TMA4250 Spatial Statistics
Project 2: Point pattern data and point processes

Spring 2023

Introduction

This project contains problems related to point processes. Relevant functions may be found in the `R` library `spatial`.

You will work with three real-world point pattern datasets from the `R` package `MASS`:

- biological cell data, available in `cells.dat`
- redwood tree data, available in `redwood.dat`
- pine tree data, available in `pines.dat`

Additionally, Problem 2 uses the files:

- detection probabilities, available in `obsprob.txt`
- pine tree counts, available in `obspines.txt`

Problem 1: Analysis of point pattern data

For each subproblem consider all three real-world point pattern datasets.

a) Display each point pattern and discuss its appearance. Relate the observed appearance and behaviour to real processes in nature.

b) Compute the empirical L-function for each of the point patterns, the function `Kfn` can be used. *Read help file (`?Kfn`) to understand exactly what the function computes.* Display the L-functions for each point pattern and discuss their appearance.

Give the expression for the theoretical L-function for a homogeneous Poisson point process. Display the empirical L-function for each of the point patterns together with the

corresponding theoretical L-function for a homogeneous Poisson point process, and discuss whether a homogeneous Poisson RF appears to be a suitable model for each of the point patterns.

c) For each dataset, assume that the point pattern arose from a homogeneous Poisson point process. Generate 100 realizations of the homogeneous Poisson point process conditional on the number of points being equal to the observed number of points. For each realization, compute the associated L-function. Display 90% (pointwise) prediction intervals for the empirical L-function (based on 100 realizations) jointly with the corresponding empirical L-function for each point pattern. Discuss, for each point pattern, whether a homogenous Poisson point process appears to be a reasonable model.

Problem 2: Remote sensing of trees

Consider a $300\,\mathrm{m} \times 300\,\mathrm{m}$ observation window in a pine tree forest. The locations of pine trees in the observation window are remotely sensed by a satellite. The data provided by the satellite is counts of pine trees inside $10\,\mathrm{m} \times 10\,\mathrm{m}$ grid cells for a regular 30×30 grid of the observation window. However, due to partly cloudy weather, the detection probabilities for individual trees vary across the observation window.

Denote the true (and **unknown**) number of pine trees by N_{ij} and the detected number of pine trees by M_{ij}, for grid cells $i, j = 1, \ldots 30$. Let $\boldsymbol{N} = (N_{1,1}, \ldots, N_{30,1}, \ldots, N_{30,30})^{\mathrm{T}}$ and let $\boldsymbol{M} = (M_{1,1}, \ldots, M_{30,1}, \ldots, M_{30,30})^{\mathrm{T}}$. The detection probability is considered fixed in each grid cell and is denoted by α_{ij}, for grid cells $i, j = 1, \ldots, 30$, and we let $\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{30,1}, \ldots, \alpha_{30,30})^{\mathrm{T}}$.

The detection probabilites are given in `obsprob.txt` together with (x, y)-coordinates of the centroids of the grid cells, and the observed counts are given in `obspines.txt`.

a) Display the counts of detected pine trees and the observation probabilities. Assume that the numbers of detected pine trees in the grid cells are independent conditional on the true numbers of pine trees in the grid cells. Specify the observation model (i.e., $\boldsymbol{M}|\boldsymbol{N}$) and its joint probability mass function (i.e., the probability mass function of $\boldsymbol{M}|\boldsymbol{N}$).

b) Assume *a priori* that the pine trees follow a homogeneous Poisson point process with intensity λ. Specify the discretized latent model and derive its probability mass function (i.e., describe how to find the distribution of $\boldsymbol{N}$ and derive the probability mass function of $\boldsymbol{N}$).

c) The estimator $\hat{\Lambda}_1 = \frac{1}{300^2} \sum_{i,j} N_{ij}$ is unbiased for λ, but requires the unknown true counts (and cannot be used). Determine an unbiased estimator $\hat{\Lambda}_2 = C \cdot \sum_{i,j} M_{i,j}$, where

C is a function of α_{ij}, $i, j = 1, \ldots, 30$. Calculate the estimate $\hat{\lambda}$ for the dataset using the estimator $\hat{\Lambda}_2$.

Generate three realizations, for $\lambda = \hat{\lambda}$, of the discretized true counts $\boldsymbol{N}$. For each simulation of $\boldsymbol{N}$, simulate a point pattern by simulating the locations of the points within the grid cells independently from uniform distributions. Display the realized point patterns, and discuss why it is expected that the realizations behave differently than the observed counts in the pine tree dataset.

d) Derive the probability mass function for $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$. Explain that the resulting probability mass function indicates that the point process of undetected points is an inhomogeneous Poisson point process and specify its intensity function.

Generate three realizations for $\lambda = \hat{\lambda}$, where $\hat{\lambda}$ is the estimate from **c)**, of the posterior distribution of the observed counts $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$. For each simulation, simulate a point pattern by simulating the locations of the points within the grid cells independently from uniform distributions. Display the realized point patterns, and discuss the similarities and differences between the realizations in **c)** and the realizations in **d)**.

e) For $\lambda = \hat{\lambda}$, where $\hat{\lambda}$ is the estimate from **c)**, generate i) 500 realizations of $\boldsymbol{N}$, and ii) 500 realizations of $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$. For i) and for ii), compute the average value of the realizations. This estimates the *a priori* expected value ($\mathrm{E}[\boldsymbol{N}]$) and the *a posteriori* expected value ($\mathrm{E}[\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}]$). Estimate also *a priori* standard deviations (from $\boldsymbol{N}$) and *a posteriori* standard deviations (from $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$).

Display estimated expected true counts and estimated standard deviations of true counts in a 2×2 display of figures. Compare the figures and explain the differences. ***NB:*** *Make sure the two figures of estimated expectations have the same (and suitable!) color scale, and the two figures of estimated standard deviations have the same (and suitable!) color scale.*

Problem 3: Point process with clustering

This problem concerns the redwood tree dataset listed above.

a) Consider a Neyman-Scott (parent-daughter) process, where the number of daughter points of a parent is distributed according to a Poisson distribution, and locations of daughter points are generated according acconding to $\mathcal{N}_2(\boldsymbol{y}, \sigma^2 \mathbf{I}_2)$, where $\boldsymbol{y}$ is the parent location. Describe shortly how the model works, specify the full set of model parameters, and interpret the model parameters.

Discuss potential border problems, caused by the use of a finite domain $W \subset \mathbb{R}^2$ when simulating realizations from the model. You will correct for such boundary effects when simulating.

Make an empirical fit of the model parameters to the redwood tree data. This fit need only be done by inspecting the tree pattern and guessing the model parameter values from your intuitive understanding of their impact on the pattern. Evaluate your guess by simulating 100 realizations from the Neyman-Scott process and producing a 90% prediction interval of the empirical L-function under the chosen parameter values.

Iterate your guestimate procedure to improve the fit and try to make the empirical L-function of the data consistent with the 90% prediction interval. List the final guestimates of the model parameters and justify them by displaying the empirical L-function of the data together with the 90% prediction interval. Discuss the results.

Display the redwood tree data set next to three realizations from the guestimated Neyman-Scott model. Comment on the display.

Problem 4: Repulsive point processes

This problem concerns the biological cell dataset listed above.

a) We plan to model the point pattern using a Strauss process with a fixed number of points n and pair potential function

$$\phi(r) = \begin{cases} \beta, & r \leq r_0, \\ 0, & r > r_0. \end{cases}$$

Describe the full set of model parameters and interpret them.

Describe the potential border problems caused by using a bounded observation window $W \subset \mathbb{R}^2$ when simulating from the model. We will ignore boundary problems when simulating in this problem.

Make an empirical fit of the model by guessing parameter values based on their interpretation and the behaviour seen in the dataset. Simulate 100 realizations with the parameters you chose, create a 90% prediction interval for the empirical L-function, and compare to the empirical L-function of the dataset. *Either implement your own sampler, or use the function* `Strauss` *from the* `R` *package* `spatial` *with parameter* $c = \exp(-\beta)$. *See help file (*`?Strauss`*) to understand how to use the function.*

Iterate your guestimate procedure to improve the fit and try to make the empirical L-function of the dataset be consistent with the 90% prediction interval. List the final model parameters guestimates and justify them by displaying the 90% prediction interval and the empirical L-function of the data. This model does not fit the dataset perfectly, discuss the results and suggest how one could change the model to improve the fit.

Display the biological cell data set next to three realizations from the guestimated Strauss process. Comment on the display.