Spatial Statistics for Environmental Epidemiology

Peter Diggle

(Medical Statistics Unit, Lancaster University)

p.diggle@lancaster.ac.uk

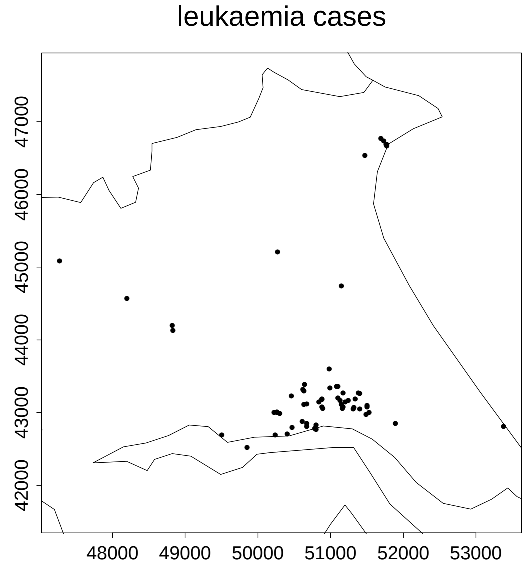
Last updated on: March 2, 2000

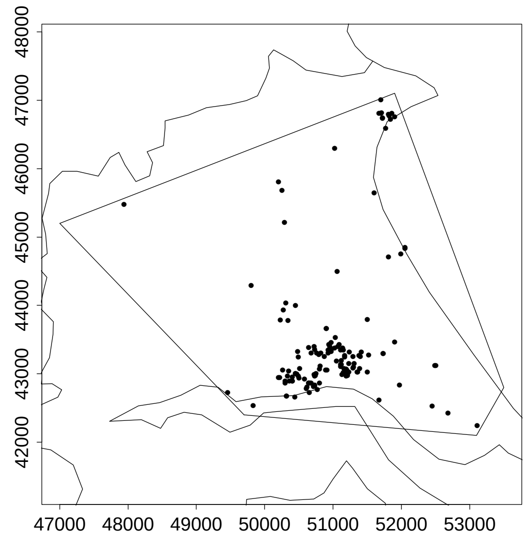
Summary

This short course covers a range of methods and techniques of spatial data analysis used in environmental epidemiology. Specific topics include: testing for spatial clustering, investigating variation in risk around a point source, constructing smooth maps of spatial variation in risk. Both point data (individual level) and spatially aggregated data (area level) are discussed.

1 Motivating examples

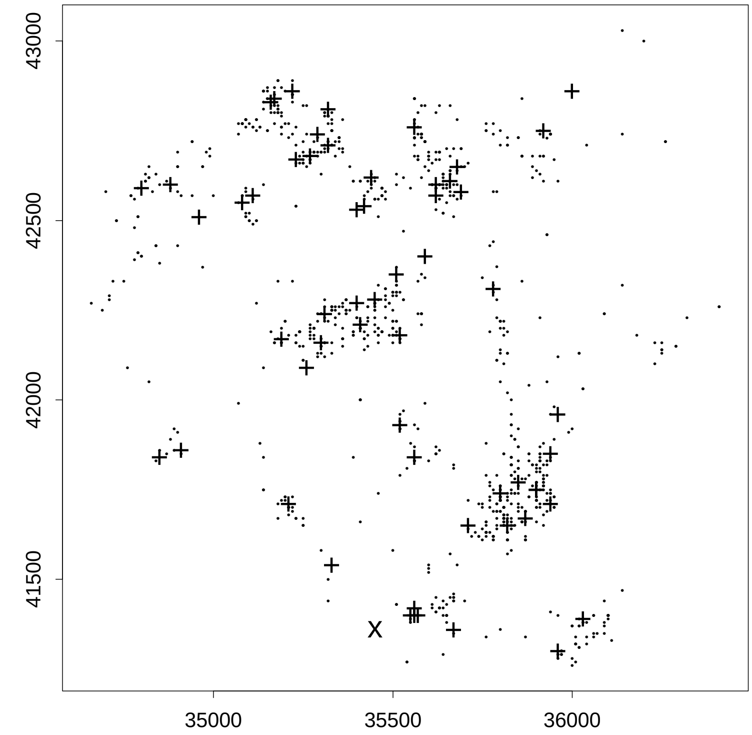
Example 1. Childhood leukaemia in Humberside (from Cuzick and Edwards, 1990).





- residential locations of all known cases of childhood leukaemia in Humberside, England, over the period 1974-82;
- residential locations of a random sample from the birth register over the same area and time-period.

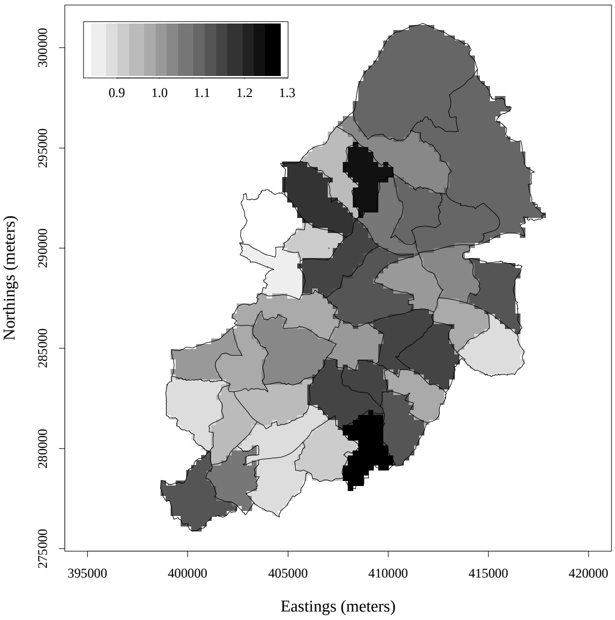
Example 2. Lung and larynx cancers in Chorley-South Ribble (from Diggle, Gatrell and Lovett, 1990).



The map shows

- all known cases of lung cancer in Chorley-Ribble, England (dots);
- all known cases of larynx cancer in the same area (small crosses);
- the location of a now-disused industrial incinerator (large cross)

Example 3. Colorectal cancer in Birmingham (from Kelsall and Wakefield, 2000).



- the raw data are counts of the numbers of cases of colorectal cancer in regions A_i corresponding to 36 electoral wards in Birmingham, England.
- the map shows smoothed estimates of relative risk in each ward, adjusted for the age-sex mix of the population in each ward

Substantive questions

- Do cases show a *surprising* tendency to cluster together?
- Does the risk of disease vary spatially?
- Is disease risk elevated near a particular location?

Testing or estimation?

- All of the above substantive questions can be expressed as hypotheses to be tested.
- But rejection of the null is only the first stage.
- We will usually want to *estimate* spatial effects.
- And we should ideally do so after taking account of non-spatial risk factors.

2. Point processes

A **spatial point process** is a stochastic process, a realisation of which consists of a countable set of points $\boldsymbol{x}_i$ in the plane.

We often call these points **events** to distinguish them from arbitrary points $\boldsymbol{x}$ in the plane.

We write $N(A)$ for the random variable which represents the number of events in a planar region A,

$$N(A) = \#(\boldsymbol{x}_i \in A).$$

- The process is *stationary* if, for any integer k and regions $A_i : i = 1, \ldots, k$ the joint distribution of $N(A_1), \ldots, N(A_k)$ is invariant to *translation* by an arbitrary amount $\boldsymbol{x}$.
- The process is *isotropic* if, for any integer k and regions $A_i : i = 1, \ldots, k$ the joint distribution of $N(A_1), \ldots, N(A_k)$ is invariant to *rotation* through an arbitrary angle θ, i.e. no directional effects.

2.1 Point process intensities

How should we define point process analogues of the mean and covariance structure for real-valued processes?

Let $d\boldsymbol{x}$ denote a small region containing the point $\boldsymbol{x}$.

Def 1. The *(first-order) intensity function* of a spatial point process is

$$\lambda(\boldsymbol{x}) = \lim_{|d\boldsymbol{x}|\to 0} \left\{ \frac{E[N(d\boldsymbol{x})]}{|d\boldsymbol{x}|} \right\}$$

Def 2. The *second-order intensity function* of a spatial point process is

$$\lambda_2(\boldsymbol{x}, \boldsymbol{y}) = \lim_{\substack{|d\boldsymbol{x}|\to 0 \\ |d\boldsymbol{y}|\to 0}} \left\{ \frac{E[N(d\boldsymbol{x})N(d\boldsymbol{y})]}{|d\boldsymbol{x}||d\boldsymbol{y}|} \right\}$$

Def 3. The *covariance density* of a spatial point process is

$$\gamma(\boldsymbol{x}, \boldsymbol{y}) = \lambda_2(\boldsymbol{x}, \boldsymbol{y}) - \lambda(\boldsymbol{x})\lambda(\boldsymbol{y}).$$

If we assume stationarity and isotropy, it follows that:

(i) $\lambda(\boldsymbol{x}) \equiv \lambda = E[N(A)]/|A|$, (constant, for all A).

(ii) $\lambda_2(\boldsymbol{x}, \boldsymbol{y}) \equiv \lambda_2(\|\boldsymbol{x} - \boldsymbol{y}\|)$ (depends only on distance)

(iii) $\gamma(u) = \lambda_2(u) - \lambda^2$.

Physical interpretation:

- $\lambda =$ expected number of events per unit area.
- $\lambda_2(u) =$?

To get a more easily interpretable quantity than $\lambda_2(u)$, proceed as follows:

Def 4 The *reduced second moment function* of a stationary, isotropic spatial point process is

$$K(s) = 2\pi\lambda^{-2} \int_0^s \lambda_2(r) r dr.$$

Theorem 1. For a stationary, isotropic, orderly process, $K(s) = \lambda^{-1}$E[number of further events within distance s of an arbitrary event]

- gives a tangible interpretation of $K(s)$,
- suggests a method of estimating $K(s)$ from data,
- hints at why an estimate of $K(s)$ would be a useful descriptor of an observed spatial pattern:
  - for clustered patterns, each event is likely to be surrounded by further members of the same cluster and, for small values of s, $K(s)$ will be relatively *large*.
  - conversely, if events are regularly spaced, each one is likely to be surrounded by empty space and, for small values of s, $K(s)$ will be relatively *small*.

A benchmark to determine what we mean by *relatively* large or small is:

Theorem 2. For a homogeneous, planar Poisson process,

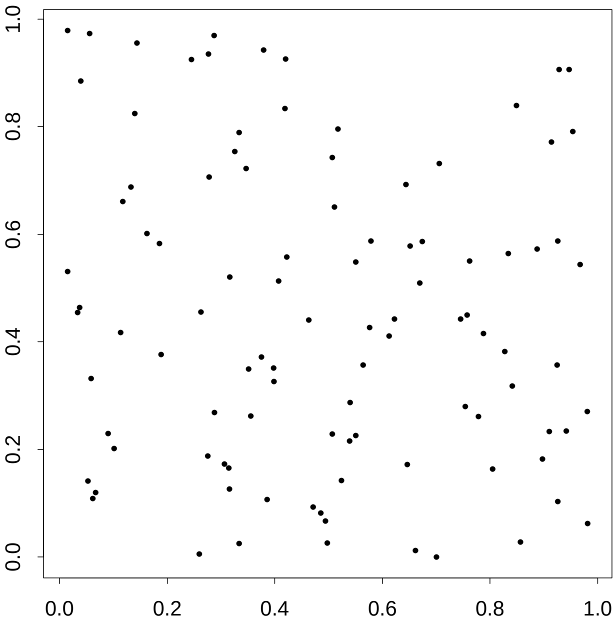
$$K(s) = \pi s^2$$

Proof

The events in a Poisson process are located independently of one another. Thus, the existence of an event at $\boldsymbol{x}$ has no bearing on the number of further events within distance s of $\boldsymbol{x}$. Since λ is the expected number of events per unit area, the expected number of further events within distance s of $\boldsymbol{x}$ is $\lambda \pi s^2$. Divide by λ to give the result.

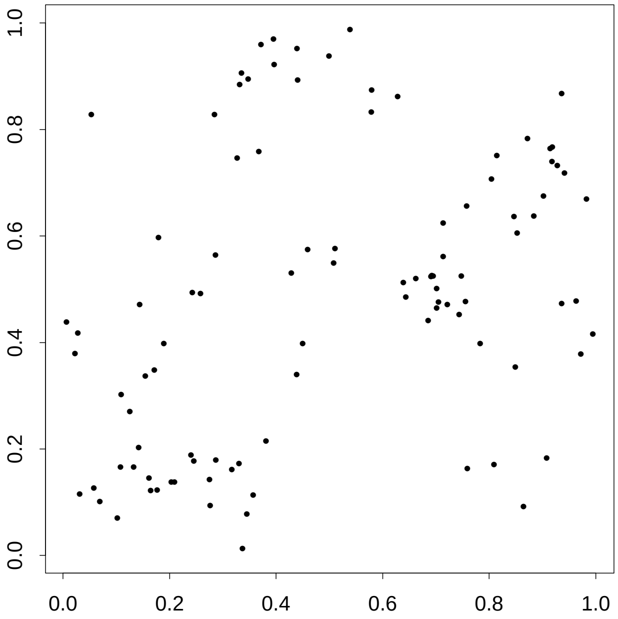
#

A homogeneous Poisson process:

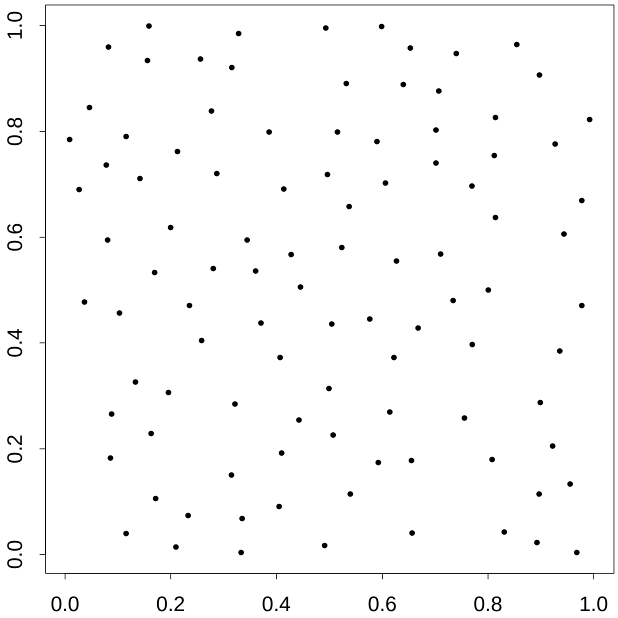


Two point processes

A clustered point process:



A regular point process:



In epidemiological studies, we sometimes can't achieve a complete ascertainment of all cases.

Def 5. A *random thinning*, P', of a point process P, is a point process whose events are a sub-set of the events of P generated by retaining or deleting the events of P in a series of mutually independent Bernoulli trials.

We can now establish the following result:

Theorem 3. $K(s)$ is invariant to random thinning.

Proof

Random thinning reduces both the intensity λ, and the expected number of further events within distance s of an arbitrary event, by the *same* multiplicative factor. Hence, their ratio $K(s)$ is unchanged.

#

Conclusion: the interpretation of an estimated K-function is robust to incomplete ascertainment of cases, provided the incompleteness is spatially neutral.

Estimating the K-function

Our objective is to estimate λ and $K(s)$ from a set of data of the form $x_i \in A : i = 1, \ldots, n\}$, for some planar region A.

Estimation of λ

Because λ is the expected number of events per unit area, we define

$$\hat{\lambda} = n/|A|$$

Estimation of $K(s)$

Similarly, because

$\lambda K(s)$ =E[number of further events within distance s of an arbitrary event]

we can construct an estimator of $K(s)$ as follows.

1. Define $E(s) = \lambda K(s)$. Let d_{ij} be the distance between the events x_i and x_j. Define

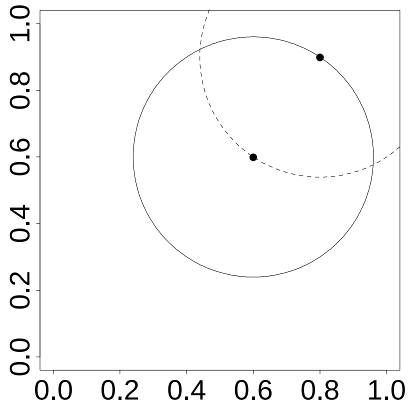
   $$\tilde{E}(s) = n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} I(d_{ij} \leq s), \qquad (1)$$

   where $I(\cdot)$ denotes the indicator function.

2. The estimator $\tilde{E}(s)$ is negatively biased because we do not observe events outside A, so the observed counts from events x_i close to the boundary of A will be artificially low.

3. Introduce weights,

   w_{ij} = reciprocal of proportion of circumference of circle, centre $\boldsymbol{x}_i$ and radius d_{ij}, which is contained in A.



4. An edge-corrected estimator for $E(s)$ is

$$\hat{E}(s) = n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq s).$$

5. Since $K(s) = E(s)/\lambda$, define

$$\begin{aligned} \hat{K}(s) &= \hat{E}(s)/\hat{\lambda} && (2) \\ &= n^{-2}|A| \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} I(d_{ij} \leq s) && (3) \end{aligned}$$

   Explicit formulae for the w_{ij} are given in Diggle (1983, p72) for A a rectangle or circle. An algorithm for an arbitrary polygon A is used in Rowlingson and Diggle (1993).

Notes

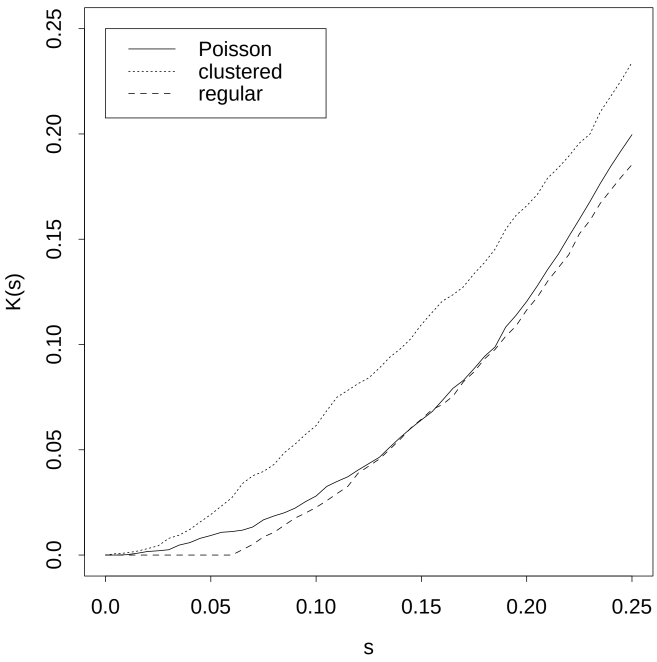
1. Typically, $\mathrm{Var}\{\hat{K}(s)\}$ tends to increase with s.

2. The dimensions of A limit the range of values of s which can be considered. In practice, the increasing variance of $\hat{K}(s)$ is a more serious limitation. As a rough guide, for data on a rectangle A, it is usually not worth trying to estimate $K(s)$ at values of s bigger than one-half the length of the shorter side of A.

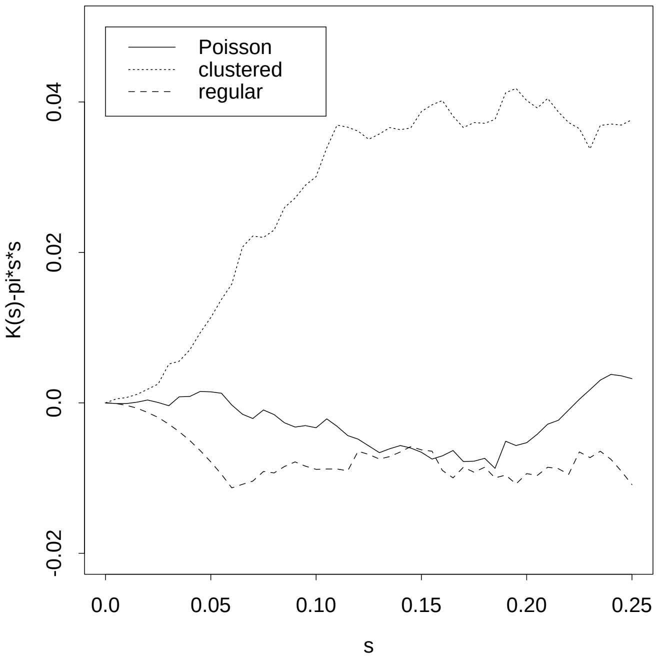
3. The sampling distribution of $\hat{K}(s)$ is largely intractable. Chetwynd and Diggle (1998) give an estimator for the variance of $\hat{K}(s)$ when the underlying process is a Poisson process. More generally, if n is large we can get a rough idea of the variability in $\hat{K}(s)$ by dividing A into two or more sub-regions, calculating the empirical variance of the estimates $\hat{K}(s)$ from the different sub-regions, and using the fact that the dominant term in the variance of $\hat{K}(s)$ is of order n^{-1}.

4. There is some technical advantage in using $n(n-1)$ rather than n^2 as the divisor in the expression (3) for $\hat{K}(s)$.

Estimates $\hat{K}(s)$ for three simulated patterns:



Estimates $\hat{K}(s) - \pi s^2$ for three simulated patterns:



3 Case-control methods

The data for a spatial case-control study consist of two point patterns:

- the locations of all known *cases* of particular disease in a geographical region A, over a defined period of time
- the locations of a sample of *controls*, selected from the population at risk:
  - completely at random
  - group-matched (eg to preserve sex-ratio)
  - individually matched

For each of the three substantive problems identifed in Section 1, we first consider the analysis of a completely random case-control study, then note the necessary modifications to analyse an individually matched study. Group-matched studies can usually be analysed by pooling results from separate analyses within each group.

3.1 Spatial clustering

- most epidemiological case-maps show apparent clustering because cases occur most often in area of high population density
- if this is the only source of spatial clustering, the same should be true of the control-map
- more formally:
  - under the null hypothesis of *no spatial clustering*, the *cases* and the *controls* are independent random samples from the *same underlying population at risk*
  - under this hypothesis, $K_1(s) = K_0(s)$
- Hence, consider $D(s) = K_1(s) - K_0(s)$

Chetwynd and Diggle (1991) propose a test of spatial clustering using the statistic

$$D = \int_0^{s_0} \{v(s)\}^{-0.5} \hat{D}(s) ds$$

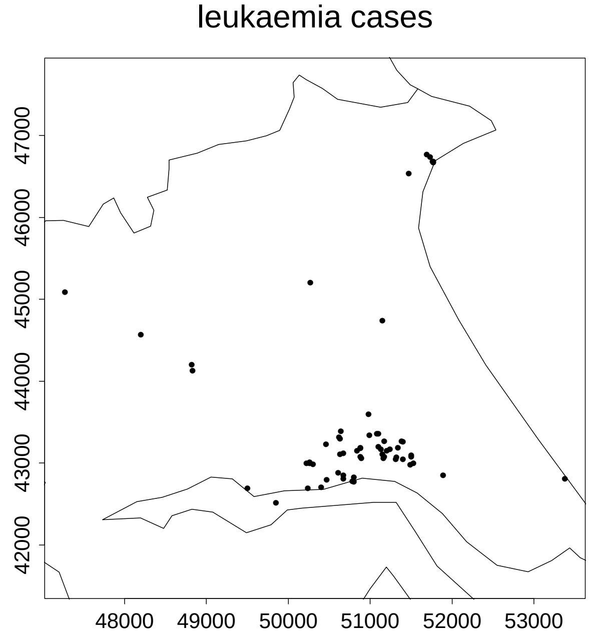
where $v(s)$ is the variance of $\hat{D}(s) = \hat{K}_1(s) - \hat{K}_0(s)$ under random permutation of the case-control labels.

Significance is assessed either by a Normal approximation or, for an exact Monte Carlo test, by simulation from the randomisation distribution under the null.

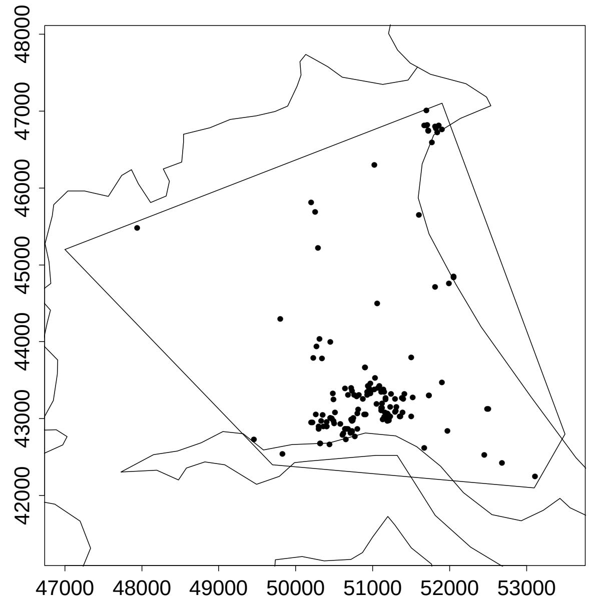
Thus, whilst the statistic is motivated by the theory of stationary point processes, the inference is design-based.

Example 1. Childhood leukaemia in North Humberside (from Cuzick and Edwards, 1990).

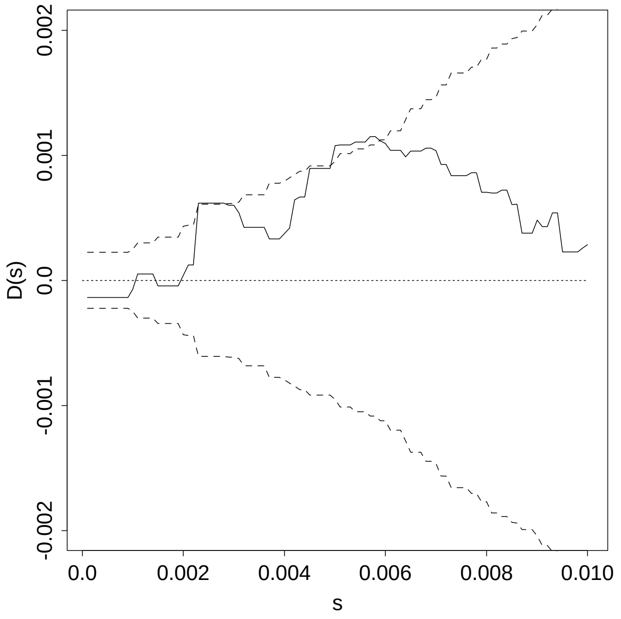
Leukaemia cases:



Leukaemia controls:



The diagram below shows the estimate $\hat{D}(s)$ for the leukaemia data, with plus and minus two standard errors under random labelling



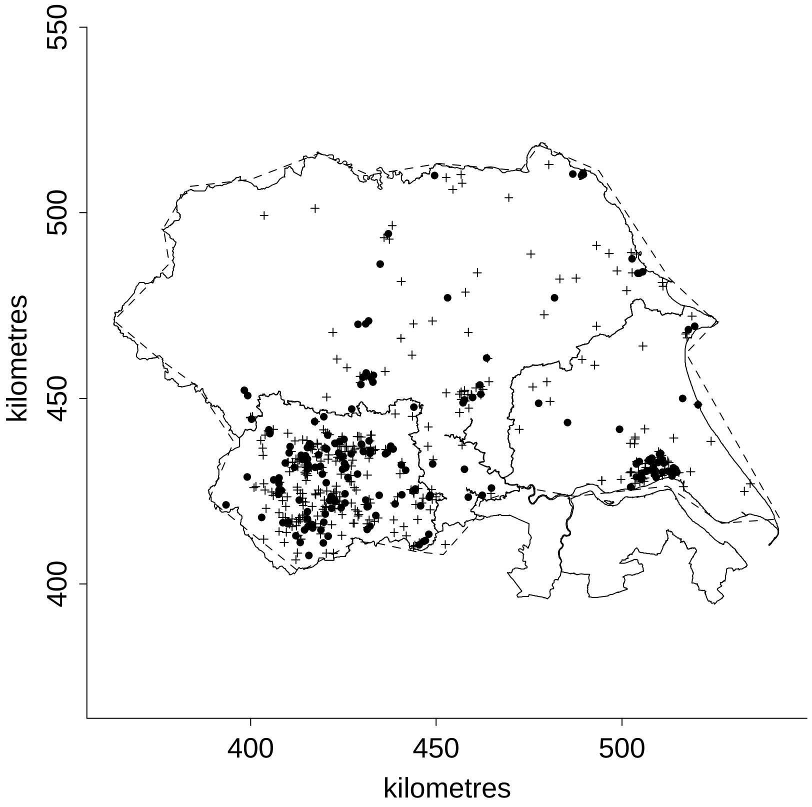
- A Monte Carlo test using the test statistic D with 99 simulated random labellings gave a p-value of 0.14.
- The Normal approximation gave a standard Normal deviate of $Z = 1.21$, corresponding to a one-sided p-value of 0.11.

Adaptation to matched case-control data

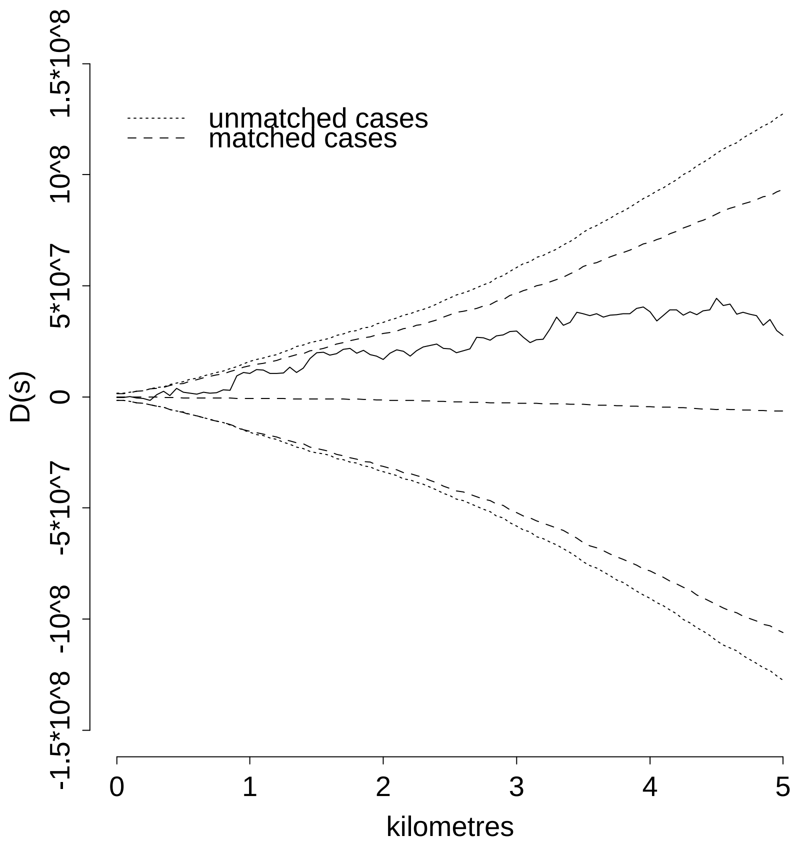
Chetwynd, Diggle, Marshall and Parslow (2000) consider the adaptation of the above method to individually matched case-control data.

- for a test of clustering, a Monte Carlo test based on D is still available, comparing the observed value of D with simulate values under random re-labellings within matched case-control sets
- for estimation, modifications are necessary because:
  - the randomisation variance of $\hat{D}(s)$ changes
  - more fundamentally, in a k-to-1 matched case-control study, $\mathrm{E}[D] \neq 0$ under the null hypothesis of no spatial clustering.

Example 4. Childhood diabetes in Yorkshire, England



- matched case-control study
- two controls per case, matched by age, sex and FHSA



- plot shows $\hat{D}(s) = \hat{K}_1(s) - \hat{K}_0(s)$, with null expectation and ±2 standard errors
- matching introduces almost no bias in this particular study ($\mathrm{E}[\hat{D}(s)] \approx 0$)
- but materially affects standard error of $\hat{D}(s)$
- test for clustering is non-significant

3.2 Spatial variation in risk

Our working model for spatial variation in risk is that:

- *cases* form a Poisson process with intensity $\lambda(x)$
- *controls* form a second, independent Poisson process with intensity $\lambda_0(x)$
- $\lambda(x) = \alpha\lambda_0(x)\rho(x)$ where
  - α is determined by the number of controls in the design
  - $\rho(x)$ represents spatial variation in risk

It follows that, conditional on both case and control locations:

- case/control *labels* are determined by a series of independent Bernoulli trials with success probabilities

  $$p(x) = \lambda(x)/\{\lambda(x) + \lambda_0(x)\} = \alpha\rho(x)/\{1 + \alpha\rho(x)\}$$

- spatial variation in risk is estimable up to a constant of proportionality

Kelsall and Diggle (1998) consider three approaches to nonparametric estimation:

1. **Density ratio**
   - Use **kernel smoothing** for separate estimation of $\lambda(x)$ and $\lambda_0(x)$
   - For cases-locations $x_i : i = 1, ..., n$,

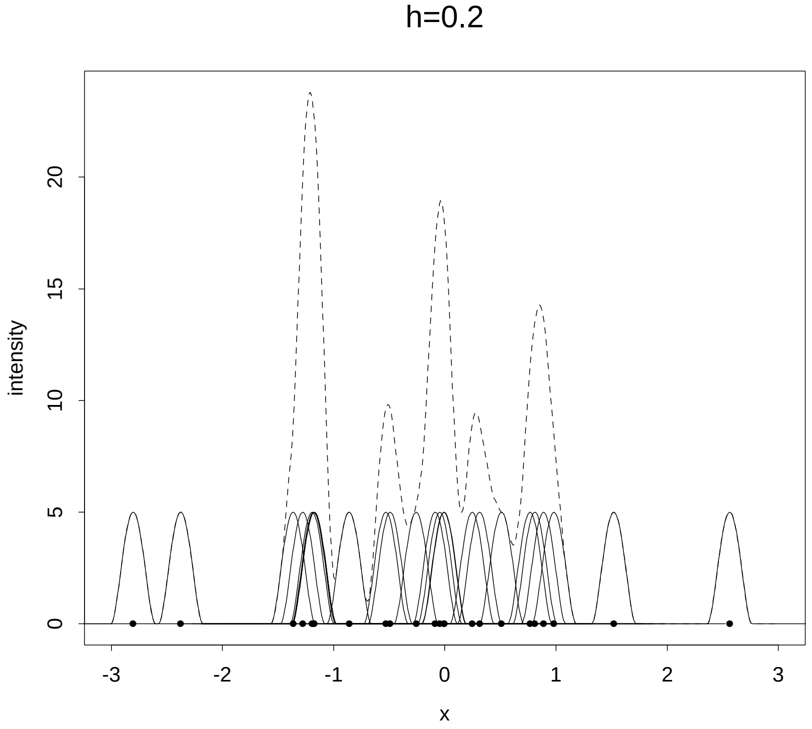
     $$\hat{\lambda}(x) = h^{-2} \sum_{i=1}^{n} W\{(x - x_i)/h\}$$

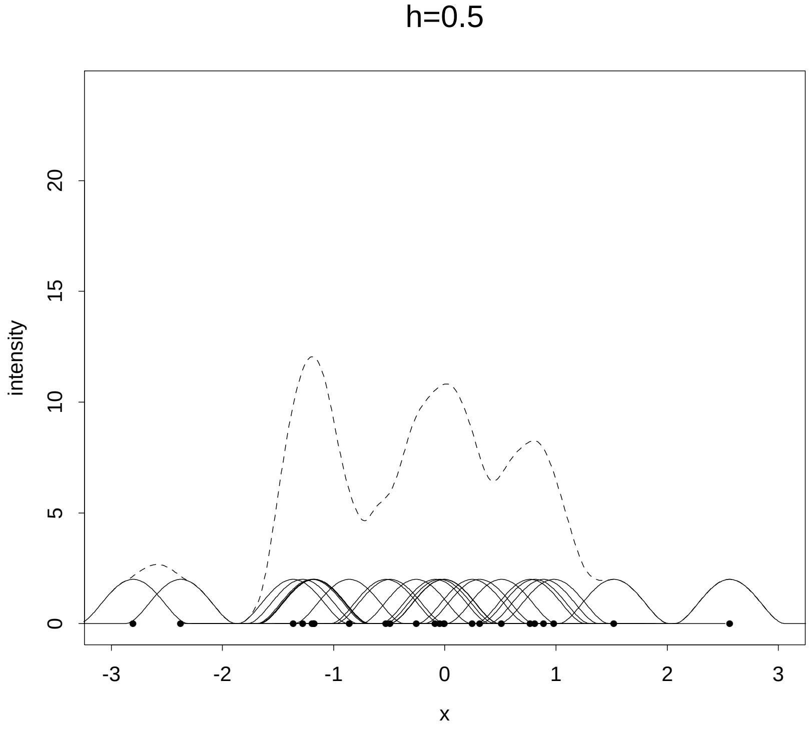
     where $W(\cdot)$ is a circularly symmetric bivariate pdf, and $h > 0$ a scalar which determines the amount of smoothing
   - $\hat{p}(x) = \hat{\lambda}(x)/\{\hat{\lambda}(x) + \hat{\lambda}_0(x)\}$

   The effect of changing h is illustrated in the following two pictures, for which the data are a random sample of size $n = 25$ from the standard Gaussian distribution.

   Each solid curve shows the kernel centered on an individual observation. The dashed curve is the kernel estimate.

Kernel smoothing:





2. **Binary regression**

- For n cases and $m = N - n$ controls, let $Y_i = 1/0$ for case/control respectively at location x_i.
- Define weights,

$$w_i(x) = W\{(x - x_i)/h\} / \sum_{j=1}^{N} W\{(x - x_j)/h\}$$

- Kernel estimator is

$$\hat{p}(x) = \sum_{i=1}^{N} w_i(x) Y_i$$

3. **Generalized additive model**

- Model assumption is

$$\log[p(x)/\{1 - p(x)\}] = u(x)'\beta + g(x)$$

  where $u(x)$ is vector of known risk factors and $g(x)$ models smooth residual spatial variation
- Log-likelihood is of the form

$$\sum_{i=1}^{n} \log p(x_i) + \sum_{i=n+1}^{N} \log\{1 - p(x_i)\}$$

- Fitting algorithm is adapted from Hastie and Tibshirani (1990), and includes kernel smoothing step for $\hat{g}(x)$ within iteratively weighted least squares

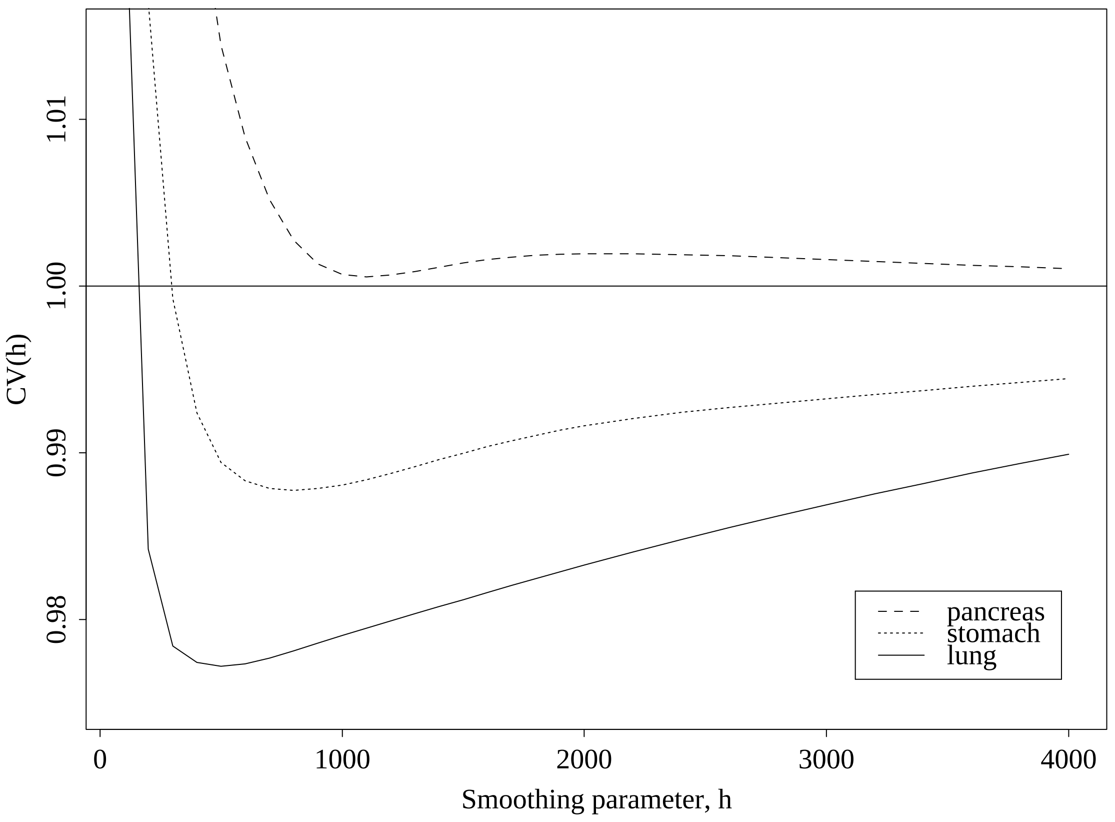
Choosing the amount of smoothing

- in kernel smoothing, the amount of smoothing is determined primarily by the band-width, h
- Kelsall and Diggle (1998) recommend a cross-validation method to choose h
- in the binary regression version, with no explanatory variables, this is defined as follows:
  - for each h, let $\hat{p}^i(x_i)$ be the estimate of $p(x_i)$ using all data *except* Y_i;
  - choose h to maximise

$$CV(h) = \sum_{i=1}^{n} \log \hat{p}^i(x_i) + \sum_{i=n+1}^{N} \log\{1 - \hat{p}^i(x_i)\}$$

- with explanatory variables included, use a similar idea within the GAM iterations for $\hat{g}(x)$

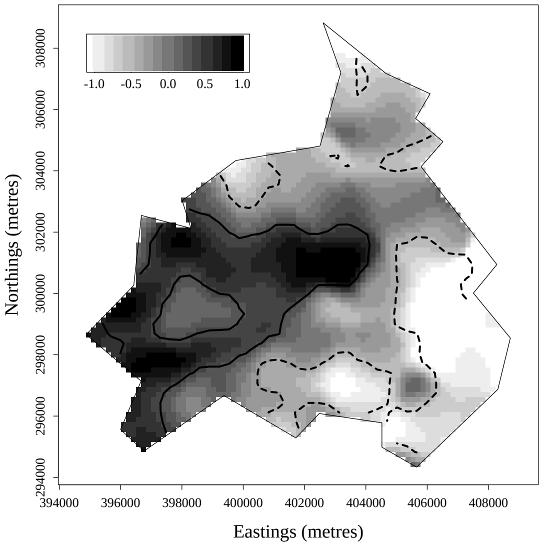
Example 5. Lung and stomach cancers in Walsall (from Kelsall and Diggle, 1998).



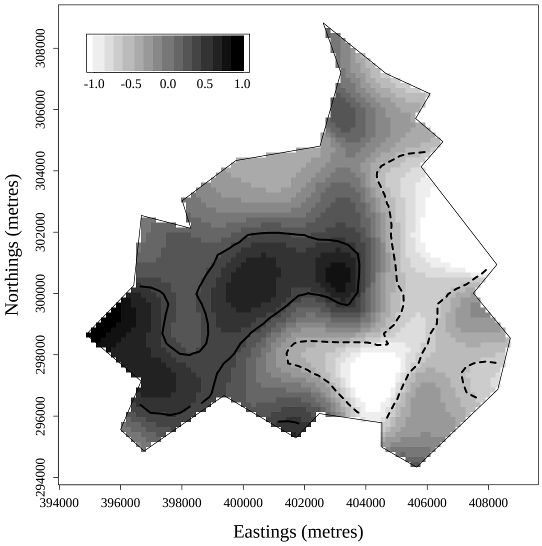
- well-defined minimum of $CV(h)$ for lung and stomach cancer data
- no well-defined minimum for (rare) pancreas cancer implies no strong evidence for spatial variation in risk for pancreas cancer

Estimated risk surfaces for:

- lung cancer



- stomach cancer



- similar pattern of variation for both diseases
- solid and dashed lines identify boundaries of regions where risk is significantly higher or lower, respectively, than average.

Adaptation to matched case-control data

Jarner, Diggle and Chetwynd (2000) consider the adaptation of Kelsall and Diggle's GAM methodology to individually matched data.

- Recall that in an unmatched case-control study, the control intensity is

  $$\lambda_0(x) = \alpha\lambda(x)\rho(x)$$

  In a k-to-1 matched design, the control intensity corresponding to the ith case is

  $$\lambda_0(x) = \alpha_i\lambda(x)\rho(x)$$

  and we need to eliminate the nuisance parameters α_i.

- To achieve this, we condition on the $(k+1)$ locations within each of n matched case-control sets, and let P_{ij} denote the probability that the jth member of the ith matched set is the case. Then

  $$P_{ij} = p(x_{ij}) / \sum_{r=1}^{k+1} p(x_{ir})$$

  where

  $$p(x_{ij}) = \alpha_i\rho(x_{ij}) / \{1 + \alpha_i\rho(x_{ij})\}$$

  and the P_{ij} depend only on $\rho(\cdot)$ as required.

- The log-likelihood is now of the form

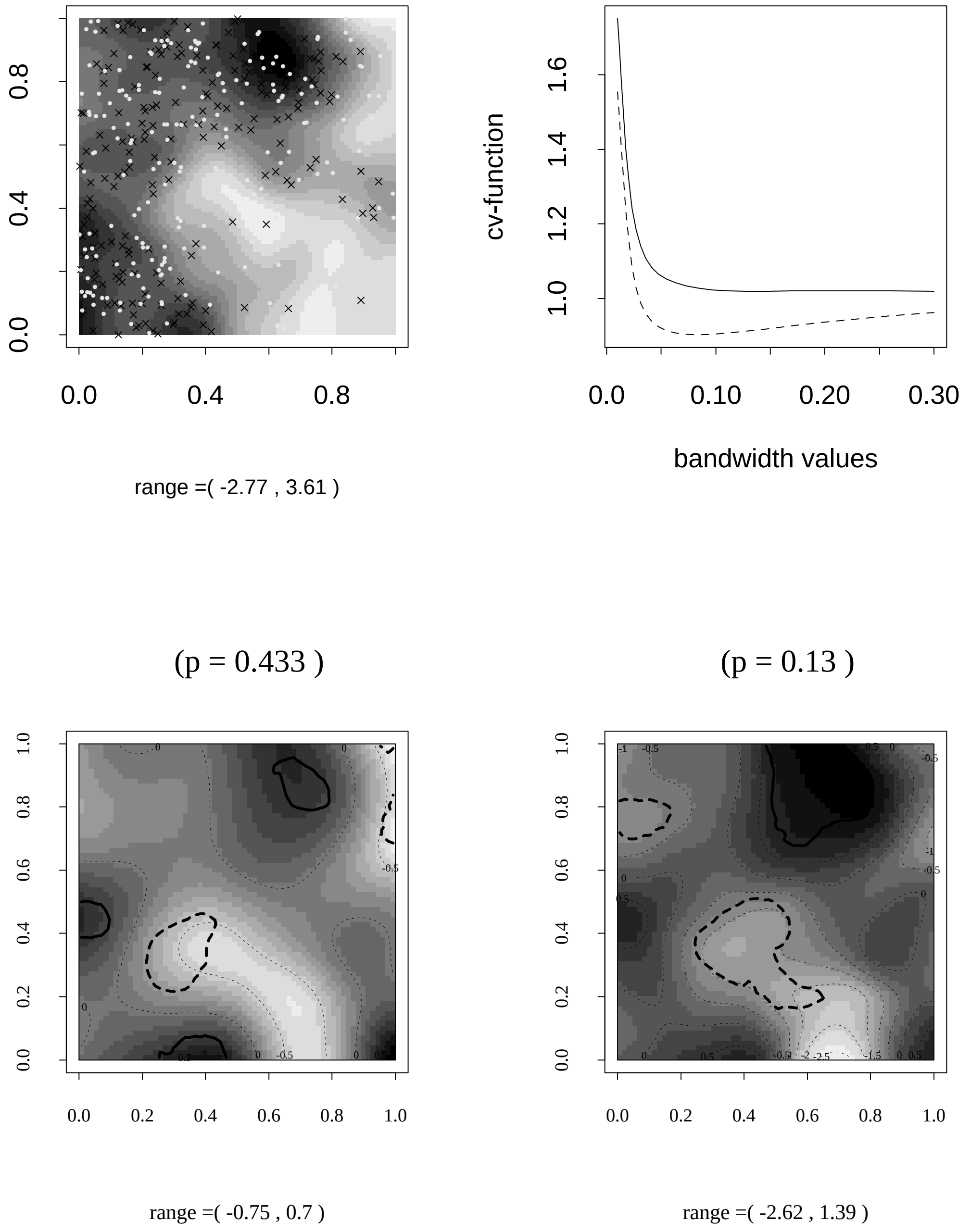
$$\sum_{i=1}^{n} \log P_{i1}$$

- The GAM fitting algorithm is easily adapted to this form of the likelihood.

A simulated example (see diagram on page 36))

- *top-left panel*
  - 200 cases (dots) generated from a Poisson process with spatially varying intensity $\lambda(x)$ (grey-scale image);
  - square study region divided into 3 by 3 grid of square sub-regions;
  - one individually matched control (crosses) per case, uniformly distributed over the sub-region containing its matched case;
- *top-right panel*
  - cross-validation plots for band-width selection, assuming matched (dashed line) or unmatched (solid line) controls, respectively;
  - incorrect assumption of unmatched conrols leads to over-smoothing, and consequent under-estimation of spatial variation in risk;

- *bottom-left panel*
  - estimated risk surface for (incorrect) unmatched analysis;
  - contour lines identify pointwise significant departures from average risk;
- *bottom-right panel*
  - estimated risk surface for (correct) matched analysis;
  - contour lines again identify pointwise significant departures from average risk;
  - note increased range of estimated risk by comparison with unmatched analysis

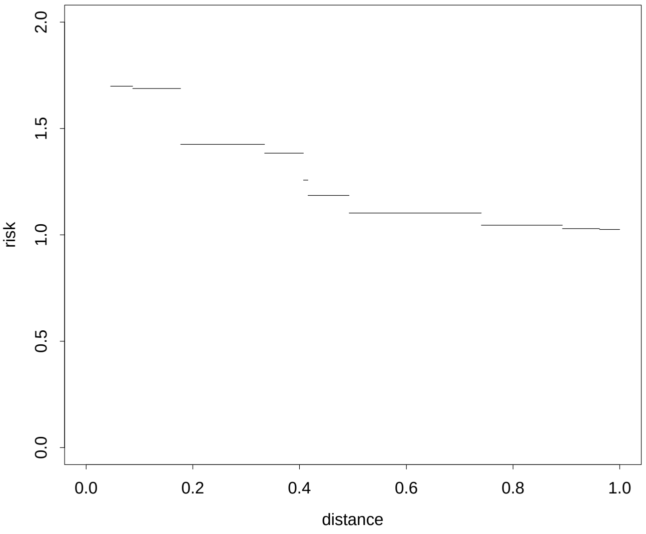

0.0
0.4
0.8
0.0
0.4
0.8
range =(-2.77 , 3.61)
cv-function
1.0
1.2
1.4
1.6
0.0
0.10
0.20
0.30
bandwidth values
(p = 0.433)
(p = 0.13)
0.0
0.2
0.4
0.6
0.8
1.0
range =(-0.75 , 0.7)
range =(-2.62 , 1.39)

3.3 Point source problems

- the methods of Section 3.2 are appropriate when there is no prior hypothesis about the source of any localised variation in risk
- when the concern is of possible elevation in risk around a particular point source, x_0 say, more tightly constrained modelling of $\rho(\cdot)$ may be justifiable

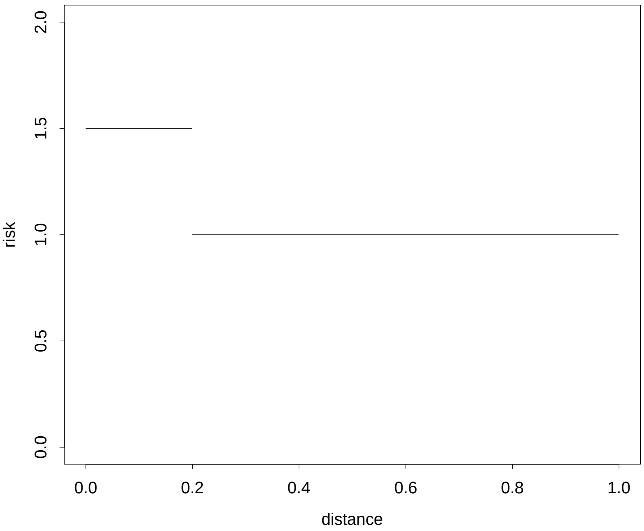
Some options:

- **Isotonic regression** (Stone, 1988)



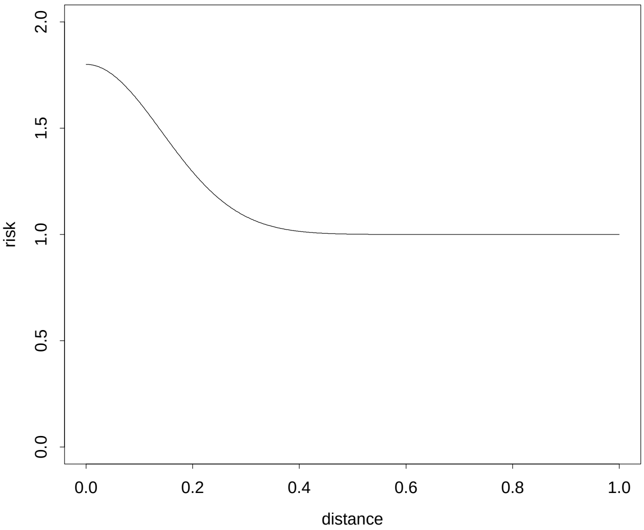
$\rho(x) = \rho(||x - x_0||$, monotone non-increasing

- **Near vs far** (Elliott et al, 1992)



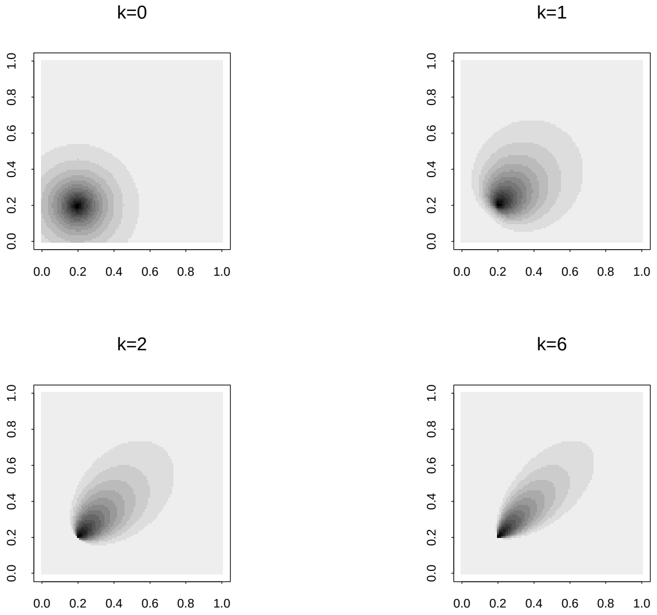
$$\rho(u) = \begin{cases} 1+\alpha & : \ u \leq \delta \\ 1 & : \ u > \delta \end{cases}$$

- **Isotropic Gaussian** (Diggle and Rowlingson, 1994)



$$\rho(u) = 1 + \alpha \exp\{-(u/\delta)^2\}$$

• A directional plume model



$$f(d,\theta) = 1 + \alpha \exp(-[d \exp\{\kappa \cos(\theta - \phi)\}/\beta]^2)$$

Interpretation of model parameters:

– α : elevation in risk at source

– β : rate of decay of risk with distance from source

– ϕ : direction of plume

– κ : degree of directional concentration of plume

Example 6. Asthma in north Derbyshire, England.

Diggle and Rowlingson (1994) fit the isotropic Gaussian model to data from the following case-control study of asthmatic symptoms in elementary schools in north Derbyshire.

- the study population consisted of all children attending one of 10 schools in the area;
- schools were stratified according to whether the headteacher had previously reported concern about the apparently high level of asthmatic symptoms in the school;
- four potential sources were considered – here, we look only at two:
  - a coking works (point source)
  - the main road network (line source)
- additional binary covariates included:
  - household includes at least one cigarette smoker?
  - child suffers from hay fever?
- overall risk was modelled multiplicatively, with separate terms for each of the two sources, and for log-linear covariate adjustment

Likelihood ratio comparisons:

Risk factors in model	2×log-likelihood	Parameters
None	-1165.9	2
Coking works	1160.7	4
Coking works, main roads	1160.6	6
Coking works, smoking	1159.4	5
Coking works, hay fever	1127.6	5
Hay fever only	1132.5	3

Conclusions:

- hay fever is biggest single risk factor
- proximity to coke works increases risk, with or without prior adjustment for hay fever
- no significant association with main roads, or with cigarette smoking

Adaptation to matched case-control data

Diggle, Morris and Wakefield (2000) consider the fitting of point source models to individually matched case-control data.

Recall that the generic form of the model for matched data is that the matched control intensity corresponding to the ith case is

$$\lambda_0(x) = \alpha_i \lambda(x) \rho(x)$$

where $\lambda(x)$ is the case intensity.

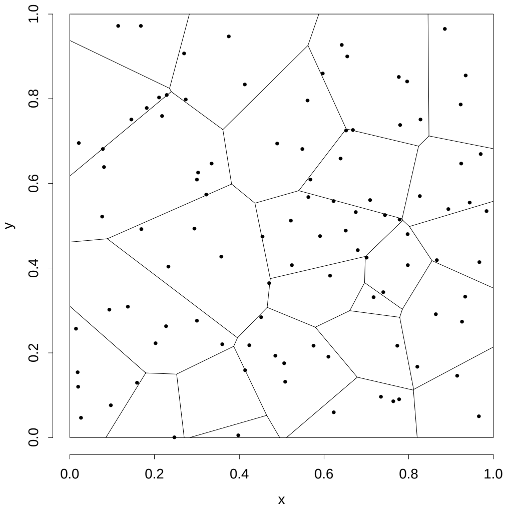
The conditioning argument used arlier in the non-parametric setting again eliminates the nuisance parameters α_i

- DMW compare likelihood-based and Bayesian methods of inference.
- In a Bayesian setting, the use of a conditioning argument to eliminate the nuisance parameters α_i is apparently controversial.

Methods for spatially aggregated data

- case-control studies are expensive
- identifying individual cases with specific point locations is problematic
- disease and population data are often routinely collected as counts in geographical regions (census tracts, counties,...)
- how should we analyse data of this kind?

4.1 Poisson regression modelling



- Let $A_i : i = 1, ..., n$ be a partition of a study region A into sub-regions
- Let Y_i denote the number of cases in A_i
- Suppose cases from a Poisson process with intensity $\lambda(x)$
- Then, the Y_i are mutually independent, $Y_i \sim \text{Poiss}(\mu_i)$, where

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

Poisson regression modelling takes as starting point the model

$$Y_i \sim \text{Poiss}(\mu_i)$$

and incorporates covariate information at the area-level by a log-linear model

$$\log \mu_i = u_i'\beta$$

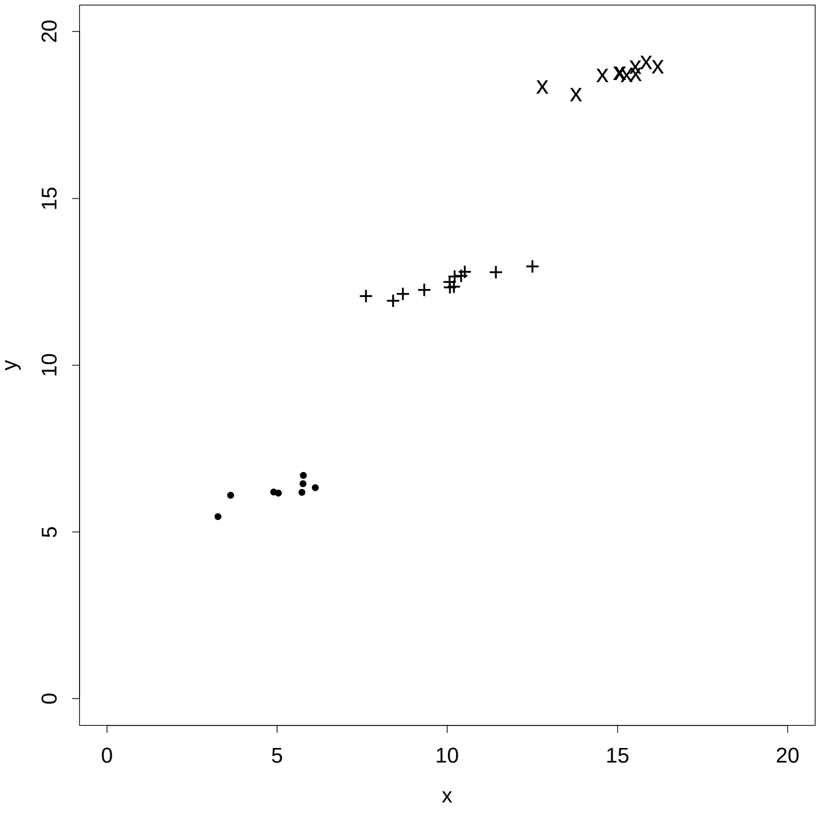
4.2 Ecological bias

When regression models are fitted to area-level data, the effect of covariates on the fitted means μ_i may or may not be the same as the corresponding effects on individual risk.

Differences between covariate effects at individual and at area levels lead to what is usually called *ecological bias*

Ecological bias as model mis-specification

- Suppose that we wish to describe the relationship between an outcome variable Y and an exposure x;
- imagine that we collect data (x_{ij}, Y_{ij}), where i denotes groups (areas), and j denotes individuals within groups.
- the following diagram shows the relationship between x and Y using synthetic data from individuals in three groups:



Consider three possible models for the above synthetic data:

1. **Common individual-level regressions**

   $$Y_{ij} = \alpha + \beta x_{ij} + Z_{ij} : Z_{ij} \sim \text{iid} \quad \text{N}(0, \sigma^2)$$
   $$\bar{Y}_i = \alpha + \beta \bar{x}_i + \bar{Z}_i : \bar{Z}_I \sim \text{id} \quad \text{N}(0, \sigma^2/n_i)$$

2. **Separate individual-level regressions**

   $$Y_{ij} = \alpha_i + \beta x_{ij} + Z_{ij} : Z_{ij} \sim \text{iid} \quad \text{N}(0, \sigma^2)$$
   $$\bar{Y}_i = \alpha_i + \beta \bar{x}_i + \bar{Z}_i : \bar{Z}_I \sim \text{id} \quad \text{N}(0, \sigma^2/n_i)$$

3. **An additional group-level covariate**

   $$Y_{ij} = \alpha_i + \beta x_{ij} + Z_{ij} : Z_{ij} \sim \text{iid} \quad \text{N}(0, \sigma^2)$$
   $$\alpha_i = \alpha + \gamma u_i + Z_i^* : Z_i^* \sim \text{iid} \quad \text{N}(0, \tau^2)$$

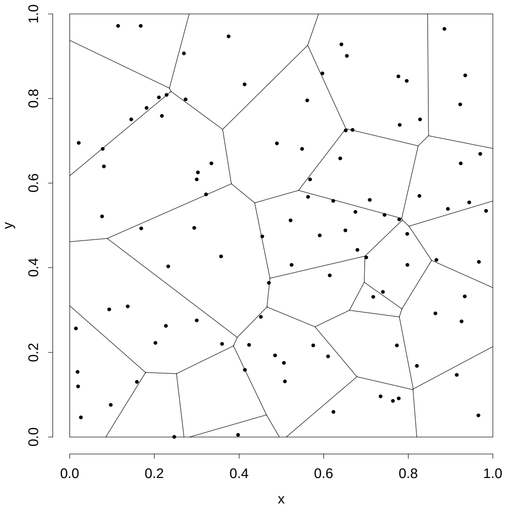
   (a) if $u_i \neq \bar{x}_i$, then
   $$\bar{Y}_i = \alpha + \beta \bar{x}_i + \gamma u_i + (Z_i^* + \bar{Z}_i)$$

   (b) if $u_i = \bar{x}_i$, then
   $$\bar{Y}_i = \alpha + (\beta + \gamma)\bar{x}_i + (Z_i^* + \bar{Z}_i)$$

Comment:

- model 1 is wrong
- model 2 is correct, but not identifiable from group-level data
- model 3a is correct, and is identifiable if u_i is known, and $\mathrm{Corr}(u_i, \bar{x}_i) < 1$
- model 3b is implicitly assumed in an ecological regression of $\bar{Y}_i$ on $\bar{x}_i$

Ecological bias for spatial count data



A_i = small areas $\qquad$ Y_i = counts

$\{z(x) : x \in \cup A_i\}$ = spatially varying risk factor

1. **Individual-level model**

   Cases form an inhomogeneous Poisson process,

$$\lambda(x) = \lambda_0(x) \exp\{\beta z(x)\}$$

2. **Area-level model**

$$\begin{aligned} Y_i &\sim \text{id} \quad \text{Poisson}(\mu_i) \\ \mu_i &= \int_{x \in A_i} \lambda_0(x) \exp\{\beta z(x)\} dx \\ &\neq \bar{\lambda}_{0i} \exp(\beta \bar{z}_i) \end{aligned}$$

For problems of this kind, the author's preferred strategy is to:

- *specify* the model at the individual level
- *derive* the resulting joint probability distrbution for area-level data
- *check* that parameters of interest are identifiable from area-level data
- make the required *inferences*

See, for example, Prentice and Sheppard (1995), Sheppard and Prentice (1995).

4.3 Extra-Poisson variation

- a widely used diagnostic for the goodness of fit of a Poisson regression model is

$$X^2 = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$$

  - for a well-fitting model, $X^2 \sim \chi^2_{n-p}$ where p is the number of fitted regression parameters;
  - When $X^2 >> n-p$, nominal standard errors from the Poisson regression model are too small.

- Provided the Y_i are independent, approximately valid inferences about β are obtained if we multiply nominal standard errors by a factor $\sqrt{X^2/(n-p)}$.
- But in spatial setting, independence is not guaranteed.

5. Real-valued spatial stochastic processes

A **spatial stochastic process** is a collection of random variables Y_i with associated spatial locations x_i.

In particular applications, any of three distinct scenarios may apply:

1. the pairs (x_i, Y_i) may be sampled from an underlying continuous process $\{Y(x) : x \in \mathbb{R}^2\}$
2. each Y_i may be derived from an underlying continuous process $Y(x)$ by integration,
$$Y_i = \int_{A_i} Y(x)dx$$
3. the random variables Y_i may only be defined at a fixed set of locations x_i

Scenarios 1 and 3 define **continuous spatial variation** and **discrete spatial variation**, respectively.

In practice, scenario 2 is often treated as if it were the same as scenario 3.

Historically, models and methods for discrete spatial variation, and for continuous spatial variation, were developed independently. A possible reason is that they were motivated by different areas of application.

5.1 Continuous spatial variation

Most models for continuous spatial variation are based on a **stationary Gaussian process**,

$$S(x) \sim \mathrm{SGP}\{\sigma^2, \rho(u)\}$$

- $\mathrm{E}[S(x)] = 0$
- $\mathrm{Var}\{Y(x)\} = \sigma^2$
- $\mathrm{Corr}\{Y(x), Y(x-u)\} = \rho(u)$
- $\{S(x_1), ..., S(x_n)\} \sim \mathrm{MVN}$

Typical model for measurements Y_i at locations x_i is

- $Y_i = \mu(x_i) + S(x_i) + Z_i : i = 1, ..., n$
- $\mu(x) = \Sigma_{k=1}^{p} z_k(x)\beta_k$
- $Z_i \sim \mathrm{iid} \quad \mathrm{N}(0, \tau^2)$

Models of this kind are often called **geostatistical models** (Cressie, 1991). This is an indirect reference to their historical development in connection with spatial prediction problems in the mining industry.

Typical geostatistical problem: use data Y_i from locations x_i to predict

$$T = \int_A S(x)dx$$

Generalized linear geostatisical models

A different way to think about the geostatistical model is in its conditional form:

$$Y_i | \{S(x) : x \in \mathbb{R}^2\} \sim \text{id} \quad \text{N}\{\mu(x_i) + S(x_i), \tau^2\}$$

This led Diggle, Moyeed and Tawn (1998) to embed the model within the class of generalized linear mixed models (Breslow and Clayton, 1993), to allow for example:

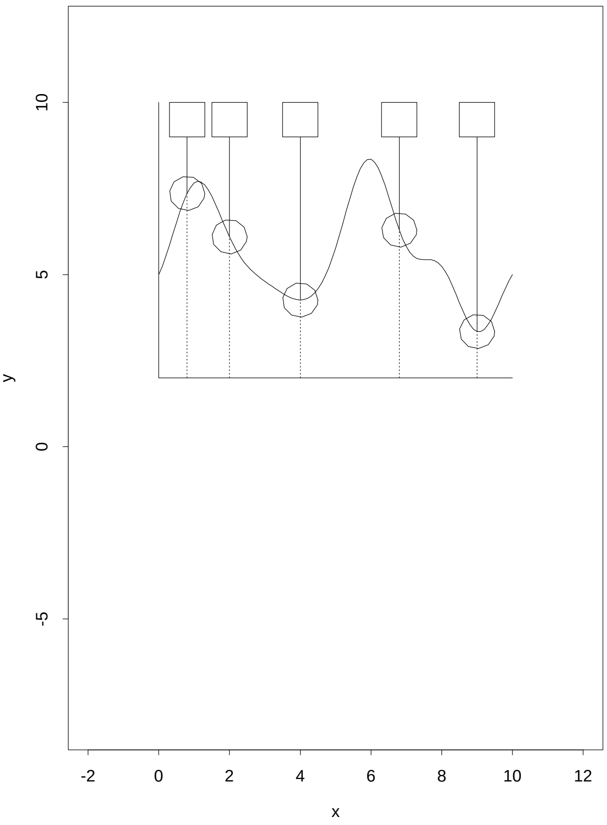
- **Poisson log-linear geostatistical models**

  $S(x) \sim \text{SGP}\{\sigma^2, \rho(u)\}$
  $\alpha_i = \Sigma_{k=1}^{p} z_k(x_i)\beta_k + S(x_i)$
  $Y_i | S(x_i) \sim \text{Poiss}\{\exp(\alpha_i)\}$

- **Binary logistic-linear geostatistical models**

  $S(x) \sim \text{SGP}\{\sigma^2, \rho(u)\}$
  $\alpha_i = \Sigma_{k=1}^{p} z_k(x_i)\beta_k + S(x_i)$
  $Y_i | S(x_i) \sim \text{Bernoulli}[\exp(\alpha_i)\}/\{1 + \exp(\alpha_i)\}]$

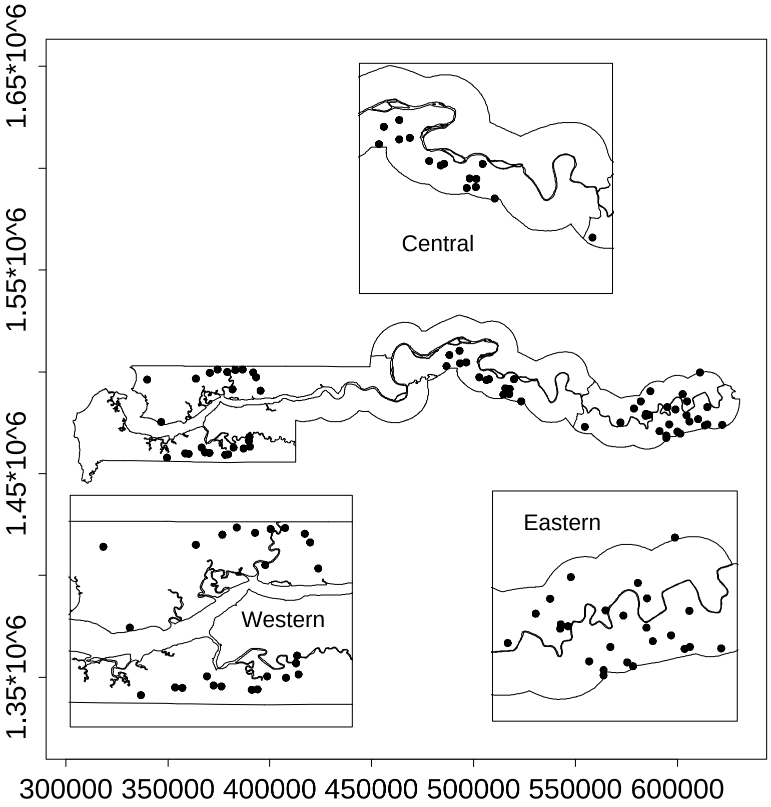
The structure of the DMT model is illustrated (in one spatial dimension) by the following diagram:



Over the last ten years, geostatistical models have been used in many different areas of application, including environmental epidemiology.

Example 7. Childhood malaria in the Gambia

A survey was conducted in village communities throughout the Gambia, as shown in the following map:



For each child in the survey, the following information was recorded:

- Y_{ij} = presence/absence of malarial parasites in blood-sample, for jth child in ith village
- age, sex, bednet use
- satellite-derived vegetation green-ness index

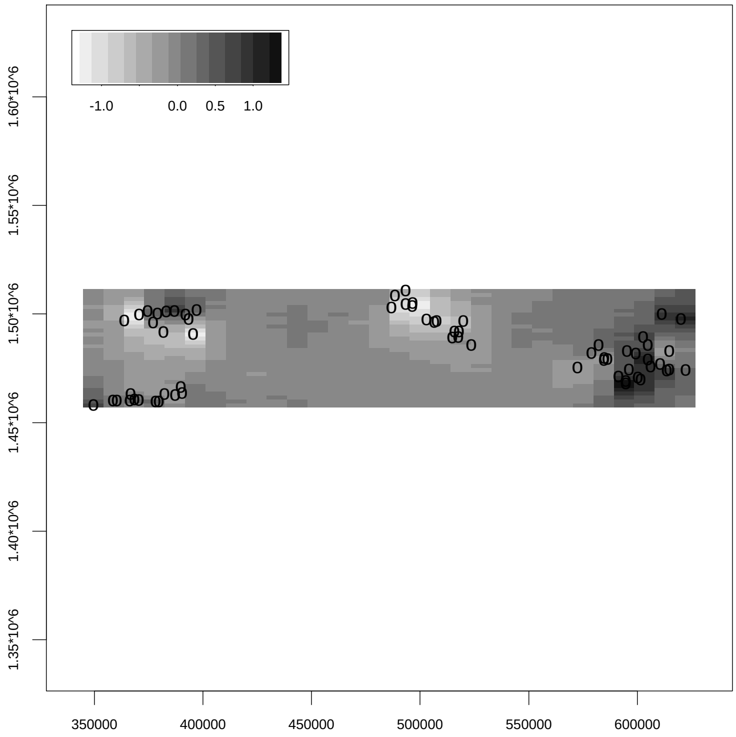
A logistic regression model, allowing for residual spatial variation *and* residual non-spatial variation between villages, is:

$$\mathrm{logit}\mathrm{P}(Y_{ij} = 1) = \sum_{k=1}^{p} z_{ijk}\beta_k + U_i + S(x_i)$$

where $U_i \sim \mathrm{N}(0, \nu^2)$ and $\{S(x) : x \in \mathbb{R}^2\}$ is a stationary Gaussian process.

Results:

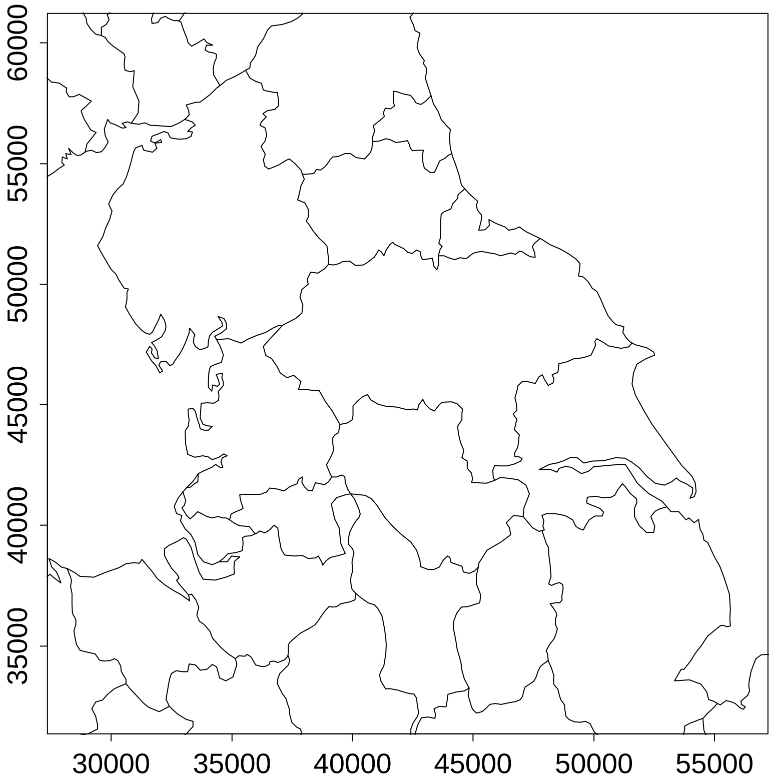
- inclusion of either U_i or $S(x_i)$ term materially affects inferences about β_k
- map of $\hat{U}_i$ against village locations suggests strong spatial structure
- map of $\hat{S}(x)$ shows spatial smoothing of unexplained spatial variation in risk



5.2 Discrete spatial variation

In models of **discrete spatial variation**, the geographical space under study is regarded as a fixed set of spatial sampling units, typically defined by a partitioning of a continuous region into politically defined sub-regions.

In the following example, the region is the north of England and the sub-regions are counties.



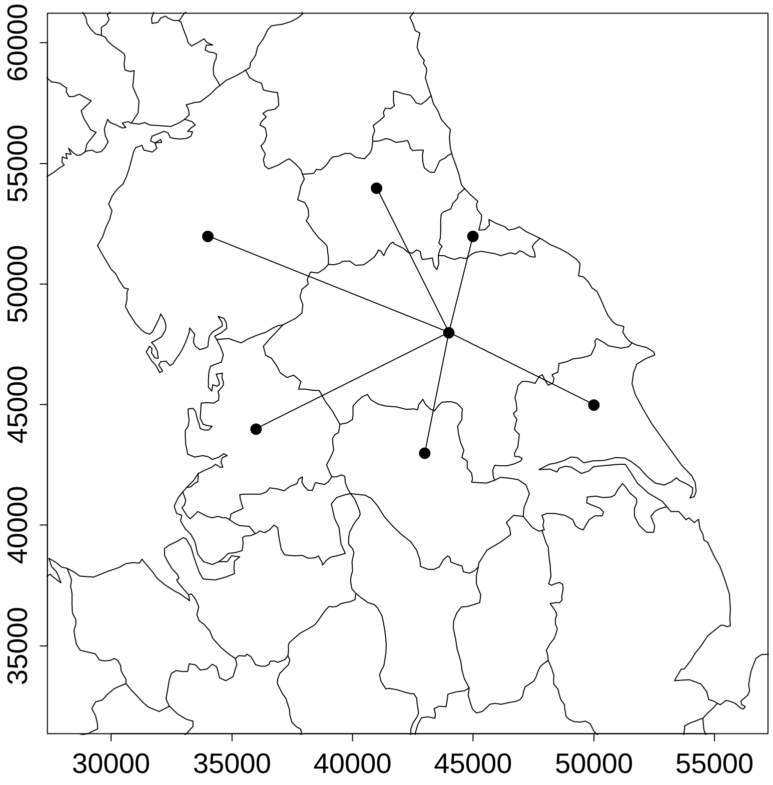
- Models for discrete spatial variation are usually defined in terms of their so-called full conditional distributions, incorporating notions of “local” dependence between spatial units.

- More formally, if $Y_i : i = 1, ..., n$ denote a set of outcome variables associated with each of n spatial units, the model is specified by the n univariate distributions

$$[Y_i|Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n]$$

- Note that a mutually consistent specification of the full conditionals involves non-obvious constraints on the allowable forms of distribution, which are set out in the celebrated Hammersley-Clifford Theorem (Besag, 1974).

- Typically, simplifying assumptions are made so that only a few of the $n - 1$ terms in the conditioning set play any part.
- The following shows one example of how this might be done, based on the counties of northern England.



Example 8. Lip cancer in Scotland

This example was originally analysed in Clayton and Kaldor (1987), with further comment and analysis in Clayton and Bernardinelli (1992) and in Breslow and Clayton (1993).

- spatial units are the counties of Scotland;
- the response from each county is Y_i, the number of cases during the years 1975-1980 inclusive;
- let R_i denote the risk for county i, and N_i the size of the population at risk
- a natural model to fit to the data is that

$$Y_i | R_i \sim \text{Poiss}(N_i R_i)$$

- an available covariate is x_i, the percentage of the population in each county who are engaged in agriculture, fishing or forestry.

100000
90000
80000
70000
60000
10000
20000
30000
40000
50000

To model residual spatial variation in risk, after adjusting for the available covariate, we assume that

$$\log R_i = \log N_i + \alpha + x_i\beta + S_i$$

where the S_i follow a discrete spatial variation model in which:

- two counties are ***neighbours*** if they share a common boundary
- the full conditionals of county i depend only on the neighbours of county i
- $Y_i|\text{neighbours} \sim \text{N}(m_i, v_i)$ where
  - $m_i =$ mean of S_j from counties j which are neighbours of county i;
  - $v_i = \sigma^2/m_i$, where $m_i =$ number of neighbours of county i

Note that this specification corresponds to an improper joint distribution for $(S_1, ..., S_n)$, with joint pdf

$$f(s-1, ..., s_n) \propto \exp\{-\sum_{i\sim j}(s_i - s_j)^2/(2\sigma^2)\}$$

where $i \sim j$ indicates that counties i and j are neighbours.

5.3 Including both spatial and non-spatial variation

When there is a “natural” statistical model for the data, such as the Poisson model for count data or the binomial for binary data, it is often the case that:

- after adjusting for all known covariates, the residual variation is bigger than can be explained by the assumed distributional model
- but this extra-variation may or may not be spatially structured

As in our analysis of the Gambia malaria data, a cautious strategy is to contemplate a model with both spatial and non-spatial sources of extra-variation.

Example 8 (continued). Lip cancer in Scotland

Replace the previous model for risk by

$$\log R_i = \log N_i + \alpha + x_i\beta + U_i + S_i$$

where S_i follow a discrete spatial variation model as before, and U_i are mutually independent $\mathrm{N}(0, \nu^2)$.

The following table of results is adapted from Breslow and Clayton (1993). Three different models are considered:

1. Poisson regression (include neither U_i nor S_i terms)
2. non-spatial extra-Poisson variation (include U_i)
3. spatial extra-Poisson variation (include S_i)

Model	Estimates ± standard errors			
	α	$\beta \times 10$	σ	ν
1	-0.54 ± 0.07	0.74 ± 0.06	—	—
2	-0.44 ± 0.16	0.68 ± 0.14	—	0.60 ± 0.08
3	-0.18 ± 0.12	0.35 ± 0.12	0.73 ± 0.13	—

1. Poisson regression model seriously underestimates standard errors of α and β;
2. non-spatial extra-Poisson variation model fixes up standard errors, point estimates only change slightly;
3. spatial extra-Poisson variation model gives comparable standard errors to model 2 but materially changes point estimates.

5.4 Continuous vs discrete spatial variation models?

- Sometimes, spatial variation is genuinely discrete.
- More often, discrete models are used pragmatically.

Advantages of continuous spatial variation models include:

- they are more natural than discrete spatial variation models
- their interpretation does not rely on (arbitrary?) definitions of sub-regions

Advantages of discrete spatial variation models include:

- data are often only available from spatially aggregated sampling units
- inference (via MCMC) is computationally much easier than for continuous spatial variation models

6 Concluding remarks

- almost everything we have discussed in this course has its analogue in a space-time setting
- methods for space-time data are less well developed than for purely spatial (or purely temporal) data, but this situation is changing rapidly
- this course has included very little detailed discussion of **inference** for spatial data:
  - for exploratory analysis of case-control data, non-parametric methods are widely used, and hypothesis testing can be based on the randomisation distribution induced by the study design;
  - in other settings, parametric modelling assumptions are widely used, and inference uses likelihood-based methods, whether classical or (increasingly) Bayesian;
  - the Bayesian paradigm is particulary well suited to problems involving predictive inference for latent spatial processes, because it naturally adjusts for parameter uncertainty in the constuction of prediction intervals.

- software to implement most of the methods described in this course is freely available:
  - point patterns: **Splancs**

    `www.maths.lancs.ac.uk/Software/Splancs`
  - continuous spatial variation: **geoS**

    `www.maths.lancs.ac.uk/~ribeiro/geoS.html`
  - discrete spatial variation: BUGS

    `www.mrc-bsu.cam.ac.uk/bugs`

Splancs and **geoS** are add-ons to **Splus**.

BUGS is stand-alone software.

7. References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *Journal of the Royal Statistical Society* B **36**, 192–225.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9–25.

Chetwynd, A.G. and Diggle, P.J. (1998). On estimating the reduced second moment measure of a stationary spatial point process. *Australian and New Zealand Journal of Statistics*, **40**, 11-15.

Chetwynd, A.G., Diggle, P.J., Marshall, A. and Parslow, R. (2000). Investigation of spatial clustering from matched and stratified case-control studies. (submitted)

Clayton, D. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology*, eds. P. Elliott, J. Cuzick, D. English and R. Stern, 205–20. Oxford : Oxford University Press.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardised relative risks for use in disease mapping. *Biometrics* **43**, 671–81.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* New York : Wiley.

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with Discussion). *Journal of the Royal Statistical Society*, B **52**, 73–104.

Diggle, P.J. (1983). *Statistical Analysis of Spatial Point Patterns.* London : Academic Press.

Diggle, P.J. and Chetwynd, A.G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**, 1155–63.

Diggle, P.J., Gatrell, A.C. and Lovett, A.A. (1990). Modelling the prevalence of cancer of the larynx in part of Lancashire : a new methodology for spatial epidemiology. In R.M. Thomas (ed.) *Spatial Epidemiology*, London Papers in Regional Science No. 21. London : Pion.

Diggle, P.J., Morris, S.E. and Wakefield, J.C. (2000). Point-source modelling using matched case-control data. *Biostatistics*, **1** (to appear).

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998). Model-based Geostatistics (with Discussion). *Applied Statistics* **47** 299–350.

Diggle, P.J. and Rowlingson, B.S. (1994). A conditional approach to point process modelling of raised incidence. *J. R. Statist. Soc. A* **157**, 433–40.

Elliot, P., Westlake, A., Hills, M., Kleinschmidt, I., Rodrigues, L., McGabe, P., Marshall, K. and Rose, G. (1992). The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom. *Journal of Epidemiology and Community Health*, **46**, 345–9.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models.* London : Chapman and Hall.

Jarner, M.F., Diggle, P.J. and Chetwynd, A.G. (2000). Nonparametric estimation of spatial variation in risk from matched case-control data. (in preparation)

Kelsall, J.E. and Diggle, P.J. (1998). Spatial variation in risk: a nonparametric binary regression approach. *Applied Statistics* **47**, 559–73.

Kelsall. J.E. and Wakefield, J.C. (2000). Modelling spatial variation in disease risk. *Journal of the American Statistical Association* (submitted).

Prentice, R.L. and Sheppard, L. (1995). Aggregate data studies of disease risk factors. *Biometrika*, **82**, 113–25.

Rowlingson, B.S. and Diggle, P.J. (1993). SPLANCS : Spatial point pattern analysis code in S-plus. *Computers in Geosciences*, **19**, 627-55.

Sheppard, L. and Prentice, R.L. (1995). On the reliability and precision of within- and between-population estimates of relative rate parameters. *Biometrics*, **51**, 853–63.

Stone, R.A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in Medicine*, **7**, 649–60.