SESSION: Putative Source Analysis

Key Issues:

Objectives of these studies are:

- To assess the effect of a pollution source or sources on the health status of an area
- To make inferences about the existence of pollution in an area
- To respond retrospectively to *alarms*

Scenario 1: a local community is concerned about the effect of an incinerator on the health of the community.

Scenario 2: A cluster of disease incidence is thought to exist in a local area and it is to be assessed and analysis of its relation to (putative) local pollution sources is to be made

Scenario 3: A known pollution source (i.e. health hazard) is to be monitored for its effect on the a local population

In 1 and 2, usually a retrospective assessment of the incidence is required. In 3, a prospective analysis can be planned.

Study Design Issues

In what follows, we consider a delimited geographical study area or window within which data concerning disease occurrence and exposure to the pollution source are collected. Issues concerning the strategic aims of the study must be considered prior to detailed consideration of the appropriate study region and data collection requirements.

Retrospective and Prospective Studies

During the 1980’s, a number of studies of disease occurrence in geographical regions around putative sources of risk were carried out. Most of these were ‘reactive’, in that suspicion of a health risk, due to the past operation of a pollution source, instigated a review of the historical evidence for a link between disease incidence and exposure to the source. In essence, a *retrospective* study of disease occurrence was carried out. In some cases, continued monitoring of the source was also recommended or initiated. However, solely *prospective* studies of sources are seldom encountered. These two approaches and their respective strengths and weaknesses are well-known in the epidemiological literature.

Such studies of effects of pollution have a number of limitations, however. First,

- typically the emission characteristics of a source are not recorded for a suitable time period. Retrospective data on emissions may not be available and prospective monitoring data is expensive to collect over a long time period for a wide range of substances of interest.
- Often, no direct information is available on correlation between emission and disease occurrence.
- Furthermore, exposure and disease data are often collected by separate groups at different levels of resolution (even in prospective studies).

- the nature of available data may be limited for particular diseases or health status indicators, or for particular time periods.
- Often, nationally-collected data rather than data from a specially designed study must be utilised. In some cases, the level of resolution in available data constrains the analysis considerably. For example, some diseases are reported only as counts from postal zones or census enumeration districts and not as exact addresses due to confidentiality. In that case, methods based on analysis of counts rather than point events are appropriate.

Inevitably, such regionalisation leads to some loss of information. For example, very small clusters cannot be detected if they occur within a large census tract as the aggregate disease rate for the tract as a whole may not differ from the background disease rate. Only if the spatial pattern of events occurs at a larger scale than the measurement unit will it be detectable in regionalised data.

- for chronic outcomes like cancers, the temporal lag between exposure and an event of interest may be on the order of years or decades. Mobility of individuals over such a time period can confound exposure-outcome relationships and cause prohibitive costs in prospective studies over large areas.

Study Region Design

The design of a study region or window is of great practical importance. Usually, a study will concern the distribution of events (e.g. incident disease cases) within a fixed map area of given size and shape. The choice of size and shape can have considerable impact on study results and, while often it is not possible to choose the most appropriate region, some consideration should be given to these issues.

Region Size

A study region should be defined which is of sufficient size that any effects of a putative source can be measured adequately. As it is often not possible to assess, a priori, the spatial scale of pollution effects, it is therefore important that a large region including the pollution source should be used. In many published studies a region is defined and the total incidence in the region is analysed (compared to external ‘control’ regions). The Lenihan report (1985) provides an example of this approach. If a region is specified which is larger than the true pollution range then a localised effect within some part of the region may be diluted. On the other hand, a small region may truncate the evidence and not represent the complete effects in the population. In addition, the use of multiple region sizes may still induce problems in data analysis if a pollution effect occurs at a spatial scale different than those considered.

- In previous studies, sizes of region, in radial units from a source, vary from less than 1 kilometre to 10 kilometres. Most study windows have areas between 10 and 100 square kilometres. Often, the size of region is defined by a natural break in the underlying population. For example, the boundary of a town or physical barriers such as rivers, mountains, or coastlines may affect the region size (and shape). Practical data acquisition problems may limit the region size.

Furthermore, exposure and outcome data may be available for different regions.

Region Shape

- When one assesses exposure to a single pollution source, and one assumes that distance is a surrogate for exposure, then a circular region centred on the source yields the least sampling bias for detecting directional trends, in that sampling is equal in all directions. Square, rectangular, or other polygonal regions do not provide such unbiasedness. Of course, if the putative source is not central to the region then a circular window has no advantage. If population structure dictates the region shape and size then a polygonal region may have to be adopted, although some advanced statistical techniques can be used to allow for population sparse regions in regular windows.
- When one examines multiple pollution sources, a rectangular or polygonal region should suffice. However, one should make some effort to provide ‘similar’ sampling detail in all directions from the sources in

case directional differences are present.

Replication and Controls

Few studies examine replicated realisations of disease events around pollution sources.

The main use of replication in such studies should be to provide estimates of variability not available from single realisations. An alternative use of replication is to study other areas where potential pollution sources exist but where no evidence has been demonstrated for adverse health links to the source or sources.

- If substantial hypotheses concerning an individual source are to be examined then control areas may be of some use. However, the use of replication to provide increased sample size by pooling, without examination of variability, only provides evidence for hypotheses concerning the sources in general, and not as individual sites. Local effects, which may be ‘unusually’ marked at an individual site may be swamped in such a pooled sample.
- In any study of disease incidence within a

population, one must take some account of population structure. A standard epidemiological case-control design can be used where individuals are selected as controls and matched to cases with respect to confounding factors (e.g. age and occupation). Another standard approach in the conventional analysis of small area count data involves the use of strata-specific standardised rates to represent the ‘background’ population effect. The ratio of observed count to expected count, based on such a standardisation, can be used as a crude estimate of region-specific relative risk.

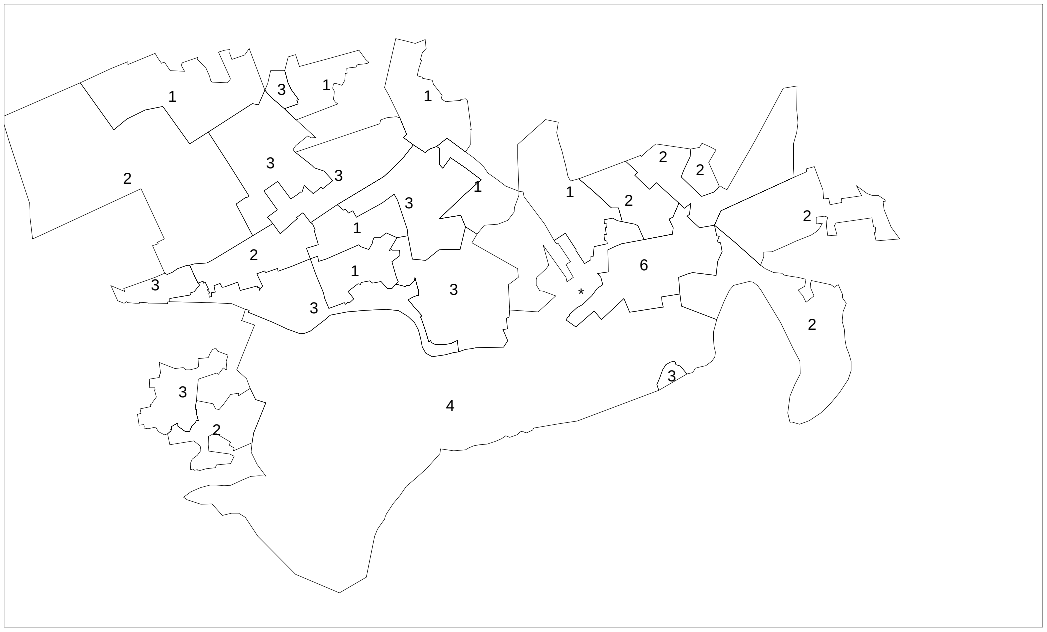
- An alternative approach is to utilise a disease or group of diseases which is thought to represent the ‘at risk’ population in the area but is usually unaffected by the type of pollution being considered. This approach is designed for point event data where a ‘background’ point event map of a ‘control’ disease is available. This method could also be used with count data, where counts of ‘case’ and ‘control’ diseases are available.

- The goal is to find a ‘control’ disease which affects the same population with respect to possible confounding variables (e.g. age, occupation, smoking, etc.) yet is unrelated to the exposure of interest. While the existence of such a ‘control disease’ is subject to epidemiological debate, if such data are available, the statistical foundation of the methods is sound.

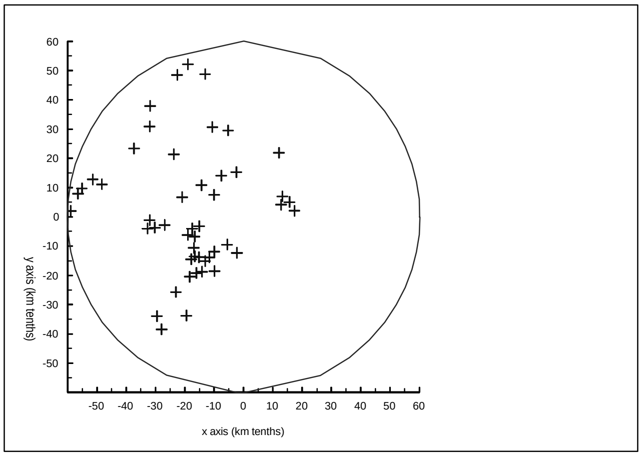
———————————————————-

A number of studies utilise data based on the spatial distribution of such diseases to assess the strength of association with exposure to a pollution source. Raised incidence near the source, or directional preference related to a dominant wind direction may provide evidence of such a link. Hence, the aim of the analysis of such data is usually to assess the effect of specific spatial variables rather than general spatial statistical modelling. That is, the analyst is interested in detecting patterns of events near (or exposed to) the focus and less concerned about aggregation of events in other locations. The former type of analysis has been named *‘focussed clustering*’ by Besag and Newell. The latter is often termed *’non-focussed’ clustering.* To date, most pollution-source studies concentrate on incidence of a single disease (e.g., childhood leukaemia around nuclear power stations or respiratory cancers around waste-product incinerators).

The types of data observed can vary from disease-event locations (usually residence addresses of cases) to counts of disease (mortality or morbidity) within census tracts or other arbitrary spatial regions. An example of a data set consisting of residential locations is provided in the Armadale figure. The locations of respiratory cancer cases around a steel foundry (0,0) in Armadale central Scotland for the period of 1968-1974. In this example, the distribution of cases around the central foundry is to be examined to assess whether there is evidence for a relation between the locations and the putative source (the foundry). In the other figure, the counts of respiratory cancer for the period of 1978-1983 in Falkirk central Scotland are displayed. A number of putative sources of health hazard are located in this area, most notably an metal processing plant (*).



+ local foundry site



Armadale

Inference Problems

The primary inferential problems arising in putative-source studies are

- post hoc analyses,
- multiple comparisons.

The well-known problem of *post hoc analysis* arises when prior knowledge of reported disease incidence near a putative source leads an investigator to carry out statistical tests or to fit models to data to ‘confirm’ the evidence. Essentially, this problem concerns bias in data collection and prior knowledge of an apparent effect. Both hypothesis tests and study-region definition can be biased by this problem. However, if a study *region* is thought *a priori* to be of interest because it includes a putative pollution source, one does not suffer from post hoc analysis problems if the internal spatial structure of disease incidence did not influence the choice of region.

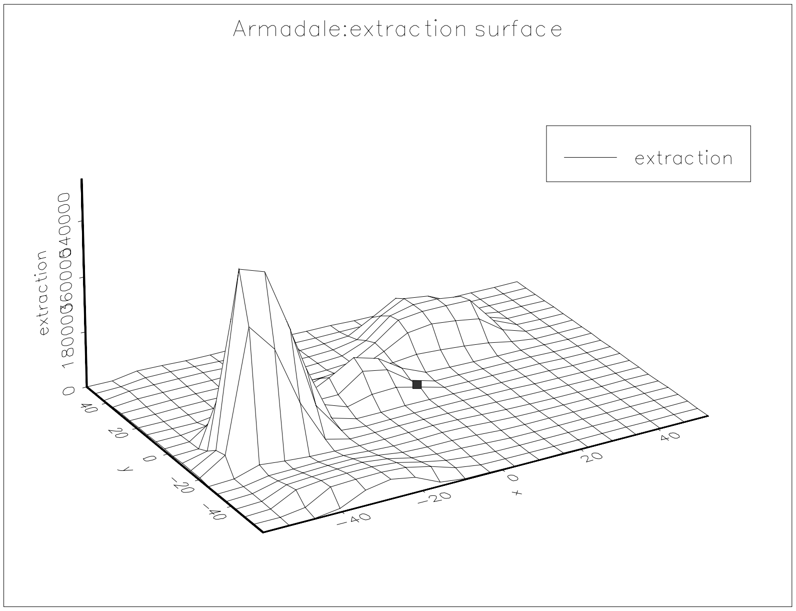
The *multiple-comparison problem* has been addressed in several ways. Bonferonni’s inequality may be used to adjust critical regions for multiple comparisons but the conservative nature of such an adjustment is well known. The use of cumulative p-value plotting has been proposed to assess the number of diseases yielding evidence of association with a particular source. An alternative approach is to specify a general model for the incidence of disease or diseases. Such an approach can often avoid multiple comparison problems.

Exploratory Techniques

The use of exploratory techniques is widespread in conventional statistical analysis. However, in point-source analysis one must exercise care about how subsequent analysis is influenced by exploratory or diagnostic findings. For example, if exploratory analysis isolates a cluster of events located near a pollution source, then this knowledge could lead to a post hoc analysis problem, namely, inference based on a model specifically including such a cluster is suspect.

For case-event data, one can employ standard point process methods to explore data structure. For example, the intensity (i.e. points per unit area) of events can be mapped and viewed as a contoured surface, usually using density estimation.

If the intensity of controls is also mapped, then it is useful to assess whether the cases demonstrate an excess of events beyond that demonstrated by the controls (e.g. in areas of increased risk). Controls could consist of randomly selected individuals from the population at risk (perhaps matched on confounding factors), or a ‘control disease’ as mentioned above. A higher relative intensity of ‘cases’ to ‘controls’ near a pollution source, as compared to far away, may support a hypothesis of association.



In the case of tract-count data, a variety of exploratory methods exist. One can use representation of counts as surfaces and incorporate expected count standardisation (e.g. through a standardised mortality/morbidity ratio (SMR)). While mapping regional SMRs can help isolate excess incidence, estimates of SMRs from counts in small areas are notoriously variable, especially for areas with few persons at risk. Various methods have been proposed to stabilise these small area estimates. Two different approaches are based on nonparametric smoothing and empirical Bayes 'shrinkage' estimation.

In general, the use of nonparametric relative risk estimation, particularly combined with Monte Carlo evaluation of excess risk, is a powerful tool for the initial assessment of risk elevation. Care must be taken, however, not to prejudice further inference by the *a posteriori* focussing of analysis.

Models for Point Data

In this section, we consider a variety of modelling approaches available when data are recorded as a point map of disease case events.

- Event locations often represent residential addresses of cases and take place in a heterogeneous population that varies both in spatial density and in susceptibility to disease.
- Define the first-order intensity function of the process as $\lambda(\mathbf{x})$, which represents the mean number of events per unit area (local density) in the neighbourhood of location $\mathbf{x}$. This intensity may be parameterised as :

$$\lambda(\mathbf{x}) = \rho . g(\mathbf{x}) . f(\mathbf{x};\boldsymbol{\theta})$$

where $g(\mathbf{x})$ is the ‘background’ intensity of the population at risk at $\mathbf{x}$, and $f(\mathbf{x};\boldsymbol{\theta})$ is a parameterised function of risk relative to the location of the pollution source. The focus of interest for assessing associations between events and the source is inference regarding parameters in $f(\mathbf{x};\boldsymbol{\theta}),$ treating $g(\mathbf{x})$ as a nuisance function. The log-likelihood of m events in A(the study region), conditional on m, is (bar a constant) :

$$\sum_{i=1}^{m} \log f(\mathbf{x}_i;\boldsymbol{\theta}) - m \log \int_A g(\mathbf{x})\, f(\mathbf{x};\boldsymbol{\theta}) d\mathbf{x}.$$

Here, parameters in $f(\mathbf{x};\boldsymbol{\theta})$ must be estimated as well as $g(\mathbf{x})$.

- Inferential problems arise when $g(\mathbf{x})$ is estimated as a function and then apparently regarded as constant in subsequent inference concerning $\lambda(\mathbf{x})$. One solution to this problem is to incorporate the estimation of the background smoothing constant in $\widehat{g}(\mathbf{x})$, by the use of a prior.

The specification of $f(\mathbf{x};\theta)$

It is important to consider the appropriate form for the function $f(\mathbf{x};\theta)$, which usually describes the exposure model used in the analysis of the association of events to a source. Define the location of the source as $\mathbf{x}_0$. Usually the spatial relation between the source and disease events is based on the polar coordinates of events from the source: $\{r, \phi\}$, where $r = \|\mathbf{x} - \mathbf{x}_0\|$, and ϕ is the angle measured to the source. It is important to consider how these polar coordinates can be used in models describing pollution effects on surrounding populations.

- In many studies, only the distance measure (r) has been used as evidence for association between a source and surrounding populations. However, it is dangerous to pursue distance-only analyses, when considerable directional effects are present. The reason for this is based on elementary exposure modelling ideas, which are confirmed by more formal theoretical and empirical exposure studies. It is clear, that differential exposure may occur with change in distance *and* direction, particularly around air pollution sources (such as incinerator stacks or foundry chimneys). Indeed the wind regime which is prevalent in the vicinity of a source can easily produce considerable differences in exposure in different directions. Such directional preference or anisotropy can lead to marked differences in exposure in different directions and hence to different distance exposure profiles. Hence the collapsing of exposure over the directional marginal of the distribution could lead to considerable mis-interpretation, and in the extreme to *Simpson’s* paradox. In the extreme case, a strong distance relationship with a source may be masked by the collapsing over directions, and this can lead to erroneous conclusions.

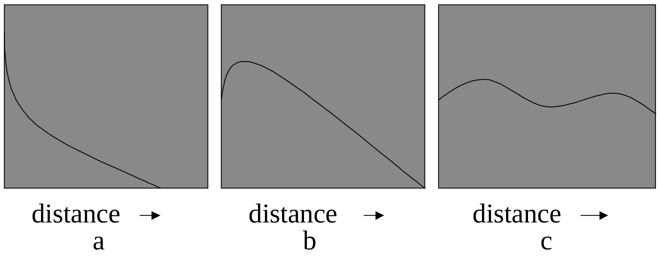
- The importance of the examination of a *range* of possible indicators of association between sources and health risk in their vicinity is clear. The first criterion for association is usually assumed to be evidence of a decline in disease incidence with increased distance from the source. Without this distance-decline effect, there is likely to be only weak support for an association. However, this does not imply that this effect should be examined in isolation. As noted above, other effects can provide evidence for association, or could be nuisance effects which should be taken into consideration so that correct inferences be made. In the former category are directional and directional-distance correlation effects, which can be marked with particular wind regimes. In the latter category are peaked incidence effects, which relate to *increases* of incidence with distance from the source. While a peak at some distance from a source can occur, it is also possible for this to be combined with an overall underlying decline in incidence, and hence is of importance in any modelling approach. This peaked effect is a nuisance effect, in terms of association, but it is clearly important to include such effects. If they were not included then inference may be erroneously made that no distance-decline is present, when in fact a combination of distance-decline and peaked incidence is found. Further nuisance effects which may be of concern are e.g. random effects related to individual *frailty*, where individual variation of susceptibility is directly modelled or where general heterogeneity is admitted.

- A general approach to modelling exposure risk is to include an appropriate selection of the above measures in the specification of $f(x;.)$. First it is appropriate to consider how exposure variables can be linked to the background intensity $g(\mathbf{x})$. We define $f(\mathbf{x};\theta) = m\{f^*(\mathbf{x})'\alpha\}$, where $m\{.\}$ is an appropriate link function, and $f^*(\mathbf{x})$ represents the design matrix of exposure variables which is evaluated at $\mathbf{x}$. The link function is usually defined as $m\{.\} = 1 + \exp\{.\}$, although a direct multiplicative link can also be used. Usually each row of $f^*(\mathbf{x})$ will consist of a selection of the variables

$$\{r, \log(r), \cos(\phi), \sin(\phi), r\cos(\phi),$$

$$r\sin(\phi), \log(r)\cos(\phi), \log(r)\sin(\phi)\}.$$

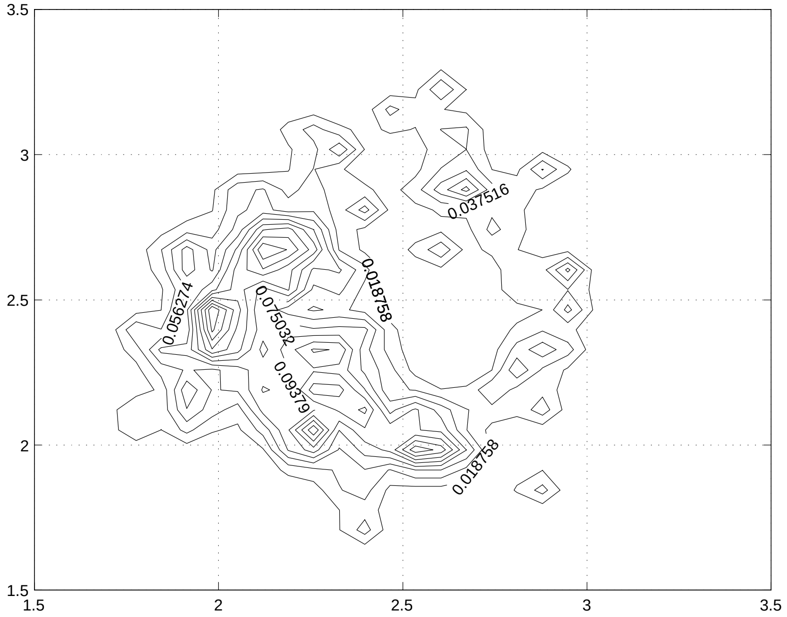
- The first four variables represent distance-decline, peakedness, and directional effects, while the latter variables are directional-distance correlation effects. The directional components can be fitted separately and transformations of parameters can be made to yield corresponding directional concentration and mean angle. The figure below displays different distance-related exposure models which could be used to specify $f(\mathbf{x};\boldsymbol{\theta})$. Note that nuisance effects of peakedness and heterogeneity appear in b) and c).



Distance-risk relationships

- Further examination of dispersal models for air pollution, suggest that the spatial distribution of outfall around a source is likely to follow a convolution of Gaussian distributions where in any particular direction there could be a separate mean level and lateral variance of concentration (dependent on r).
- As a parsimonious representation of these effects it is possible to use a subset of the exposure variables listed above to describe

this behaviour. The following figure displays the result of a simulation for a model which involves both peaked and distance-decline components and directional preference. Time-averaged exposure can be thought to lead to patterns similar to that depicted. Here a NW direction of concentration is apparent and the simulated exposure intensity surface was obtained from a 5 parameter model for the distance and directional components. Note that averaging over the directional marginal of this distribution will lead to considerable attenuation of increased risk at distance from the source due to the anisotropic distance relations found.



Estimation

The parameters of the models discussed above, can be estimated by maximum likelihood, conditional on $\widehat{g}(\mathbf{x})$. In fact, it is possible to use GLIM or S-PLUS for such model fitting. Bayesian models can be sampled using MCMC methods.

Hypothesis Tests

Note that both likelihood-ratio (LR) and score tests are available in statistical software packages (such as GLIM, GENSTAT or S-PLUS).

- Tests of monotonic radial decline assume that distance acts as a surrogate for exposure. Many proposed tests are based on radial decline models in point data and tract-count data. Simple radial decline tests can have low power when nonmonotone effects, such as those discussed above, are present.
- The collection of data and spatial modelling of exposure levels should lead to increased power to detect pollution effects. Unobserved heterogeneity may be included as random effects following the generalized linear mixed models. Alternatively, the heterogeneity may be formulated in terms of nuisance parameters. Lawson and Harrington (1996) examined Monte Carlo tests, in a putative source setting, when spatial correlation is present and can be estimated as a nuisance effect under the null hypothesis.

Models for count data

For a variety of reasons, outcome data may be available only as counts for small census regions rather than as precise event locations. As a result, a considerable literature has developed concerning the analysis of such data.

- Analysis based on regional counts is ecological in nature and inference can suffer from the well-known ‘ecologic fallacy’ of attributing effects observed in aggregate to individuals.
- extreme sparseness in the data (i.e., large numbers of zero counts) can lead to a bimodal marginal distribution of counts or invalidate asymptotic sampling distributions.

While the above factors should be taken into consideration, the independent Poisson model may be a useful starting point from which to examine effects of pollution sources. Often, a log-linear model parameterisation is used, with a modulating value e_i, say, which acts as the contribution of the population of subregion i to the expected deaths in subregion $i,\ i = 1, \ldots, p$. Usually the expected count is modelled as

$$E(n_i) = \lambda_i = e_i . m(f_i^{\prime}\alpha)),\ i = 1, \ldots, p.$$

Here, the $e_i,\ i = 1, \ldots, p,$ act as a background rate for the i-th subregion. The function $m(\bullet)$

represents a link to spatial and other covariates in the $p \times q$ design matrix F, whose rows are $f_1', \ldots, f_p'$. The parameter vector α has dimension $q \times 1$. Define the polar coordinates of the subregion centre as (r_i, θ_i), relative to the pollution source. Often, the only variable to be included in F is r, the radial distance from the source. When this is used alone, an additive link such as $m(\bullet) \quad = 1 + \exp(.)$, is appropriate since (for radial distance decline) the background rate (e_i) is unaltered at great distances. However, directional variables (e.g., $\cos\theta, \sin\theta, r\cos\theta, \log(r)\cos\theta$, etc.) representing preferred direction and angular-linear correlation can also be useful in detecting directional preference resulting from preferred directions of pollution outfall.

This model may be extended to include unobserved heterogeneity between regions by introducing a prior distribution for the log relative risks ($\log\lambda_i,\ i = 1, \ldots,\ p$). This could be defined to include spatially uncorrelated or correlated heterogeneity. The empirical and full Bayes methods described above often take this approach.

Estimation

One may estimate the parameters of the log-linear model above, via maximum likelihood through standard GLM packages, such as GLIM or S-PLUS. Using GLIM, the known log of the background hazard for each subregions, {log(e_i), $i = 1, \ldots, p\}$, is treated as an ‘offset’. A multiplicative (log) link can be directly modelled in this way, while an additive link can be programmed via user-defined macros.

Log-linear models are appropriate if due care is taken to examine whether model assumptions are met. For example, Lawson(1993) suggests the use of Monte Carlo tests for goodness of fit. If a model fits well, then the standardised model residuals should be approximately independently and identically distributed (i.i.d.). One may use autocorrelation tests, again via Monte Carlo, and make any required model adjustments. If such residuals are not available directly, then it is always possible to compare crude model residuals to a simulation envelope of m sets of residuals generated from the fitted model (parametric bootstrap).

Bayesian models for count data can be posterior-sampled via MCMC methods, and a variety of approximations are also available to provide empirical Bayes estimates.

Hypothesis Tests

Most of the existing literature on regional counts of health effects of pollution sources is based on hypothesis testing. Stone (1988) first outlined tests specifically designed for count data of events around a pollution source. These tests are based on the ratio of observed to expected counts cumulated over distance from a pollution source. The tests are based on the assumption of independent Poisson counts with monotonic distance ordering. A number of case studies have been based on these tests.

- While Stone’s test is based on traditional epidemiologic estimates (i.e. SMRs), the test is not uniformly most powerful (UMP) for a monotonic trend. If a UMP test exists, it is a score test for particular clustering alternative hypotheses. Unfortunately, these forms of alternative commonly arise in small-area epidemiological studies.

- Lawson(1993) developed a distance-effect score test versus a non-monotone, peaked alternative, and also suggested tests for directional and directional-distance effects within a log-linear model framework.

A cautionary note should be sounded in relation to the use of tests for putative source locations. The results of recent power studies carried out on a range of distance-decline tests have shown that:

'….many current tests of focussed clustering

often have poor power for detecting

the small increases in risk

often associated with environmental exposures'

This supports the fundamental need to examine a range of approaches to putative sources analysis within one study as well as a range of association variables.

Modelling versus Hypothesis Testing

Conclusions

- The analysis of small area health data around putative hazard sources has developed now to a stage where some basic issues are resolved and basic methods are in place.
- There is still considerable lack of agreement on a number of key issues relating to basic methods and also a number of underdeveloped areas worthy of further consideration.
- Perhaps the most contentious area of basic methodology is that of exposure modelling and how this should be carried out in the small area context. I believe that some degree of sophistication in exposure modelling should be attempted, as naive use of simple exposure models (e.g. distance-only models) can lead to erroneous conclusions.
- Both directional and distance-related effects should be included in any analysis, unless there are good reasons *not* to do so.

Armadale example

- 49 cases of respiratory cancer in Armadale, central Scotland in a six year period

- CHD control disease available as well as 18 expected rates
- Study region defined by the town limits
- Case event analysis carried out using point process models and extraction mapping

Results

parameter	CHD control
grand mean	2.78 (0.207)
r	-
$\cos\theta$	-0.935 (0.275)
$\sin\theta$	-0.331 (0.217)
r $\cos\theta$	-
r $\sin\theta$	-
null Deviance	92 (78)
Deviance (df)	73 (76)
AIC	583.6
Armadale: results	

The results suggest that there is no distance decline effect but a directional effect. This may be due to the large extraction peak extending into the south-west of the area.

Session: The

Armadale Example:a Case Study in Environmental Epidemiology

Background

- Armadale is a small industrial town in central Scotland

- During 1968-1974 a large increase in respiratory cancer (ICD code 162) mortality was recorded
- It was dubbed the Armadale Epidemic
- Early-mid 1960s : a local foundry in central Armadale had operated new industrial processes, including ore boiling
- Increase in cancer risk hypothesised to be

linked to emissions from the central foundry

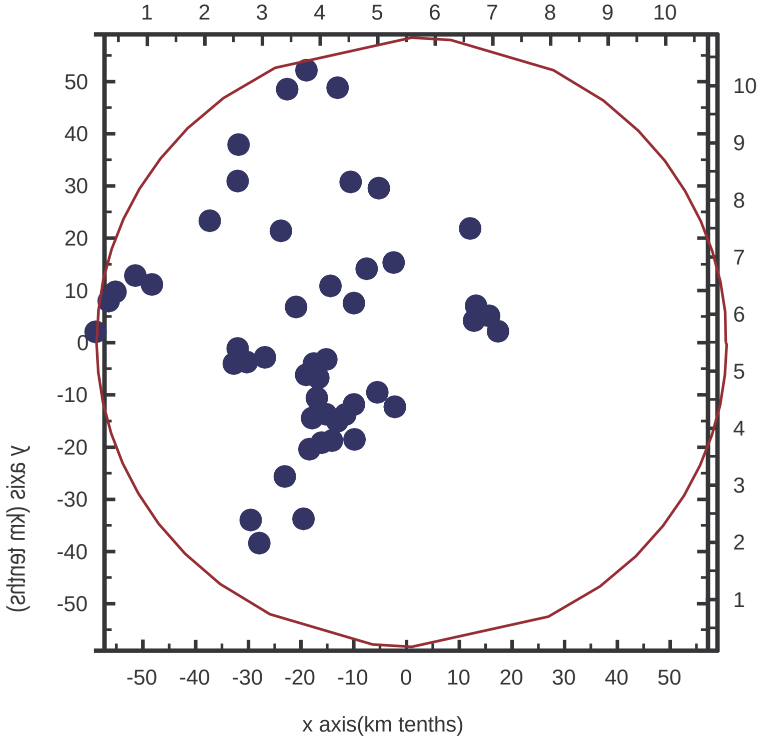
- Unusually short latency period for this cancer was hypothesised to be related to tumour promoters being emitted by the foundry
- Respiratory cancer reduced considerably after 1974......this appears to be related to changes in air pollution controls at the foundry.
- Spatial distribution of the disease may help to assess the link to a source.

The Data Example

- In the period of study (1968-1974) 49 deaths

  occurred. This is a large number for such a small town. A retrospective questionnaire survey established that these deaths were not lifestyle-related or occupational in nature.

Armadale: case event locations


1
2
3
4
5
6
7
8
9
10
50
40
30
20
10
0
-10
-20
-30
-40
-50
10
9
8
7
6
5
4
3
2
1
-50
-40
-30
-20
-10
0
10
20
30
40
50
x axis(km tenths)
y axis (km tenths)

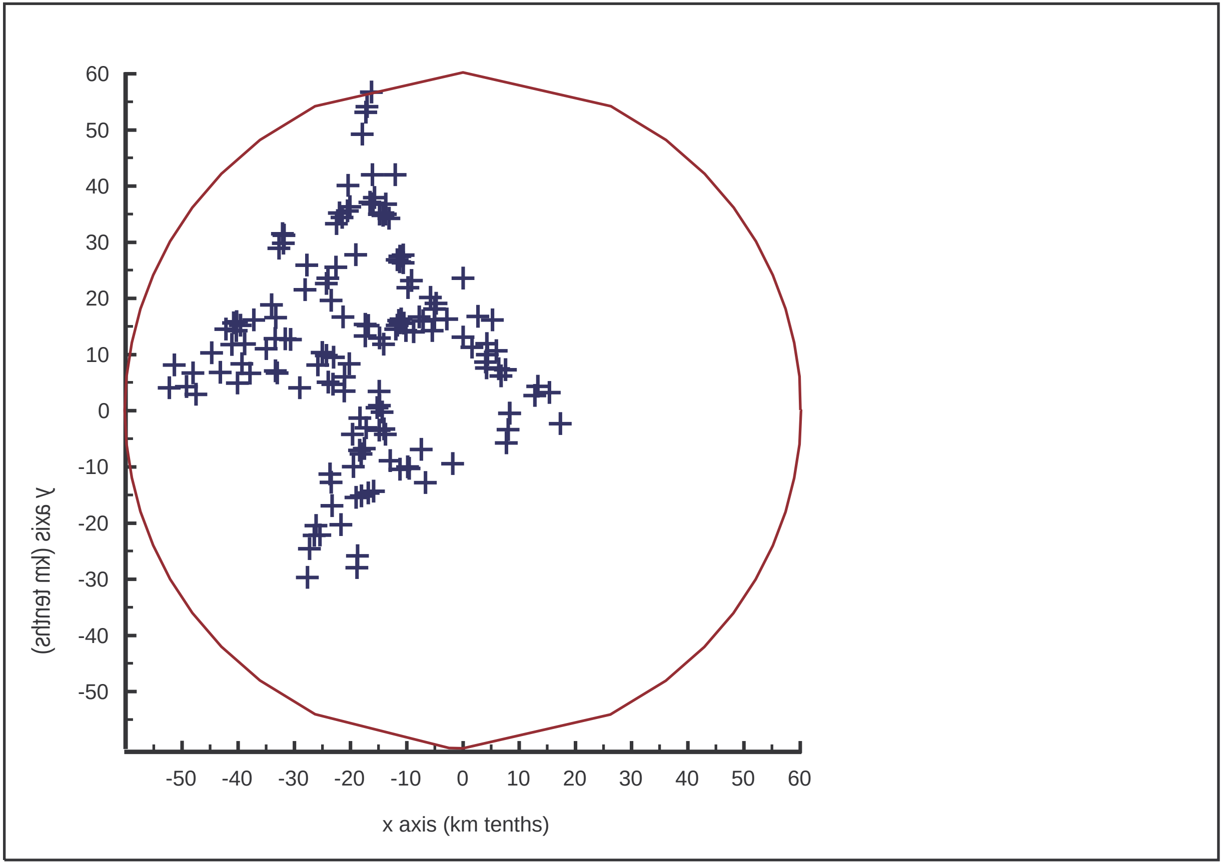
Armadale: control event locations

- A control disease (coronary heart disease:

  CHD) was used as a control for respiratory cancer
- Question of matching?

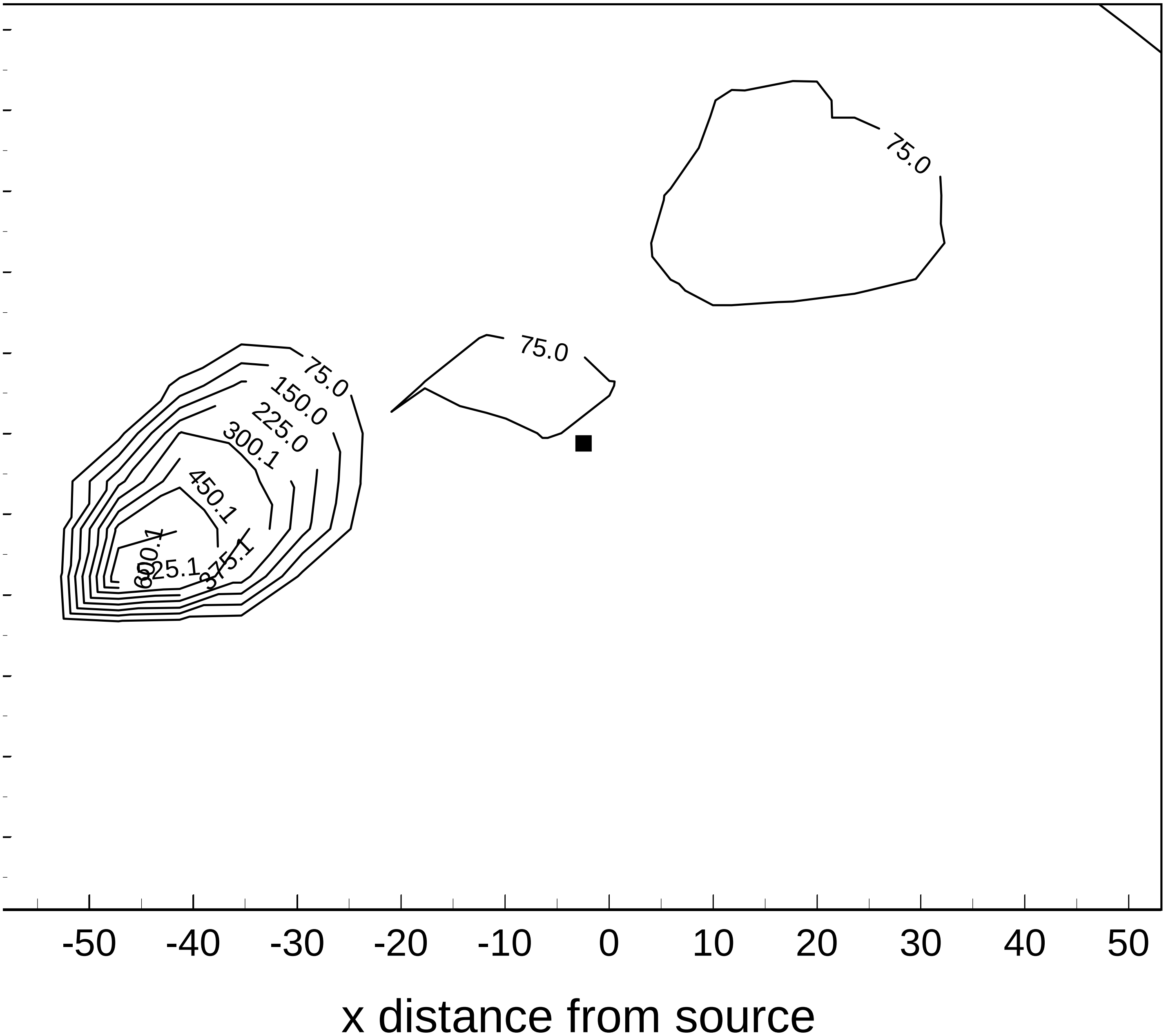
Exploratory Methods

- extraction mapping: relative risk assessment


60
50
40
30
20
10
0
-10
-20
-30
-40
-50
y axis (km tenths)
-50
-40
-30
-20
-10
0
10
20
30
40
50
60
x axis (km tenths)

Extraction map of Armadale

Armadale extraction surface (CHD)


75.0
75.0
75.0
150.0
225.0
300.1
450.1
600.1
525.1
375.1
-50
-40
-30
-20
-10
0
10
20
30
40
50
x distance from source

Modelling and Hypothesis Testing

- models can have distance and directional components
- A general approach to modelling exposure risk is to include an appropriate selection of measures in the specification of $f(x;.)$. First it is appropriate to consider how exposure variables can be linked to the background intensity $g(\mathbf{x})$. We define $f(\mathbf{x};\theta) = m\{f^*(\mathbf{x})'\alpha\}$, where $m\{.\}$ is an appropriate link function, and $f^*(\mathbf{x})$ represents the design matrix of exposure variables which is evaluated at $\mathbf{x}$. The link function is usually defined as $m\{.\} = 1 + \exp\{.\}$, although a direct multiplicative link can also be used. Usually each row of $f^*(\mathbf{x})$ will consist of a selection of the variables

$$\{r, \log(r), \cos(\phi), \sin(\phi), r\cos(\phi), r\sin(\phi)\}.$$

The first four variables represent distance-decline, peakedness, and directional effects, while the latter variables are directional-distance correlation effects

Results

parameter	CHD control
grand mean	2.78 (0.207)
r	-
$\cos\theta$	-0.935 (0.275)
$\sin\theta$	-0.331 (0.217)
r $\cos\theta$	-
r $\sin\theta$	-
null Deviance	92 (78)
Deviance (df)	73 (76)
AIC	583.6
Armadale: results	

- The above results are for CHD control and the best subset model for any of the distance and angular variables specified.
- Only the directional terms are significant, and the grand mean.
- This suggests that distance is not a significant contributor

parameter	expected deaths
grand mean	**3.064 (0.405)**
r	**0.034 (0.018)**
$\cos\theta$	**-**
$\sin\theta$	**-**
$r\cos\theta$	**-0.001 (0.014)**
$r\sin\theta$	**-0.02 (0.008)**
null Deviance	**77 (78)**
Deviance (df)	**73 (76)**
AIC	**583.6**
Armadale: expected death	

- If expected deaths in 18 census districts are used instead of CHD then a different picture emerges. There is still no distance decline but a different directional effect appears.
- What if we include heterogeneity (uncorrelated and correlated)...does this make a difference

- The results of such an analysis (reported elsewhere) suggest that a model including a distance effect is relevant and some directional components are also present, once the unobserved heterogeneity is accounted for. There is a strong indication that uncorrelated heterogeneity is present although there appears to be little evidence of autocorrelation here

Issues

1) choice of control

2) sensitivity to smoothing

3) evidence of a link

4) data quality

5) disease to study

Open Questions

- edge effects

- what happens at edges?

- smoothing

-is it necessary?

- random effects

-related to smoothing: are they necessary?

- public health implications of mapping:

-maps, clusters, ecological analysis

- the need for space-time methods and surveillance

-what about the dynamic picture?

- epidemic modelling?