Point pattern data

ST733 – Spatial Statistics

Point Pattern Data

- Here the response is not the value at a location, it is the location itself
- Examples:
- Common objectives:

Outline

- Definitions and notation
- Testing for a completely random sample
- Point process models
- Hot spot detection

Notation and definitions

- Let $\mathbf{s}_i \in \mathcal{R}^2$ be the location of the i^{th} observation
- We are interested only in observations that fall in the domain of interest $\mathcal{D} \subset \mathcal{R}^2$
- For arbitrary region $\mathcal{B} \subset \mathcal{D}$, let $N(\mathcal{B})$ be the number of observations in $\mathcal{B}$ and $|\mathcal{B}|$ be the size of $\mathcal{B}$
- The covariate vector $\mathbf{X}(\mathbf{s}) = [X_1(\mathbf{s}), ..., X_p(\mathbf{s})]^T$ is assumed to be available for any $\mathbf{s} \in \mathcal{D}$

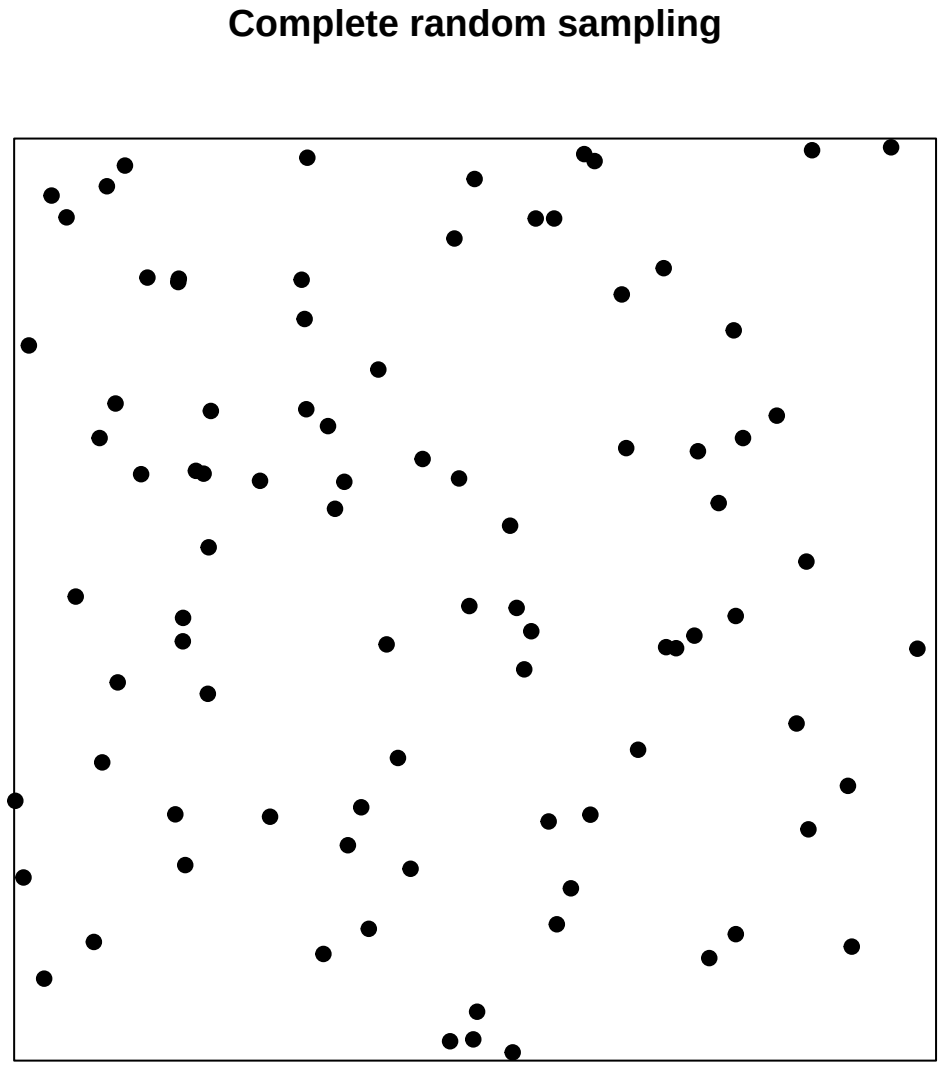
Notation and definitions

- A spatial point pattern is
- A multivariate spatial point pattern is
- A marked spatial point pattern is

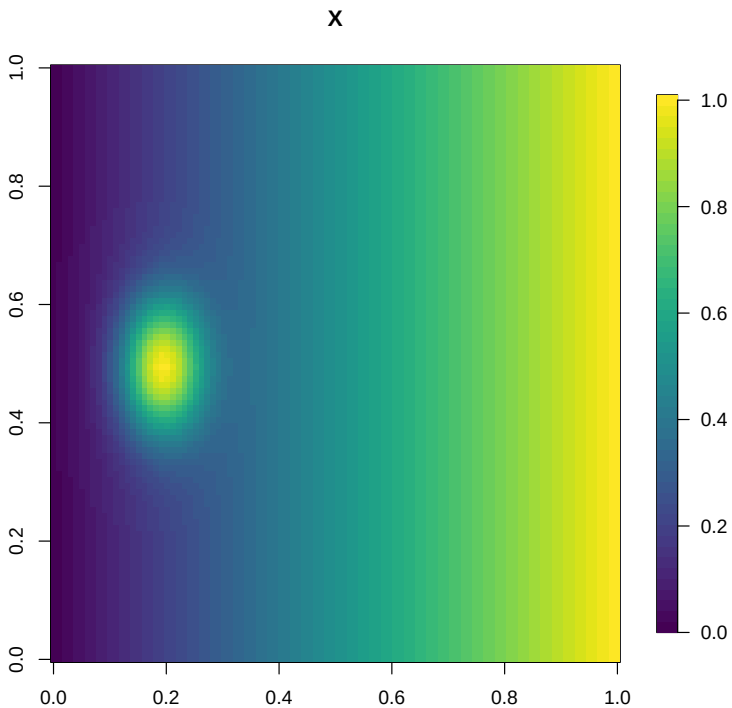
Classifications of spatial point patterns

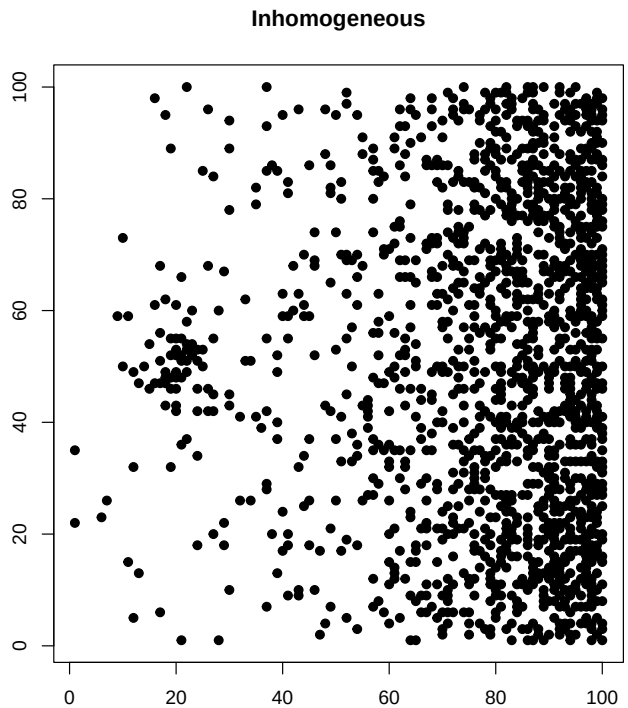
- Completely random:
- Inhomogeneous:
- Clustered:
- Regular:

SPP random sample

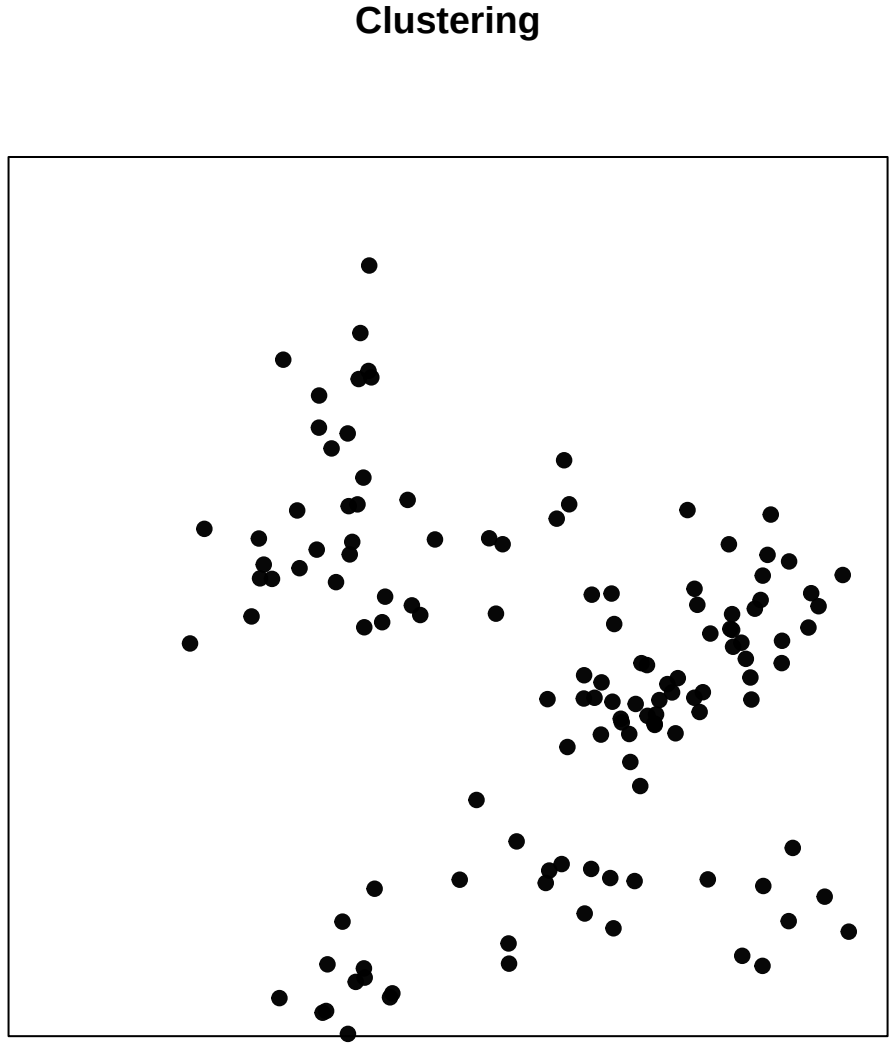


SPP random sample

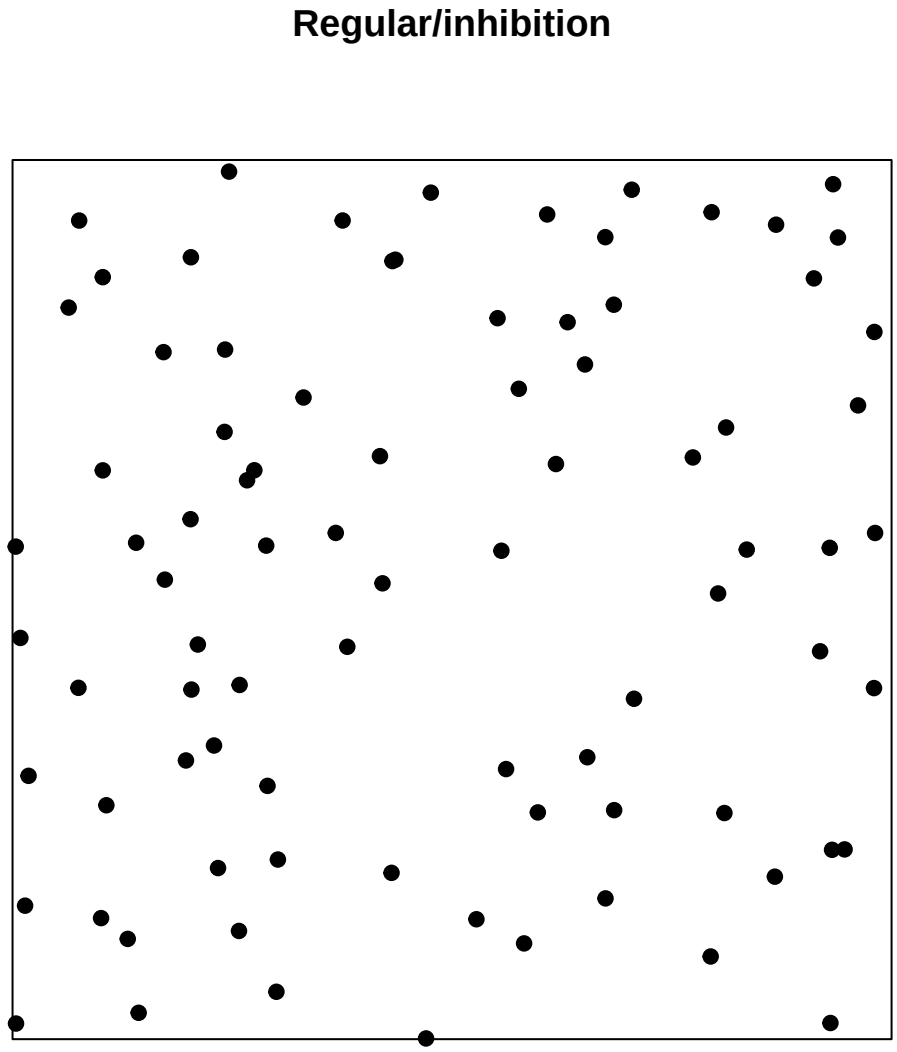




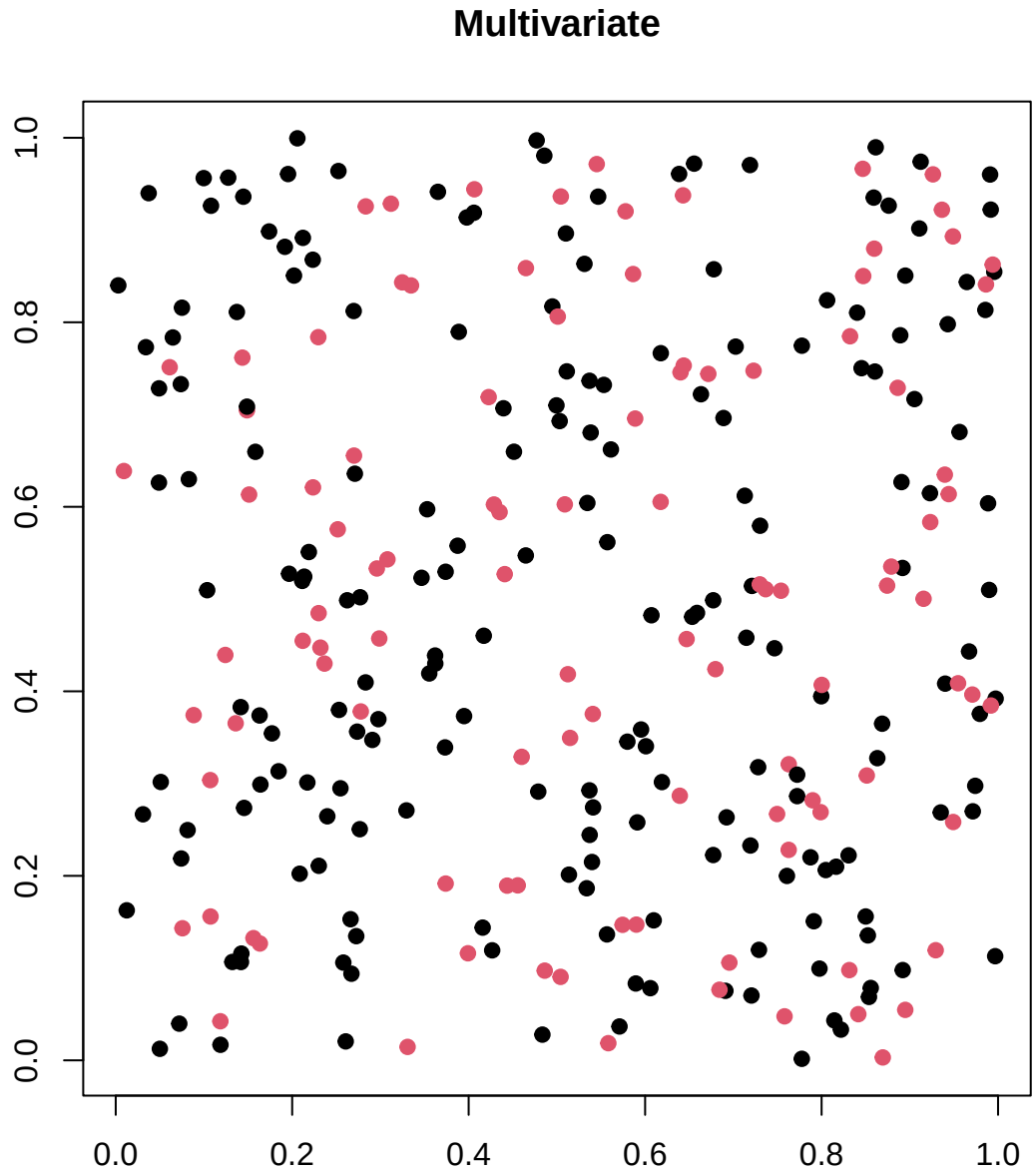
SPP random sample



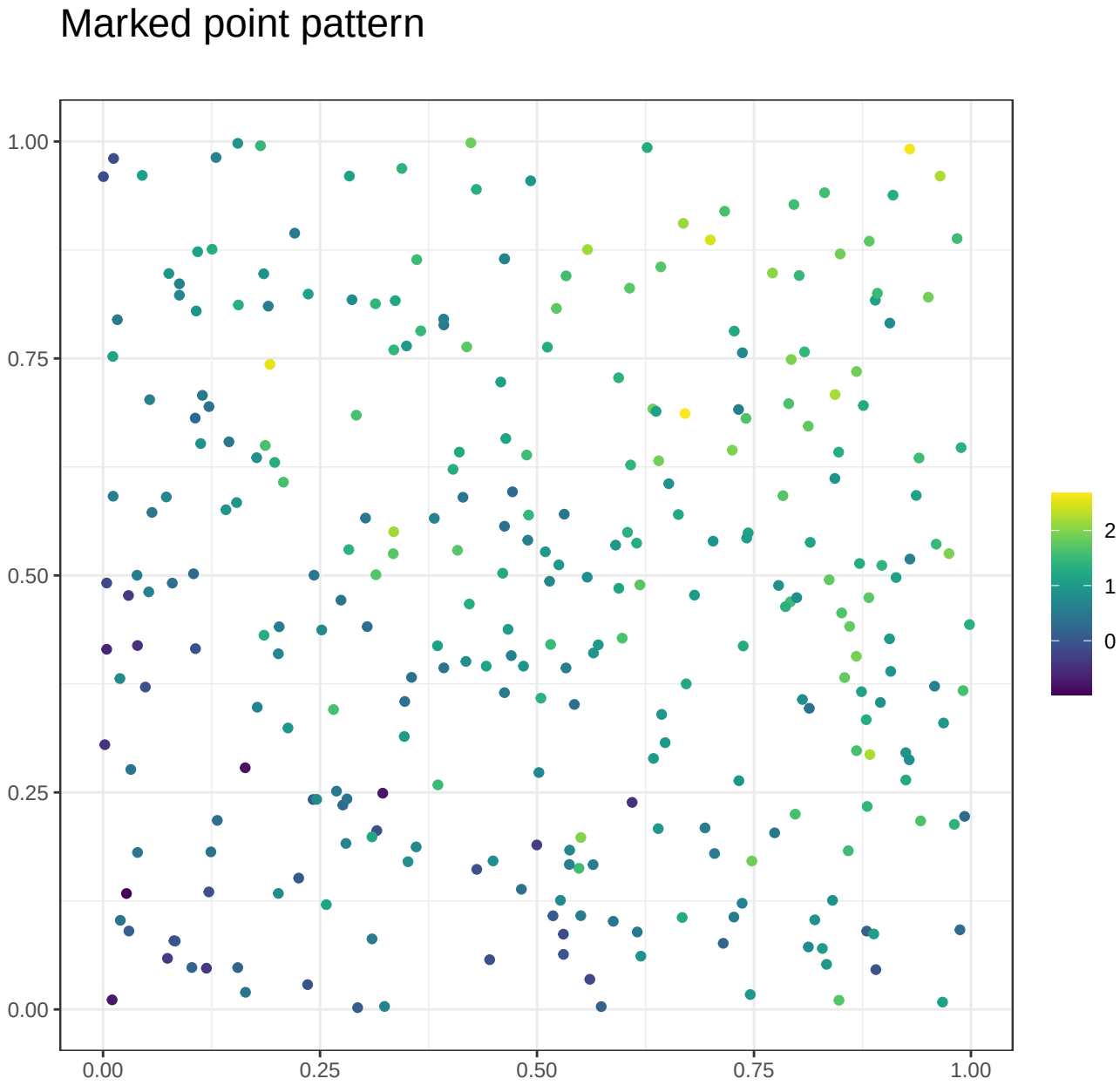
SPP random sample



SPP random sample



SPP random sample



Simple tests for completely random sampling

- **Quadrat**: Split the domain into m subregions and do a χ^2 test that the mean is the same in each subregion

- **Clark-Evans**: For each point, compute the distance to the nearest neighbor and do a test that this mean equals the value under the null

Ripley's K function

- Ripley's K function is used for more than just testing if the sample is completely random, it is analogous to the variogram in geostatistics
- Let $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ be the distance between samples i and j
- Ripley's empirical K function is

Ripley's K function

- Under completely random sampling

- Therefore, it is common to plot

Ripley's K function

- Under completely random sampling, clustering and inhibition we expect

- Test of completely random sample

Models - Definitions

- Let $d\mathbf{s}$ be a small region around $\mathbf{s}$, $d\mathbf{s} = \{\mathbf{t}; ||\mathbf{t} - \mathbf{s}|| < \epsilon\}$
- First-order intensity (related to the mean):

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

- Second-order intensity (related to the covariance):

$$\lambda_2(\mathbf{s}, \mathbf{t}) = \lim_{|d\mathbf{s}|, |d\mathbf{t}| \to 0} \frac{E[N(d\mathbf{s})N(d\mathbf{t})]}{|d\mathbf{s}||d\mathbf{t}|}$$

Models - Definitions

- Mean function: $\mu(\mathcal{B}) = \mathrm{E}[N(\mathcal{B})]$

- First-degree orderliness:

$$\lim_{|d\mathbf{s}|\to 0} \frac{\mathrm{Prob}(N(d\mathbf{s}) > 1)}{|d\mathbf{s}|} = 0$$

- Second-degree orderliness:

$$\lim_{|d\mathbf{s}|,|d\mathbf{t}|\to 0} \frac{\mathrm{Prob}(N(d\mathbf{s}) > 1, N(d\mathbf{t}) > 1)}{|d\mathbf{s}||d\mathbf{t}|} = 0$$

Models - Definitions

- Stationarity: Probability statements about $N(\mathcal{B})$ are invariant to shifts in $\mathcal{B}$

- Stationarity implies:
  - $\lambda(\mathbf{s}) = \lambda$ for all $\mathbf{s}$
  - $\lambda_2(\mathbf{s}, \mathbf{t}) = \lambda_2(\mathbf{s} - \mathbf{t})$ for all $\mathbf{s}$ and $\mathbf{t}$

Models - Poisson process (PP)

- Like the GP for geostatistics, the PP is the fundamental construct of point pattern analysis
- A spatial point pattern is a Poisson process on $\mathcal{D}$ if
- These conditions imply $N(\mathcal{B}) \sim \text{Poisson}(\mu(\mathcal{B}))$ for any $\mathcal{B}$

Models - Homogeneous Poisson process

- This is the simplest case with $\lambda(\mathbf{s}) = \lambda$ for all $\mathbf{s} \in \mathcal{D}$
- This implies $\mu(\mathcal{B}) = \lambda|\mathcal{B}|$ and $N(\mathcal{B}) \sim \text{Poisson}(\lambda|\mathcal{B}|)$
- And equivalent representation is
- The process is stationary and sampled completely at random
- The sufficient statistic for the single parameter is n and $\hat{\lambda} = n/|\mathcal{D}|$ is the MLE

Models - inhomogeneous Poisson process (IPP)

- An IPP is non-stationary with spatially-varying intensity $\lambda(\mathbf{s}) \geq 0$
- For an IPP, $N(\mathcal{B}) \sim \text{Poisson}(\mu(\mathcal{B}))$ where the mean is the integrated intensity
- Given the number of events n, $s_1, ..., s_n \overset{iid}{\sim} f(\mathbf{s})$ where

Models - inhomogeneous Poisson process (IPP)

- Given the intensity function, the locations are independent, i.e., there is no clustering or inhibition

- However, it can be very difficult to distinguish between an IPP and a cluster (or inhibition) process

Models - thinned Poisson process

- Say $\lambda(\mathbf{s}) = Z(\mathbf{s})\lambda$ where $Z(\mathbf{s}) \in [0, 1]$ is the thinning process

- We can generate an IPP as a thinned homogeneous Poisson process:

Models - Connection with areal data

- Say that $\mathbf{s}_1, ..., \mathbf{s}_n$ are the event locations
- For privacy concerns you are only given the county that includes each event
- Let $\mathcal{B}_i$ be the region defining county i and $Y_i = N(\mathcal{B}_i)$ be the number of events in county i
- If the locations follow an IPP with intensity $\lambda(\mathbf{s})$ then the areal data are distributed as
- How would you analyze these data?

Models - IPP with spatial covariates

- As in Poisson regression, covariates can be related to the log intensity function

$$\log[\lambda(\mathbf{s})] = \beta_0 + \mathbf{X}(\mathbf{s})\boldsymbol{\beta}$$

- The interpretation is the coefficient is the log relative intensity

Models - IPP with spatial covariates

- The likelihood (given n) is

- The denominator is tough to compute

Models - IPP with spatial covariates

- If there is one binary covariate we can compute the maximum likelihood in closed form
- Say n_x is the number of samples with $X(\mathbf{s}) = x$ and $n = n_0 + n_1$
- Define A_x as the area with $X(\mathbf{s}) = x$ so $|\mathcal{D}| = A_0 + A_1$
- The maximum likelihood estimator is:

Models - IPP with spatial covariates

- One approach is to approximate the integral with a sum
- Let $\mathcal{B}_1, ..., \mathcal{B}_m$ be a partition of $\mathcal{D}$ and assume that $X(\mathbf{s}) = X_j$ for all $\mathbf{s} \in \mathcal{B}_j$
- Then the likelihood is approximately
- MLE or Bayes proceeds as usual.[1]

[1] In economics this is called a discrete choice model and in ecology this is called a habitat selection process. In general it is a multinomial logistic regression model.

Models - Clustering

- The most intuitive clustering model is the Poisson cluster process
- Let $\mathbf{v}_1, ..., \mathbf{v}_m$ follow a homogeneous Poisson process, and
- Can be fit using the EM algorithm or MCMC (it's a finite mixture of normals)

Models - Log-Gaussian Cox process

- Another way to induce clustering in to model the intensity using a Gaussian process
- The Cox process takes $\theta \sim$ GP and

$$\log[\lambda(\mathbf{s})] = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \theta(\mathbf{s})$$

- The likelihood is really difficult to deal with:
- Grid approximations and a basis expansion approximation for θ make estimation possible

Model - Inhibition

- The Strauss model is
- The full conditional distribution is
- The role of the parameters

Model - Inhibition

- Model fitting is hard
- Pseudolikelihood (i.e., the product of full conditional distributions) is the most common approach
- Data generation is even hard, often resorting to MCMC

Hot spot detection using the scan statistic

- This is another way to test for a completely random sample, but the alternative is that there is a “hot spot”
- For example, there might be a small region with a higher rate of nausea than the rest of the domain, and this could be used to detect food poisoning
- The hypotheses to be tested are

Hot spot detection using the scan statistic

- The likelihood ratio statistic is

- Searching over all possible regions is impossible, so we usually restrict to a simple class, like circles

Hot spot detection using the scan statistic

- Approximating the optimal test statistic by scanning over a grid of centers and radii gives the scan statistic

- The p-value comes from:

Preferential sampling

- Spatial point pattern models are used to adjust for preferential sampling