Small area estimation of incidence of cancer around a known source of exposure with fine resolution data

E Kokki, J Ranta, A Penttinen, E Pukkala, J Pekkanen

Abstract

Objectives—To describe the small area system developed in Finland. To illustrate the use of the system with analyses of incidence of lung cancer around an asbestos mine. To compare the performance of different spatial statistical models when applied to sparse data.

Methods—In the small area system, cancer and population data are available by sex, age, and socioeconomic status in adjacent “pixels”, squares of size 0.5 km × 0.5 km. The study area was partitioned into sub-areas based on estimated exposure. The original data at the pixel level were used in a spatial random field model. For comparison, standardised incidence ratios were estimated, and full bayesian and empirical bayesian models were fitted to aggregated data. Incidence of lung cancer around a former asbestos mine was used as an illustration.

Results—The spatial random field model, which has been used in former small area studies, did not converge with present fine resolution data. The number of neighbouring pixels used in smoothing had to be enlarged, and informative distributions for hyperparameters were used to stabilise the unobserved random field. The ordered spatial random field model gave lower estimates than the Poisson model. When one of the three effects of area were fixed, the model gave similar estimates with a narrower interval than the Poisson model.

Conclusions—The use of fine resolution data and socioeconomic status as a means of controlling for confounding related to lifestyle is useful when estimating risk of cancer around point sources. However, better statistical methods are needed for spatial modelling of fine resolution data.

(*Occup Environ Med* 2001;58:315–320)

Keywords: disease mapping; point source; spatial random field model

Many environmental exposures originate from point sources such as factories, power plants, oil refineries, and dump areas. Because the point sources are often easy to recognise, they can be associated with strong negative attitudes and even fears.¹ A suspicion of increased risk of cancer in a region can cause remarkable psychological and economic consequences.² Therefore, public health authorities should be

able to respond rapidly to questions about potentially increased risks of disease. Due to the progress in geographical information systems (GIS), methods providing rapid initial answers to questions on risks of disease in an area have become realistic.³

The Finnish Cancer Registry and National Public Health Institute have developed a small area system, which can be used to compare the risk of cancer in a freely selected area of interest with a given reference area.⁴ In the United Kingdom, The Small Area Health Statistics Unit has developed a similar system.⁵ Many analyses have been done for this purpose with study areas defined by concentric circles,^{6–14} with bandwidths beside sources of pollution^{15–17} or with confirmatory partitions based on other covariates.¹⁸

Several statistical methods are available for analyzing risk of disease around point sources—for example, those of Elliot *et al*,^{3, 19} and Lawson *et al*.^{20, 21} The conventional approach is to calculate the standardised incidence (or mortality) ratios (SIR or SMR). Such standardised ratios are often subject to smoothing.

The confidence intervals (CIs) for SIRs (and SMRs) are calculated with assumed independence when applying the Poisson regression.⁴ However, the existence of spatially correlated factors which are not found but affect the risk of cancer cannot be excluded. Therefore, lack of spatial autocorrelation is often a strong simplification. Poisson regression with neighbouring counts as covariates²² is not a solution because it is mathematically inconsistent. Therefore, hierarchical methods as suggested by Clayton and Kaldor²³ and Besag *et al*²⁴ are more realistic approaches in spatial modelling of non-infective diseases. If exact coordinates for cases and reliable estimates for the population density exist, then point process modelling is also plausible.^{25, 26}

In this paper, we first describe the small area system developed in Finland. Then, we describe the data and define the confirmatory partition used in our illustration. With the full bayesian approach closely related to Besag *et al*, we derived the spatial random field model for estimation of risks of cancer with fine resolution data.²⁴ We compared the results obtained by means of the spatial random field model, which uses fine resolution data with the ones obtained by other spatial methods that use aggregated data. Our main objective was not in the estimation of the environmental effects but

Unit of Environmental Epidemiology, National Public Health Institute, PO Box 95, FIN-70701 Kuopio, Finland
E Kokki
J Ranta
J Pekkanen

Department of Statistics, University of Jyväskylä, PO Box 35, FIN-40351 Jyväskylä, Finland
E Kokki
A Penttinen

Rolf Nevanlinna Institute, University of Helsinki, PO Box 4, FIN-00014, University of Helsinki, Finland
J Ranta

Finnish Cancer Registry, Liisankatu 21 B, FIN-00170 Helsinki, Finland
E Pukkala

Correspondence to:
Mr E Kokki
Esa.Kokki@ktl.fi

Accepted 15 January 2000

in comparing the statistical methods to choose the best approach for the small area system.

Methods

THE SMALL AREA SYSTEM

The Finnish Cancer Registry often has to consider questions of whether there is an excess number of cases of cancer in a given geographic area. Previously, the analyses could only be undertaken at the level of the municipalities, which in Finland have a median population of 4800 and a median area of 380 km². To carry out analyses in smaller areas, the Finnish Cancer Registry and National Public Health Institute started to develop a system called SMASH (small area statistics on health).

The latitude and the longitude with an accuracy of 10 m for the centre of each building in Finland are given by the local housing officials. These data have been linked with the Central Population Register to obtain coordinates of the home address of each Finn. Based on these data, numbers of inhabitants by sex, age, and socioeconomic classification in squares of size 0.5 km×0.5 km (pixels) are calculated for all of Finland by Statistics Finland. Currently, SMASH holds population data for the years 1980, 1990, and 1997, separately for both sexes, for age groups (0–4, 5–9, . . . , 80–84, and >85) and for socioeconomic groups (farmers, upper clerical workers, lower clerical workers, skilled workers, unskilled workers, and others). The number of inhabited pixels was 197 761 in 1980, 197 169 in 1990, and 195 870 in 1997, covering about 16% of the land area of the country.

The SMASH also holds data on cancers obtained from the Finnish Cancer Registry.

Figure 1 Population density and confirmatory partition of the study area. The size of the circle is proportional to the number of inhabitants (1–782) in each 0.5 km × 0.5 km pixel. The rectangles divide the study area into the risk areas A_1 , A_2 , and A_3 . The asbestos mine is located in the centre of the study area.

Current data consist of all the cases of cancer diagnosed in 1981–97 in Finland, which are registered by the Finnish Cancer Registry. The metric coordinates of residence were linked for the individual cases of cancer by Statistics Finland. The place of residence is known on 31 December in 1980 and 1990 and on 1 January in the year of diagnosis of the cancer. Sex, age, and socioeconomic status of cases are known in the same categories as in population data.

The purpose of SMASH is to compare incidence of cancer in the area of interest with that of a selected reference area. The area, cancer, period of diagnoses, date of living in the area, age groups, socioeconomic classes, and the reference area can be chosen freely within the precision of resolution of the system (0.5 km×0.5 km). Examples are: the whole of Finland, one of the five large geographical areas used in cancer treatment administration (east, south, southwest, west, or north), old towns, new towns, rural areas, or any combination of individual areas.

In routine use, SMASH enables the user to calculate the SIRs for a selected area—that is,

$$\text{SIR} = \frac{\sum_i O_i}{\sum_i E_i}$$

where O_i is the number of observed cases of cancer in pixel i which belongs to the area of interest. The expected number of cases of cancer E_i is counted in pixel i separately for each sex, age, calendar year, and socioeconomic group by multiplying the class specific incidence of cancer for sex, age, and socioeconomic status in the reference area with the class specific number of inhabitants in pixel i for sex, age, and socioeconomic status. The CIs for the SIRs are calculated with the Byar's approximation formula with the Poisson assumption.²⁷

CASE STUDY MATERIAL

Exposure to various types of asbestos results in an increased risk of lung cancer.²⁸ Anthophyllite asbestos mining continued between years 1918–75 in Paakkila in the eastern part of Finland. The study area is a square of size 50 km×50 km around the former asbestos mine consisting of a rural area with three built up areas. The present analyses used population data for 31 December 1980.

The cancer data consist of those people with diagnosed lung cancer (seventh revision of the international classification of diseases (ICD-7) 162) between 1981–97, registered by the Finnish Cancer Registry, and whose place of residence was within the study area on 31 December 1980. For calculation of the expected number of cases (stratified by sex, age, and socioeconomic class) all the rural areas in Finland were selected as the reference area.

In the present case study we used an initially defined confirmatory partition (fig 1). The risk area A_1 was defined as a 4 km×4 km square around the asbestos mine. The area was not centred at the mine, because prevailing wind direction in this area is from the south or southwest. The bandwidth beside the main roads away from the mine formed the risk area

Table 1 The descriptive statistics in the sub-areas

The risk area	Population	Cancers n	All pixels n	Pixels with inhabitants n (%)	Pixels with cancer n (%)
A_1	194	6	64	24 (38)	4 (6)
A_2	2124	24	654	193 (30)	18 (3)
A_3	17507	154	9282	1834 (20)	116 (1)

A_2 . The exposure in the risk area A_2 was supposed to be higher than in the rest of the study area, because of transportation of asbestos to and away from the mine. The rest of the study area formed the risk area A_3 .

SPATIAL ANALYSIS

As in the standard use of SMASH today, we started with a non-spatial analysis of aggregated data, where the cases O_k in the areas A_1 , A_2 and A_3 were assumed to follow the Poisson distribution with mean $E_k \lambda_k$. The maximum likelihood estimate of the relative risk under this model is simply $SIR_k = O_k / E_k$. The CIs were estimated according to Byar's approximation formula.²⁷

As a second approach, we applied a full bayesian model to the aggregated data in each confirmatory area. In bayesian analysis initial knowledge of the phenomenon can be expressed in terms of a prior distribution. This can be chosen to represent informative expert knowledge, if available. We chose a non-informative distribution for the relative risks. The prior mean for relative risks was 1, and the variance was very large. A mean of 1 corresponds to the expected $SIR=1$. The posterior distribution of relative risks provided a summary of our initial knowledge and the information drawn from data. With the WinBUGS

software, the posterior distribution was approximated by the numerical Markov chain Monte Carlo (MCMC) sampling method.²⁹ We calculated the posterior estimates of relative risks with bayesian CIs. Equal tail intervals were calculated; the lower limits were at 2.5% and the upper limits were at the 97.5% percentile. Posterior probabilities were used to compare the risk levels between the confirmatory areas.

As a third approach, we applied the empirical bayesian method introduced by Clayton and Kaldor²³ also to the aggregated data. In this approach the prior distribution for relative risk was derived from data. We chose a gamma distribution, with a mean equal to the mean of the SIRs, and a variance equal to the sample variance of the SIRs. We again calculated the posterior risk estimates with equal tail 95% CIs with WinBUGS. In empirical bayesian analysis any posterior probabilities cannot be calculated, because the posterior distribution is estimated.³⁰

As a fourth approach, we used a spatial random field model. The derivation of this new model is described in the appendix. By contrast with the three other approaches, data were not aggregated, but estimation was done with the 0.5 km×0.5 km pixels. This model is closely related to the one presented by Besag *et al.*²⁴ The model is based on the assumption that two nearby pixels are more alike for exposure than any two pixels farther apart. This was likely to be the case, because the underlying exposure variable was the amount of asbestos particles emitted from the local asbestos mine. An analysis of the data from the register at the pixel level requires the use of a smoothed spatial statistical model, because the numbers of inhabitants and cases of cancer in pixels is very low and therefore, the risk estimates are prone to large random variation. The present spatial random field model was too complex to be calculated with WinBUGS. Hence, the program used was written by ourselves with the language C. As in a case of the full bayesian approach, we calculated the posterior probabilities and the posterior estimates of relative risks with bayesian equal tail CIs.

Results

The spatial distribution of population at risk is shown in figure 1. The number of inhabited pixels in the study area was 2051, which was 21% of the total number of pixels (table 1). The number of inhabitants varied from 1 to 782 in the inhabited pixels, whereas the total number of inhabitants was 19 825 in the whole study area. The spatial distribution of cases of lung cancer is shown in figure 2. The number of pixels with cases of cancer was 138, with 1–5 cases in each (totally 184 cases of cancer in the whole area).

We modelled the risk of lung cancer in the inhabited pixels. The spatial random field model did not work with the fine resolution data without some stabilising methods. In the derivation of the random field, equation (A.3), we started with eight neighbouring pixels. Because of the large number of isolated pixels we needed to

Figure 2 Spatial variation of cases of cancer and confirmatory partition of the study area. The size of the circle is proportional to the number of cases of cancer (1–5) in each 0.5 km × 0.5 km pixel. The rectangles divide the study area into the risk areas A_1 , A_2 , and A_3 . The asbestos mine is located in the centre of the study area.

Table 2 The area estimated relative risks (RRs (95% CIs)) given by the maximum likelihood method (SIR), the full bayesian model (FB), the empirical bayesian model (EB), the spatial random field model with the order condition (RF1), the spatial random field model with the fixed effect for A_3 (RF2)

Model	A_1		A_2		A_3	
	RR	95% CI	RR	95% CI	RR	95% CI
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
RF1	2.32	1.27 to 5.94	1.31	0.99 to 1.81	0.98	0.88 to 1.10
RF2	4.07	1.44 to 7.43	1.45	1.06 to 2.02	1.20	*

*In RF2, the area effect for the sub-area A_3 was fixed to 1.20, so no 95% CI was calculated.

enlarge the neighbourhood to cover 20 neighbouring pixels. Also, the smoothing parameter, κ in equation (A.3), was problematic because the data contain only a small amount of information on the smoothing level. Hence we had to set restrictions at baseline on the variations of κ to obtain a stable Markov chain in the MCMC simulation. We tried to keep the restrictions as weak as possible. Because of problems with convergence, an informative distribution was also selected for the variances of the effect of area, δ_k s, to allow for large variations in the effects of area, ξ_k s in equation (A.2). This was still not enough to achieve convergence, and we had to impose some restrictions on the effects of area. We applied two approaches. Firstly, based on SIRs we assumed that the risk of disease decreases with distance from the point source. bayesian methods allow order restrictions to be assigned to the model.³⁰ So we assigned an order $\xi_1 \geq \xi_2 \geq \xi_3$. The joint baseline distribution was thus $p(\xi_1, \xi_2, \xi_3) \propto p(\xi_1)p(\xi_2)p(\xi_3)1_{\xi_1 \geq \xi_2 \geq \xi_3}$, where the indicator function $1_{\xi_1 \geq \xi_2 \geq \xi_3}$ is 1 if $\xi_1 \geq \xi_2 \geq \xi_3$, and zero otherwise. The order was also supported by the analyses of the aggregated data. Secondly, because we were not so interested in the effect of distant area ξ_3 , we fixed it to be equal to the SIR in the sub-area A_3 .

The estimates of the parameters based on area are shown in table 2. Maximum likelihood estimates (SIRs) of relative risks indicate that the risk of cancer even within the area A_3 was significantly higher than in rural areas on average. The estimates given by the full bayesian model with non-informative distributions were the same, with slightly narrower CIs, as expected. In the empirical bayesian model, the estimate of relative risk in the area A_1 decreased but in the other areas they stayed almost the same. This is typical for the empirical bayesian method due to the shrinkage effect. The spatial random field model with the effect of A_3 fixed gave slightly higher estimates than SIRs, but the intervals were narrower. Instead, the ordered spatial random field model gave clearly lower estimates.

With the full bayesian approaches we can simulate the exact posterior distribution to compare the area specific risks to each other or to the mean risk in rural areas, all this in terms of posterior probabilities (table 3). The risk of lung cancer decreased when the distance from the asbestos mine increased. On the other hand, in two areas nearest to the asbestos mine the risks of lung cancer were still higher than in rural areas in Finland in 1981–97 on average. In the most distant area, A_3 , the risk of lung

Table 3 Comparison of posterior probabilities given by the full bayesian (FB) and spatial random field (RF1, RF2) models: RR_k denotes the estimated relative risk in the subarea A_k , $k=1,2,3$

	FB	RF1	RF2
$Pr(RR_1 > RR_2 O)$	0.96	1.00	0.96
$Pr(RR_1 > RR_3 O)$	0.98	1.00	0.99
$Pr(RR_2 > RR_3 O)$	0.77	1.00	0.86
$Pr(RR_1 > 1 O)$	0.99	1.00	1.00
$Pr(RR_2 > 1 O)$	0.95	0.97	0.99
$Pr(RR_3 > 1 O)$	0.99	0.32	*

*In RF2, the area effect for the sub-area A_3 was fixed to 1.20, so no posterior probability was calculated.

cancer was higher than in rural areas on average by the full bayesian model, but not by the ordered spatial random field model.

Discussion

A small area system with fine resolution data is expected to be more informative than conventional aggregated data for estimation of risks of cancer. On the other hand sparse data cause methodological and computational problems. The spatial random field models developed for conventional small areas could not be straightforwardly adapted to the present system of small areas with very few inhabitants and with no or very few neighbours. Therefore, better statistical methods have to be developed for analyzing fine resolution data on incidence of cancer.

Public health authorities need to be able to react rapidly to suspected increases in risks of cancer due to the environmental pollution around point sources.³ In SMASH, the system developed in Finland, calculation of SIRs can be done within a couple of hours. The system also allows the study area to be chosen freely with an accuracy of 0.5 km. The main weakness currently in the interpretation of the results from SMASH is that exposure assessment is solely based on place of residence at one time and there is no accounting for migration or daily commuting. Socioeconomic status is known to be a good proxy of risk factors strongly related to social class, such as smoking, healthy diet, etc.³¹ However, often data on other confounding variables would also be needed for reliable inference. Therefore comparisons by area, as in SMASH, can only provide preliminary results on risks of cancer, which then can be used to help to decide on the need for further investigations, which often include extensive fieldwork.³

Due to the low numbers in the present analyses, indirect standardisation and SIRs were the only feasible method of analyses. The SIRs produce a valid comparison between the area of interest and the reference area. However, due to the indirect standardisation used, two SIRs are mutually comparable only if there is no difference in stratum specific relative risks.³² In the present study, the strata that were adjusted for indirect standardisation were sex, age, and socioeconomic status. The only major difference between study areas was the lower proportion of farmers in sub-area A_1 . However, the relative risks between study areas among farmers and non-farmers were quite similar. Therefore, the SIRs in the present study are comparable.

We found an increased risk of lung cancer around the former asbestos mine. It is known that asbestos miners in Paakkila have a 2.8-fold increased risk of lung cancer.³³ In the present case study we were not able to use the occupational exposure of the ex-miners as a covariate or confounder, because our broad standard socioeconomic classification did not include an indicator of former work with asbestos. If the matrix on population and cancer data had been developed especially for studying the problem of environmental exposure to asbestos, the population in each pixel could be classified into asbestos miners and non-miners by linking the personal identifiers with the existing cohort of asbestos miners.³³ It should also be noted that our data may not best exemplify a study on environmental exposures, because the main exposure to asbestos is occupational and at one geographical location (in the factory). Holopainen *et al*³⁴ calculated the SIRs for lung cancer in 1953–95 in the municipality of Tuusniemi, where the asbestos mine is. They found that the SIR of lung cancer among the total population was 1.08, among occupationally exposed subjects it was 3.0, and among non-occupationally exposed people it was 0.89. In the present analyses, the risk estimates in the area A_1 were similar to the SIR among miners and the estimate in the area A_3 were similar to the SIR among inhabitants of the Tuusniemi district. Therefore, occupational exposure seems to be the main cause of the increased risk of lung cancer among inhabitants living near the asbestos mine.

Georeferenced data are often spatially autocorrelated.³⁵ Spatial statistical modelling, as in the present spatial random field model, is then a useful approach because it accounts for spatial autocorrelation. One of the strengths of the spatial random field model is that it exploits information from the neighbouring pixels for smoothing the risks towards the local mean. Estimation uncertainty of risk at the pixel level is also possible, because the model provides exact estimates of the precision for each pixel—that is, posterior variances. Statistical strength of the spatial random field model and the full bayesian model is that the assumption of the Poisson process is usually valid conditionally, in a hierarchical model in a case of non-infectious diseases. Another strength of the spatial random field model and other bayesian methods is that prior knowledge about the subject of analysis can be used if available. Also, the interpretation of the bayesian CI is convenient, because it considers the actual probability that the value of an unknown variable lies between the limits of the interval. When the posterior distribution is represented by a numerical sample generated by MCMC, the computation of approximate posterior probabilities of functions of the variables in the model is straightforward.

The spatial random field model presented by Besag *et al*³ is commonly used for mapping incidences of disease, and it has proved to be one of the best approaches to mapping disease.²¹ The estimation of the spatial random field model is,

however, quite laborious and we also found several problems with convergence when the model was used with sparse data. If data are very few or pixels have only few neighbours, smoothing of the risk estimates can be difficult. This can be partly overcome by enlarging the neighbourhood area with some reasonable weighting. However, this should be done carefully, because at the same time one loses precision in reference area. Another way to help the spatial random field model to converge is to choose informative distributions for hyperparameters. With fine resolution data other restrictions may still be needed. In this paper we applied two types of restrictions. Ordering the effects of area did not work. Fixing one of the three effects of area proved to be a useful limitation, as the model converged. Also the sizes of the estimates were reasonable compared with SIRs, and the estimates had narrower CIs.

In approaches with aggregated data, it should be noted that aggregation is a way of smoothing the data. After aggregation, estimation or mapping of risk of cancer with finer resolution is impossible, because the area risks are the only information available. After all, if the data are very sparse, then aggregation into larger areas is reasonable, but it is an open question how large the areas should be.

The approach of estimating the SIRs is easy and fast. Main statistical problems with SIRs are associated with the Poisson assumption and with the lack of spatial autocorrelation. Although the sum of independent numbers of observed cases is Poisson distributed, this is no longer the case if spatial autocorrelation exists and the resulting CIs may be wrong. For reliable inference, the SIRs require quite large areas—that is, large populations, because when the expected number of cases is small the CIs of SIRs can become very wide. The requirement of aggregation to large areas may be a contradiction in the study of point sources. However, SIRs are useful summaries of data at an early stage of modelling.

In conclusion, the small area system, SMASH, developed in Finland is useful for estimation of risks of cancer in a small area due to the use of fine resolution data and availability of socioeconomic status, which is a major confounder. On the other hand, sparse data can cause stabilising problems in the computing. The spatial random field model applied in previous small area studies did not work straightforwardly with fine resolution and sparse data. Therefore, better statistical methods are needed for spatial modelling of fine resolution data.

The Academy of Finland through the project “Small area analyses of cancer incidence around a point source (42559)” and The Ellen and Artturi Nyyssönen Foundation have supported the work of EK. The European Community through the project “A European environment and health information system for exposure and disease mapping and risk assessment (VS/1999/5279 (99CVF2–606))” has supported the work of JR. The Academy of Finland has supported AP through a senior researcher grant (385).

Appendix:

Derivation of the spatial random field model

The risk of cancer in pixel i is denoted by θ_i , $i = 1, \dots, I$. Hence, the collection $(\theta_1, \dots, \theta_I)$ seen at the pixel

locations is called a disease map and is the objective of statistical inference. We assume, that given $(\theta_1, \dots, \theta_I)$, the observed cases O_i are conditionally independent and follow the Poisson distribution with mean:

$$(A.1) \quad \theta_i = E_i h_i \zeta_i.$$

Covariate information on sex, age, and socioeconomic status was considered through the expected number of cases E_i . The dummy variable h_i contained the effect of distance measured from the fixed point source. The unmeasured covariate $\zeta = (\zeta_1, \dots, \zeta_I)$ was specified as a spatially correlated random field. The role of this random effect was in the smoothing of the risks of cancer and in the exchange of information between nearby pixels. The likelihood function can be written as a product:

$$L(O|E, h, \zeta) = \prod_{i=1}^I \frac{(E_i h_i \zeta_i)^{O_i}}{O_i!} \exp(-E_i h_i \zeta_i).$$

In the analysis, we present the distance effect as a log linear model with the confirmatory partition described in methods section, so that:

$$(A.2) \quad h_i = \exp\{\xi_1 1_1(i) + \xi_2 1_2(i) + \xi_3 1_3(i)\},$$

where $1_k(i)$ is 1 if pixel i was located in the sub-area k , and zero otherwise. The confirmatory partition is convenient to use with SMASH because the source of exposure of pollutant is not always a point source. For example, a fixed source can be a river, power line, or a dump area. The sub-areas are also allowed to be asymmetric. Conditional on E , h , and ζ the model (A.1) is loglinear with:

$$\log(\theta_i) = \log(E_i) + \eta_i + \xi_1 1_1(i) + \xi_2 1_2(i) + \xi_3 1_3(i),$$

where E_i was treated as an offset and $\eta_i = \log(\zeta_i)$.

We assumed that the effect of area (effect distance from the point source) varies around unity in the risk areas. Therefore, baseline distributions for ζ_i s were assumed to be normal with zero mean and variance δ_s . We assigned inverse gamma distribution with fixed parameters as the prior of variances δ_s .

The prior derivation for the random field $\eta = (\eta_1, \dots, \eta_I)$ resembles the derivation by Besag *et al.*²⁴. We assumed that η is an intrinsic gaussian random field defined by the joint improper distribution:

$$(A.3) \quad p(\eta_1, \dots, \eta_I | \kappa) \propto \kappa^{I/2} \exp\left\{-\frac{\kappa}{2} \sum_{i,j} w_{ij} (\eta_i - \eta_j)^2\right\},$$

where κ has the role of a smoothing parameter and w_{ij} is a weight relating the neighbouring pixels i and j . The prior for κ was assumed to be gamma distribution with fixed variables. We chose the weights w_{ij} to be one for all neighbouring pixels.

The posterior distribution was calculated by MCMC simulation. In the derivation of the Markov chain we applied a single site Metropolis updating scheme for η_i s and ζ_i s and Gibbs updating for κ and δ_s these being conventional choices. For further information, see Gilks *et al.*³⁶

- 1 Bender AP, Williams AN, Johnson RA, *et al.* Appropriate public health responses to cluster: the art of being responsibly responsive. *Am J Epidemiol* 1990;132:48–52.
- 2 Guidotti TL, Jacobs P. The implications of an epidemiological mistake: a community's response to a perceived excess cancer risk. *Am J Public Health* 1993;83:233–9.
- 3 Elliot P, Cuzick J, English D, *et al.*, eds. *Geographical and environmental epidemiology: methods for small area studies*. Oxford: Oxford University Press, 1992.
- 4 Pekkanen J, Pukkala E, Vahteristo M, *et al.* Cancer incidence around an oil refinery as an example of a small area study based on map coordinates. *Environ Res* 1995;71:128–34.
- 5 Aylin P, Maheswaran R, Wakefield J, *et al.* 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. *J Public Health Med* 1999;21:289–98.

- 6 Hills M. Some comments on methods for investigating disease risk around a point source. In: Elliot P, Cuzick J, English D, *et al.*, eds. *Geographical and environmental epidemiology: methods for small area studies*. Oxford: Oxford University Press, 1992:231–7.
- 7 Michaelis J, Keller B, Haaf G, *et al.* Incidence of childhood malignancies in the vicinity of West German nuclear power plants. *Cancer Causes Control* 1992;3:255–63.
- 8 Urquhart J. Studies of disease clustering: problems of interpretation. In: Elliot P, Cuzick J, English D, *et al.*, eds. *Geographical and environmental epidemiology: methods for small area studies*. Oxford: Oxford University Press, 1992:278–86.
- 9 Viel J-F, Pobel D. Incidence of leukaemia in young people around the La Hague nuclear waste reprocessing plant: a sensitivity analysis. *Stat Med* 1995;14:2459–72.
- 10 Dummer TJB, Dickinson HO, Pearce MS, *et al.* Stillbirth rates around the nuclear installation at Sellafield, North West England: 1950–89. *Int J Epidemiol* 1998;27:74–82.
- 11 Dolk H, Vrijheid M, Armstrong B, *et al.* Congenital anomalies near hazardous waste landfill sites in Europe. In: Lawson A, Biggeri A, Böhning D, *et al.*, eds. *Disease mapping and risk assessment for public health*. Chichester: Wiley, 1999:383–94.
- 12 Hoffmann W, Schlattmann P. An analysis of the geographical distribution of leukaemia incidence in the vicinity of a suspected point source: a case study. In: Lawson A, Biggeri A, Böhning D, *et al.*, eds. *Disease mapping and risk assessment for public health*. Chichester: Wiley, 1999:395–410.
- 13 Waller LA, Poquette CA. The power of focused score tests under misspecified clusters models. In: Lawson A, Biggeri A, Böhning D, *et al.*, eds. *Disease mapping and risk assessment for public health*. Chichester: Wiley, 1999:257–70.
- 14 Sharp L, McKinney PA, Black RJ. Incidence of childhood brain and other non-haematopoietic neoplasms near nuclear sites in Scotland, 1975–94. *Occup Environ Med* 1999;56:308–14.
- 15 Verkasalo PK, Pukkala E, Hongisto MY, *et al.* Risk of cancer in Finnish children living close to power lines. *BMJ* 1993;307:895–9.
- 16 Verkasalo PK, Pukkala E, Kaprio J, *et al.* Magnetic fields of high voltage power lines and risk of cancer in Finnish adults: nationwide cohort study. *BMJ* 1996;313:1047–51.
- 17 Harrison RM, Leung P-L, Somerville L, *et al.* Analysis of incidence of childhood cancer in the West Midlands of the United Kingdom in relation to proximity to main roads and petrol stations. *Occup Environ Med* 1999;56:774–80.
- 18 Trepka MJ, Heinrich J, Krause C, *et al.* The internal burden of lead among children in a smelter town: a small area analysis. *Environ Res* 1997;72:118–30.
- 19 Elliot P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Stat Methods Med Res* 1995;4:137–59.
- 20 Lawson A, Biggeri A, Böhning D, *et al.* *Disease mapping and risk assessment for public health*. Chichester: Wiley, 1999.
- 21 Lawson AB, Biggeri AB, Boehning D, *et al.* Disease mapping models: an empirical evaluation. *Stat Med* 2000;19:2217–41.
- 22 Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B* 1974;36:192–236.
- 23 Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987;43:671–81.
- 24 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;3:1–59.
- 25 Diggle PJ. A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society A* 1990;153:349–62.
- 26 Lawson AB, Waller LA. A review of point pattern methods for spatial modelling of events around sources of pollution. *Environmetrics* 1996;7:471–87.
- 27 Breslow NE, Day NE. *Statistical methods in cancer research. Vol II. The design and analysis of cohort studies*. Lyon: IARC, 1987.
- 28 International Agency for Research on Cancer. *Monographs on the evaluation of carcinogenic risks to humans, suppl 7*. Lyon: IARC, 1987.
- 29 Spiegelhalter DJ, Thomas A, Best NG. Computation on bayesian graphical models. *Bayesian Statistics* 1996;5:407–25.
- 30 Carlin BP, Louis TA. *Bayes and empirical Bayes methods for data analysis*. London: Chapman and Hall, 1996.
- 31 Pukkala E. *Cancer risk by social class and occupation. A survey of 109 000 cancer cases among Finns of working age*. Basel: Karger, 1995.
- 32 Rothman KJ, Greenland S. *Modern epidemiology*, 2nd ed. Philadelphia: Lippincott-Raven, 1998.
- 33 Meurman LO, Pukkala E, Hakama M. Incidence of cancer among anthophyllite asbestos miners in Finland. *Occup Environ Med* 1994;51:421–5.
- 34 Holopainen M, Pukkala E, Piirainen J, *et al.* Asbestille altistumisen terveysvaikutukset Paakkilassa. In: Lestinen P, ed. *Neljännet geokemian päivät 12.-13.11.1997 Kuopio - Laajat tiivistelmät*. Kuopio: Vuorimiesyhdistys, 1997:69–72. (In Finnish.)
- 35 Smans M, Estève J. Practical approaches to disease mapping. In: Elliot P, Cuzick J, English D, *et al.*, eds. *Geographical and environmental epidemiology: methods for small area studies*. Oxford: Oxford University Press, 1992:141–50.
- 36 Gilks WR, Richardson S, Spiegelhalter DJ. *Markov chain Monte Carlo in practice*. London: Chapman and Hall, 1996.