TMA4250 - Project 2

Jakob Heide, Bendik Waade

2024-03-04

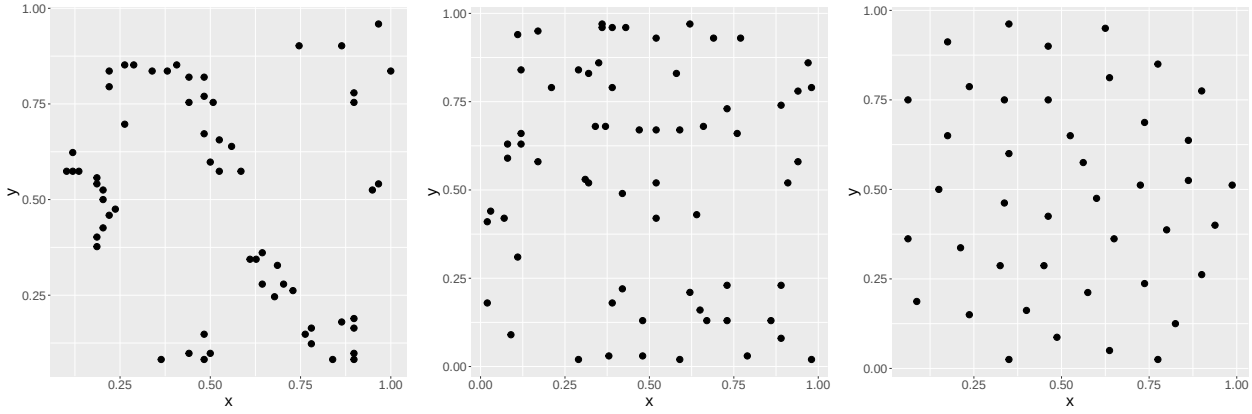
In this project, we consider three real-world point pattern datasets,

- redwood tree data: 62 observations in the observation window $[0,1] \times [0,1]$
- pine tree data: 42 observations in the observation window $[0,1] \times [0,1]$
- biological cell data: 42 observations in the observation window $[0,1] \times [0,1]$

Part 1 - Analysis of point pattern data

Point pattern visualization

We begin by displaying each of the point patterns.



(a) Redwood (b) Pines (c) Cells

Figure 1: Point patterns

In Figure 1a, the point pattern seems to indicate clustering. This might be related to how redwood trees produce offspring and how they grow - for example, some tree types prefer to grow close together so they can share root systems, in order to make them more resilient. In Figure 1c, the point pattern shows repulsion. Natural systems like cell systems tend towards (local) states of minimal energy. For example, if the cells carry similar charges, then the cells will repulse each other and the positioning of the cells will tend towards a state where the distance between cells is maximized. The point pattern in Figure 1b seems to indicate randomness. One possible reason for this is that some trees or plants rely on wind (or animals) to spread their seeds, which might result in a seemingly random spread of the seeds.

The L-function

To quantify the repulsion or clustering of a stationary point process, we can use the L-function, which is defined on $\mathbb{R}^2$ as

$$L(r) = \sqrt{\frac{k(r)}{\pi}}, \quad r \geq 0, \tag{1}$$

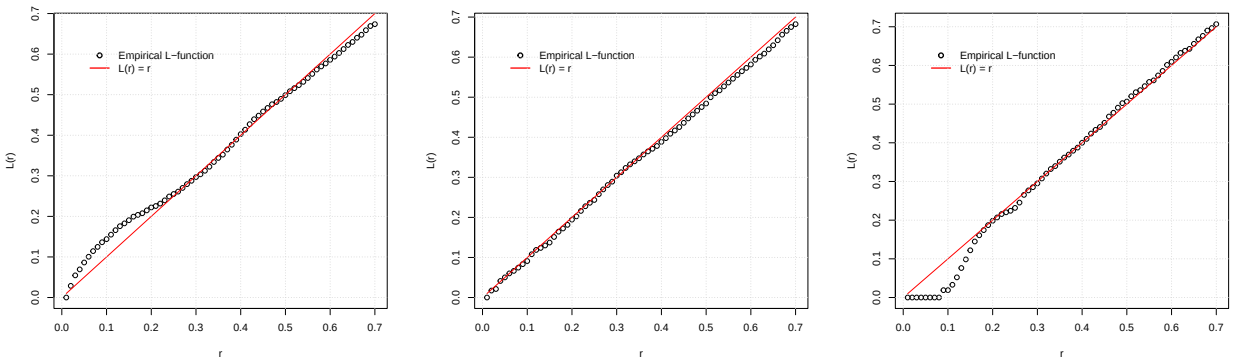
where $k(r)$ is Ripley's K-function. The K-function is defined for a stationary point process N with intensity λ as

$$k(r) = \frac{1}{\lambda} \mathrm{E}_{\mathbf{0}}[N(b(\mathbf{0}, r) \backslash \{\mathbf{0}\})], \quad r \geq 0$$

where the subscript in $\mathrm{E}_{\mathbf{0}}$ denotes the assumption that there is a point in $\mathbf{x} = \mathbf{0}$, and $N(b(\mathbf{0}, \mathbf{r}))$ denotes the number of points in a ball (on $\mathbb{R}^2$, a circle) centered in $\mathbf{0}$ with radius r. For a homogeneous Poisson point process, the K-function becomes $k(r) = \pi r^2$, which gives the L-function

$$L(r) = r, \quad r \geq 0$$

If we replace $k(r)$ in Equation (1) by the empirical K-function $\hat{k}(r)$, we obtain the empirical L-function. We use the function `Kfn` from the library `spatial`, and plot the empirical L-function for each of the point patterns, along with the L-function for a homogeneous Poisson point process.



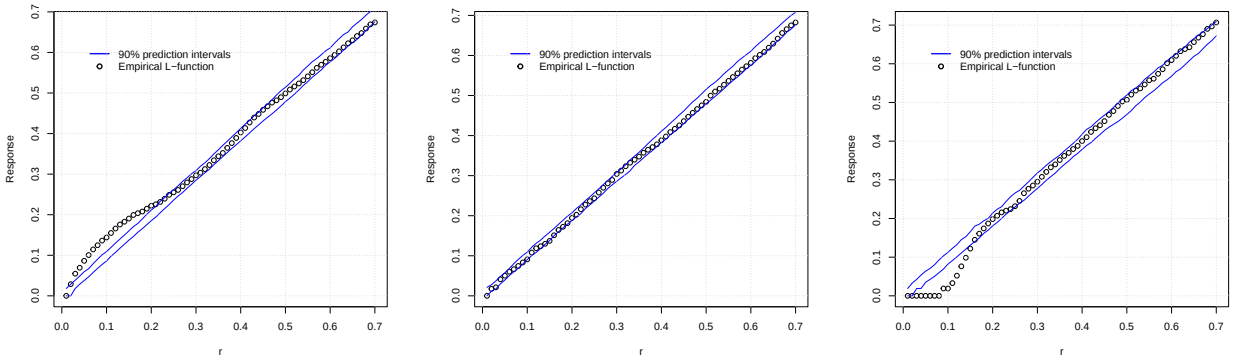
(a) Redwood (b) Pines (c) Cells

Figure 2: Empirical L-functions

In Figure 2a, we see that the empirical L-function lies above the L-function for a homogeneous Poisson point process, for $r = 0$ to $r \approx 0.2$. This indicates that we have some clusterization and that a homogeneous Poisson process is a bad choice for modelling this data. In Figure 2b, the empirical L-function matches pretty well with the theoretical L-function for the homogeneous Poisson process, although there might be some slight deviation for $r > 0.5$. A homogeneous Poisson process appears to be a suitable model. In Figure 2c, the empirical L-function lies below the theoretical L-function for $r < 0.2$, which indicates repulsion and that a homogeneous Poisson process is a bad choice of model.

Prediction intervals

However, it can be hard to tell how much deviation from the line $L(r) = r$ we need in order to conclude that the homogeneous Poisson point process is a bad choice for a model. Therefore, for each dataset, we simulate 100 realizations of a homogeneous Poisson point process with intensity equal to the number of points in the dataset, and calculate the 5% and 95% quantiles, to be used as the lower and upper limit of the 90% prediction interval. The results are shown in Figure 3.



(a) Redwood (b) Pines (c) Cells

Figure 3: Empirical prediction intervals, with the empirical L-functions for each data set.

We see that our conclusions from before for each of the point patterns are supported. In Figure 3a and Figure 3c, more than 10% of the points lie outside the estimated 90% prediction interval, which violates the assumption of a homogeneous Poisson point process. In Figure 3b, almost all the points lie within the prediction interval, which indicates that a Poisson point process is a reasonable model. However, using only 100 realizations to estimate a 90% prediction interval seems a bit thin - to gain more precise estimates, we could increase the number of realizations.

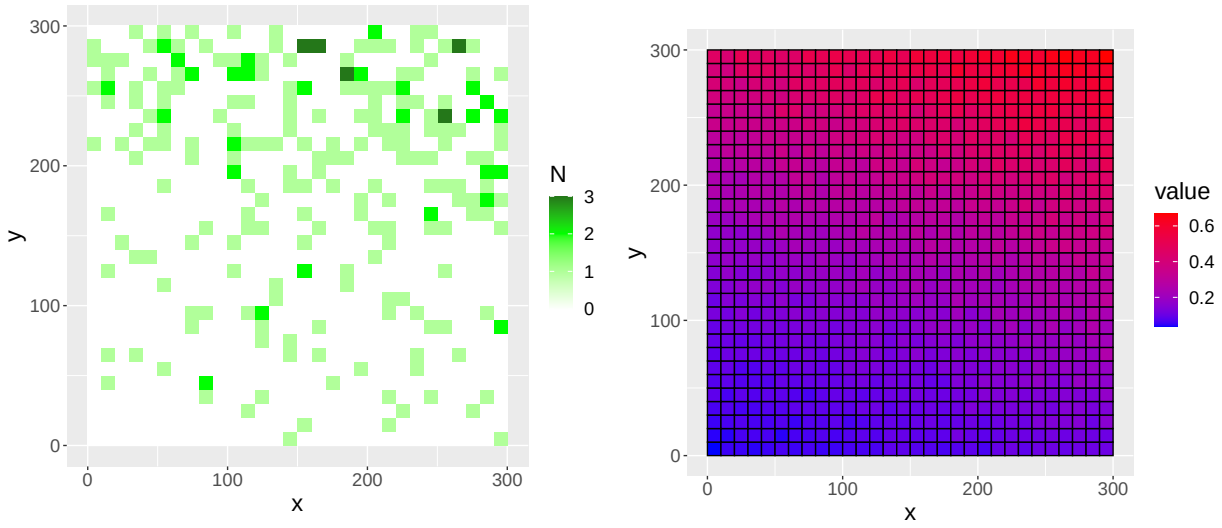
Part 2 - Remote sensing of trees

We consider a $300\text{m} \times 300\text{m}$ observation window, where the locations of pine trees are observed by a satelite. The satelite counts the pine trees in a regular 30×30 grid where each cell is $10\text{m} \times 10\text{m}$ and has an associated detection probability.

For all the cells $i, j = 1, \ldots, 30$, we denote by M_{ij} the detected number of pine trees, by N_{ij} the true number of pine trees, and by α_{ij} the detection probability. Let $\mathbf{M} = (M_{1,1}, \ldots, M_{1,30}, \ldots, M_{30,,30})$, $\mathbf{N} = (N_{1,1}, \ldots, N_{1,30}, \ldots, N_{30,30})$ and $\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{1,30}, \ldots, \alpha_{30,30})$. In addition, let $\mathbf{m} = (m_{1,1}, \ldots, m_{30,30})$ and $\mathbf{n} = (n_{1,1}, \ldots, n_{30,30})$ be realized values of $\mathbf{M}$ and $\mathbf{N}$, respectively.

Data visualization

We display the data in Figure 4.



(a) Pine count observations

(b) Observation probabilities

Figure 4: Data visualization

In Figure 4a, we see that the number of observed pine trees increases towards the top of the grid, that is, with increasing values of y. In Figure 4b, we see that the observation probabilites increase with increasing values of y and slightly with increasing values of x.

Observation model

We assume that the pine trees are detected independently of each other. Consider the conditional $\mathbf{M}|\mathbf{N}$, where the true amount of trees is given. For $i, j = 1, \ldots, 30$, the probability of observing M_{ij} trees is then binomially distributed, where N_{ij} is the number of trials and α_{ij} is the success probability. The observation model is

$$M_{ij}|N_{ij} \sim \text{Bin}(N_{ij}, \alpha_{ij}),$$

with the probability mass function

$$f_{M_{ij}|N_{ij}}(m_{ij}|n_{ij}; \alpha_{ij}) = \binom{n_{ij}}{m_{ij}} \alpha_{ij}^{m_{ij}} (1 - \alpha_{ij})^{n_{ij} - m_{ij}}, \quad m_{ij} = 0, 1, \ldots$$

Since the observations from one grid cell are independent from the observations of another grid cell, we can write the joint probability mass function as

$$f_{\mathbf{M}|\mathbf{N}}(\mathbf{m}|\mathbf{n}; \boldsymbol{\alpha}) = \prod_{i,j=1}^{30} \binom{n_{ij}}{m_{ij}} \alpha_{ij}^{m_{ij}} (1 - \alpha_{ij})^{n_{ij} - m_{ij}}, \quad m_{ij} = 0, 1, \ldots \tag{2}$$

Prior model

We assume that the number of pine trees follow a homogeneous Poisson point process with intensity λ. The observation window $W = [0, 300] \times [0, 300]$ is discretized as before by a regular 30×30 grid. For $i, j = 1, \ldots, 30$, we define the following:

- The center of the cell, i.e. the centroid: $\mathbf{s}_{ij} = ((i - \frac{1}{2}) \cdot 10, (j - \frac{1}{2}) \cdot 10)^T$.
- The grid cell: $B_{ij} = \mathbf{s}_{ij} + (-5, 5] \times (-5, 5]$
- Grid cell counts: $N_{ij} = N(B_{ij})$

Inside a grid cell, the number of pine trees is Poisson distributed with rate 100λ. Since the observations in different cells are independent from each other, we can write the probability mass function of $\mathbf{N}$ as the product of the probability mass functions for N_{ij}, $i, j = 1, \ldots, 30$. The joint probability mass function is thus

$$f_{\mathbf{N}}(\mathbf{n}; \lambda) = \prod_{i,j=1}^{30} f_{N_{ij}}(n_{ij}; \lambda)$$

$$= \prod_{i,j=1}^{30} \frac{(100\lambda)^{n_{ij}}}{n_{ij}!} \exp(-100\lambda), \quad n_{ij} = 0, 1, \ldots \tag{3}$$

Estimator for λ

Since $\mathbf{N}$ follows a Poisson point process with intensity λ, we know that the observed counts $\mathbf{M}$ follow a Poisson point process with intensity $\lambda\boldsymbol{\alpha}$. The joint probability mass function for $\mathbf{M}$ becomes

$$f_{\mathbf{M}}(\mathbf{m}; \boldsymbol{\alpha}, \lambda) = \prod_{i,j=1}^{30} \frac{(100\lambda\alpha_{ij})^{m_{ij}}}{m_{ij}!} \exp(-100\lambda\alpha_{ij}), \quad m_{ij} = 0, 1, \ldots \tag{4}$$

To estimate the intensity λ, we would preferably use the estimator

$$\hat{\Lambda}_1 = \frac{1}{300^2} \sum_{i,j} N_{ij},$$

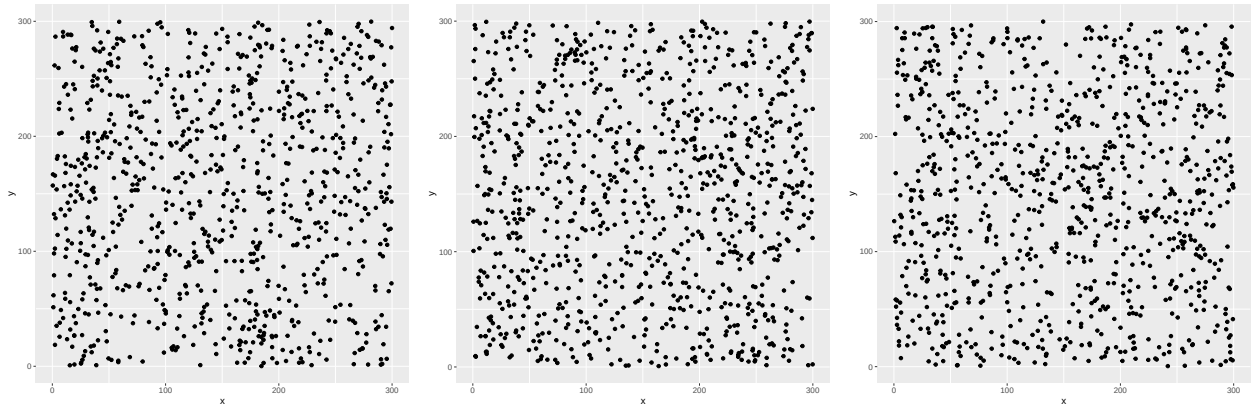
which is unbiased. However, it requires the true counts N_{ij} for $i, j = 1, \ldots, 30$, which we do not have access to. We therefore turn to the estimator

$$\hat{\Lambda}_2 = C \sum_{i,j} M_{ij},$$

which is unbiased for $C = 1/(100 \sum_{ij} \alpha_{ij})$, since

$$\mathrm{E}[\hat{\Lambda}_2] = C \sum_{i,j} \mathrm{E}[M_{ij}] = 100C\lambda \sum_{ij} \alpha_{ij},$$

We generate three realizations of the discretized true counts $\mathbf{N}$, using the estimator $\hat{\Lambda}_2$. For each realization, we place the n_{ij} points uniformly in each grid cell $i, j = 1, \ldots, 30$, and display the point patterns in Figure 5.



(a) One realization (b) One realization (c) One realization

Figure 5: Realizations of $\mathbf{N}$

In Figure 5, we see that the realizations of $\mathbf{N}$ behave differently than the observed counts in Figure 4a. The reason for this is that $\mathbf{N}$ are assumed to follow a homogeneous Poisson point process, but the observed counts $\mathbf{M}$ follow an inhomogeneous Poisson process, where the intensity varies from grid cell to grid cell, due to the observation probabilities $\boldsymbol{\alpha}$.

Posterior distribution

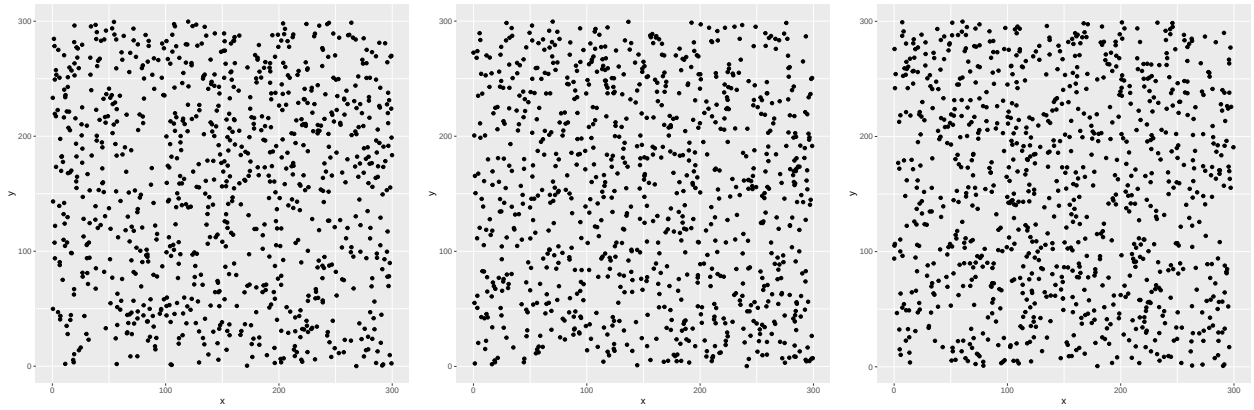
To find the probability mass function for the conditional $\mathbf{N}|\mathbf{M} = \mathbf{m}$, we use Bayes' rule in addition to the definition of conditional density, which gives

$$f_{\mathbf{N}|\mathbf{M}=\mathbf{m}} = \frac{f_{\mathbf{M}|\mathbf{N}} f_{\mathbf{M}}}{f_{\mathbf{N}}}$$

We know $f_{\mathbf{M}|\mathbf{N}}$ from Equation (2), $f_{\mathbf{N}}$ from Equation (3), and $f_{\mathbf{M}}$ from Equation (4). After inserting the three expressions and carrying out some algebra, we find

$$f_{\mathbf{N}|\mathbf{M}=\mathbf{m}}(\mathbf{n}|\mathbf{m}; \boldsymbol{\alpha}, \lambda) = \prod_{i,j=1}^{30} \frac{(100\lambda(1-\alpha_{ij}))^{n_{ij}-m_{ij}}}{(n_{ij}-m_{ij})!} \exp(-100\lambda(1-\alpha_{ij})), \quad n_{ij} - m_{ij} = 0, 1, \ldots, \tag{5}$$

i.e. the unobserved trees $(\mathbf{N} - \mathbf{M})|\mathbf{M}$ follow a Poisson distribution with intensity $\lambda(1 - \boldsymbol{\alpha})$. To create realizations, we sample from this distribution and add the number of observed trees, M_{ij}, to all grid cells $i, j = 1, \ldots, 30$.



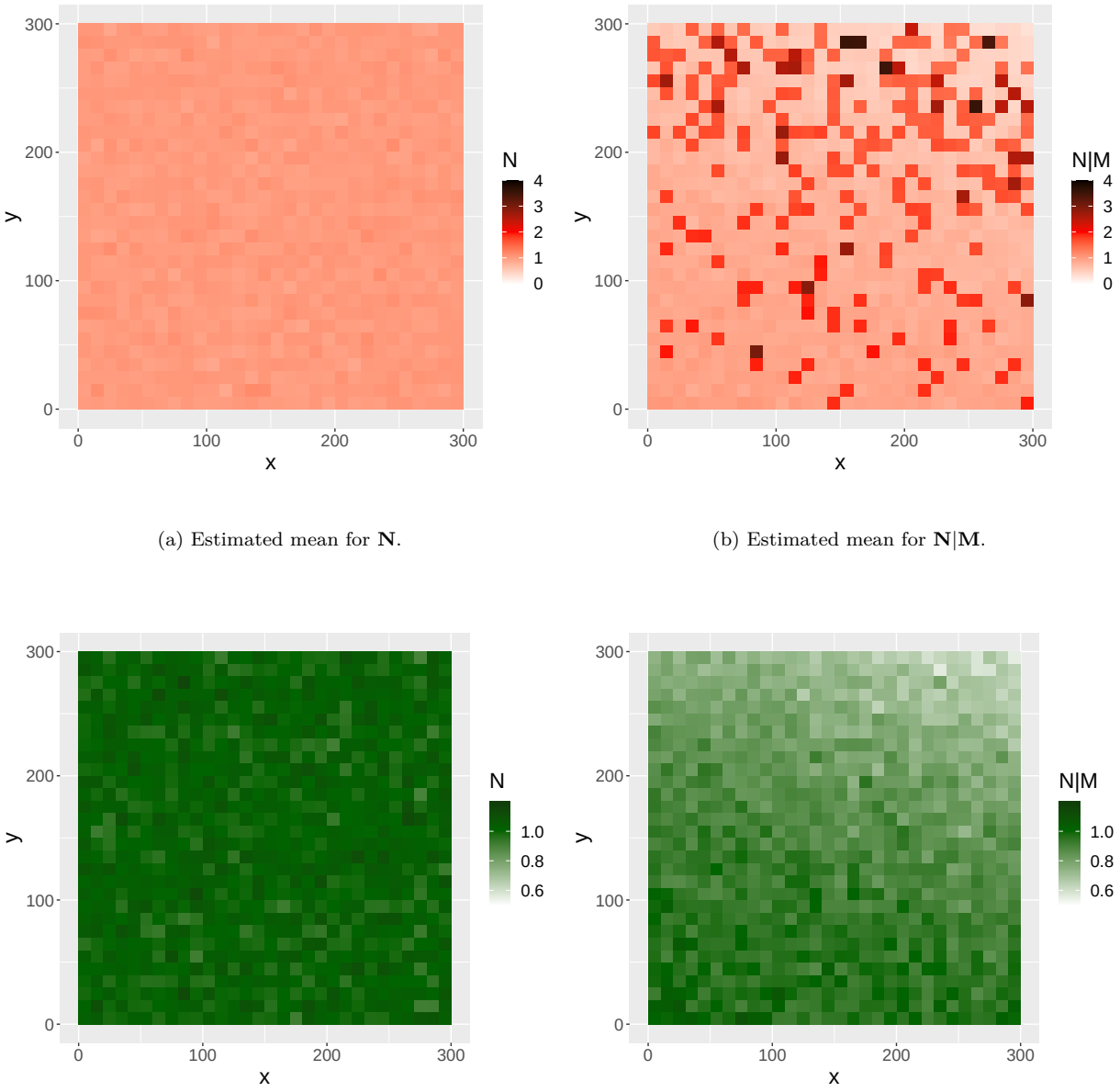
(a) One realization (b) One realization (c) One realization

Figure 6: Realizations of $\mathbf{N}|\mathbf{M}$

When we compare Figure 6 with Figure 5, we see that the realizations produced are very similar. The fact that conditioning the true counts $\mathbf{N}$ on the observed counts $\mathbf{M}$ does not change the realizations significantly indicates that the choice of prior for $\mathbf{N}$ was a good choice - namely that $\mathbf{N}$ follows a homogeneous Poisson point process.

Estimating the mean and standard deviation in each cell

We simulate 500 realizations of $\mathbf{N}$ and $\mathbf{N}|\mathbf{M}$. In both cases, we calculate the mean and the standard deviation in each grid cell, using the realizations.



(a) Estimated mean for $\mathbf{N}$.

(b) Estimated mean for $\mathbf{N}|\mathbf{M}$.

(c) Estimated standard deviation for $\mathbf{N}$.

(d) Estimated standard deviation for $\mathbf{N}|\mathbf{M}$.

Figure 7: Estimated means and standard deviations.

In Figure 7a and Figure 7b, we see the estimated means for $\mathbf{N}$ and $\mathbf{N}|\mathbf{M}$, respectively. The mean values for $\mathbf{N}$ seem to be constant, but for $\mathbf{N}|\mathbf{M}$, the mean values are much higher in the grid cells where we have observed trees. In Figure 7c and Figure 7d, we see that the standard deviation for $\mathbf{N}$ is approximately constant across the grid, but the standard deviation for $\mathbf{N}|\mathbf{M}$ is smaller towards areas where the observation probabilities are higher.

Part 4 - Repulsive point processes

We consider the biological cell count data displayed in Figure 1c, which showed signs of repulsion. We will attempt to model the data using a Strauss process with a fixed number of points and the pair-potential function

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

The model parameters to be considered are

- r_0: when points are less than r_0 apart, there is an interaction between them. When points are further than r_0 apart, the interaction between them are 0.
- β: this parameter determines the strength of the interaction between points, e.g. $\beta = \infty$ would give the Gibbs hard-core process, where points cannot be closer than r_0. A smaller β increases the probability that points are closer together.

Potential border problems arise when using a bounded observation window $W \subset \mathbb{R}^2$. If there exists a point outside of W that is less than r_0 away from the boundary, there might be some point within W that should be affected by the point outside W. We will ignore this potential boundary issue.

We make a rough guess of the parameters by looking at Figure 1c. The points are spaced relatively equally apart, but it is clear that some are closer than others, so we set $\beta \approx 10$. For r_0, we look at the points at the bottom from $x = 0.5$ to $x = 0.75$, and estimate $r_0 \approx 0.13$. Using these two parameters, we simulate 100 realizations and create an estimated 90% prediction interval using the empirical L-function of each realization, which we compare to the empirical L-function of the data. The results are shown in Figure 8.

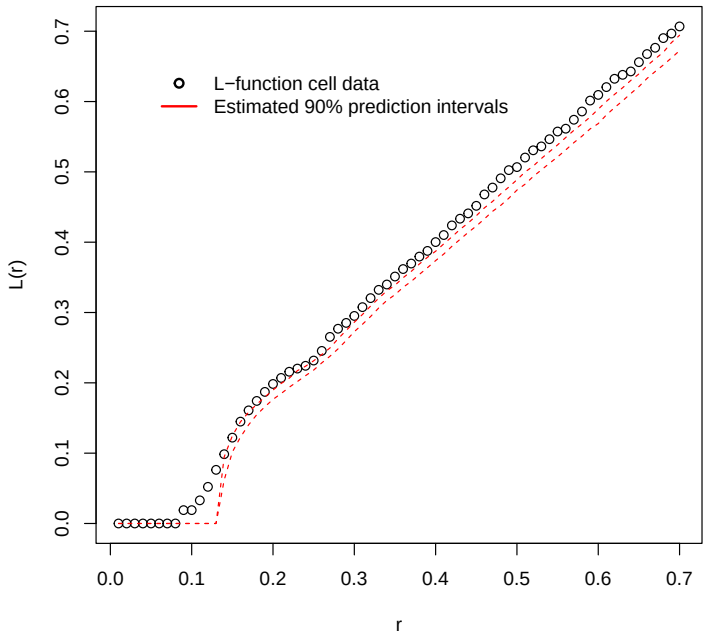


Figure 8: Estimated prediction intervals, with the empirical L-function of the data.

The estimated 90% prediction interval does not fit the empirical L-function very well. We therefore make a guess on new values for the parameters, by setting $\beta = 6$ and $r_0 = 0.11$. We repeat the procedure from above and display the new results in Figure 9.

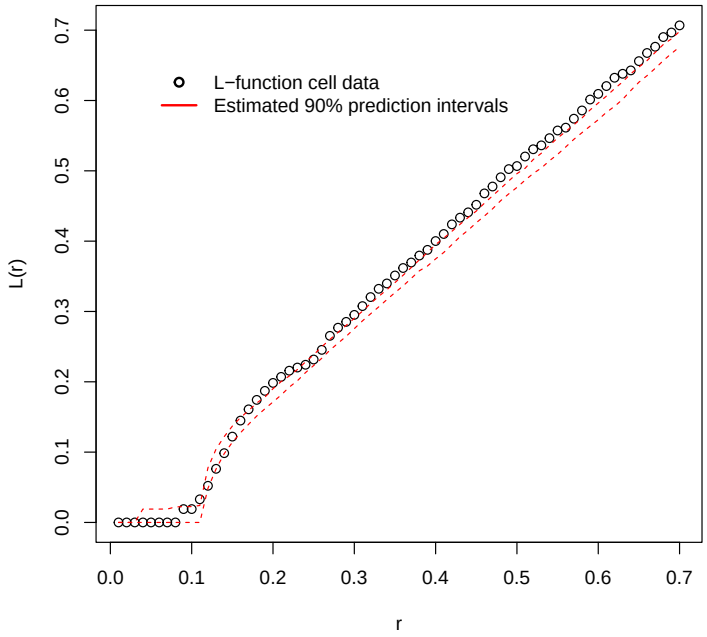
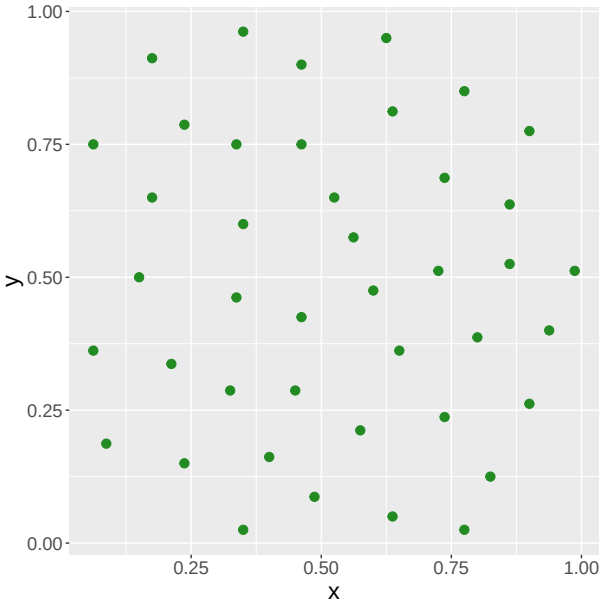


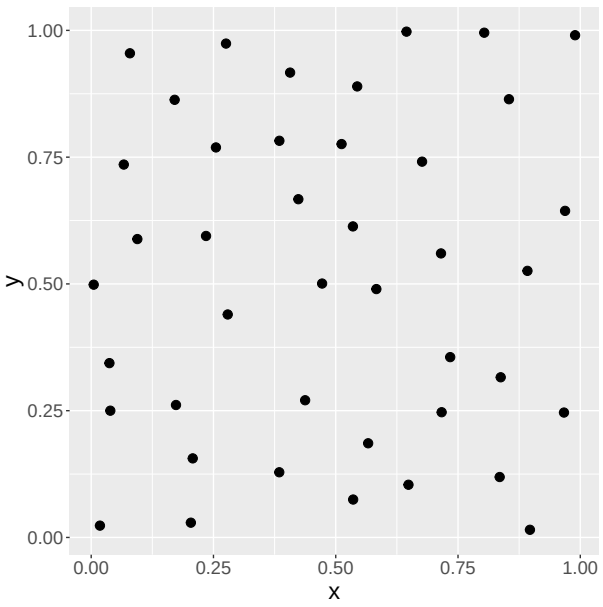
Figure 9: Estimated prediction intervals, with the empirical L-function of the data.

In Figure 9, we see that the 90% prediction intervals fails to capture most of the points in the empirical L-function of the data - however, after some trial and error, these were the best parameters we were able to find. This leads us to conclude that a Strauss process is too simple of a model for modelling the cell data. One possible modification would be to include more parameters to model the interaction between cells. Another modification we might consider making is to include boundary effects, since the data most likely has these effects.

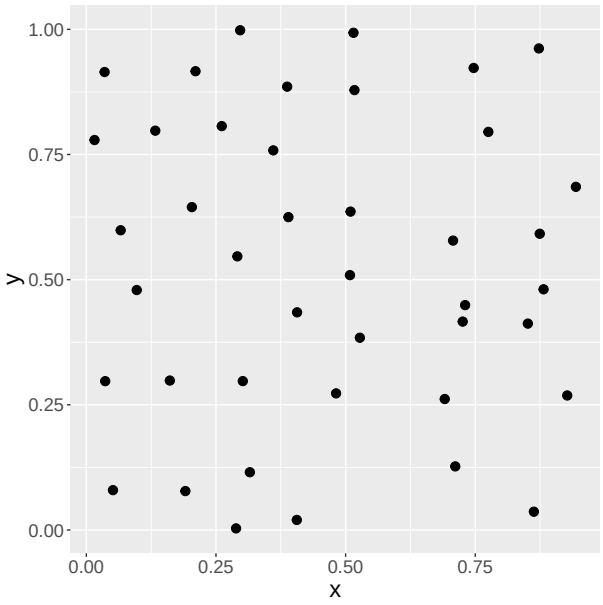
In Figure 10, the biological data set is displayed, along with three realizations from the Strauss process with $\beta = 6$ and $r_0 = 0.11$.



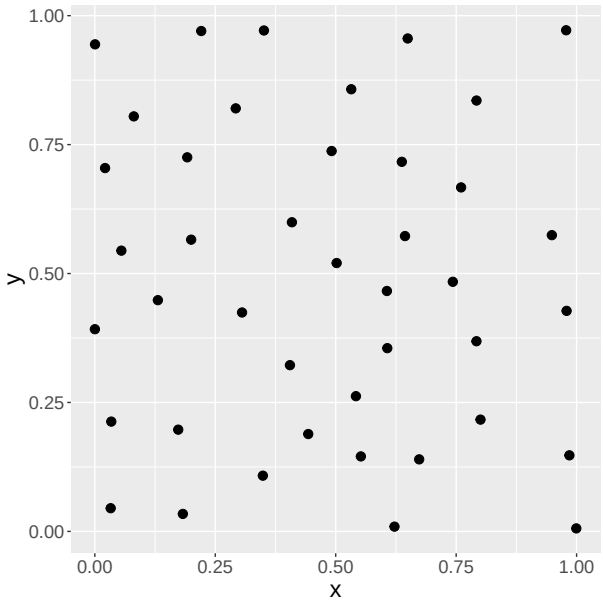
(a) Biological cell data



(b) One realization



(c) One realization



(d) One realization

Figure 10: The biological cell data compared to three different realizations of the Strauss process.

The realizations are rather similar to the data. A few points might be a bit closer together in the realizations than in the data, but overall, the Strauss process seems to generate realizations which are similar to that of the biological cell data.