Spatial point pattern data - Part I

Applied Spatial Statistics

Spatial point pattern data

- The response is a spatial location, $\mathbf{s}_i = (s_{i1}, s_{i2})$
- Example: $\mathbf{s}_i$ is the location of a hurricane landfall
- Example: $\mathbf{s}_i$ is the location of cancer diagnosis
- Example: $\mathbf{s}_i$ is the location of a homicide
- Example: $\mathbf{s}_i$ is the location of a sighting of an endangered species

Spatial point pattern data

- Analysis of point pattern data is fundamentally different than point-referenced or areal data
- For example, it does not make sense to model $\mathbf{s}_i$ as Gaussian or Poisson
- We need completely new terminology and methods

Objectives

- Estimate the response density, i.e., the PDF of $\mathbf{s}_i$
- Test if locations arise as a completely random sample
- Test/model interactions (clustering/repulsion) of samples
- Estimate the effect of covariates on the response density

Outline

- Notation and terminology
- Ripley's K function, which is analogous to the variogram
- Tests for a completely random sample
- Statistical models for spatial point patterns

Notation and terminology

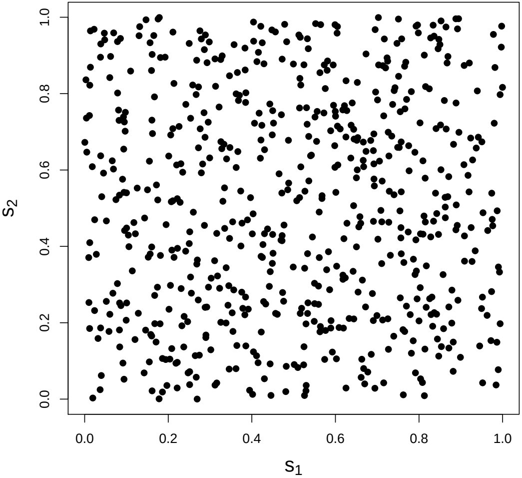
- The response for observation i is $\mathbf{s}_i = (s_{i1}, s_{i2})$ where s_{i1} is longitude and s_{i2} is latitude
- We are interested only in events in a sampling window, $\mathcal{D} \in \mathcal{R}^2$, e.g., $\mathcal{D}$ is North Carolina
- Later we will use covariates $\mathbf{X}(\mathbf{s}) = \{X_1(\mathbf{s}), ..., X_p(\mathbf{s})\}$, e.g., elevation at $\mathbf{s}$
- The spatial point pattern is $\mathbf{s}_1, ..., \mathbf{s}_n$ where n and $\mathbf{s}_i$ are random

Notation and terminology

- A **multivariate point pattern** has multiple types of events
- Example: $\mathbf{s}_1, ..., \mathbf{s}_n$ are locations of pine trees and $\mathbf{t}_1, ..., \mathbf{t}_m$ are locations of oak trees
- A **marked point pattern** includes a measurement (mark) taken at each response
- Example: $\mathbf{s}_1, ..., \mathbf{s}_n$ are locations of pine trees and $Y_1, ..., Y_n$ are their heights
- We will not address either of these types of point patterns

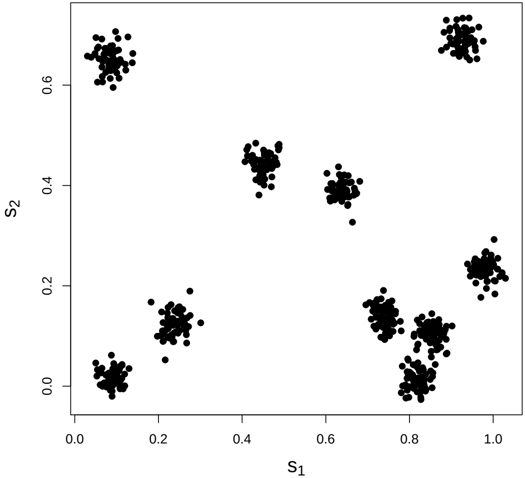
Classifications of spatial point patterns

In a **completely random sample** (CRS) the locations are independent and uniformly distributed



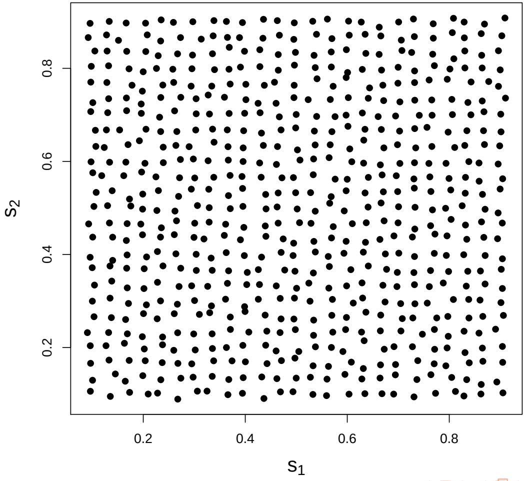
Classifications of spatial point patterns

In a **clustered** sample the locations are attracted to each other



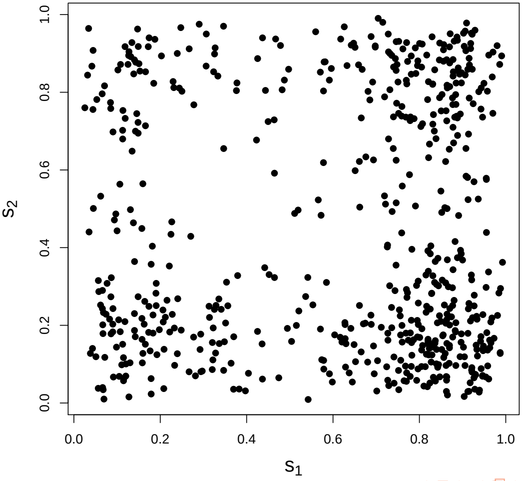
Classifications of spatial point patterns

In a **regular** (aka, repulsion or inhibition) sample the locations are repulsed by each other



Classifications of spatial point patterns

In an **inhomogeneous** sample the density of locations varies with space (hard to distinguish from clustered)



Ripley's K function

- Clustering and inhibition are two types of interactions between locations
- They are analogous to positive and negative correlation
- Ripley's K function is a graphical tool (like the variogram) used to study interactions
- It guides model building
- It also provides a way to test for interactions

Ripley's K function

- Let $d_{ij} = ||\mathbf{s}_i - \mathbf{s}_j||$ be the distance between $\mathbf{s}_i$ and $\mathbf{s}_j$
- Denote the sample proportion of the pairs of sites within t of each other as

$$p(t) = \frac{1}{n^2} \sum_{i \neq j} I(d_{ij} < t),$$

where $I(d < t) = 1$ if $d < t$ and $I(d < t) = 0$ if $d \geq t$

- Ripley's K function at distance t is

$$K(t) = |\mathcal{D}| p(t)$$

where $|\mathcal{D}|$ is the area of the sampling window

Ripley's K function

- Under CRS, the expected value is $\mathrm{E}\{K(t)\} = \pi t^2$
- Clustering at distance t gives $K(t) > \pi t^2$
- Inhibition at distance t gives $K(t) < \pi t^2$

Hypothetical K functions