Journal of Epidemiology and Community Health 1995;49(Suppl 2):S20–S27

Disease risk near point sources: statistical issues for analyses using individual or spatially aggregated data

Peter Diggle, Paul Elliott

Abstract

Study objective **– To examine the statistical issues involved in the analysis of disease risk near point sources of environmental pollution, where data are held at both the individual and group (areal) level. To explore these issues with reference to possible socioeconomic confounding.**

Design **– Statistical review.**

Setting **– Point sources of environmental pollution.**

Main results **– Except in very specific circumstances unlikely to hold in practice, aggregation of data to the areal level will lead to bias in the estimation of disease risk.**

Conclusions **– There is no easy solution to the analysis of spatial data when some covariates (for example, age and sex of cases) are known at individual level, whereas others (for example, populations, age–sex distributions, small area deprivation indices) are known only at the areal (ecological) level. The underlying assumptions inherent in the analysis of these data need to be explicitly recognised in order to understand better the limitations of the available methodology as well as to inform interpretation of results. Ideally, the data should be kept as disaggregated as possible, to maximise the information available and minimise potential for bias.**

(*J Epidemiol Comm Health* 1995;49(Suppl 2):S20–S27)

This paper addresses the statistical and interpretive issues arising from a common problem in small area studies of environment and health: how to proceed when the necessary data (health, population, pollution, confounders) are available at different levels of spatial aggregation? What assumptions are implicitly made when data are aggregated spatially before analysis? What effects will this have on the results?

The resurgence of interest in the investigation of disease risk in small areas near point sources of environmental pollution[1] follows identification of a cluster of childhood leukaemia cases near the Sellafield nuclear plant in 1983.[2] Although a range of statistical methods has been developed to deal with such problems, none is ideal and interpretation is complex.[3] Often routine sources of data are used, which are subject to important limitations, errors, and possible biases. The health data are susceptible in varying degrees, to inaccuracy, diagnostic and coding variation, and, for cancer registrations and congenital malformations in particular, to incompleteness and duplication. Population data, necessary for calculation of disease risks, are usually obtained in aggregate form from national census (every 10 years in the UK) and may not reflect well local population structure and migration patterns during the inter-censual years. Worse still, exposure data may not be available at all – for example, historical pollution patterns around a putative source, and some proxy for exposure (such as distance) may have to be used. These potential sources of error are unlikely to be spatially neutral and could substantially bias small area studies of environment and health.

In this paper, we shall be concerned with two further issues in small area analyses. First is the problem of the availability of data for small area studies, alluded to above. The second, given these problems of data aggregation and measurement, is how to deal appropriately with the potentially major confounding effects of social deprivation.

In the UK, we are fortunate that many health data, including mortality and cancer registrations, are available at an individual level through use of the postcode of residence; but other data, including information on the population at risk (from census) and possible social and other confounding variables, may only be available for areal units such as census wards (10 000 people approximately) or enumeration districts (440 people). In the classic approach to epidemiology, relationships between health outcome, exposure, and potential confounders are investigated at the level of individuals – that is, in longitudinal, case-control, or cross sectional studies. But because of data availability problems, this approach is often not feasible or practicable in studies of pollution and health near a point source. The prospect of obtaining individual level data (including use of population based controls) may be remote, as purpose designed studies to obtain these data are notoriously expensive and time consuming.[4]

Unfortunately, as we shall show in the analysis of aggregated data it is not valid to assume that relationships found at the aggregated (areal) level will hold for individuals within those areas – that is, the so called *ecological fallacy*.[5]

How then should we deal with potential socioeconomic confounding and its effects on estimates of risk near point sources of environmental pollution, as individual meas-

Department of Mathematics and Statistics, Lancaster University
P Diggle

Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine
P Elliott

Correspondence to:
Professor P Diggle.

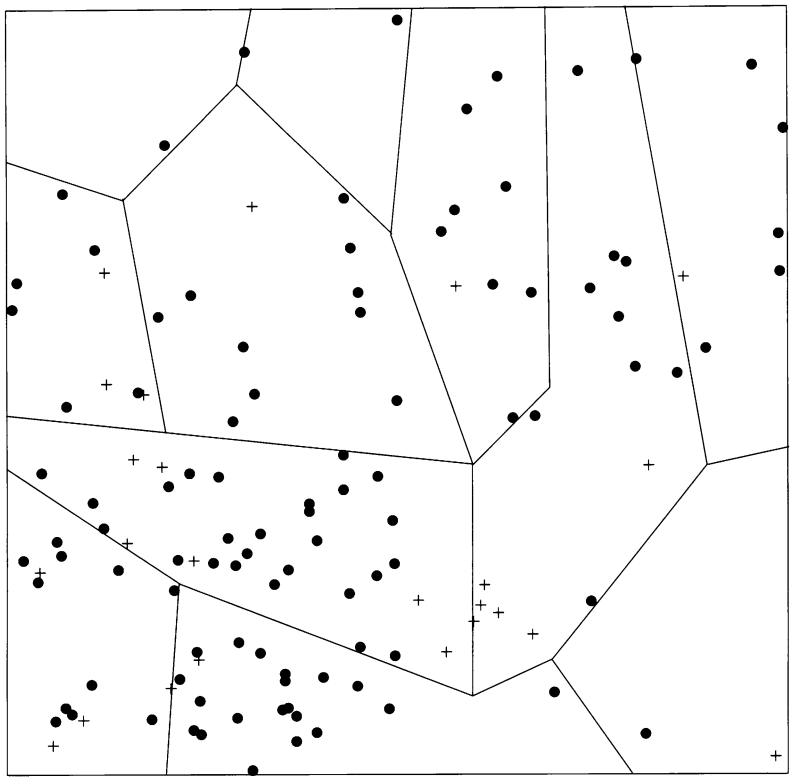

Figure 1 Hypothetical data on the spatial incidence of disease, in a unit square region with a relatively high population density in the lower left quarter square, and a cluster of cases in the lower right quarter square. Crosses represent 25 individual cases, dots 100 individual controls. Straight lines define a subdivision of the unit square study region into ten polygonal subregions.

ures of social class are rarely available? Often, use is made of proxy measures, which in the UK can readily be obtained at the small area scale. For example, areal deprivation measures – such as Carstairs or Townsend scores – are commonly based on census small area statistics variables such as unemployment rates, household overcrowding, and car ownership.[6] That deprivation is a true confounder of small area studies around a point source is apparent from the following: (i) it is well documented that for many diseases, including common cancers such as stomach and lung, occurrence of disease may be up to two- to threefold higher in areas of deprivation compared with affluent areas; and (ii) many sources of pollution are located in poor or deprived areas.[6] Therefore, an analysis of disease risk near a source that ignores the possibility of socioeconomic confounding, may reach seriously misleading or false conclusions about the effects of emissions on health.

Our objective in this paper is to present an idealised model of spatial variation in disease, formulated at the individual level. We are then able to examine the assumptions implicitly made, and the problems that arise, when we are forced (through non-availability of data) to aggregate to the group (areal) level. These problems are concealed if we simply analyse the data at the group level from the start. The model incorporates explicit assumptions about the pattern of risk associated with proximity to a point source, in the absence of information on pollution levels and individual exposure. It also incorporates adjustment for relevant explanatory variables measured at the individual level.

We end by discussing the "real world" problem mentioned at the start of this section of how we might try to deal with situations in which some variables, such as health, can be measured at the individual level while others, such as socioeconomic confounders, are available only at group level.

An idealised point process model

A *spatial point process* is a statistical model for determining the locations of events of interest in a geographical region. In this paper, we shall discuss point processes very informally. Mathematically precise descriptions can be found in a number of textbooks including Cox and Isham[7] or, at a more advanced level, Daley and Vere-Jones.[8] In the epidemiological context, the events represent reference locations (typically place of residence at diagnosis or death) of all known individual *cases* of a disease, although the appropriateness of this may be open to debate. In particular, no allowance is made for local commuting patterns, work habits, etc (for example, in studies of air pollution and health), nor are effects of migration into or out of the area taken into account. We shall initially assume that a second set of events is also observed, representing the corresponding reference locations for a set of *controls* selected at random from the population at risk. The data for analysis can then be presented as a dot map of both types of event within some designated study region, R say (fig 1).

We shall assume that each type of event can be modelled as a partial realisation of an *inhomogeneous Poisson process*. This model incorporates a spatially varying *intensity function*, the intensity at a point x being defined as the expected number of events per unit area in a small neighbourhood around x, but allows no direct interactions amongst the events. In other words, the events are assumed to occur independently in space, with the possibility of an event occurring being unaffected by the occurrence of other events nearby. The model would therefore be inappropriate for studying the spatial distribution of an infectious disease, but is reasonable as a model for non-infectious diseases whose spatial distribution reflects the spatial distribution of the population at risk and of any relevant demographic, socioeconomic, or environmental risk factors.

We let $\lambda(x)$ and $\lambda_0(x)$ denote the intensity functions of the case and control processes, respectively, and $r(x) = \lambda(x)/\lambda_0(x)$ the *risk function*. In practice, we would want to assume that the overall risk derives from one or more measured risk factors, and we shall consider how to deal with this in a later section. For the time being, we assume only that there is a spatial distribution of the population at risk, represented by the control intensity function $\lambda_0(x)$, and a spatially varying overall risk function, $r(x)$, which together determine the spatial

variation in the case intensity function $\lambda(x)$ via the equation

$$\lambda(x) = \lambda_0(x)r(x). \quad (1)$$

A model based interpretation of aggregation bias

Ecological bias arises when estimates of the variation in risk between groups of individuals exposed to different average levels of a risk factor are wrongly interpreted as estimates of the variation in risk between individuals.[9] Greenland and Morgenstern[9] give numerical examples which illustrate its potentially severe effects. They also make the important point that ecological bias can arise for a range of reasons, including spatial aggregation and confounding. The bias which arises as a direct consequence of spatial aggregation, and which we will call *aggregation bias*, is therefore a special case of the more general phenomenon of ecological bias. Note, however, that aggregation bias and ecological bias are sometimes taken as synonyms.[9] In the present context, the practical relevance of aggregation bias is that it is often difficult to obtain reliable data on the locations of individual controls. Suppose, instead, that we know only the numbers of individuals at risk, say $N_1, \ldots, N_m$, in a designated partition of the study region into subregions, $A_1, \ldots, A_m$ (fig 1). For example, the subregions may correspond to local government wards or census enumeration districts. We can then base our analysis not on individual locations, but on counts of the numbers of cases in the subregions. Under the assumed Poisson process model, the number of cases in A_i, say Y_i, follows a Poisson distribution with expectation

$$\mu_i = \int_{A_i} \lambda(x)dx = \int_{A_i} \lambda_0(x)r(x)dx \quad (2)$$

and counts in different subregions are mutually independent. If we were able to evaluate the integrals which appear on the right hand side of equation (2), an ecological analysis of the aggregated counts Y_i would give consistent estimates of parameters in any assumed regression model for the spatial variation in the individual level risk, $r(x)$. This is rarely feasible, except when the only risk factors under consideration are categorical variables and the numbers at risk within each category are known. For example, the subregional counts $N_1, \ldots, N_m$ may be further subdivided by age and sex.

The more usual way in which the analysis proceeds is by invoking (often implicitly) a spatially aggregated version of equation (1), in which we assume that the Y_i follow independent Poisson distributions with expectations

$$\mu_i^* = \mu_{0i}r_i,$$

where μ_{0i} represents the size of the population at risk in the subregion A_i, and r_i the assumed risk in subregion A_i. We can then replace μ_{0i} by the observed population count N_i (or, as noted above, by an appropriately weighted average of subcounts in different risk groups such as age and sex) to obtain

$$\mu_i^* = N_i r_i. \quad (3)$$

The theoretical justification for this is that, under the assumed Poisson process model, conditional on the numbers of people at risk in the m subregions A_i, the numbers of cases follow independent binomial distributions with numbers of trials N_i and probabilities r_i, and the Poisson distribution is a good approximation to the binomial because the r_i are small. Note that if the r_i are not small, we would need to work with the exact binomial distributions, or Normal approximations, rather than with the Poisson approximation.

As noted above, a variant of (3) is to replace the population count N_i by an expected number of cases, E_i say, calculated as a weighted sum of numbers in different risk categories. The implicit assumption in this analysis is the following. Imagine that the number of cases, Y_{ij}, and corresponding numbers at risk, N_{ij}, were available for each of c risk categories $j = 1, \ldots, c$ within each of m subregions $i = 1, \ldots, m$. Furthermore, suppose that the risk factors, p_j say, are known for each of the c risk categories, and that our objective is to investigate the effects of further possible risk factors at the subregional level. We would then model the counts Y_{ij} as independent, Poisson distributed random variables with expectations $\mu_{ij} = N_{ij}r_{ij}$, where the risks r_{ij} are factorised as $r_{ij} = p_j r_i$. In practice, we do not observe the individual Y_{ij}, but only the subregional totals $Y_i = \Sigma_{j=1}^{c} Y_{ij}$. But it then follows that these subregional totals are also independent, Poisson distributed random variables with means $\mu_i = \Sigma_{j=1}^{c} \mu_{ij} = E_i r_i$, as required.

Analyses based on equation (3) are usually called *Poisson regression* methods, because the risks r_i are typically specified by a regression model in which the explanatory variables define the characteristics of the corresponding subregions (see section below). One objection to this approach is that the inferences then depend on the definition of the subregions, which are usually entirely arbitrary with respect to the disease under investigation. A more fundamental problem is that if we describe both population density and risk only at the subregional rather than the individual level, we are liable to fall foul of the ecological fallacy. A theoretical explanation of this is as follows.

Under the assumed Poisson process model for the spatial distribution of individual cases, the correct model for the Y_i is that they are mutually independent, Poisson distributed random variables, with expectations μ_i given by equation (2). A spatially aggregated Poisson regression analysis would assume that the Y_i are mutually independent and Poisson distributed, but with expectations given by equation (3), for suitably defined risks r_i. Aggregation bias now manifests itself because in general, $\mu_i \neq \mu_i^*$. Conditions under which $\mu_i = \mu_i^*$ – that is, no aggregation bias – are derived in the appendix.

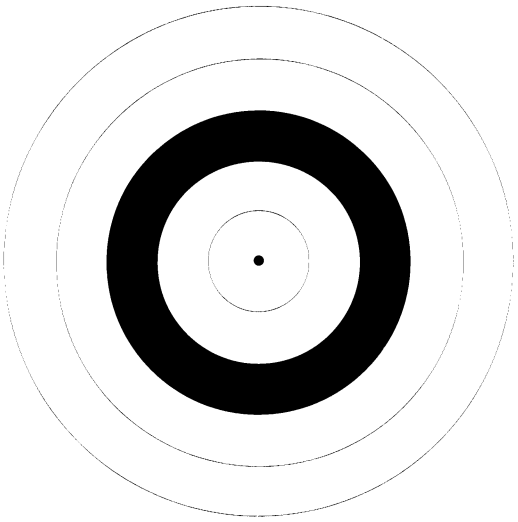

Figure 2 A circular region R *of unit radius, divided into* m = 5 *subregions* A_i *by a series of equally spaced, concentric circles. The subregion* A_3 *is shaded.*

These occur if, and only if, at least one of the following statements is true:

- $\lambda_0(x)$ is constant within A_i,
- $r(x)$ is constant within A_i,
- $\lambda_0(x)$ and $r(x)$ are spatially uncorrelated – that is, the risk of disease at any point is unrelated to the density of the population at risk at that point.

None of these conditions seems likely to hold in practice. However, they are more likely to hold approximately if the subregions are made as small as is practicable. Of course, if the ecological bias operated in opposite directions within different subregions, there might be some overall cancellation, but this seems highly implausible in practice.

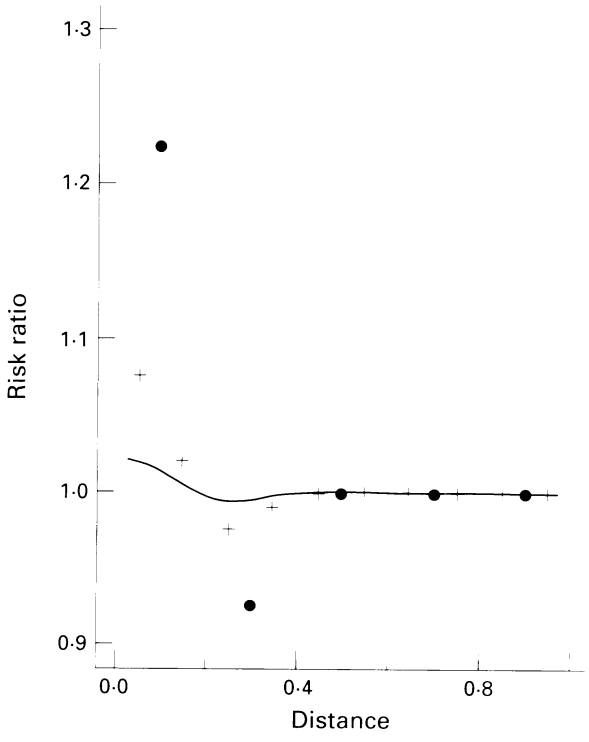

Figure 3 Ratios of notional to true average risk against distance from point source when a circular study region of unit radius, with a point source at its centre, is partitioned into n *circular subregions. Results are shown for* m = 5 *(solid dots),* m = 10 *(crosses), and* m = 20 *(continuous line). See text for details of the mathematical model used.*

We now give a specific numerical example of the effects of aggregation bias, based on an idealised but qualitatively plausible model for the variation in population density and risk around a point source. We consider a single point source to be located at the centre of a circular region R of unit radius, divided into m subregions A_i by a series of equally spaced, concentric circles (fig 2). Thus, A_i consists of all locations whose distance from the point source is between $(i-1)d$ and id, where $d=1/m$. Using x to denote distance from the point source, we model the variation in population density by the linear function

$$\lambda_0(x) = a + bx$$

and the variation in risk by

$$r(x) = \rho\{1 + \alpha \exp(-\beta x^2)\}$$

as in Diggle and Rowlingson.[9] By adjusting the values of a, b, α, and β this model can mimic a range of situations which might arise in practice. In particular, choosing positive values for a, α and β, and a negative value for b corresponds to the common situation in which both population density and risk decrease with increasing distance from the point source. Straightforward integration gives the population in the subregion A_i as

$$\lambda_{0i} = \pi a d^2(2i-1) + 2bd^3(3i^2 - 3i + 1)/3$$

and, from (2), the mean number of cases in A_i as

$$\begin{aligned}\mu_i = \pi\rho[&ad^2(2i-1)\\ &+ 2bd^3(3i^2-3i+1)/3\\ &+ a\alpha\beta^{-1}\{\exp(-\beta d^2(i-1)^2)\\ &- \exp(-\beta d^2 i^2)\} + \alpha b\{\mathcal{J}(id)\\ &- \mathcal{J}((i-1)d)\}/2],\end{aligned}$$

where

$$\begin{aligned}\mathcal{J}(z) = &-z\beta^{-1}\exp(-\beta z^2)\\ &+ \pi^{1/2}\beta^{3/2}\{\Phi(z(2\beta)^{1/2}) - 0{\cdot}5\}\end{aligned}$$

and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

For any values of the model parameters, we can now compare the true average risk μ_i in the subregion A_i with the notional risk r_i which we would ascribe to A_i if we assumed, incorrectly, that all individuals in A_i were subject to the same risk. Figure 3 shows the risk ratio, r_i/μ_i plotted against $(i-\frac{1}{2})/m$, for $m=5$, 10, and 20, $a=5$, $b=-5$, $\alpha=5$, and $\beta=40$ (the ratio does not depend on ρ). The aggregation bias at small distances from the putative point source is marked when $m=5$, but almost negligible when $m=20$. The bias is negligible at large distances, whatever the value of m, because, as shown in figure 4, the risk in this particular example is essentially constant for distances $x \geq 0{\cdot}4$.

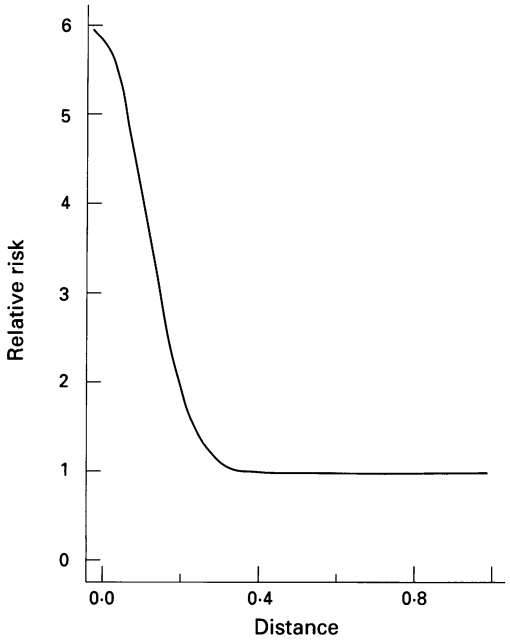

Figure 4 True relative risk as a function of distance, for the mathematical model used in figure 3.

We conclude that spatial aggregation is undesirable on two counts – arbitrariness introduced by the choice of subregions and bias in estimation of risk. However, we recognise that spatial aggregation is often forced upon us by the unavailability of control or covariate data at the individual level, and note that the practical consequences of aggregation bias can be diminished by using the smallest subregions which are practicable in a particular application.

Regression modelling of risk

The usual way to recognise multiple risk factors in models for spatially aggregated data is via the following log-linear regression equation. Let r_i denote the risk in the i^{th} subregion and z_{ij}: $j=1,\ldots,p$ the values for each of p explanatory variables attached to the i^{th} subregion. We then model the risk in the i^{th} subregion as

$$r_i = \exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j z_{ij}\right\}. \qquad (4)$$

By combining the regression equation (4) with (3) and the Poisson distributional assumption for the numbers of cases Y_i, we obtain the class of *Poisson regression* models, within which inferences about the effects of putative risk factors can be made using standard methodology for generalised linear models.[10]

For modelling at the individual level, we revert to a description of risk at an arbitrary point, x. Now, a log-linear regression equation for risk would take the form

$$r(x) = \exp\left\{\beta_0 + \sum_{j=1}^{p} \beta_j z_j(x)\right\}, \qquad (5)$$

where $z_j(x)$ represents the value of the j^{th} explanatory variable at the point x. We shall describe in the next section the technicalities involved in fitting this model to case-control data. For the moment, we simply point out that the explanatory variables $z_j(x)$ could be of several logically different kinds.

The most straightforward situation arises when $z_j(x)$ is a characteristic of an individual whose reference location is x. Examples include the age, sex, social class, or occupation of the individual concerned. Although an explanatory variable of this kind is not strictly a characteristic of a point x in geographical space, it is standard, and not unreasonable, practice to adjust for its effect at the individual level using the regression in equation (5).

It is slightly less obvious how we should deal with explanatory variables which directly describe the location x, for example, distance from a putative point source, or height above sea level. Although mathematically straightforward, incorporation of this kind of explanatory variable into the regression in equation (5) raises several operational difficulties. Firstly, interpretation of the corresponding regression coefficient requires us to believe that it is the value of the variable at a point which affects risk, whereas it might be more realistic to acknowledge that individuals in effect occupy a finite "territory" around their reference location, so that $z_j(x)$ should ideally be replaced by a spatial average of values $z_j(y)$ over points y in some neighbourhood of x. Secondly, spatial explanatory variables of this kind often cannot be measured directly at every location x, but their values must be estimated by some kind of interpolation procedure. This would be true, for example, of measurements of air pollution. From a statistical point of view, any such interpolation introduces measurement error, which in the simple case will result in attenuation of the corresponding regression coefficient towards zero.[11] Thirdly, a special difficulty applies to the explanatory variable "distance from putative point source", which is of particular interest in this context. The regression relationship for this variable cannot be log-linear, for the following reason. In modelling the distance based effect of a putative point source we would typically assume that risk decreases with increasing distance from the source and approaches the background level of risk at distances so large that the source no longer has any conceivable influence. However, in a log-linear model the assumed risk would automatically tend to be zero as distance from the point source increases.

In applications where the potential influence of the point source extends throughout the study region, this would not be a crucial objection, since it is clearly unwise to extrapolate *any* fitted model beyond the range of the data. However, such studies would be open to criticism on the grounds that they would not be capable of reliably estimating the true relationship between risk and distance from the point source in question. For example, if the true risk were approximately constant within 5 km of a point source, and thereafter declined rapidly to reach background level by 10 km, a study confined to a circular region of radius

5 km around the source would necessarily conclude that there was no relationship between risk and distance.

A third kind of explanatory variable is one which is only defined at the subregional level, for example, indices of social deprivation. A pragmatic way forward in such cases is to assume that the same value of $z_j(x)$ holds for every point x within a given subregion. The resulting discontinuities between subregions make little physical sense, but the alternative of deriving the value of $z_j(x)$ from a moving geographical window centred on x is entirely impracticable. The underlying statistical issue is again one of dealing with explanatory variables whose values are measured imprecisely, but with the added complication that the "measurement errors" are themselves highly structured spatially.

We return to these questions in the final section. For the time being, we allow ourselves the luxury of assuming that any relevant explanatory variable can be assigned a value for any location x at which we have either a case or a control.

A point process model for the association between risk and one or more point sources

We first summarise a modelling proposal in Diggle and Rowlingson[12] (henceforth DR). They note that if data on individual case locations $x_1, \ldots, x_n$ and control locations $x_{n+1}, \ldots, x_{n+n_0}$ are generated by Poisson processes with intensity functions $\lambda(x)$ and $\lambda_0(x)$ related by equation (1), then conditional on all $n+n_0$ locations x_i, the labels of the locations into cases and controls are mutually independent, with the probability that an event at x is a case given by

$$p(x) = \lambda(x)/\{\lambda(x)+\lambda_0(x)\} = r(x)/\{1+r(x)\}. \quad (6)$$

The important feature of equation (6) is that the conditioning on the locations x_i has removed the control intensity function $\lambda_0(x)$ from the problem – which is an appropriate thing to do precisely because it is of no interest to us. For a single point source, DR assume that

$$r(x) = f\{d(x)\}\exp\left\{\beta_0 + \sum_{j=1}^{p}\beta_j z_j(x)\right\}, \quad (7)$$

where $d(x)$ denotes the distance from x to the point source. Their specific implementation assumes that

$$f(d) = 1 + \theta\exp(-\phi d^2). \quad (8)$$

We emphasise that the particular algebraic form of (8) is neither crucial nor compelling. Its important features are that the parameters θ and ϕ are readily interpretable as elevation in risk at the point source and rate of decay in risk with squared distance from the source, and that $f(d)$ approaches the "neutral" value 1 at large distances. Although the binary regression model defined by (7) and (8) is inherently non-linear in θ and ϕ, its associated likelihood function is easily written down, and likelihood based inference can be implemented using a general purpose numerical optimisation routine.

For multiple sources, DR suggest a modification of (7) to

$$r(x) = [f\{d_1(x)\}\ldots f\{d_s(x)\}] \times \exp\left\{\beta_0 + \sum_{j=1}^{p}\beta_j z_j(x)\right\} \quad (9)$$

in which $d_k(x)$ denotes the distance from x to the k^{th} of s point sources, and the effects of the different sources are here assumed to be multiplicative. This is a strong assumption. However, as with the choice of the function $f(\cdot)$, the multiplicative model could be replaced by whatever other model was thought to be appropriate in a given application.

The simplest version of (9) is one in which all sources are governed by the same parameter values for θ and ϕ. If preferred, these can be replaced by source-specific parameters θ_k and ϕ_k, or grouped to reflect point sources of two or more qualitatively different kinds. Note that in all of these variants of the DR model, the parameter β_0 is an artefact of the relative numbers of cases and controls. It provides no information about the overall risk, and the models seek only to describe variation in relative risk with distance after allowing for other explanatory variables.

If individual control data are not available, the risk equation (9) can be incorporated into the Poisson regression formulation by simply replacing all of the $z_j(x)$ by spatially averaged versions z_{ij} and defining all locations within a given subregion to be the same distance from any given point source. For example, we could define d_{ik} to be the distance from the centroid of the i^{th} subregion to the k^{th} point source. As noted in section 3, this device introduces aggregation bias. We intend to explore the practical consequences of this simple approach in a future paper.

Discussion

In this final section, we raise two general issues which arise when using these models in practice – dealing with a mixture of explanatory variables at individual and at group level, which is most commonly the case in practice, and recognising the possibility of residual spatial correlation in the data. In each case, the options for dealing with the problem are different depending on whether we have control data at the individual level or spatially aggregated information on the size of the population at risk. We assume that case data are always available at the individual level (typically the postcode of residence in UK studies).

For individual level control data, the ideal situation is when all relevant explanatory variables are available at the individual level. In the present context, this would include, for

example, information on the individual's social class. As noted earlier, for variables which purport to describe the conditions of an individual's environment (for example, distance from a point source), there may be substantive difficulties even when the analysis is technically straightforward. If a relevant explanatory variable is available only as an average value for a subregion (for example, deprivation score), the options would seem to be the following.

- Pretend that the average applies to every location x within the subregion, and attach the average value to each individual accordingly.
- Recognise that if the explanatory variable, $z(x)$ say, is not constant within subregions, then $z(x) = \bar{z}_i + U(x)$, where $U(x)$ represents the local deviation of $z(x)$ from its subregional average value $\bar{z}_i$; we then need to think about appropriate statistical models for the deviations $U(x)$ which might, for example, be assumed to be a spatially smooth stochastic process. This is analogous to the empirical Bayes procedure used by Clayton and Kaldor[13] to estimate smooth spatial variation in risk.
- Aggregate everything to the subregional level and proceed as outlined at the end of the section above.

For spatially aggregated control data (for example, population counts for census wards or enumeration districts), the simplest option is the last of the above – that is, aggregate everything to the subregional level. As indicated earlier, the effect of this is to induce aggregation bias. To minimise this effect, we recommend using the smallest available areal units, for example, enumeration districts rather than electoral wards. Some covariates, such as average social class of head of household, based on a 10% sample, may be reliably available only at the ward level. We would then assign the same value of the covariate to each enumeration district within each ward. Note that if all relevant covariates are only available at ward level, this simply recovers the ward level analysis.

A slightly more subtle, but still approximate, approach would be to apply a variant of the integration equation (2), taking outside the integral sign the multiplicative factors arising from explanatory variables recorded at the subregion level (including the populations sizes N_i), and numerically integrating the remainder to give expected numbers of cases

$$\mu_i = N_i \exp\left(\sum_{j=0}^{p_1} \beta_j z_{ij}\right) \int_{A_i} \exp\left\{\sum_{j=p_1+1}^{p} \beta_j z_j(x)\right\} dx.$$

We turn now to the issue of residual spatial correlation, by which we mean that even after adjustment for the effects of explanatory variables, the numbers of cases in spatially adjacent subregions seem to be correlated. This represents a departure from the assumed Poisson process model. For non-infectious diseases, the most likely source of residual spatial correlation is the omission of one or more relevant explanatory variables which are themselves smoothly varying over the study region. In principle, this can be accommodated by the device of treating the combined effects of all such explanatory variables as an unobserved stochastic process, $U(x)$ say, the effect of which is to induce spatial correlation into the model for the data. This has a strong connection to generalised linear mixed models.[13 14]

In summary, there is no easy solution to the analysis of spatial data when some covariates (for example, age and sex of cases) are known at individual level, whereas others (for example, population, age–sex distributions, small area deprivation indices) are known only at the areal (ecological) level. However, we urge strongly that the underlying assumptions inherent in the analysis of such data are explicitly recognised, in order to understand better the limitations of the available methodology, as well as to inform the interpretation of results. Our current view is that the data should be kept as disaggregated as possible, both to maximise the information available and to minimise the potential for bias, and the entire analysis conducted at this minimal level of spatial aggregation.

Appendix

CONDITIONS FOR ABSENCE OF AGGREGATION BIAS

Recall that our model for cases and controls is that controls form a Poisson process with spatial intensity function $\lambda_0(x)$, and cases an independent Poisson process with spatial intensity $\lambda(x) = \lambda_0(x) r(x)$. Furthermore, for a partition of the study region into subregions A_i, with respective areas denoted by $|A_i|$, we define:

$$\mu_i = \int_{A_i} \lambda_0(x) r(x) dx$$

and $\mu_i^* = \lambda_{0i} \bar{r}_i$ where

$$\lambda_{0i} = \int_{A_i} \lambda_0(x) dx$$

and

$$\bar{r}_i = |A_i|^{-1} \int_{A_i} r(x) dx.$$

If Y_i denotes the number of cases in subregion A_i, an ecological analysis assumes that the expectation of Y_i is $E(Y_i) = \mu_i^*$, whereas in fact $E(Y_i) = \mu_i$.

Now, if we pick a point x uniformly at random within A_i, we induce a bivariate probability distribution for the pair of random variables $L = \lambda_0(x)$ and $R = r(x)$. Furthermore, L has expectation $E(L) = |A_i|^{-1} \lambda_{0i}$ and R has ex-

pectation $E(R) = \bar{r}_i$, whence:

$$\mu_i^* = |A_i| E(L)E(R).$$

Similarly, the product LR has expectation

$$E(LR) = |A_i|^{-1} \int_{A_i} \lambda_0(x) r(x) dx$$

whence

$$\mu_i = |A_i| E(LR).$$

Thus, the ecological bias induced by the spatial aggregation is given by

$$\begin{aligned} \mu_i^* - \mu_i &= |A_i| \{E(L)E(R) - E(LR)\} \\ &= -|A_i| \operatorname{Cov}(L, R). \end{aligned} \quad (10)$$

We call the covariance on the right hand side of equation (10) the *spatial covariance* between $\lambda_0(x)$ and $r(x)$, and conclude that the ecological bias is zero – that is, $\mu_i^* = \mu_i$, only when the spatial covariance is zero. This is trivially satisfied if either $\lambda_0(x)$ or $r(x)$ is constant throughout A_i, in which case an ecological analysis is both unbiased and fully efficient. Otherwise, it can only be true if $\lambda_0(x)$ and $r(x)$ are spatially uncorrelated – that is, if there is no linear association, however weak, between population density and risk.

1 Elliott P, Hills M, Beresford J, *et al.* Incidence of cancer of the larynx and lung near incinerators of waste solvents and oils in Great Britain. *Lancet* 1992;**339**:854–8.
2 Black D. *Investigation of the possible increased incidence of cancer in West Cumbria.* London: HMSO, 1984.
3 Elliott P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research* 1995;**4**:139–61.
4 Gardner MJ. Childhood leukaemia around the Sellafield nuclear plant. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small area studies.* Oxford: Oxford University Press, 1992;291–309.
5 Selvin HC. Durkheim's "Suicide" and problems of empirical research. *American Journal of Sociology* 1958;**63**:607–19.
6 Jolley D, Jarman B, Elliott P. Socio-economic confounding. In: Elliott P, Cuzick J, English D, Stern R, eds. *Geographical and environmental epidemiology: methods for small-area studies.* Oxford: Oxford University Press, 1992;115–24.
7 Cox DR, Isham V. *Point processes.* London: Chapman and Hall, 1980.
8 Daley DJ, Vere-Jones D. *An introduction to the theory of point processes.* New York: Springer, 1988.
9 Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. *Int J Epidemiol* 1990;**18**:269–74.
10 McCullagh P, Nelder JA. *Generalized linear models* (2nd ed). London: Chapman and Hall, 1989.
11 Fleiss JL. *The design and analysis of clinical experiments.* New York: Wiley, 1986.
12 Diggle PJ, Rowlingson BS. A conditional approach to point process modelling of raised incidence. *JR Stat Soc A* 1994; **157**:433–40.
13 Clayton D, Kaldor J. Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* 1987;**43**:671–81.
14 Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J Am Stat Assoc* 1993;**88**: 9–25.

Open discussion

ELLIOTT – The intensity function λ_0 that drops out of the analysis is contained in the controls; but, of course, as you say, Professor Diggle, you are then just moving the problem one step down the line and asking the epidemiologist or the data collectors to produce the population distribution for you. The problem is not resolved.

DIGGLE – No, it becomes the question, "Can you get what you regard from your practical expertise as an appropriate set of controls?" If you do there is no need to estimate the functions. They are just in the model to make your assumptions explicit. The thing that you are interested in, which is the risk, drops out of the analysis naturally. I am not trying to say that things should be done this way, but I do think that it behoves statisticians to model the truth as best they can and then ask if the model is fittable to the data they have and if not what *should* be fitted to the data? You should not start by saying, "These are the data I have and here is a model that looks as though it might fit". You should ask what is actually happening on the ground, model that, and then ask whether you can derive a statistical distribution for the data. If you can ... fit it, and if not ... then you begin the more delicate business of balancing practicality against idealism. As with many modelling exercises in my experience, this approach does not create problems, it reveals them, and it forces you to address them. Sometimes when you do not model the process you may try to tell yourself that you have solved the problem by putting in something like an extra Poisson variation parameter. I say "What does it mean? – it has no spatial interpretation".

STAINES – Your presentation was very illuminating, but it did not seem to address a real problem that occurs when the effects you are looking for are not individual effects. I think there is a large body of social theory as well as practical experience that suggests that many effects result from living in a particular area or in a particular community and these are inherent in area based effects. Individual level measurements will not address them.

DIGGLE – If my $Z(x)$ is piece-wise constant and if $Z(x)$ is really equal to 0·3 everywhere in an areal region, that is the value it will have in my model. If you believe you have covariates at that level they should go in the model at that level.

STAINES – If the effect is constant that would be both agreeable and astonishing because clearly the geographical boundary problem is actually insurmountable. What I am saying is different. I believe, for various reasons, that there are real effects which are not individual level effects – they are related to where you live, the social structure within which you live, and the community in which you live. This is not captured particularly intelligently by existing boundaries because the evidence available suggests that boundaries of enumeration districts are completely arbitrary. They mean absolutely nothing. Ward boundaries, although probably more meaningful in a conceptual sense, perhaps because of political rigging down the centuries, are still not very satisfactory. There is not a good way of approaching this problem, to capture areal level effects.

DIGGLE – No model can address and analyse data you have not got, but if you hypothesise that in some area everybody is exposed to the same effect because it is an effect of living in that area, then you should define that area and the statistician should analyse whether in fact your hypothesis is correct. I would then do that by attaching the same value $Z(x)$ to all individuals living within that area. But I would not wish to impose the use of arbitrary areas on you. That was my other point – I want the areas to be medically defined, not politically defined.