HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Spatial Data Mining
Spatial modelling

Antti Leino ‹antti.leino@cs.helsinki.fi›

Department of Computer Science

Spatial dependence

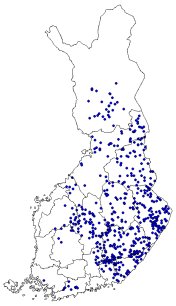
- *Everything is related to everything else, but nearby things are more related than distant things*
- This is usually true even for spatially discrete phenomena
  - Typically depend on underlying factors that are
    - numerous
    - not easy to measure
    - spatially continuous
  - In other words, spatial correlation is an approximation
  - Still, a useful one

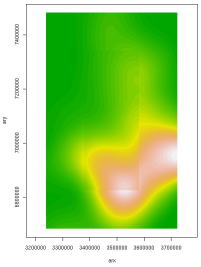
Different scales

- Sometimes useful to divide spatial dependence in two
- First order effects
  - Differences in intensity
  - Other large-scale variation
- Second order effects
  - Correlation between neighbouring places
  - Other small-scale variation

First order variation

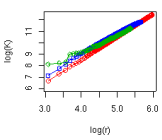
- Distribution of the name *Mustalampi* 'Black Pond'
- Kernel estimate of the intensity





Second order variation

- Again, the lake name *Mustalampi*
- *K* function
  - A measure for attraction between neighbouring instances
  - Red: theoretical value for no attraction
  - Blue: estimated value, constant intensity
  - Green: estimated value, variable intensity



First or second order effects?

- Same phenomenon can be modelled as either
  - Small-scale variation in intensity
  - Large-scale spatial autocorrelation
- In other words,
  - First order methods can be used for detailed study
  - Second order methods can be used at low resolutions
- Distinction between first and second order effects is largely a decision during modelling
  - Choice has to be based on the goals of the study

Dealing with space

- No (a priori) direction
  - Correlations in a two-dimensional space
  - Not reasonable to assume that correlation is directional
- Hence: no obvious definition for
  - neighbourhood in point patterns
  - proximity in area data
- Boundary effects
  - Observations do not typically cover all the phenomenon
  - In reality, correlation reaches to the unseen areas
  - This is not available for analysis

Background concepts

- Statistics commonly has certain methodological assumptions
  - Null hypothesis: the phenomenon is completely random
  - Goal: prove that the null hypothesis is invalid
  - Usually: phenomena follow the normal distribution
- What does this mean for spatial data?
  - Complete *spatial* randomness
  - Suitable probability distribution

Modelling spatial randomness

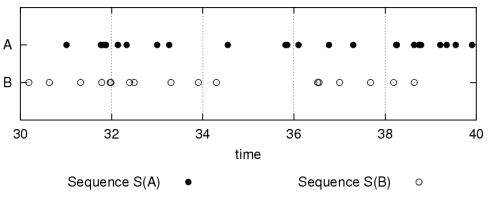
- Spatial stochastic process
- Statistical model for a spatial phenomenon
- Represented by the joint probability distribution of a set of random variables
  - $\{X(\mathbf{s}), \mathbf{s} \in \mathscr{R}\}$ for point data
  - $\{Y(\mathscr{A}), \mathscr{A} \subseteq \mathscr{R}\}$ for area data
- Normally only one realisation is observed
  - The actual values of the variable in each location

Modelling point patterns

- Randomness: the Poisson process
- Independent events happening with a constant intensity λ
- In its basic form one-dimensional
  - E.g. time
  - The probability of an event happening during an equal-sized time slot is uniform
- The expected number of events in a time slot $\mathrm{E}(X(t)) = \lambda t$

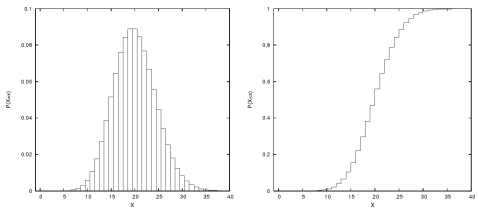
Poisson process: example

- Two time sequences generated from a Poisson process with $\lambda = 2$
  - A (•): 24 events
  - B (○): 17 events



Poisson process: example

- Probability distribution of the expected value of events
  - $\lambda = 2$, $t = 10$
  - $X(t) \sim \mathrm{Poisson}(20)$



Poisson process: from one to two dimensions

- Easy to extend the Poisson process to a two-dimensional case
- Again, constant intensity λ
- The expected number of events in region A depends on the intensity and the area of A: $\mathrm{E}(X(A)) = \lambda|A|$
- The spatial Poisson process is a model of what would happen if the events were independent from each other
  - No first order variation
  - No second order effects

First order variation: intensity

- Instead of constant intensity λ an intensity function

$$\lambda(s) = \lim_{|\mathrm{d}s|\to 0} \frac{\mathrm{E}(X(\mathrm{d}s))}{|\mathrm{d}s|}$$

  - $\mathrm{d}s$ a neighbourhood of point s
  - $\mathrm{E}(X(\mathrm{d}s))$ the expected number of points in this neighbourhood
  - $|\mathrm{d}s|$ the size of the neighbourhood
- The intensity at point s can be viewed as the »density» of events in an infinitely small neighbourhood of s

Using the intensity function

- A Poisson process can use the intensity function instead of a constant intensity
- Such a heterogeneous Poisson process models the first order variation of a point pattern
- The expected number of events in a region $\mathscr{A}$

$$\mathrm{E}(X(\mathscr{A})) = \int_{\mathscr{A}} \lambda(s)\mathrm{d}s$$

Estimating intensity

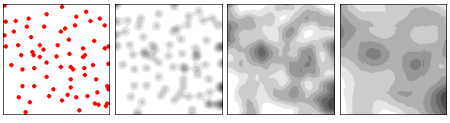
- Kernel estimation
- Represent each point by a symmetrical two-dimensional density function, e.g. normal distribution
- Estimate the intensity function as the sum of these density functions

$$\hat{\lambda}_\tau(s) = \frac{1}{\delta_\tau(s)} \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{s - s_i}{\tau}\right)$$

  - $s_1, \ldots, s_n$ event points
  - k kernel function
  - $\tau > 0$ bandwidth
  - $\delta_\tau(s)$ edge correction

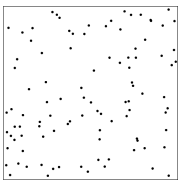
Kernel estimation

- Bandwidth defines how far from each point the effect reaches
- In effect, it specifies how detailed the variation in intensity is



Simulating a Poisson process

- Homogeneous Poisson process: two phases
  1. Number of events in area $\mathscr{A}$: $n \sim \mathrm{Poisson}(\lambda|\mathscr{A}|$
  2. The locations for the events can be obtained from a uniform distribution over $\mathscr{A}$



- Similarly for a heterogeneous Poisson process
  1. λ not constant
  2. Locations from a non-uniform distribution

Measuring second order effects

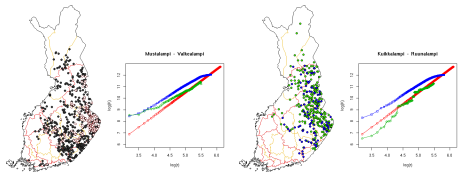
- Nearest neighbour measures
  - G(h): probability that the distance from a random event to the nearest other event $\leq h$
  - F(h): probability that the distance from a random location to the nearest event $\leq h$
- If events are clustered, $G(h) < F(h)$
- Only shows very small-scale attraction / repulsion
- Something else is required for scales larger than the nearest neighbour distance

K function

- Measure for second order effects
- Basic case: constant λ, one point pattern
  - $\lambda K(h)$ = expected number of other events within radius h of a random event
  - For a homogeneous Poisson process $K(h) = \pi h^2$
- Also possible to measure $K^{inhom}(h)$ for a heterogeneous point pattern
- For two point patterns
  - $\lambda_j K_{ij}(h)$ = expected number of events of type j within radius h of a random event of type i

K function: example

- Two pairs of lake names
  - *Mustalampi* 'Black Pond' – *Valkealampi* 'White Pond'
  - *Kuikkalampi* 'Diver Pond' – *Ruunalampi* 'Gelding Pond'
- Spatial distributions and K functions
  - Blue line: homogeneous K_{ij}
  - Green line: heterogeneous K_{ij}^{inhom}



Modelling second order variation

- Poisson cluster process
- Start with a Poisson process
  - Normally, a homogeneous process
  - In principle, heterogeneous also possible, but difficult to estimate
- This process generates »parents»
- Each parent generates a random number of »daughters»
  - Distributed independently around the parent
  - These are the actual events

Spatially continuous phenomena

- Observations from distinct points in space
- This time, measurements of a spatially continuous variable $\{Y(\mathbf{s}), \mathbf{s} \in \mathscr{R}\}$
- Goal: model the behaviour of Y across $\mathscr{R}$
- Again, useful to divide variation into first and second order effects

First order properties of continuous data

- Mean value surface $\{\mu(\mathbf{s}), \mathbf{s} \in \mathscr{R}\}, \mu(\mathbf{s}) = E(Y(\mathbf{s}))$
- Normal statistical regression problem
  - Linear regression of $Y(\mathbf{s})$ with spatial coordinates $\mathbf{s}_x, \mathbf{s}_y$
  - Trend surface analysis
  - More sophisticated methods available
- Goal: interpolate the value of Y between the observation points
  - $Y(\mathbf{s}) = \mu(\mathbf{s})$

Second order effects in continuous data

- Usually better to assume $Y(\mathbf{s}) = \mu(\mathbf{s}) + U(\mathbf{s})$
  - $\mu(\mathbf{s})$ global trend
  - $U(\mathbf{s})$ spatially correlated residual, with $\forall \mathbf{s} \in \mathscr{R} : \mathrm{E}(U(\mathbf{s})) = 0$
- $U(\mathbf{s})$ can be used to model second order effects
- Common assumption: $U(\mathbf{s})$ is stationary
  - $\mathrm{E}(U(\mathbf{s}))$ and $\mathrm{Var}(U(\mathbf{s}))$ constant
  - $Cov(U(\mathbf{s}), U(\mathbf{s}'))$ depends only on $\mathbf{h} = \mathbf{s}' - \mathbf{s}$
  - In other words, the same in different parts of $\mathscr{R}$
- Often also isotropic
  - $Cov(U(\mathbf{s}), U(\mathbf{s}'))$ depends only on $|\mathbf{h}|$
  - In other words, the same in all directions

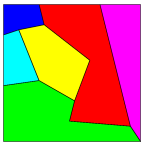
Predicting with second order effects

- If the residual process $\{U(\mathbf{s}), \mathbf{s} \in \mathscr{R}\}$ is spatially correlated, it is possible to give better estimates than $Y(\mathbf{s}) = \hat{\mu}(\mathbf{s})$
- Kriging: $\hat{Y}(\mathbf{s}) = \hat{\mu}(\mathbf{s}) + \hat{U}(\mathbf{s})$
- Various methods for this
  - Beyond the scope of this course
  - No general criterion for choosing, beyond »see what works»
- Bottom line: modelling both first and second order effects gives reasonably good predictions

Proximity in area data

- Proximity matrix W

$$w_{ij} = \begin{cases} 1 \text{ if } \mathscr{A}_i \text{ and } \mathscr{A}_j \text{ share a border} \\ 0 \text{ otherwise} \end{cases}$$



	A	B	C	D	E	F
A	0	1	0	1	1	0
B	1	0	1	0	1	1
C	0	1	0	0	0	1
D	1	0	0	0	1	1
E	1	1	0	1	0	1
F	0	1	1	1	1	0

- More elaborate measures for proximity possible

First order variation

- Simple option: moving averages
  - Replace the value for each area by the averages of its neighbours

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} y_j}{\sum_{j=1}^{n} w_{ij}}$$

- Convert to point data
  - E.g. represent each area by its centre
  - Perform kernel estimation
- Median polish
  - For regular grids
  - Represent each grid cell as

$$y_{ij} = \mu + r_i + c_j + \varepsilon_{ij}$$

  - r_i, c_j row and column trends, ε_{ij} random error

Second order effects

- Moran's I statistic: spatial correlation

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) \left(\sum\sum_{i \neq j} w_{ij}\right)}$$

  - Varies between -1 and $+1$, no autocorrelation when $I = 0$
- Geary's C statistic: variance of the difference of neighbouring values

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - y_j)^2}{2 \left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) \left(\sum\sum_{i \neq j} w_{ij}\right)}$$

  - Varies between 0 and 2, no autocorrelation when $C = 1$

Summary

- Lots of statistical methods for spatial modelling
- Different methods for point patterns, area data and continuous data
  - Some related to each other
- If still interested, take a course in spatial statistics