Spatial Statistics in Epidemiology SoSe2009 Wednesday: Point processes

Michael Höhle[1]

[1]Department of Statistics
Ludwig-Maximilians-Universität München

8 July 2009
IBE, LMU München

Outline of todays lecture(s)

1. Introduction to Spatial Point Processes
2. Point process theory
3. Point process models and inference
4. Application in epidemiology

Outline

1. Introduction to Spatial Point Processes
2. Point process theory
3. Point process models and inference
4. Application in epidemiology

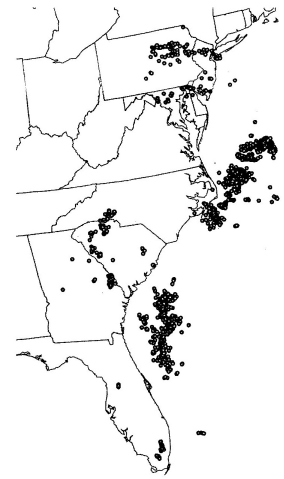
Goals for today's module

- Get a basic understanding of spatial point process theory: Intensity and Poisson process.
- Provide an appetizer for further reading in e.g. Gatrell et al. (1996) or Diggle (2003).
- Use R! Packages for the analysis of spatial point process data are `splancs` and `spatstat` – both available from CRAN.

Point Process data (1)

- Notation:
  - Z Random locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ of events in $D \subset \mathbb{R}^2$
  - Y Possible mark of events, $Y = (Y_1, \ldots, Y_n)$ with Y_i having given support (continuous, discrete, categorical)
  - $\mathbf{x}(\mathbf{s})$ vector of covariates continuously observed in S (if available)
- Contrary to *geostatistical data*, focus is now on *where* the event occurs
- Spatial point processes are an extension to temporal point processes known from e.g. survival analysis.

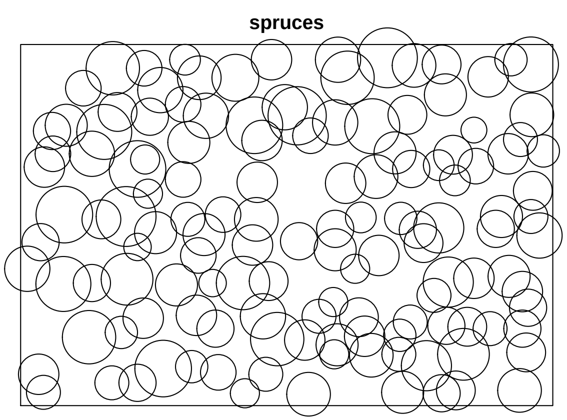
Example: Location of lightnings, 17-20 April 2003



Location of lightning strikes within approximately 200 miles of US east coast mentioned in (Schabenberger and Gotway, 2005).

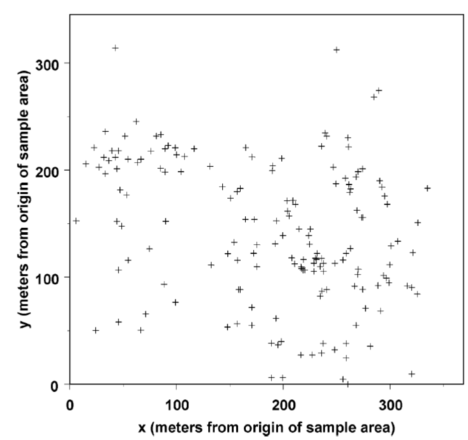
Example: Location of Norwegian spruce trees

- Point pattern of tree locations in a 56 x 38 metre sampling region in a natural forest stand in Saxonia, Germany.
- Each tree is marked with its diameter at breast height.



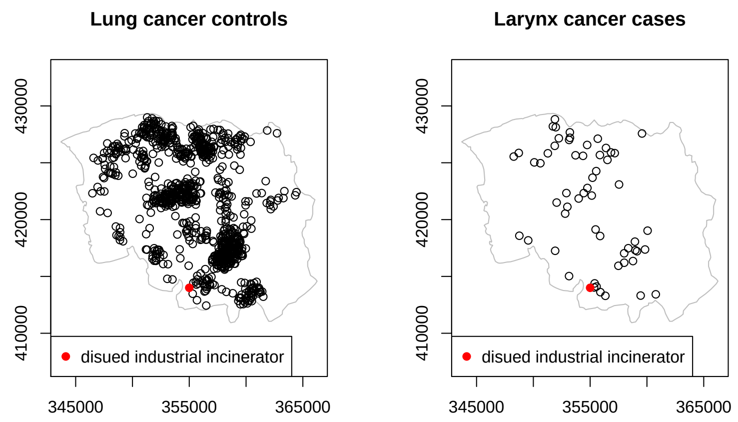
Example: UXO locations

Location of unexploded ordnance device (UXO) at the former US military training range Fort Ord, CA, USA described in Macdonald and Small (2006)



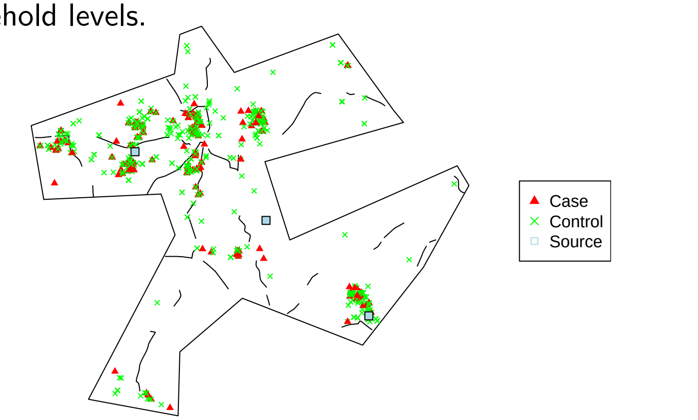
Example: Cancer cases in Chorley-Ribble, UK, 1973-1984

- Investigation of a risk elevation around a putative source of environmental pollution in Diggle (1990).
- 57 larynx cases in Ribble Health Authority, Lancashire, England. Collateral information on 917 lung cancers cases.



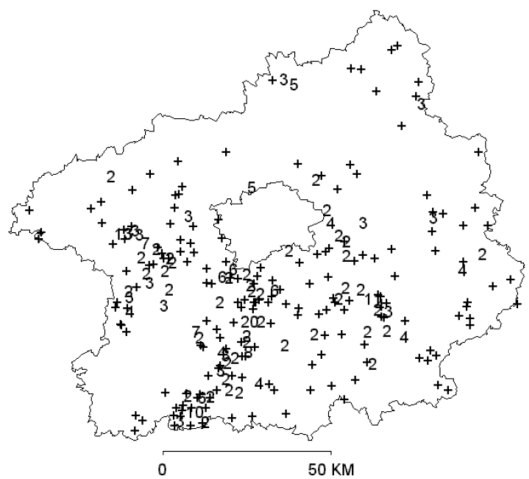
Example: Asthma in North Derbyshire

- Case-control study on the incidence of asthma in children in North Derbyshire described in Diggle and Rowlingson (1994).
- 215 cases and 1076 controls from 10 schools.
- Risk factors: Proximity to main roads and three putative pollution sources. Other relevant covariates at the individual and household levels.



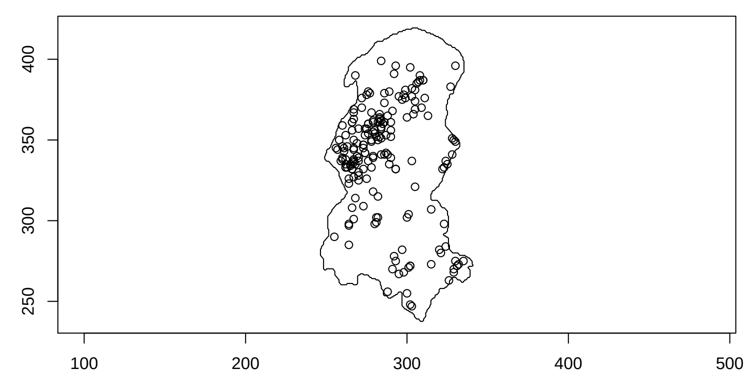
Example: Location of infection of tick-borne encephalitis

Data reported in Benes et al. (2005) on the locations of infection of 446 cases of tick-borne tick-borne encephalitis in Central Bohemia, Czech Republic.



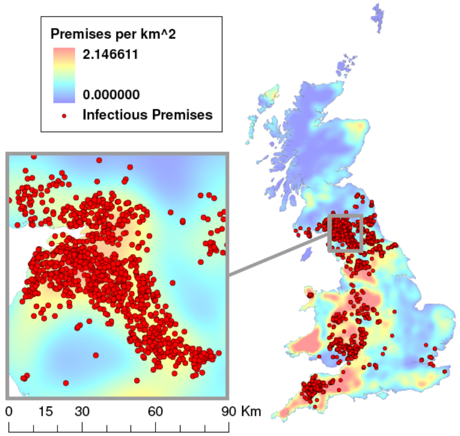
Example: Burkitt's lymphoma

Spatio-temporal data on Burkitt's lymphoma cases in West Nile district district of Uganda 1960-1975.



Example: Infected premises during 2001 FMD epidemic

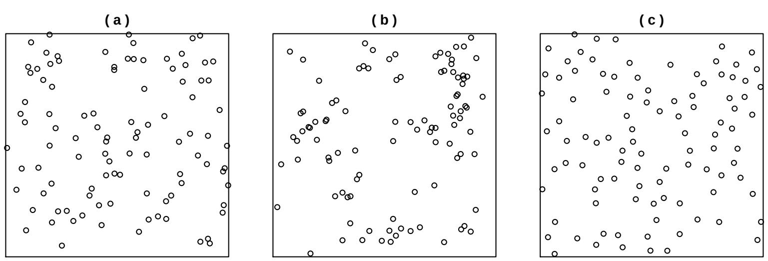
Location of infected farms during the 2001 foot and mouth epidemic in the UK. Taken from Jewell et al. (2009).



Outline

Types of point patterns (1)

- Realizations of a (a) completely random pattern, (b) a Poisson cluster process and (c) a regular process with inhibition
- All patterns have $n = 100$ points on $[0, 1] \times [0, 1]$.



Types of point patterns (2)

- *Complete spatial random* (CSR) pattern:
  - The average number of events per unit area is homogeneous throughout D.
  - The number of events $N(A)$ in a subregion A is Poisson distributed.
  - The number of events $N(A_1)$ and $N(A_2)$ in two non-overlapping patterns A_1 and A_2 are independent.
- *Clustered pattern*: the average distance between an event $\mathbf{s}_i$ and its nearest neighbour event is smaller than the same average distance in a CSR pattern.
- *Regular pattern*: the average distance between an event $\mathbf{s}_i$ and its nearest neighbour is larger than in a CSR pattern.

Testing for complete spatial randomness (1)

- Null-hypothesis: Observed pattern is a realization of a homogeneous Poisson process.
- Partition study region D into non-overlapping regions (quadrants) $A_1, \dots, A_k$ of equal size. Specifically, use a rectangle with r rows and c columns.
- Let n_{ij} be the observed number of counts in cell (i,j), $i = 1, \dots, r$, $j = 1, \dots, c$.
- The expected number of events of quadrant under the null-hypothesis is $\overline{n} = n/(r \times c)$.

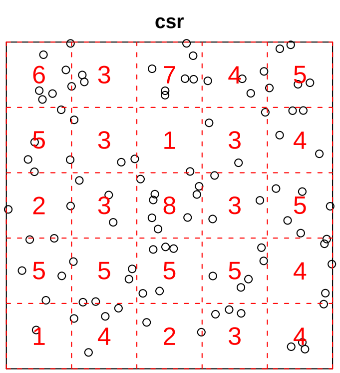
Testing for complete spatial randomness (2)

- Use a χ^2-test to investigate the null hypothesis based on the test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \overline{n})^2}{\overline{n}}.$$

- Under the null hypothesis $X^2 \sim \chi^2(rc - 1)$.
- Alternatively, one could use a Monte Carlo test to compute the p-value. $\rightarrow$ R

Testing for complete spatial randomness (3)




```
> quadrat.test(csr, nx = 5, ny = 5)

        Chi-squared test of CSR using quadrat counts

data:  csr
X-squared = 17, df = 24, p-value = 0.8487
```


Research questions formulated in Gatrell et al. (1996)

- Is observed clustering mainly due to natural background variation in the population from which events arise?
- Over what spatial scale does any clustering occur?
- Are clusters associated with proximity to specific features of interest, such as transport arteries or possible point sources of pollution?
- Are events that aggregate in space also clustered in time?

Bernoulli and Binomial Processes

- Let $\nu(A)$ measure the area of a region $A \subset D$.
- Assume a single event $\mathbf{s}$ is distributed in D such that

$$P(\mathbf{s} \in A) = \frac{\nu(A)}{\nu(D)}$$

  for all sets $A \subset D$. This process is termed a *Bernoulli process*
- The superposition of n Bernoulli processes is termed a *Binomial process*. If $N(D) = n$ then for $A \subset D$

$$N(A) \sim \text{Bin}\left(n, \frac{\nu(A)}{\nu(D)}\right).$$

First-order intensity function

- First-order intensity of a spatial point process

$$\lambda(\mathbf{s}) = \lim_{\nu(d\mathbf{s})\to 0} \frac{E(N(d\mathbf{s}))}{\nu(d\mathbf{s})}$$

- Interpretation: $\lambda(\mathbf{s})\nu(d\mathbf{s})$ for a small area $d\mathbf{s}$ describes the probability for an event in $d\mathbf{s}$.
- The above definition of intensity for a process in $\mathbb{R}^2$ (space) is similar to the definition the hazard rate for a process in $\mathbb{R}$ (time).
- Intensity function for the binomial process? (✎)

Homogeneous Poisson process

A spatial point process is called *homogeneous Poisson process* if it has the following two properties:

(i) $N(A) \sim \text{Po}(\lambda\nu(A))$, where $0 < \lambda < \infty$ denotes the constant intensity function and $A \subset D$.

(ii) If A_1 and A_2 are two disjoint subregions of D, then $N(A_1)$ and $N(A_2)$ are independent.

The homogeneous Poisson process acts as reference process defining complete spatial randomness (CSR).

Inhomogeneous Poisson process

If the intensity function $\lambda(\mathbf{s})$ varies spatially, then property (i) of the homogeneous Poisson process is violated, but (ii) may still hold.

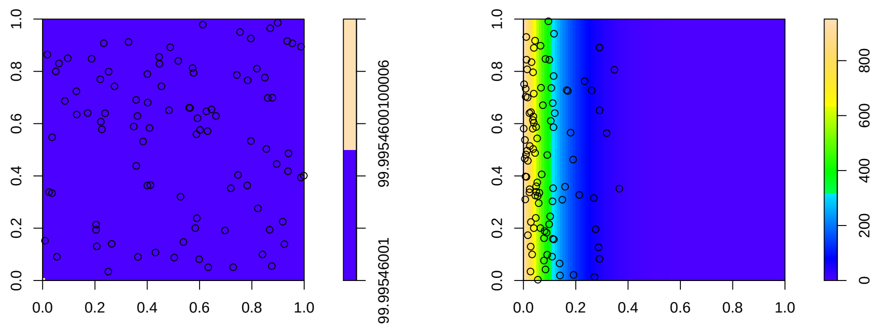
The *inhomogeneous Poisson process* is defined by the following two properties:

(i) $N(A) \sim \text{Po}(\int_A \lambda(\mathbf{s})d\mathbf{s})$, where $0 < \lambda(\mathbf{s}) < \infty$ is the intensity of the process at $\mathbf{s} \in D$.

(ii) If A_1 and A_2 are two disjoint subregions of D, then $N(A_1)$ and $N(A_2)$ are independent.

Simulation from Poisson process

Two Poisson process realizations on the unit square having the same $\int_D \lambda(\mathbf{s}) d\mathbf{s} = 99.9955$.



(a) Homogeneous

(b) Inhomogeneous $\lambda(\mathbf{s}) = 1000 \exp(-10 s_x)$

Aside: Kernel density estimation (1)

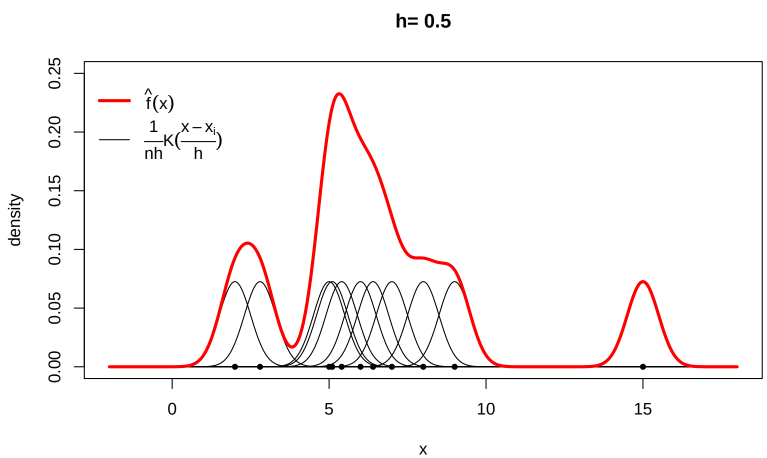
- Given an independent and identical distributed sample $(x_1, \ldots, x_n)$ of a univariate random variable X, *kernel density estimation* is a non-parametric procedure to estimate the density $f(\cdot)$ of X.

$$\hat{f}(x) = \frac{1}{n\,h} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- The function $K(\cdot)$ is called the *kernel* and is a probability density function.
- The *bandwidth* controls the smoothness of the resulting density function estimator.

Aside: Kernel density estimation (2)

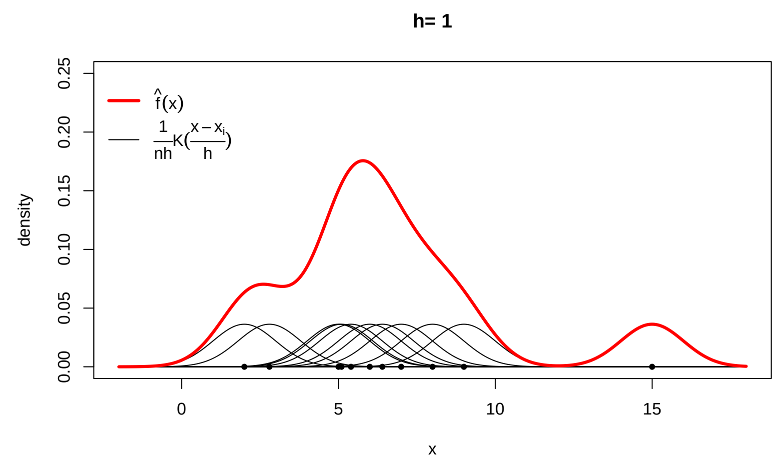
Graphical illustration of univariate kernel density estimation for a sample of size $n = 11$ using a standard normal kernel :



Intro Theory Models and inference Application References

Aside: Kernel density estimation (2)

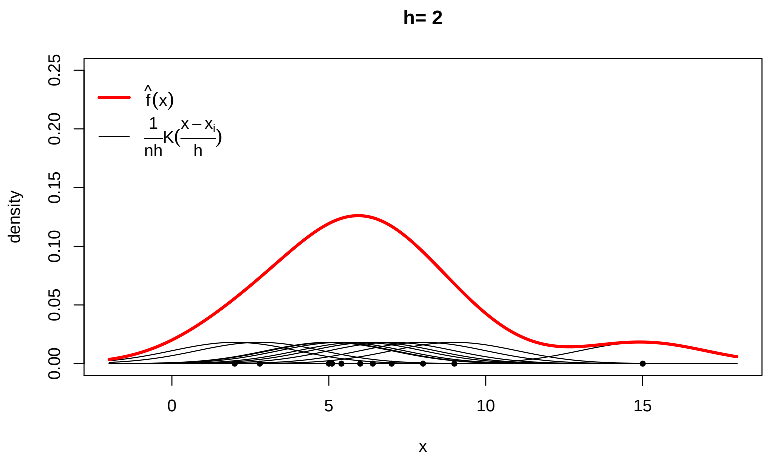
Graphical illustration of univariate kernel density estimation for a sample of size $n = 11$ using a standard normal kernel :



Michael Höhle SpatialEpi2009 27/ 48

Aside: Kernel density estimation (2)

Graphical illustration of univariate kernel density estimation for a sample of size $n = 11$ using a standard normal kernel :



Kernel estimation of the intensity function (1)

- Estimating the intensity of a spatial point pattern is similar to estimating a bivariate probability density.
- Kernel smoothing is a technique to estimate a bivariate probability density function of a random variable $\mathbf{Y}$ based on a sample $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$.
- An estimate of the density at $\mathbf{y}$ can be found as the (weighted) number of sample realizations within a certain distance h from $\mathbf{y}$:

$$f(\mathbf{y}) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{\mathbf{y}_i - \mathbf{y}}{h}\right),$$

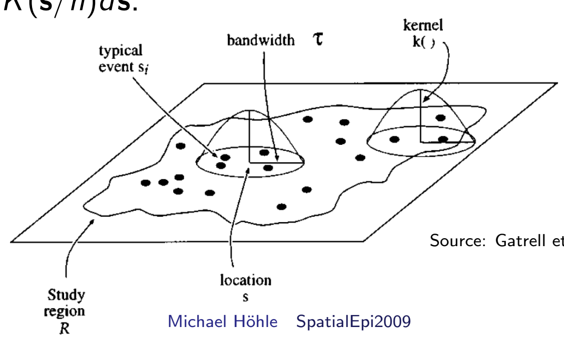
where $K(\cdot)$ is a kernel function and h is the *bandwidth*. $\rightarrow$

Kernel estimation of the intensity function (2)

- The expression for kernel smoothing of the intensity function at location $\mathbf{s}_0$ becomes

$$\hat{\lambda}(\mathbf{s}) = \frac{1}{h^2} \sum_{i=1}^{n} K\left(\frac{\mathbf{s}_i - \mathbf{s}}{h}\right)$$

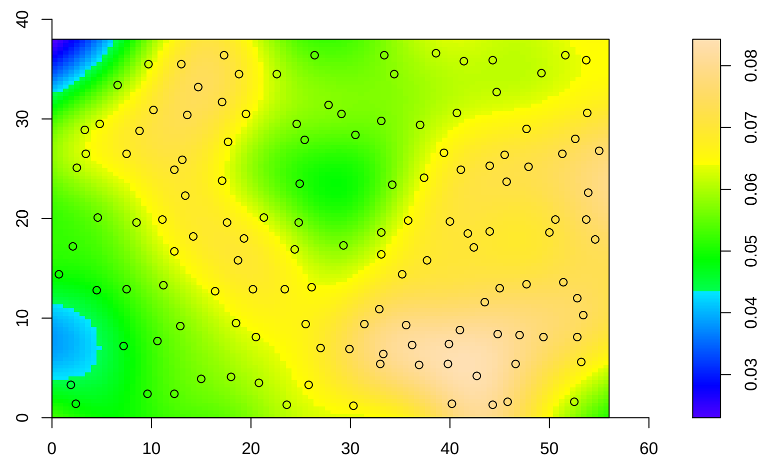
- One can use an additional edge correction by dividing with $h^{-2} \int_D K(\mathbf{s}/h) d\mathbf{s}$.



Source: Gatrell et al. (1996)

Kernel estimation of the intensity function (3)

Kernel estimated $\hat{\lambda}(\mathbf{s})$ for the Norwegian spruce data using the `density{spatstat}` function.



Second-order intensity

- The second-order properties of a point process involve the relationship between number of events in two subregions of D
- The *second-order* intensity of a spatial point process is defined as

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = \lim_{\nu(d\mathbf{s}_i), \nu(d\mathbf{s}_j) \to 0} \left\{ \frac{E\left[N(d\mathbf{s}_i) N(d\mathbf{s}_j)\right]}{\nu(d\mathbf{s}_i) \cdot \nu(d\mathbf{s}_j)} \right\}$$

- Interpretation: For small $d\mathbf{s}_i, d\mathbf{s}_j$

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) \cdot \nu(d\mathbf{s}_i) \cdot \nu(d\mathbf{s}_j)$$

  is approximately the probability that we will have an event in $d\mathbf{s}_i$ *and* an event in $d\mathbf{s}_j$.

Stationary and isotropic process

- We call a point process *stationary* if the intensity is constant over D, i.e. $\lambda(\mathbf{s}) = \lambda$, $s \in D$, and $\gamma(\mathbf{s}_i, \mathbf{s}_j) = \gamma(\mathbf{s}_i - \mathbf{s}_j)$, i.e. depends only on direction and distance.
- A stationary point process termed *isotropic* if

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = \gamma(||\mathbf{s}_i, \mathbf{s}_j||),$$

  i.e. the second-order properties only depend on distance.

K-function

- If the process is stationary and isotropic an alternative characterization of the $\gamma(\mathbf{s}_i, \mathbf{s}_j)$ is the *K-function* of the point process:

  $\lambda K(h)$ is the expected number of extra events within distance h from an arbitrary event.
- Mathematical definition:

$$K(h) = \frac{2\pi}{\lambda^2} \int_0^h x\gamma(x) dx$$

- $K(h)$ for a homogeneous Poisson process? (✎)

Estimating the K-function (1)

- For a stationary process the natural estimator of the constant intensity function $\lambda(\mathbf{s}) = \lambda$ is

$$\hat{\lambda} = N(D)/\nu(D).$$

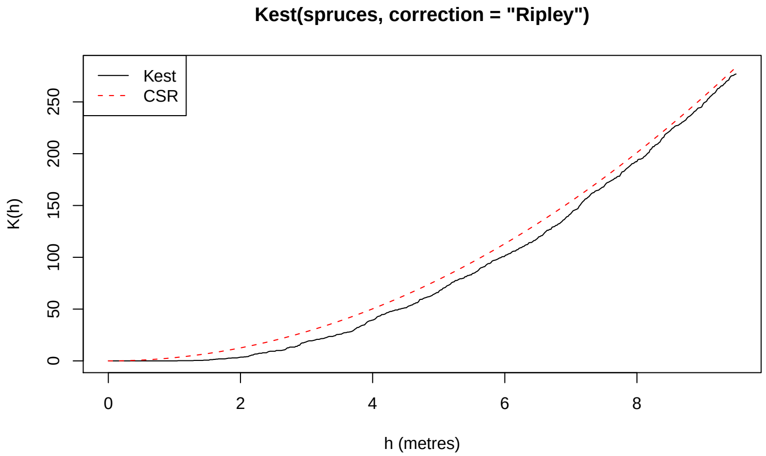
- A naive moment estimator of $E(h)$, i.e. the expected number of extra events within distance h, is

$$\hat{E}(h) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} I(||\mathbf{s}_i - \mathbf{s}_j|| \leq h).$$

  Then $\hat{K}(h) = \hat{\lambda}^{-1} \hat{E}(h) \to$ R: `Kest{spatstat}`.
- Because events outside the observation region D are not observed, this estimator is negatively biased $\to$ edge correction.

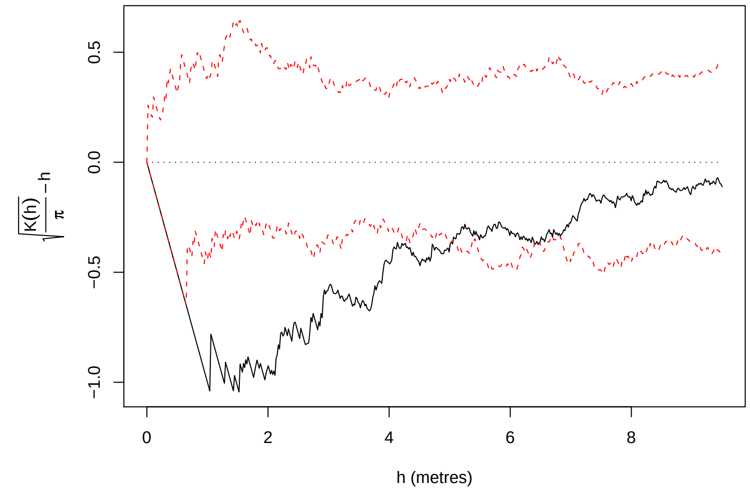
Estimating the K-function (2)

Reconsidering the Norwegian spruce data. Furthermore, we compare with 99 simulated K-functions under the null model.



Estimating the K-function (2)

Reconsidering the Norwegian spruce data. Furthermore, we compare with 99 simulated K-functions under the null model.



Outline

Inhomogeneous Poisson process model (1)

- A regression approach for point process modelling aims at parameterizing $\lambda(\mathbf{s})$ by covariates:

$$\lambda(\mathbf{s}) = \exp(\mathbf{x}(\mathbf{s})'\beta),$$

  where $\mathbf{x}(\mathbf{s})$ denotes the p-dimensional covariate vector at location $\mathbf{s} \in D$.
- Contrary to geostatistical data, the covariates have to be known at each $\mathbf{s} \in D$.

Inhomogeneous Poisson process model (1)

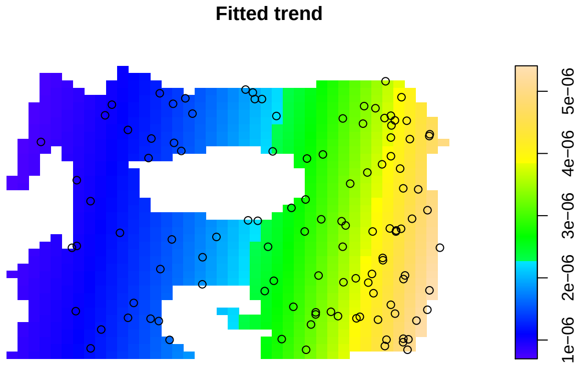
- The loglikelihood of an observed point patterns $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$ is given by:

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \lambda(\mathbf{s}_i; \boldsymbol{\theta}) - \int_D \lambda(\mathbf{s}; \boldsymbol{\theta}) d\mathbf{s}$$

- Find the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ using numerical optimization.
- By clever handling of the numerical integration the optimization problem can be translated into one which can be solved by a Poisson GLM.

Inhomogeneous Poisson process model (1)

```
> data(demopat)
> pfit <- ppm(demopat, ~polynom(x, y, 1), Poisson())
> plot(pfit, se = FALSE)
```



Outline

Spatial variation of relative risk (1)

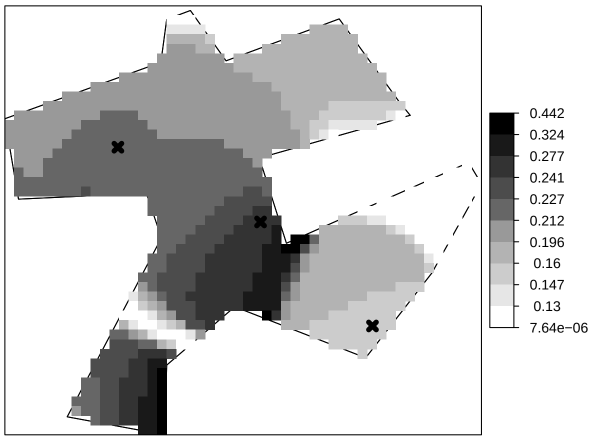
- Comparison of spatial risk by the relative intensities

$$\rho(\mathbf{s}) = \frac{\lambda_{\text{case}}(\mathbf{s})}{\lambda_{\text{control}}(\mathbf{s})}$$

- Compute kernel estimators of intensity functions $\hat{\lambda}_{\text{case}}(\mathbf{s})$ and $\hat{\lambda}_{\text{control}}(\mathbf{s})$. Then determine ratio $\hat{\rho}(\mathbf{s})$.
- Visualization of $\hat{\rho}(\mathbf{s})$ provides descriptive plot for generating hypotheses.
- The example is taken from Bivand et al. (2008, Chapter 7).

Spatial variation of relative risk (2)

Application to the asthma data in North Derbyshire:



Spatial variation of relative risk (2)

- Under the null hypothesis of equal spatial distribution

$$\rho(\mathbf{s}) = \rho_0 = \frac{n_1}{n_0} = 0.20$$

- A Monte Carlo test for the hypothesis of equal spatial risk is described in Bivand et al. (2008). In the asthma example the null hypothesis can not be rejected ($p = 0.69$).

Point source pollution (1)

- Diggle and Rowlingson (1994) investigate point source pollution by formulating the following intensity model:

$$\lambda_1(\mathbf{s}) = \rho\lambda_0(\mathbf{s}) f(\mathbf{s} - \mathbf{s}_0; \boldsymbol{\theta}),$$

  where ρ measures the overall number of events per unit area, $\lambda_0(\mathbf{s})$ is the spatial variation of the underlying population and $f(u; \boldsymbol{\theta})$ is a function of the distance between between the point and the source located at $\mathbf{s}_0$.
- The distance function could e.g. be a decaying function

$$f(u; \boldsymbol{\alpha}, \boldsymbol{\beta}) = 1 + \alpha \exp(-\beta u^2),$$

  where $\beta \in \mathbb{R}$ and $\alpha \geq 0$ describes the exposure effect.

Point source pollution (2)

- Estimation can be performed by maximum likelihood with kernel estimated $\hat{\lambda}_0(\mathbf{s})$.
- Alternative: Condition on the location of cases and controls and determine the probability of becoming a case at location $\mathbf{s}$

$$p(\mathbf{s}) = \frac{\lambda_1(\mathbf{s})}{\lambda_0(\mathbf{s}) + \lambda_1(\mathbf{s})} = \frac{\rho f(\mathbf{s} - \mathbf{s}_0; \boldsymbol{\theta})}{1 + \rho f(\mathbf{s} - \mathbf{s}_0; \boldsymbol{\theta})}.$$

- Estimation of ρ and $\boldsymbol{\theta}$ by maximum likelihood $\rightarrow$ R: `tribble{splancs}`.
- The framework also handles multiple pollution sources and allows for additional covariates.

Point source pollution (3)

```
> case <- ifelse(spasthma$Asthma == "case", 1, 0)
> d2source2 <- as.matrix(sqrt(spasthma$d2source2))
> RHO <- ncases/ncontrols
> m.dr <- tribble(ccflag = case, vars = d2source2, rho = RHO, alphas =
+     betas = 1)

Call:
tribble(ccflag = case, vars = d2source2, alphas = 1, betas = 1,
    rho = RHO)
Kcode = 2

Distance decay parameters:
        Alpha     Beta
[1,] 1.305824 25.14672

rho parameter : 0.163395847627903

     log-likelihood : -580.495955916672
null log-likelihood : -581.406203518987

        D = 2(L-Lo) : 1.82049520462942
```


Literature I

Benes, V., Bodlák, K., Møller, J., and Waagepetersen, R. (2005). A case study on point process modelling in disease mapping. *Image Analysis and Stereology*, 24:159–168.

Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). *Applied Spatial Data Analysis with R*. Springer.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A*, 153(3):349–362.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Arnold, 2nd edition.

Diggle, P. J. and Rowlingson, B. S. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society, Series A*, 157(3):433–440.

Gatrell, A. C., Bailey, T., Diggle, P. J., and Rowlingson, B. S. (1996). Spatial point pattern analysis and its application in geographical epidemiology. *Trans Inst Br Geogr NS*, 21:256–274.

Jewell, C. P., Kypraios, T., Neal, P., and Roberts, G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 4(2):191–222.

Literature II

Macdonald, J. A. and Small, M. J. (2006). Assessing sites contaminated with unexploded ordnance: Statistical modeling of ordnance spatial distribution. *Environmental Science and Technology*, 40(3):931–938.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC.