# SPATIAL SMALL AREA ANALYSES OF DISEASE RISK AROUND SOURCES OF ENVIRONMENTAL POLLUTION: MODELLING TOOLS FOR A SYSTEM USING HIGH RESOLUTION REGISTER DATA

ESA KOKKI

To be presented, with the permission of the Faculty of Mathematics and Science of the University of Jyväskylä, for public criticism in Auditorium S212, on May 15th, 2004, at 12 o'clock noon.

*To Anu, Oskari and Sakari*

# Abstract

The overall aim of this study was to further develop the existing small area system in estimation of regional variation of disease risk around a putative source of environmental pollution. The more specific aims were i) to search advanced statistical methods, suitable for the system; ii) to implement the methods as a part of the system; iii) to evaluate the use of register data; and iv) to test and validate epidemiological methods within the system.

The data stored in the system are aggregated into 500 m × 500 m squares. High spatial resolution is informative when local variation is studied but brings up methodological challenges due to the autocorrelation and the sparseness of data (low population counts and low number of cases). The development of methodology was illustrated with the case studies on relative risk of cancer in a rural municipality with the problems in drinking water, and in a vicinity of a former waste site, a former asbestos mine, and a polluted river.

In a simple case, the relative risk of cancer was estimated with the Poisson regression model with the regional effects. As a more sophisticated model, the hierarchical Markov model was exploited. The hierarchical Markov model as such turned out to be inconsistent with sparse high resolution data. Three constraining methods to improve the behaviour of the model were suggested. Due to the demanding calculation of the hierarchical Markov model, an extension of the Poisson regression model based on the change-point model was exploited as another simpler method.

The effect of adjustment for the socioeconomic status and the choice of the reference area as confounding factors were investigated. Also the different standardizing methods were compared in the estimation of the relative risk.

The system with high resolution data is useful in rapid estimation of relative risk around a putative source of pollution. The strength of the system is that the study area can be defined freely with an accuracy of 500 metres. However, if data tends to be sparse, the classical methods, like standardized incidence ratios (SIR) or Poisson regression models, may give uncertain estimates. This can be overcome by using smoothing methods. Secondly, if data is doubted to be spatially autocorrelated, the classical methods may give incorrect results, and some more sophisticated methods are needed. In this thesis some useful statistical methods were suggested to solve the problems mentioned.

# Acknowledgements

This study was carried out at the Unit of Environmental Epidemiology, Department of Environmental Health, National Public Health Institute during the years 1998–2004. The computer facilities of the Laboratory of Data Analysis, Department of Mathematics and Statistics, University of Jyväskylä, have also been used during this work. I am thankful for the facilities available for this work. I would like to thank the National Land Survey of Finland, the Population Register Centre, the Statistics Finland and the Finnish Cancer Registry for providing data for the SMASH software. Many people have had an important role in developing my knowledge, skills and ideas of studying small area statistics on health. In particular, I wish to thank:

Professor Antti Penttinen from the University of Jyväskylä and Research Professor Juha Pekkanen from the National Public Health Institute for their tireless supportive supervision and for providing me an opportunity to work with this interesting multidisciplinary project;

Reader Lars Järup from the Imperial College London, United Kingdom (Associate Professor from Karolinska Institutet, Stockholm, Sweden) and Professor Esa Läärä from the University of Oulu, the official reviewers of this thesis for their constructive and critical comments that helped me to improve my manuscript;

Mrs. Tuula Blåfield for the revision of the language of this thesis;

Docent Eero Pukkala from the Finnish Cancer Registry for excellent cooperation in the development of the SMASH system;

Mr. Esa Kauniskangas from the Geological Survey of Finland for coding the Avenue scripts for the SMASH software;

Mr. Jari Väisänen and Ph.Lic. Anne Korhonen for their assistance in increasing my skills on coding language C;

Dr. Jukka Ranta for all his guidance which increased my understanding of Bayesian modelling;

Docent Pia Verkasalo for all her guidance which increased my understanding of cancer epidemiology, and for her important work in the development of the SMASH system;

Research Professor Terttu Vartiainen and Mr. Hannu Kiviranta for coauthoring;

All the personnel of the Department of Environmental Health for keeping up an encouraging working atmosphere;

All the personnel of the Unit of Environmental Epidemiology, and, especially, Mrs. Anna Maria Helppi for assisting in practical issues;

Mr. Pekka Tiittanen for his never-ending preparedness to assist in statistical issues;

All the players in our floorball team, without you I could not serve the goals;

All my friends, with whom I have been able to share few moments in deep water and countless joyful and relaxing moments;

My mother Mirja, my siblings Pentti, Pauli, Helvi, Hannu and Seija, and my father-in-law Hannes and mother-in-law Taimi for their encouragement and concrete help during all these years;

And, finally, my wife Anu for all her patience and love, and my sons Oskari and Sakari for giving me such love, happiness and joy.

This study was financially supported by ”The Finnish Research Programme on Environmental Health” -project (n:o 42559) of the Academy of Finland, ”A European Environment and Health Information System for Exposure and Disease Mapping and Risk Assessment” -project (n:os VS/1999/5279 (99CVF2-606), SI2.291820 (2000CVG2-605), SI2.329122 (2001CVG2-604)) of the European Union, and the grants of the Ellen and Artturi Nyyssönen foundation and the Paavo Koistinen foundation.

Kuopio, March 2004

*Esa Kokki*

## List of original publications

This thesis is based on the following five articles:

[I] Kokki E, Pukkala E, Verkasalo PK, Pekkanen J. Small Area Statistics on Health (SMASH): A System for Rapid Investigations of Cancer in Finland. In Briggs DJ, Forer P, Järup L, Stern R (Eds.). *GIS for Emergency Preparedness and Health Risk Reduction.* Kluwer Academic Publishers, Dordrecht, 2002; pp. 255–266.

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

[III] Kokki E. Constrained Bayesian modelling of disease risk around a point source. Publications of the Laboratory of Data Analysis, No. 5, University of Jyväskylä, 2003.

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

[V] Verkasalo PK, Kokki E, Pukkala E, Vartiainen T, Kiviranta H, Penttinen A, Pekkanen J. Cancer risk near a polluted river in Finland. Submitted, 2004.

In what follows, these articles will be referred to according to their Roman numerals. The original articles in this thesis have been reproduced with the permission of the publishers.

# Contents

# 1 Introduction

Many environmental exposures originate from point sources like industrial plants, dumping areas, or oil refineries. Such sites are often recognizable and may cause concern or even fears [1]. Only a suspicion of an increased risk of cancer in a region can cause remarkable psychological and economical consequences [2]. In order to respond to these concerns, rapid methods are needed to produce an initial estimate of a potentially increased risk of disease around a known point source. Finnish Cancer Registry and National Public Health Institute have been developing a small area system called Small Area Statistics on Health (SMASH). The main use of SMASH is in comparison of cancer incidence in an arbitrary selected area of interest in Finland with the incidence in a given reference area. In the present application oriented work, the statistical methodology of SMASH is considered.

Spatial epidemiology means the investigation (describing and understanding) of the geographical variation in disease risk, especially with respect to variations in environmental exposures at the small area level. Elliott *et al.* [3] distinguishes four types of study in spatial epidemiology: 1) disease mapping; 2) geographical correlation studies; 3) assessment of risk in relation to a point or line source; 4) cluster detection and disease clustering. Same methods can be applied in many of the types of study listed above, although the phrasing of a question is quite different. *Disease mapping* is usually used for descriptive purposes. Crude or adjusted morbidity or mortality rates are mapped for summarizing the spatial variation in disease risk [4]. In *geographical correlation studies* spatial variation in some external factors is related to health outcomes. For example, environmental exposure maps or ecological variation in lifestyle factors can be used in studies focusing on aetiological questions [5],[6]. In *point or line source studies* the source is considered to present a potential environmental hazard, and an increased risk is suspected in the vicinity of the source. When well-defined hypotheses are set, the interpretation of results may be more straightforward than in the case of no *a priori* hypothesis, for example, if the reason for a study is the worries of local population [7],[8]. *Cluster detection* is usually used for monitoring purposes. In this case, no hypotheses are set in advance, so the interpretation of the results meets difficulties. Cluster detection may be used for the detection of the raised incidence of disease, or for descriptive purposes, i.e., giving clues to aetiology [9],[10].

The focus of this work is on type 3 problems, because the aim of SMASH is to assist in the study of the possible increase or decrease in cancer incidence in a pre-specified study area. SMASH is neither a monitoring system producing disease atlases by methods of disease mapping nor "fishing hot spots" being areas with increased cancer incidences. *The aim of the system is to transform register data into epidemiological information for decision-making.*

A system similar to SMASH has been developed in the UK. Small Area Health Statistics Unit [11], Imperial College London, provides a system called Rapid Inquiry Facility (RIF) as a national facility for small area disease mapping and the rapid initial assessment of apparent disease clusters around a point source [12]. Along the EUROHEIS project [13], RIF has been implemented also in Sweden, in Spain, and in the Netherlands. There are geographical information systems for public health analyses also in the US, but they are not nationwide [14],[15].

A recent tendency is to apply the Bayesian framework in the analysis of spatial health data [3],[4]. Two typical assumptions of spatial analysis of non-infectious diseases, conditionally independent prevalence and a possibility of non-perceptible spatial correlation, naturally lead to fully Bayesian (see, e.g., [16]–[18]) or empirical Bayesian methods (see, e.g., [19],[20]). The Bayesian framework gives possibilities to solve many problems faced in spatial modelling. For example, the Bayesian methods can take into account spatial patterns in a disease. Moreover, they allow the calculation of exact probabilities to be used in the comparison of disease risks between several areas.

Register data are of high quality in Finland [21]. Through the unique personal identifiers, the linking of information from several registers is straightforward and technically easy. Examples of high quality health data are registers on cancers, on the cause of death and on birth. Population data including, among others, data on residence, on education and on economical activity are also of high quality. Data on residence is available in metric coordinates in Finland. The centre point of building as the place of residence can be georeferenced with an accuracy of 1 metre at its best. The high resolution of residence data is valuable in the exposure assessment but, at the same time, it brings up such methodological challenges as sparseness and spatial autocorrelation in the risk estimation.

This study concentrates on improving the usefulness of the small area system SMASH. The main focus will be on the development of statistical methods for analyzing high resolution small area data around given sources of pollution.

# 2 Statistical and epidemiological issues in point source studies

Several statistical methods and tests for the analysis of disease incidence in the vicinity of a point source of possible environmental risk can be found in the literature, but they are usually designed for individual level data, see, e.g., [7],[8],[22]–[35]. In the following, the emphasis is on reviewing statistical issues in the point source studies on cancer based on area level data. Later, in Section 5, the methodology is reviewed in more detail in connection with the development of modelling tools for SMASH. In the following subsections, the case studies during the recent decade and some epidemiological issues on ecological studies will be discussed.

## 2.1 Small area studies on cancer in the vicinity of the source of pollution

The observation of increased incidence of childhood leukaemia in the vicinity of a British nuclear installation in 1984 [36] triggered several similar and other point source studies worldwide. In Table 1, a sample of case studies of cancer incidence or mortality on an ecological level in the recent decade are listed. Finally, in Section 2.1.4, these case studies are discussed critically.

### 2.1.1 Childhood leukaemia and lymphoma around nuclear sites

Nuclear power plants in Germany [37], in England and Wales [40], in Scotland [45],[49], and in Spain [48] and one nuclear waste reprocessing plant in France [43] have been objects of investigation during the recent decade. These studies have mainly been focused on childhood (children under age of 15 years) leukaemia and lymphomas.

Michaelis *et al.* [37] studied incidence of leukaemia and lymphoma but also cancer of all sites. The study areas were defined around all the nuclear power plants in former West-Germany. They were comprised of communities with at least one third of them within 5 km, 10 km and 15 km from a source. As an overall relative risk estimator pooled standardized incidence ratios (SIR) were calculated. SIR is calculated as a ratio of the observed and the expected number of cases, for more information see Formula (2.3) in Section 2.2.5. Here the value SIR=1.00 means the standard, values larger than 1.00 indicate an

Table 1: A sample from small area studies on cancer in the vicinity of the source of pollution in 1992–2003.

| Reference | Source of pollution | Small areas (ED = enumeration district with ca 400 people) | Cancers studied, age (if not in all ages), study period | Methods used |
|---|---|---|---|---|
| Michaelis *et al.*, 1992 [37] | 18 nuclear power plants in former West-Germany | communities within 15 km | total cancer, leukaemia, lymphoma in children, 1980–90 | pooled SIR, RR in study vs. control area |
| Selvin *et al.*, 1992 [38] | large microwave tower in USA | all census tracts in San Francisco | leukaemia, brain, lymphomas in $<$ age of 21, 1973–88 | RR in $\leq$ 3.5 km vs. others |
| Bhopal *et al.*, 1994 [39] | coking works in England | all EDs in South Tyneside | lung, others than lung, respiratory, circulatory, 1981–89 | SMR |
| Bithell *et al.*, 1994 [40] | 23 nuclear sites in England & Wales | electoral wards within 25 km | leukaemia, lymphoma in children, 1966–87 | SIR, Stone's tests [22],[24] linear risk score test [7] |
| Lyons *et al.*, 1995 [41] | petrochemical plant in Baglan Bay, Wales | EDs within 3 km | leukaemia, lymphoma in $<$ age of 25, 1974–91 | SIR |
| Sans *et al.*, 1995 [42] | petrochemical plant in Baglan Bay, Wales | EDs within 7.5 km | total, larynx, leukaemia, 1974–84 | SIR, Stone's tests |
| Viel *et al.*, 1995 [43] | nuclear waste reprocessing plant in La Hague, France | cantons within 35 km | leukaemia in $<$ age of 25, 1978–92 | SIR, SMR, Stone's test, mapping with kernel regression smoothing [28] |
| Elliott *et al.*, 1996 [44] | 72 municipal solid waste incinerators in Great Britain | EDs within 7.5 km | total, lung, stomach, colorectal, liver, and 9 other subtypes, 1974–87 | SIR, Stone's tests |
| Sharp *et al.*, 1996 [45] | 7 nuclear sites in Scotland | EDs within 25 km | leukaemia, lymphoma in children, 1968–93 | SIR, Stone's tests, linear risk score test |
| Michelozzi *et al.*, 1998 [46] | waste disposal site, waste incinerator, and oil refinery in Rome, Italy | census tracts within 10 km | total, liver, larynx, lung, kidney, lymphatic and haematopoietic 1987–93 | SMR, Stone's test |
| Harrison *et al.*, 1999 [47] | main roads and petrol stations in UK West Midlands | EDs within 100 m | leukaemia in children, 1990–94 | SIR |
| López-Aberte *et al.*, 1999 [48] | 7 nuclear power plants, 5 nuclear fuel facilities in Spain | towns within 30 km | leukaemia, lymphoma, myeloma, 1975–93 | SMR, RR in towns vs. matched control towns |
| Sharp *et al.*, 1999 [49] | 7 nuclear sites in Scotland | EDs within 25 km | non-haematopoietic in children, 1975–94 | SIR, Stone's test |
| Dickinson *et al.*, 2003 [50] | railways in England & Wales | electoral wards within 20 km | leukaemia, lymphoma in children, 1966–87 | RR in 3 groups of exposure |
| Reynolds *et al.*, 2003 [51] | point sources of air pollutants in USA | all census tracts in California | total, leukaemia in children, 1988–94 | RR in 4 groups of exposure |

increased risk whereas values smaller than 1.00 stand for a decrease in the relative risk. For total cancer, SIR within 5 km from a source was 0.99, for leukaemia 1.10, and for lymphoma 0.78. Also a matched control area for every exposed area was defined, and relative risks (RR) in these pairs were calculated. For total cancer, the relative risk (with 95% confidence interval) was 1.13 (0.81–1.60), for leukaemia 1.44 (0.81–2.79), and for lymphoma 1.67 (0.33–18.48).

Bithell *et al.* [40] studied the incidence of leukaemia and non-Hodgkin's lymphoma combined in the vicinity of 23 nuclear sites in England and Wales. Six possible prospective sites were used as controls. Study areas were defined to include electoral wards with the population centre within 25 km from a site. The relative risk estimators varied from 0.77 to 1.30 in the study areas and from 0.85 to 1.18 in the control areas. In the assessment of the existence of the decreased risk of cancer they developed a linear score test and compared it with Stone's tests, the maximum likelihood ratio test [22] and the Poisson maximum test [24]. Decreased risk was observed for two study sites and one control site.

Viel *et al.* [43] studied the incidence of leukaemia among young people ($<$ age of 25) around a nuclear waste reprocessing plant in France. The study area was divided into three parts by the distance from the source (<10 km, 10–19 km, 20–35 km). All the "cantons" with at least 50% of area inside the boundaries were included. The average population count in cantons was 6 000. As an estimate of the relative risk, SIRs were calculated. Within 10 km SIR was 2.8 (0.8–7.2). The possible increase in risk was tested by means of the Poisson maximum test [24], resulting in minor increase in risk. The relative risk was mapped with kernel regression smoothing [28].

In two separate investigations, Sharp *et al.* studied the combined incidence of childhood leukaemia and non-Hodgkin's lymphoma [45], and non-haematopoietic cancers [49] near the nuclear sites in Scotland. All the enumeration districts (ED) within 25 km from a site were included in the investigation. The number of enumeration districts around the sources varied from 54 to 3 064, while corresponding population counts were from 3 000 to 180 000. The study area around each source was divided into 25 bands and the sites were studied separately. In the first study [45], six tests were evaluated: the maximum likelihood ratio test [22], the linear risk score test based on rank and distance [7],[40], the Poisson maximum test [24], a variance of the Poisson maximum test based on the minimum $p$-value [26], and SIR. SIR for leukaemia and non-Hodgkin's lymphoma combined varied from 0.84 (0.61–1.14) to 1.99 (0.91–3.77) around the sources. In the second study [49], the maximum likelihood test [22] and SIR were applied. Age, sex, deprivation categories, the 5-level urban-rural residence indicator were considered in the calculation of expected numbers. Stone's test for central nervous system tumours around one site found an increased relative risk estimate. For central nervous system tumours SIR varied from 0.47 (0.09–1.38) to 1.22 (1.02–1.44) around the sources. For other malignant neoplasms SIR varied from 0.32 (0.04–1.17) to 1.51 (0.83–2.54).

In a Spanish study [48], people of all ages were included in the study population, and the mortality of leukaemia, lymphomas and myeloma in the vicinity of nuclear power plants and nuclear fuel facilities were investigated. 489 towns within 15 km and 30 km from a source were applied as study areas. Further 477 control towns matched by population size and socioeconomic status were defined. For confidentiality, population smaller than 5 000 had to be estimated. As an estimate of risk the standardized mortality ratios (SMR) and risk ratios between study towns and control towns were calculated. The changes in risk were studied by comparing the position before and after the start-up date of the sites. SMR within 15 km from any nuclear power plant varied from 0.69 (Hodgkin's disease) to 1.47 (leukaemia, under age of 25 years). The risk ratio with 95% confidence interval varied from 0.92 (0.45–1.87) (non-Hodgkin's lymphoma) to 1.62 (0.73–3.58) (myeloma).

### 2.1.2 Waste incinerator plants

In Great Britain, all public incinerators for burning waste were studied by the Small Area Health Statistics Unit [44]. Recent incinerators (operation started after 1975) were excluded, and overlapping sites were considered as a multi-site group. At the first stage, 20 randomly selected incinerators were combined. The incidence of total cancer, and thirteen sub-types in the vicinity of the incinerators was investigated. At the second stage, 52 other incinerators were combined and the incidence on cancers having relative risk significantly higher than the one in the first stage was studied. SIRs were adjusted for the socioeconomic status. For estimating the residual confounding unrelated to the incineration, a pre-incinerator period was analyzed. For the decline in risk at some distance from the source, unconditional and conditional likelihood tests [24],[40] were performed in 8 bands. To allow multiple testing, the level of significance was set to 0.0025. SIRs varied from 0.83 (0.63–1.09) (nasal cancer) to 1.29 (1.10–1.51) (liver cancer). Stone's tests found significantly increased relative risks for seven out of thirteen cancers.

In Rome, Italy, a waste incinerator plant combined with a waste disposal site and an oil refinery was an object of investigation [46]. Mortality of all causes, total cancer and nine sub-types were analyzed separately for males and females. Small areas used in studies were census tracts with 480 inhabitants on an average. SMRs adjusted for age and socioeconomic status were calculated in areas with distance 0–3 km, 3–8 km and 8–10 km from the source. Also Stone's test in 9 bands with 1 km interval was performed [22]. SMRs for men varied from 0.82 (0.03–40.9) (leukaemia) to 2.76 (0.31–93.4) (kidney cancer), and for women from 0.55 (0.02–27.1) (lung cancer) to 1.37 (0.05–67.9) (leukaemia). Out of eighteen Stone's tests, an increased risk was found only for larynx cancer among males.

### 2.1.3 Other sources of pollution

*A large microwave tower* in San Francisco, USA, was considered as a point source [38]. Incidences of leukaemia, brain cancer and lymphatic cancers were studied among people less than 21 years of age. The census tracts with population count 342 on an average were divided into two groups by the distance of 3.5 km from the source. The risk ratio for leukaemia was 0.73, for brain cancer 1.16, and for Hodgkin's disease and non-Hodgkin's lymphoma combined 1.17, none of them differing significantly from the expected value.

Bhopal *et al.* [39] studied excess ill health in people living near *a coking works*. The site was built in 1937 and enlarged in 1980. Between 1940 and 1960 housing estates were built near the site. Among many health indicators, mortality from lung cancer and other cancers, the respiratory system combined and circulatory diseases were analyzed. A total of 349 enumeration districts, with 456 inhabitants on an average, were categorized to classes of high and low exposure areas and the control area. Two separate categorizings were used, one based on perceived exposure and the other based on modelled exposure. Multiple comparisons were considered in the determination of the significance level. SMRs varied from 0.88 (other than lung cancer, both sexes) to 1.23 (lung cancer, females), the first-mentioned differing significantly from the expected value.

In 1995, *a petrochemical plant* in Baglan Bay, Wales, was an object of two separate studies published consecutively in the same journal. Lyons *et al.* [41] repeated an earlier investigation produced by a TV company. The SIRs of leukaemia and lymphomas combined among people less than 25 years of age within circles with radius of 1.5 km and 3 km from the source were estimated as 1.37 and 1.22, respectively. In the Small Area Health Statistics Unit, Sans *et al.* [42] performed more detailed analyses studying people of all ages.

The incidence of total cancer, leukaemia and larynx cancer was estimated by SIRs. The study area within 3 km and 7.5 km from the source was divided into 8 bands, and 75 separate Stone's tests were performed. Within 3 km the SIR for total cancer was 1.07 (1.01–1.14), for leukaemia 0.99 (0.66–1.51), and for larynx cancer 1.44 (0.91–2.27). Stone's test produced significant results only for total cancer.

Harrison *et al.* [47] studied the incidence of leukaemia in children living near a main road or a petrol station. The "main road" was defined by having 23 400 vehicles a day on average. SIRs were calculated in all the enumeration districts within 100 m from the main road (1.48 (0.65–2.93)) and from the petrol station (1.16 (0.74–1.72)) separately and from both (0.81 (0.16–2.38)).

Dickinson *et al.* [50] studied the national rail network in England and Wales. The relative risk of leukaemia and non-Hodgkin's lymphoma among children was estimated. The electoral wards with population-weighted centroid within 20 km from the railways were categorized in three groups separately by a railway proximity function and by a railway density function. The median size of the wards was 6 km$^2$ with at least 100 children. The rate ratios for the risk were estimates by the Poisson regression, unadjusted and adjusted for population mixing, stratified by the Townsend deprivation category. In the highest category of railway proximity, the rate ratio was 1.04 (0.97–1.01) compared to the lowest group, and in the category of railway density 1.05 (0.99–1.13).

Reynolds *et al.* [51] studied all the point sources emitting hazardous air pollutants (mobile sources, area sources and point sources separately and combined) in California, USA. Children under 15 years of age formed a study population, and the incidence of total cancer and leukaemias was investigated. Hazardous air pollutant exposure scores were defined by combining cancer potency factors with modelled outdoor hazardous air pollutant concentrations. The census tracts were categorized by exposure: low exposure (below 25 percentile), medium exposure (25–74 percentiles), high exposure (75–89 percentiles), and very high exposure (above 89 percentile). The risk ratios adjusted for age, sex and race/ethnicity were estimated applying Poisson regression in these four groups. Also a cubic spline curve (see, e.g., [52]) for the rate ratio was fitted. The highest risk ratios were found around point sources in the "very high exposure" group compared to the "low exposure group" 1.13 (1.03–1.23) for total cancer, and 1.32 (1.11–1.57) for leukaemia.

### 2.1.4 A critique

In most of these case studies a possible increase in risk around a source of pollution was studied applying Stone's test or one of its counterparts. A decrease in the risk of disease with an increase in the distance from the source of pollution is tested against the null hypothesis of a uniformly distributed risk. A problem of these tests is that a definition of the study area may affect the result. For example, the selection of too large a study area may support the null hypothesis too frequently.

As an estimate of the relative risk, the standardized incidence ratio or the risk ratio were often calculated over the areas aggregated from the small areas which stand for the spatial unit. The aggregation is problematic if the counts of the cases between the small areas are correlated. The aggregation of the dependent areas may lead to a biased estimation of variance, which is discussed later in Section 2.3.1 in more detail. This is called the problem of the "change of support".

In most of the case studies the study population contained only children, while adults may also be exposed. Therefore an investigation of the effect of the point source on cancers may not use all the possible information. The rarity of childhood cancer may also cause inferential problems. The public concern may encourage the media to report about these diseases and cause a problem of *post hoc* study (see Section 2.3.4). However, a strength of studying children is that the place of residence is a better surrogate of exposure for children than for adults. One reason for studies on children may be that radiation has been identified to be a main external cause of childhood cancers [53].

A further problem in exposure assessment is the accuracy in the definition of exposed study areas. The definition of the exposed areas suffers from the fact that the small areas used as a spatial unit were based on administrative boundaries. Large administrative units are likely to swamp the local effect. When the study area was defined by drawing a circle for computation convenience, also the rest of the administrative areas (70% at most) around the circle was included in the study (e.g., [37],[40],[41],[43],[50]).

One inferential problem is multiple testing (see Section 2.3.5). In many cases Stone's tests (usually both unconditional and conditional tests) were performed and relative risk estimates for several cancers were calculated, but only in the studies by Bhopal *et al.* [39] and by Elliott *et al.* [44] was the problem taken into account in the determination of the significance level.

Another inferential problem is that in many case studies a ”significant” increase in risk was observed for total cancer but not for sub-types of cancer, mainly due to the increased efficiency of the tests. Total cancer is seldom the interesting disease in point source studies. It is not easy to interpret, and the given source is usually assumed to have an effect on specific sub-types of cancer. For the testing of the possible increased risk in sub-types of cancer, additional information, for example from other cancers or from neighbouring areas, is required to strengthen the inference due to the rarity of these diseases.

In this work, our goal is to obtain a more informative result than the possible existence of a trend in risk around a source of pollution. The idea is to develop models allowing the risk estimation not only in the areas aggregated from the small areas but also in these small areas. Especially, the methods paying attention to the possible spatial correlation are of interest. The accuracy of the determination of exposure areas is not a problem in our approach, because we are not tied to administrative boundaries. Our spatial units are based on metric coordinates and a high-resolution grid.

## 2.2 Epidemiological issues

Small area analyses are valid for descriptive studies such as assessing the (increased or decreased) incidence of cancer in selected areas. Many methods developed for disease mapping can easily be adapted to descriptive investigations. If the data are comprehensive enough, the observed numbers of cases are known accurately enough and the expected numbers can be calculated reliably. On the other hand, if the objective of a small area analysis is in the study of causality, many possible problems may cause major difficulties in the inference. Some of these are considered in the following.

### 2.2.1 Long latency of cancers

Small area studies are usually performed when rapid answers are needed to questions on the possible excess in disease risk near a source of exposure. The most commonly analysed health outcome is cancer, although association with the exposure may be weak. The time period needed to develop cancer after the start of exposure, called the latency period, is probably several years or even decades [54], whereas the exposure may have started only a short while ago. This problem is smaller in childhood cancers or some other endpoints, such as birth outcomes, where the exposure can also be more accurately timed.

### 2.2.2 Exposure assessment

A major weakness of the majority of point source studies is the lack of an adequate exposure measure. Exposure assessment based on geographical location is doubtful, because total exposure has usually more complicated pathways, for example, due to occupational exposure or exposure from traffic or any other environment. However, the place of residence is usually utilized as a measure of an environmental exposure, which is often criticized. For example, in 60% of the 45 papers reviewed by Huang and Batterman [55], estimated exposure based on the place of residence was quantified with monitoring measures.

With some exposures, such as magnetic fields or radiation, the place of residence may be relevant for exposure especially among children who spend most of the time at home [37],[40],[41],[43],[45],[47],[49],[51],[56]–[59]. When considering exposures like air pollution, the exposure assessment is more complicated even in studies on children. Studies on children can be improved if data on the change of the place of residence can be utilized, like in Dickinson *et al.* [50].

For a population of working age, data on the workplace may also be needed for a good exposure assessment. Occupational exposures may be multiple compared to the exposure at home either due to the nature of work or due to the location of the workplace, see, e.g., [60] or [61]. For example, if the workplace is downtown and the dwelling place is located in suburban areas, the differences may be remarkable (see, e.g., the study on $NO_2$ levels in three European cities [62]). Also the exposure may be different in the daytime and in the night-time due to the traffic [63]–[65].

Migration is also a problem in exposure assessment [66]. It is usually very difficult to model the effect of the change in the place of residence to the exposure period on the ecological level, because data on individuals are not available. Good-quality data on residential and occupational history may be needed for calculating the time spent in various places and hereby estimating the exposure.

Exposure assessment is one major problem within SMASH. At present, the assessment is based purely on the distance from a putative source. There are no exposure data nor any facilities to model exposure in the system. There are data neither on the change of the place of residence nor on occupational history in SMASH.

### 2.2.3 Ecological bias

Ecological bias (or fallacy) is a major limitation of ecologic analysis for causal inference, see, for example, [6],[31],[66]–[72]. The ecological bias refers to the situation where the estimated effect of the exposure on the ecological level does not correctly reflect the effect of the exposure on the individual level. In ecological analyses there is no or only limited information on the variability of exposure and covariates on the individual level in the areas to be compared, so this heterogeneity can not be taken into account. This heterogeneity can produce bias in the ecological analyses, as in any other epidemiological study, if there is confounding, selection bias, or misclassification of disease or exposure. As an example, in a study on the effect of the socioeconomic status on the risk of cervical cancer, women with a low standard of living in a well-to-do environment had a high risk of cervical cancer [73].

The most common reason leading to the ecological bias are differences between study areas in the prevalence of other risk factors or effect modifiers, that is confounding or effect modification between areas. This problem is most clear in the studies which compare two study areas (e.g., [38],[48]), where the possible difference in disease rates is attributed to the environmental exposure under study. However, the exposure to other environmental exposures or socioeconomic factors may also be very different between the areas.

In most extreme situations, the ecological bias may reverse the sign of the statistic describing dependence, as reported by Lagarde and Pershagen [74]. In their example of residential radon exposure and the lung cancer incidence, the direction of association in the aggregated data was opposite to the one given by the individual level data.

In SMASH, the possibility of ecological bias still remains. We do not have tools or facilities to handle the problem. This should be taken into account in the interpretation of results.

### 2.2.4 Confounding factors

Confounding can cause difficulties in the interpretation of all epidemiological studies. Confounding occurs when covariate is a risk factor for the disease and correlated with the exposure in the population at risk, but not affected by the exposure or the disease [75]. Typical confounders are sex, age and socioeconomic status. Other potential confounders may be, for example, occupation or lifestyle factors.

Sex and age are available in registers and therefore easy to take into account. Instead, covariates such as life style factors are seldom extensively registered. Variables describing socioeconomic factors are sometimes available. In the UK, socioeconomic classification is usually based on material deprivation reflecting wealth and income. For example, the "Carstairs index" is based on unemployment, lack of car access, percentage of individuals in low social class, and overcrowding [76]. In Finland, descriptions of socioeconomic differences are usually based on education and in some cases on occupation (see, e.g., [77]), and they are available in registers.

The correlation between socioeconomic factors and disease occurrence is widely recognized, see, e.g., [76] or [77]. Also, the correlation between socioeconomic factors and exposures has also been reported, see, e.g., the study on exposures to $NO_2$ [78]. Confounding by socioeconomic factors is a major potential source of bias in spatial epidemiological studies [67]. This is especially true in point source studies, as sources of pollution tend to be located in socioeconomically disadvantaged areas [32]. The lack of adjustment for socioeconomic factors may lead to artificially high risk-estimates, that is socioeconomically determined variations in the risk are wrongly attributed to environmental hazards. While the possibility of the over-adjusting for socioeconomic status should be noted. This may lead to an underestimate of the relative risk.

A good selection of the reference area is one way to control the confounding. As mentioned earlier, exposure may differ in urban and rural areas. Cancer incidence rates are also known to vary with geographical region (see, e.g., [79]), reflecting the effects of genetics or various lifestyle factors such as diet, smoking or alcohol consumption. In Finland, also the differences in rates between old towns, other towns and rural municipalities have been reported (see, e.g., [80],[81]).

The best way to control the confounding would be *randomization* of exposure. However, it is impossible in small area studies, and hence other solutions have to be employed [82]. One possibility is to use *restriction*, in which the study is aimed at specific classes of possible confounding variables. For example, we investigated farmers in the study of the effect of river pollutants on health [V]. In *stratification*, investigations are confounder-class specific. For example, in Article [V] we analyzed the incidence of total cancer in the classes of several possible confounders. The fourth possibility to control the confounding is to apply *modelling*. This means that several covariates, if available, can be controlled easily in the same model. It is also possible to employ several methods mentioned simultaneously.

Socioeconomic classification is available in data held by SMASH (see Section 4.1.1). The effect of using the socioeconomic class as an adjusting variable is reported in Article [I] and in Section 6.2. There are also several alternatives for the selection of the reference area (see Section 4.4). The effect of the choice of the reference area is also reported in Article [I] and in Section 6.2.

### 2.2.5 Direct and indirect standardization

In standardization, the effect of the confounding variable on the risk in the study population is removed by utilizing the known distribution of the confounder in a standard population. Standardization of rates is a basic tool in demography (see, e.g., [83],[84]) as well as in epidemiology (see, e.g., [82],[85]–[89]). The most frequently used methods in epidemiology are usually called as *direct standardization* and *indirect standardization*, while there is nothing "direct" or "indirect" about these standardizations [85]. There are just different weights as can be seen in the following. A general formula for the calculation of standardized rates can be written as:

$$\frac{\sum_j \frac{o_{sj}}{n_{sj}} w_j}{\sum_j \frac{o_{rj}}{n_{rj}} w_j}, \tag{2.1}$$

where $o_{sj}$ and $n_{sj}$ are, respectively, the observed number of cases and population count (or person-years) for category $j$ in the study area, $o_{rj}$ and $n_{rj}$ the observed number of cases and population count for category $j$ in the reference area, and $w_j$ is the weight (prevalence or proportion) for covariate category $j$, derived from the standard population.

When using direct standardization the question is: What should the incidence rate be in the study population if the distributions of the confounding variables were the same as in the reference population? The most commonly used directly standardized measure of the disease risk is the standardized risk ratio (SRR). If the reference area is taken as the standard, the formula for calculating of SRR takes the form [82]:

$$\mathrm{SRR} = \frac{\sum_j \frac{o_{sj}}{n_{sj}} \frac{n_{rj}}{\sum_j n_{rj}}}{\sum_j \frac{o_{rj}}{n_{rj}} \frac{n_{rj}}{\sum_j n_{rj}}} = \frac{\sum_j \frac{o_{sj}}{n_{sj}} n_{rj}}{\sum_j o_{rj}}, \tag{2.2}$$

where $o_{sj}, n_{sj}, o_{rj}$ and $n_{rj}$ are as before. Because the standard population serves as the reference population, the SRRs in several study populations are comparable. Direct standardization suffers from the problem of instability, because covariate specific incidence rates have to be calculated in the study area, where population counts are often small.

When standardizing indirectly the question is: How many cases would there have been in the study population if its morbidity were the same as in the reference population? The most commonly used indirectly standardized measures are the standardized incidence or mortality ratios (SIR or SMR). SMR has been in service at least since 1786, when mortality was compared in different occupational groups in the UK [86]. In indirect standardization, the study population is the standard, that is the formula for calculating of SIR takes the form [82]:

$$\text{SIR} = \frac{\sum_j \frac{o_{sj}}{n_{sj}} \frac{n_{sj}}{\sum_j n_{sj}}}{\sum_j \frac{o_{rj}}{n_{rj}} \frac{n_{sj}}{\sum_j n_{sj}}} = \frac{\sum_j o_{sj}}{\sum_j \frac{o_{rj}}{n_{rj}} n_{sj}}\,, \tag{2.3}$$

where $o_{sj}, n_{sj}, o_{rj}$ and $n_{rj}$ are as in (2.1). SIRs (or SMRs) are not comparable between different study areas, because in indirect standardization the study population itself is used as the standard. However, comparisons between each study area and the reference area are valid. In contrast to SRRs, SIRs and SMRs can be calculated when the covariate specific number of cases is not available in the study population. Indirectly standardized ratios are stable, because covariate specific incidence rates have to be known only in the reference population, which is usually sufficiently large.

A vast literature exists on comparisons of standardization methods, see, e.g., [82],[85]–[96]. In most publications, direct standardization is the recommended method because the SIRs in different study areas are not mutually standardized [82],[90]–[92],[95]. However, the general similarity of numerical results provided by direct and indirect standardization has also been observed [86],[87],[93],[94]. Breslow and Day [86] concluded that SRRs and SMRs give substantially different results if both (i) covariate specific population proportions and (ii) covariate specific risk ratios differ substantially, and (iii) the differences in (i) and the ratios in (ii) correlate.

Also other standardizing methods have been suggested. Breslow and Day [86] reviewed various weighting methods to improve standardization. Lee [96] suggested a new method: the *harmonically weighted ratio*, which can be used for external as well as internal comparisons.

Despite the theoretical problems with indirect standardization, it usually gives valid results in empirical applications. For example, Goldman and Brender compared the methods in ranking mortality in different populations and concluded that ”SMRs may be used to compare different geographical areas” [94]. The results of the comparison of direct and indirect standardization we achieved in the writing process of Article [V] will be shown in Section 6.2. Due to these considerations and the instability in our system, we use indirect standardization in the modelling.

### 2.2.6 Data quality

The formats of register data limit both the design of the study as well as statistical modelling. There are several problems in data quality, which may cause limitations in the presentation of ecologic studies, see, e.g., [35],[66] or [67].

The population counts or the estimation of person-years at risk usually rely upon certain points in time, most commonly upon the time of census [54]. The interpolation of the population counts or the person-years between censuses may be difficult, due to defective information on births, mortality or migration. Strong assumptions or modelling are required to obtain reliable intercensal counts, see, e.g., [97]. For example in the UK, population data are available in censuses. Instead, in Finland, population data are registered nowadays for every month.

For health data, there can be either under- or over-ascertainment, depending on diagnostic accuracy, completeness of registration, and possible duplication [98]. Also, geocoding of the health data (as well as of population data) may differ remarkably between areas [98].

The different levels of spatial data may cause problems in modelling. For example, in the UK, data are available on different levels of aggregation, and the extrapolation of data from level to another may cause difficulties, as reported by Small Area Health Statistics Unit [99]. In one example of incorrect linking, 53% of postcodes were put into a wrong enumeration district [98].

Data from different periods may also be disparate [98]. The administrative spatial units change in time. Boundaries of municipalities have changed and some municipalities may have been joined. In addition, for example in the UK, the extraordinary proliferation of different area codes is a practical problem. The numbers of inhabited postcode areas have been removed and have been possibly given to some new area. Diagnostic coding has changed many times. In addition of the problems due to changes, all the different versions of International Classification of Diseases, ICD-7 (from the year 1955), ICD-8 (1967), ICD-9 (1977) and ICD-10 (1992) are still in use in some countries. Also, the coding for the socioeconomic status has changed in time, at least in Finland.

The quality of the Finnish register data is generally high, see, e.g., [21] or the discussion in Section 4.1. In SMASH, we use grids based on metric coordinates as spatial units, which do not change in time. The individual level cancer data are aggregated in these grids in analyses of SMASH, so we do not have the extrapolation problem. However, SMASH holds data on population at risk only for three years, that is we need to extrapolate the population counts for remaining years (or months). In some cases, the data on the socioeconomic status may differ for cases and the population at risk, which will be discussed in Section 4.2.

### 2.2.7 Weak associations

The major problem in studies on environmental pollution around point sources is that the expected increases in the risk are usually modest. Typically, the estimates of relative risks in environmental studies are below 1.50 [100]. For example, in the case studies reviewed in Section 2.1, only in two cases were the highest estimates of relative risks over 2.00. Viel *et al.* [43] observed four leukaemia cases and reported SIR=2.80 within 10 km from the nuclear waste reprocessing plant. Michelozzi *et al.* [46] observed two cases of kidney cancer among males and reported SIR=2.76 within 3 km from the multiple sources of air pollution. In all other cases the highest reported relative excess in the risk at the source was in the range 0.04–0.99, that is, relative risks of 1.04–1.99.

Because there are risk factors (e.g., smoking or other lifestyle factors) that are more predictive for cancer than environmental factors, the potential for the confounding is strong. Also, the environment includes usually a very large number of low-level intercorrelated exposures, which often occur in complex mixtures [100]. If the association between exposure from a point source and cancer seems to be weak, inference on causality should be done cautiously. Reference values for the lower bound of acceptability of the relative risk for ecological studies have been introduced based on the investigations of the biases that occur in estimates when confounding variables are incorporated in the analysis [72]. For example, for lung cancer the relative risk in excess of 1.40, and for bladder and stomach cancer 1.20, is unlikely to be an artifact due to the uncontrolled confounding [72].

## 2.3 Statistical issues

Typical statistical problems in small area estimation are spatial autocorrelation, overdispersion, and sparseness of data (for example, small size of population or few cases per aggregation unit). The point source studies have also inferential problems such as problems of *post hoc* studies, multiple testing and determination of size and shape of a study area.

### 2.3.1 Spatial autocorrelation

Many routinely used methods assume that the observable outcomes follow independently and identically some relatively simple probability distribution. Contrary to this, spatial small area data are usually dependent and heterogeneous [101]. Spatial autocorrelation means correlation between counts in nearby areas. Several approaches have been suggested for measuring spatial autocorrelation (see, e.g., [102]). If the disease risk is estimated by a traditional approach in the form of the standardized incidence ratio SIR $= o/e$, where $o$ and $e$ are the observed and the expected number of cases, respectively, the spatial autocorrelation can be estimated, for example, by Moran's $I$ statistics [103] as

$$(2.4) \qquad I = \frac{n \sum_i \sum_{i'} W_{ii'}(\mathrm{SIR}_i - \overline{\mathrm{SIR}})(\mathrm{SIR}_{i'} - \overline{\mathrm{SIR}})}{(\sum_i \sum_{i'} W_{ii'}) \sum_i (\mathrm{SIR}_i - \overline{\mathrm{SIR}})^2}, i, i' = 1, \ldots, n, i \neq i',$$

where $n$ is the number of areas, $W_{ii'}$ is a measure of proximity of areas $i$ and $i'$. The variance of SIR is usually calculated under the assumption of independency [86], which may lead to biased estimates of confidence intervals.

The boundaries of small areas are usually administrative, or they are based on metric coordinates or another information not related to health data. Changes in disease incidence and relative risk do not follow externally determined boundaries of the small areas. Therefore small area statistics may not be assumed to be self-evidently independent between adjacent areas. The observed counts for rare diseases in small areas are usually assumed to follow the Poisson distribution [86],[104]. Problems appear when mutually dependent areas are aggregated. The sum of Poisson distributed counts is Poisson distributed too only if the counts are independent. The calculation of confidence intervals for a SIR assuming the counts to be independent and Poisson distributed thus does not lead to statistically strictly correct results.

It should be noted that spatial autocorrelation allows "borrowing of strength", that is, utilizing information from nearby or similar small areas.

The problems due to spatial autocorrelation will be one of the main focuses in this work. In an example case of Articles [II] and [III], we calculated Moran's $I$ and observed the typical problem of small area statistics on disease incidences being spatially correlated. In Section 5.3 models allowing and employing spatial autocorrelation will be considered.

### 2.3.2 Sparseness of data

Sparseness of data will be encountered with rare diseases (low disease counts) or data of high resolution (low population counts).

The heterogeneity of population density in the study area is a major problem of point source studies [34]. Many classical tests for point source studies may produce spurious results for such data [105]. Study areas usually contain both urban and rural areas. For example, in SMASH almost all small areas in the cities are inhabited whereas this is not the case in rural areas. In our case studies of rural study areas (Sulkava [I] and Paakkila [II]–[IV]), the inhabited small areas covered 20–25% of the area, whereas in the whole of Finland inhabited areas cover 16% of the total area. The problem of heterogeneity of population density is especially important in the case of high resolution data where the boundaries are not administrative. In SMASH, the number of empty small areas, isolated areas and areas with a few inhabited neighbouring areas becomes high. Also, area-specific population counts differ remarkably within both cities and rural areas. For example, in a single small area in Helsinki the population count was 3 000, whereas at Sulkava the total population count in all together 599 small areas was 4 400, the area-specific counts varying from 1 to 298 [I].

The uncertainty in the estimated incidence of disease is increased in areas with low population counts. For example, when using high spatial precision, the uncertainty in the estimates usually becomes large. If the increase in the risk of disease is estimated under the Poisson assumption by calculating the SIRs, the variances of the SIRs are roughly proportional to the inverse of the expected numbers of cases and, simultaneously, to the inverse of population counts in the area (see, e.g., [86],[104]). This causes difficulties in interpretation (see, e.g., [106]).

Low population counts may also affect the precision of the estimated relative risk of disease. The most extreme risk estimates are usually observed in these areas [32]. In areas with low population counts one more case can increase the relative risk estimate remarkably, while in areas with high population counts a change of disease count by one does not affect so much. If there are no cases in an area, the raw risk estimate refers to a false zero risk.

High spatial resolution reduces the exposure misclassification and the resulting bias in the estimated measures of the effect but increases the above effects due to lower population counts. The objective of small area analyses is to find a compromise between these effects. This is especially true when dealing with point source data. If, as usual, the areas with highest exposure near the point source are small, the uncertainty of the risk estimates in these most interesting areas is the largest. If the density of the population at risk is high, then high spatial resolution may be used, but when the density is low, the use of high resolution is often not possible, which leads to exposure misclassification and may lead to the underestimation of the effects of the exposure.

Smoothing methods (see, e.g., [52],[107]) are usually suggested to solve the problems of sparseness (see, e.g., [106]). Analyses of data with small amount of information get more feasible, if the information either from the neighbouring areas, other similar areas or from the whole data are utilized. The problem which remains is the level of smoothing. In particular, if the smoothing level will be estimated from the data, difficulties will be met.

The problems induced by the sparseness of data will be another main focus in this work. Later, in Section 5, smoothing methods are reviewed in more detail.

### 2.3.3 Overdispersion

Overdispersion or extra-variation in the data with respect to a given model exists when the observed variance of the number of cases is larger than the one derived from the model. One obvious source of overdispersion in area-level data is the violation of the assumption that covariates are constant within each area. A potentially more complex source of overdispersion arises through the failure to recognize residual spatial variation in the relative risk after adjustment for all known risk factors. If both the cases and the population at risk were measured without error, overdispersion might have either spatial or non-spatial components or both. In the literature, it is recommended to take into account the possibility of overdispersion. Several tests of heterogeneity (e.g., [102],[108]) and the methods for considering the overdispersion (e.g.,

[104],[109]) are introduced.

Due to the sparse high resolution data the population counts are usually very low in spatial units of SMASH. This implies that overdispersion is not such a major problem in SMASH. Despite that, a method accommodating overdispersion with spatial and non-spatial component will be introduced in Section 5.3.

### 2.3.4 *Post hoc* or a priori investigation?

Statistical inference on the results of point source studies becomes difficult if no hypothesis on an exposure source is set. It is a question of *post hoc* study if the excess number of cases is recognized before any possible environmental source of exposure and this observed excess leads to a search of a putative point source. This means a clustering illusion. For example, individual cases of disease may be noted and then the boundaries of the study areas are drawn. Prior knowledge of the disease incidence near a putative source leads an epidemiologist to carry out analyses to confirm the evidence. This problem may produce bias in data collection or in the definition of the study area. Also the hypothesis testing can be biased by this problem. For example, in testing the trend with distance from the source, controlling the size of the study area can lead to a desired result [23]. Data on more than one source, over different time periods, at different ages, and separately for men and women are usually recommended [32]. This may present a problem of *multiple testing*, which will be discussed in Section 2.3.5.

If the study area is thought *a priori* to be of interest because it includes a possible source of pollution, there is no *post hoc* problem if the internal spatial distribution of health data does not influence the choice of the area [28]. After all, the interpretation of studies based on the reported disease incidence or on the concern of several cases of disease around a putative source should be cautious.

To try to avoid the problems of *post hoc* studies, we have developed a protocol which should be read before any analysis done by SMASH. In the protocol, the basics (possibilities and restrictions) of small area studies are told. Due to the protocol a proposer of an investigation should pay attention to the reasons and the justification of a possible case of environmental emergency.

### 2.3.5 Multiple comparisons

The other inferential problem of point source studies is multiple comparisons. Often several cancer types in several areas around the point source are analysed (see, e.g., [39],[40],[44]–[46],[49]), but the problem of multiple comparisons is passed by. Only in two out of six previously mentioned case studies ([39],[44]) was the multiple testing considered. The assumption of the independence of the tests is usually doubtful. If statistical significance is set at 0.05, by the rules of probability one out of twenty tests produces a significant result by chance. The methods addressing this problem have been discussed by Greenland and Rothman [110]. By modelling techniques the problem of multiple comparisons may also be avoided [34].

Multiple comparisons are also a problem of studies analysed with SMASH (see Article [V] or Tables 2 and 3 in Section 4.5). There is no systematic practice to handle this problem. Instead, it should be noted in the interpretation of the results.

### 2.3.6 Determination of a study area

The determination of a study area may have a major influence on the results [111]. Without high-quality exposure measurements the boundaries are usually drawn arbitrarily. It is quite difficult to know how large an area to consider. For example, in routine cases in Small Area Health Statistics Unit, the area within 7.5 km from the source (see, e.g., [12] or case studies [42],[44]) is used for investigation. Often the boundaries are drawn based on the availability of population data. The immediate problem is to determine the population size in the exposed areas. Who are exposed? Those who live in the area, or those who have lived (say 10 years) in the area, or those whose workplace or school is in the area? The determination of a study area based on health data leads to problems of *post hoc* studies described earlier. The selection of too large a study area may lose the real effect. While with too small an area the null hypothesis may be rejected too infrequently for rare events.

In the case studies analysed by SMASH, the proposer of an investigation is required to define the exposed area which is used as the study area (see Section 4.4). In the determination of the population at risk, we are limited by the availability of data. As a geocoded data SMASH holds only the place of residence. Alternative choices of selecting the time of the place of residence will be presented in Section 4.2.

# 3 Aims of the study

The overall objective of this thesis is to further develop the small area system SMASH. The data stored in the system are aggregated into squares of size 500 m × 500 m. High spatial resolution is informative but brings up new methodological challenges in the risk estimation. High resolution is valuable in the risk estimation, but at the same time creates sparse data especially in rural areas. SMASH, including a database and modelling tools, concurrently requires and sets restrictions on the methodological development. This system forms the frame of this development.

The more specific goals of the present work are:

(i) *To search suitable and reliable advanced statistical methods for analyzing the regional variation of cancer risk around a putative source of pollution within SMASH.* Advanced statistical methods are searched to solve the problems of spatially autocorrelated data, sparseness of data, and other possible restrictions due to the high resolution of data. [II–IV]

(ii) *To implement the methods as a part of SMASH.* In the implementation computation algorithms of the methods are coded with suitable programming language. [II–V]

(iii) *To evaluate the use of register data.* In the evaluation the aim is to find out what can be achieved with the register data and when more laborious and costly individual level data should be introduced. [I–V]

(iv) *To test and validate epidemiological methods within SMASH.* Small area epidemiological methods are tested by means of case studies. The standardization methods are compared. The effect of different confounders is also studied. [I,V]

# 4 Small Area Statistics on Health (SMASH)

The Finnish Cancer Registry and the Unit of Environmental Epidemiology in the National Public Health Institute have been developing a system called SMASH for small area epidemiology purposes. The objective of SMASH is to provide the user with an easy, quick, accurate and reliable method for a preliminary analysis of the area-level cancer incidence. The main aim of SMASH is to give an answer to a question like whether there are more cancer cases in a prefixed area than what would be obtained by the knowledge on the reference population.

The need for the existence of a system like SMASH is mainly in getting a preliminary answer in administratively problematic situations. Usually, health authorities and communal decision-makers require an initial answer quickly. For example, in 1998, the knowledge on the building of the housing estate on a former waste site raised concern in the local population at Myllypuro, a suburb of Helsinki. The Environment Centre of the City of Helsinki required an investigation about the possibilities of connections between the high concentrations of toxic substances and diseases. An individual level case-control study was built up and the results were available after some months [112]. Meanwhile, SMASH-based results were produced within a few days [I]. Initial answers for informing the local population were available considerably quickly. Before the existence of SMASH that kind of rapid analysis was not possible: The Finnish Cancer Registry was able to produce geographical analyses at the municipality level only [113].

At an early stage of the development of SMASH, cancer incidence was modelled by the classical Poisson regression. In the model applied by Pekkanen *et al.* [114], the distance from the point source, age and sex were the independent variables. The lack of control of spatial correlation was recognized, and the development of SMASH continued.

## 4.1 Register data bases as sources of information

Minimal data requirements for producing rapid initial risk estimates are health data, denominator (or control) data, and data on putative sources of pollution (point sources, roads, railways, power lines, etc.). All data should be georeferenced in a way which allows spatial linking. Key requirements for *health data sets* are some form of diagnostic code, date of diagnosis, date of birth, and sex of person. Because diagnostic coding (e.g., (ICD)) varies through time (see Section 2.2.6), it is essential that codes should be matched. *Denominator data sets* are required for calculating rates of diseases within geographical areas. In many cases the denominator data will be population data. In some cases the denominator may be data on births or data on hospital admissions depending on health data examined. Denominator data must include date and must be broken down by sex and age to enable the estimation of age-sex specific rates. Other *environmental data* (geography, land use, forests, rivers, lakes, etc.), *meteorological data* (e.g., prevailing wind direction and wind speed) and *data on confounding variables* (e.g., data on socioeconomic class or data on life style factors) would further improve geographical information systems as a risk assessment tool.

Easy linkage of register data is possible in Finland. All Finnish residents since 1 January 1967 have a unique personal identification number. Those personal identifiers enable the linking of data bases straightforwardly and fast.

As a geographical system SMASH applies the National Grid Coordinate System [115]. In this system Finland is divided into four projection zones based on the central meridian of these zones. The projection zones have the central meridian at 3 degree-intervals starting from the 21-degree longitude. The respective numerical values of these central meridians are given at 1 000-kilometre-intervals starting from 1 500 kilometres, where the first number is the ordinal number of the particular zone. The longitude in each zone is the distance from the central meridian. The latitude is the distance from the equator.

### 4.1.1 Population data

*(i) Population data and geographical residential data*

Statistics Finland has been collecting population census data since 1950 every ten years. Since 1975 also *intermediate population censuses* have been collected. Date of birth, sex, marital status, mother tongue, citizenship, religion, and regular place of residence are registered as demographic variables. Since 1970, the personal identifiers have been registered.

The Population Information System of the Population Register Centre holds data on Finnish residents and on buildings and apartments in Finland. Data are supplied by the administrators of the building inspections of the municipalities. The centre points of all buildings in Finland are coded to provide the coordinates in latitude and longitude. This data base is available from the year 1970 onwards. The quality and accuracy of coordinate data vary by municipality. In 1970, the accuracy was principally 10 metres, in some municipalities in East-Finland and in North-Finland up to 500 metres. The administrators of municipalities have been responsible for the continuous updating of the data base. Since 1980, the coordinates could be provided with an accuracy of one metre. An extensive cross-checking was performed in the turns of the decades 1980 and 1990. In the cross-checking in 1999 the Population Register Centre observed that more than 99% of the residences were recorded in The Population Information System.

These register data bases can be linked using unique identification numbers to obtain co-ordinates of the home address of each Finnish resident. In a validation study in the year 1990, 97% of the 96 000 Finns surveyed lived in the same building as was recorded in the register [116]. Accordingly, the quality of georeferenced population data obtains high ranking.

*(ii) Data on social class*

The population census data collected by Statistics Finland hold data on education and economical activity. Based on these data, socioeconomic classification can be easily obtained.

Data on the educational level and the educational status were asked by questionnaires up to 1970. From 1975 to 2000 data were collected from registers of the Population Register Centre.

Up to 1985, variables describing economical activity (principal activity, occupational status, livelihood, occupation, site of work place, socioeconomic status) were asked by questionnaires. From 1990 to 2000 the variables were collected from the registers of Statistics Finland and Population Register Centre. Data on income were collected from registers from 1970 to 2000.

*(iii) Population data in SMASH*

Based on the data described above, numbers of inhabitants by sex, age and socioeconomic class in squares of size 500 m × 500 m are calculated for the whole of Finland, and these data form an input to SMASH. Currently, SMASH holds population data for the years 1980, 1990 and 1997. The population counts in intermediate years are estimated as the weighted average of the existing data. The inhabited squares cover around 16% of the land area of the country. Age is categorized into eighteen 5-year age-groups (0–4, 5–9,...,80–84 and 85–). The socioeconomic classification used in SMASH is based on the socioeconomic status and the education level. As discussed above, there are also other variables in population censuses describing economic activity, but the socioeconomic status and the education level have been chosen to describe the socioeconomic class of each Finnish resident. The classification of socioeconomic activity follows broadly the one applied by Pukkala [77] based on the socioeconomic status code of Statistics Finland. The difference is that the farmers, foresters and fishermen are separated in a group of their own and the manual workers are divided into separate groups of "skilled" and "unskilled" manual workers. The socioeconomic classification used in SMASH comprises six categories:

1) farmers (employers, self-employed persons, employees), foresters and fishermen;
2) other employers and self-employed persons, and upper clerical workers;
3) lower clerical workers;
4) skilled manual workers;
5) unskilled manual workers;
6) others.

The temporarily unemployed persons have been classified according to their latest occupation. Those not in labour force (e.g., housewives and students) are categorized by the reference persons of the household-dwelling unit. Skilled manual workers are those workers having at least lower intermediate education grade, that is, upper secondary school, vocational (technical, economical, etc.) school, vocational high school or university. The group "others" includes pensioners, students, pupils, employees with unknown vocation, and the unemployed without a reference person in the family.

The classification of socioeconomic status for the population in 1980, 1990 and 1997 is derived from the most recent census, that is from years 1980, 1990 and 1995, respectively. For pensioners, the last non-pensioner socioeconomic status is used, derived by searching the previous census data up to the 1970 Population Census.

### 4.1.2 Cancer data

*(i) Registration of cancer cases*

The Finnish Cancer Registry has been collecting data on all incident cancer cases and all cancer deaths in Finland since 1953. Since 1961, reporting is been compulsory. All hospitals, medical practitioners and institutions with hospital beds are obliged to notify the Finnish Cancer Registry of all cancer cases that come to their attention. Pathological and cytological laboratories send information on all tissue and cytological specimens with a diagnosis of cancer. Statistics Finland sends a report whenever cancer is mentioned on the death certificate. The quality of cancer data is very high. More than 99% of the cancers are registered by the Finnish Cancer Registry [117].

*(ii) Cancer data in SMASH*

Health data held in SMASH currently consist of all cancer cases diagnosed in Finland between 1981 and 2000. A person can have more than one cancer, that is, same person may be more than once in the data. The total number of cases is 460 565 (373 660 malignant tumors). The place of residence for cancer patients on 31 December in 1980, 1990, and on 31 December the year before the cancer diagnosis is defined through the coordinate system described above. Sex, age and socioeconomic status of all cases are are classified in the same way as in population data. The classification of the socioeconomic status is derived from the most recent census (1980, 1985, 1990, 1995 or 2000).

Through the personal identifiers, data on population, cancer, residence and socioeconomic class have been linked by Statistics Finland.

### 4.1.3 Environmental data

At present, SMASH holds neither data on putative sources of pollution nor other register data describing exposure levels. In routine applications of SMASH, the existence and the location of a source is usually given by the proposer of an investigation (see Section 4.4).

As environmental data, SMASH holds digitized maps such as the General Map at 1:400 000 scale and the PerusCD including a basic map at 1:20 000 scale. Those maps are both in the raster form. SMASH also holds digitized data on all the rivers and lakes in Finland. These digitized data are produced by the National Land Survey of Finland. The inclusion of further environmental data is case sensitive.

## 4.2 Different types of cohorts

In SMASH, there are two alternative cohort definitions. In the *cross-sectional cohort*, the study population is based on the place of residence either in 1980 or 1990. In the *dynamic cohort* the place of residence the year before the diagnosis is used for the study population.

In studies based on the cross-sectional cohort, both the observed number of cases and the population counts come from the cross-sectional population in 1980 (or 1990). The cross-sectional population is followed irrespective of whether or not they later move out of the area.

The strength of the dynamic cohort is that it gives an accurate answer to the main administrative question asked from SMASH in an environmental emergency, namely: Does cancer occur more frequently near a suspected source of pollution than could be expected? Also, if the assumption of dynamic populations is true, that is, the population moving in and out of the area is similar, the approach gives a good estimate of the incidence of cancer in the area.

There may be strong temporal changes in the amount of exposure due to the point source. Large changes, for example, may be introduced to the point source to reduce emissions, especially if there is a concern about an increased risk of disease in the area. In this situation, the place of residence at the start of the follow-up in 1980 (or 1990) is probably a better marker of exposure than the place of residence the year before diagnoses for diseases with long latency like cancer. On the other hand, for diseases with short latency such as childhood cancers, using the dynamic cohort provides a flexible tool to link changes in exposure to changes in the risk of these outcomes.

When using the dynamic cohort, population counts for each year are not available but are subject to interpolation of using the population counts in 1980, 1990 and 1997. On the other hand, the observed cancer cases are classified according to the place of residence the year before cancer is diagnosed. A similar weakness is met with the socioeconomic status; it is not necessarily defined at the same time. For persons with cancer, the socioeconomic status is derived from the previous population census (in 1980, 1985, 1990 or 1995), while for the population at risk, the socioeconomic status is interpolated from the data in 1980, 1990 or 1995. In the cross-sectional cohort, the data on the numerator (cancer cases) and on the denominator (population at risk) are strictly comparable.

## 4.3 Estimation of risk in routine analyses

In routine applications of SMASH, the standardized incidence ratio (SIR) with the 95% confidence interval is calculated as an estimate of cancer risk. SIR is computed by dividing the observed number of cases ($o$) by the expected number of cases ($e$) over the study period, namely:

$$\text{SIR} = \frac{o}{e}, \tag{4.1}$$

where the observed number of cases is the sum of the cases in the study area. The expected number of cases is calculated according to [82]:

$$e = \sum_j n_s(j) o_r(j) / n_r(j), \tag{4.2}$$

where $n_s(j)$ is a population count in the study area, and $o_r(j)$ is the number of observed cases and $n_r(j)$ the population count in the reference area in sex, age and socioeconomic status specific covariate group $j$. Due to the lack of data on mortality, the formulation based on person years is not possible.

In Breslow and Day [86], there is a table of coefficients for the accurate calculation of confidence intervals of SIRs for the selected number of cases under the assumption of independency and under the Poisson model. There are also approximation formulas for this of which Byar's formula seems to be the most accurate and accurate enough for the calculation of 95%-confidence interval when the number of observed cases is at least 20. In SMASH, the confidence interval is calculated by Byar's formulas when the number of cases is at least 20. If the number is 20 or less, the coefficients tabulated by Breslow and Day [86] are used. Byar's approximation formula, giving the lowest limit of the 95%-confidence interval, is

$$\text{SIR}_L = \frac{o}{e}\left(1 - \frac{1}{9o} - \frac{1.96}{3o^{1/2}}\right)^3. \tag{4.3}$$

Respectively, the upper limit is

$$\text{SIR}_U = \frac{o+1}{e}\left(1 - \frac{1}{9(o+1)} + \frac{1.96}{3(o+1)^{1/2}}\right)^3. \tag{4.4}$$

## 4.4 Routine analyses

In routine analyses the possible selections are made on study area, cancer, years of diagnosis, cohort type, age-groups, socioeconomic status and reference population. The socioeconomic status can be ignored.

A study area can be defined freely with an accuracy of 500 metres within Finland. Health outcome can be the total cancer, some group of cancers or specific cancer type from the following list. To overcome the problems of different versions of ICD codes over time the Finnish Cancer Registry has developed the coding of their own, which does not directly follow any ICD coding (see, e.g., [77]).

*Mouth or pharynx:* lip, tongue, salivary glands, other mouth, pharynx;

*Digestive organs:* oesophagus, stomach, small intestine, colon, rectum or rectosigmoid, liver, gallbladder or bile ducts, pancreas, other digestive organs;

*Respiratory organs:* nose or sinuses, larynx or epiglottis, lung or trachea, pleura, mediastinum;

*Breast*

*Female genital organs:* cervix uteri, corpus uteri, other uterus, ovary, other female genital;

*Male genital organs:* prostate, testis, other male genital;

*Urinary organs:* kidney, bladder or ureter or urethra;

*Circulatory organs:* non-Hodgkin's lymphoma, Hodgkin's disease, multiple myeloma, leukaemia;

*Other organs:* melanoma of the skin, skin (non-melanoma), eye, nervous system, thyroid gland, other endocrine glands, bone, soft tissues, other or unspecified.

A reference population represents routinely the same sex, age-groups and socioeconomic status strata. For example, farmers cannot be compared to other groups in a routine analysis. Instead, there is a range of choices for an area defining the reference population. The reference area can be the whole of Finland or one of the five large geographical areas based on the districts of the university hospitals: Helsinki (South), Turku (South-West), Tampere (West), Kuopio (East), and Oulu (North). The reference can also be a population of old towns (founded before 1906), other towns, rural municipalities, or any combination of individual municipalities.

The software used in routine analyses is an ArcView GIS 3.2 application running under the Windows NT 4.0 Workstation.

To try to avoid the problems of *post hoc* studies, we have developed a protocol with a form which should be read and filled in before any analysis. In the protocol, the basics (possibilities and restrictions) of small area studies are told. Due to the protocol a proposer of an investigation should pay attention to reasons and justification of a possible case of environmental emergency. In the form, the description of the problem, the hypothesis including the assumed exposure and the putative source of pollution in addition to the choices mentioned in the previous section are asked.

The high resolution of data brings up the questions of confidentiality. The personal identifiers have been deleted from data after linkages and before loading in SMASH. In the data loaded in SMASH, the date of birth is available by year. In a case of sparsely inhabited areas the results are shown in such a way that a person cannot be identified. The data are stored in the password protected folders on the hard disk of a desktop computer. The backup disks of the data are also under lock and key.

## 4.5 Case studies in 1995–2003

In this subsection case studies analysed with SMASH between the years 1995–2003 are described. In Tables 2 and 3, a short description of all case studies are presented. A more detailed description of the cases of Sulkava, Myllypuro, Paakkila and Kymijoki can be found in the articles this thesis is based on [I],[II],[V].

The first case study was a pilot on coordinate-based data [114]. This study was carried out before the SMASH software was available. Total cancer and leukaemia in 1983–86 were investigated around the oil refinery in Porvoo, South-Finland, because an excess of risk in Porvoo was reported [118]. The ratio of the observed and the expected number of cases was modelled using the classical Poisson regression. No association was found between the distance from an oil refinery and the risk of total cancer or leukaemia.

The next case studies were carried out with SMASH software. In a small village of Pahkala, concern was raised about several cancers. This case was a typical *post hoc* study, because it was a disease driven and not an exposure driven case study like the other case studies. In a municipality of Hattula, a local physician was concerned about the possible increased cancer risk in

Table 2: Case studies carried out with SMASH in 1995–1999.

| Study area,<br>year of the study,<br>source of pollution,<br>reference | Short description of the study (reason, cancers, years of diagnosis) | Main results (risk in total cancer, unless else stated) |
|---|---|---|
| Porvoo,<br>1995,<br>oil refinery,<br>[114] | Empirical example for research purposes. Total cancer, leukaemia, 1983–86. | No increased risk with distance from source. |
| Pahkala,<br>1997,<br>no specified source | Concern about several cancer cases in a small village. Total cancer, 1981–92. | 16% increase. |
| Hattula,<br>1998,<br>no specified source | Local physician concerned about cancers in area with high concentrations of arsenic. Total cancer, 1981–92. | 6% increase. |
| Säkylä,<br>1998,<br>problems in drinking water | Municipality administrators concerned about cancers in citizens using drinking water with high concentration of chloro ethene. Total cancer, liver, non-Hodgkin's lymphoma Hodgkin's disease, multiple myeloma, leukaemia, 1981–92. | 3% increase. |
| Myllypuro,<br>1999,<br>former waste site,<br>[I] | Concern about health effects in a housing estate. Total cancer, digestive organs, lung, leukaemia urinary organs, total cancer among children, 1981–92. | 9% increase in total cancer and 102% increase in lung cancer when compared to all old towns. 1% and 6% increase when compared to the another similar area. |

Table 3: The case studies carried out with SMASH in 2000–2003.

| Study area, year of the study, source of pollution, reference | Short description of the study (reason, cancers, years of diagnosis) | Main results (risk in total cancer, unless else stated) |
|---|---|---|
| Parkano & Kihniö,<br>2000,<br>problems in drinking water | Municipality administrators concerned about cancers in citizens. High concentrations of radon, iron and manganese in drinking water. Total cancer, lung, kidney, skin, 1981–97. | 2% increase. |
| Sulkava,<br>2000,<br>problems in drinking water,<br>[I] | Citizens of municipality concerned about cancers. High concentrations of chloro phenols in drinking water. Total cancer, colon, kidney, sarcoma, non-Hodgkin's lymphoma, Hodgkin's disease, leukaemia, 1981–97. | 80% increase in kidney cancer in 1990s. |
| Olkiluoto & Loviisa,<br>2000,<br>nuclear power plants | Empirical example for research purposes. Leukaemia, 1981–97. | No increased risk around either of the nuclear power plants. |
| Paakkila,<br>2001,<br>former asbestos mine,<br>[II] | Empirical example for a methodological articles. Lung cancer, 1981–97. | 280% increase in a vicinity of the mine. |
| Kurikka,<br>2003,<br>fur dressing chemical factory | Local physician concerned about cancers in area with high concentrations of chloro ethene in soil. Total cancer, liver, pancreas, bladder, kidney, oesophagus, non-Hodgkin's lymphoma, 1981–2000. | 15% increase. |
| Kymijoki,<br>2003,<br>dioxin polluted river,<br>[V] | First part of a sophisticated multi-approach study. Total cancer and selected subtypes among farmers, 1981–2000. | 13% increase near the river. |

areas with high concentrations of arsenic. In a municipality of Säkylä, municipal administrators paid attention to the health effects of high concentration of chloro ethenes, measured in local drinking water. In a federation of municipalities of Parkano and Kihniö, administrators were worried about high concentration of radon, iron and manganese in drinking water. In a town of Kurikka, a local physician was anxious about the the possible increased cancer risk in areas with high concentrations of chloro ethenes, measured in soil around a local fur dressing chemical factory. In all these cases the cancer incidence did not differ from the normal level, and the proposers of the analyses were calmed.

A case study on leukaemia around the nuclear power plants in Olkiluoto and Loviisa was performed by the request of researchers from the Radiation and Nuclear Safety Authority of Finland (STUK). The incidence of leukaemia was investigated in the circles around the nuclear power plants. As a result, no increased risk was detected and no association was found between distance from the plant and risk of leukaemia.

## 4.6 Problem of using cumulative incidences

SMASH does not hold data on an individual level but on the ecological level. Therefore we are not able to use methods applying person-years but population counts in the calculation of incidences, see Formula (4.2). In the other words, we do not have incidence rates but *cumulative incidence* [82] or *incidence proportion* [119]. The applying of the cumulative risk is not a problem if the assumption of similarity of competing risks of death in the exposed and in the reference group is valid. However, the cumulative incidence may be biased for a lengthy period, because the probability of dying increases and the observed number of cases simultaneously decreases. For example, the decrease of RRs with time periods in farmers (see Table 4 in Article [V]) may be caused by this phenomenon.

## 4.7 Hybrid adjustment - a problem of indirect standardization

Relative risks in indirect standardization cannot be estimated reliably for covariates used in the calculation of the denominator. This is called a *hybrid adjustment* [89]. This problem was met during the writing process of Article

[V]. In this particular case, the expected number of cancer cases was calculated in classes of covariates such as sex, age, socioeconomic status and the time period. With the log-linear model we estimated the risk ratios for the covariates and also for spatial covariates describing the effect of distance from the river and the sea. In doing this, the model with the expected number of cases as the denominator (indirect standardization) gave incorrect estimates of the risk ratio for those used in the calculation of the denominator. For example, women had a higher relative risk of cancer than men, while reality is commonly assumed to be the opposite. In spite of that, for the covariates describing the effect of distance from the river and from the sea, the estimates of risk ratios were the same as given by the model with population count as the denominator.

## 4.8 A call for methodological tools

The methodology used in routine applications of SMASH assumes that the observed number of cancer cases is independently Poisson distributed, which in many cases turns out to be a simplification. Because the methods do not consider the possible spatial autocorrelation, this may cause errors in the estimation of confidence intervals. In addition, the routinely used methods, like SIRs, cannot satisfactorily deal with the problem of the sparseness of data.

One convenient approach to overcome the problem of spatial autocorrelation is to employ models with fixed effects or spatial covariates. Also the models with random effects or spatial random effects are possible to use. The latter requires applying the Bayesian framework. The problem of sparseness requires some smoothing methods, that is methods which can utilize information either from the neighbourhood or from other similar areas or from the whole data.

# 5 Borrowing strength in small area analysis

## 5.1 A need for smoothing methods

For a non-infectious disease, with a known closed population and fixed follow-up time without censoring and disease risk, the counts of disease may be considered as independent binomial random variables. For rare diseases such as cancers, the binomial distribution may be approximated by the Poisson distribution [86],[104]. The counts $o_{ij}$ in an area $i$ and a covariate strata $j$ may be considered as independent Poisson distributed random variables, that is,

$$o_{ij} \sim \text{Poisson}(\theta_{ij}), \tag{5.1}$$

where $\theta_{ij} = e_{ij}\lambda_{ij}$. Here $e_{ij}$ stands for the expected number of cases derived from a reference population and assumed to be observed without error and $\lambda_{ij}$ stands for the relative risk. An advantage of the Poisson assumption under the independence is that we can sum over strata $j$ and obtain

$$o_i \sim \text{Poisson}(\theta_i),\ \theta_i = \sum_j \theta_{ij} \tag{5.2}$$

for $o_i = \sum_j o_{ij}$. Now the maximum likelihood estimator of the relative risk $\hat{\lambda}_i = o_i/e_i$ is equal to $\text{SIR}_i$. This model can be presented in terms of a simple generalized linear model

$$\log \theta_i = \log e_i + \log \lambda_i, \tag{5.3}$$

where the first term is considered as an offset variable.

A typical problem in the estimation of the relative risk of a rare disease for small areas is the precision of the estimate. The most extreme risk estimates are usually observed in areas with low population counts. In those areas an additional case can increase the relative risk estimate remarkably. Secondly, if there are no cases, which occurs with a non-negligible probability, a zero risk estimate is obtained. This is in any case wrong, because the risk is always over zero.

In addition, for small areas the variance of SIR easily becomes large, because the variance is approximately proportional to the inverse of the expected number of cases. The high variance of the SIRs has the effect that the analysis is inefficient and the interpretation difficult.

The problem of the Poisson model (5.1) is that the assumption of independence usually does not hold, that is, the observed number of cases are spatially correlated. One reason for this claim is that the observed counts $o_{ij}$ are affected by one or several spatially correlated factors which have not been observed.

In order to overcome the problems of spatial aurocorrelation and low precision of the estimate, *smoothing* or *"borrowing strength"* methods are usually employed. The idea in smoothing is to borrow the information from other small areas for the estimation of the relative risk. The information can be borrowed either from similar areas, from nearest areas (local smoothing) or from all areas in the study area (global smoothing). Smoothed estimates of the relative risk can be received by smoothing either the estimator itself or the numerator and the denominator of the estimator separately. Several methods for smoothing are available, see, e.g., [27],[52],[107],[120],[121]. Those methods are not specific for point source studies, but many of them can be applied in this specific context. When smoothing methods are utilized, the possibility of over-smoothing must be noted. With over-smoothing possibly remarkable findings may be faded out.

## 5.2 Distance-based models

### 5.2.1 Parametric distance-based models

A vast literature exists on parametric models based on the distance from a point source, see, e.g., [8],[33],[34] or [109]. When parametric models are used in the calculation of predictions (or fits), these predicted values can be considered as smoothed values. In these models the relative risk is usually considered as a function of distance ($d_i$) from the source, that is,

$$\lambda_i = f(d_i), \tag{5.4}$$

with $f(d_i) \to 1$ as $d_i \to \infty$. In the case of aggregated count data, the distance from a source $d_i$ is measured from the centroid of the area $i$. These models make the estimates smooth by treating all the areas with the same distance from a source identically and by taking advantage of the information from all those areas. We utilized a popular distance-based function introduced by Diggle [25]:

$$f(d_i) = 1 + \alpha \exp\{-(d_i/\beta)^2\}, \tag{5.5}$$

where $d_i$ is the distance between area $i$ and the point source, $\alpha$ is the proportional increase in risk, and $\beta$ measures the rate of decay with increasing distance from the source [III]. Other functions can be also utilized, see, e.g., [8],[30],[33]–[35],[109],[122].

### 5.2.2 Regional effect model

A simple non-parametric simplification of the distance-based model (5.4) is to exploit spatial fixed effects. The idea of this method is to divide the whole study area into (say) $k$ sub-areas, which are compounded of underlying small areas. Then we aggregate counts in small areas into these sub-areas, in other words we fix the relative risk at the sub-area level $k$. So we obtain the log-linear presentation

$$\log \theta_i = \log e_i + \beta x_k, \tag{5.6}$$

where an area $i$ is within a sub-area $k$ and $x_k$ is an regional effect related to the sub-area $k$. If the observed numbers of cases are assumed to be independently Poisson distributed random variables, this is an instance of the *Poisson regression model*. We applied a *full Bayesian version* of this method [II]. We aggregated small areas and calculated the relative risk at the level of sub-areas.

If the relative risk is presumed to vary also according to some regional effect ($x_k$) and a set of $J$ ecological covariates ($z_{ij}$), then we obtain a generalization of (5.6)

$$\log \theta_i = \log e_i + \beta x_k + \sum_{j=1}^{J} \gamma_j z_{ij}, \tag{5.7}$$

where $\beta = 0$ if $i \notin k$. We applied the model (5.7) with fixed effects to a line source problem [V].

In these simple cases (5.6–5.7) smoothing arises from the consideration of the marginal distribution of the relative risk and taking advantage of information from the whole data at the sub-area level (a group of small areas). These estimators reduce the problem of uncertainty of estimates in small areas. The possible spatial autocorrelation is somehow considered through the fixed regional effects and through adjustment for the covariates. If strong spatial correlation exists, the regional effect model (5.7) suffers from invalid summing. Because, the summing is not correct, representing the SIR as a raw estimate of relative risk is usually irrelevant.

## 5.3 "Borrowing strength" models

The models described in the previous subsection are not able to handle the problem of spatial correlation, because these models assume that observations are independently and identically distributed. A direct consequence is that the inference based on results may not be correct. Two main approaches to deal with correlated counts are Markov type modelling developed for image analysis (see, e.g., [123]) and, especially, hierarchical (or mixed) models (see, e.g., [16] or [124]).

### 5.3.1 Auto-Poisson model

An extension of the Poisson regression allowing the dependence of the observations is the so called *auto-Poisson model* introduced by Besag [125]. The idea of this method is to utilize the neighbouring information on cases. Here, a neighbourhood is defined through a symmetric relation "$\sim$". For example, all small areas $i'$ which share a common boundary with a small area $i$ can form the neighbourhood, $i' \sim i$, for area $i$. The observations are assumed to be conditionally Poisson distributed, such that

$$o_i | o_{i', i' \sim i} \sim \text{Poisson}(\theta_i). \tag{5.8}$$

Now the regression model can be written as

$$\log \theta_i = \log e_i + \beta x_k + \sum_{i' \sim i} \beta_{i'} o_{i'} + \sum_{j=1}^{J} \gamma_j z_{ij}, \tag{5.9}$$

where $\sum_{i' \sim i}$ means the summing over the neighbourhood of area $i$, $\beta_{i'}$ is an autoregressive coefficient representing the spatial autocorrelation. The problem when using this model is that the joint distribution for $o$ may be defined only when the autoregressive coefficients are below zero [126].

### 5.3.2 Truncated auto-Poisson model

To overcome this problem, Ferrándiz *et al.* [126] have proposed a choice which they called a *truncated auto-Poisson model*. They truncated the Poisson distributions to a limited range. In fact, they restricted Poisson counts to be smaller than the living populations in the small areas. This requires the use of direct standardization.

A major weakness of the (truncated) auto-Poisson model is that it is not strong in modelling global patterns of risk although it may treat local dependence adequately.

### 5.3.3 Empirical Bayesian model

Clayton and Kaldor [19] introduced an empirical Bayesian modelling approach for disease mapping. In their model, the observations are assumed to be conditionally independent given the relative risks, that is,

$$o_i | e_i, \lambda_i \sim \text{Poisson}(e_i \lambda_i). \tag{5.10}$$

From this model, the posterior expectations $\{\lambda_i\}$ given $\{o_i\}$ may be estimated. The prior distribution for the relative risk is commonly assumed to be Gamma$(\alpha, \beta)$. The hyperparameter estimates $\hat{\alpha}$ and $\hat{\beta}$ are derived from the marginal likelihood. The empirical Bayes point estimates of the relative risk can be written as

$$\hat{\lambda}_i = \frac{o_i + \hat{\alpha}}{e_i + \hat{\beta}}. \tag{5.11}$$

The estimator of the relative risk (5.11) is a compromise between data and a prior mean of the distribution of relative risks. If the number of the observed cases is large, the estimate is close to the SIR, while with small numbers the estimate is close to the prior mean of the relative risks. We applied the empirical Bayesian method for data aggregated at the sub-area level [II].

### 5.3.4 Hierarchical hidden Markov models

One advanced way to overcome the problem of possible spatial correlation is to add a spatially structured random effect into the model. The idea of adding this effect is to allow similarity of neighbouring estimates. The observed numbers of cases are again assumed to be conditionally independent given the relative risks. Now the regression model takes the form

$$\log \theta_i = \log e_i + \beta x_k + \sum_{j=1}^{J} \gamma_j z_{ij} + u_i, \tag{5.12}$$

where $u_i, i = 1, ..., I$, denote the spatially structured random effect. This may be referred to as *a local smoothing*. The dependence of the $I$-dimensional random variable $u = (u_1, ..., u_I)$ can be defined either by a *joint model* of $u$ or by a *conditional model* with the univariate conditional distribution for $u_i|u_{i'}$, where $i' \neq i$ and $i, i' = 1, ..., I$.

In joint modelling when using a Gaussian Markov random field, dependence is introduced by the correlation (or covariance) matrix of the multivariate distribution of the random variable. Wakefield *et al.* [104] review definitions for dependence, in which the correlations of areas $i$ and $i'$ depend on the distance between the areas.

In the conditional specification, an explicit definition of the whole covariance matrix is not required. The matrix of weights describing the association between the neighbouring estimates and the diagonal of the covariance matrix is required. This model is called a *conditional autoregressive (CAR) model.*

The hierarchical autoregressive model was introduced by Besag *et al.* [127] as an extension of the Poisson regression model with two random components, $u_i, v_i$,

$$\log \theta_i = \log e_i + \beta x_k + \sum_{j=1}^{J} \gamma_j z_{ij} + u_i + v_i, \tag{5.13}$$

where $u_i$ is as in the previous model (5.12) and $v_i$ is an unstructured (or heterogeneity) random effect. Now, in addition of the modelling of the spatial dependence, the overdispersion is modelled through the unstructured random effect.

The hierarchical Markov model has become very popular and widely accepted in spatial modelling. For the spatially unstructured effect $v_i$, Besag *et al.* [127] set a conventional prior with zero mean with variance $\delta_v^2$. Instead, for the structured effect $u_i$ they introduced a "borrowing strength" prior, an *intrinsic Gaussian autoregression prior*, given by

$$p(u_1, ..., u_I|\kappa) \sim \kappa^{I/2} \exp\left\{ -\frac{\kappa}{2} \sum_{i \sim i'} w_{ii'}(u_i - u_i')^2 \right\}, \tag{5.14}$$

where $\kappa$ has a role of a smoothing parameter and $w_{ii'}$ is a weight relating the neighbouring areas $i$ and $i'$. The idea of the prior (5.14) is to penalize large deviations of relative risk estimates between neighbouring areas. In the paper by Besag *et al.* [127], another joint improper distribution defining the random fields is presented, too.

Compared to the auto-Poisson model (5.9), one of the strengths of the hierarchical Markov model (5.12) in the Bayesian approach is that conditioning by the relative risk the data are controlling for scale patterns and the random field serves a local smoother.

A difficulty with the hierarchical Markov model is that it is usually unclear how to choose the weights $w_{ii'}$, the neighbourhood $i \sim i'$ and the smoothing parameter, $\kappa$, in the prior (5.14). This is especially true in the case of sparse data with empty areas. The most general solution to define the neighbourhood is to take the areas sharing a common boundary (e.g., [19],[127]). This is reasonable if the areas are of similar size (as the squares based on metric coordinates of SMASH). An alternative solution is to use the distance between either area or population centroids in the definition. The definition of weights is usually dependent on the definition of the neighbourhood. If the distance-based neighbourhood definition is exploited, the weights may naturally be based on the distance between the neighbouring areas. The definition of smoothing is instead quite difficult, because spatial data often have a small amount of information concerning the level of smoothing [128].

We applied a version (5.12) of the hierarchical Markov model [II],[III]. We also exploited the prior (5.14) with the weights equal to one [II],[III]. In Article [II], we set a hyperprior for the smoothing parameter $\kappa$. Instead, in Article [III], we treated $\kappa$ as fixed, as did Besag *et al.* [127].

### 5.3.5 Constrained hierarchical Markov model

According to an empirical evaluation [120], the hierarchical Markov model is rather robust and performs well in comparison with its competitors. In addition, the hierarchical Markov model (5.13) and its extensions are widely applied in disease mapping (see, e.g., the books by Elliott *et al.* [3] or Lawson *et al.* [4] or [5],[129]–[137]). Its exploitation with spatial small area data, especially with high resolution data, may be labourious. The posterior computing by means of the Markov chain Monte Carlo (MCMC) simulation method (e.g., [138],[139]) may be instable [II]. This feature, lack of convergence, has been noticed also elsewhere [128].

Several possibilities are available to overcome the problem of instability in the computing of the posterior. One possibility is to *increase the information*. One way to increase the information is to *enlarge the neighbourhood* used in the autoregressive prior ($i \sim i'$ in 5.14). We first defined the neighbourhood as in the most common cases in the literature, the adjacent areas defined the neighbourhood of an area [II]. In our case we had squares, so eight neighbouring squares were selected. With high resolution data there were several isolated squares, so the enlargement of the neighbourhood consisting of twenty squares was needed. Ranta and Penttinen [137] applied an adaptive enlargement of the neighbourhood. The neighbourhood was enlarged to cover at least three inhabited areas. This adaptive method may be appropriate in disease mapping but not in point source studies, because in these studies the distance from the source is the main objective of inference and the nearest neighbourhood should be equal sized. Another way to increase the information is to apply *the joint modelling* of two or more outcomes if available. Knorr-Held and Best [140] exploited the shared-component model in studying relative risks of oral cavity and oesophageal cancer for males in Germany.

Another possibility to overcome the problem of instable simulation is to *improve the computing algorithm*. The algorithm applied in Articles [II] and [III] was tailored in language C (see, e.g., [141]), because our version of the hierarchical Markov model model was too complex to be calculated with the popular Bayesian software WinBUGS [142]. Because the writer of the algorithm was not a professional C programmer, the algorithm may not be as sophisticated as possible. Our algorithm was a modification of single site Metropolis updating, which may also suffer from low mixing. More clever and better mixing algorithms may be possible. For example, Knorr-Held and Rue [143] introduced an algorithm based on *block updates*. They studied the low mixing of MCMC in connection of the hierarchical Markov model advocating simultaneous block updating of the parameters and hyperparameters. Breslow *et al.* [144] have proposed to use approximation methods instead of MCMC simulation.

The improvement of the convergence of the MCMC algorithm may be possible through *controlling the posterior expectations of estimates by the priors*. The auto-regression prior has been defined in an improper way, which may be problematic with high resolution data. The computational problem in connection with the hierarchical Markov model (5.12) has usually been linked with the problem of the smoothing level. A solution proposed is to choose the prior in a proper way, (see, e.g., [130]). We had to set a quite *informative prior* for the smoothing parameter [II].

One possibility to overcome the problem of instability is to restrict the model. In the Bayesian framework this can be formulated by *constraining the priors*. We suggested three constraining methods in the estimation of the relative risk in three sub-areas [III]. The idea of the first method was to fix the relative risk estimate in the farthest sub-area. An interpretation of this constraining method is similar to the typical regression model where the sub-area with a fixed effect is a control and the other areal risks are compared with this control. The idea of the second method was to order the sub-area specific risk estimates a priori. The idea of the third method was to relax the previous ordering. We implemented Markov chain priors for the relative risks on the sub-area level.

### 5.3.6 Change-point model

One simple modification of the Poisson model is a hierarchical one with a change-point prior (see, e.g., [145]–[147]) for the relative risk. The change-point model is a special case of partition models (e.g., [148],[149]). The observed number of cases is modelled by

$$o_i|\lambda, e, s, k \sim \text{Poisson}(e_i\lambda_{s_k}), \tag{5.15}$$

where $k$ is the number of change-points and $s$ is the positions of the changes. In this model, unknown parameters are the number of change-points, their locations, and the relative risk between them. The relative risk is assumed to be a piecewise constant function of the distance from the source. We treated the number of change points and their locations as nuisance parameters (see, e.g., [16]), because they were not of main importance [III]. When MCMC simulation is applied in the estimation of the posterior for the unknowns, the result the estimate of the relative risk is a smooth curve (see, e.g., [150],[151]). This is a result from a sample of marginal posterior distribution with varying estimate for the location of change-points.

# 6 Empirical findings

## 6.1 Statistical prelude

The methodological suggestions to solve the main statistical problems, spatial autocorrelation and sparseness of data, were considered in the previous section. In addition to these, the change-point model offers also a solution to the problem of the determination of the study area.

To determine a small enough study area is not a problem, because the change-point model estimates the high risk areas. Due to the random number of change-points, the high risk areas can be found even if the trend is fading away with increasing distance.

In addition, the investigator cannot interfere in the sub-area selection within the study area. The determination of the sub-area small enough to confirm a priori presumptions is not possible because the model defines the high risk areas. To determinate the sub-area large enough is neither a problem, because the model smooths the estimate of the relative risk in the areas with low expected counts.

## 6.2 Epidemiological findings

In what follows, the findings of the development of epidemiological methods are briefly discussed. In the next section the empirical results will be compared.

During the writing process of Article [V], we compared the results provided by direct and indirect standardization. The unpublished result was that direct and indirect standardization gave very similar results in this particular case.

In Article [I], we compared the effect of reference areas on estimates of relative risks. In the case study concerning Sulkava, SIR was closer to one when the reference area was the district of Kuopio University Hospital (East-Finland) or all the rural municipalities than when the reference area consisted of the whole of Finland. Also for the Myllypuro data, the known urban-rural differences and the differences between different parts of Finland in cancer rates appeared. SIR was highest when compared with the whole of Finland, but it became closer to one when the reference was chosen to be the district of Helsinki University Hospital (South-Finland) or the old towns. Finally, SIR was almost one when the reference was the town of Helsinki.

In Article [I], we explored the effect due to adjusting for the socioeconomic status. In a rural municipality of Sulkava, adjusting for the socioeconomic status had no important effect on SIRs. During the writing process of Article [V], we also explored this effect. The unpublished results showed no important differences whether the adjusting was applied or not. Instead, at Myllypuro with the dynamic cohort adjusting for the socioeconomic status gave higher SIR, 1.20 (1.03–1.39), than without adjustment, 1.06 (0.92–1.23).

At the moment, no evidence is available concerning the comparison of the use of area level data in SMASH and individual data. In Article [I], we compared the results given by SMASH with the results of individual level study by Pukkala and Pönkä [112]. The results of comparison are only suggestive, because the population at risk was defined in a different way and the study periods were not the same. For total cancer, SIR unadjusted for the socioeconomic status in the individual study was higher, 1.20 (0.97–1.48), than by SMASH, 1.06 (0.92–1.23). By chance, with adjustment for the socioeconomic status in SMASH, the SIRs were equal, 1.20 (1.03–1.39), as in the individual level study without adjusting.

If the use of register data is evaluated for modelling, our experience is that with the currently available data in SMASH not much more can be achieved. The hierarchical Markov model is complex enough, possibly too complex, for analysing the health data in SMASH [II],[III]. Instead, additional data (available in Finnish registers) may be valuable for more advanced modelling of register data.

## 6.3 Empirical findings from the comparison of methods

As an application, a classical point source example was used [II]–[IV]. The relative risk of lung cancer around the former asbestos mine at Paakkila in the eastern part of Finland was estimated. The exposure to asbestos results in an increased risk of lung cancer [152]. Asbestos mining continued between the years 1918–75 at Paakkila. Therefore, a high incidence of lung cancer cases was expected, which would be good for our purposes in the development of methodology.

*(i) Fixed sub-areas*

When the relative risk was estimated by the Bayesian modelling of SIRs [II], the empirical Bayesian modelling of SIRs [II], and the constrained hierarchical

Markov model [III], the whole study area (square of size 50 km × 50 km) was divided into three fixed sub-areas based on the assumed exposure levels without a prior knowledge of the spatial distribution of lung cancer cases. The first area beside the mine was a 4 km × 4 km square. It was not centered at the mine, because the prevailing wind direction was from the south or from the south-west. The bandwidth beside the main roads from the mine excluding the first area formed the second area. The exposure in this area was supposed to be higher than in the rest of the study area, because of the transportation of asbestos away from the mine. The rest of the study area was supposed to be the least exposed area.

The number of inhabited 500 m × 500 m squares was 2 051 (21% of the total). The number of inhabitants varied from 1 to 782 in those squares, whereas the total population count was 19 825 (194 beside the mine, 2 124 in the second area, and 17 507 in the rest of the area) in the whole area in 1980. The number of squares with cancer cases was 138 (1%), with 1–5 cases in each. The total number of lung cancer cases diagnosed between 1981–97 was 184 (6 in the first, 24 in the second, and 154 in the third area).

Table 4: The estimated areal relative risks with 95% intervals given by the maximum likelihood method (SIR), the full Bayesian model (FB), the empirical Bayesian model (EB), the model constrained with fixed area effect in $A_3$ (FE), the model constrained with the strict order condition (SOC) and the model with relaxed order condition (ROC). The interval of SIR is an equal tail 95% confidence interval, and the Bayesian intervals are equal tail 95% credible intervals.

| Model | Area nearest to mine | | Second area | | Farthest area | |
|---|---|---|---|---|---|---|
| | Risk | 95% interval | Risk | 95% interval | Risk | 95% interval |
| SIR | 3.80 | 1.39 to 8.28 | 1.44 | 0.92 to 2.13 | 1.20 | 1.02 to 1.41 |
| FB | 3.80 | 1.40 to 7.39 | 1.44 | 0.92 to 2.07 | 1.20 | 1.02 to 1.40 |
| EB | 3.13 | 1.38 to 5.62 | 1.48 | 0.97 to 2.11 | 1.21 | 1.03 to 1.41 |
| FE | 3.47 | 0.85 to 9.77 | 1.55 | 1.03 to 2.42 | 1.20 | † |
| SOC | 4.13 | 2.00 to 8.24 | 1.50 | 1.17 to 1.97 | 1.13 | 1.01 to 1.27 |
| ROC | 3.72 | 1.89 to 9.01 | 1.47 | 1.11 to 2.32 | 1.22 | 1.06 to 1.42 |

† The area effect was fixed to 1.20, hence no confidence interval was calculated.

The estimates of the parameters based on the three exposure areas are shown in Table 4. Maximum likelihood estimates (SIRs) were calculated as reference point estimates. The uncertainty of the SIRs is expressed in terms of the equal-tail 95% confidence interval. The results concerning Bayesian estimators are expressed in terms of posterior medians instead of means due to the skewness of the distributions. The uncertainty is described by equal-tail 95% credible intervals. These Bayesian intervals are strictly comparable with each other but not with the confidence interval of the SIRs. The Bayesian interval can be regarded as a probability interval, while the (frequentistic) confidence interval can be interpreted only in relation to repeated sequence of similar inferences.

The SIR corresponding to the area nearest to the mine was the highest, SIR=3.80, with a wide confidence interval, (1.39–8.28), due to lowest risk population counts and consequently lowest number of cases. The estimates given by the fully Bayesian model and the relaxed order restricted constraining method showed similar high risks for this area, above 3.70, with quite wide 95% intervals. In the empirical Bayesian model, the estimate was smaller but had also a narrower interval (1.38–5.62) than the estimate given by the fully Bayesian model. This is typical for the empirical Bayesian method due to the shrinkage effect [20]. Also the constraining method with one fixed area effect gave a slightly smaller estimate, 3.47, with the widest 95% interval (0.85–9.77). On the contrary, the strict order restricted constraining method produced a slightly higher relative risk estimate, 4.13. The 95% intervals were smaller using the strict order restriction on the risk estimates (1.89–8.03).

For the second area, SIR was 1.44 (0.92–2.13). The fully Bayesian model, the empirical Bayesian model and the model with relaxed order condition produced risk estimates of the same size. The 95% interval of the last mentioned model was widest (1.11–2.32) among these. Other constrained methods gave higher estimates, above 1.50. Again the intervals were smallest using the strict order restriction, (1.17–1.97).

For the rest of the area, SIR was 1.20 (1.02–1.41). Now, the fully Bayesian model, the empirical Bayesian model and the model with relaxed order condition produced the same risk estimates. Instead, the model with strict order condition produced a smaller risk estimate, 1.13. The 95% intervals were quite similar among all of the Bayesian methods, around 1.00 to 1.40.

*(ii) Random sub-areas*

When the relative risk around the mine was estimated with the change-point model, the study area was divided into 47 nested zones by circles with radius of 1.5, 2.0,. . . ,24.0 and 24.5 km [IV]. All the zones in the study area were inhabited. The range of inhabitants was from 15 to 1 911 in the zones, while

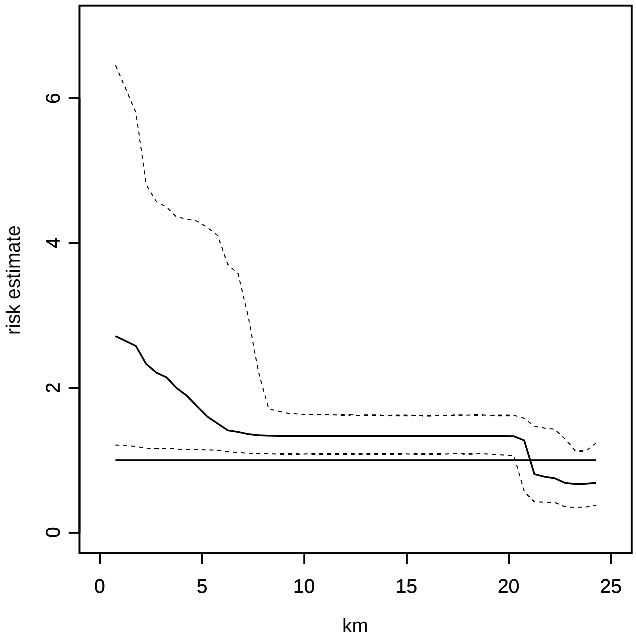

Figure 1: Estimated relative risk of lung cancer with the increasing distance from the former asbestos mine. The solid line describes the posterior median of relative risk, and the dotted lines the 95% credible interval. The straight line describes the reference, relative risk = 1.00.

the total number of inhabitants was 17 194 in the whole study area. The number of zones with cancer cases was 41, with 1 to 19 cancer cases in each (153 cancer cases in the whole area).

The results concerning Bayesian estimators are expressed in terms of posterior medians. The uncertainty is described by an equal-tail 95% credible interval. Because the idea of the change-point method is to let the model show the number and the location(s) of the change-points, we let the number vary according to the design starting from 0. We concluded that the reasonable number of change-points was two. The smooth curve of the Bayesian estimate of the risk is shown in Figure 1.

The estimate near the asbestos mine was 2.44 (1.17–6.07). At the farthest zone the risk estimate was 0.77 (0.41–1.52).

# 7 Computational issues and software

During the process of this thesis the following softwares were used. SMASH software was used in the routine analyses. It runs under ArcView GIS 3.2 for Windows running under Microsoft Windows NT 4.0 Workstation on desktop computer [I],[V]. The avenue code for SMASH software on ArcView GIS was written in Geological Survey of Finland in Kuopio. The software and data are stored in the password protected directories on Windows NT, and on OpenVMS 7.3 server of the National Public Health Institute, Helsinki. Data management (checking and cleaning) was carried out mainly using SAS (versions 6.0–8.2) for VMS and SAS 8.02 for Windows. Further data management (integrating) was carried out with ArcView GIS. In the modelling we used WinBUGS 1.3 running under Windows NT [II],[IV], SAS for Windows [V], and Sun WorkShop 6 update 2 C 5.3 C-compiler running under Solaris 8 4/01 on Sun Enterprise 250 server of the University of Jyväskylä for self-written C code [II],[III]. In the visualization we used ArcView GIS, and S-PLUS (versions 3.3–2000) for Windows. This text is produced with WinEdt 5 and MiKTeX 2.3 running under Windows NT.

# 8 Discussion

## 8.1 Methodological choices

Models for analysing the regional variation of cancer risk around a putative source of pollution were suggested. In the following the findings in the development of methodology are discussed.

### 8.1.1 Models for estimation of disease risk around a source of pollution

Let us consider first modelling of cancer risk in fixed sub-areas obtained as a result of aggregation from small areas, which were the smallest spatial units available. In the simplest case, if the spatial autocorrelation can be assumed to be weak within sub-areas, a Poisson regression model with regional effect (5.7) can be exploited. In the case of the river Kymijoki, we estimated the sub-area specific relative risks with this model [V].

As a more sophisticated model, we applied the hierarchical Markov model (5.12). The effect of the point source was considered again through the fixed regional components ($x_k$ in (5.12)). However, the hierarchical Markov model turned out to be inconsistent with sparse high resolution data to some degree. The posterior computation was unstable [II]. The MCMC samples in the simulation of the posterior distribution did not converge in a proper way, and we could not reach identifiable estimates for the relative risks. We suggested three constraining methods, which all stabilized the posterior simulation and resulted in identifiable estimates of the relative risk [III].

Finally, we suggested a simple extension of the Poisson regression model based on a change-point prior (5.15) [IV]. In the case of the change-point model the effect of the point source was estimated in the random areas around the source.

A general advantage of the use of Bayesian models is that the Poisson assumption is usually valid conditionally. Another general advantage of Bayesian models is that prior knowledge about the subject can be used through prior distributions, if available. Also, the MCMC simulation of the Bayesian model is a convenient tool for inference on the behaviour of the relative risk, because the Bayesian interval is an actual probability interval. This simplifies the testing of the hypothesis like "the relative risk in area A is higher than in area B".

The hierarchical Markov model with constraining seemed to be superb if the comparison of the methods is based on the capability to solve the main problems, spatial autocorrelation and sparseness of data (discussed in the next two subsections). The cost in the use of the model is the demanding and time-consuming calculation. The hierarchical Markov model and its extensions are extremely developed. Because the high resolution data set restrictions, the development of the complex models may be useless if the models are more complex than data.

The change-point model is an alternative to the utilization of tests for the linear trend in the risk [7],[22],[24],[40] and the calculation of SIRs. The change-point model assess simultaneously the estimate and the trend of the relative risk. In addition, the change-point model detects the possible non-linearity in risk and increase or decrease in risk with increasing distance from the source.

### 8.1.2 Consideration of spatial autocorrelation

In the case of the Poisson regression model with fixed regional effects, the possible spatial autocorrelation is somehow considered through the fixed regional effects and through the adjustment for the covariates.

The possible spatial autocorrelation was estimated by calculating Moran's $I$ statistics (2.4) in the example case of Articles [II],[III] and [IV]. There were high correlations even between squares quite far away (20 km) from each other. The incidence of lung cancer in rural area around the former asbestos mine proved to be spatially autocorrelated.

In order to allow for the possible existence of spatial autocorrelation, we first exploited our version of the hierarchical Markov model. We had to set some restrictions to reach identifiable results. The main strength of the hierarchical Markov model was the ability to consider spatial autocorrelation. Another strength of the model was that it could simultaneously estimate the relative risk in the small areas and in the fixed sub-areas.

In a change-point model the spatial autocorrelation is not taken into account at the level of aggregation to concentric sub-areas. However, spatial autocorrelation between these sub-areas is considered through the model.

### 8.1.3 Smoothing sparse data

The Poisson regression model with fixed regional effects compensates the sparseness of data by smoothing. The covariates describe the spatial effect. In the estimation of these covariates the whole data in the study area are applied.

When applying the hierarchical Markov model, the smoothing effect comes from two sources. First, the spatially structured effect smooths locally from the neighbouring small areas, even over the boundaries of the sub-areas. Secondly, the modelling in the Bayesian framework smooths globally the observed and the expected number of cases. The local smoothing is suitable for point source studies.

The change-point model can also be considered as a smoother. The smoothing effect comes from the assumption that the risk between change-points is constant, that is, equal in adjacent annuli not separated by a change-point. The model with such an assumption is of course an extreme simplification, but model averaging results in a smooth curve.

Altogether, the possible problems of over-smoothing should be recognized. With over-smoothing possibly remarkable findings may be diluted or erased.

### 8.1.4 Validation of epidemiological methods

Along with the case studies analysed with SMASH, the knowledge on limitations and possibilities has increased notably. The comparison of standardizing methods has encouraged to continue applying indirect standardization, although a typical problem, hybrid adjustment, was realized and should be noted. The importance of the selection of the appropriate reference area, being a major confounding factor, was also realized. Although in the case studies the adjusting for the socioeconomic status had no important effect, the adjustment should be decided case by case, as in some situations it may be very important.

## 8.2 SMASH compared to RIF

The Small Area Health Statistics Unit (SAHSU) [11], Imperial College London, has been developing a system quite similar to SMASH called Rapid Inquiry Facility (RIF) [12]. SAHSU has powerful facilities for the development of tools and methodology for small area analyses. SAHSU was established by the Government in 1983 and has currently a staff of 10 members (1 professor and 6 doctors).

### 8.2.1 Comparison of data

RIF holds national cause-specific data on deaths, births, cancers, hospital admissions and congenital anomalies at a postcode level. Instead, population data and socioeconomic data are available at the enumeration districts level. Among these, RIF also holds a range of geographical and environmental data. The amount of data is much more massive than in SMASH.

Although occupation and education explain quite well the social differences in Finland, the division of socioeconomic classes based on these may not be good enough, because the adjustment showed no important effect in case studies [I],[V]. The British practice, considering effects such as size of household or flat owning, may improve the socioeconomic classification. Accordingly, it is recommended to estimate the relative risks both with and without adjustment for the socioeconomic status.

Data management within RIF is not without problems. Because data are available at different levels of aggregation, the extrapolation of data from one level to another is required. This may cause difficulties as reported by Briggs *et al.* [99]. Instead, aggregation is not a problem within SMASH, because all data are based on metric coordinates and small area units.

Another problem of British data is that administrative units are subject to change in the course of time. Instead, in SMASH using of grid based on metric coordinates as the spatial unit ensures the study area being always up-to-date. The boundaries of small areas do not change in time.

A further problem in RIF is that the accuracy of the determination of study areas suffers from the fact that the small areas used as a spatial unit are based on administrative boundaries. Large administrative units are likely to swamp the local effect. When the study area is defined by drawing a circle, also the rest of the administrative areas around the circle is included in the study area. In SMASH, the study area may be defined freely and reliably, where administrative boundaries do not restrict the area selection.

The nature of spatial high resolution of Finnish data sets some restrictions. The using of high resolution data may be problematic in sparsely populated areas, because the population counts get low. SMASH has empty and isolated small areas, which causes problems in applications. This is a contrast to the British postcode level data where all spatial units have inhabitants, 17 households on average.

In conclusion, SMASH has some strengths of its own. While learning the best characteristics of RIF, SMASH can be improved, for example, by including further health or environmental data or further data describing socioeconomic activity.

### 8.2.2 Comparison of software and methods

Due to the huge amount of data, RIF has been running under a network of three Sun Sparc servers [12]. At the present, SAHSU is developing a laptop version of RIF, which can utilize a part of the data loaded into the laptop. Instead, the data held by SMASH is stored on a desktop computer, under which the software is also running. If necessary, the data and software can be used also on a laptop.

The possibilities in the interpretation of results are more sophisticated within RIF. Due to the more versatile reporting characteristics, the background information can be applied more efficiently in the interpretation. For example, the distributions of sex, age and socioeconomic status can be easily produced. RIF can also produce statistically smoothed maps on disease and on socioeconomic status, which are not currently possible in SMASH.

RIF has the facility to utilize the methods using person-years, because of holding data on mortality. This is not possible in SMASH. The choice of the denominator of the risk ratios is more versatile in RIF, in addition to the population data, for example, the counts of births can be applied as a denominator.

Instead, RIF does not offer such a wide possibility to choose the reference area as SMASH does. The possibilities to choose the reference area based on the type of municipalities or on individual municipalities are not possible in RIF.

To avoid the problem of *post hoc* studies RIF uses the *a priori* standard "near" and "far" bands of 0–2 km and 2–7.5 km around the point source. These were selected arbitrarily and a useful compromise was achieved [12]. After all, it may be questionable to use that division into regions in every situation of environmental emergency. In SMASH, there is no standard definition of "near" or "far" areas. Instead, the change-point model offers one clever alternative solution to this problem through the modelling of location of risk areas.

Again, it can be concluded that by learning the best characteristics from methodology used in RIF, the SMASH can be improved further. For example, reporting the background information or producing the directly standardized estimates can be easily implemented also in SMASH.

## 8.3 Back to the future

SMASH can be applied as a rapid system for the analysis of the relative risk of any registered health outcome. It would be interesting to add, for example, data on birth outcomes or data on mortality into the system. The usefulness of SMASH may appear on its best in ecological investigations of congenital or childhood diseases where the place of residence is a better surrogate of exposure than with cancers.

Due to the lack of high quality exposure assessment the results given by SMASH are only preliminary. This means that no strong causal associations can be concluded. The adding of data on environmental exposures or modelling tools may further improve SMASH. The knowledge of all possible sources of pollution may help in the interpretation, for example, in a case of multiple sources. The tools for modelling dispersion or plumes can improve the exposure assessment, for example, in the case studies of air pollution.

Although the occupational and educational factors explain the most of the socioeconomic differences in Finland, the improvement of socioeconomic classification should be considered. For example, considering covariates such as data on income, living conditions (e.g., flat owning, size of household) or family conditions (e.g., marital status), may improve the use of socioeconomic classification as a confounding factor.

The reporting characteristics of SMASH are limited. The report in routine analyses produces only the selections made and the result. No background information, for example on covariates, for helping the interpretation of the result is produced. The implementation of RIF [12] along the EUROHEIS-project [13] will improve currently quite modest reporting in SMASH.

It is possible to create a link between the ArcView software, under which SMASH runs, and some other softwares, for example, statistical software S-PLUS. So, implementing the methods suggested in the present work within the SMASH software is not unthinkable. Another developmental possibility of applying the suggested methods is to sweeten the ArcView software such that the output data can be easily exploited in the statistical softwares for further analyses.

The validation of small area level analysis by comparing to individual level analyses is also on the list of future works. Reporting about the case studies may be on more solid ground if we have any idea how reliable the ecological studies are, although the level of reliability is case sensitive.

In the development of statistical methods there are several interesting alternatives to improve the methodology related to SMASH. For example, point processes (see, e.g., Diggle [153] or Møller and Waagepetersen [154]) or the shared-component model (see, e.g., Knorr-Held and Best [140]) may be plausible methods to estimate the relative risk at the small area level around a point source.

## 8.4 Concluding remarks

SMASH with high resolution data is useful in the rapid estimation of the relative risk around a putative source of pollution. The strength is the possibility to define the study area freely with an accuracy of 500 metres. If data tend to be sparse, the classical methods like SIRs or Poisson (regression) models, may give incorrect and uncertain estimates. This can be overcome by using smoothing methods. If strong spatial autocorrelation is in doubt, the classical methods may give incorrect results, and some more sophisticated methods are needed in further investigations. In this work some useful statistical methods were suggested. After all, it should be in mind that the small area level estimates are only preliminary. If small area studies imply the possible excess risk, further more detailed individual level studies are always needed.

# References

[1] Bender AP, Williams AN, Johnson RA, Jagger HG. Appropriate public health responses to cluster: the art of being responsibly responsive. *American Journal of Epidemiology*, 1990; 132: S48–S52.

[2] Guidotti TL, Jacobs P. The implications of an epidemiological mistake: A community's response to a perceived excess cancer risk. *American Journal of Public Health*, 1993; 83: 233–239.

[3] Elliott P, Wakefield JC, Best NG, Briggs DJ (Eds.). *Spatial Epidemiology - methods and applications.* Oxford University Press, Oxford, 2000.

[4] Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R (Eds.). *Disease mapping and risk assessment for public health.* Wiley, Chichester, 1999.

[5] Richardson S. Statistical methods for geographical correlation studies. In Elliott P, Cuzick J, English D, Stern R (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press: Oxford, 1992; pp. 231–237.

[6] Richardson S, Monfort C. Ecological correlation studies. In [3]; pp. 205–220.

[7] Bithell JF. Statistical methods for analysing point-source exposures. In Elliott P, Cuzick J, English D, Stern R, (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press: Oxford, 1992; pp. 221–230.

[8] Morris SE, Wakefield JC. Assessment of disease risk in relation to a pre-specified source. In [3]; pp. 153–184.

[9] Alexander FE, Boyle P (Eds.). *Methods of investigating localised clustering of disease.* IARC, Lyon, 1996.

[10] Lawson AB, Denison DGT (Eds.). *Spatial cluster modelling.* Chapman & Hall/CRC: Boca Raton, 2002.

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

[12] Aylin P, Maheswaran R, Wakefield J, Cockings S, Järup L, Arnold R, Wheeler G, 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: 89–98.

[13] Cockings S, Järup L. A European health and environment information system for exposure and disease mapping and risk assessment. In Briggs DJ, Forer P, Järup L, Stern R (Eds.). *GIS for Emergency Preparedness and Health Risk Reduction.* Kluwer Academic Publishers, Dordrecht, 2002; pp. 207–226.

[14] Heineman EF. A GIS for researchers and the community, from the Long Island reast cancer study project. *Epidemiology*, 2002; 13: 724.

[15] Wendt RD, Hall HI, PriceGreen PA, Dhara VR, Kaye WE. Evaluating the sensitivity of hazardous substances emergency events surveillance - A comparison of three surveillance systems. *Journal of environmental health*, 1996; 58: 13–17.

[16] Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis.* Chapman & Hall: London, 1995.

[17] Congdon P. *Bayesian statistical modelling.* Wiley, Chichester, 2001.

[18] Congdon P. *Applied Bayesian modelling.* Wiley, Chichester, 2003.

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

[20] Carlin BP, Louis TA. *Bayes and empirical Bayes methods for data analysis.* Chapman & Hall: London, 1996.

[21] Valkonen T, Koskinen S, Martelin T (Eds.). *Rekisteriaineistot yhteiskunta- ja terveystutkimuksessa.* Gaudeamus, Helsinki, 1998. [In Finnish]

[22] Stone RA. Investigations of excess environmental risks around putative sources: statistical problems and proposed test. *Statistics in Medicine*, 1988; 7: 649–660.

[23] Hills M, Alexander F. Statistical methods used in assessing of the risk of disease near a source of possible environmental pollution: a review. *Journal of the Royal Statistical Society, Series A*, 1989; 152: 353–363.

[24] Bithell JF, Stone RA. On statistical methods for analysing the geographical distribution of cancer cases near nuclear installations. *Journal of Epidemiology and Community Health*, 1989; 43: 79–85.

[25] Diggle PJ. A point process modelling approach to raised incidence of a rare phenomen in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 1990; 153: 340–362.

[26] Urquhart J. The investigation of leukaemia incidence around sites of special interest. *Nuclear Energy*, 1991; 30: 21–26.

[27] Hills M. Some comments on methods for investigating disease risk around a point source. In Elliott P, Cuzick J, English D, Stern R (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press: Oxford, 1992; pp. 231–237.

[28] Lawson AB. On the analysis of mortality events around a prespecified fixed point. *Journal of the Royal Statistical Society, Series A*, 1993; 156: 363–377.

[29] Lawson AB, Williams FLR. Armadale: A case-study in envronmental epidemiology. *Journal of the Royal Statistical Society, Series A*, 1994; 157(2): 285–298.

[30] Bithell JF. The choice of the test for detecting raised disease risk near a point source. *Statistics in Medicine*, 1995; 14: 2309–2322.

[31] Diggle P, Elliott P. Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data. *Journal of Epidemiology and Community Health*, 1995; 49 (Suppl 2): S20–S27.

[32] Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research*, 1995; 4: 137–159.

[33] Lawson AB, Waller, LA. A rewiev of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics*, 1996; 7: 471–487.

[34] Lawson AB, Biggeri A, Williams FLR. A rewiev of modelling approaches in health risk assessment around putative sources. In [4]; pp. 231–245.

[35] Wakefield JC, Morris SE. The Bayesian modeling of disease risk in relation to a point source. *Journal of the American Statistical Society*, 2001; 96 : 77–91.

[36] Black D. *Investigation of the possible increased incidence of cancer in West Cumbria.* HMSO, London, 1984.

[37] Michaelis J, Keller B, Haaf G, Kaatsch P. Incidence of childhood malignancies in the vicinity of West German nuclear power plants. *Cancer Causes and Control*, 1992; 3: 255–263.

[38] Selvin S, Schulman J, Merrill DW. Distance and risk measures for the analysis of spatial data: a study of childhood cancers. *Social Science & Medicine*, 1992; 34 (7): 769–777.

[39] Bhopal RS, Phillimore P, Moffatt S, Foy C. Is living near a coking works harmful to health? A study of industrial air pollution. *Journal of Epidemiology and Community Health*, 1994; 48: 237–247.

[40] Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. *British Medical Journal*, 1994; 309: 501–505.

[41] Lyons RA, Monaghan SP, Heaven M, Littlepage BNC, Vincent TJ, Draper GJ. Incidence of leukaemia and lymphoma in young people in the vicinity of the petrochemical plant at Baglan Bay, South Wales, 1974 to 1991. *Occupational and Environmental Medicine*, 1995; 52: 225–228.

[42] Sans S, Elliott P, Kleinschmidt I, Shaddick G, Pattenden S, Walls P, Grundy C, Dolk H. Cancer incidence and mortality near the Baglan Bay petrochemical works, South Wales. *Occupational and Environmental Medicine*, 1995; 52: 217–224.

[43] Viel J-F, Pobel D, Carré A. Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis. *Statistics in Medicine*, 1995; 14: 2459–2472.

[44] Elliott P, Shaddick G, Kleinschmidt I, Jolley D, Walls P, Beresford J, Grundy C. Cancer incidence near municipal solid waste incinerators in Great Britain. *British Journal of Cancer*, 1996; 73, 702–710.

[45] Sharp L, Black RJ, Harkness EF, McKinney PA. Incidence of childhood leukaemia and non-Hodgkin's lymphoma in the vicinity of nuclear sites in Scotland, 1968–93. *Occupational and Environmental Medicine*, 1996; 53: 823–831.

[46] Michelozzi P, Fusco D, Forastiere F, Ancona C, Dell'Orco V, Perucci CA. Small area study of mortality among people living near multiple sources of air pollution. *Occupational and Environmental Medicine*, 1998; 55: 611–615.

[47] Harrison RM, Leung P-L, Somervaille L, Smith R, Gilman E. Analysis of incidence of childhood cancer in the West Midlands of the United Kingdom in relation to proximity to main roads and petrol stations. *Occupational and Environmental Medicine*, 1999; 56: 774–780.

[48] López-Abente G, Aragonés N, Pollán M, Ruiz M, Gandarillas A. Leukaemia, lymphomas and myeloma mortality in the vicinity of nuclear power plants and nuclear fuel facilities in Spain. *Cancer Epidemiology, Biomarkers & Prevention*, 1999; 8: 925–934.

[49] Sharp L, McKinney PA, Black RJ. Incidence of childhood brain and other non-haematopoetic neoplasms near nuclear sites in Scotland, 1975–94. *Occupational and Environmental Medicine*, 1996; 53: 823–831.

[50] Dickinson HO, Hammal DM, Dummer TJB, Parker L, Bithell JF. Childhood leukaemia and non-Hodgkin's lymphoma in relation to proximity railways. *British Journal of Cancer*, 2003: 88, 695–698.

[51] Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Smith DF. Childhood cancer incidence rates and hazardous air pollutants in California: An exploratory analysis. *Environmental Health Perspectives*, 2003; 111: 663–668.

[52] Hastie TJ, Tibshirani RJ. *Generalized additive models.* Chapman & Hall, London, 1990.

[53] Draper GJ, Parkin DM. Cancer incidence data for children. In Elliott P, Cuzick J, English D, Stern R (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press: Oxford, 1992; pp. 63–71.

[54] dos Santos Silva I. *Cancer epidemiology: Principles and methods.* IARC, Lyon, 1999.

[55] Huang Y-L, Batterman S. Residence location as a measure of environmental exposure: a review of air pollution epidemiology studies. *Journal of Exposure Analysis and Environmental Epidemiology*, 2000; 10: 66–85.

[56] Verkasalo PK, Pukkala E, Hongisto MY, Valjus JE, Järvinen PJ, Heikkilä KV, Koskenvuo M. Risk of cancer in Finnish children living close to power lines. *British Medical Journal*, 1993; 307: 895–899.

[57] Trepka MJ, Heinrich J, Krause C, Schultz C, Lippold U, Meyer E, Wichmann H-E. The internal burden of lead among children in a smelter town - a small area analysis. *Environmental Research*, 1997; 72: 118–130.

[58] Cheng KK, Day NE, Cartwright R, Craft A, Birch JM, Eden OB, McKinney PA, Peto J, Beral V, Roman E, Elwood P, Alexander FE, Chilvers CED, Doll R, Greaves M, Goodhead D, Fry FA, Adams G, Gilman E, Skinner J, Williams D, Deacon J, Mott M, Muir K, Law G, Simpson J. Childhood cancer and residentialproxomity to power lines. *British Journal of Cancer*, 2000; 83 (11): 1573–1580.

[59] Reynolds P, Von Behren J, Gunier RB, Goldberg DE, Hertz A, Smith D. Traffic patterns and childhood cancer incidence rates in California, United States. *Cancer Causes & Control*, 2002; 13 (7): 665–673.

[60] Rom WN (Ed.). *Environmental and occupational medicine.* Little, Brown and Company, Boston, 1983.

[61] Yassi A, Kjellström T, de Kok T, Guidotti TL. *Basic environmental health.* Oxford University Press, Oxford, 2001.

[62] Kousa A, Monn C, Rotko T, Alm S, Oglesby L, Jantunen MJ. Personal exposures to $NO_2$ in *EXPOLIS*-study: relation to residential indoor outdoor and workplace concentrations in Basel, Helsinki and Prague. *Atmospheric Environment*, 2001; 35: 3405–3412.

[63] Pearson RL, Wachtel H, Ebi KL. Distance weighted traffic density in proximity to a home is a risk factor for leukaemia and other childhood cancers. *Journal of the Air & Waste Management Association*, 2000; 50 (2): 175–180.

[64] Boudet C, Zmirou D, Poizeau D. Fraction of PM2.5 personal exposure attributable to urban traffic: A modeling approach. *Inhalation Toxicology*, 2000; 12: 41–53, Suppl. 1.

[65] Heinonen-Guzejev M, Vuorinen HS, Kaprio J, Heikkila K, Mussalo-Rauhamaa H, Koskenvuo M. Self-report of transportation noise exposure, annoyance and noise sensitivity in relation to noise map information. *Journal of Sound and Vibration*, 2000; 234 (2): 191–206.

[66] Morgenstern H. Ecologic studies. In [88]; pp. 459–480.

[67] Elliott P, Wakefield JC. Bias and confounding in spatial epidemiology. In [3]; pp. 68–84.

[68] Biggeri A, Divino F, Frigessi A, Lawson AB, Böhning D, Lesaffre E, Viel J-F. Introduction to spatial models in ecological analysis. In [4]; pp. 181–192.

[69] Best N. Bayesian ecological modelling. In [4]; pp. 193–201.

[70] Wakefield J, Elliott P. Issues in the statistical analysis of small area health data. *Statistics in Medicine*, 1999; 18: 2377–2399.

[71] Greenland S. Ecologic versus individual sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 2001; 30: 1343–1350.

[72] Wakefield J. Sensitivity analyses for ecological regression. *Biometrics*, 2003; 59: 9–17.

[73] Hakama M, Hakulinen T, Pukkala E, Saxén E, Teppo L. Risk indicators of breast and cervical cancer on ecological and individual levels. *American Journal of Epidemiology*, 1982; 116: 990–1000.

[74] Lagarde F, Pershagen G. Parallel analyses of individual and ecologic data on residential radon, cofactors and lung cancer in Sweden. *American Journal of Epidemiology*, 1999; 149: 268–274.

[75] Rothman KJ, Greenland S. Precision and validity in epidemiologic studies. In [88]; pp. 115–134.

[76] Carstairs V. Socio-economic factors at areal level and their relationships with health. In [3]; pp. 51–67.

[77] Pukkala E. *Cancer risk by social class and occupation. A survey of 109,000 cancer cases among Finns of working age.* Contributions to Epidemiology and Biostatistics, Vol 7. Karger, Basel, 1995.

[78] Rotko T, Kousa A, Alm S, Jantunen M. Exposures to nitrogen dioxide in *EXPOLIS*-Helsinki: microenvironment, behavioral and sociodemographic factors. *Journal of Exposure Analysis and Environmental Epidemiology*, 2001; 11: 216–233.

[79] Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, Hakulinen T, Becker N, Stabenow R, Bjarnadottir K, Stengrevics A, Gurevicius R, Glattre E, Zatonski W, Men T, Barlow L. *Cancer Atlas of Northern Europe.* Cancer Society of Finland publication No. 62, Helsinki, 2001.

[80] Koivusalo M, Pukkala E, Vartiainen T, Jaakkola JJK, Hakulinen T. Drinking water chlorination and cancer - A historical cohort study in Finland *Cancer Causes & Control.* 1997; 8 (2): 192–200.

[81] Hellen H, Hakola H, Laurila T, Hiltunen V, Koskentalo T. Aromatic hydrocarbon and methyl tert-butyl, ether measurements in ambient air of Helsinki (Finland) using diffusive samplers. *Science of the Total Environment*, 2002; 298 (1-3): 55–64.

[82] Rothman KJ. *Modern epidemiology.* Little, Brown and Company: Boston, 1986.

[83] Benjamin B. *Demographic analysis.* George Allen and Unwin: London, 1968.

[84] Pollard AH, Yusuf F, Pollard GN. *Demographic techniques. 3rd edition.* Pergamon: Sydney, 1990.

[85] Miettinen OS. *Theoretical epidemiology. Principles of occurence research in medicine.* John Wiley & Sons: New York, 1985.

[86] Breslow NE, Day NE. *Statistical methods in cancer research. Vol. 2. The design and analysis of cohort studies.* IARC: Lyon, 1987.

[87] Estève J, Benhamou E, Raymond L. *Statistical methods in cancer research. Vol. 4. Descriptive epidemiology.* IARC: Lyon, 1994.

[88] Rothman KJ, Greenland S. (Eds.). *Modern epidemiology - 2nd edition.* Lippincott-Raven: Philadelphia, 1998

[89] Woodward M. *Epidemiology. Study design and data analysis.* Chapman & Hall: New York, 1999.

[90] Wolfenden HH. On the methods of comparing the mortalities of two or more communities, and standardization of death rates. *Journal of the Royal Statistical Society*, 1923; 86: 399–411.

[91] Yule GU. On some points relating to vital statistics, more especially statistics of occupational mortality. *Journal of the Royal Statistical Society*, 1934; 97: 1–84.

[92] Miettinen OS. Standardization of risk ratios. *American Journal of Epidemiology*, 1972; 96: 383–388.

[93] Pickle LW, White AA. Effect of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine*, 2001; 23: 40-46.

[94] Goldman DA, Brender JD. Are standardized mortality ratios valid for public health data analysis? *Statistics in Medicine*, 2000; 19: 1081–1088.

[95] Julious SA, Nicholl J, George S. Why do we continue to use standardized mortality ratios for small area comparisons? *Journal of Public Health Medicine*, 2001; 23: 40-46.

[96] Lee WC. Standardization using the harmonically weighted ratios: internal and external comparisons. *Statistics in Medicine*, 2002; 21: 247–261.

[97] Arnold RA, Diamond ID, Wakefield JC. The use of population data in spatial epidemiology. In [3]; pp. 30–50.

[98] Staines A, Järup L. Health event data. In [3]; pp. 15–29.

[99] Briggs DJ, de Hoogh C, Hurt C, Maitland I. *Geographical variations in populations living around landfill sites. SAHSU Report 2002.1*, Small Area Health Statistics Unit, Imperial College of Science, Technology and Medicine, London, 2001.

[100] Pekkanen J, Pearce N. Environmental epidemiology: Challenges and opportunities. *Environmental Health Perspectives*, 2001; 109: 1–5.

[101] Smans M, Estève J. Practical approaches to disease mapping. In Elliott P, Cuzick J, English D, Stern R (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press, Oxford, 1992; pp. 141–150.

[102] Wakefield JC, Kelsall JE, Morris SE. Clustering, cluster detection, and spatial variation in risk. In [3]; pp. 128–152.

[103] Moran PAP. The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B*, 1948; 10: 243–251.

[104] Wakefield JC, Best NG, Waller L. Bayesian approaches to disease mapping. In [3]; pp. 104–127.

[105] Tango T. Comparison of general tests for spatial clustering. In [4]; pp. 111–117.

[106] Jarup L, Best N. Editorial comment on Geographical differences in cancer incidence in the Belgian province of Limburg by Bruntix and colleagues. *European Journal of Cancer*, 2003; 39: 1973–1975.

[107] Simonoff JS. *Smoothing methods in statistics.* Springer, New York, 1996.

[108] Alexander F, Cuzick J. Methods for the assessment of disease clusters. In Elliott P, Cuzick J, English D, Stern R (Eds.). *Geographical and environmental epidemiology: methods for small-area studies.* Oxford University Press, Oxford, 1992; pp. 238–250.

[109] Diggle PJ. Overview of statistical methods for disease mapping and its relationship to cluster detection. In [3]; pp. 87–103.

[110] Greenland S, Rothman KJ. Fundamentals of epidemiologic data analysis. In [88]; pp. 201–229.

[111] Hertz-Picciotto I. Environmental epidemiology. In [88]; pp. 555–584.

[112] Pukkala E, Pönkä A. Increased incidence of cancer and asthma in houses built on a former dump area. *Environmental Health Perspectives*, 2001; 109: 1121–1125.

[113] Teppo L, Pukkala E, Hakama M, Hakulinen T, Herva A, Saxén E. Way of life and cancer incidence in Finalnd. A municipal-based ecological analysis. *Scandinavian Journal of Social Medicine*, 1980; Supplement 19.

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

[115] Maanmittauslaitos. *Suomen kartasto, Vihko 112, Suomen kartoitus.* Suomen maantieteellinen seura, Helsinki; 1984. [In Finnish]

[116] Statistics Finland. Evaluation study of census 1990. In *Population census, Vol. 9.* Statistics Finland, Helsinki, 1994.

[117] Teppo L, Pukkala E, Lehtonen M. Data quality and quality control of a population-based cancer registry. Experience in Finland. *Acta Oncologica*, 1997; 33: 365–369.

[118] Pukkala E, Teppo L. Too much leukaemia in Porvoo area? *Suomen Lääkärilehti*, 1992; 47: 3431–3433. [In Finnish]

[119] Rothman KJ, Greenland S. Measures of disease frequency. In [88]; pp. 29–46.

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

[121] Kelsall J, Diggle P. Spatial variation in risk of disease: a non-parametric binary regression approach. *Journal of the Royal Statistical Society, Series C*, 1998; 47: 559–573.

[122] Diggle PJ, Morris SE, Wakefield JC. Point-source modelling using mathced case-control data. *Biostatistics*, 2000; 1: 89–105.

[123] Winkler G. *Image analysis, random fields and dynamic Monte Carlo methods.* Springer, Berlin, 1995.

[124] Clayton DG. Generalized linear mixed models. In [138]; pp. 275–302.

[125] Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 1974; 36: 192–236.

[126] Ferrándiz J, López A, Sanmartín P. Spatial regression models in epidemiological studies. In [4]; pp. 203–215.

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

[128] Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine*, 2000; 19: 2279–2294.

[129] Clayton DG, Bernardinelli L, Montomoli C. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 1993; 22: 1193–1202.

[130] Besag JE, Green PJ, Higdon DM, Mengersen KL. Bayesian computation and stochastic systems (with discussion). *Statistical Science*, 1995; 10: 3–66.

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

[132] Best NG, Waller LA, Thomas A, Conlon E, Arnold RA. Bayesian models for spatially correlated disease and exposure data. In Bernardo J., Berger J., Dawid A., Smith A. (Eds.) *Bayesian Statistics 6.* Oxford University Press, Oxford, 1998. pp. 131–156.

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

[134] Langford IH, Leyland AH, Rasbash J, Goldstein H. Multilevel modelling of the geographical distributions of diseases. *Applied Statistics*, 1999; 48: 253–268.

[135] Osnes K, Aalen OO. Spatial smoothing of cancer survival: a bayesian approach. *Statistics in Medicine*, 1999; 18: 2087–2099.

[136] Sun D, Tsutakawa RK, Kim H, He Z. Spatio-temporal interaction with disease mapping. *Statistics in Medicine*, 2000; 19: 2015–2035.

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

[138] Gilks WR, Richardson S, Spiegelhalter DJ. *Markov chain Monte Carlo in practice.* Chapman & Hall: London, 1996.

[139] Robert CP, Casella G. *Monte Carlo Statistical Methods.* Springer & Verlag: New York, 1999.

[140] Knorr-Held L, Best NG. A shared component model for detecting joint and selective clustering of two diseases *Journal of the Royal Statistical Society, Series A*, 2001; 164, 73–85.

[141] Kelley A, Pohl I. *A book on C. Programming in C. Fourth edition.* Addison–Wesley: Reading, 1998.

[142] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility *Statistics and Computing*, 2000; 10: 325–337.

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

[144] Breslow N, Leroux B, Platt R. Approximate hierarchical modelling of discrete data in epidemiology. *Statistical Methods in Medical Research*, 1998; 7: 49–62.

[145] Raftery AE, Akman VE. Bayesian analysis of a poisson process with a change-point. *Biometrika*, 1986; 73: 85–89.

[146] Carlin BP, Gelfand AE, Smith AFM. Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 1992; 41: 389–405.

[147] Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 1995; 82: 711–732.

[148] Denison DGT, Holmes CC, Mallick BK, Smith AFM. *Bayesian methods fo nonlinear classification and regression.* Wiley, Chichester, 2002.

[149] Ferreira JTAS, Denison DGT, Holmes CC. Partition modelling. In [10]; pp. 125–145.

[150] Arjas E, Gasbarra D. Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. *Statistica Sinica*, 1994; 4: 505–524.

[151] Arjas E, Heikkinen J. An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics*, 1997; 12: 385–402.

[152] IARC. *Monographs on the evaluation of the carcinogenic risks to humans, suppl 7.* IARC, Lyon, 1987.

[153] Diggle PJ. *Statistical Analysis of Spatial Point Patterns.* Academic Press, London, 1983.

[154] Møller J, Waagepetersen R. *Statistical Inference and Simulation for Spatial Point Processes.* Chapman & Hall, London, 2003.

# Yhteenveto - Summary in Finnish

Tämän työn yleisenä tavoitteena oli kehittää olemassaolevaa pienaluejärjestelmää, kun arvioidaan sairauden suhteellisen riskin alueellista vaihtelua mahdollisen päästölähteen ympäristössä. Yksityiskohtaisempina tavoitteina oli i) etsiä järjestelmään sopivia kehittyneitä tilastollisia menetelmiä; ii) ohjelmoida menetelmät osaksi järjestelmää; iii) arvioida rekisteriaineiston käyttömahdollisuuksia; ja iv) vahvistaa järjestelmän epidemiologisia menetelmiä.

Pienaluejärjestelmässä aineisto on jaettu 500 m × 500 m -ruutuihin. Aineiston alueellinen hienojakoisuus on informatiivinen, kun tarkastellaan paikallisia eroja, mutta aiheuttaa samaan aikaan alueellisesta riippuvuudesta ja aineiston harvuudesta (vähän väestöä ja tapauksia) aiheutuvia menetelmällisiä haasteita. Menetelmien kehittelyssä on tarkasteltu syövän suhteellista riskiä mm. entisen kaatopaikan, entisen asbestikaivoksen ja saastuneen joen ympäristössä sekä talousvesiongelmaisessa maaseutukunnassa.

Yksinkertaisimpana syövän suhteellisen riskin arviointimenetelmänä käytettiin aluevaikutustermillä varustettua Poisson-regressiomallia. Monimutkaisempana mallina sovitettiin hierarkkista Markovin mallia, joka sellaisenaan osoittautui kelpaamattomaksi harvalle hienojakoiselle aineistolle. Kolmea rajoitusmenetelmää ehdotettiin parantamaan hierarkkisen Markovin mallin käyttäytymistä. Koska hierarkkisen Markovin mallin laskenta osoittautui työlääksi, yksinkertaisempana arviointimenetelmänä sovitettiin vielä muutospistemalliin perustuvaa kehitelmää Poisson-regressiomallista.

Sekoittavien tekijöiden vaikutusta tuloksiin tarkasteltiin tutkimalla sekä eri vertailualueiden vaikutusta että sosioekonomisen aseman vaikutusta vakioivana tekijänä. Lisäksi tarkasteltiin eri vakiointimenetelmien vaikutusta syövän suhteellisen riskin arvioinnissa.

Hienojakoiseen aineistoon perustuva pienaluejärjestelmä on hyödyllinen, kun halutaan saada nopea alustava arvio syövän suhteellisesta riskistä mahdollisen päästölähteen ympäristössä. Järjestelmän vahvuus on, että tutkimusalue voidaan valita vapaasti 500 metrin tarkkuudella. Tosin harvalla aineistolla perinteiset menetelmät, kuten vakioitu esiintyvyyssuhde tai Poisson-regressiomalli, voivat tuottaa epävarman arvion syövän suhteellisesta riskistä. Ongelman ratkaisemiseksi voidaan käyttää tasoitusmenetelmiä. Toisaalta perinteiset menetelmät saattavat antaa virheellisiä tuloksia, jos tutkittavassa aineistossa on alueellista riippuvuutta. Tässä tapauksessa voidaan käyttää monimutkaisempia malleja. Tässä väitöskirjassa on ehdotettu joitain sopivia tilastollisia menetelmiä edellämainittujen ongelmien ratkaisemiseksi.

# Original Publications

[I]

Kokki E, Pukkala E, Verkasalo PK, Pekkanen J. Small Area Statistics on Health (SMASH): A System for Rapid Investigations of Cancer in Finland. In Briggs DJ, Forer P, Järup L, Stern R (Eds.). *GIS for Emergency Preparedness and Health Risk Reduction.* Kluwer Academic Publishers, Dordrecht, 2002; pp. 255–266.

[II]

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

[III]

Kokki E.

Constrained Bayesian modelling of disease risk around a point source. Publications of the Laboratory of Data Analysis, No. 5, University of Jyväskylä, 2003.

[IV]

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

[V]

Verkasalo PK, Kokki E, Pukkala E, Vartiainen T, Kiviranta H, Penttinen A, Pekkanen J. Cancer risk near a polluted river in Finland. Submitted, 2004.