SEEC stats toolbox seminar series:

Basics of Point Pattern Analysis

Mzabalazo Z. Ngwenya

Centre for Statistics in Ecology, Environment and Conservation (SEEC)

SEEC - Statistics in Ecology, Environment and Conservation

Department of Statistical Sciences
University of Cape Town

Outline

1. Introduction
   - Statistical formulation
   - Types of point patterns
   - Common scientific questions
2. Characterizing Point Processes
   - First-order intensity
   - Second-order intensity
3. A simple example of point pattern analysis in R
   - Exploratory data analysis
   - Modeling and inference

Preamble

- The classical methods of point pattern analysis largely developed outside the field of statistics.
- These methods were developed for relatively small data sets and rest heavily on the assumption of stationarity of the point process though stationarity is often an unreasonable assumption.
- The classical approach also relies heavily on non-parametric summary statistics for making inferences.
- Developments in computing coupled with advances in computational statistics have moved spatial point analysis away from non-parametric methods and assumptions of stationarity to flexible likelihood based methods where the assumption of stationarity can be relaxed.
- This modern approach to spatial point pattern analysis allows for more statistical rigor.

1.1 Statistical formulation

Point process

An unordered countable set of locations $\mathbf{s} = \{s_1, \ldots, s_n\}$ in a defined d-dimensional study region $W \in \mathbb{R}^d$ at which certain events have been recorded.

Assumptions:

- The point process $\mathbf{S}$ extends throughout $\mathbb{R}^d$ but is only observed inside W.
- W the sampling window is fixed and known.
- Number of points is random.
- Locations of points is random.

- Simple point pattern - have locations of events only

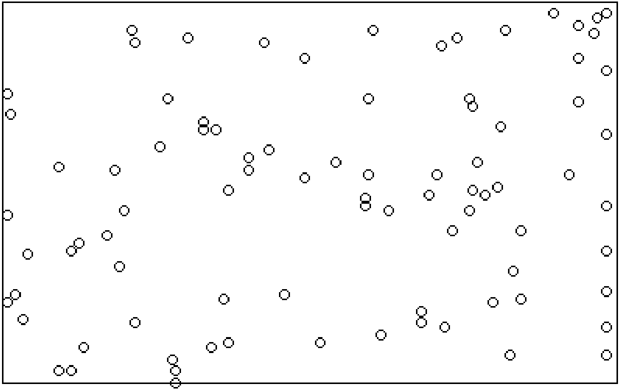


Figure 1: Locations of Japanese black pine saplings in a square sampling region in a natural forest.

- Marked point pattern - have location and magnitude of events

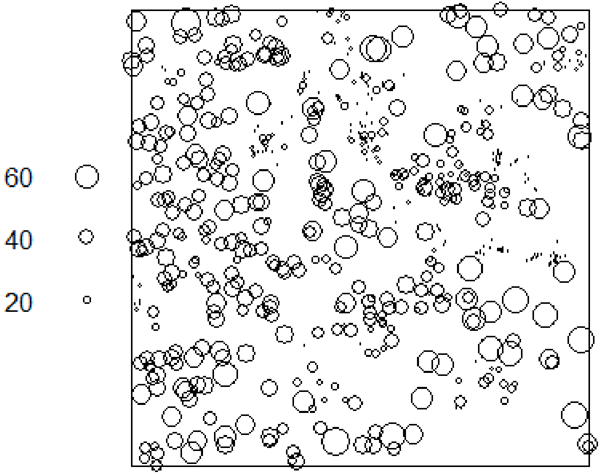


Figure 2: Locations of Longleaf pine trees shown using circles proportional to tree diameters.

- Multi-type point pattern - consist of locations of different types of events

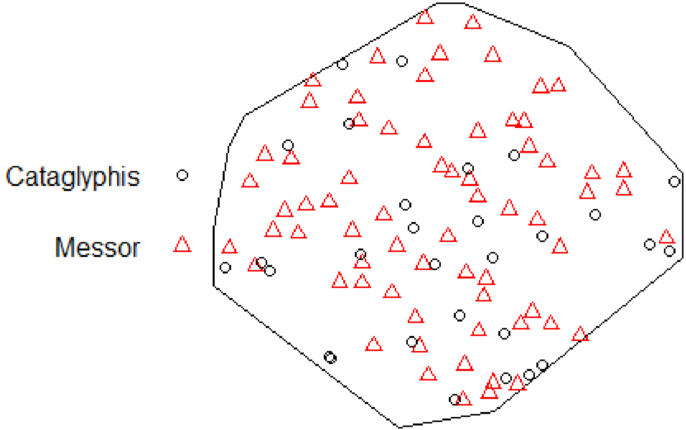


Figure 3: Location of the nests of two ant species.

1.3 Common scientific questions

1. Is the point pattern random?
2. What is the spatial density of points?
3. What are the inter-point relationships?

In order to answer these questions one needs to characterize the process which necessitates the determination of the **first-order** and **second-order** intensities of the process.

2.1 First-order intensity

- First order intensity, $\lambda(\mathbf{s})$ relates to the intensity of the process in space, that is, the spatial intensity.
- The intensity maybe either homogenous over W or inhomogeneous
- $\lambda(\mathbf{s})$ can be either estimated via kernel or likelihood methods.
- Kernel estimates of are frequently useful when no information of the mechanisms driving the spatial intensity exist.
- Likelihood methods for the estimation of $\lambda(\mathbf{s})$ are often used in situations where some knowledge on the form of the trend or mechanisms driving the process exist.

Kernel estimation of $\lambda(\mathbf{s})$:

- In kernel estimation of intensity the Gaussian kernel $exp(||s||^2/2)$ is predominantly used.
- If the point process is assumed to be stationery with homogenous intensity a fixed bandwidth is chosen.
- If the intensity is inhomogeneous the kernel estimator may be made adaptive by using various bandwidth values.

Likelihood estimation of $\lambda(\mathbf{s})$:

- If information about the process exists a parametric point process models is first fitted giving $\lambda(\mathbf{s}) = exp(\mathbf{F}\beta)$, this is the likelihood approach.
- The following classes of point process models are usually assumed depending on what is known/suspected of the process;

Second-order property	Point process model	**spatstat** function
CSR	Poisson process	ppm(data $\sim$ 1)
Inhibition	Gibbs	ppm(data $\sim$ trend , interaction = " ")
Aggregation	Inhomogeneous Poisson Cox/Inhomogenous cluster Gibbs	ppm(data $\sim$ trend) kppm(data $\sim$ trend, clusters = " ") ppm(data $\sim$ trend, interaction = " ")

Interaction specification:

- For Gibbs models modeling inhibition, the inter-point interaction should be models using the area-interaction model or the Geyer saturation model.
- For Gibbs models modeling inhibition all other interaction models specify a clustering process (see ?ppm)
- For Cox cluster processes the `clusters` argument should be one of `"Thomas"` (default), `"Cauchy"`, `"VarGamma"`, `"LGCP"`

2.2 Second-order intensity

- Second-order describes inter-point relationships i.e spatial randomness, regularity and aggregation.
- Second order properties of a point process can be inferred using one of three approaches
  1. Quadrant methods - (highly subjective)
  2. Fitted models - (if likelihood approach to model fitting is taken parameters can be estimated objectively following probability theory)
  3. Calculation of theoretical envelopes of summary statistics using simulation (classical approach)

- The summary statistics that are commonly used with the simulation approach are

1. Ripley's K-function, $K(r)$[1];

   For typical point of the process $K(r)$ computes the expected number of other points of the process within a distance r of that point.

2. Pair-correlation function, $g(r)$[2];

   For any set of points separated by distance r values of $g(r) = 1$ indicate Complete Spatial Randomness (CSR), $g(r) > 1$ indicates clustering of points at that distance and values of $g(r) < 1$ indicate inhibition or regularity

[1] `Kest()`

[2] `pcf()`

- These functions are usually computed for various values of r which are then plotted.
- The analogs of the $K(r)$ and $g(r)$ functions for inhomogeneous processes are respectively
  - The L-function a.k.a inhomogeneous K-function; `Kinhom()`, `Lest()`
  - The inhomogeneous pair correlation function; `pcfinhom()`
- To perform inference using these summary statistics simulation envelopes[3] corresponding to Poisson Process i.e. CSR are computed for that statistic.

  The statistic is also then computed for the observed data which is the compared to the simulation envelope.

[3] `envelope()`

- For statistical rigor the following approach is taken in the analysis and modeling process;

  1. Firstly non-parametric methods and the simulation based approach to inference are used for exploratory data analysis (EDA).
  2. Using the insight gained from the EDA parametric point process models are then fitted and the best model is selected, evaluated and then subsequently used for inference.

3.1 Exploratory data analysis

- We use the non-parametric methods for investigating the first-order intensity
- Preliminary inference on the second-order characteristics of the process are drawn using summary statistics in conjunction with simulation.

Characteristic	R code
first-order intensity	density.ppp(nztrees)
second-order intensity	plot(envelope(nztrees, fun = Kinhom, correction ="best", global = TRUE), main = NULL)

3.1.1 Non-parametric estimation of the first order intensity

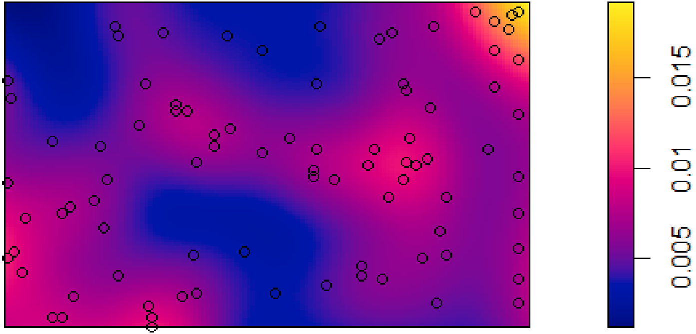


Figure 4: Fixed Kernel estimate of spatial intensity.

3.1.2 Non-parametric statistic for second-order intensity

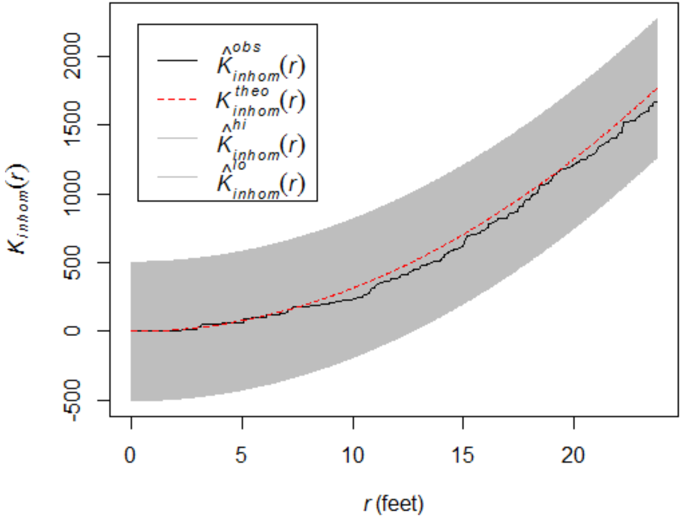


Figure 5: K inhomogeneous function and associated quantities

- It appears as if the point process is inhomogeneous
- Though inhomogeneous it appears that we have a Poisson point process i.e. an inhomogeneous Poisson point process.

3.2.1 Model fitting

- Using the insights gathered from the EDA with the non-parametric methods we are now able to specify and fit various parametric models.
- Four models were fitted
  - Homogenous Poisson model - (the null model)
  - Inhomogeneous Poisson (I.P) model with linear trend (L.T.)
  - Inhomogeneous Poisson (I.P) model with quadratic trend (Q.T.)
  - Cox model with Neyman-Scott clustering mechanism specified by a Cauchy kernel; linear trend (L.T.)

Model	Description	R code
0	Poisson	ppm(nztrees ∼ 1)
1	I.P.- L.T	ppm(nztrees ∼ x+y)
2	I.P. - Q.T.	ppm(nztrees ∼ polynom(x,y,2))
3	Cox - L.T.	kppm(nztrees ∼ x+y, cluster = "Cauchy")

Model 1:

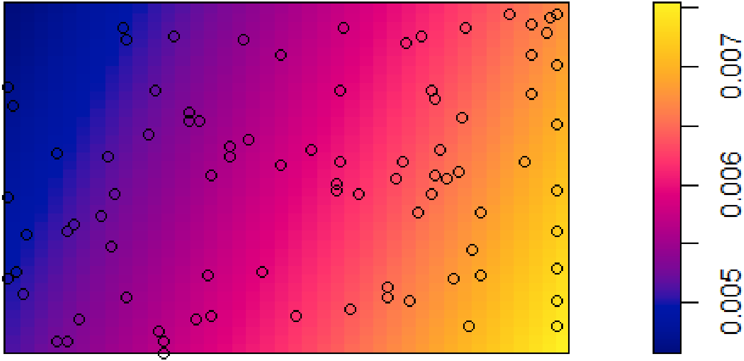


Figure 6: Spatial intensity estimate for fitted inhomogeneous Poisson model with linear trend.

Model 2:

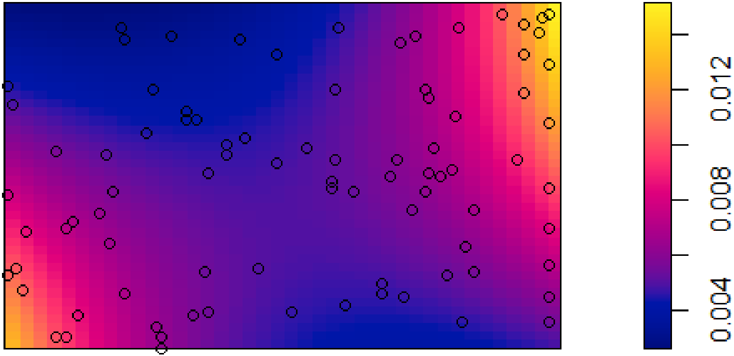


Figure 7: Spatial intensity estimate for fitted inhomogeneous Poisson model with quadratic trend.

Model 3:

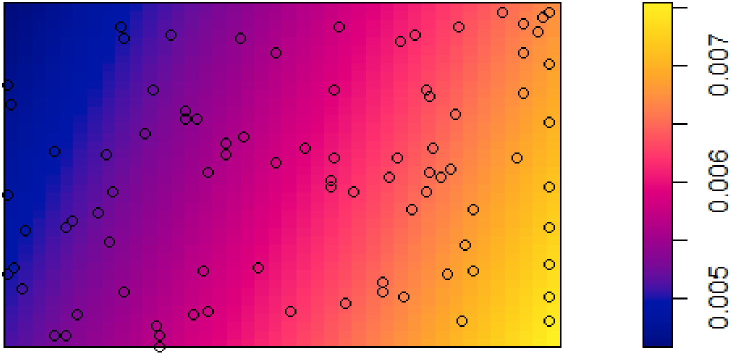


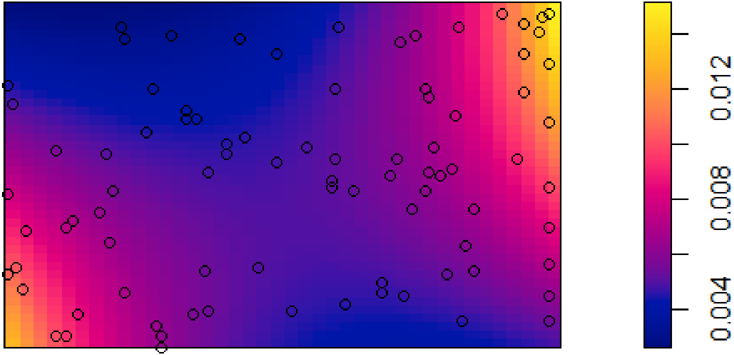
Figure 8: Cox model with Neyman-Scott clustering mechanism specified by a Cauchy kernel; linear trend.

3.2.2 Model evaluation and diagnostics

```
AIC(fitted.model)
```

Model	Description	AIC
0	Poisson	1056.354
1	Inhomogeneous Poisson - linear trend	1059.137
2	Inhomogeneous Poisson - quadratic trend	1056.238
3	Cox - linear trend	2455.198

Model 2:



```
diagnose.ppm(poisson.mdl.2, which = "smooth")
```

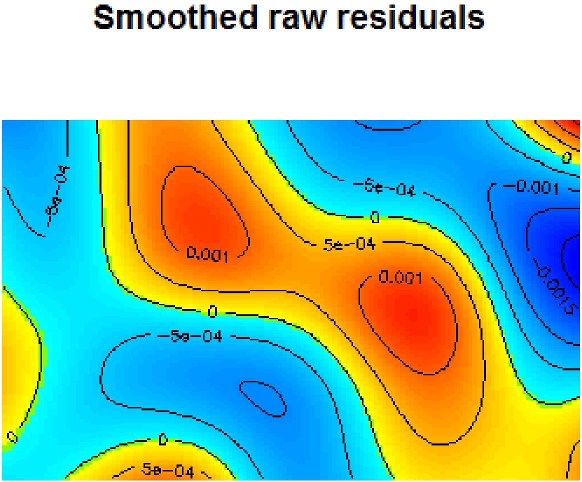
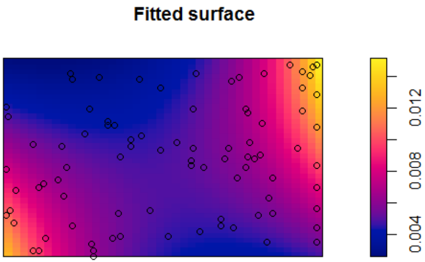
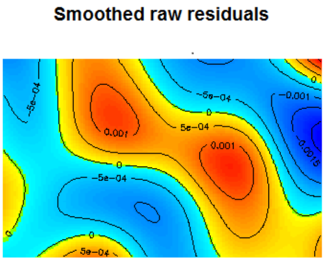


Figure 9: Residuals from the best fitting model.






```
diagnose.ppm(poisson.mdl.2, which = "smooth")
Model diagnostics (raw residuals)
Diagnostics available:
smoothed residual field
range of smoothed field =  [-0.002166, 0.001624]
```

3.2.3 Inference

- We use the best fitting model to make inference

```
> print(poisson.mdl.2)
Nonstationary Poisson process

Log intensity:  ~x + y + I(x^2) + I(x * y) + I(y^2)

Fitted trend coefficients:
(Intercept)             x             y        I(x^2)
-4.2948849770 -0.0222001974 -0.0089257067  0.0001007236
I(x * y)        I(y^2)
0.0001955263 -0.0000881984
```

- From the fitted model we can infer that
  - The point process is a Nonstationary Poisson process
  - The (estimated) first order intensity (spatial trend/ systematic component of the point process) is

$$\hat{\lambda}(\mathbf{s}) = exp(\mathbf{F}\hat{\beta}) = exp(-4.295 - 0.0222x - 0.009y + 0.0001^2 + 0.0001x * y - 0.00008y^2)$$

- Lets look at the scenario where there is clustering; we interpret the Inhomogeneous cluster process model, the Cox model

```
> print(cox.mdl)
Inhomogeneous cluster point process model
Fitted to point pattern dataset 'nztrees'
Fitted by maximum Palm likelihood
rmax = 23.75
weight function: Indicator(distance <= 11.875)

Log intensity:  ~x + y

Fitted trend coefficients:
(Intercept)            x            y
-5.277292606  0.002590988 -0.001221149

Cluster model: Cauchy process
Fitted cluster parameters:
kappa      scale
215.026320   7.291613
```

- We can obtain an estimate of the first-order intensity as before; $\hat{\lambda}(\mathbf{s}) = exp(\mathbf{F}\hat{\beta}) = exp(-5.277 + 0.003x - 0.001y)$
- We can also obtain parameters on the clustering process, the stochastic component of the point process, these are $\kappa = 215.026$ and $r = 7.292$.
    - This means the clustering process has an intensity of 215.026
    - The scale parameter, r, indicates the average radius of each of the clusters
- Some models have the additional parameter ν. This provides an estimate of the average number of offspring for each parent point. (therefore $\nu + 1$ will give you average number of points for each cluster)

Bibliography

Baddeley, A. and Turner, R. (2005).
Spatstat: an R package for analyzing spatial point patterns.
Journal of Statistical Software, **12**, 1-42.

Cressie, N. A. C. (1993)
Statistics for Spatial Data (revised edition), John Wiley and Sons, New York.

Diggle, P.J. (2003).
Statistical Analysis of Spatial Point Patterns(second edition), Hodder Arnold, London. edition, 2003.

J., Penttinen A., Stoyan H. and Stoyan D.
Statistical Analysis and Modelling of Spatial Point Patterns, John Wiley and Sons, Chichester.

Møller J. and Waagepetersen R.P. (2008)
Modern spatial point process modelling and inference (with discussion).
Scandinavian Journal of Statistics, **34**, 643-711