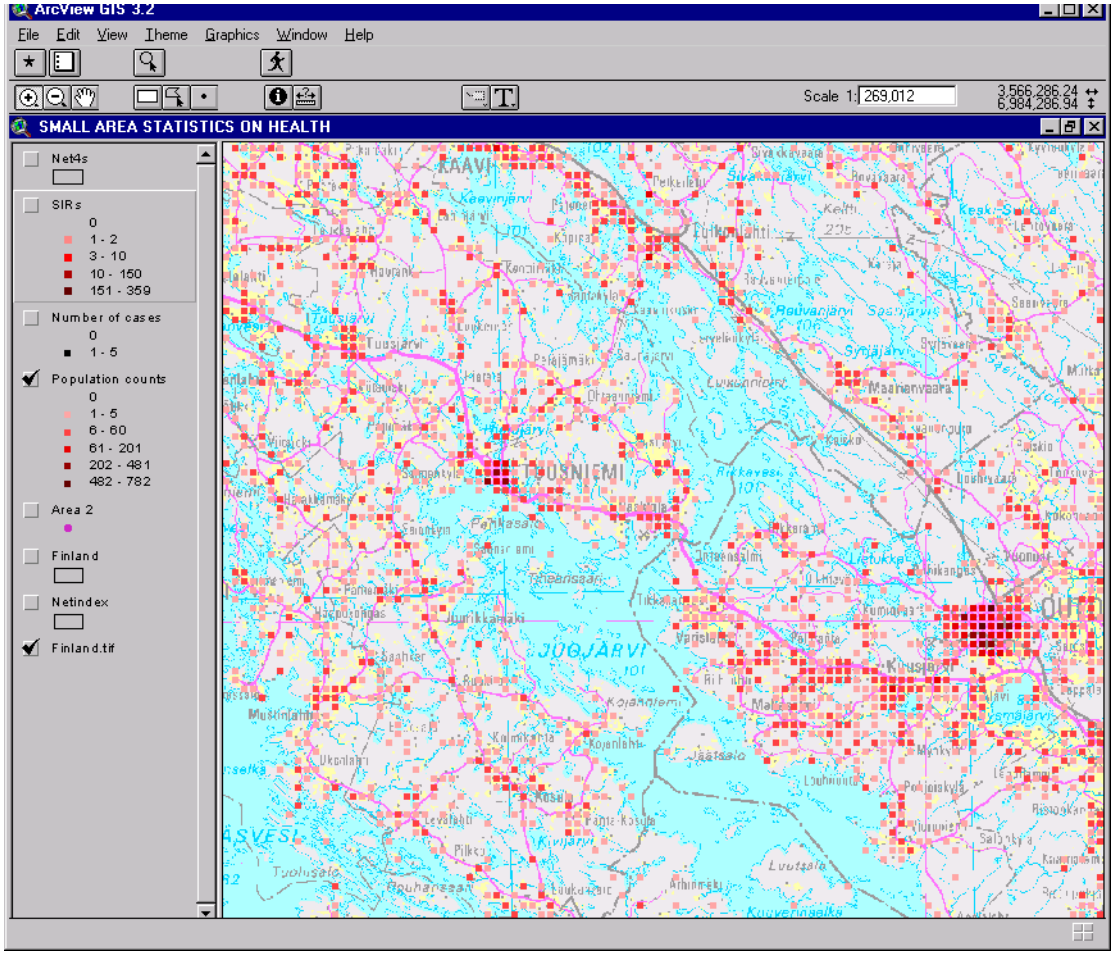
Implementation of a small area system

In Finland, as in many other countries, a local small area system called “Small Area Statistics on Health” (SMASH) (Kokki et al. 2001, Kokki et al. 2002A) had already been initiated before the EUROHEIS - project. The Finnish system provided very similar analytic approaches as the early versions of the UK system called "Rapid Inquiry Facility" (RIF) and, therefore, there was little need to completely rebuild the Finnish system at that early stage.

The SMASH system holds, in squares of size 0.5 km x 0.5 km, register-data on cancer cases in 1981-2000 and on population in 1980, 1990, and 1997 received from the Finnish Cancer Register and the Central Population Register. These data are

georeferenced by metric co-ordinates of residence received from the Statistics Finland. Both, the health data sets, and the population data are broken by sex, age and socio-economical status. Age has been classified in groups of periods of 5-years (0-4,5-9, …,80-84, 85+). Socio-economical status has been classified in 6 classes: (1) farmers, (2) upper clerical workers, (3) lower clerical workers, (4) skilled workers (5) unskilled workers and (6) others. Typical features of high resolution spatial population data in Finland are spatially inhomogeneous populating, small population counts and a large number of inhabited cells (around 80 %), see Figure 4.

Figure 4. A typical sample of high-resolution population data in Finland. Inhabited squares of size 500 m * 500 m are coloured from dark red to light red in five categories (482-782, 202-481, 61-201, 6-61, and 1-5 inhabitants in square). Rest of the area is uninhabited (around 80%).



The SMASH - system was introduced in ISEA/ISEE 2002 conference (Kokki et al. 2002B) and has proven to be useful for rapid investigations. Since the early days of RIF, the number of researchers contributing to its development has grown and its development has enhanced remarkably. We are currently implementing a RIF system in Finland, planning to compare the two systems, and hoping to create an even better tool for risk communication by combining the best characteristics of both systems.

Development of statistical methods for analyses of high resolution and sparse data

The methodological development of SMASH has been focused on solving the problems of sparse high-resolution data: 1) instability in (Bayesian) computation; 2) areas of low information (sparse data); 3) empty pixels. The behaviour of sparse data was demonstrated with lung cancer risk estimates in three areas around former asbestos mine.

A hierarchical Bayesian model based on the idea of borrowing strength was exploited. The idea of this model is in division of risk in terms of area and clustering component. The latter takes into account the possible spatial autocorrelation. With sparse data, the MCMC simulation of the model converged slowly (Kokki et al. 2001). Four methods to improve convergence were investigated (Kokki et al. 2003A): I) a prior for the sum of area level risks with the corresponding SIR as a mean, II) a fixed risk in one of the three areas, III) order restriction for risks, IV) order with high probability. All the methods speeded up the convergence, but the computation was still slow.

As a simpler method a change-point model was applied (Kokki et al. 2003B). In this model the risk is expressed as a non-parametric function of distance from the source. The idea is that the risk changes unknown times within an unknown distance from a source. This model is simple to implement, and the MCMC calculation results in a smooth risk curve.

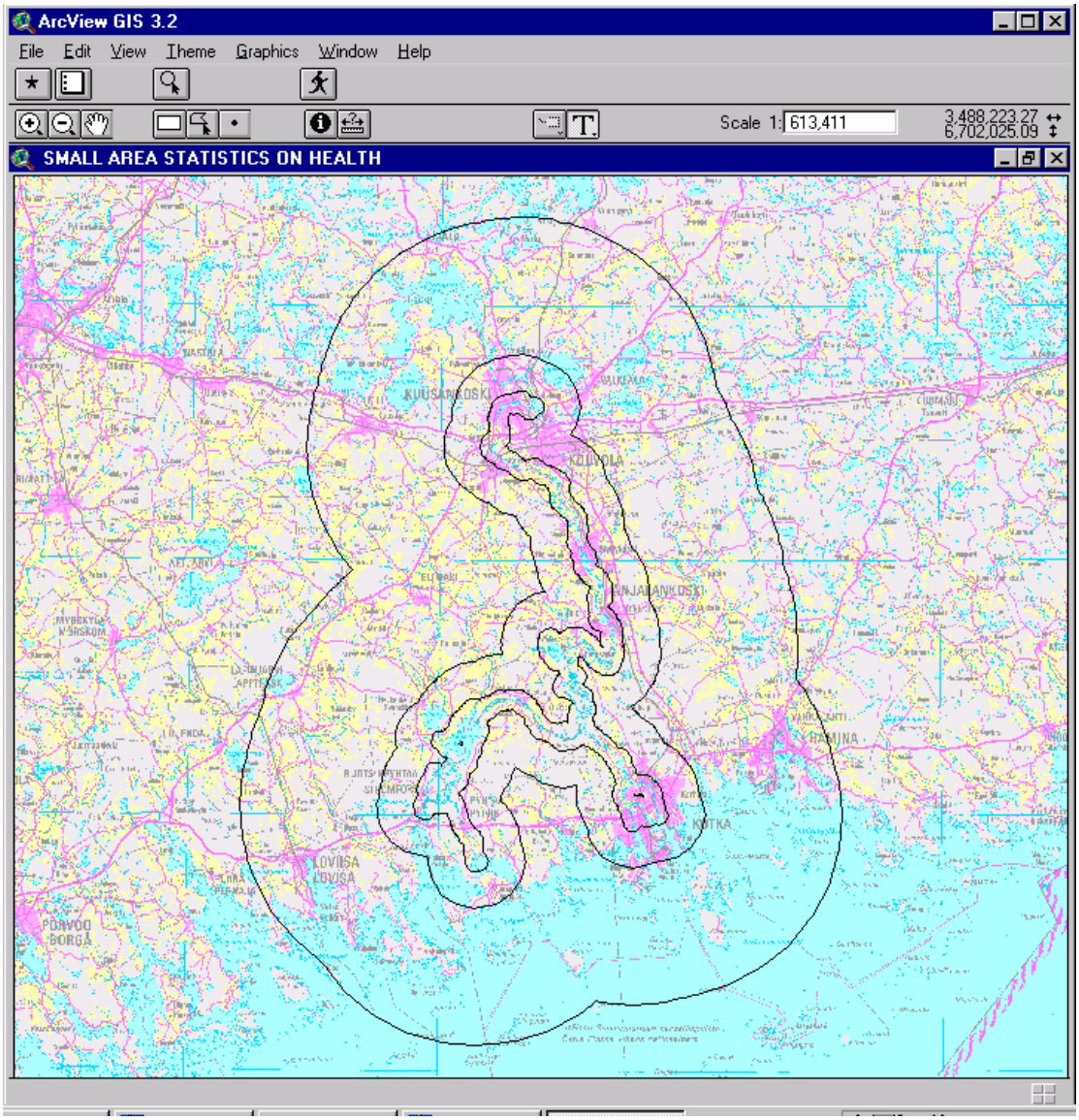
Case study

The overall aim of year three was in evaluation of implemented systems for point-source analyses and disease and exposure mapping using country-specific case studies. The Finnish partner investigated the possible increase in cancer risk following exposure to chlorophenols and PCDD/F emanating from a heavily contaminated river.

In the River Kymijoki, the environmental levels of dioxins and furans, deriving from the production of a chlorophenol product called Ky-5 from 1939 to 1984, are among the highest in the world (i.e., between 0.5 and 350 ng g-1 (I-TEQ)). This study investigates the risk of total cancer and selected cancer subtypes in farmers living near the River Kymijoki, presuming that highest dioxin exposures occur closest to the river.

Exposure assessment was based on the distance of residence to the River Kymijoki in 1980. SMASH was used to create data sets of three cohorts of farmers (living closer than 1 km, 1 to 4 km, or 5 to 19 km from the river, see Figure 2) and their incident cancers between 1981 and 2000. Expected numbers were counted based on incidence rates in all farmers in Finland. Estimates of relative risk (RR) were obtained from Poisson – regression models using observed and expected numbers of cancer and adjusting for sex, age and calendar period.

Figure 5. The study areas in a vicinity of the river Kymijoki (with distances of 0-4 km, 5-9 km, and 10-19 km from the bank of the river) defining the three cohorts.



The RRs for total cancer in farmers were 1.00, 0.97, and 1.21 by decreasing distance to the river; the respective numbers of observed cases (and 95% confidence intervals, CI) were 388 (reference), 134 (0.80-1.18) and 100 (0.97-1.51). For farmers living closest to the river, the RRs were statistically significantly increased during 1981-1990 (RR = 1.40) but not during 1991-2000 (1.09).

Likewise, the RRs were statistically significantly increased for farmers aged less than 45 years (RR = 1.81) in 1980, but not for those aged 45 to 59 years (1.09), or those aged 60 years or more (1.04). Increased (but statistically non-significant) RRs were observed for cancers of the testis, nervous system (significant), breast, pancreas, bladder, liver, ovary, thyroid, prostate, and rectum, and for soft tissue sarcoma, malignant non-melanoma of the skin, Hodgkin's disease and leukaemia.

The results are suggestive of increased cancer risk in farmers living close to the River Kymijoki. However, GIS can only give first approximations of risk and the results need to be confirmed in individual-level studies. The fact that the risk was increased during 1981-1990 but not during 1991-2000 suggests that any increase in risk may already be fading away.

Future

It is our experience that we have reached the end of the first phase of system development. On the way, we have implemented the SMASH system, which has proved a useful tool in assessing cancer risks in freely selected areas in Finland. The EUROHEIS project has been very useful in contributing to the creation of a basic network of researchers. In the future we wish to continue system development, and to implement and use a more comprehensive and even better small area system. Nevertheless, we have come to understand that it is crucially important to consider both potentials and restrictions of the available data and analytical methods and find it very important to create evidence-based guidelines of interpretation for users of such systems as well as to end-users of study results.

Ireland

Alan Kelly
Imanol Montoya
Elaine Hand
Conor Teljeur
Small Area Health Research Unit
Trinity College Dublin

Small Area Health Research Unit
Trinity College Dublin
IRELAND

Summary

This report has been divided in two parts. In the first part, we report on the analysis of data from Ireland, Spain and Finland, in which health outcomes and socio-economic measures were available. A Bayesian smoothing method was applied to each of the outcomes in order to, on one hand, provide a reliable estimate of the Standardised Mortality Ratio (SMR) and, on the other, to ascertain if any of the available socio-economic variables had a significant association with the outcome. This is to establish whether or not it is important to control for socio-economic effects in area-level modelling, and additionally, to ascertain if this holds for a variety of health outcomes from selected Partner countries.

The principal aim of the second part of the report was to consider key statistical methods for spatial point-level data and to develop a realistic simulation environment for assessing practical utility of these methods. Most of our work reported here has focused on the use of Kulldorff's spatial scan statistic attempting to show its lack of power when there may be several or indeed a large number of putative sources of pollution. A new variant on the spatial scan statistic - potentially more appropriate under these circumstances - is proposed. We confine our report to the simulations exercise rather than reporting on a specific Irish application due to issues of confidentiality. A case study based on data kindly supplied by our Danish partner, has been used as an initial text of the method. This latter effort is intended to complement the primary attention of EUROHEIS to area-level modelling where the possibility exists to conduct an analysis with spatial point-level data.

Part I. Analysis of Irish, Spanish and Finnish Data

Partner data sets with relevant health outcomes plus some local measure of socio-economic status at a small-area level were available for regions of Spain, Finland and Ireland only.

1. OVERVIEW OF THE DATA SETS

Ireland

The provisional analyses reported herein relate to 493 DEDs (District Electoral Divisions, i.e. small areas similar to Wards in the UK) of Dublin County and County Borough. Three different health outcomes (*All cause mortality, Heart diseases and Lung cancer*) were analysed taking into account the geographical location and the national deprivation index of each DED.

Spain

The Spanish partner provided vary detailed data for two different regions: Granada and Valencia.

The data from Granada contains mortality (*all cancer and breast cancer*) and a range of socio-economic variables (age, income per capita, illiteracy rates, unemployment rates, activity rate and a rural-urban indicator). All data are at the municipality postcode level covering 168 postcodes.

The data from Valencia contains mortality figures for *leukaemia, cancer of the stomach, bladder and prostate, cerebrovascular and ischaemic heart disease*, and a number of socio-economic variables (proportion of workers in different occupational groups, illiteracy and activity rates). All data are at the municipally postcode level covering 540 postcodes.

Finland

The Finish partner provided data sets relating to regional population and regional *cancer incidence*. Information in both data sets is coded to x and y spatial grid co-ordinates. Both data sets included a socio-economic status variable and a municipality code. The data were aggregated to grid square of 5-km ^{2} in extent giving 60 areas in total.

2. THE BAYESIAN MODELLING FRAMEWORK

A Bayesian smoothing method was applied to each of the data sets in order to, on one hand to provide a suitable estimate of the Standardised Mortality Ratio (SMR) at the small-area level (see details and justification in our first report) and, on the other, to determine if any of the available socio-economic variables were significant related to the health outcome, the nature of such a relationship and whether it would be useful to generalise certain conclusions across the three countries.

The Bayesian Modelling Framework - brief overview:

If the observed count of events in an area i (O_i) is assumed to be Poisson distributed with an expected rate given by

$\mu_i = \rho_i E_i$, where ρ_i is the unknown area-specific relative risk (RR), then this leads to the maximum likelihood estimator

$\rho_i = O_i / E_i$ which is simply the traditional standardised ratio (SMR) for area i.

However, if the ρ_i are not equal (i.e. there is extra-Poisson variation) then this estimator will perform poorly. This will arise when analysing intrinsically rare outcomes or data geo-coded to small geographic areas.

We must model the μ_i (as $\log_e \mu_i$ for convenience) as a Poisson regression model allowing for the additional variation:

$$O_i \sim \text{Poisson}(\mu_i)$$

$$Log_e(\mu_i) = Log_e(E_i) + \alpha + H_i + S_i + \beta x_i$$

The α term is just the overall mean. The second and third terms are random effects and serve to decompose the extra-Poisson variation into:

i) the H_i are independent normally distributed random effects with zero mean and variance λ^2.

This term represents the spatially unstructured extra-Poisson variation- risk is allowed to vary by area (this reflects heterogeneity)

ii) the S_i are also normally distributed random effects but with a mean determined by the mean of the adjacent S_i and variance κ^2.

This term represents local spatially structured variation – risk is allowed to vary smoothly across neighbouring areas (This is reflective of a tendency to spatial clustering)

The last term βx_i allows for dependence upon one or more explanatory. β is the regression coefficient for such covariates.

3. MAIN RESULTS

A tabular summary of the main findings by region is presented below for the given outcomes. Of particular interest is the significance or otherwise of the socio-economic/deprivation term.

Ireland - Dublin County & County Borough

Table 1. All cause mortality, Dublin

Bayesian model with heterogeneity, clustering term and deprivation index						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	-0.364	0.05056	0.002173	-0.4652	-0.3657	-0.2715
Deprivation index	0.112	0.01918	8.66E-04	0.07506	0.1125	0.1508
Heterogeneity	4.653	0.3622	0.008864	3.978	4.647	5.435
Clustering	2229	2174	172	128.4	1546	8292

Table 2. Heart disease, Dublin

Bayesian model with heterogeneity, clustering term and deprivation index						
Parameter	mean	sd	MC error	2.50%	median	97.50%
Constant	-0.3708	0.05982	0.00224	-0.4913	-0.3713	-0.251
Deprivation index	0.1184	0.02289	8.04E-04	0.0729	0.1185	0.1628
Heterogeneity	4.619	0.4571	0.01182	3.772	4.603	5.59
Clustering	2292	2309	202	69.43	1551	8714

Table 3. Lung cancer, Dublin

Bayesian model with heterogeneity, clustering term and deprivation index						
Parameter	mean	sd	MC error	2.50%	median	97.50%
Constant	-0.6293	0.07647	0.003235	-0.7832	-0.6249	-0.4883
Deprivation index	0.2204	0.02794	0.001378	0.1659	0.2198	0.2757
Heterogeneity	6.056	1.125	0.04735	4.254	5.896	8.635
Clustering	1075	1533	168.9	18.16	413.3	5829

In the various models shown above, the level of relative risk is significantly positively associated with the level of deprivation. This means that the higher levels of relative risk are associated with higher levels of deprivation.

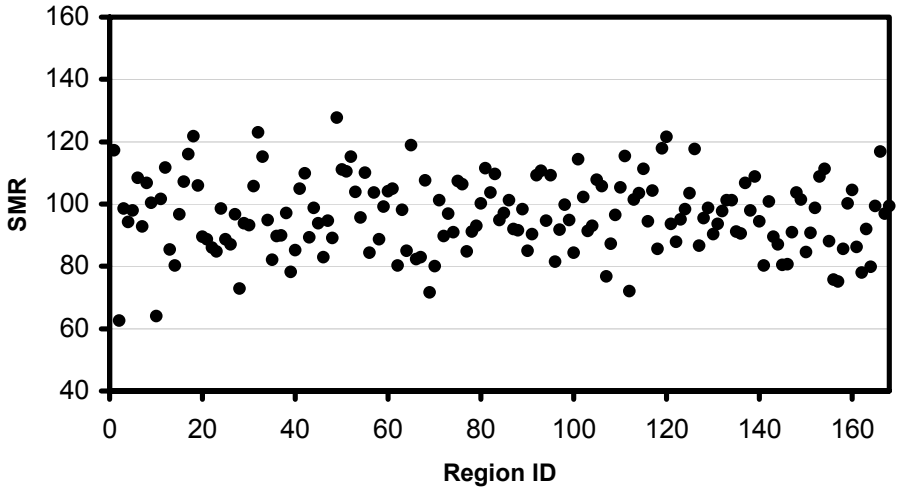
Another important aspect evident from these analyses is that the clustering effect is not significant when the deprivation index is included in the model. This reflects the fact that the deprivation index is already highly *spatially clustered*, confirming that neighbouring areas tend to experience similar levels of deprivation in this region. So it is reasonable to conclude that the tendency for high (or low) levels of relative risk to cluster reflects the tendency for deprived (or affluent) areas to cluster, i.e. “black spots” cluster!

Spain - Granada

In this case information on spatial neighbourhood was not available. Therefore, a simpler (Bayesian) global smoothing model was employed.

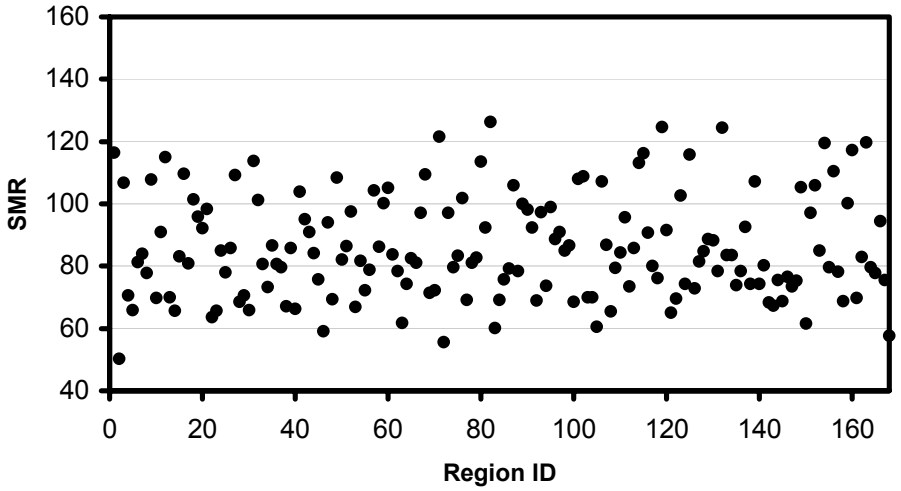
In the analysis of 'all cancers' in Granada, several covariates were individually found to have a statistically significant role. These were: *Age, Income per capita and Unemployment level.* Interestingly, in the analysis of breast cancer mortality only *illiteracy* was found to be significant.

Figure 6. SMRs for 'all cancers' by postcode showing a two-fold variation by area (Granada).



The regions with the highest SMR for 'all cancers' were Cúllar-Vega, Calicasas, Armilla, Padul and Fuente Vaqueros. And the regions with the lowest SMRs were Alamedilla, Algarinejo, Gor, Murtas and Busquistar.

Figure 7. SMRs for breast cancer showing a greater than two-fold variation by area. (Granada).



The regions were the highest SMRs for breast cancer were Huetor Vega, Otura, Pulianas, Villamena and Granada And the regions with the lowest SMRs were Alamedilla, Guadahortuna, Zagra, Cortes de Baza and Illora.

The main results of each of the analyses can be seen in the following tables.

Table 4. Mortality from 'All cancers' (Granada).

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	-0.03935	0.09875	0.01564	-0.2571	-0.03758	0.1258
Age rate	-0.0857	0.02833	0.002984	-0.1418	-0.08561	-0.03107
Income	3.19E-04	1.30E-04	2.05E-05	9.22E-05	3.08E-04	6.07E-04

Illiteracy	3.33E-04	7.44E-04	8.07E-05	-0.00111	3.18E-04	0.001852
Unemployment	-0.00962	0.003681	3.47E-04	-0.01696	-0.00951	-0.00267
Activity rate	0.001222	0.004344	6.03E-04	-0.00754	0.00112	0.009777
Rural-Urban	3.70E-04	7.19E-04	7.60E-05	-9.92E-04	3.50E-04	0.001903

These results give a strong indication that the probability of dying of cancer differs between *Rural* and *Urban* areas. Rural areas, in which the proportion of older individuals is greater, with a lower income per capita and a higher rate of unemployment, constitute the areas with the lowest SMRs.

Not surprisingly, there are significant interdependencies between the socio-economic variables and this has implications for the interpretation of the results. For example, in the model with all predictors present, the "*% of Farmers*" is not statistically significant, but in the bivariate model with "*% of Farmers*" alone this variable is statistically significant (with a negative coefficient). This inevitably complicates the interpretation of the findings and local knowledge becomes critical to a correct understanding of model results.

The next table shows the correlation matrix between all the variables and the SMRs:

Table 5. Correlation matrix (SMR all cancers).

	SMR	Age rate	Income	Illiteracy	Unemployment	Activity Rate	% Farmers
SMR	1.000						
Age rate	-0.277	1.000					
Income	0.386	-0.262	1.000				
Illiteracy	-0.133	0.065	-0.351	1.000			
Unemployment	-0.323	0.146	-0.438	0.195	1.000		
Activity Rate	0.316	-0.502	0.595	-0.304	-0.299	1.000	
% Farmers	-0.223	0.400	-0.429	-0.204	0.342	-0.488	1.000

Table 6. Breast cancer.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	0.5494	0.4629	0.05161	-0.3125	0.5978	1.365
Age rate	-0.01058	0.1604	0.01182	-0.3505	-0.00691	0.2901
Income	-1.53E-04	5.18E-04	4.99E-05	-0.00115	-1.48E-04	8.89E-04
Illiteracy	-0.0071	0.003544	1.95E-04	-0.01431	-0.00702	-3.53E-04
Unemployment	-0.0324	0.01764	9.40E-04	-0.06752	-0.03293	0.002708
Activity rate	-0.00108	0.01706	0.001381	-0.03268	-0.00133	0.03282
Rural-Urban	-0.00329	0.0032	1.56E-04	-0.00985	-0.00325	0.002867

Table 7. Correlation matrix (SMR breast cancer).

	SMR	Age rate	Income	Illiteracy	Unemployment	Activity Rate	% Farmers
SMR	1.000						
Age rate	-0.108	1.000					
Income	0.113	-0.262	1.000				
Illiteracy	-0.122	0.065	-0.351	1.000			
Unemployment	-0.147	0.146	-0.438	0.195	1.000		
Activity Rate	0.129	-0.502	0.595	-0.304	-0.299	1.000	
% Farmers	-0.098	0.400	-0.429	-0.204	0.342	-0.488	1.000

For both 'All Cause' and Breast cancers, age is negatively related to relative risk - in the former case, significantly so. For Breast cancer outcomes, only Illiteracy rates appear to be significantly (negatively) associated with outcome. The lack of association of risk for Breast cancer with specific socio-economic factors has also been observed in Irish data.

Spain- Valencia

As in the previous analyses, there was no neighbourhood matrix information available, a (Bayesian) global smoothing model was employed. Table 8 shows a summary of the significant covariates (grey cells) for the Valencia data by outcome ('+' indicating positive association and '-' indicating a negative association).

Table 8. Valencia data analysis – result summary.

	Activity level	**Illiteracy rate**	**Employment Type**			
			% employed in Service	% employed in Construction	% employed in Agriculture	% employed in Industry
Ischaemic Heart disease	-		+	+	+	-
Bladder Cancer	-					
Leukaemia	-					
Prostate cancer	-					
Stomach cancer	-					

From our analyses, only '*Activity Level'* was found significant to be consistently significantly associated (negative coefficient) with these outcomes. Due to the sign of the coefficient for this factor, postcodes with low activity levels will experience higher relative risks for these cancers.

'Illiteracy' was not found to be significant in this region, although the *'% Employment'* by sector (Service, Construction, Agriculture and Industry) did have a significant effect in relation to mortality from Ischaemic Heart disease (positive coefficient for 3 of 4 categories), showing that areas in which with the highest percentage of people working in the percentages employed in the Service, Construction and Agriculture sectors will experience elevated risks for hearth disease, whereas by contrast, areas with high percentages employed in the Industry sector appear to experience the lowest SMRs.

The next tables show the main results for each of the analyses and the correlation matrix between all the variables and the SMRs:

Table 9. Ischaemic heart disease.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	0.2158	0.4861	0.08638	-0.3245	0.0198	1.214
Activity level	-0.01942	0.01098	0.001959	-0.03529	-0.01787	-0.00147

Illiteracy	8.29E-04	0.001194	1.92E-04	-0.00175	0.00108	0.00268
% Service	0.00941	0.003797	6.76E-04	4.53E-04	0.0103	0.01455
% Construction	0.01457	0.003352	5.13E-04	0.006755	0.01487	0.02041
% Agriculture	0.01194	0.002522	4.04E-04	0.006232	0.01245	0.01559

Table 10. Correlation matrix ischaemic heart disease.

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
SMR	1.000						
Activity Level	-0.431	1.000					
Illiteracy Rate	0.083	-0.110	1.000				
% Service	-0.103	0.294	-0.077	1.000			
% Construction	0.140	-0.190	-0.008	0.002	1.000		
% Agriculture	0.195	-0.367	0.128	-0.561	-0.152	1.000	
% Industry	-0.194	0.245	-0.074	-0.242	-0.281	-0.578	1.000

Table 11. Bladder cancer.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	0.7016	0.4679	0.08282	-0.3587	0.8545	1.242
Activity level	-0.02037	0.005345	9.35E-04	-0.02586	-0.02185	-0.00429
Illiteracy	-8.84E-04	0.001826	2.18E-04	-0.00408	-9.38E-04	0.003156
% Service	0.005022	0.004337	7.36E-04	-0.00302	0.004445	0.01409
% Construction	-7.90E-04	0.006552	7.79E-04	-0.01223	-0.00102	0.0126
% Agriculture	0.007097	0.003982	5.28E-04	0.00027	0.007022	0.0152

Table 12. Correlation matrix bladder cancer.

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
SMR	1.000						
Activity Level	-0.131	1.000					
Illiteracy Rate	-0.034	-0.110	1.000				
% Service	-0.037	0.294	-0.077	1.000			
% Construction	-0.015	-0.190	-0.008	0.002	1.000		
% Agriculture	0.093	-0.367	0.128	-0.561	-0.152	1.000	
% Industry	-0.067	0.245	-0.074	-0.242	-0.281	-0.578	1.000

Table 13. Leukaemia.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	1.364	0.4257	0.07428	0.1011	1.425	2.011
Activity level	-0.02131	0.007312	0.001281	-0.03341	-0.02204	-0.00264
Illiteracy	0.002143	0.001762	1.53E-04	-0.00131	0.002116	0.005717
% Service	-0.0056	0.003396	5.36E-04	-0.01328	-0.00575	0.001083
% Construction	-0.00195	0.007456	7.77E-04	-0.01673	-0.00203	0.0126
% Agriculture	-0.00509	0.003882	3.39E-04	-0.01294	-0.00503	0.001946

Table 14. Correlation matrix leukaemia

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
SMR	1.000						
Activity Level	-0.090	1.000					

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
Illiteracy Rate	0.114	-0.110	1.000				
% Service	-0.099	0.294	-0.077	1.000			
% Construction	0.070	-0.190	-0.008	0.002	1.000		
% Agriculture	0.088	-0.367	0.128	-0.561	-0.152	1.000	
% Industry	-0.045	0.245	-0.074	-0.242	-0.281	-0.578	1.000

Table 15. Prostate cancer.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	1.414	0.7468	0.1335	-0.1151	1.628	2.261
Activity level	-0.03046	0.01105	0.001967	-0.04153	-0.03529	-0.0029
Illiteracy	0.001723	0.001683	2.15E-04	-0.00153	0.001756	0.00508
% Service	2.39E-04	0.00373	6.27E-04	-0.00593	5.33E-04	0.006754
% Construction	-0.00479	0.007844	0.001174	-0.01862	-0.00485	0.01268
% Agriculture	0.01295	0.003907	5.89E-04	0.005803	0.01298	0.02037

Table 16. Correlation matrix prostate cancer.

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
SMR	1.000						
Activity Level	-0.145	1.000					
Illiteracy Rate	-0.009	-0.110	1.000				
% Service	-0.152	0.294	-0.077	1.000			
% Construction	0.049	-0.190	-0.008	0.002	1.000		
% Agriculture	0.148	-0.367	0.128	-0.561	-0.152	1.000	
% Industry	-0.057	0.245	-0.074	-0.242	-0.281	-0.578	1.000

Table 17. Stomach cancer.

Bayesian global model with heterogeneity						
Parameter	Mean	sd	MC error	2.50%	median	97.50%
Constant	0.5116	0.3707	0.06549	-0.09339	0.5398	1.174
Activity level	-0.01198	0.005799	0.001019	-0.02087	-0.01343	-8.15E-04
Illiteracy	-9.24E-04	0.001347	1.41E-04	-0.00355	-8.70E-04	0.001627
% Service	0.001764	0.002174	3.40E-04	-0.00312	0.002012	0.005401
% Construction	-0.00521	0.005668	7.45E-04	-0.01671	-0.00523	0.00541
% Agriculture	0.007798	0.002625	2.77E-04	0.002435	0.00776	0.01261

Table 18. Correlation matrix stomach cancer.

	SMR	Activity Level	Illiteracy Rate	% Service	% Construction	% Agriculture	% Industry
SMR	1.000						
Activity Level	-0.134	1.000					
Illiteracy Rate	-0.062	-0.110	1.000				
% Service	-0.042	0.294	-0.077	1.000			
% Construction	0.040	-0.190	-0.008	0.002	1.000		
% Agriculture	0.066	-0.367	0.128	-0.561	-0.152	1.000	
% Industry	-0.057	0.245	-0.074	-0.242	-0.281	-0.578	1.000

Finland

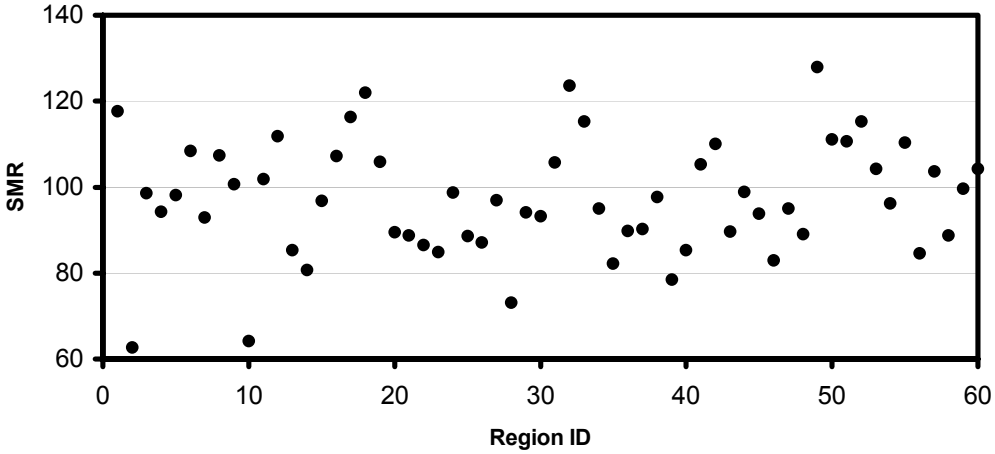
The data supplied by the Finland partner included geographic co-ordinates at a very fine level, and so the data were aggregated to 60 x 5 km^2 grid-squares and the

standardised mortality ratios computed accordingly. Mortality for all-cause cancer was provided.

Socio-economic status variables for each individual was supplied as part of the outcome data set, at 5 year intervals from 1970 to 1995. However, in the population data file, only one SES variable was available for specific years (i.e., 1980, 1990 and 1997). As the outcome data was aggregated, the median SES value, from the 1997 and 1990 population data, was used. When the 1997 SES median values where examined it was found that approximately 58% of the 60 areas had a socio-economic status of 3-4 while 11% of the areas had an extreme SES of 1 or 6. An almost identical distribution was found for the 1990 median SES values.

A Bayesian global smoothing method was employed in the analysis of the cancer data. Neither the median SES for 1990 nor the median SES for 1997 proved to have a significant impact on the Bayesian-adjusted SMRs. This result must of course be treated with considerable caution, given that the population data does not necessarily correspond on a one-to-one basis with the cancer data as noted above. An important consideration that could impact on the analysis in the change in SES for some patients during the period and this is something that it was not possible to examine in this study, but should be possible for the Finnish partner to investigate in due course.

Figure 8. SMR for all cancer showing an approximate 2-fold variation in relative risk.



.

In summary

The first point to make is the limited availability of outcome/socio-economic data supplied to this partner for purposes of investigating the general issue of the role and importance of controlling for deprivation (or possible the constituent elements of a deprivation index) in small-area level modelling.

There is of course considerable international evidence (reviewed in our first report) of the role of material and social deprivation in connection with risk of chronic disease. There is also explicit support for this finding in relation to area-level analyses of chronic diseases - while controlling for deprivation - in small-area level modelling since the development of material deprivation indices in the UK during the early

1990s and more recently in Ireland. There have been limited efforts to formally construct locally tailored deprivation indices in other European countries to date, although in the context of this project, the partners expressed considerable interest in the concept. In practice, this has proved problematic given their limited access to the necessary detailed area-level date for their own regions/countries. Where data have been provided (by our Spanish partner for 2 regions and by our Finnish partner for one region and our own data for 1 region) the extent and appropriateness of the variables supplied has not been ideal. By definition, a deprivation index is constituted for a limited number (typically four or five) of socio-economic variables (as discussed in more detail in our first report). This precluded the development of such an index for the Finnish data, with only a single socio-economic variable available. In the case of the two Spanish regions, several potentially relevant socio-economic variables were made available for consideration. Unfortunately, an attempt to form a single index from these (specific to each region) was not successful in that the percentage of total variation accounted for by the first principal component fell far short of a reasonable value (around 70% - again, please refer to our first report for a discussion on this point) to allow for the formation of a useful index, hence the reliance on the 'raw' original socio-economic variables in the analyses reported above.

It is very noticeable, from the results of the analyses of the mortality outcomes for the two Spanish regions that the role of the individual socio-economic variables is far from consistent. Not only does the sign of the coefficients in the various models tend to change for most of these predictors, but the coefficients may or may not be statistically significant. The inter-relationships between these variables and - as a set of predictors - their association with the various outcomes, is clearly complex and also evidently dependent on urban-rural differences. This warrants further investigation locally.

The anticipated relationship between area-level relative risk of disease and deprivation is more evident and interpretable in the context of the Irish data reported here, and we have no reason to doubt that this finding is generalisable to other countries should an appropriate series of socio-economic data become available at the small-area level.

Part II. Applying spatial cluster detection methods: sense and sensibility

The principal aim of this report was to consider all the available statistical methods for spatial point-level data and to develop a realistic simulation environment for assessing practical utility of these methods.

Most of the work has focused on Kulldorff's spatial scan method showing its lack of power when there may be several or indeed large number of clusters. A new method for these situations has been proposed.

1. POINT-LEVEL DATA VS. AREA-LEVEL DATA

Area-level data arise when information on disease cases is presented only in a spatially aggregated form, typically in counts of the number of cases in each of a set of sub-regions which partition the study region. On the other hand, point-data would consist of the reference locations for every member of the population at risk. This is seldom feasible. An alternative is to conduct a case-control study, in which we record the locations of all know cases, and a random sample of non-cases, termed controls.

Area-level data has been more frequently used because it was easier to obtain. In general, administrative systems based on address are used. Typically some link between addresses and a set of area codes is available. In the United Kingdom for example, postcodes have been extensively used for this purpose. Similar systems exist in several other countries.

Another reason for the use of area-level data has been that covariate information is not often available for point data.

Therefore, since area-level data has been more frequently used, most of the statistical methods for spatial cluster detection have been developed for this kind of data.

However some problems arise with the use of area-level data:

- Boundary selection arises as a problem whenever there is a range of levels at which data can be mapped and analysed. For example, in the United Kingdom, census data are available at enumeration district, ward, district, county, region and country level, among others. While choice of area for analysis may be forced by limitations in existing data, or by computational considerations, it is important to be aware that this is not an insignificant choice, and that it is possible to get different results at different levels.

- Administrative boundaries can and do change over the time to reflect changes in the population. Therefore, for studies extending over any long period of time it may be necessary to decide on a fixed set of boundaries.

- While UK postcodes area advertised as being accurate on average to ±100 m, manually imputed grid references, still a significant proportion of the total,

have a nominal accuracy of ± 400m. This is large enough to put some postcodes in the wrong ward, and occasionally the wrong district, or even in the wrong county.

Nowadays point-level data is being collected more often than before. In the UK, a system (AddressPoint) exists with nominal 1 m precision, although it is not available currently for routine analyses. In Sweden, addresses databases are available commercially, covering most urban areas.

2. AVAILABLE SPATIAL CLUSTER METHODS FOR POINT-LEVEL DATA

Main spatial cluster methods for point-level data:

General methods: explore clustering without pre-determined hypotheses about cluster location

I. **Global methods:** detect clustering throughout the study area regardless of their specific location or spatial extent.

- *Ripley's K function*
- *Cuzik & Edwards*

II. **Local methods:** detect clustering limited to geographically restricted areas within the study.

- *Kulldorff's Scan Statistic*
- *Tango's test*

Focused methods: detect clustering around specific locations.

- *Diggle's method*

3. KULLDORFF'S SPATIAL SCAN METHOD

Most of the work done in this project has been using Kulldorff's scan statistic. It has become a popular method for detection and evaluation of disease clusters, and it is now used by many health departments and academic epidemiologists both nationally and internationally.

Some of the benefits of its uses are that it can be used for area-level data and individual-level data, it can be used to detect spatial, temporal and space-time clusters, it is a local test, which will pinpoint the most likely area in which there may be a cluster but it can be also used as a focused test to some extend, using the location of the putative sources.

Method

The purely spatial scan statistic imposes a circular window on the area. The window is in turn centred on each of several possible grid points positioned throughout the study region. For each grid point, the radius of the window varies continuously in size

from zero to some upper limit. The likelihood function is maximized over all these windows, and the window with the maximum likelihood constitutes the most likely cluster. Its distribution under the null hypothesis is obtained by repeating the same analytic exercise employing Monte Carlo simulation.

4. SIMULATION ENVIRONMENT DEVELOPED

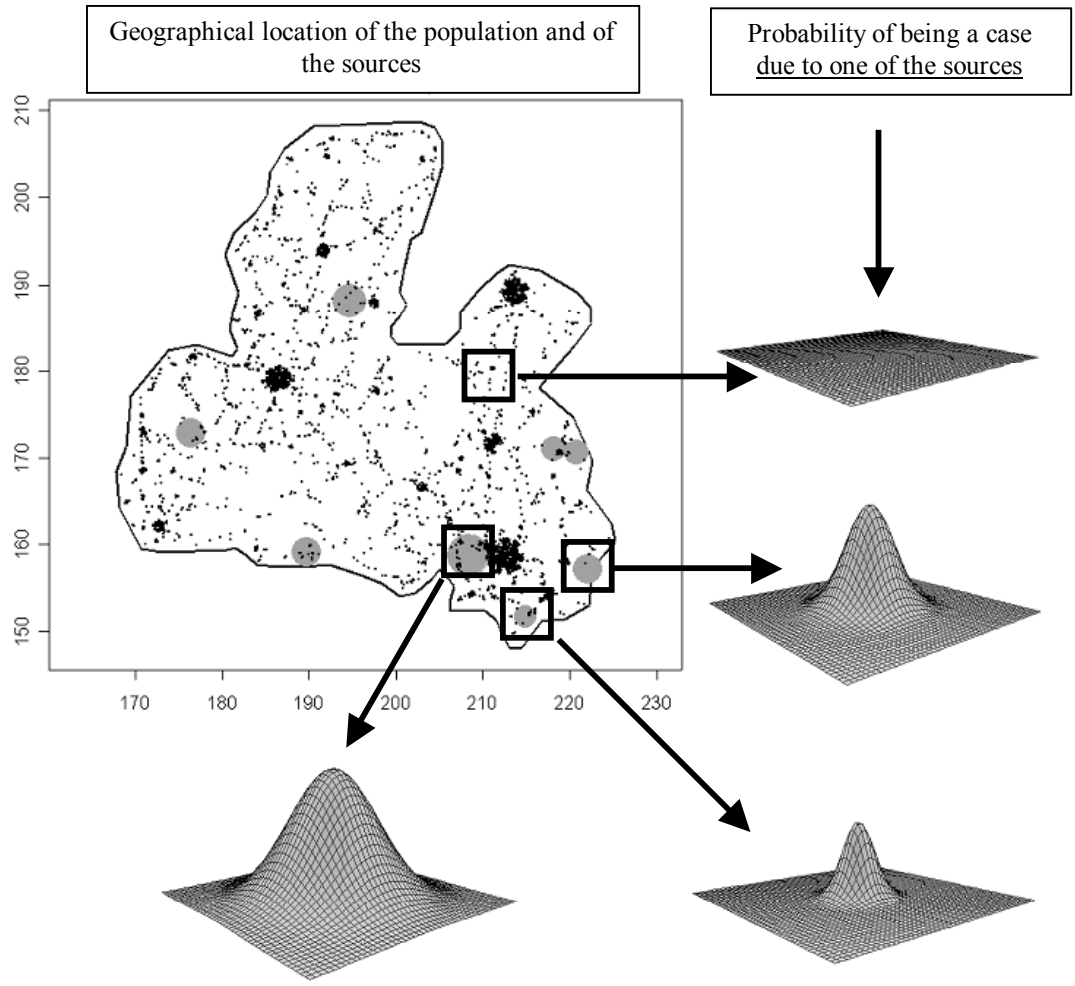
One of the aims of this project was to assess the reliability of the available statistical methods, in particular *Kulldorff's scan test statistic,* when there may be several or indeed a large number of clusters.

A geographical area (modelled on a county in Ireland) was created with 55,681 individuals. Households were distributed realistically. Roads, mountains, lakes, towns, villages were introduced.

From this population, 100 case-control data sets were simulated. In the simulation algorithm[1] several parameters were taken into account: number of cases and controls, number of parent locations, dispersion parameter, cluster cases/ random cases ratio, distance between parent locations and distance from a household to a parent location.

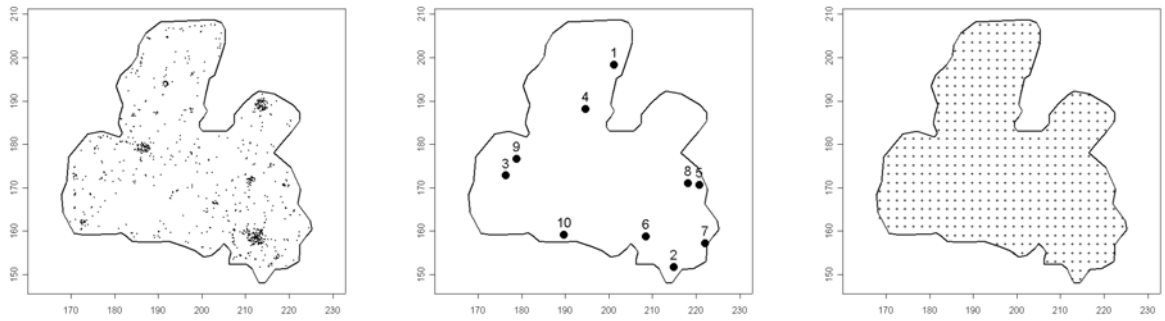
[1] *See Appendix 1A. Simulation algorithm*

Figure 9. Geographical area in which cases and controls will be selected and probability of being a case due to one of the sources.



Then, to each data-set, Kulldorff's scan test statistic was performed using SaTScan software. This test was applied in three different ways (Figure 10). The first one using the geographical coordinates of the cases and controls, the second one using the geographical coordinates of the parent locations and the third one, using the geographical coordinates of the centres of a grid dividing the area.

Figure 10. Data used for Kulldorff's scan test statistic. (Geographical location of the cases, geographical location of the parent locations and geographical location of the centres of a grid dividing the area)



5. PRELIMINARY RESULTS OF POWER ESTIMATION

Kulldorff's scan statistic main results[2]

To make the results the simplest as possible, the location of the cases was classified in rural, urban and mixed. Urban data was found in areas in which the density of population was relatively high, for example in towns. Rural data was found in areas in which the density of population was relatively low and mixed data a combination of these last two.

- No clustering in the data: No significant clusters were found. However several non-significant clusters were pinpointed because the relatively risk in these areas was high.

- Small number of clusters in the data: The only situation in which significant clusters were found was when the parent location was in an urban area. When the clusters where located in rural areas Kulldorff's scan test statistic did not detect any significant cluster[3].

- Large number of clusters in the data: Significant clusters were found. One, two or three significant clusters were pinpointed, always in urban areas. Other possible, but non-significant clusters, were also detected.

Kulldorff's scan statistic is designed to find the most likely cluster, even though there could be other clusters. It is based on the likelihood ratio test, which tests the null hypothesis against the alternative hypothesis that there is only one cluster. Due to

[2] *See Appendix 1B. Graphical example of Kulldorff's scan statistic test using SaTScan*
[3] *See Appendix 1C. Kulldorff Scan Statistic difficulties finding a significant cluster in small-populated areas*

this, the p-values of the secondary clusters should be considered as being rather conservative. (SaTScan User Guide)

Another result is that due to the circle shape used in Kulldorff's scan test, it tends to identify, as the most likely cluster, a much larger cluster than expected from the observed disease map by absorbing other clusters and neighbouring regions with non-elevated risk of disease occurrence.

A key finding of this study using Kulldorff's method is that when there are a large number of clusters, it cannot identify the location of most of them, although it could locate some of the urban clusters.

6. OVERVIEW OF A PROPOSED NEW METHOD

Introduction

A focused method is being developed using the idea behind the scan statistic but applying it to a data set when we think that there may be several or indeed a large number of clusters.

Data used

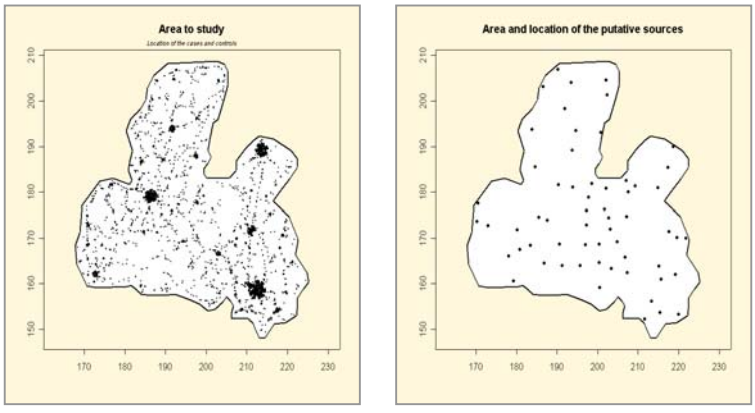
The same 100 simulated data sets used for assessing Kulldorff's scan statistic method.

Method

Suppose that $x_i : i = 1,\ldots,C$, denote the locations of all known C cases of a disease within a geographical region and that $x_i : i = C+1,\ldots,N$, denote the locations of N-C controls, defined to be a random sample from the population at risk.

Let $s_i : i = 1,\ldots,S$, denote the locations of the S putative sources. (Figure 11)

Figure 11. Geographical location of cases, controls and putative sources.



Then we generate an infinite number of circular zones with equal radius around the putative sources, letting the radius of these circles vary from zero upwards. Each time that we change the radius a new area will be generated containing all the circular zones around the sources (Figure 11). Let Z be all the areas generated.

Let c_z denote the number of cases in an area z and n_z the number of individuals, cases and controls, in the same area. The probability of being a case in this area z is p and q denotes the probability for all individuals outside the zone.

For the Bernoulli model, we can express the likelihood function as

$$L(z,p,q) = (c_z / n_z)^{c_z} (1 - c_z / n_z)^{(n_z - c_z)} ((C - c_z)/(N - n_z))^{(C - c_z)} (1 - ((C - c_z)/(N - n_z)))^{(N - n_z) - (C - c_z)} \quad (1)$$

Following the Kulldorff and Nagarwalla methodology, we use the likelihood ratio test statistic, which is

$$LRT(z) = \frac{\sup_{z \in Z, p > q} L(z,p,q)}{\sup_{p=q} L(z,p,q)} \quad (p,q \in [0,1]) \qquad (2)$$

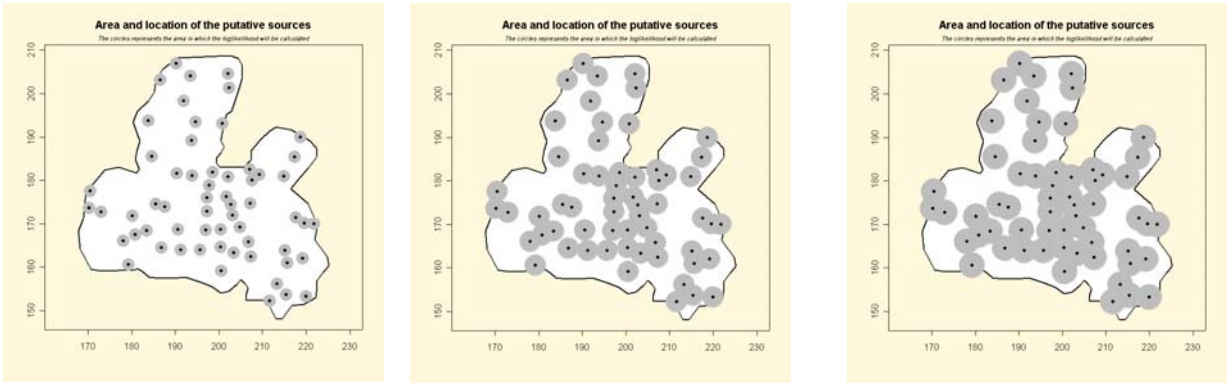
As the most likely area we will pick the area $\hat{z} \in Z$,for which the likelihood ratio test statistic is maximized. This area is the less likely to have occurred by chance and it also gives the most likely radius in which the effect of the sources can be experienced. We can write the test statistic as

$$\lambda = \max_z LRT(z) \qquad (3)$$

As a difference from Kulldorff and Nagarwalla we are not interested on evaluating if this area is the area with the highest likelihood ratio test in the whole geographical region.

We are interested on testing if the risk around these *S* putative sources is higher than the one expected in the area generated around any *S* points in the whole region. Let *u* denotes the area generated around any *S* points in the region of study.

Figure 12. Three (grey) areas containing all the circular zones with equal radius around the putative sources.



The alternative hypothesis is $H_1 : p_z > p_u$, where p_u is the probability of being a case in the area *u*. The null hypothesis is that $H_1 : p_z = p_u$.

The distribution of λ has no simple analytical form, and thus we use the Monte Carlo method to sample from the exact distribution of λ. Supposing that *j* simulations are done, then *j* sets of putative sources will be generated randomly in the same geographical region, with two conditions: The number of random putative sources must be equal to the real number of sources, *S;* and that the minimum distance between the random sources cannot be less than the minimum distance between the real ones. This last condition will ensure that the overlapping effect between two sources in the random sets will be similar to the real one. With each set the likelihood ratio test statistic will be maximized as in the real case by generating infinite areas changing the radius.

The p-value is obtained through Monte Carlo hypothesis testing, by comparing the rank of the maximum likelihood ratio test from the real set of putative sources with the maximum likelihood ratio tests from the random sets.

Illustration

Proposed method main results

Bear in mind that in all the data sets there were 1000 random controls and 1000 random cases.

- Data sets with clustering using the real location of the sources

Number of parent locations	Total number of cluster cases in the whole data	Significant clustering around the sources
from 10 to 80	from 40-50 upwards	Yes (98%)
less than 10	from 40-50 upwards	Most of the times (91%)
less than 10	Less than 40-50	Sometimes (63%)

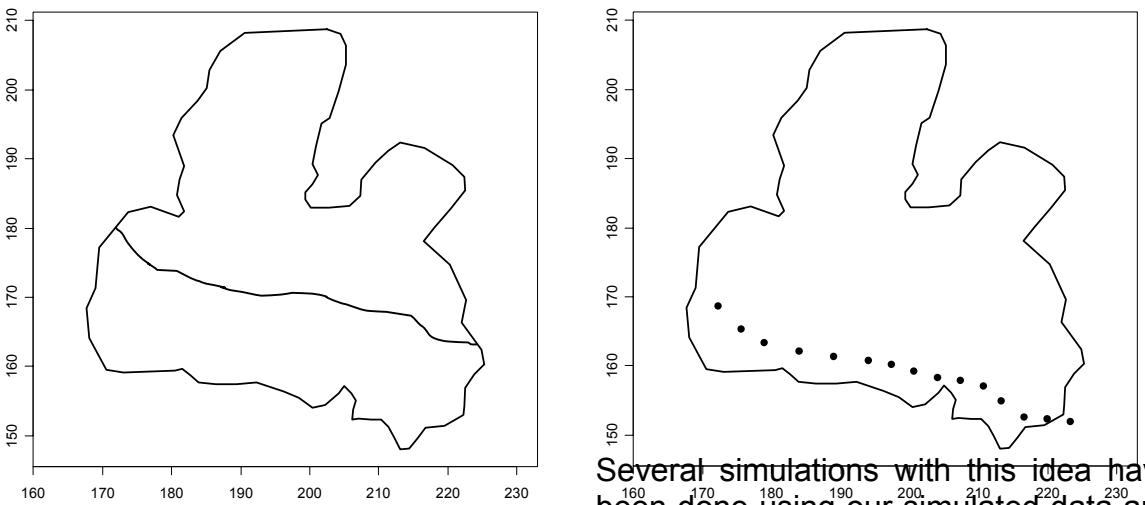
- Data sets without clustering using the location of random sources

It almost never rejects the null hypothesis of no clustering around the sources. From 90 simulations done until now, a significant p-value have been obtained only in 3 of them.

Applications

This focused test could be applied to a data set when we think that there may be a large number of clusters, but also in cases such as clustering close to roads and rivers. In these last cases, Kulldorff's scan statistic test has been proved to be inefficient due to the circle shape used. As an example we could think of our study area, in which we want to know if there is any kind of clustering along a road. We could select as sources for our test 10 or 20 points along the road and perform the test (Figure 13).

Figure 13. Selection of test points for tst pf clustering along a road.



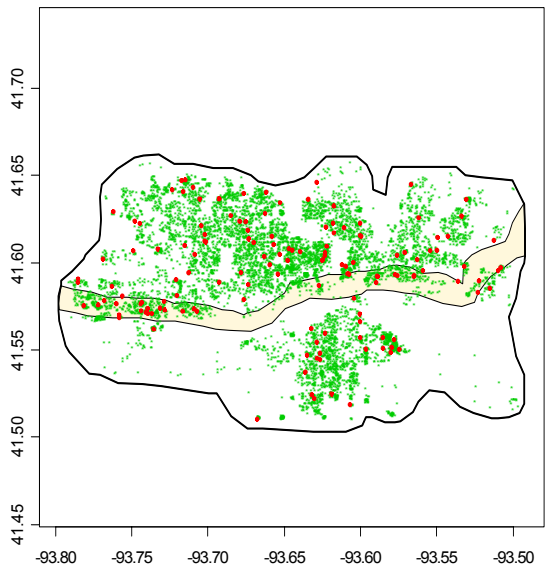
Several simulations with this idea have been done using our simulated data and a data set with a non-uniform distribution

of 8,689 people at risk and until now the results are very satisfactory.

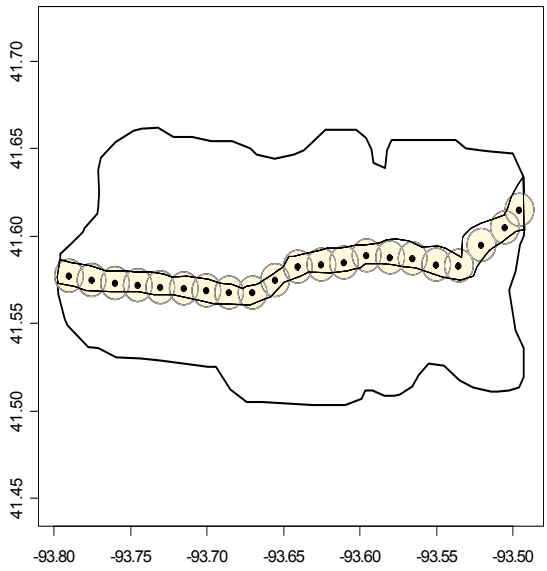
This last data is the same that the one used by Geoffrey H. Smith in his paper *"Disease Cluster Detection Methods: The impact of Choice of Shape on the Power of Statistical Tests"*[4] (*http://www.cobblestoneconcepts.com/ucgis2summer/smith/SMITH.HTM*).
In his paper 72 deaths were simulated from the 8.689 individuals at-risk following a modification of the uniform probability distribution such that each person at risk outside the hot spot has a risk or 1 and those inside the hot spot have a risk of 2. The hot spot contains 1020 of the at-risk population within its boundaries.

In our simulations we have followed the same idea of more probability of death inside the hot spot. Simulations have been done with a total number of deaths (cases) of 36, 72, 144, 288,576,1152 and 2304 (Figure 14).

Figure 14. Illustration of simulation.



Location of the population, cases and hot spot



Location of the sources for the test and estimated radius

[4] A poster presented at the UCGIS 2001 Summer Assembly

From 100 simulations done for each number of deaths the results are:

Number of deaths in the data	Power of the new method (The Relative risk within the hot spot is equal to 2)	Power of the new method (The Relative risk within the hot spot is equal to 3)	Power of the new method (The Relative risk within the hot spot is equal to 4)
72	0.41	0.59	0.70
144	0.62	0.78	0.88
288	0.92	0.95	0.97
576	0.98	0.99	1
1152	0.99	1	1
2304	1	1	1

KEY STUDY: DANISH DATA

This is an example in which both methods have been applied.

Data sets

Population data set
Data set with the location of 109,879 individuals from the year 1986 to 1998. Due to the change of address of some individuals, the total number of rows in the data set was 251,134.

Main variables:

CPR	unique 10 digit Central Population Register No. Digit number 1 & 2 is date of birth, number 3&4 is the month and number 5&6 is the year, last Digit indicates the gender by: use of an even digit indicates a female and an odd one indicates a male
Tilflyttet	date of moving in or birth
Fraflyttet	date of moving out or dead
X,Y	indicate the geographical coordinates
Datoforskl	difference in days between moving in and out, i.e. a simple calculator built into the system

Cases data set (All cancers except skin cancers)
Two data sets with the location of cases from the year 1986 to 1998. The first one with no lower limit for time of residence (1,769 cases) and the second one with a residence time of 2 or more than 2 years (1,112 cases).

Sources location

- Kolding Affaldskraftvarmeværk (**Municipal Incinerator**)
  Coordinates: X: 268.668,82 Y: 120.424,37

- UNISCRAP A/S Genvindingsindustri (**Car Scraping Plant**)
  Coordinates: X: 268.893,71 Y: 120.485,88

- STENA Aluminium (**Aluminum Recycling Plant**)
  Coordinates: X: 268.422,37 Y: 120.239,36

Table 19 Main types of Cancer.

Cancer type	Number of cases	Percentage of cases
Lung	171	15.38 %
Breast	162	14.57 %
Colon	88	7.91 %
Bladder	68	6.12 %
Prostate	65	5.85 %
Melanoma	39	3.51 %
Rectum	39	3.51 %
Brain	36	3.24 %
Kidney	35	3.15 %
Cervix	34	3.06 %
Stomach	33	2.97 %
…		
TOTAL	1112	100%

Spatial cluster analysis

Looking at the geographical map Figure 15) it was found that all the selected cases came from a small region. Consequently, for the spatial cluster analysis only the population from this region were selected (controls).

Figure 15. Population, cases and sources location

The main interest of this analysis is to check if there is any kind of clustering around the sources: Municipal Incinerator, Car Scraping Plant, and Aluminum Recycling Plant.

Two different statistical methods will be applied. Kulldorff's scan statistic and our proposed new method. Kulldorff's test will be applied in two different ways. The first one using the geographical coordinates of the cases and controls (as a local test) and the second one using the geographical coordinates of the sources locations (as a focused test).

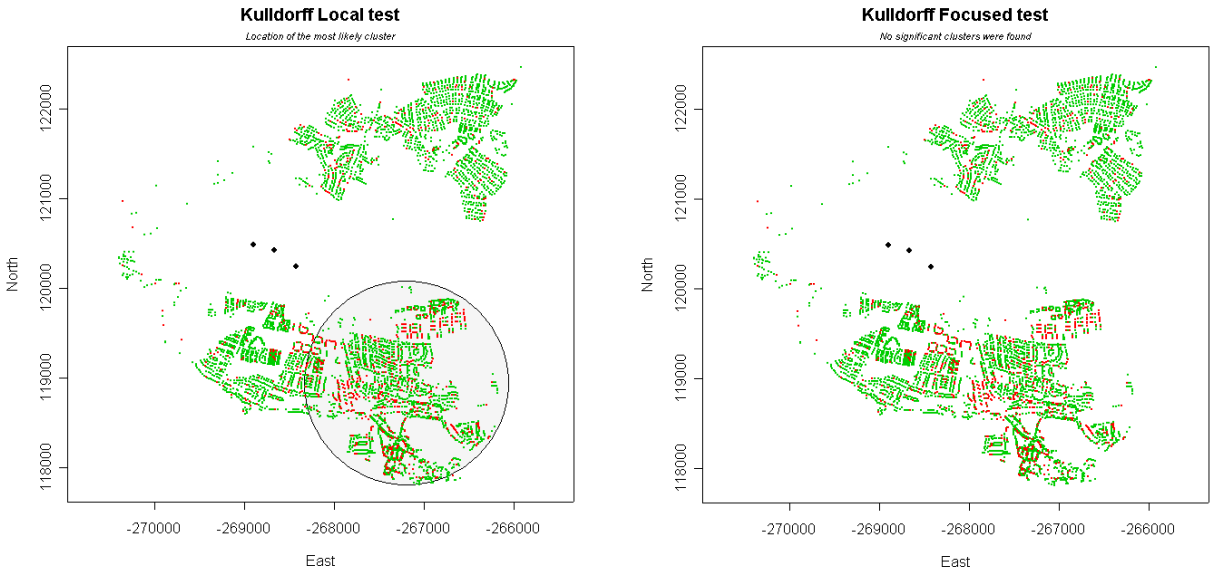
The first analysis will be done using the whole data set of cases. Then, the most frequent types of cancer will be individually analysed.

Only the population and cancer cases with a residency time of 2 or more years will be used for these analyses.

All types of cancers

Using Kulldorff's scan statistic as a local test, one significant cluster was found in the southeast area of the map (Figure 16). However no significant clustering around the sources was found using Kulldorff's scan statistic as focused test. (Test p-value = 0.275).

Figure 16. Kulldorff's scan test output.



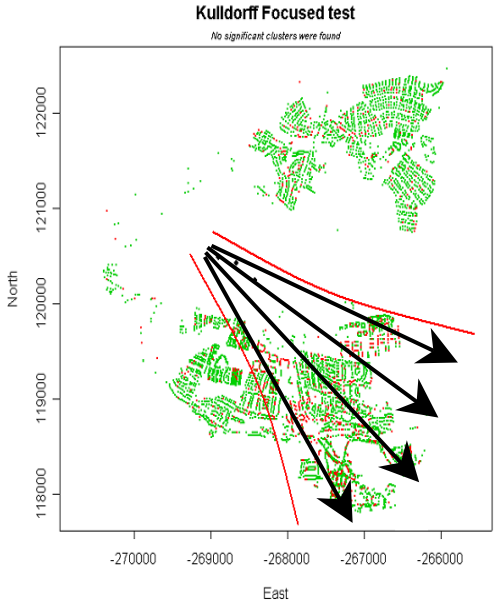
From our focused method, a p-value equal to 0.332 was obtained. Therefore, the null hypothesis of equal probability of being a case around the sources and around any other points of the area could not be rejected.

This may suggest that although there is clustering in the area, this clustering it is not close to the sources. Bear in mind that both methods try to find more cases than the expected ones in circular zones around the sources.

However, looking at how the population is geographically distributed, some hypotheses about the clustering found could be made in relation to the effect of the sources in the risk surface. For example, if the prevalent wind in the region comes from the northwest, then we would expect to find a significant cluster (more cases of cancers then the expected) in the southeast area, as it already happens (

Figure 17).

Figure 17. Hypothesis for the clustering due to the putative sources in the southeast region.



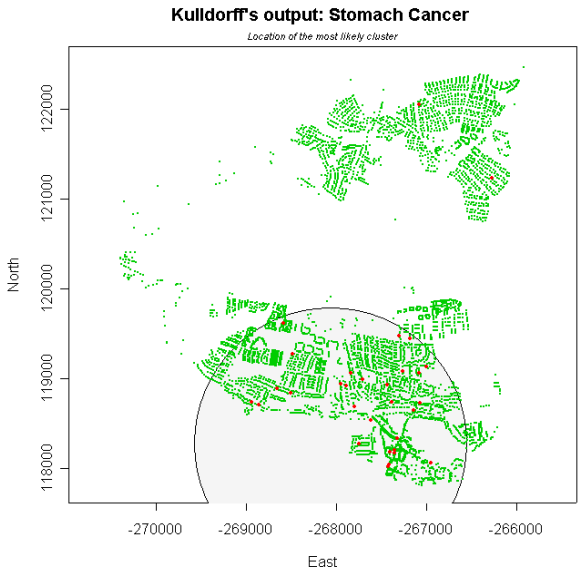
Other reasons for the clustering in the southeast region could be found, however, since we do not have any knowledge about the geographical characteristics of the area, the prevalent wind or what is really happening in that area of Denmark, we cannot assure anything.

Most frequent types of cancer individually

The most frequent types of cancer were individually analysed. These were Lung, Breast, Colon, Bladder, Prostate, Melanoma, Rectum, Brain, Kidney, Cervix and Stomach Cancers.

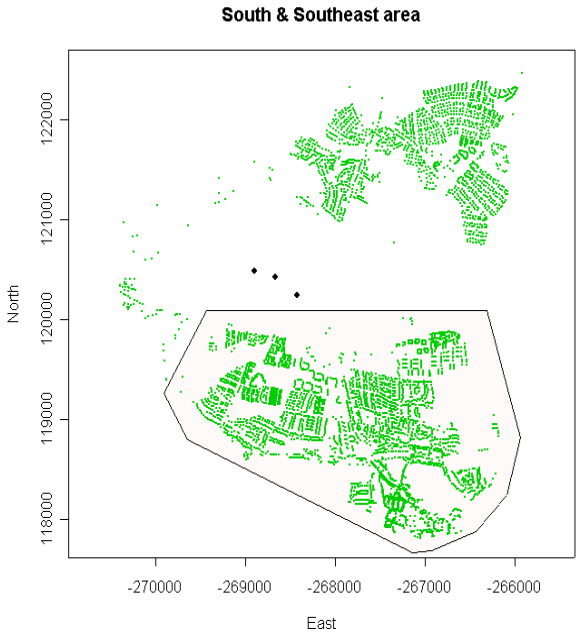
In most of them no significant evidence of clustering was found. Only in the analysis of the Stomach Cancer a significant cluster was found, again in the south area (Figure 18). From the 33 valid cases of stomach cancer, 28 of them were allocated in the south area, while there were only 5 of them in the north one.

Figure 18. Stomach cancer.



However, looking at the south & southeast area (Figure 19) it was found that in most of the types of cancer, there is more probability of developing a cancer in this area than outside it.

Figure 19. South and southeast area.



Conclusions

The main interest of this analysis was to check if there is any kind of clustering around the sources: Municipal Incinerator, Car Scraping Plant, and Aluminum Recycling Plant. An important aspect for this analysis was that all the sources are very close geographically but far away from the location of the population.

No clear evidence of clustering around the sources was found using Kulldorff's scan statistic or our proposed new method. However, *some hypotheses about the clustering* found in the southeast region were made. Still, since we did not have any knowledge about the geographical characteristics of the area, the prevalent wind or what is really happening in that area of Denmark, we cannot assure them.

Italy

Marco Martuzzi

WHO European Centre for Environment & Health
Rome
ITALY

Background

Health impact assessment (HIA) is of growing interest in the field of public health in Europe. Scientific knowledge on the adverse effects of several environmental factors on the population is in some cases substantial, but often-regulatory policies fail to reflect such knowledge adequately. In addition the Amsterdam Treaty of the European Union (Article 152) states that definition and implementation of all new Community policies should ensure a high level of human health protection, which can be achieved through HIA.

A key element to support the formulation of public health policies based on the available evidence is to develop rigorous methods to translate experimental, toxicological and epidemiological information into accurate estimation of the overall impact on the health of the population. So far, most of these exercises have been based on relatively simplistic methods, where the health impact is measured by direct derivation of attributable risks, based on available dose-response estimates and exposure profile of the population (often an average measure), and not using statistical models. Uncertainties in the estimates of the impact are normally evaluated on the basis of the confidence intervals attached to the dose-response function, and other sources of uncertainty that can be of great importance are ignored, or acknowledged only qualitatively. In addition, concerns of “double counting” of events in multiple health endpoints assessments, often suggest limiting studies to a partial set of adverse health consequences. As a result, impact assessments are often said to be “conservative”, and little or no attempt is made to verify that the necessary assumptions are met and/or to carry out quantitative evaluation of the true systematic and random error involved.

In addition to that, the attention towards estimates dealing with the same problems in terms of potential years of life lost has been growing, in order to identify more specific and sensible population groups and develop a more structured approach in public health policies so that targets can be established and goals achieved.

Using data on urban air pollution and health, estimates of the effects of shifting risk factor distributions towards a counterfactual, rather than the difference between “exposed” and “unexposed”, have been calculated. Statistical methods based on (i) a Bayesian methodology (short-term approach) and (ii) years of life lost models (long-term approach) have been explored. In the first approach assessment of health risks related to air pollution levels for the eight major Italian cities has been calculated, disaggregated by sex and class ages; in the second approach health effect delays and different scenarios in reduction of average pollutants concentration have been considered as well. Both approaches lead to the calculation of the *attributable burden* in a given population due to a specific exposure (PM10 ambient concentration in our case), and, consequently, to the portion of disease burden (the *avoidable burden)*

that could be reduced in the exposed population if causative exposure were eliminated.

The report is organized as follows: section 2 illustrates short-term (both conventional and Bayesian approach) and long-term (life-table approach) methodologies, section 3 summarizes the results of the quantitative estimates, section 4 provides the conclusions and section 5 (Appendix 2. Italy, page 105) accurately describes model details.

Methods

Short term approach

To estimate urban population exposure from the existing monitoring data, relevant monitoring stations for the eight major Italian cities have been selected and daily pollutants data from 1998 and 1999 have been considered. Because of the unavailability of dispersion models, average values from the two years empirical distributions have been calculated and used as *proxy* of the current population exposure. When available, values from stations measuring PM10 have been used; otherwise, estimates have been made using TSP data, applying correction coefficients derived from literature or directly calculated from available Italian data. When daily data were not available they have been estimated by using a Poisson distribution imposing the constraints of monthly and weekly averages of PM10 values. To calculate the attributable proportion of health effects from air pollution, a relative risk value has been obtained by pooling the estimates of the available studies through a meta-analytic approach. Conservative dose-response coefficients have been selected to estimate the attributable number of cases, so that the adopted relative risk value is the lower confidence limit of the meta-analysis relative risk.

Conventional models

The model normally applied to calculate the attributable proportion **(A)** of health effects of a given exposure to a risk factor (air pollution in the example discussed here) is the following one:

$$\mathbf{(1)} \qquad \mathbf{A = (RR - 1) / RR}$$

Where **RR** is the Relative Risk estimate obtained from the available literature.

The central estimate of the necessary relative risks and 95% confidence bounds were derived using a meta-analytic approach (Künzly et al., 1999).

To calculate the number of cases due to air pollution (**E**), the following formulation is used:

$$\mathbf{(2)} \qquad \mathbf{E = A * B * C * P}$$

Where

B = population baseline rate of the given health effect;
C = relevant change in exposure;
P = relevant exposed population for health effect.

The population baseline rate is the proportion of the exposed population that would experience the health outcome assuming a baseline level (or no effects level) of air pollution.

This can be calculated as:

(3) **B = Bo / [1 + (RR-1)*C]**

Where

Bo = observed rate (all causes mortality rate for population over 30 years, in our case) of the health effect under current exposure.

Bo and **P** are obtained from available health statistics and from census data.

The Bayesian model

A Bayesian probability model, in which most of the quantities are stochastic variables, was defined and applied to the data to explore the relationships between the distribution of air pollutants (PM10) and health effects.

The model consists of a full joint probability distribution of all the unobserved (i.e. parameters and missing values) and observed quantities (PM10 data). This distribution is conditioned by data and *posterior* distributions for parameters and unobserved quantities are obtained. From each of these *h* conditional posterior distributions, alternatively for each unknown parameter and keeping fixed the remaining ones, one value is generated through a large number (*n,* e.g. 10 000) of simulations. It has been demonstrated that, under general conditions, every set of simulations (chain) can be considered as a sample extracted from the joint distribution of probability. The estimate for the unknown parameter is derived from these values through descriptive statistics (means, medians, percentiles). Some statistical tests are used to prove the convergence of the different chains to the true value.

The algorithm used to estimate the unobserved quantities belongs to the family of Monte Carlo simulation iterative methods and it is a procedure of numerical integration known as Gibbs sampling.

Through this Monte Carlo simulation method it is possible to calculate attributable risks and number of cases with 95% credibility intervals for every single city and for a joint variable “sum” of all the cities. By disaggregating the outcome by sex and age-class attributable number of deaths for every city, every age-class (from 30 to >95 years) and sex have been obtained. Calculations have been made using different baselines (20, 30 and 40 $\mu g/m^3$) for PM10 ambient concentrations.

By using this Bayesian approach a more accurate assessment of the uncertainty in the data is possible: the width of the 95% credibility intervals also depends on the variability of the eight cities empirical distributions of pollutants and not only from the variability of the relative risk obtained from the meta-analysis. In addition to that, a joint variable, with its own variability, can be calculated for each sex.

To get the final values for every single city and for the total of the cities a sufficient number of simulations (10 000 for each chain let us say, without considering in the analysis the first 5 000) is done through the Monte Carlo method described above.

Details of the models, and procedures followed for model fitting are given on page 105.

Long term approach: calculation of years of life lost (YLL)

The same air pollution data has been utilized for the calculation of an estimate of the total number of YLL under different future scenarios by utilizing a classical life-table approach, with no Bayesian implications. Only average annual data of air pollutants and population and mortality data for a starting year (1998, in our case) disaggregated by one-year age-class are needed.
The dose-response function is assumed to be linear and proportional increases in relative risk are not age-dependent.

Applying that coefficient to the 1998 population-at-risk the total number of deaths associated to air pollution that can be avoided by reducing the level of PM10 pollution can be calculated. Total number of YLL, under specific assumptions, can be calculated as well, utilizing the life-table methods described in a recent work (Hurley et al., 2000).

Mortality for each age-class can be described in terms of continuous mathematical survival functions that can show the pattern of survival probability from 100% (at birth) to 0% (at some advanced age, >95 in this case). The concept of the hazard, i.e. the instantaneous probability of dying at a specified age conditioned on not dying before that age, has been utilized. The survival function shows the cumulative effects of the hazards over time, but, unlike this one, is non-increasing with age, while the hazards can increase or decrease over time. Hazard rates as well are continuous functions of age and they can be estimated like the ratio of deaths to mid-year population. If the periods under exam are reasonably short (e.g. one-year periods) half the deaths over the whole period of time can be assigned to the first half of period, so that on entry population for each period can be calculated.

Relationships between hazard and survival function and between number of deaths d, initial or entry population e, mid-period population m and hazard h in the period can be summarized in simple mathematical relationships as follows:

$$h = d/m$$
$$m = e - d/2$$
$$e = m + d/2$$

The probability of survival **s** for each period, conditioned on entering that period is

$$s = (2 - h) / (2 + h)$$

while the inverse relationship with hazard h is

$$h = 2*(1 - s) / (1 + s)$$

Cumulating these conditional probabilities it is possible to predict the mortality patterns of the initial population under study.

The Italian mid-year populations and the number of deaths, by sex and one-year age-class, in 1998 have been summarized in tables like Table 20. For the eight major

Italian cities the same data elaboration have been made, for the whole population and for people older than 30 years.

Table 20. Mid-year population and number of deaths by sex and age-class (example).

Age	Mid-year population (m), i=1,2, ,95+			Deaths (d), I=1,2, ,95+		
(years)	Men	Women	Total	Men	Women	Total
age 0	m_0	m_0	m_0	d_0	d_0	d_0
age 1	m_1	m_1	m_1	d_1	d_1	d_1
------	---	---	---	---	---	…..
age 95+	m_{95+}	m_{95+}	m_{95+}	d_{95+}	d_{95+}	d_{95+}
Total	Sum($m_{i,\ men}$)	Sum($m_{i,women}$)	Sum($m_{i,total}$)	Sum($d_{i,\ men}$)	Sum($d_{i,women}$)	Sum($d_{i,total}$)

In order to predict future mortality for the population that has been followed up and to describe future scenarios under different hypothesis of reduction of PM10 pollution some strong assumptions have been made. The first assumption is that the hazards have been kept constant over time; in addition to that, the number of births in years after 1998 has been maintained at 1998 levels, no assumption about different hazards for migrant population has been made and the net migration has been considered equal to zero.

Under these strong hypothesis (even though different set of hazards can be considered for every different year) an initial scheme can be filled (Table 21) in which, for every year and for every age-class a vector of births, on-entry populations (the population for 1999 will be the first one depleted by 1998 mortality with survivors one year older) and hazards have been inserted. Along the diagonal of the hazards, which starts where age zero row meets year 1998 column, the 1998 population mortality experience can be followed. For each year of the follow up the hazards for every age-class are the same as in 1998 population but, with this approach, once available or estimated, they can easily be replaced with other different mortality patterns.

Table 21. Organization of data for life-table method.

Age	***Entry pop***	***1998 Births***	***1999*** $b_1=e_0$	***2000*** $b_2=e_0$	***---*** ---	***j*** $b_i=e_0$	***---*** ---	***2091*** $b_{93}=e_0$	***2092*** $b_{94}=e_0$	***2093*** $b_{95+}=e_0$
0	e_0	e_0	h_0	h_0	---	h_0	---	h_0	h_0	h_0
1	e_1	e_0	h_1	h_1	---	h_1	---	h_1	h_1	h_1
2	e_2	e_0	h_2	h_0	---	h_2	---	h_2	h_2	h_2
--	---	---	---	---	---	---	---	---	---	---
i	e_i	e_0	h_i	h_i	---	h_i	---	h_i	h_i	h_i
--	---	---	---	---	---	---		---	---	---
93	e_0	e_0	h_{93}	h_{93}	---	h_{93}	---	h_{93}	h_{93}	h_{93}
94	e_0	e_0	h_{94}	h_{94}	---	h_{94}	---	h_{94}	h_{94}	h_{94}
95+	e_0	e_0	h_{95+}	h_{95+}	---	h_{95+}	---	h_{95+}	h_{95+}	h_{95+}

Once Table 21 has been filled, deaths or number of person years (py) can be calculated and inserted in Table 22. The latter are calculated summing up the survivors at the end of each period (one year lived each) and part of the dead (estimated as half a year per death). So, in each cell of Table 22 the number of person years lived will be calculated by inserting the number of persons entering the period minus half the number of death in the same period (e – d/2). The total of person years lived by 1998 generation has been calculated with the sum of the values of all cells (every cell value is equal to the same value in person years) under the 1998 diagonal and on the diagonal (the shaded portion of Table 22). Dividing this sum with the size of the initial population the average years of life expected at entry can be estimated while the contribute of the born after 1998 is described by the upper part of the same table.

Table 22. Schematic layout for calculation of predicted person-years (py) gained with PM10 reductions.

Age	***Year***								
	1998	***1999***	***2000***	***---***	***j***	***---***	***2091***	***2092***	***2093***
0	py	py	py	---	py	---	py	py	py
1	py	py	py	---	py	---	py	py	py
2	py	py	py	---	py	---	py	py	py
--	---	---	---	---	---	---	---	---	---
i	py	py	py	---	py	---	py	py	py
--	---	---	---	---	---		---	---	---
93	py	py	py	---	py	---	py	py	py
94	py	py	py	---	py	---	py	py	py
95+	py	py	py	---	py	---	py	py	py

A second set of these tables can be easily calculated replacing the hazard for all causes mortality with the hazard that has been observed applying our meta-analytic relative risk coefficient to the original instantaneous mortality rates. To different reductions of air pollution levels other hazard values will correspond and a different mortality profile will be observed, so that various mortality scenarios can be considered. In other words, utilizing this way of organizing data, under different scenarios of reduction of PM10 pollution, the number of years of life that could be gained can be easily known. This kind of calculations can be made also for people born after 1998 and followed up along their life, by updating continuously the vector of hazards and populations, once forecasted values improve or once historical mortality patterns become available.

Different scenarios of possible reductions in ambient PM10 particulates concentration have been simulated, starting from a reduction of 5 μg/m3 up to 40 μg/m3 (5, 10, 15, 20, 25, 30, 35 and 40 μg/m3) even though the latter hypothesis (from 25 μg/m3 up to 40 μg/m3) are quite unrealistic.

The assumption under which these results have been calculated is that, after a reduction in PM10 concentrations, the altered hazard stays constant over time and that this reduction causes an immediate effect on the hazards, starting from the structure of population and mortality for the year 1999 (let us say, the full effect of reduction is immediate, there is not any kind of delay).

Alternative hypothesis can be made, if delayed effects are taken in account. In this analysis effects for delays of 5, 10, 20 and 30 years have been considered. In the first case, for instance, the 1998-generation experiments the reduction in PM10 concentration only after the fifth year, i.e. the hazards stay constant until then, and only after the fifth year decrease. In this way people from zero to five years old in 1998 are not followed until 2003 (and then followed up for the rest of their life, conditioned on surviving) while people older than 90 years are not followed at all.

In the following chapter some examples of these results, disaggregated by sex, reductions in PM10 concentration and delays, are shown in Table 26 and Table 27.

Calculations for the eight major Italian cities have been made in two different ways. The first one is the way showed until now. The second one has been added in order to compare the new results with the ones obtained in a previous WHO report (Martuzzi et al., 2002) which studied the gain in lives that the populations of the eight towns could gain once the PM10 levels have been brought to thresholds of 20, 30 or 40 μg/m3). This second kind of calculations has been made once that the mortality data have been updated with historical rates in place of the estimated ones, utilized in the WHO study, and after that different methodologies (geometrical averages in place of arithmetical ones for the two years distribution of pollutants) have been utilized. At last, only populations over 30 years old have been considered.

These health gap results have been disaggregated by sex, by pollutants concentration abatement to the three thresholds and by different delays to full effect in the tables of the results section.

Results

Bayesian model

Several models were used to explore some of the available options, as follows.

Model 1: Pr 24 results updated with population historical data - variability of Relative Risk; use of arithmetical mean;

Model 2: WinBugs, first version - same variability of Model 1; variability of pollutants distribution not included; use of arithmetical mean;

Model 3: WinBugs, second version - same variability of Models 1 and 2 plus variability of the pollutants distribution; use of arithmetical mean;

Model 4: WinBugs, third version - same variability of Model 3; use of geometrical mean; first 1 000 iterations thrown away.

Model 5 (a, b and c): WinBugs, forth version - updating of model 4 with historical mortality rates (not estimated ones); same variability of model 3 but data disaggregated by sex and age classes; two chains of simulate values calculated; 10 000 iterations for each chain, first 5 000 not considered in the analysis; initial values inserted and not generated by the model itself; Gelman-Rubin convergence test applied.

The results obtained using the above-mentioned different approaches have been shown in where the number of attributable cases estimated by the different models, using three different counterfactual exposure levels (i.e. alternative exposure distribution used as baseline for estimating the burden of disease caused by the exposure distribution of interest) for PM10 concentrations: 20, 30 and 40 μg/m3 has been reported. It can be seen that central estimates are not substantially affected by the choice of the model also when more in-depth analyses have been introduced. The substantial differences on the average effects depend on more updated mortality and population data used to calculate the estimates and on the more correct use (on the methodological side) of the geometrical mean as a synthetic indicator. For example, using 20 μg/m3 as a counterfactual exposure (first part of Table 23), the last developed model (5c) estimates that around 6000 extra deaths are attributable to PM10 in the eight Italian cities each year. While this kind of information is valuable to describe the health benefits associated with abatement measures (and can therefore help formulate evidence-based policies), it is also crucial to be able to evaluate the degree of uncertainty that surrounds the estimates. The tables show the 95% credibility intervals attached to the estimates. Unlike the central estimates themselves, credibility limits are heavily affected by the choice of the model: in particular, models that allow for more sources of variability produce wider credibility intervals (from 2616 in model 3 to 8517 in model 5c).

Disaggregating by sex, age-class (single classes or grouped in two classes, working people (between 30 and 65 years old) and not working one (more than 65 years old)) allow us to compare results between the two groups and among the eight cities and can help the administrators to identify priorities in health impact assessment decision making process.

Table 23. Comparison of results among different Bayesian models.

Baseline 20 - sum of the 8 cities	mean	CI 95% (LL)	CI 95% (UL)	2.5 perc	median	97.5 perc	CI width	N° of Simul.	Notes	Type of mean used
1 – No Bugs: var of RR (correct), population > 30	5143	3890	6396	-	-	-	2505	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	5105	-	-	3768	5120	6384	2616	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	4812	-	-	1110	4787	8642	7530	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	5018	-	-	1909	4752	9708	7799	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	3025	-	-	1227	2894	5649	4422	10000**	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	2979	-	-	1084	2854	5426	4342	10000**	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	6003	-	-	2463	5746	10980	8517	10000**	Variability of RR and C	Geometrical

Table 23 continued. Comparison of results among different Bayesian models.

Baseline 30 - sum of the 8 cities	mean	CI 95% (LL)	CI 95% (UL)	2.5 perc	median	97.5 perc	CI width	N° of Simul.	Notes	Type of mean used
1 – No Bugs: var of RR (correct), population > 30	3496	2598	4394	-	-	-	1797	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	3479	-	-	2540	3488	4391	1851	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	3158	-	-	-745	3151	7026	7771	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	3376	-	-	2046	3107	8198	6152	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	2040	-	-	190	1905	4620	4430	10000*	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	2005	-	-	198	1880	4512	4314	10000*	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	4045	-	-	386	3781	9155	8769	10000**	Variability of RR and C	Geometrical
Baseline 40 - sum of the 8 cities	mean	CI 95% (LL)	CI 95% (UL)	2.5 perc	median	97.5 perc	CI width	N° of Simul.	Notes	Type of mean used
1 – No Bugs: var of RR (correct), population > 30	1765	1269	2262	-	-	-	992	0	Variability of RR	Arithmetical
2 - Bugs: var of RR, population > 30	1764	-	-	1248	1767	2257	1009	10000	Variability of RR	Arithmetical
3 - Bugs: var of RR and C, population > 30	2051	-	-	-2779	1453	5340	8119	10000	Variability of RR and C	Arithmetical
4 - Bugs: var of RR and C, population > 30	1642	-	-	-1800	1387	6552	8332	10000*	Variability of RR and C	Geometrical
5a - Bugs: var of RR and C, men, one year age-classes	1000	-		-963	878	3671	4634	10000**	Variability of RR and C	Geometrical
5b - Bugs: var of RR and C, women, one year age-classes	976	-	-	-935	866	3566	4501	10000**	Variability of RR and C	Geometrical
5c - Bugs: var of RR and C, total, one year age-classes	1977	-	-	-1896	1746	7245	9141	10000**	Variability of RR and C	Geometrical

**In the forth model developed with WinBugs, 10 000 simulations have been made but the first 1 000 have not been used in the analysis.*

***In the fifth model, 10 000 simulations have been made for both chains but the first 5 000 have not been used in the analysis.*

Years of life lost

All the simulations have been based on the Pope meta-analysis coefficients in which, taking the lower confidence limit as our central estimate for the relative risk, for a reduction of 10 μg/m3 in PM10 concentration a relative risk RR of 1.026 has been obtained. This corresponds to a reduction factor in hazard equal to 1/1.026 = 0.9747. All the reduction factors in hazards for the studied range of potential pollution reductions, with 95% "uncertainty" intervals (it is incorrect defining them confidence intervals, a lot of variability different from the relative risk one should be included in their computation) have been reported in Table 24.

Table 24. Hazard impact factors for a range of potential pollution reductions.

Reduction in PM10 pollution concentration, μg/m^3	Assumed reduction factor in hazards		
	Central estimate	95 % uncertainty interval	
		Lower limit	Upper limit
5	0.9872	0.9792	0.9955
10	0.9747	0.9588	0.9911
15	0.9622	0.9388	0.9867
20	0.9500	0.9193	0.9822
25	0.9378	0.9001	0.9779
30	0.9259	0.8814	0.9735
35	0.9141	0.8630	0.9691
40	0.9024	0.8450	0.9648

The values in Table 24 correspond to an immediate effect on the hazards of the air pollution reductions, without considering any delayed effect. By applying all these coefficients to the 1998 Turin population (e.g.), average gain in life (years) can be obtained. These information, starting from values of life expectance at birth of 75.82 for men and 82.01 for women (1998 data collected from ISTAT mortality tables), have been collected in Table 25 which shows the same gain calculated in terms of life-years.

Table 25. Prediction of expectation of life and average gain in expectation, under various reductions in hazards, Turin, no delay, all the ages, 1998.

Reduction in PM10 concentration, μg/m^3	RR	Assumed reduction factor in hazards	Men		Women	
			Expected life (years)	Life gained (years)	Expected life (years)	Life gained (years)
0	-	None	75.82	-	82.01	-
5	1.013	0.9872	75.93	0.1096	82.10	0.0902
10	1.026	0.9747	76.04	0.2216	82.19	0.1819
15	1.039	0.9622	76.15	0.3337	82.28	0.2734
20	1.053	0.9500	76.27	0.4457	82.37	0.3645
25	1.066	0.9378	76.38	0.5577	82.47	0.4553
30	1.080	0.9259	76.49	0.6696	82.56	0.5459
35	1.094	0.9141	76.60	0.7816	82.65	0.6361
40	1.108	0.9024	76.71	0.8934	82.74	0.7260

Table 25 shows, e.g., that a reduction in PM10 concentration levels of 10 μg/m3, which causes a reduction in all the age hazards of 0.9747, would cause a gain in expected life for men equal to 0.2216 years (from 75.82 to 76.04 years), i.e. about two and a half months. The same calculations made for women would increase the expected life of 0.1819 years, i.e. about two months. For all the reduction factors the effect is always higher for men.

Similar calculations have been made in Table 26 that shows the same gain calculated in terms of life-years.

Table 26. Predicted total gain in life-years (thousands) under various assumed reductions in PM10 pollution, Turin, all the ages, 1998.

Reduction in PM10 concentration, μg/m^3	Years of life gained (thousands), no delay		
	Men	Women	Total
5	47.8	42.7	90.5
10	96.7	86.1	182.8
15	145.6	129.4	275.0
20	194.5	172.5	367.0
25	243.4	215.5	458.9
30	292.3	258.4	550.6
35	341.1	301.1	642.2
40	389.9	343.6	733.5

It can be easily noticed that at higher reductions of PM 10 pollution correspond a higher number of gained person-years for both sexes, here reported without uncertainty intervals for space reasons. In
Table 27 the same kind calculations have been showed, introducing the delaying factors in the analysis.

Table 27. Predicted total gain in life-years (thousands) under various assumed reductions in PM10 pollution by delay to full effect, total, Turin, all the ages, 1998.

Turin – Total	Delay				
Reduction	0	5	10	20	30
5	90.51	83.48	75.86	59.67	44.26
10	182.82	167.81	152.05	119.40	88.55
15	275.00	252.02	228.14	179.08	132.81
20	367.02	336.11	304.12	238.69	177.02
25	458.90	420.06	380.00	298.22	221.21
30	550.61	503.88	455.76	357.69	265.33
35	642.17	587.56	531.40	417.06	309.42
40	733.55	671.08	606.91	476.35	353.45

As a result the gain in person-years, at the end of the calculations, will be obviously less than the one with no delay, because the exposed population is greater. For the same reason a declining trend in gain of person-years will be observed for increased delays. By applying a 30-year delay the gain that can be obtained almost halves the no delay benefit for every kind of PM10 reduction.

To compare the results obtained for the eight towns in the Bayesian section, the same outputs for population older than 30 years have been produced, for future scenarios of pollutants concentration abatement to 20, 30 and 40 μg/m3. The results have been reported in Table 28.

Table 28. Turin, population older than 30 years, gain in life-years, reduction to baseline 20, 30 and 40, males and females, by delay to full effect.

Turin - Total	Delay				
Reduction to	0	5	10	20	30
20	405.62	357.91	308.55	211.01	125.34
30	285.45	252.05	217.38	148.67	88.30
40	164.93	145.94	126.02	86.23	51.22

Discussion

This kind of comparative risk assessment methodology used to calculate health gaps carried out during EUROHEIS project indicates that it is feasible to apply methods that allow estimating attributable risks with appropriate treatment of several sources of systematic and random error. The methods have been explored using real data from a participating country. The methodology has been identified and tested on these data. Results suggest that it is possible to develop a tool that can be used in conjunction with data on environment and health in order to:

- derive more accurate estimates of health impact;
- further assist health professionals involved with health impact assessment;
- facilitate the evaluation of the concrete public health implications;
- allow effective communication with the general public and decision makers alike;
- enhance the scientific basis for decision making;
- assist in the development of appropriate policies.

The methods presented here should be further developed through:

- Bayesian models which take into account further sources of variability (short term approach) and not linear relationship in the dose-response function for higher concentrations of pollutants;
- applications and calculations of disability-adjusted life years (DALYs), to combine into a single unit mortality and morbidity data.

The Netherlands

J.Kwekkeboom
K.Huijsmans
C.Van Wiechen
B.Staatsen

National Institute for Public Health and the Environment Centre for Environmental Health Research
A. Van Leeuwenhoeklaan 9
P.O. Box 1
3720 BA Bilthoven, The Netherlands

Background

In the Netherlands, several environmental health studies have been carried out using geographical information systems (GIS), particularly in the field of air quality, noise and living quality. These studies were carried out as separate activities. For each study, environmental and health databases from various data-holders were linked, Until now, this linkage did not happen on a routinely basis. Consequently it is difficult to carry out rapid inquiries into environmental health problems, if needed.

The EUROHEIS project effort has focused on the development of an integrated information system for assessment of (spatial) relations between environment and health. Recently, the Netherlands has joined the EUROHEIS project group. Within the National Institute of Public Health and The Environment, a project team was established as well as a multidisciplinary support group. The Rapid Inquiry Facility (RIF) was introduced to the Ministry of Housing, Spatial Planning and the Environment and to the Dutch Health Council. All concerned agreed that the RIF and its associated methodology could advance the ongoing work in the field of environmental health

Aim

Our study consists of the following three phases:

1. Implementation of RIF:
    - Software installation;
    - Incorporation of health, population and geographical data;
    - Explore the potential of adding environmental data (noise exposure data) to the system.

2. Perform analyses:
    - Disease mapping. The first step is to compare disease maps created by RIF with disease maps created in earlier geographical studies to investigate whether RIF-disease mapping works properly for the Dutch situation.

3. Explore (methodological) development of RIF:
    - Include life-style data (smoking, obesity, etc.);
    - Statistical development;
    - Develop an organisational framework for applying RIF.

Since it was the first year that we joined EUROHEIS, we focused on the first phase, i.e. implementation of the system in The Netherlands

Implementation

Implementation of Software

The UK version of the RIF was installed on a standalone PC, using Windows XP Pro, Oracle 9i personal and Arcview 3.2a packages. Several technicalities had to be worked out before the RIF software could be installed successfully.

Incorporation of data

For testing purposes, a relatively small subset of data has been incorporated into the system The geographical extent is limited to an area of 55x55 km. Health and population data for a limited time period as well as a limited area have been loaded into the system. Inclusion of all available data-sets is foreseen for the very near future.

Detailed digitised geographic information and associated data is readily available for the Netherlands and generally of good quality. Age and sex specific population data is currently available at postcode (4) level for the whole country for the time period 1995-2002. The age structure does not exactly meet the requirements of the RIF. Population is currently not available for age 0 and age 1 to 4. at postcode 4 level. Due to privacy regulations the population numbers are rounded to 5, which could introduce a larger margin of error!

Currently available health data mainly consists of age and sex specific hospital admissions covering the period 1980-2000. This data is not very reliable for the period before 1990 and only available at postcode 4 level from 1991 onwards. Age and sex specific mortality data at postcode 4 level is currently available for the years 1995 to 1997.

A deprivation index for the whole country is only available for the year 1995. There is no data at postcode 4 level available on life-style factors. The integration of available data into the RIF software has proven to be a time consuming task but it is expected that the datasets mentioned above will be loaded into the RIF/Oracle database by June 2003

Exploring environmental data

Exposure data of aircraft noise modelled by the Dutch National Aerospace Laboratory are available and of good quality. These data concern modelled annual average exposure levels in the vicinity of Amsterdam Airport Schiphol. Recently these exposure levels have been combined with x,y co-ordinates of all residential addresses for the period1998–2001 using GIS. As a result a database of average noise levels within postcodes (4 position) has been established. Results showed that

most of the variance of noise was due to contrasts between postcodes and not due to contrasts within postcodes or between years (Van Wiechen, 2003). These results make us confident in the usefulness of using the aggregated aircraft noise exposure data in a RIF setting.

Reference

Van Wiechen et al. (2003) Monitoring environmental health around the international airport Schiphol. Ostersund, Sweden, 30-31.3.2003 (conference abstract).

Spain

Juan Ferrándiz
Virgilio Gomez
Pasqual Milvaques

Department of Statistics, University of Valencia, Spain

Ricardo Ocaña
Carmen Sanchez-Cantalejo
Antonio Daponte

Andalusian School of Public Health (EASP), Granada, Spain

RIF installation

After completion of the Spanish version of the RIF, it has been implemented in the Department of Epidemiology of the Valencian Regional Health Authority (DGSP hereafter as the acronym for *Dirección General de Salud Pública*) under two settings:

i. ***Standalone PC,*** requiring the inclusion of ORACLE and ARCVIEW packages jointly with the RIF in a isolated PC running Windows 2000 OS, and

ii. ***RIF satellites*** of a central database, where many PCs are connected to the central ORACLE database held in a UNIX machine of the DGSP computer center. Data is generated and updated in this central database while individual RIF packages are installed in different PCs according to staff specific tasks within the Department of Epidemiology

Code debugging

The RIF is now being tested in routine surveillance tasks in the DGSP. It is being subject to a massive trial use with these routine problems. Minor code inconsistencies are being detected and considered for debugging. We are about to include as well some improvements involving small changes in the written code, which will lead to a more friendly handling of the tool and more useful presentations.

Installation/updating wizard

We are considering to make a widespread offer of the RIF to all epidemiological units within the DGSP. In order to facilitate installation tasks in different PCs running windows 2000 we are developing a wizard module based on typical needs detected in the DGSP past experience. This wizard module looks as well at future systematic updates of the built-in database.

Statistical extensions

From the regular use of the RIF in the DGSP for surveillance problems we have detected the need for some statistical procedures whose inclusion will increase the usefulness of this tool.

Analysis with covariates

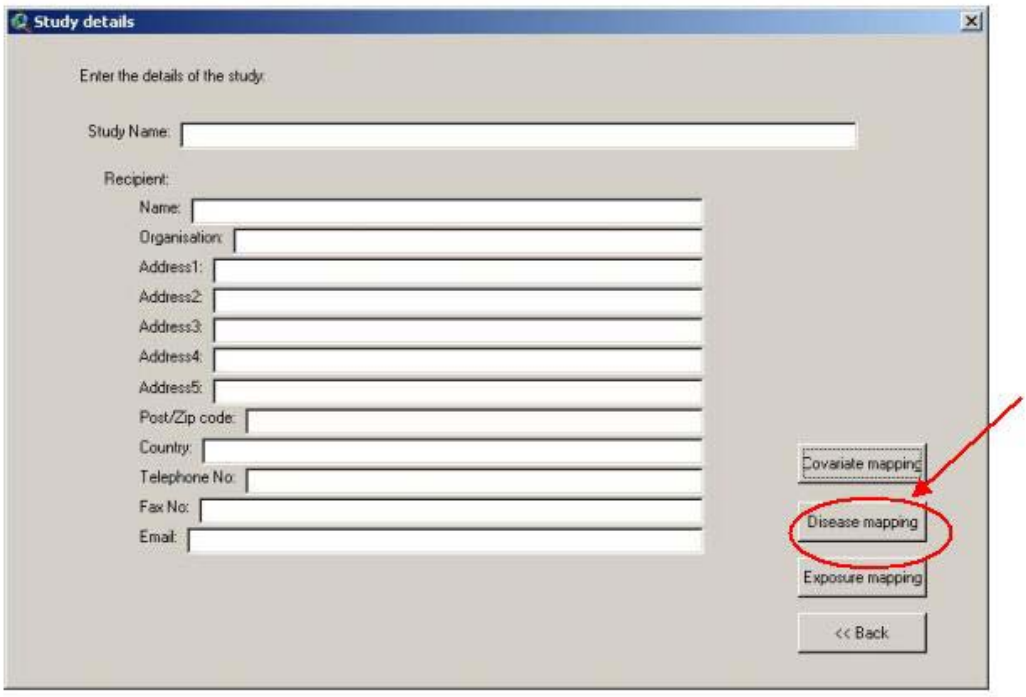
Most of the Spanish envisaged case studies have to do with exposure to different contaminants that, although distributed with a clear geographical structure, are not reducible to the *point-source analysis* format. Instead of measuring exposure by distance to contamination origins we consider the level of these contaminants in the

geographical units as a measure of exposition, much in the usual conception of ecological regression problems.

Thus we have added a new option within the RIF menu, offering this type of analysis besides those previously stated. Nevertheless we have kept mathematical complexity at the same level it was originally by performing this new analysis with the same computational techniques as those used with deprivation indexes
The main steps of this new analysis are:

1. To compare maps of disease and covariates to ascertain possible relationships
2. To compare mortality/morbidity rates of regions defined by grouping geographical units with similar levels of the selected covariates. This allows verification of the possible effects of these covariates.
3. To perform standardization of rates taking levels of the covariate as an added criterion to stratify the population (as we do with a deprivation index). In this way we can see how filtering the effects of covariates changes the resulting rates or maps.

Figure 20. Sample screen showing RIF module added by Spanish partner.



Detailed description of instructions to perform this kind of analysis can be found in the attached technical report entitled '*Implementation of covariate studies in an Epidemiological G.I.S'* by J. Ferrándiz and V. Gómez.
(http://matheron.estadi.uv.es/investigar/tr15-02.ps).

This additional option to the basic RIF has been incorporated into the version installed in the DGSP. It has been used in the case-studies to be presented in the Östersund EUROHEIS confernce.

Cluster detection

Disease cluster detection is another important issue in regular epidemiological surveillance. In the aforementioned use of the RIF in the DGSP we have experimented with some methods of automatic detection as a complementary tool of disease mapping.

We are preparing a statistical module written in the R statistical programming language (http://cran.r-project.org/) called 'Dcluster'. Main standard clustering measures as well as focused cluster tests are included. A more detailed description can be found in the enclosed technical report 'DCluster: Detecting disease clusters with R' by V. Gómez-Rubio, J. Ferrándiz and A. López. In addition to its regular use as a extension of the R environment, we are working out the necessary scripts for it to be called online from any RIF session. It has been presented in the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) held in Vienna on March 20-22, 2003 and constitutes the object of our contribution to the Östersund Euroheis meeting entitled *'R.I.F. output analysis for the detection of disease clusters'* by Ferrándiz J., López A. and Gómez-Rubio V. We intend

Missing data imputation

Some datasets, in particular those related to environmental data, present missing entries making their use in future studies difficult. The statistical models accounting for these nonexistent data are too complex to be incorporated in the RIF. First, because these models require computer-intensive statistical techniques and expert tuning of the computing process that are beyond the current capabilities of a 'rapid' tool like the RIF. Second because these complex models have to be tailored to the particular purpose of each study.

The traditional solution to this problem is the imputation of missing data. That means the estimation of lacking observations by fitting a good model to the observed data. Future studies will take estimated data as if they had been observed and will apply standard techniques of analysis.

The problem remains of how the ignored uncertainty about the true values of missing data could affect the conclusions of the analyses. This is a topic still under current research and we have applied modern procedures to perform this task. In particular we have fitted spatiotemporal models to our datasets to consider all the relevant information provided by the temporal sequences of measures in each region.

We have compared different models and approaches and we have used the best model in each particular dataset. A first report in Spanish is now in press in *Boletín de la SEIO* (Spanish Statistical Society Bulletin) and constitutes the object of our contribution to the Östersund meeting entitled '*Spatio Temporal Imputation in Environmental Data sets'* by Abellán C., Ferrándiz J. and López A.

The imputed values have been incorporated into the RIF in order to complete the registered series of data. Then, the complete data have been used in the case-studies which are to be presented in the Östersund EUROHEIS conference.

Deprivation index

In addition to the study made by the Irish partner, we have developed an autochthonous deprivation index based on a bibliographic revision on Spanish

previous experiences and its adequacy to the envisaged case-studies. It has been incorporated in the Spanish version of the RIF.

Case studies

Several case studies have been used to test RIF adequacy in routine epidemiological surveillance tasks. They correspond to real current concerns of the DGSP in Comunidad Valenciana. We have performed a rapid exploratory study of each problem using the RIF as well as deeper analyses with advanced statistical techniques based on hierarchical models and Bayesian methods.

These advanced methods are too demanding in computing time and statistical expertise to be incorporated in a 'rapid' tool like the RIF. The purpose of the duplication is to evaluate the RIF power in reaching useful conclusions when used as a standard tool.

Cardiovascular and cerebrovascular mortality in Comunidad Valenciana and quality of drinking water (calcium and magnesium contents)

This case-study focuses on the protective effect of calcium and magnesium contents of drinking water from public supplies against mortality from cardiovascular diseases. Disease mapping and covariate analysis have been performed using the RIF and, from its output, the protective hypothesis seems plausible.

To get more insight into the problem and take advantage of the historical data at our disposal we have performed the statistical analysis via hierarchical Bayesian spatio-temporal models as well. A preliminary study, that of *cerebrovascular mortality*, is now in press in *Environmetrics*, (http://www.nrcse.washington.edu/ties/).

A map of the distribution of cerebrovascular disease mortality in men and women, in the region of Valencia, is shown in Figure 21, and areas with low and high risks are shown in Figure 22.

We are now extending the study to the remaining cardiovascular diseases and working out the comparison of results obtained from RIF and from the more sophisticated analyses.

Figure 21. Smoothed Standardised Mortality Ratios. Cerebrovascular disease (ICD9 430--438). Total mortality (Males+Females) Comunidad Valenciana (Spain)

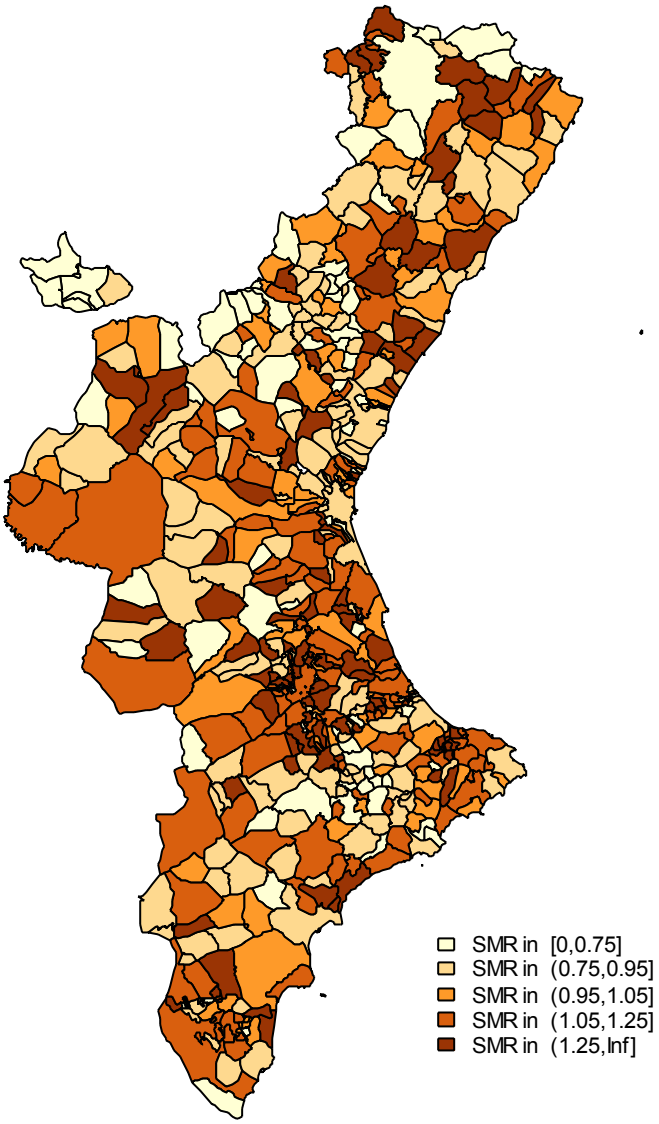
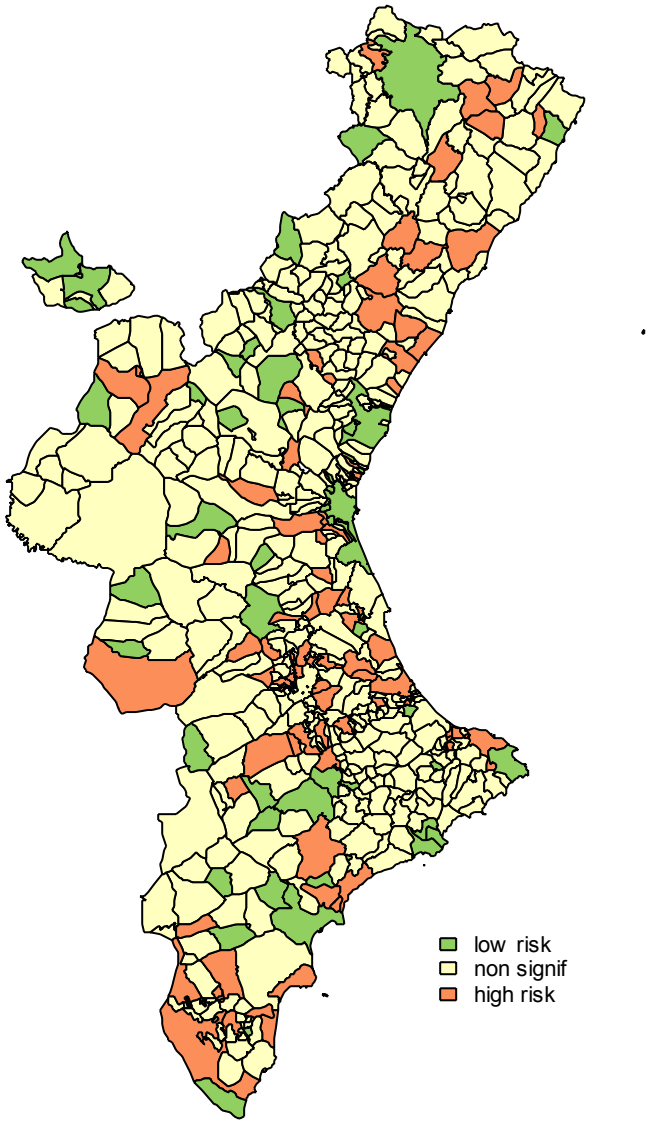


Figure 22. Significant 95% Confidence Intervals for Smoothed Standardised Mortality Ratios Cerebrovascular disease (ICD9 430--438) Total mortality (Males+Females) Comunidad Valenciana (Spain).



Cancer mortality (prostate, bladder, colon) in Comunidad Valenciana and nitrate concentration in drinking water

Nitrate concentration in drinking water from public supplies are high in Valencian municipalities due to intense agricultural activities and acute exploitation of underground aquifers. Measuring its effect on mortality/morbidity from cancer diseases located in the digestive system is an important question of public health concern.

The main problem with the standard use of the RIF is the existence of abundant missing data in the available historical series. This component of drinking water is regularly analysed in a municipal scale since 1991, and there still remain some inconsistencies and lack of registers.
The first task has been dataset completion by imputation of missing data as described in paragraph 3. Then we have performed disease mapping and covariate

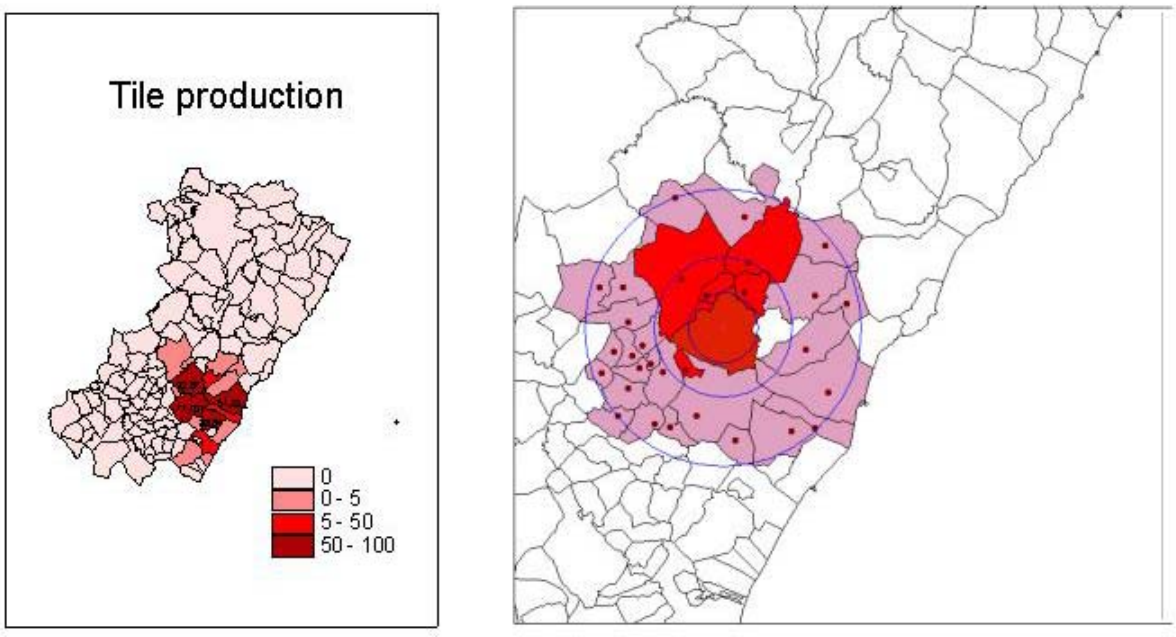
analysis with the RIF. The case of stomach cancer is shown in our contribution to the Östersund Euroheis meeting entitled '*Spatio Temporal Imputation in Environmental Data sets*' by Abellán C., Ferrándiz J. and López A. We are now working out more sophisticated analyses via Bayesian methods on spatial hierarchical models in order to compare their results with those provided by the RIF

Health impacts of air pollution by tile industry in Castellón

Tile industry has experienced an important growth in a region around Alcora in the last two decades. It comprises 18 municipalities within Castellón province. Air pollution has been a consequence and it has induced citizen demands of epidemiological surveys.

Figure 23 shows the tile production and corresponding exposed areas as defined for anlaysis. The DGSP has undertaken several lines of investigation focusing on respiratory and cardiovascular mortality and morbidity to ascertain if there is an increased risk in this region and if it is related to air pollution. RIF covariate analysis as well as the use of hierarchical spatiotemporal models constitute one of the first studies performed as the main part of a first internal report addressed to the Regional Health Authority. A summary description of the results reached are shown in our contribution to the Östersund EUROHEIS meeting entitled '*Health impact assessment of the ceramic industry in Castellón*' by Ferrándiz J., López A., Martínez-Beneito M.A., Vanaclocha H. and Zurriaga O.

Figure 23. Tile industry production and definition of exposure areas.



Dissemination

Spanish user's manual

Spanish RIF user's manual is a necessary instrument for encouraging the use of this tool among epidemiologists within the DGSP and other similar institutions of the remaining autonomous regions in Spain. We need a friendly manual explaining the purpose of every kind of analysis, the way it is performed within the RIF environment and how to interpret the resulting tables and maps. Clear writing and appropriate images are crucial. The current version of this manual integrates disease mapping as well as covariate analysis. The style is practical and self-contained.

Spanish partner meeting

The Spanish EUROHEIS partner involves many institutions in two autonomous regions: Andalucia and Valencia (see http://matheron.uv.es/euroheis). We hold regular meetings in order to coordinate our activity. Last meeting was held in in Valencia on January 2003 and we invited statisticians and epidemiologists from the DGSPs of Sevilla and Madrid, as well as from the University of Barcelona, interested in the development of Health Geographic Information Systems. It was planned as a first contact to extend the reach of the Spanish group bringing up new opportunities of collaboration.

Sweden

Christina Reuterwall

Research and Development Unit
Jämtland County Council
P.O. Box 602
S-832 23 Frösön
SWEDEN

Niklas Hammar

Department of Epidemiology
Institute of Environmental Medicine
at Karolinska Institutet
Stockholm
SWEDEN

Introduction

The UK RIF system was adapted for use with Swedish data and has been installed on a laptop computer. The system holds cause-specific data on deaths, cancer registrations and hospital admissions using *"Stockholm's läns basområden"* to locate cases. Each record represents an event occurring to an individual and includes date of birth, sex, cause and/or procedure and, most crucially, a geo-reference in the form of the variable *baskod99*. Denominator data is currently obtained from the central population register (CPR) of Statistics Sweden.

The full implementation of the RIF in Stockholm has been finalised in close cooperation between the Stockholm and the London partners in the project. The data for the case study has been prepared and organized in Stockholm. The analyses have, however, mainly been carried out at SAHSU in London. This was necessary due to the serious illness of one staff member in Stockholm combined with the fact that both the senior investigators in Stockholm moved to new positions during the project time. One of the senior investigators (CR) has, however, continued to work with the project from distance. The distance work has been complicated by strong security restrictions in the new computer environment, and that has slowed down interaction with the staff members in Stockholm. On the other hand, it has generated extended request for collaboration with the EUROHEIS project, including implementing the RIF in other parts of Sweden (currently in one of the EU support areas).

CR has also been part of the organizing committee for the EUROHEIS Conference (March 2003) including being a local representative for the Conference.

The Stockholm RIF system holds a range of geographical, socio-economic and environmental data, all of which are geographically referenced. Using the RIF these datasets can be integrated, analysed and displayed. Further details of the data held within the Swedish system are given in Table 29.

The final preparations for the case study revealed some features of the database handling routines, which need to be adjusted for to achieve smooth implementation of the RIF in any EU country. This was very useful information to gain in the early stage of transfer of the RIF to new users/user countries.

Table 29. Swedish RIF databases.

GIS:
Stockholm county is divided into:
6 hospital administration areas
24 communities (where Stockholm community is one)
18 community-units in Stockholm community (counts equal to the other 23 communities)
1412 baseunits.

Socioec index:
Townsends index. Min-value=-5.2, max value=21.2, for all geographical levels.

Environmental data:
NO_2-values from 1995, for all geographical levels.
Originally values were based on a model from the Environmental department in Sockholm city. The resolution in the model was 100x100 up to 2000x2000 m, the highest resolution in central city areas and the lowest resolution in the countryside. The predicted value from the square that contained the centroid for each base unit was used.

Population	Time period:	1990-2000
	Textfile:	pop_ora_9000.txt
	Geog units:	baskod99, (kdel99), komdel99, kom99, so99
	Records:	509,168
Heart	Time period:	1990-99
	Textfile:	hi_ora_9095.txt
	Geog units:	baskod99, kdel99, komdel99, kom99, so99
	Icd-codes:	1990-96: icd9, 1997-99: icd10
	Records:	56,260
Mortality	Time period:	1990-98
	Textfile:	dod_ora_9098.txt
	Geog units:	baskod99 (not 1996 and 1998), (kdel99), komdel99, kom99, so99
	1998:	ages in fiveyear-classes, not oneyear
	Icd-codes:	1990-96: icd9, 1997-98: icd10
	Records:	142,000
Cancer	Time period:	1990-99
	Textfile:	ca_ora_9099.txt
	Geog units:	baskod99, kdel99, komdel99, kom99, so99
	Icd-codes	1990-99: icd9
	Records:	75,357
Incare	Time period:	1990-2000
	Textfile:	slv_ora_9000.txt
	Geog units:	baskod99 (not 1996), (kdel99), komdel99, kom99, so99
	Icd-codes	1990-1996: icd9, 1997-2000: icd10
	Records:	3,058,243

Case study: Acute myocardial infarction (AMI) and socio-economic deprivation in Stockholm County

Acute Myocardial Infarction (AMI) is a serious public health problem. Around 40,0000 people in Sweden suffer from myocardial infarction each year. This study aimed to illustrate the disease mapping potential of the Rapid Inquiry Facility for showing the geographical relationship between acute myocardial infarction in people aged under 75 and socio-economic deprivation as well as a potential geographical relation to air pollution.

The study base consisted of the population of Stockholm County aged under 75 years old between 1990–1999. Within the Stockholm County Acute Myocardial Infarction (AMI) register, new cases AMI were identified by combining information of hospital discharges and deaths in accordance with a method that was developed and evaluated previously [1 and 2]. From the National Cause of Death Register, all those with AMI (International Classification of Diseases (ICD), ninth revision, code 410) as the underlying or contributory cause of death among residents in Stockholm County during the study period were selected. Similarly, all patients discharged with a diagnosis of AMI (ICD ninth revision code 410) were obtained from the Stockholm county hospital discharge register. As a main rule, if two registration dates for the same person differed more than 28 days two cases of AMI were recorded, otherwise the two registrations were considered to belong to the same AMI episode. If a person had not been registered for a hospital discharge due to AMI for at least eight years, the case was considered to be a first AMI. This classification of first events has been evaluated in previous studies [3].

An evaluation of the diagnostic quality has been carried out for all first AMI cases 45–70 years old that occurred during 1992–1994 [4]. During these years a population based case referent study was carried out in Stockholm with an aim to identify all first AMI cases in the population 45–70 years old, the Stockholm Heart Epidemiology Program (SHEEP) [5]. The cases of this population identified in the present study were compared to the SHEEP cases by record linkage. For those cases not identified by both methods, medical records were examined with regard to fulfilment of the diagnostic criteria accepted by the Swedish Association of Cardiologists in 1991. According to these criteria, an AMI was considered to be present if there were typical symptoms in combination with typical enzyme changes or appearance of a new pathologic Q-wave on ECG. For non-hospitalised fatal cases, death certificates were examined to see whether or not autopsy had been performed and to check the diagnostic information by comparing the diagnostic codes in the National Cause of Death Register to the diagnoses stated on the death certificate.

Townsend index of relative deprivation

The Townsend index is a measure of relative deprivation developed in Britain 1988 by Peter Townsend, Peter Phillmore and Alastair Beattie[6]. This index was used to determine relative material deprivation in the Stockholm metropolitan area 1990-1993. The level of analysis were small residential areas (Stockholm's läns basområden). The index is a summary measure combining: unemployment, car ownership, home ownership, overcrowding (see Table 30).

Unemployment and overcrowding were transformed to a log scale, as the distribution was skewed. Then z scores were estimated of the four indicators (by subtracting the

overall mean and dividing by the standard deviation). All areas were in that way assigned a value that reflects the deviation of the value from the mean. The four values were summed to create the index score.

In areas with a missing value on one of the indicators (n=29), a method of imputation was used. A regression model was fitted with available data and based on the regression equation predictions were made of the missing values. The Townsend score range from –5.2 to 21.2 (Q1: –3.0, median: –0.3, Q3: 2.7). In 1990 Stockholm county comprised 1132 small residential areas, 982 urban and 150 rural. 124 areas had less than 20 inhabitants and were not included due to the instability of proportions based on small numbers. Finally, the level of material deprivation was calculated for 859 small residential areas.

Table 30. Indicators in Townsend deprivation index.

Unemployment	The percentage of economically active residents aged 16-64 who are unemployed
Car ownership	The percentage of private households who do not posses a car.
Home ownership	The percentage of private households not owner occupied.
Overcrowding	The percentage of private households with children* with > 2 person per room (kitchen and 1 room uncounted)

*Originally, households without children are included, but that information is not available in Sweden.

Methods

We used the HEART data supplied from the Stockholm County AMI register to carry out the analysis using the Swedish Rapid Inquiry Facility (RIF).

The population studied was the whole of Stockholm County. Standardised Incidence Ratios (SIRs) adjusted for age, sex and socioeconomic deprivation were calculated for all first AMIs over the years 1990-2000 by 24 communities (where Stockholm community is one) and 18 community-units in Stockholm community. Expected figures were derived from the whole Stockholm County Area.

Results

There were 15,902,471 person years within the study period 1990-1999. There was ana average of 1,590,000 residents within the Stockholm County area. There were 56,260 records of acute MIs within the database.

Figure 24 shows the Stockholm County area covered by the study with the individual communities and community units. Figure 25 shows the socio-economic deprivation by community using Townsend Quintiles. The darker areas are the most deprived. Figure 26 shows smoothed relative risk of first AMI, indirectly standardised by age and sex for men and women aged under 75 over 1990-99.

Figure 27shows smoothed relative risk for the same population and time period, but adjusted for socio-economic deprivation using the Townsend score. Both maps have been smoothed using empirical Bayes smoothing.

Figure 24. Geographical units (KOMDEL99) in the Stockholm County area.

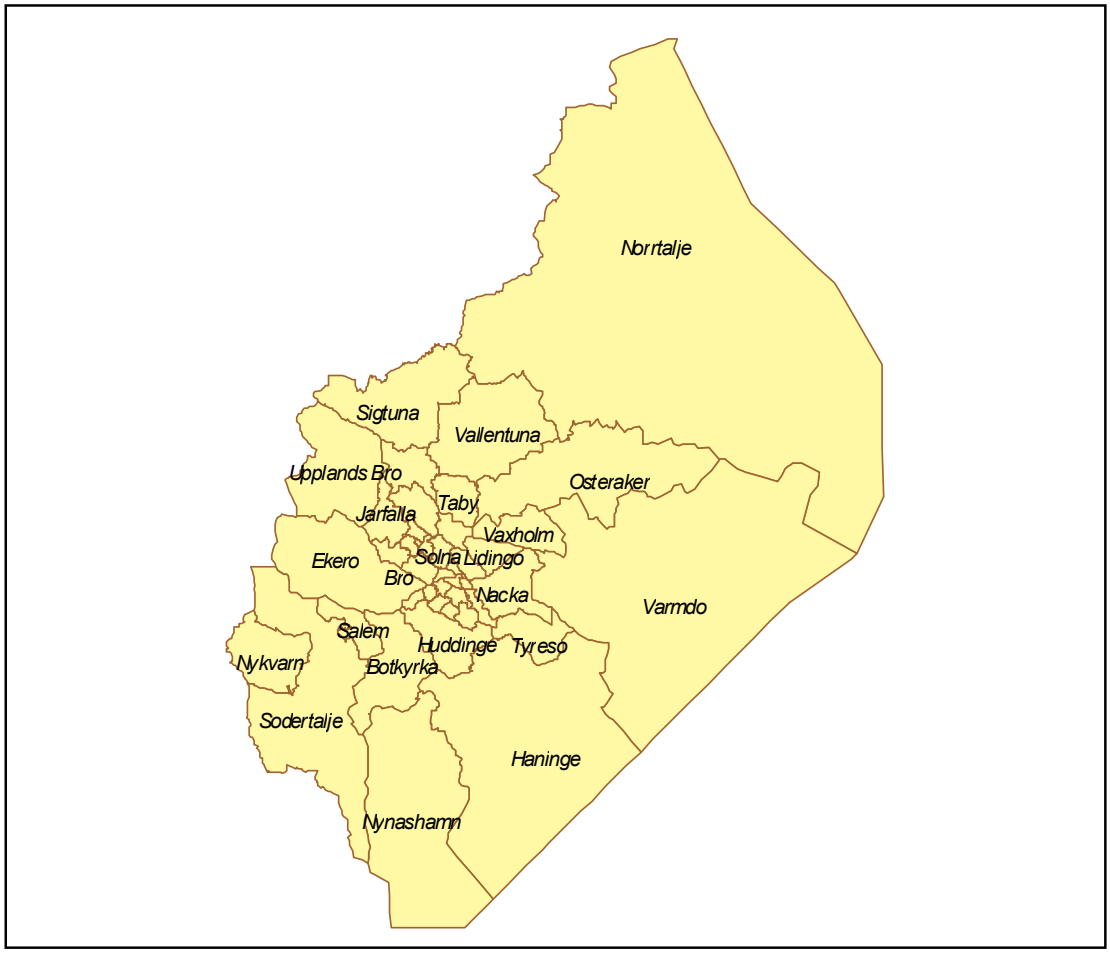


Figure 25. Socio-economic deprivation in quintiles within Stockholm County area. (Darker areas are most deprived).

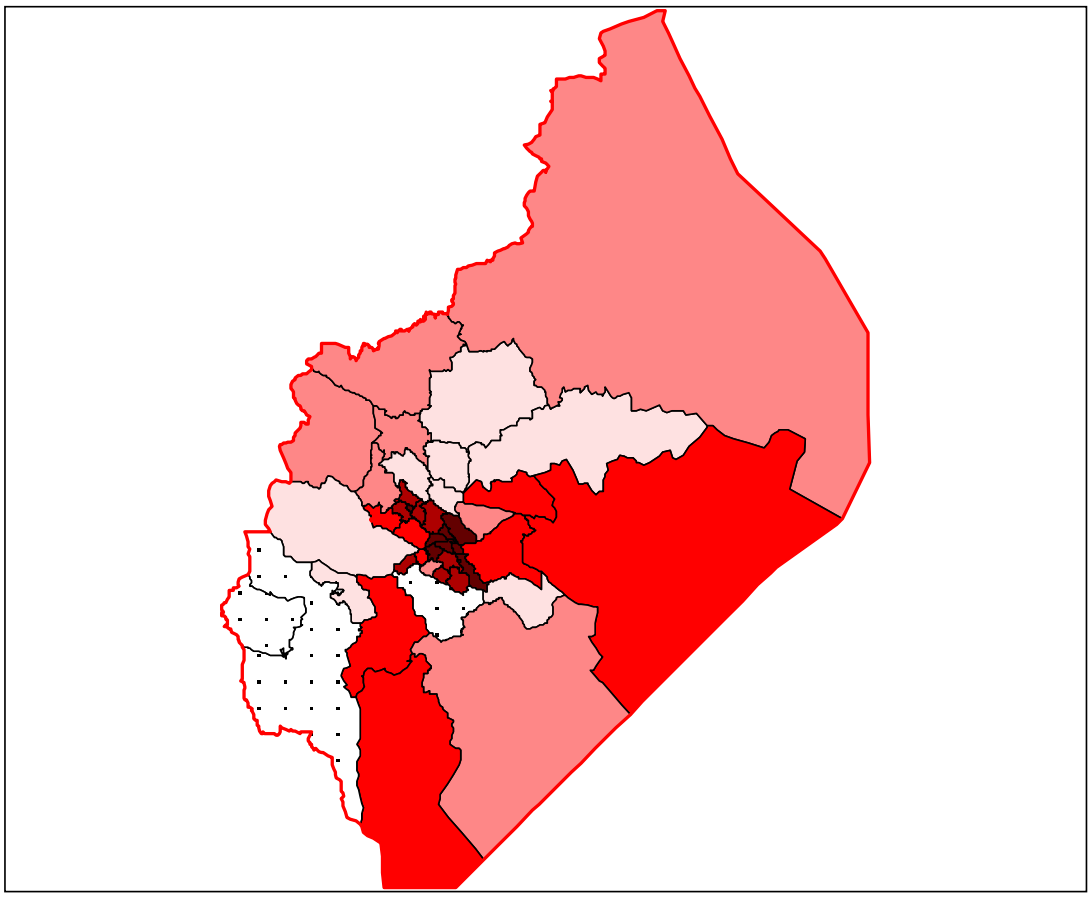


Figure 26. Smoothed Relative Risk, indirectly standardised by age and sex for men and women aged under 75, 1990-99.

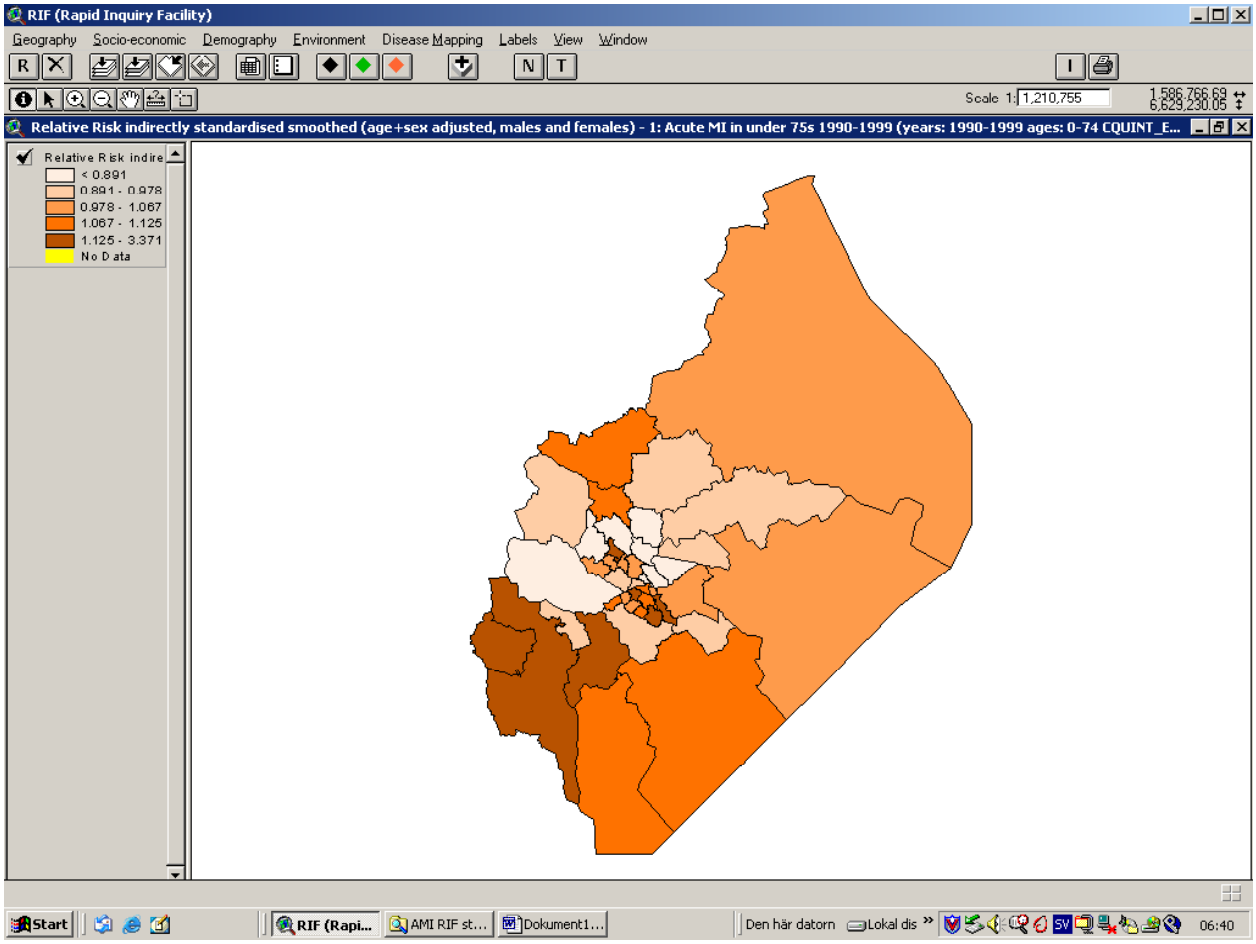


Figure 27. Smoothed map of Relative Risk indirectly standardised by age, sex and socio-economic deprivation for men and women aged under 75, 1990-99.

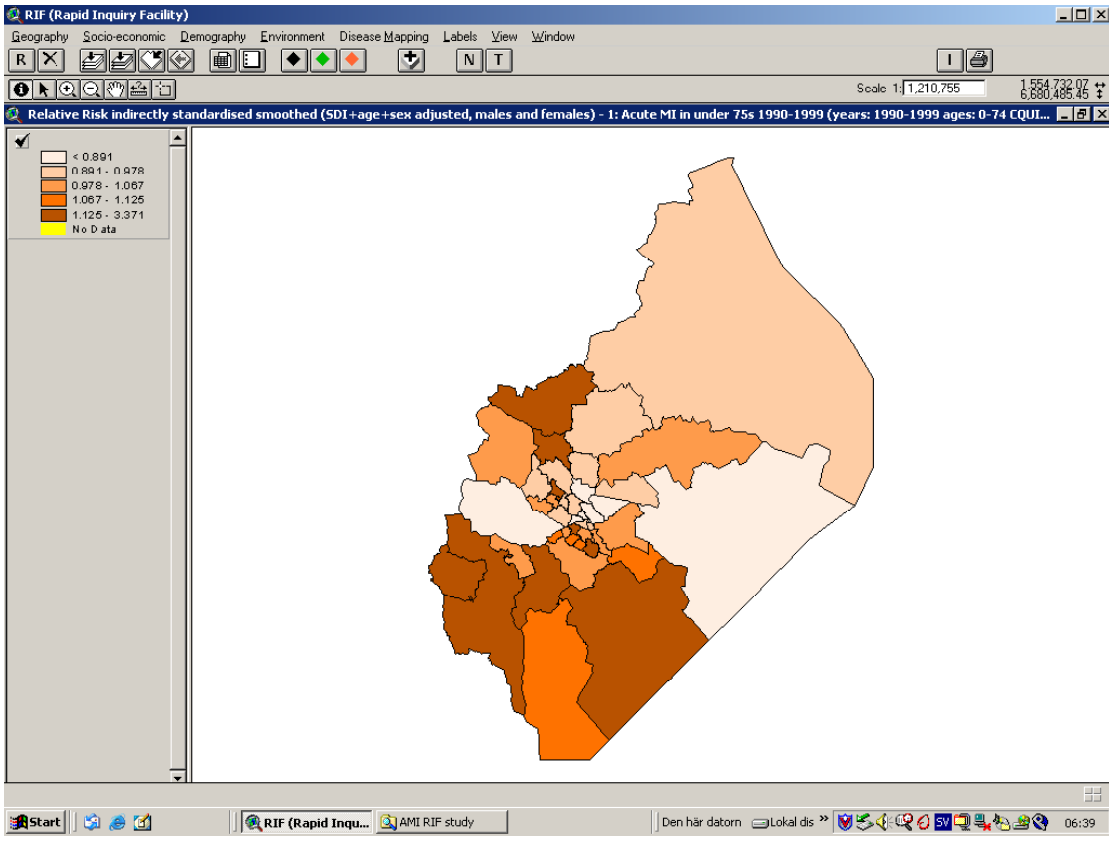
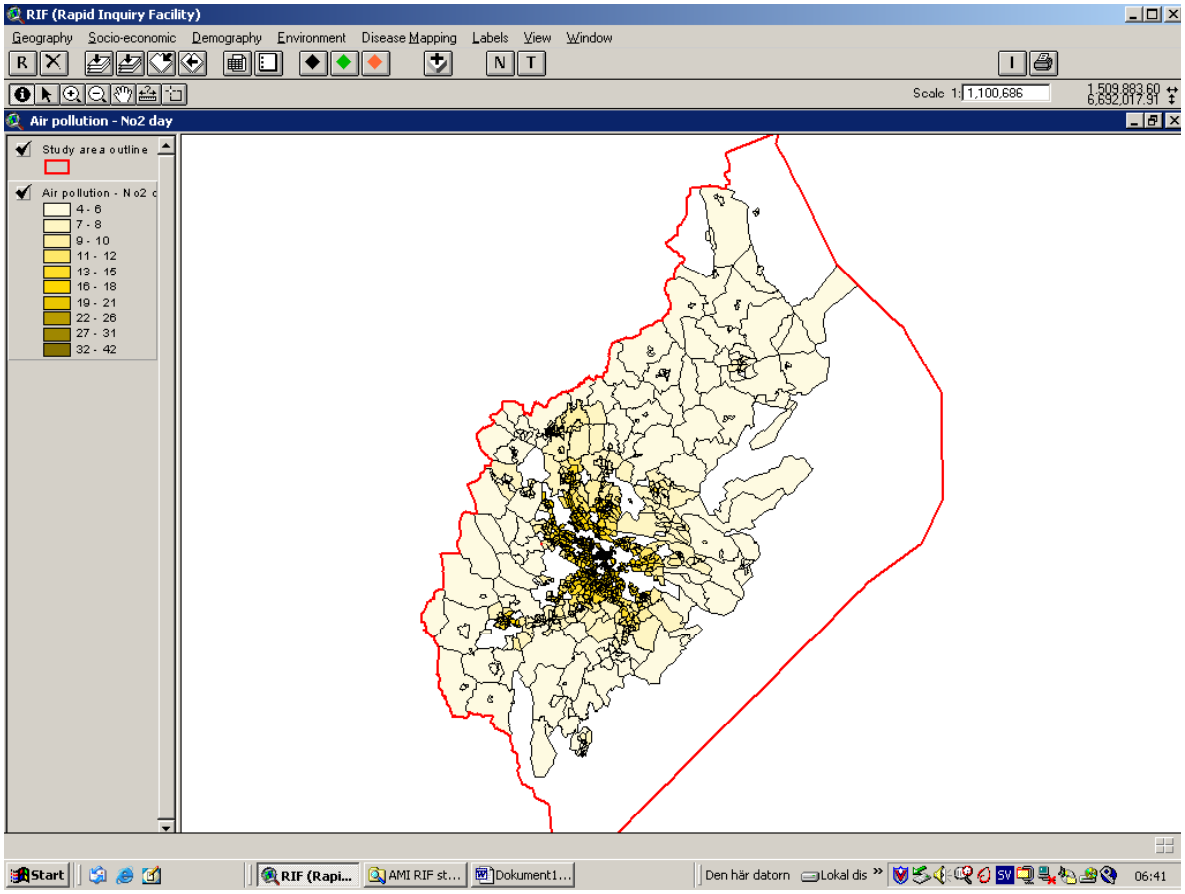


Figure 28. Map of NO_2 concentrations in the Stockholm County area.



Conclusions

This case study demonstrates the utility of the RIF in mapping not only patterns of disease within a study area, but other factors such as socio-economic deprivation and air pollution. Further work is required to disentangle the relationship between these factors and Acute MI.

References

1. A. Ahlbom, Acute myocardial infarction in Stockholm — a medical information system as an epidemiological tool. *Int. J. Epidemiol.* **7** (1978), pp. 271–276.

2. N. Hammar, C. Nerbrand, G. Ahimark *et al.*, Identification of cases of myocardial infarction: hospital discharge data and mortality data compared to myocardial infarction community registers. *Int. J. Epidemiol.* **20** (1991), pp. 114–120.

3. Hammar N. Myocardial infarction in southern and mid-Sweden. Regional differences, occupational differences, and time trends. Doctoral dissertation. Stockholm: Karolinska Institutet, 1992.

4. A. Gustavsson, E. Sandberg, N. Hammar, C. Reuterwall, J. Hallqvist and A. Ahlbom, Identifiering av förstagångsfall av hjärtinfarkt i en definierad befolkning: en jämförelse av två metoder. *Hygiea* **106** 2 (1997), p. 125.

5. C. Reuterwall, J. Hallqvist, A. Ahlbom *et al.*, Higher relative, but lower absolute risks of myocardial infarction in women than in men: analysis of some major risk factors in the SHEEP study. *J. Int. Med.* **246** (1999), pp. 161–174.

6. Townsend P, Phillmore P, Beattie A. Health and deprivation, inequality and the north. Croom Helm. 1988.

United Kingdom

Dr Paul Aylin
Dr Lars Jarup

Small Area Health Statistics Unit (SAHSU)
Dept. Epidemiology and Public Health
Imperial College London
UK

Implementation and dissemination

The first phase of the project was designed to examine the feasibility of installing a system in each of the partner countries, based on the Rapid Inquiry Facility (RIF) developed at Imperial College. The RIF combines geographical information software and health and environmental databases in an easy to use exploratory tool, which allows the assessment of risk and disease mapping at a small area level. Further development of the RIF was carried out to facilitate the transfer to the other partners.

The second phase of the project allowed the development and installation of a RIF to varying degrees in partner countries, taking into account the feasibility study. The UK partner made available the software and its expertise and experience to facilitate installation and setting up of systems using the RIF program in Spain, Sweden and the Netherlands. This has sometimes involved modifying the original code to accommodate different datasets. We summarise the development of this system in the next section. Further documentation has been produced to aid installation and support of the system (Annex 1).

The objective for this, the third phase of the project, is to demonstrate the usefulness of the RIF in answering questions concerning environmental health risks, utilising the system within the context of improving public health, preventing human illness and diseases, and obviating sources of danger to health. Within this section we report on a case study commissioned by East and North Herts Health Authority to investigate the incidence of cancer within a geographical area exposed high levels of bromates in the drinking water.

We also report on work, which has developed as a direct result of the collaborative network of the EUROHEIS project. It does not form part of the formal work packages agreed under the EUROHEIS contract; rather, it is illustrative of the 'added-value' that can arise from such projects. The Department of Geography at the University of Southampton, the driving force for this work, was not a formal partner in the EUROHEIS programme during 2002-2003, but its lead investigator, Samantha Cockings, was heavily involved in initial development of the GIS aspects of the UK RIF at Imperial College, London, and in the feasibility stages of the EUROHEIS project.

The UK partner has also taken the lead on the development and maintenance of the EUROHEIS website, which includes details of the project, project partners, interim reports, meeting proceedings and a demonstration of the RIF.

The development of the Rapid Inquiry Facility (RIF)

The Small Area Health Statistics Unit (SAHSU) within the UK department holds national cause-specific data on deaths, births, cancer registrations, hospital admissions, and congenital anomalies, using the postcode of residence to locate cases to within 10–100 metres. Each record represents an event occurring to an individual and includes date of birth, sex, cause and/or procedure and, most crucially, a geo-reference in the form of a postcode. In 2000, there were around 1.6 million residential postcodes in use in the UK containing, on average, around 14 unique addresses each.

Denominator data is currently obtained from the 1991 Census at enumeration district (ED) level for age and sex in five year age-bands. Using annual births and deaths and a migration factor based on the estimate for the whole health authority, mid-year enumeration district population estimates have been derived up to 2000.

SAHSU also holds a range of geographical, socio-economic and environmental data, all of which are geographically referenced. Using in-house database, statistics and GIS technology and expertise, these datasets can be integrated, analysed and displayed. The resultant store of data is large (as shown in Table 1) and requires substantial hardware to cope with the analysis.

Table 31. Data held at the Small Area Health Statistics Unit in 2002

Dataset	Years	Years	No. records	Total (Mb)
Cancer	1974–1998	25	6,783,072	1,218
Births	1981–2000	18	14,400,364	1,426
Deaths	1981–2000	18	12,657,648	1,664
Hospital Admissions:				
England (HES)	1992–1999/2000	7	76,421.934	20,500
Scotland	1992–1999/2000	7	5,740,948	772
Wales (PEDW)	1991–2000	9	7,241,139	1,420
Northern Ireland (HIS)	1991/2–1994/5	4	1,566,000	357
Populations	1981, 1991 census		1,672,000	472
Socio-economic variables	1981, 1991 census		1,971,000	103
Postcode data			2,041,000	688
Geographical data	[GIS datasets}			5,000
Other data				18,000

HES Hospital Episode Statistics
PEDW Patient Episode Database for Wales
HIS Hospital Information System

The Rapid Inquiry Facility was originally developed in the mid-1990s, with the aim of facilitating the calculation of disease risk around a point source of pollution. A customised system was developed, based around the SAHSU database. The system was able to calculate, relatively rapidly, the indirectly Standardised Mortality Ratio (SMR) and the Standardised Incidence Ratio (SIR), by dividing the observed number of health events by the expected number (calculated based on a set of reference

rates). Reference rates were pre-calculated for speed, using rates from the UK Standard Region within which the study area was located as the reference. Using the Carstairs' index, disease risks were also adjusted to allow for the potential confounding effect of deprivation. The Carstairs' index is a small area deprivation measure derived from UK Census variables (overcrowding, access to a car, unemployment and social class of head of household), which has been shown to be strongly predictive of mortality and cancer incidence.

For estimating the risk surrounding a point source, concentric bands (usually of radius 2km and 7.5km) were drawn around the source (specified either as a national grid reference or as a postcode which was then converted to a grid reference). The enumeration districts, which had their population-weighted centroid falling within the bands, were then selected to form the study area and the risk was calculated for each of the bands. To acknowledge sampling variability, 95% confidence intervals for these risks were also calculated. The system could also generate contextual maps of the study area, together with basic disease maps.

The version of the Rapid Inquiry Facility described above and in Aylin et al. (1999) was fairly restrictive for users as it was developed within a Unix environment and required that they acquire a number of software packages and development tools including Tcl/Tk, ORACLE PL/SQL, ArcInfo AML, C and HTML. A key part of the EUROHEIS project was to make the Rapid Inquiry Facility more generic so that users from other countries would be able to adapt and implement the system. The latest version therefore employs two relatively common and affordable software packages: ArcView 3.2 and Oracle (a copy of Oracle Personal is sufficient) and the code has been written more generically to enable it to operate on datasets that are in the appropriate, pre-specified, format. The Rapid Inquiry Facility is therefore now platform independent and more exportable for other users. Additional functionality has also been developed, including the ability to calculate directly standardised rates and extended output options such as the generation of contextual and disease maps in .jpg and .wmf format, and tabular or text-based output in .html or comma separated format. A further additional feature is the production of smoothed maps of disease risk. For small area disease mapping, large differences in health risk between small areas may arise simply due to chance, even when several years of data are used. This is particularly true when the numbers of cases are very small (for example typically, an electoral ward will have fewer than 10 deaths from heart disease in the under 75s per year). The system applies empirical Bayesian smoothing to the risk estimates giving a more robust estimate of the 'true' ward relative risks than the raw SMRs.

Case study - An investigation into cancer incidence in areas exposed to high levels of bromate in East and West Herts

Dr Paul Aylin, Pauline Savigny, Susan Hodgson (SAHSU)

Introduction

Recent WHO guidelines have suggested a standard of no more than 25μg/litre of bromates in drinking water with a future EU standard of 10μg/litre.[1] In preparation for these standards, during routine monitoring, Three Valleys Water company discovered unusually high levels of bromates (above 150μg/litre) in water supplies around the Hatfield area and concentrations as high as 3000μg/litre in several private boreholes at Sandridge. This contamination is likely to have come from a chemical plant in Sandridge that operated from 1958 to 1972, manufacturing sodium and potassium bromates.

There is currently no conclusive epidemiological data that potassium bromate is a human carcinogen.[2] Most cases of human poisoning from bromate are due to the accidental or intentional ingestion of home permanent wave solutions, which contain 2-10% bromate. In children serious poisonings have been reported following ingestion of 60-120ml 2% potassium bromate (~46-92mg bromate per kg body weight per day for a 20kg child). Lethal oral doses of bromate are estimated to be 154-385mg/kg body weight.[3]

There is evidence in experimental animals of the carcinogenicity of potassium bromate. In rats, oral administration produces renal tubular tumours (adenomas and carcinomas) and thyroid follicular tumours in rats of each sex, and peritoneal mesotheliomas in males. In mice, oral administration produces a low incidence of renal tubular tumours in males, and in hamsters the incidence was also marginally increased. The International Agency for Research on Cancer classifies potassium bromate as possibly carcinogenic to humans (Group 2B).[4]

There are very limited data on the developmental and reproductive effects of bromate. A single, short-term assay exposing male and female rats to sodium bromate before and during gestation showed no developmental toxicity, however a decrease in epididymal sperm concentration was found in males.[4] The reproductive effects of potassium bromate were also evaluated in a study in which rats and mice were fed flour treated with 15mg of potassium bromate per kg over five and eight generations respectively. No effects on reproductive performance or survival were observed in either species.[3] Despite evidence in experimental animals of the carcinogenicity of potassium bromate, conclusive epidemiological data on its human carcinogenicity is still lacking, and the International Agency for Research on Cancer classifies it as possibly carcinogenic to humans.[4]

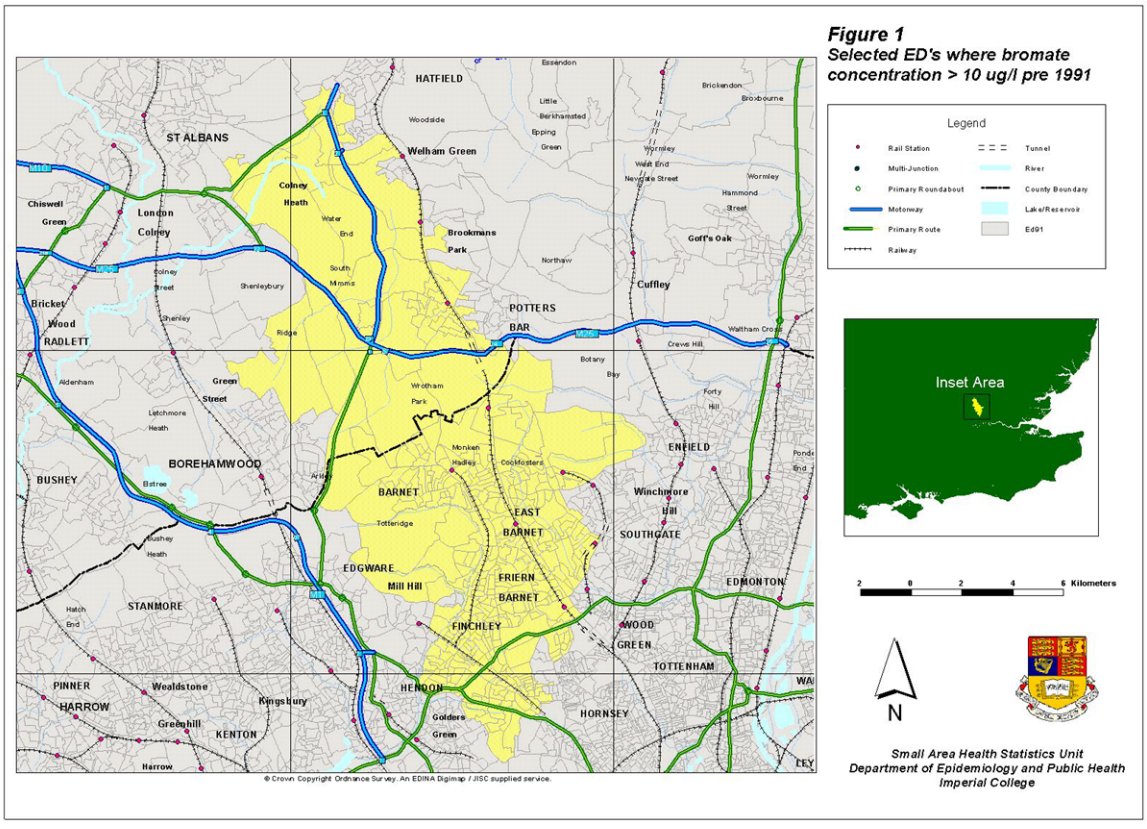
This study uses routinely collected health data to investigate the incidence and relative risk of selected cancers in the affected area in East and West Herts and is an update of a report originally commissioned by the East and North Herts Health Authority to include more recent cancer registration data.

Methods

We used cancer incidence data supplied by the Office for National Statistics (ONS) and held by the Small Area Health Statistics Unit (SAHSU) to carry out the analysis using the SAHSU Rapid Inquiry Facility (RIF).[5]

Two approaches were taken in defining the exposed population. The first involved selecting enumeration districts falling within contaminated water supply boundaries. Lists of Enumeration Districts (EDs) with population centroids within the boundaries of the affected water supply areas were obtained from Three Valleys Water. Contaminated water supplies were classified as those where the bromate concentrations exceeded the proposed EU standard of 10µg/litre in the period pre 1991 (Figure 29). This area included Potters Bar West, Arkley Reservoir, South Hatfield, Bullens Green and North Mymms; The second approach used distance from the Sandridge plant as a proxy for exposure, including EDs whose population centroids were within a 2km and 7.5km radius.

Figure 29. Bromate concentrations in water supplies.



Standardised Incidence Ratios (SIRs) adjusted for age, sex and socioeconomic deprivation[6] were calculated for all malignant neoplasm (ICD9 140-208; ICD10 C00-C97), malignant neoplasm of the thyroid gland (ICD9 193; ICD10 C73), malignant kidney neoplasm (ICD9 189; ICD10 C64), malignant neoplasm of the bladder (ICD9 188; ICD10 C67) and leukaemia (ICD9 204-208; ICD10 C91-C95) over the years 1979-1998. Expected figures were derived from the South East government region population.[5]

Results

Within the contaminated water supply boundaries there were on average throughout the 20 year study period, 89,837 men and 97,066 women giving a total of 3,738,043 person years. There were 15,312 observed malignant neoplasm registrations in the population living within the contaminated water supply boundaries between 1979 and 1998 (Table 32). After adjusting for age, sex and deprivation, there was a statistically significant low risk for all malignant neoplasms (SIR 85, 95% CI 83-85 for both men and women). For men and women combined there were no statistically significant excess risks for any of the cancers examined.

Table 32. Standardised Incidence Ratios (SIR) of cancer registrations within contaminated water supply boundaries, adjusted for age, sex and deprivation 1979-1998.

Endpoint	Sex	Observed	Expected	SIR	95% CI
All malignant neoplasm	Males	7517	8845.25	85	83-86
	Females	7795	9255.92	84	82-86
	Persons	15312	18101.17	85	83-85
Malignant neoplasm of the Thyroid gland	Males	20	18.68	107	65-165
	Females	44	53.7	82	60-110
	Persons	64	72.38	88	68-113
Malignant neoplasm of the kidney	Males	160	182.53	88	75-102
	Females	99	109.66	90	73-110
	Persons	259	292.18	89	78-100
Malignant neoplasm of the bladder	Males	499	611.15	82	75-89
	Females	207	244.48	85	74-97
	Persons	706	855.64	83	77-89
Leukaemia	Males	203	224.66	90	79-104
	Females	183	196.6	93	81-108
	Persons	386	421.26	92	83-101

Within 2km of the old Sandridge chemical factory there were on average throughout the 20 year study period, 5,455 men and 5,530 women giving a total of 219,693 person years. Within 2-7.5km there were on average, 73,249 men and 75,626 women giving a total of 2,977,501 person years. There were 11,944 observed cancer registrations in the population living within 7.5 km of the factory (Table 33).

Table 33. Standardised Incidence Ratios (SIR) of cancer registrations within 0-2km and 2-7.5km distance bands around Sandridge, adjusted for age, sex and deprivation, 1979-1998.

		Distance Band					
		0-2km			2-7.5km		
Endpoint	Sex	Observed	SIR	95% CI	Observed	SIR	95% CI
All malignant neoplasms	Males	360	88	79-97	5512	80	78-83
	Females	373	89	81-99	5699	82	80-85
	Persons	733	88	82-95	11211	81	80-83
Malignant neoplasm of the thyroid gland	Males	2	195	24-705	11	73	37-131
	Females	5	178	58-415	34	84	58-117
	Persons	7	182	73-376	45	81	59-109
Malignant neoplasm of the kidney	Males	5	55	18-128	149	103	87-121
	Females	2	41	05-149	72	87	68-110
	Persons	7	50	20-103	221	97	85-111
Malignant neoplasm of the bladder	Males	29	106	71-152	355	75	68-84
	Females	4	41	11-105	138	78	66-92
	Persons	33	89	61-125	493	76	70-83
Leukaemia	Males	10	90	43-165	154	88	75-103
	Females	7	79	32-164	117	80	67-96
	Persons	17	85	50-137	271	85	75-95

The SIRs indicate no excess risk for any of the specified conditions; there was an overall lower risk for all cancers in populations living within 2km of the plant and in those living 2.-7.5km away (SIR 88, 95% CI 82-95 and SIR 81. 95% CI 80-83 respectively). For the population living in the 2.-7.5km area, a low risk of bladder neoplasm for males and females (SIR 75, 95% CI 68-84 and SIR 78, 95% CI 66-92 respectively) and of leukaemia for females (SIR 80, 95% CI 67-96) were observed after adjusting for age, sex and deprivation.

Discussion

Overall there did not appear to be any excess risk of malignant neoplasms in people living within the contaminated water supply area or within 7.5km of the old factory site.

The plant was operational between 1958 and 1972 and reliable cancer incidence was only available for the period 1979 to 1998. Thus there are 41 years between the plant becoming operational and the last cancer registration. There might be expected to be some migration of people into and out of the area during this time which will tend to attenuate any observed link between exposure and health risks, as exposed individuals will move out of the area and unexposed individuals move in.

Rapid Inquiry Facility (RIF) reports are made available to authorities to inform their own investigations and are intended to provide a rapid initial screen of the relevant health statistics related to a particular point source or area. They use the postcode to give information on the geographic location of cases and to link to the underlying socio-demographic population statistics (which allows, for example, analyses to be carried out based on distance from a point source). Given the relatively rapid turnaround, the RIF reports are necessarily based only on data held routinely on the Small Area Health Statistics Unit (SAHSU) database, without any scope for further checking of the data. The report will not account for errors in data supplied to SAHSU, which may include under-registration, missing or duplicate postcodes, duplicate registration, errors in coding and classification of cancer cases. In addition, problems associated with cluster investigation may affect RIF reports. These include difficulties in defining boundaries that could affect the observed rates, the risk of multiple comparisons and small numbers of events. RIF reports therefore need to be interpreted with caution and with expert local knowledge. The ecological nature of this type of RIF report implies that caution should be employed when making causal inferences based on observed associations. However, in this study there appeared to be little evidence to support the hypothesis of an excess risk in cancer incidence in populations exposed to relatively high levels of bromate in this area.

Conclusion

Despite exposure to levels of bromate in drinking water above current and proposed guidelines, people living within the study region do not appear to have a higher risk of developing cancer. Indeed, the investigation suggests that for all cancers and most of the diagnoses looked at here, the risk is considerably lower than one might expect. Our initial report (based on data only up to 1993) was submitted to the HA where it was very well received. The report had very little mention in the local press. There is however still some local concern amongst residents.

Small area investigations can be useful in rapidly establishing whether or not there is an association between a pollution source and related health outcomes.

Acknowledgements

Our thanks to Dr Marian McEvoy of East and North Herts Health Authority who commissioned the report and advised on study parameters the Three Valleys Water Company who provided boundaries for contaminated water.

References

[1] European Union 1998. Council Directive 98/83/EC on the quality of water intended for human consumption

[2] Giri U, Iqbal M & Athar M. Potassium bromate (KbrO3) induces renal proliferative response and damage by elaborating oxidative stress. Cancer Letters 135 (1999) 181-188

[3] Guidelines for drinking-water quality 2nd ed. Vol 2. Health criteria and other supporting information. Geneva, World Health Organisation, 1996. pp121-131 & 822-828. (http://www.who.int/water_sanitation_health/dwq/gdwq2v1/en/index1.html accessed July 2003).

[4] IARC monographs 1999, vol 73 pp481.
(http://193.51.164.11/htdocs/monographs/vol73/73-17.html accessed July 2003)

[5] Aylin, P., Maheswaran, R., Wakefield, J., Cockings, S., Jarup, L., Arnold, R., Wheeler, G., and Elliott, P. A National Facility for Small Area Disease Mapping and Rapid Initial Assessment of Apparent Disease Clusters Around a Point Source: the UK Small Area Health Statistics Unit. Journal of Public Health Medicine 1999;21(3):289-98.

[6] Carstairs, V. and Morris, R. Deprivation and Health in Scotland. Health Bulletin 1990;48(4):162-75.

Towards zone design methods for environment and health studies

Samantha Cockings
Department of Geography
Southampton University
UK

Context

This section reports on work, which has been developed as a direct result of the collaborative network of the EUROHEIS project. It does not form part of the formal work packages agreed under the EUROHEIS contract; rather, it is illustrative of the 'added-value' that can arise from such projects. The Department of Geography at the University of Southampton, the driving force for this work, was not a formal partner in the EUROHEIS programme during 2002-2003, but its lead investigator, Samantha Cockings, was heavily involved in initial development of the GIS aspects of the UK RIF at Imperial College, London, and in the feasibility stages of the EUROHEIS project. Since moving to the University of Southampton, Ms Cockings has continued her involvement with the project by advising partners on the adoption, adaptation and implementation of RIF-style systems in various countries. Further, she has developed a new strand of research concerned with the design of areas for environment and health studies, which has considerable potential for use in the EUROHEIS project. This work is the focus of the research reported in this section.

Automated zone design

Many environmental health studies use geographically aggregated data, either because individual level data do not exist or because confidentiality restrictions prohibit their use. Geographical areas (for example, health authorities, Census regions or municipalities) may also be used for the calculation of rates, for visualisation or mapping purposes, or to aid area level policy and planning decisions. The problem is that the design of such geographical areas (in terms of their size, shape and the characteristics of the population that they contain) can have a significant influence on any relationships observed and therefore may impact on policy decisions made using such areas. Openshaw (1984) terms this the modifiable areal unit problem (MAUP). Openshaw also describes automated zoning procedures (AZP) for designing zones given a set of design constraints (Openshaw, 1977) and this work is further developed in Openshaw and Rao (1995). These procedures use the principle of iterative recombination of building blocks (the smallest set of zones available for a study) into output areas so as to maximise the value of some pre-defined objective function. In more recent years, Martin (2002) and Martin et al (2001) have developed and implemented automated zone design methods for the collection and dissemination of (population) Census data in the UK. Automated zone matching (AZM) software has been developed by Martin (2003), which enables the AZP technique to be applied in other applications where zone design is required.

To date though, there have been few attempts to explore the potential usefulness of such techniques for the design of zones for epidemiological studies.

Automated zone design for environment and health studies

We suggest two key areas where automated zone design techniques can contribute to epidemiological studies. First, they enable us to explore the sensitivity of our results to the MAUP. One of the problems of using areas for analysis in epidemiological studies is that we generally have little or no feel for how much our results are dependent on the specific set of areal units employed. By enabling us to repeatedly redesign sets of zones at specified scales (determined for instance by population size) and recording the observed relationships, automated zone design techniques offer the ability to explore the sensitivity of our results to changes in the size of the units employed (the scale element of MAUP). Furthermore, by designing various sets of zones at the same scale, we can also explore the aggregation element of the MAUP – essentially we can record how our results vary simply by placing the boundaries differently at a given scale. Second, zone design techniques can contribute significantly to epidemiological studies through the design of purpose specific sets of zones for such studies. In this context, we perceive four key areas where zone design techniques may be used: (i) where achieving stable estimates is problematic, for instance in areas of widely varying population density: zones of homogeneous populations may be designed based on minimum threshold and/or target populations; (ii) where we wish to explore spatial patterns of disease: we may design zones of homogeneous disease rates (or similar measures), (iii) where we need to ensure that our zones conform to pre-specified boundaries or barriers, such as higher-level administrative boundaries or geographical features in the landscape: we may build these features into our zone design procedures, and finally, and perhaps most importantly, (iv) where we wish to analyse the relationships between risk factors and health outcomes: we may design zones which represent homogeneous levels of either risk or confounding factors. In all of these design scenarios good epidemiological, statistical and geographical principles and methodologies are essential. There is clearly a concern, as with any automated technique (such the production of maps using GIS), that it may be possible for naïve users to create inappropriate sets of zones and, essentially, to create any pattern or set of results that they wish. However, we suggest that, if used appropriately, the potential benefits of using zone design for spatial analysis in epidemiological studies are great. At the very least, their use highlights the fact that all zoning systems are ‘imposed’ and that they should not be considered neutral for the purposes of analysis.

Empirical examples

We have started to explore the potential usefulness of automated zone design for environment and health studies through the use of empirical examples. To date, we have focused on their usefulness in studies involving pre-aggregated data. This situation is typical of countries, such as the UK, where concerns over disclosure control prohibit the release of individual level data and data are instead aggregated geographically to ensure specified (population)

threshold levels. Preliminary results from this research were presented at the EUROHEIS Conference (Cockings and Martin, 2003) and at the GISRUK 2003 Conference (Martin and Cockings, 2003).

In countries where individual level data are available, there are different scientific and technical challenges. As a direct result of collaborations established through the EUROHEIS project, we have been working with the Danish partner to explore the usefulness of zone design techniques for studies using individual level data. Preliminary research has been based around the use of zone design techniques in an epidemiological investigation into the effects of airborne toxins from a specified point source on the health of the local population. Early thoughts on the possibilities of such applications were presented at an invited seminar on GIS and Health in Copenhagen (Cockings, 2003).

Future work

It is hoped that the work with the Danish partner will continue under the future proposed EUROHEIS programme. Furthermore, discussions with existing EUROHEIS partners indicate that there is a scientific and practical need for the development of zone design techniques for environment and health studies in Sweden, the Netherlands and Finland. Accordingly, the University of Southampton has proposed to become a partner in the EUROHEIS project. We propose to develop tools to design zones for use in environment and health studies. These tools will enable partners to:

1) Explore the sensitivity of observed relationships to different zone designs (both in terms of the scale of the zones and in the placement of the zones' boundaries)
2) Explore the feasibility of designing purpose-specific zones for a given analysis.

The tools will be useful both for partners with individual level data where there is a need to aggregate data according to pre-defined criteria, and for partners with pre-aggregated data where there is a desire to explore alternative aggregations. Zone design criteria will include population size, population threshold and homogeneity in terms of the built and social environment.

References

Cockings, S. (2002) 'Exploring zone design methods for spatial epidemiological studies'. *Seminar on the use of GIS and socio-economic data in health studies*, ESRI, (Informi GIS), Copenhagen, Denmark, November.

Cockings, S. and Martin, D. (2003) 'Zone design methods for epidemiological studies', *EUROHEIS/SAHSU International Conference on Health and Environment*, (in press).

Martin, D. (2002) 'Geography for the 2001 Census in England and Wales', *Population Trends*, 108, 7-15.

Martin, D. (2003) 'Developing the automated zoning procedure to reconcile incompatible zoning systems', *International Journal of Geographical Information Science,* 17, 181-196.

Martin, D. and Cockings, S. (2003) 'Zone design for health and environment research', *Proceedings of the GIS Research UK Conference,* 69.

Martin, D., Nolan, A. and Tranmer, M. (2001) 'The application of zone-design methodology in the 2001 UK Census', *Environment & Planning A*, 33, 1949-1962.

Openshaw, S. (1977) 'A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling', *Transactions of the Institute of British Geographers*, *New Series*, 2, 459-472.

Openshaw, S. (1984) *The Modifiable Areal Unit Problem*, CATMOG 38, Norwich: Geo Books.

Openshaw, S. and Rao, L. (1995) 'Algorithms for reengineering 1991 Census geography', *Environment and Planning A*, 27, 425-446.

Dissemination

EUROHEIS website

http://www.med.ic.ac.uk/divisions/60/euroheis/homepage.htm

will soon be transferred into

www.euroheis.org

Meetings and conferences

EUROHEIS Conference

March 2003: EUROHEIS Conference, Östersund, Jämtland, Sweden. Conference programme including submitted abstracts is included as **Annex 2.** Proceedings from the conference are included in **Annex 3.**

Internal project meetings

September 2000: Department of Epidemiology, Imperial College London, UK

January 2001: Department of Epidemiology, Imperial College London, UK

September 2001: Garmisch-Partenkirchen, Germany

April 2002: Escuela Andaluza de Salud Pública (EASP), Granada, Spain

September 2002: Department of Epidemiology, Imperial College London, UK

Other meetings and conferences

EUROHEIS presentations were also made at the following international conferences

August 2000: ISEE Conference 2000, Buffalo, USA

> Cockings S, Jarup L, Aylin P, Elliott P, Poulstrup A, Reuterwall C, Pekkanen J, Martuzzi M, Ferrandiz J, Staines A, Richardson S. A European health and environment information system for disease and exposure mapping and risk assessment. Abstract 12th conference of the International Society for Environmental Epidemiology, Buffalo, USA, 2000.

April 2001: NATO Advanced Research Workshop GIS for Emergency Preparedness and Health Risk Reduction Budapest, Hungary. NATO Science Programme 2001

September 2001: ISEE Conference 2001, Garmisch-Partenkirchen, Germany

> Aylin P, Jarup L, Cockings S. EUROHEIS (European Health and Environment Information System). Abstract 13th conference of the International Society for Environmental Epidemiology, Garmisch, Germany, 2001.

September 2001: APHEIS-EUROHEIS workshop, Garmisch-Partenkirchen, Germany

August 2002: ISEE Conference 2002, Vancouver, Canada

> Symposium: Health and Environment Information Systems - a European Approach
> Chair: Dr Lars Järup
> Co-Chair: Dr Arne Poulstrup
> Monday 12 August 10:00 - 12:00
>
> EUROHEIS (European Health and Environment Information System) - applications and case studies Aylin, P, Cockings, S, Ferrándiz, J, Hammar, N, Järup, L, Kelly, A, Martuzzi, M, Poulstrup, A, Pekkanen, J
>
> Use of exposure simulation models and health registers integrated with GIS Poulstrup, A, Hansen, HL
>
> Contribution of geographical studies to environment and health decision making Martuzzi, M
>
> Risk of cancer and proximity of residence to a river with high sediment levels of dioxins Verkasalo, PK, Kokki, E, Pukkala, E, Kiviranta, H, Vartiainen, T, Pekkanen, J
>
> Injury mortality in young people: urban-rural differences Staines, A, Boland, M, Scallan, E, Fitzpatrick, P, Laffoy, M, Kelly, A
>
> Exploring zone design methods for a small-area environmental epidemiological study Cockings, S, Poulstrup, A, Martin, D, Hansen, HL
>
> Small area statistics on health (SMASH) - a system for rapid investigations of cancer in Finland Kokki, E, Penttinen, A, Pukkala, E, Verkasalo, PK, Pekkanen, J

September 2002: Zurriaga O. "Sistemas de Información en Salud y Medio Ambiente. Situación actual y perspectivas de futuro. El proyecto Euroheis". *XIII Escuela de Verano de Salud Pública. Mahón. Septiembre 2002.*

November 2002: EUPHA Conference 2002, Dresden, Germany (http://www.eupha.org/html/2002abstractprizes.html).
The EUROHEIS contribution was awarded the prize for best abstract:

> Aylin P. EUROHEIS (European Health and Environment Information System) – applications and case studies

November 2002: Zurriaga O. "Proyecto Euroheis". *Conferencia impartida dentro del curso Sistemas de Información Geográfica en la Evaluación de Riesgos Ambientales en Salud, Madrid. Noviembre 2002.*

Publications

Abellán JJ, Martínez-Beneito MA, Zurriaga O, Jorques G, Ferrándiz J, López-Quílez A. Procesos puntuales como herramienta para el análisis de posibles fuentes de contaminación. Gaceta Sanitaria 2002;**16**:445-9.

Abellán JJ, Zurriaga O, Martínez-Beneito MA, Peñalver J, Molins T. Incorporación de la metodología geoestadística a la vigilancia de la gripe en una red centinela. *Gaceta Sanitaria* 2002;**16**:324-33.

Abellán C., Ferrándiz J. and López A. Imputación espaciotemporal en estudios medioambientales. *Boletín de la SEIO* (in press).

Aylin P, Cockings S. Health and environment information systems. Contributors of GIS in Public Health - Opportunities and Pitfalls. Editor Maheswaran M. Taylor & Francis, London 2003.

Briggs, D.J., Forer, P., Jarup, L. and Stern, R. (eds) 2002 *GIS for emergency preparedness and health risk reduction.* Dordrecht: Kluwer Academic Publishers, 326 pp.

Cockings S, Jarup L. A European Health & Environment Information System for Exposure & Disease Mapping & Risk Assessment (EUROHEIS). In: *GIS for Emergency Preparedness*. Briggs D, Forer P, Jarup L, Stern R. (eds). NATO Science Series, Kluwer Academic Publishers, Dordrecht, NL. 2002, pp.207-226.

Ferrándiz J, Abellán JJ, López A, Sanmartín P, Vanaclocha H, Zurriaga O, Martínez-Beneito MA, Melchor I, Calabuig J (2002). "Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain)". In *GIS for Emergency Preparedness and Health Risk Reduction*, edited by D Briggs, P Forer, L Jarup, R Stern. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Ferrándiz J, López-Quílez A, Gómez-Rubio V, Sanmartín P, Martínez-Beneito MA, Melchor I, Vanaclocha H, Zurriaga O, Ballester F, Gil JM, Pérez-Hoyos S, Abellán JJ. Statistical relationship between hardness of drinking water and cerebrovascular mortality in Valencia: a comparison of spatiotemporal models. *Environmetrics* (in press).

Kokki E, Ranta J, Penttinen A, Pukkala E and Pekkanen J. Small area estimation of incidence of cancer around a point source of exposure with fine resolution data. *Occup Environ Med* 2001;**58**:315-320.

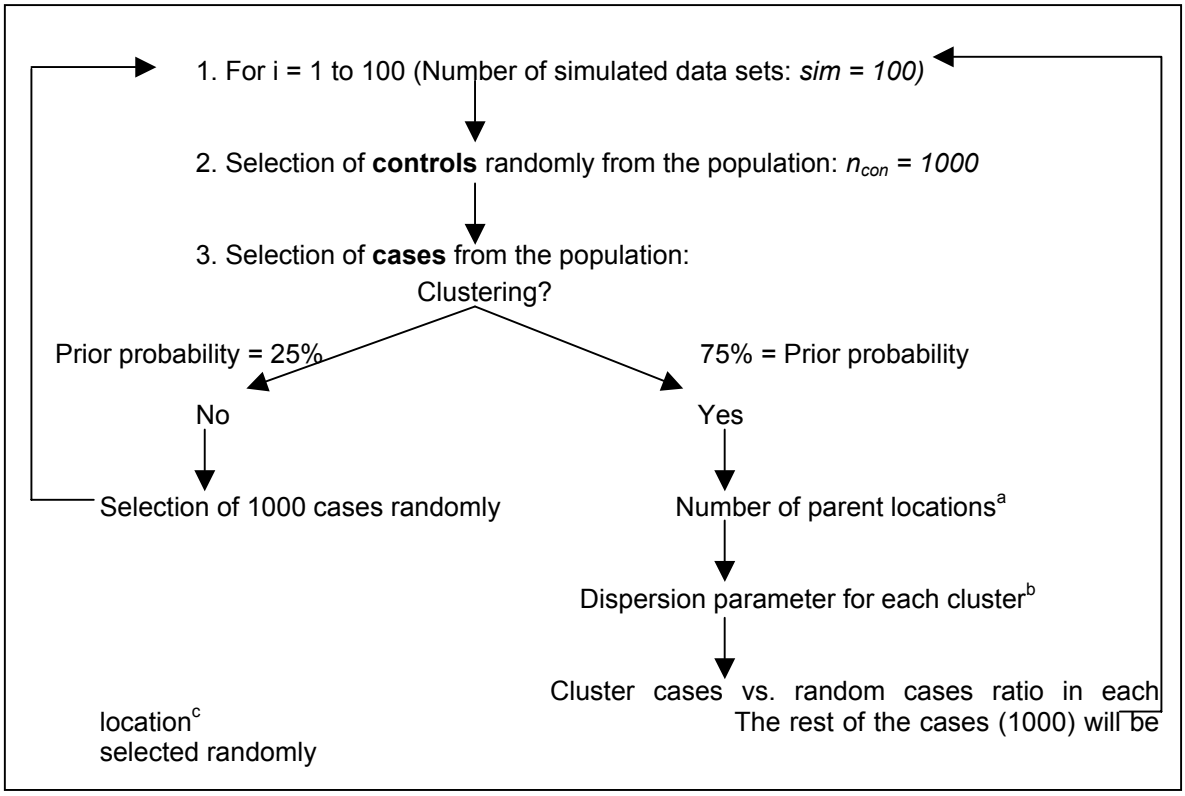
Kokki E, Pukkala E, Verkasalo PK, Pekkanen J. (2002A) Small Area Statistics on Health (SMASH) - A System for Rapid Investigations of Cancer in Finland. In Briggs D, Forer P, Järup L, Stern R (eds.): GIS for Emergency Preparedness and Health Risk Reduction, Kluwer Academic Publishers, Dordrecht.

Several other papers are submitted for publication or are in preparation. In particular, it is worth noting that six papers from the EUROHEIS conference will be included in a Mini-monograph to be published by Environmental Health Perspectives. These are:

- Paul Elliott and Dan Wartenberg. Spatial epidemiology: emerging issues.
- John Nuckols, Mary Ward and Lars Jarup. Using GIS in exposure assessment for environmental epidemiology
- Andrew Thomson, Nicky Best and Sylvia Richardson. A comparison of disease-mapping models for detecting high-risk areas.
- Pia K Verkasalo, Esa Kokki, Eero Pukkala, Terttu Vartiainen and Juha Pekkanen. Cancer risk in farmers living near a river with high sediment levels of dioxins.
- Arne Poulstrup and Henrik L Hansen. Cancer development due to airborne dioxin in a urban population assessed by use of exposure simulation models and GIS-based health registers.
- J. Ferrándiz, A. López-Quílez, V. Gómez-Rubio, P. Sanmartín, M.A. Martínez-Beneito, I. Melchor, H. Vanaclocha, O. Zurriaga, F. Ballester, J.M. Gil, S. Pérez-Hoyos and J.J. Abellán. Spatial analysis of the relationship between cardiovascular mortality and drinking water hardness.

Appendix 1. Ireland

A. Simulation algorithm



a) Number of parent locations (sources)

k	1	2	3	4	5	10	20	40	60	80
Prior probability	0.05	0.05	0.05	0.05	0.1	0.1	0.1	0.1	0.2	0.2

b) Dispersion parameter (Distance from the source in which clustered cases are located (Km))

σ	0.5Km	1 Km	1.25 Km	1.5 Km	1.75 Km	2 Km	2.25 Km	2.5 Km	2.75 Km	3 Km
Prior probability	0.05	0.05	0.2	0.2	0.2	0.1	0.05	0.05	0.05	0.05

c) Cluster cases vs. random cases ratio

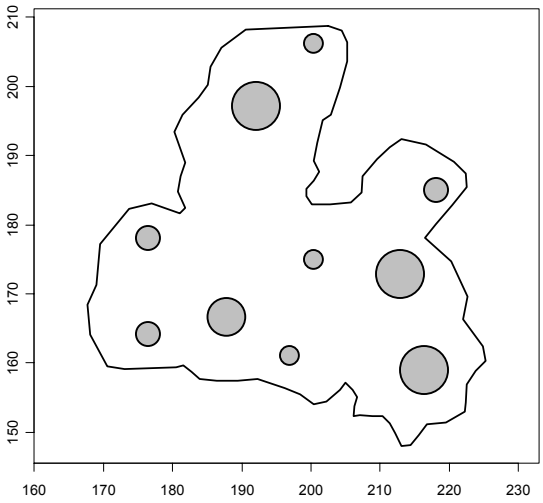
Ratio	0.2	0.4	0.6	0.8	1	1.2	1.5	2	2.5	3
Prior probability	0.05	0.05	0.1	0.15	0.2	0.15	0.1	0.1	0.05	0.05

The coordinates in the map of the parent locations will be selected randomly from the area map taking under two conditions:

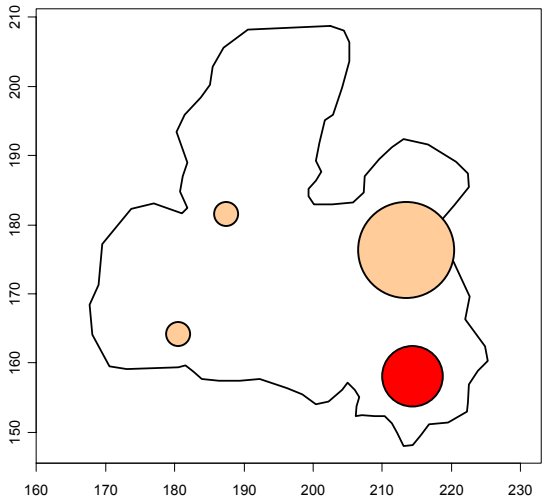
1. The distance between 2 parent locations has to be more than 2 km.

2. The distance from a household has to be less than 2 km.

B. Graphical example of Kulldorff's scan statistic results



Real location of the clusters and radius of influence



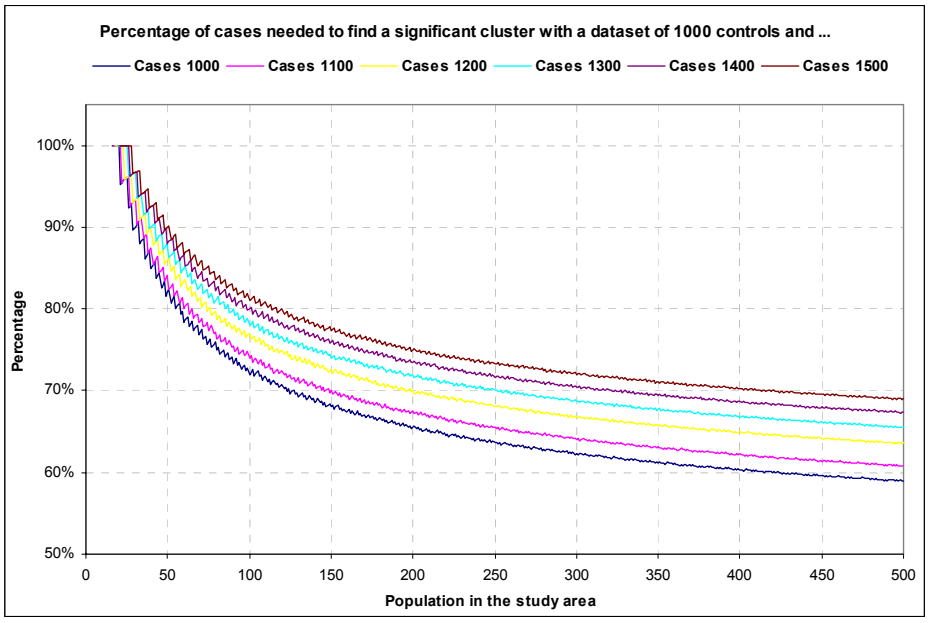
SaTScan output. In red significant cluster and in orange non significant secondary clusters

C. Kulldorff Scan Statistic difficulties finding a significant cluster in small-populated areas

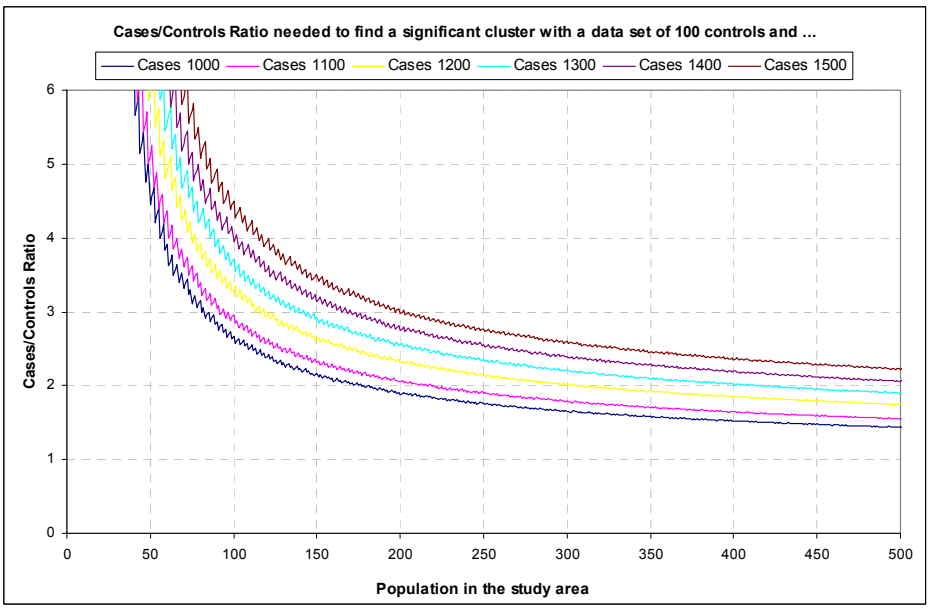
Supposing that we have 1000 controls in our data, then to find a significant cluster in a certain area, the population number (controls and cases) in this area should beat least:

Total Cases in the data	Minimum population (Controls + Cases) needed
1000 to 1100	16
1100 to 1200	17
1200 to 1275	18
1275 to 1375	19
1375 to 1475	20
1475 to 1500	21

Percentage of cases needed to find a significant cluster



Cases/Controls Ratio needed to find a significant cluster



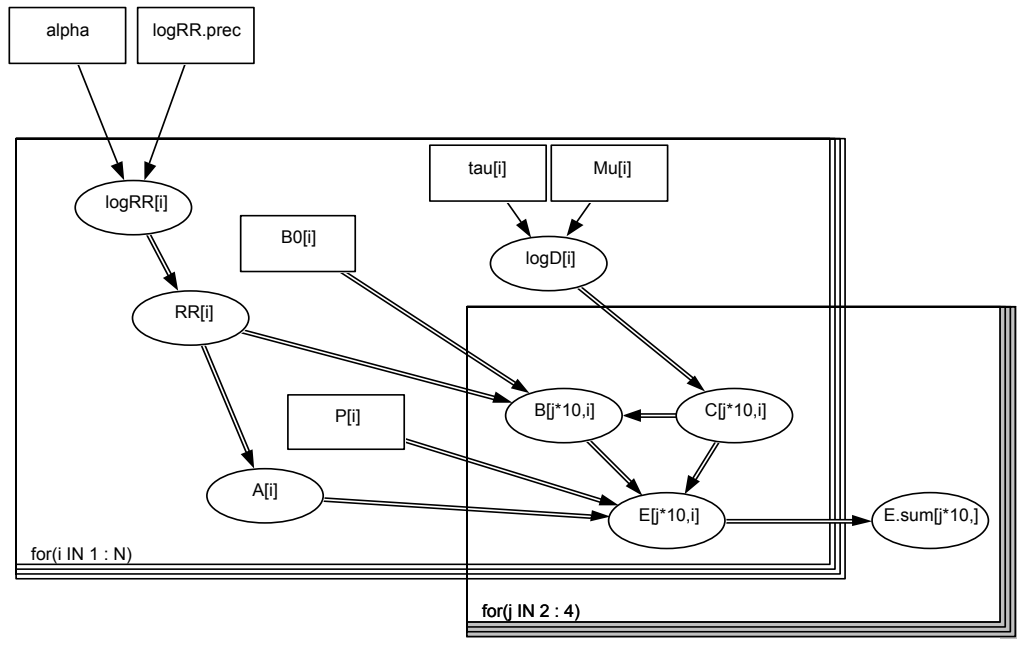
Appendix 2. Italy

Details of the Bayesian models and model fitting.

A *doodle,* i.e. a simplified graphical interface (without considering desegregations by sex and one-year age classes) of the model, has been represented in Figure 30. Ellipses represent variables that are given a distribution. They can be observed value (data) or unobserved ones (parameters); rectangles denote constant values, fixed by the design of the study, which must be specified in a file. The big rectangular boxes (called *plates)* represent cycle of operations to be executed for different set of variables (e.g. for different counterfactual exposure, sex, age-classes, towns). Links between variables can be of two types: a hollow arrow represents mathematical relations between two variables while a solid one indicates stochastic dependence. For instance, E=A*B*C*P is a mathematical (or logical) relation while logRR ~ (alpha, 1/(logRR.prec)^2) is a stochastical one.

A WinBugs file is formed by different parts that can be summed up in three sections: the *model,* containing the model code and the prior distributions, the *data,* containing the data file and its structure and the *inits*, containing the initial values.

Figure 30. The WinBugs model utilized to estimate effects of air pollution.



The code relative to the complete model, more complete than the one illustrated in the graphical interface above is as follows:

MODEL

{

To define the precision of the RR distribution:
logRR.prec <- 1/(0.0084*0.0084)

To initialize the cycle for every single city "i" (in our case N=8), for every age "k" from 30 to 95+, for every baseline "j"*10 (20,30 and 40):

```
for( k in 1 : Age ) {
for( j in 2 : Base )    {
for( i in 1 : Towns )    {

      Em[j * 10 , k , i] <- (((A[i] * Bm[10 * j , k , i]) * C[10 * j , k , i]) / 10) * Pm[k , i]
      Ef[j * 10 , k , i] <- (((A[i] * Bf[10 * j , k , i]) * C[10 * j , k , i]) / 10) * Pf[k , i]
      Bm[j * 10 , k , i] <- B0m[k , i] / (1 + (RR[i] - 1) * (C[j * 10 , k , i] / 10))
      Bf[j * 10 , k , i] <- B0f[k , i] / (1 + (RR[i] - 1) * (C[j * 10 , k , i] / 10))
      C[j * 10 ,k, i] <- exp(logD[i])-10*j
                  }
                }
              }
```

Where Em and Ef are the number of attributable deaths for males and females, disaggregated by sex, age and town.

To introduce the Relative Risk variability:

```
for( i in 1 : Towns ) {
      A[i] <- (RR[i] - 1) / RR[i]
      RR[i] <- exp(logRR[i])
      logRR[i] ~ dnorm(0.0257,logRR.prec)
```

To define the distribution shape of every single city:

```
      logD[i] ~ dnorm(Mu[i],tau[i])
      tau[i]<-1/var[i]
                }
```

To define joint variables for different baseline, for the eight cities and for each sex and for all the variables together:

```
for( j in 2 : Base ) {
      Ef.sum[j * 10] <- sum(Ef[j * 10 ,  , ])
      Em.sum[j * 10] <- sum(Em[j * 10 ,  , ])
      Esum[j] <- Em.sum[j * 10] + Ef.sum[j * 10]
            }
}
```

DATA

Data section for 8 cities, 3 baselines, 66 age classes, with: 4 matrices of dimension 66 (ages) * 8 (cities) filled with population and mortality data, 2 for each sex, 8 element vectors for average and standard deviations of the two year PM10 distribution.

list(Towns=8,Base=4,Age=66, Pm = structure(.Data =c(…),.Dim = c(66,8)),Pf = structure(.Data = c(…),.Dim = c(66,8)), B0m=structure(.Data=c(…),.Dim = c(66,8)), B0f=structure(.Data=c(…),.Dim = c(66,8)),

```
Mu=c(3.897344,3.798434,3.712465,3.681114,3.671367,3.82009,3.873678,3.715717
),
var=c(0.255965,0.124287,0.290564,0.558964,0.3126,0.25743,0.510862,0.12924))
```

INITS

Two sets of widely different initial values for stochastic "nodes" connected to the two sources of variability: the relative risk RR and the pollutants concentration. They are used to initiate two chains of 10000 simulated values.

```
list(logD = c(1,1,1,1,1,1,1,1),logRR =
c(0.001,0.001,0.001,0.001,0.001,0.001,0.001,0.001))
list(logD = c(2,2,2,2,2,2,2,2),logRR =
c(0.003,0.003,0.003,0.003,0.003,0.003,0.003,0.003))
```

Where:
N represents the number of the cities, $\mathbf{Bo_m}$ **and** $\mathbf{Bo_f}$ the above mentioned values (66*8*2 values), $\mathbf{P_m}$ and $\mathbf{P_f}$ the vector of the exposed populations (66*8*2 values), **Mu** the vector of the 8 two-years average values for each city, synthetic indicators of PM10 value, **Tau** the 8 precision vectors (where precision = 1/variance).

Alpha and **logrr.prec** are constants defined outside the data section and they are equal for all the cities. Initial values can be generated by the model itself or imposed. Two chains of 10 000 values were simulated. For every chain the first 5 000 iterations were not used in the analysis for convergence reasons in order that remaining samples are drawn from a distribution close enough to the stationary distribution. History plots have been monitored for converge assessment and a Gelman-Rubin test (Gelman et al., 1992) has been applied. It is a convergence test based on two or more parallel chains, each started from different initial values which are over-dispersed with respect to the true posterior distribution: it is based on a comparison of the within and between chain variances for each variable (essentially a classical analysis of variance). Convergence is diagnosed when the chains have "forgotten" their initial values, and the output from all chains is indistinguishable.
The steps that have to be completed in WinBugs are:

1. to compile the model (to write the code);
2. to load the data matrix;
3. to fix the number of chains;
4. to load the inits (one for each simulated chain) or to make them generate from the program itself;
5. to fix the number of simulations;
6. to fix the number of simulations that have to be "burned";
7. to run the model;
8. to make the model draw dynamic charts, trends, summary statistics, convergence test charts, density distribution charts and output files for S-Plus.

If the model is not consistent an error message is displayed. Examples of dynamic trace plot (Figure 31), shape density (Kernel) distribution (Figure 32), history plots (Figure 33)*,* sample statistics from the posterior distribution (Figure 34) and convergence test plots (Figure 35) have been reported below. More accurate convergence tests can be made importing data in S-Plus by using the CODA add-in software.

Figure 31. Dynamic trace for a two model chains (model 5). All towns, males, baseline 20.

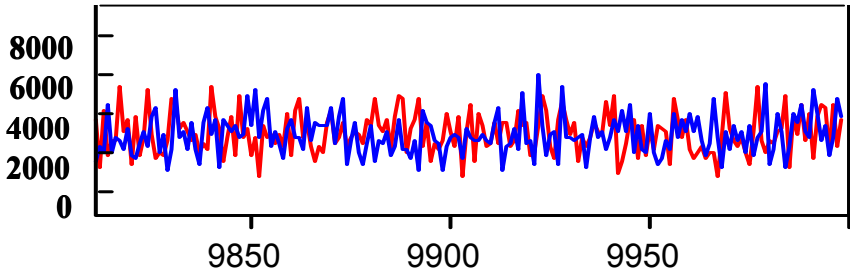


Figure 32. Shape of density (Kernel) distribution (10 000 iterations): cumulative effect for the eight cities, baseline 20 (simplified model).

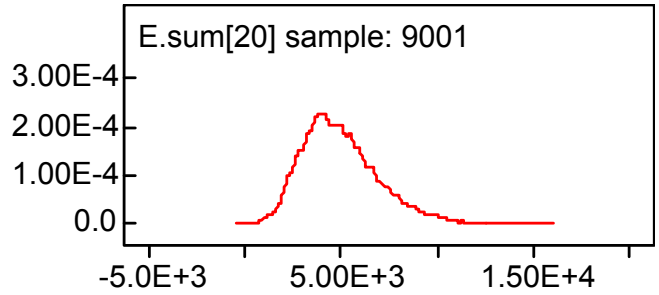


Figure 33. History plot (10 000 iterations): cumulative effect for the eight cities, baseline 40 (simplified model, model 4).

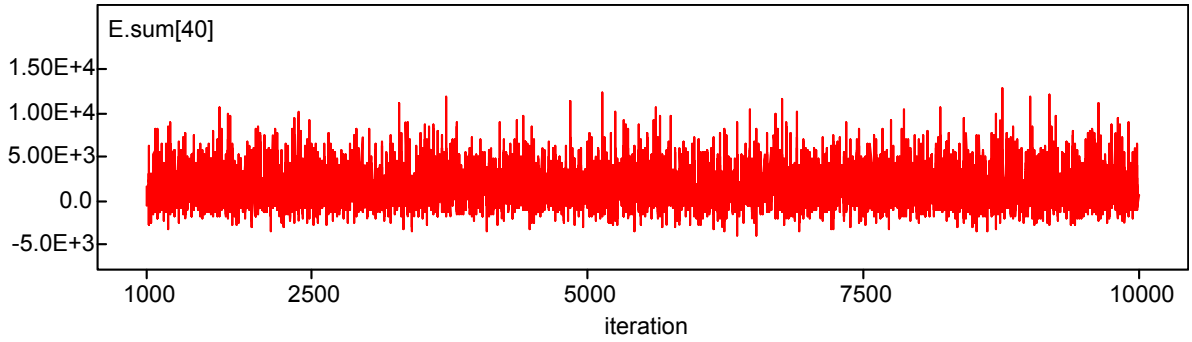
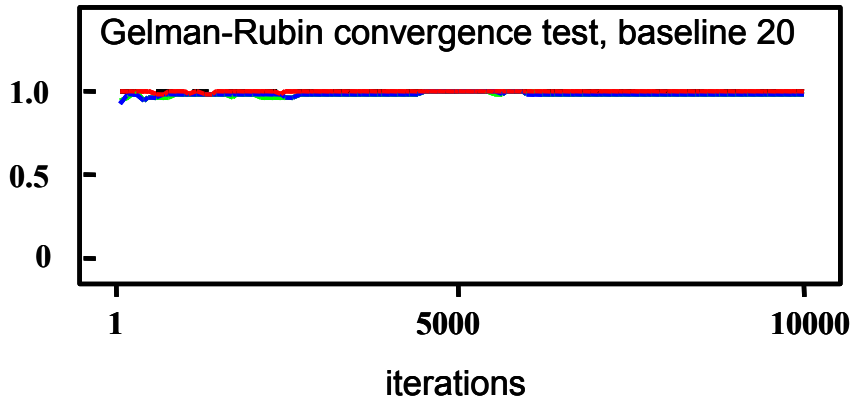


Figure 34. Means and descriptive statistics from the posterior distribution inclusive of Monte Carlo errors. Data for single cities and cumulative data, males and females together (model 4). Baseline 20, 30 and 40 (10 000 iterations).

Node	Mea	SD	MC	2.5%	Media	97.5%	Star	Sampl
E[20,1]	579.	473.8	4.22	-35.98	476.4	1752	1000	9001
E[20,2]	415.	280.9	2.94	27.66	360.3	1101	1000	9001
E[20,3]	627.	621.3	6.51	-153.6	472.2	2194	1000	9001
E[20,4]	322.	403	4.56	-146	212.7	1407	1000	9001
E[20,5]	258.	263.8	2.88	-72.49	194.6	955.3	1000	9001
E[20,6]	1790	1567	17.2	-201.8	1455	5779	1000	9001
E[20,7]	712.	750.9	7.75	-175.8	515.4	2715	1000	9001
E[20,8]	311.	227.9	2.88	1.449	268.3	871.2	1000	9001
E[30,1]	418.	482.6	4.26	-233.2	315	1607	1000	9001
E[30,2]	265.	278.5	2.97	-130.9	211.5	950.7	1000	9001
E[30,3]	385.	635.7	6.58	-469.6	237.8	1985	1000	9001
E[30,4]	211	418.2	4.79	-312.6	100.8	1328	1000	9001
E[30,5]	156.	270.3	2.93	-210.9	94.69	860.1	1000	9001
E[30,6]	1225	1598	17.6	-947.7	889.1	5341	1000	9001
E[30,7]	530.	776.2	8.03	-440	328.7	2599	1000	9001
E[30,8]	182.	225.8	2.88	-144.4	140	739.4	1000	9001
E[40,1]	248.	498.3	4.37	-475.4	151.1	1460	1000	9001
E[40,2]	107.	284.4	3.08	-340.3	61.15	802.6	1000	9001
E[40,3]	130.	662.2	6.78	-844	3.574	1767	1000	9001
E[40,4]	92.6	438.2	5.08	-504.7	-7.92	1248	1000	9001
E[40,5]	48.2	281.8	3.04	-371.9	-2.967	763.3	1000	9001
E[40,6]	629.	1655	18.4	-1812	307.4	4833	1000	9001
E[40,7]	337.	808.9	8.39	-741.3	141	2481	1000	9001
E[40,8]	47.1	231.5	2.95	-330.8	14.05	608.7	1000	9001
E.sum[2	**5018**	2000	20.7	**1909**	**4752**	**9708**	1000	9001
E.sum[3	**3376**	2046	21.2	**161.7**	**3107**	**8198**	1000	9001
E.sum[4	**1642**	2124	22.1	**-1800**	**1387**	**6552**	1000	9001

Figure 5, Legend: 1 = Turin, 2 = Genova, 3 = Milan, 4 = Bologna, 5 = Florence, 6 = Rome, 7 = Naples and 8 = Palermo; 20,30 and 40 are the baselines.

Figure 35. Gelman-Rubin convergence test plot, males and females together. attributable deaths, baseline 20



References

Gelman et al. (1992), Inference from iterative simulations using multiple sequences, Statistical Science 7, 457-511.

Hurley et al. (2000), Towards assessing and costing the health impacts of ambient particulate air pollution in the UK.

Künzly et al. (1999), Health costs due to road traffic-related air pollution. An impact assessment project of Austria, France and Switzerland: air pollution attributable cases. Prepared for the Ministerial Conference for Environ Health 1999.

Martuzzi et al. (2002), Health Impact Assessment of air pollution in the eight major Italian cities, World Health Organization - European Centre for Environment and Health, Rome.

Annex 1. RIF documentation

Rapid Inquiry Facility: Technical document (on enclosed CD).

Annex 2. EUROHEIS/SAHSU 2003 Conference programme

Most presentations given at the conference are also included in the attached CD.

Annex 3. EUROHEIS/SAHSU 2003 Conference proceedings

Extended abstracts of most papers are included in the proceedings.

Annex 4. Assessors' reports

Reports submitted by the three independent assessors.

Partners (back cover)

SAHSU, Dept. Epidemiology & Public Health, Imperial College Coordinator	UK
National Board of Health	Denmark
National Public Health Institute	Finland
Trinity College, University of Dublin	Ireland
WHO European Centre for Environment & Health, Rome	Italy
National Institute of Public Health and the Environment	The Netherlands
Dept. Statistics & Operation Research, University of Valencia	Spain
Dept. Epidemiology, Stockholm Centre of Public Health	Sweden

The Small Area Health Statistics Unit is funded by a grant from the Department of Health, Department of the Environment, Food and Rural Affairs, Environment Agency, Health and Safety Executive, Scottish Executive, Welsh Assembly Government and Northern Ireland Department of Health, Social Service and Public Safety. The views expressed in this publication are those of the authors and not necessarily those of the funding departments.