Modelling the Prevalence of Cancer of the Larynx in Part of Lancashire: A New Methodology for Spatial Epidemiology

P J DIGGLE, A C GATRELL, A A LOVETT
Lancaster University

1 Introduction
The availability of high-quality epidemiological data in Britain, notably from cancer registries, coupled with the advent of geographical information systems, has given fresh impetus to research into disease clustering. Such data are frequently available at a very fine level of spatial resolution, permitting the description and analysis of diseases in very localised regions. There is now a growing literature, with contributions from geographers, statisticians, and epidemiologists, on methods for detecting ‘clusters’ of such diseases (Elliott, 1988), and such methodologies are outlined briefly below. In this paper we follow in this tradition, but seek to go beyond the recognition of a cluster by formulating a model of disease prevalence that allows for the explicit testing of a hypothesised environmental association.

We may take a disease cluster to be “any localised spatial aggregation of cases” (Alexander et al, 1988, page 23), though we should of course recognise that diseases may be clustered in time and in space–time [as the seminal work of Knox (1964) suggests]. But, as Knox (1988) has suggested, we can approach the analysis of disease clusters in two ways. One approach is to define as a cluster any group of cases which is of sufficient size and concentration as to be unlikely to have occurred by chance. This is in essence the approach taken in Openshaw et al’s (1987) pioneering work on a geographical analysis machine (GAM). An alternative approach is to argue a priori that the cases are related to each other via a social or biological mechanism (as in a contagious disease) or as “having a common relationship with some other event or circumstance” (Knox, 1988, page 20). Such other events or circumstances might include hypothesised point sources of pollution (for instance, landfill sites or nuclear reprocessing installations) or possible linear sources of environmental impact (for example, busy roads or high-voltage electricity transmission lines). In postulating such environmental associations we would need to set up and evaluate a statistical model to evaluate the hypothesised relationship. As Bithell (1988, page 21) notes, “inferentially, we ought really to be concentrating as far as we can on prior hypotheses”. This is what the method we propose allows. We begin by reviewing some alternative approaches to the investigation of disease clusters before setting out this new method. We then explain why we have chosen to
model the prevalence of cancer of the larynx and we review what is known about the etiology of this disease. Results from applying the model to a district in Lancashire are then discussed.

2 Approaches to disease clustering

It is convenient to distinguish between approaches based on counting numbers of cases to be found in each of a set of areal units from those which treat cases as occurring at discrete sites (referenced by a pair of locational coordinates). Our own work falls squarely within the latter type of approach and we simply draw attention to the former, which often involves computing Poisson probabilities for small areas (for instance, see Barnes et al., 1987; Lovett et al., 1990) by using spatial autocorrelation tests to evaluate departures from randomness (Cliff and Haggett, 1988) or by using Bayesian statistics to estimate with use of information about other areas on the map the rates for areal units (Clayton and Kaldor, 1987).

The problem with such area-based approaches, as is well known, is that the results are very much dependent upon the given system of areal units (Openshaw, 1984). The techniques may fail to detect 'real' clusters through which one or more boundaries pass. Further, from the point of view of visualisation we almost always rely on administrative units which are arbitrary in size and shape and are not defined in terms of population at risk. This has led some authors (most recently, Selvin et al., 1988) to construct cartograms in which areal units are deformed such that the represented areas are proportional in size to population. The transformation equalizes density and the locations of disease cases may be carried across onto the transformed map. Selvin et al (1988, page 218) give a useful hypothetical example that reveals how natural 'clustering' of cases on an untransformed map arises simply because of variations in population density and how this clustering vanishes when cases are plotted on the cartogram. Investigating real diseases such as cancers, they calculate as a test statistic the average squared distance among cases and evaluate this under the null hypothesis of randomness.

This approach, then, serves as a link between area-based and point-based methods. The latter require very specific locational information, which will not always be available for reasons of confidentiality. However, when these data are available some method is needed to assess whether case 'clustering' is a function of factors other than variations in population density. One way of doing this is to use data on a control population (matched for factors other than the hypothesised 'risk' variable). Cuzick and Edwards (1990) propose an imaginative approach in which a nearest-neighbour graph is constructed on the set of all points (cases and controls together) and where the test statistic is the count of case-case joins.

There are close parallels here with a spatial autocorrelation approach.

Openshaw's GAM (Openshaw et al., 1987) is also a point-pattern analysis technique, though it relies on area-based population census data (at an enumeration district level) for computing expected prevalences. Briefly, a count is made of the number of cases within a circle of fixed radius, and if this exceeds the expected number the circle is plotted on a map. This test is performed for many overlapping circles of the same size, with the procedure repeated for circles of varying radius. Visually, a cluster of cases is recognized by a dense pattern of circles, and Openshaw has used the technique to confirm the existence of a concentration of childhood leukaemia around the Sellafield nuclear reprocessing plant and to reveal the existence of another around Gateshead. Openshaw is quite explicit in his philosophy of how to use GAM, namely as an exploratory tool, "relating the importance of specifying a hypothesis from a locationally specific to a locationally unspecific form. It is noted that lack of knowledge generally precludes more precise formulations of hypotheses in cancer studies" (Openshaw and Craft, 1988, page 35). He recommends GAM as a search tool, leading to more detailed epidemiological studies when we know which areas to target.

3 A new methodology

Our own approach is set more firmly on a foundation of statistical modelling, and we seek to test an explicit hypothesis, namely whether there is an association between a possible point source of pollution and the distribution of a cancer. We also start from the premise that methods based on discrete areal units are always subject to the criticism that the results are specific to that particular set of units. Rather than being constrained by fixed (though essentially arbitrary) area units such as wards or enumeration districts, we propose an approach based on spatial point processes (Diggle, 1983). [For a full discussion of this method, see Diggle, 1990].

Let us take a spatial distribution of diagnosed cases of some disease represented geographically as a set of points whose grid references are known residential locations. In the absence of any prior information we might suggest an homogeneous Poisson process with constant intensity, \( \lambda \), as an appropriate null model for these cases. Of course, given known spatial variations in population density this process is not reasonable a priori and we need to consider an inhomogeneous Poisson process in which \( \lambda(x) \) varies with location and whose mean is \( \int_A \lambda(x) \, dx \), where \( x \) represents location and \( A \) is the area of interest. [Note that \( x \) is a vector comprising a pair of of locational coordinates \( (x_1, x_2) \).]

Suppose that the local intensity of cases is expressed as:

\[
\lambda(x) = \rho \lambda_0(x) f(x, \theta),
\]

where \( \rho \) represents the overall prevalence of the disease and \( \lambda_0(x) \) represents the spatial variation in local intensity under the null
hypothesis. Let \( f(.) \) represent the hypothesis of interest here, namely that \( \lambda(x) \) depends upon distance from a postulated point source of pollution; \( \theta \) represents a set of parameters to be estimated. Below, we choose \( f(x; 0) = 1 \) so that \( \theta = 0 \) represents the null hypothesis of no association between local intensity of cases and the postulated source. That is to say, if there is no association the intensity of cases is the same as the intensity of controls, scaled by the overall prevalence of cases.

A log-likelihood, \( L(\rho, \theta) \), for \( \theta \) is given by

\[
L(\rho, \theta) = n \ln \rho + \sum_{i=1}^{n} \ln f(x_i, \theta) - \rho \int \lambda_0(x) f(x, \theta) \, dx .
\]  

(2)

Differentiating with respect to \( \rho \), we obtain a maximum-likelihood estimate for \( \hat{\rho} \):

\[
\frac{\partial L}{\partial \rho} = \frac{n}{\rho} - \int \hat{\lambda}_0(x) f(x, \theta) \, dx = 0 ,
\]

(3)

\[
\hat{\rho} = \frac{1}{n \int \hat{\lambda}_0(x) f(x, \theta) \, dx} .
\]

(4)

Inserting this expression for \( \hat{\rho} \) into equation (2), we get

\[
L(\theta) = -n \ln \left[ \int \hat{\lambda}_0(x) f(x, \theta) \, dx \right] + \sum_{i=1}^{n} \ln f(x_i, \theta) .
\]

(5)

The problem thus becomes one of trying to estimate the background intensity function \( \hat{\lambda}_0(x) \), as this is unknown. We could make use of census data but because these are only available for discrete areal units we prefer a different approach, that is, to use the observed spatial distribution of another, more common disease, one which matches the disease of interest in terms of age–sex structure but which, ideally, is not associated with the hypothesised point source of pollution. Our choice of the ‘control’ disease is discussed below, but let us assume for the present it is given. We need then to convert the control locations into a continuous intensity function. This is done by smoothing the point pattern of control locations.

This smoothing of point patterns has been investigated by Diggle (1985) and Berman and Diggle (1989), who propose, for any location \( x \):

\[
\hat{\lambda}_0(x) = \frac{1}{h^2} \sum_{j=1}^{m} w \left( \frac{x - y_j}{h} \right) ,
\]

(6)

where \( y_j \) \((j = 1, \ldots, m)\) are the locations of the \( m \) controls and \( w \) is a ‘kernel function’ or radially symmetric probability density function, such as the Gaussian:

\[
w(x) = \frac{1}{2\pi} \exp \left( -\frac{1}{2} x'x \right) .
\]

(7)

The amount of smoothing is determined by \( h \), which is chosen to minimise the mean square error of \( \hat{\lambda}_0(x) \). We then have sufficient information to obtain a maximum-likelihood estimate of the parameter(s) \( \theta \) and to evaluate the null hypothesis (\( H_0 \)) that \( \theta = 0 \). In principle we can say that

\[
D(\theta) = 2[L(\hat{\theta}) - L(\theta)]
\]

(8)

is approximately distributed as \( \chi^2 \), with degrees of freedom equal to the dimensionality of \( \theta \) (number of parameters to be estimated).

4 Cancer of the larynx
We seek now to test the hypothesis that cancer of the larynx in a district of Lancashire is associated with proximity to a closed-down industrial waste incinerator. We begin by outlining the background to this empirical study before reviewing what is known of the disease aetiology.

Between 1972 and 1980 an industrial waste incinerator operated at a site about 2 km southwest of the small town of Coppull in Lancashire. The incinerator was used for the disposal of a wide range of waste products, both solids and liquids (such as solvents and oils). During its period of operation there were complaints about noxious smells and about respiratory problems. The opening of the plant preceded the setting up of pollution regulations, under the Control of Pollution Act (1974). Since its closure there have been lingering worries about long-term ill-health effects, and two of us (Gatrell and Lovett) were asked to conduct a preliminary study of the distribution of cancers throughout the area (Chorley and South Ribble District Health Authority) with a view to suggesting whether there was any map evidence of possible clustering in the vicinity of the incinerator. To this end, we were given postcoded data on 6200 cases of cancer diagnosed between 1974 and 1983—patients who were resident in the District at the time of diagnosis. The full unit postcode can be associated with an Ordnance Survey grid reference by using the Central Postcode Directory—a large computer file that lists all 1.5 million unit postcodes in Britain together with corresponding grid references accurate to 100 m (for further details and commentary on the nature of this postcode matching, see Gatrell, 1989).

We found nothing of interest in most of the maps, save for a rather odd distribution of laryngeal cancer (figure 1). There were only 58 cases of this cancer during the 10-year period and the distribution was sparse, with relatively few cases in the more densely populated areas (such as south Preston, Chorley, and Leyland). However, there were 5 cases in the village of Coppull, 4 of which lay within 1 km of the incinerator and another within about 2 km. We regarded this as suspicious, particularly in view of our prior knowledge of the existence of a possible point source of pollution. It therefore seemed worthwhile to attempt to verify by using a statistical model whether or not there was any evidence for local clustering.
What evidence is there, if any, to link cancer of the larynx to proximity to an incinerator? More precisely, are there any known carcinogens which might have been released during the combustion of industrial, especially toxic, wastes?

Kleinsasser (1988) has summarised chemical environmental factors that appear to be implicated in laryngeal cancer aetiology. The evidence comes primarily from occupational studies rather than from those affected outside the workplace, and, as Kleinsasser notes, it is difficult to pinpoint individual noxious agents as causes of laryngeal cancer. However, he refers to polycyclic aromatic hydrocarbons (PAHs) as well-known carcinogens and these are by-products of combustion processes (see also Greenberg, 1988). Asbestos fibres are also considered as a causative factor (Stell and McGill, 1975), and poor disposal of any asbestos wastes may release these into the atmosphere. Kleinsasser also reports on exposure to arsenic, formaldehyde, vinyl chloride, sulphuric acid, cadmium, nickel, and chrome dust as risk factors (see also Flanders et al., 1984). Exposure to mustard gas in the field and in the workplace has been identified as a likely cause of some carcinomas. Last, thermal radiation, as experienced by those working in foundries and coke ovens, is said to be a risk factor, that is, warm air containing "all manner of gases, steam and dust, which may act as carcinogens, although individually they can hardly be identified" (Kleinsasser, 1988, page 18).

Even if we knew the precise details of substances burnt by the incinerator in question during the 1970s and if we had conducted environmental monitoring of stack gases and ground deposition to detect some of the above substances, it would be difficult to infer causal relationships. "The intensity and length of exposure, the age at the time of exposure, and other factors, such as the ubiquitous cigarette smoke, also play a part" (Kleinsasser, 1988, page 18). There is unambiguous evidence that tobacco consumption is the major risk factor in developing the cancer. Burning the tobacco releases tar containing several species of PAHs. Kleinsasser estimates that at least 90% of patients with laryngeal cancer are, or have been, active smokers. Heavy consumption of alcohol is also considered to be a promoting factor in the disease, but one which interacts with smoking to generate an increased risk where both are consumed in excess (Guenel et al., 1988).

All this is to suggest that any attempt to infer causation from a possible point source of pollution is fraught with difficulties. We do not have any information on individuals' smoking behaviour, nor much detail on likely occupational exposure. We do know the age and sex of each case and although the sample (n = 58) is small the age-sex distribution is roughly in line with that for larger samples (figure 2).

The length of time between exposure and diagnosis causes a further difficulty. This time is highly variable, ranging from perhaps 4 or 5 years to 30 years or more (Stell and McGill, 1975, page 516). If it could be demonstrated that the 5 cases close to the incinerator were long-time residents of the areas, with no history of heavy smoking and no evidence of occupational exposure, the argument in favour of the incinerator as a

**Figure 1.** Distribution of cancer of the larynx in Chorley and South Ribble District Health Authority. (Axes refer to the Ordnance Survey grid.)

**Figure 2.** Age distribution of patients with laryngeal cancer (based on European cancer registries). Source: Kleinsasser, 1988.
likely causative factor would be strengthened. We simply do not have enough evidence to argue one way or another.

5 Results
We took data on cancer of the lung (978 observations in our study area) as a variable used to construct the background intensity function, \( \lambda_0(x) \). We clearly required a variable which is similar to the disease of interest, and in terms of age and sex structure this is the case for these two cancers. Smoking, too, is a major risk factor for both diseases. Alcohol consumption is a risk factor for laryngeal but not for lung cancer, but it is difficult to argue that this introduces a major bias. Strictly speaking, the approach is not valid if lung cancer is associated with the pollution source. If it is, the lung cancer 'controls' are overmatched and the likelihood of finding a significant relationship with proximity to the incinerator is reduced. So we might argue that this particular concern is serious only if the hypothesis of no association is accepted. We reiterate the view that our approach seems preferable to one that uses appropriate age-sex matched controls drawn from areally aggregated census data.

A plot of the lung cancers (figure 3) reveals clumps of local intensity that correspond to centres of population (primarily south Preston, Leyland, and Chorley). We smoothed this point pattern according to the method outlined above, using \( h = 0.30 \) (figure 4). This is the value which minimises the estimated mean square error of \( \lambda_0(x) \), as the plot of this error against \( h \) reveals (figure 5).

In order for the model to fit [equation (1) above], we needed to specify an explicit functional form for \( f(x; \theta) \). We used:

\[
f(d) = 1 + \alpha \exp(-\beta d^2),
\]

where \( d \) is distance from the incinerator, and \( \alpha \) and \( \beta \) are parameters to be estimated. The use of a squared-distance term is arbitrary, but we felt this appropriately reflected a likely pollution-plume effect. The parameter \( \alpha \) reflects the risk at source, and \( \beta \) represents variation with distance. Clearly, if there is no distance effect then \( f(d) = 1 \) and our estimate of the intensity of laryngeal cancer at a point is simply the estimated intensity of lung cancer, scaled by \( \rho \) [equation (1)].

The likelihood function [see equation (5)] was evaluated, and maximum-likelihood estimates of \( \alpha \) and \( \beta \) were generated. The maximised log-likelihood is \(-394.593\), which is to be compared with a log-likelihood of \(-399.360\) under the null hypothesis. Evaluation of twice the difference between these values as \( \chi^2 \), distributed with 2 degrees of freedom \((\chi^2_{0.05} = 5.99)\) suggests that \( H_0 \) is to be rejected. There are many possible combinations of \( \alpha \) and \( \beta \) which generate a maximised log-likelihood close to \(-394.593\) (figure 6), and so it is difficult to be precise about an estimated function \( f(d) \). One possibility, where \( d = 25.26 \) and
\( \hat{\beta} = 0.952 \), is sketched in figure 7, but as the surface of the log-likelihood would suggest, values of \( 10 \leq \hat{\alpha} \leq 200 \) and \( 0.5 \leq \hat{\beta} \leq 2.5 \) are all plausible.

![Graph showing mean square error vs distance](image)

**Figure 5.** Mean square error, \( t \), of the background intensity function, \( \lambda_0(x) \), as a function of distance.

![Contour plot](image)

**Figure 6.** Contour plot of \( D(\alpha, \beta) = 2(L(\hat{\alpha}, \hat{\beta}) - L(\alpha, \beta)) \). Values of \((\alpha, \beta)\) within the 6.0 contour constitute an approximate 95% confidence region.

The reason for this uncertainty is the fact that there is a very marked clump of cases within about 2 km of the incinerator and little else within a radius of 5 km.

We conclude that proximity to the incinerator has some effect on the prevalence of laryngeal cancer. However, we should caution that we are indeed dealing with very small numbers. Taking the 5 cases within 2 km of the incinerator and removing one of these from the analysis, we generate a difference between the maximized and null log-likelihoods, which just fails to reach significance at the 0.05 level \( D(\theta) = 5.82 \). Removing two cases, we reduce \( D(\theta) \) to 2.96. The addition of an imaginary case in the same general area gives \( D(\theta) = 12.95 \). Clearly, then, we cannot escape the fact that we are dealing with small numbers.

![Graph showing function of distance](image)

**Figure 7.** The estimated function of distance, \( f(d) \).

### 6 Conclusions

In geographical epidemiology, there is always a possibility of plotting cases of a disease, detecting 'clusters' by eye, and looking for a possible source of environmental pollution to which one seeks to relate cases. This retrospective formulation of hypotheses is to be studiously avoided. What we have done is to suggest that the distribution of laryngeal cancer may be associated with proximity to an individual waste incinerator. Clearly, we need now to test this hypothesis in other areas. To this end we are extending the study to examine laryngeal and lung cancer in the North West Regional Health Authority as a whole and we will assess other incinerators as possible sources. Obvious candidates here are hospital incinerators—problems with which nationally have already been identified (NSCA, 1988).

We shall also explore the possibility of varying the location of the putative source of pollution within the present study area. This parallel
Openshaw's GAM approach and will generate a surface of maximum-likelihood estimates across the map. It will then be possible to assess the significance of the incinerator relative to other locations on the map.

It is also possible in principle to extend the method to incorporate other explanatory variables, where these too are measured at point locations. We see the method we have outlined here in having potentially wide applicability in geographical epidemiology, notably in testing hypotheses concerning the possible impact of nuclear installations on human health.

Acknowledgments. The data come from the North West Regional Cancer Registry, courtesy of Professor A Smith, University of Manchester. We are grateful to Doctor S Sivayocham (Specialist in Community Medicine, Chorley and South Ribble Health Authority) for his support. Doctor A Sudell, Senior Registrar in Community Medicine, Preston Health Authority, has acted as an excellent sounding board throughout this work.

References
Alexander F E, Cartwright R A, McKinney P M, 1988, “A comparison of recent statistical techniques of testing for spatial clustering: preliminary results”, in Methodology of Enquiries into Disease Clustering Ed. P Elliot (Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine, 25 Keppel Street, London WC1) pp 23–33
Bithell J, 1988, “Discussion of paper by Knox (1988)”, in Methodology of Enquiries into Disease Clustering Ed. P Elliot (Small Area Health Statistics Unit, London School of Hygiene and Tropical Medicine, 25 Keppel Street, London WC1) page 21