TMA4250 Spatial Statistics
Project 2, Spring 2022

Christian Moen, Jim Totland

Problem 1: Analysis of point pattern data

In this section, we study three point pattern data sets:

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

which are available in the R-package `MASS`.

a)

The point pattern data are displayed in Figure 1. The biological cell data, in Figure 1a, clearly shows that there is some repulsion present, as there seems to be a minimum distance between every pair of points. This is similar to the "hard-core" process discussed in the lectures. The repulsion could be due to the physical size of each cell, or there could be some electrochemical potential which makes this grid-like structure of the cells 'optimal' with respect to stability or potential energy. There seems to be no clusters of cells in the data.

The display of the red wood tree data in 1b clearly illustrates that there is some clustering effect present in this data. The number of clusters is debatable, but it is clear that some regions have trees while others do not. This could be due to favorable conditions for the trees at certain locations, e.g. good soil. Another explanation is that the trees could benefit from being closer to other trees.

The pine tree data set, displayed in 1c, reveals no clear clustering or repulsion to our eyes. The point pattern seems to follow a Poisson point process.

b)

To further investigate the characteristics of the given point processes, we consider the L-function. For a stationary point processes N with intensity λ on $\mathbb{R}^2$, the L-function is given by

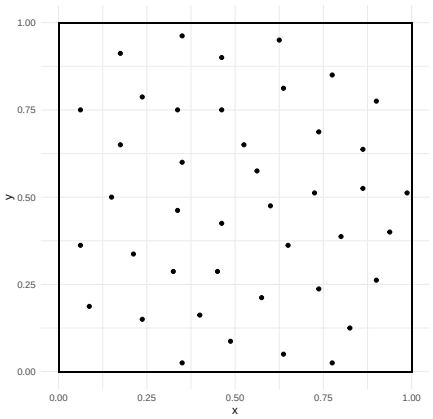
$$L(r) = \sqrt{\frac{K(r)}{\pi}},$$

where $K(r)$ is defined as

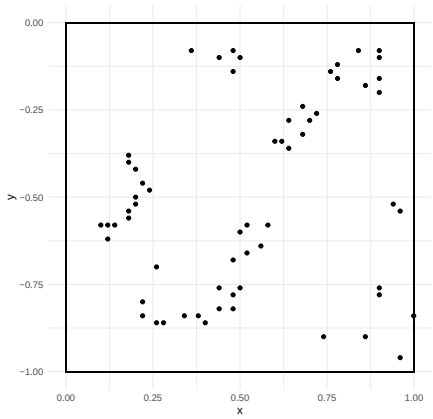
$$K(r) = \frac{1}{\lambda}\mathrm{E}_{\mathbf{0}}\left[N(b\left(\mathbf{0}, r\right) \setminus \{\mathbf{0}\})\right],$$

where $b(\mathbf{0}, r)$ is the ball centered at $\mathbf{0}$ with radius r and the index on the expectation indicates that we take the expectation conditional on the fact that we have observed a point in $\mathbf{0}$. We note that for a Poisson point process N with intensity λ on $\mathbb{R}^2$, K becomes

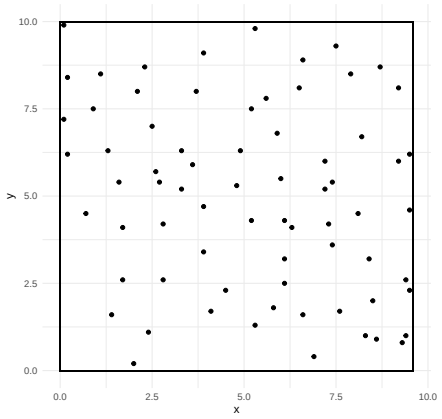
$$\begin{aligned} K(r) &= \frac{1}{\lambda}\mathrm{E}_{\mathbf{0}}\left[N(b\left(\mathbf{0}, r\right) \setminus \{0\})\right] \\ &= \frac{1}{\lambda}\mathrm{E}\left[N(b\left(\mathbf{0}, r\right) \setminus \{0\})\right] \\ &= \frac{1}{\lambda}\pi r^2\lambda = \pi r^2. \end{aligned}$$



(a) `cells.dat`

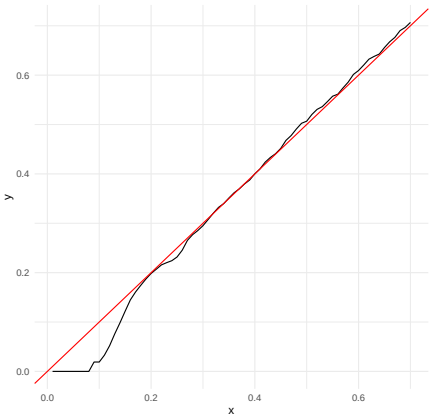


(b) `redwood.dat`

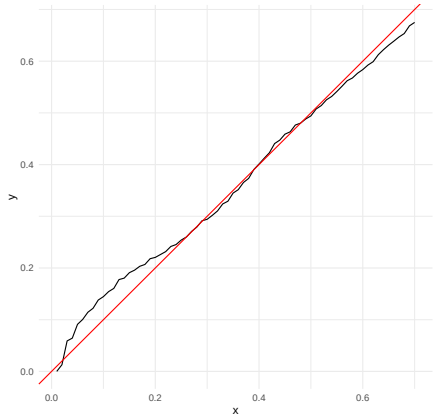


(c) `pines.dat`

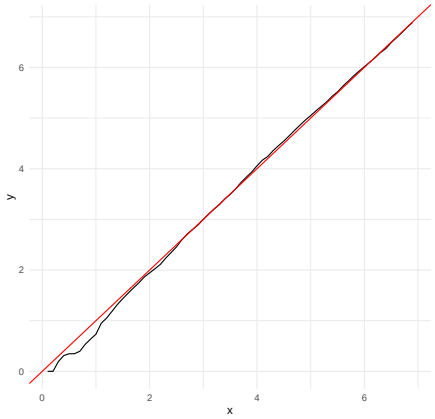
Figure 1: The three point pattern data sets described in Problem 1 displayed with rectangles illustrating the observation window.



(a) `cells.dat`



(b) `redwood.dat`



(c) `pines.dat`

Figure 2: The estimated L-function for the three point pattern data sets described in Problem 1 (black) plotted with the theoretical L-function of a homogeneous Poisson process (red).

Thus, the corresponding L-function simply becomes

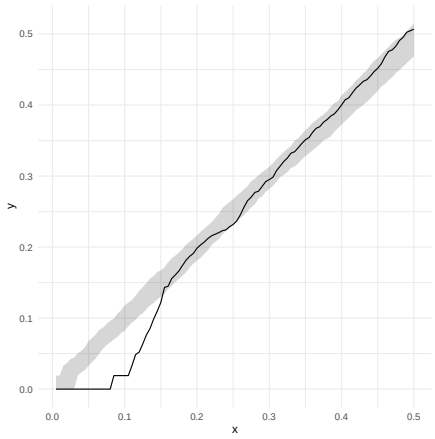
$$L(r) = \sqrt{\frac{\pi r^2}{\pi}} = r.$$

For the point process data sets, the L-function of course has to be estimated. This is done with the function `Krf` from the R-package `spatial`. The resulting estimates are plotted in Figure 2 together with the theoretical L-function of the homogeneous Poisson process.

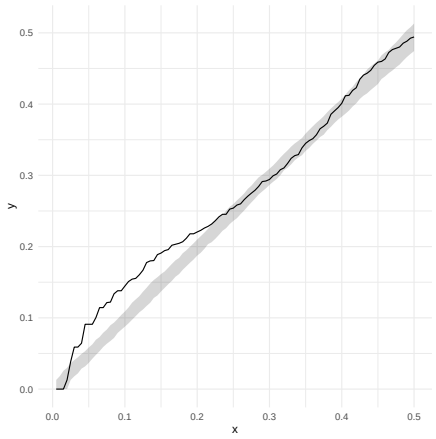
Considering the L-function of the cell data set, displayed in Figure 3a, we observe a slight 'dip' for $r < 0.2$, which indicates that short scale repulsion is present. This agrees with our observation in a), and consequently disqualifies the Poisson process as a reasonable model.

The estimated L-function for the redwood data set is given in Figure 3b and shows a 'bump' (inverted dip) for $r < 0.3$, which signals a medium scale clustering effect. This also matches our observations from a) and informs us that the Poisson process would be a poor model choice.

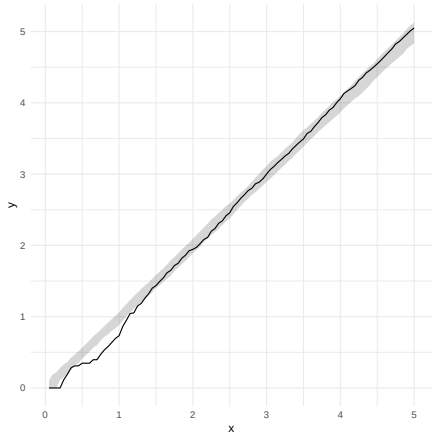
Lastly, considering the estimated L-function of the pines data set in Figure 3c, there is a small dip in the beginning, although it is not nearly as pronounced as for the cells data set. Among the point processes, the pines data set has an estimated L-function which is most similar to that of a homogeneous Poisson point process. This is also in line with what we observed in a), and arguably makes the Poisson process a suitable model for this point process.



(a) `cells.dat`



(b) `redwood.dat`



(c) `pines.dat`

Figure 3: The empirical L-function of the three point pattern data sets described in Problem 1 (black line) and approximate prediction intervals of the empirical L-functions of Poisson point processes conditional on the same amount of points.

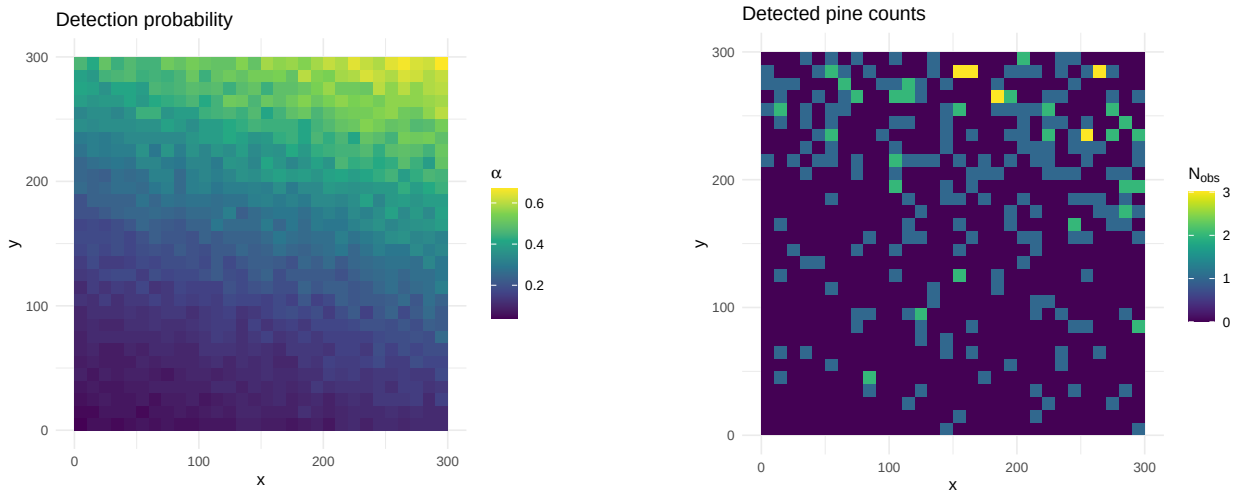
c)

For each of the datasets, we now generate 100 realizations of a homogeneous Poisson process conditional on the number of points being equal to the observed number of points in the dataset. From this, we construct 90% pointwise prediction intervals of the L-function belonging to the homogeneous Poisson processes, which are displayed jointly with the empirical L-functions of the datasets in figure 3. The findings are largely the same as in **b)**, namely that the cells and redwood datasets deviates noticeably from the prediction intervals, which indicates that the homogeneous Poisson point process is a poor model choice fot these, while for the pines dataset, the deviation from the prediction intervals is less pronounced, and the homogeneous Poisson point process might be a reasonable model.

Problem 2: Remote sensing of trees

We consider a 300 m $\times$ 300 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 m $\times$ 10 m grid cells for a regular 30 $\times$ 30 grid of the observation window. However, due to partly cloudy weather, the detection probabilities for individual trees vary across the observation window.

We denote the true (and unknown) number of pine trees by N_{ij} and the detected number of pine trees by



(a) (b)

Figure 4: The detection probabilities (a) and the detected pine counts (b) of the point process described in Problem 2.

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 $\boldsymbol{\alpha} = (\alpha_{1,1}, \ldots, \alpha_{30,1}, \ldots, \alpha_{30,30})^{\mathrm{T}}$.

a)

The observation probabilities and detected pine counts are displayed in Figure 4.

For cell number i, j, the count of detected pine trees $M_{i,j}$ given the true count $N_{i,j} = n_{i,j}$ can be viewed as a binomially distributed variable, with parameters $n = n_{i,j}$ and $p = \alpha_{i,j}$. That is,

$$f_{M_{i,j}|N_{i,j}}(m_{i,j}|n_{i,j};\alpha_{i,j}) = \binom{n_{i,j}}{m_{i,j}} \alpha_{i,j}^{m_{i,j}} (1-\alpha_{i,j})^{n_{i,j}-m_{i,j}}$$

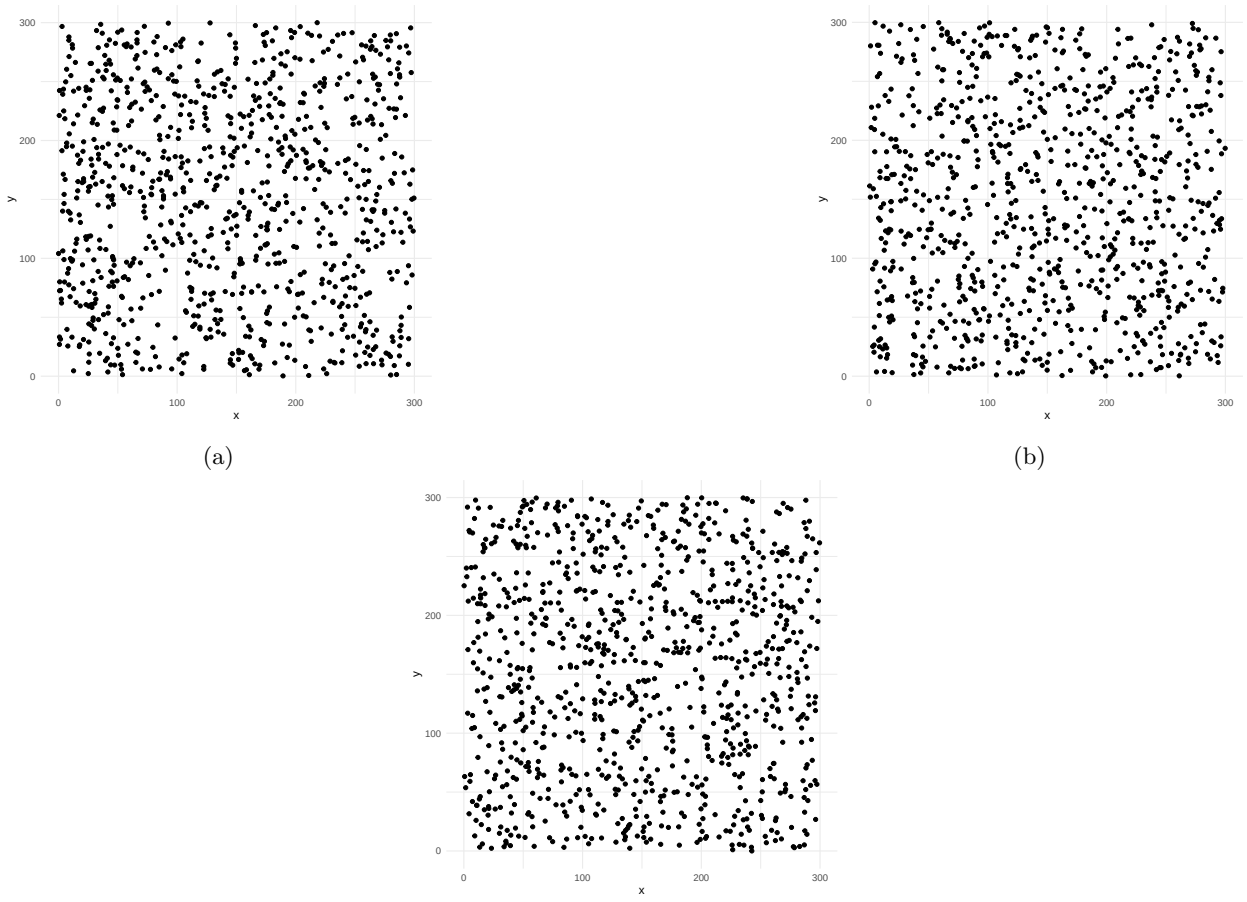
The count of detected pines given the true count in each cell are assumed to be independent of all other cells, which means that we can write the distribution of $\boldsymbol{M}|\boldsymbol{N}$ as

$$f_{\boldsymbol{M}|\boldsymbol{N}}(\boldsymbol{m}|\boldsymbol{n};\boldsymbol{\alpha}) = \prod_{i,j} f_{M_{i,j}|N_{i,j}}(m_{i,j}|n_{i,j};\alpha_{i,j}).$$

b)

We assume *a priori* that the pine trees follow a homogeneous Poisson point process with intensity λ. This means that the number of pine trees in a grid cell follows a Poisson distribution, $N_{i,j} \sim \text{Poisson}(\Delta_x \Delta_y \lambda)$. We note that the width and height of the cells $\Delta_x = \Delta_y = 10$. Since the number of pine trees in each pair of cells will be independent, the probability mass function of $\boldsymbol{N}$ can be constructed as

$$f_{\boldsymbol{N}}(n_{1,1}, \ldots, n_{30,30}) = \prod_{i,j} \frac{(100\lambda)^{n_{i,j}}}{n_{i,j}!} \exp(-100\lambda).$$



(a) (b)

(c)

Figure 5: Three realizations, (a), (b) and (c) of the Poisson point process with estimated intensity $\lambda = \hat{\Lambda}_2$.

c)

An unbiased estimator of λ is given by $\hat{\Lambda}_1 = \frac{1}{300^2}\sum_{i,j} N_{i,j}$, but calculation of this requires the unknown true counts. We therefore want to determine C such that $\hat{\Lambda}_2 = C \cdot \sum_{i,j} M_{i,j}$ is an unbiased estimator of λ. Taking the expectation yields

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

from which we see that setting $C = \left[100\sum_{i,j}\alpha_{i,j}\right]^{-1}$ gives an unbiased estimator. The given data set yields the estimate $\hat{\Lambda}_2 \approx 0.01$. Using this estimate, we generate three realizations of the true discretized counts, $\boldsymbol{N}$, and simulate a corresponding point pattern. The simulated point patterns are displayed in Figure 5. The realizations are expected to differ from the true counts because we do not take the observation probability into account.

d)

Next, we want to derive the probability mass function of $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$. Accordingly, we note that

$$[\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}] = [\boldsymbol{N}, \boldsymbol{M}][\boldsymbol{M}]^{-1} = [\boldsymbol{M}|\boldsymbol{N}][\boldsymbol{N}][\boldsymbol{M}]^{-1},$$

from which we find that the probability mass function is given by

$$\begin{aligned}
f_{\boldsymbol{N}|\boldsymbol{M}}(\boldsymbol{n}|\boldsymbol{m};\boldsymbol{\alpha}) &= f_{\boldsymbol{M}|\boldsymbol{N}}(\boldsymbol{m}|\boldsymbol{n};\boldsymbol{\alpha}) f_{\boldsymbol{N}}(\boldsymbol{n}) f_{\boldsymbol{M}}(\boldsymbol{m};\boldsymbol{\alpha})^{-1} \\
&= \prod_{i,j} f_{M_{i,j}|N_{i,j}}(m_{i,j}|n_{i,j};\alpha_{i,j}) \cdot f_{N_{i,j}}(n_{i,j}) \cdot f_{M_{i,j}}(m_{i,j};\alpha_{i,j})^{-1} \\
&= \prod_{i,j} \binom{n_{i,j}}{m_{i,j}} \alpha_{i,j}^{m_{i,j}} (1-\alpha_{i,j})^{n_{i,j}-m_{i,j}} \cdot \frac{(100\lambda)^{n_{i,j}}}{n_{i,j}!} \exp(-100\lambda) \cdot \frac{m_{i,j}!}{(100\lambda\alpha_{i,j})^{m_{i,j}}} \exp(-100\lambda\alpha_{i,j}) \\
&= \prod_{i,j} \frac{(1-\alpha_{i,j})^{n_{i,j}-m_{i,j}}}{(n_{i,j}-m_{i,j})!} (100\lambda)^{n_{i,j}-m_{i,j}} \exp(100\lambda(\alpha_{i,j}-1)).
\end{aligned}$$

If we define $\Delta_{i,j} := n_{i,j} - m_{i,j}$, i.e. the number of undetected points in cell number i, j, we retain

$$f_{\boldsymbol{N}|\boldsymbol{M}}(\boldsymbol{n}|\boldsymbol{m};\boldsymbol{\alpha}) = \prod_{i,j} \frac{[(1-\alpha_{i,j})100\lambda]^{\Delta_{i,j}}}{\Delta_{i,j}!} \exp\{-100\lambda(1-\alpha_{i,j})\},$$

which indicates that the point process of undetected points is an inhomogeneous Poisson point process with intensity given by $\lambda(1-\alpha_{i,j})$ in cell number i, j.

We simulate three realizations of $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$ by simulating the inhomogeneous Poisson process of undetected points and then adding the observed $\boldsymbol{M} = \boldsymbol{m}$. A point pattern is then simulated by simulating the locations of points within each cell independently from a uniform distribution. The results are displayed in Figure 6, and reveal no obvious difference with the simulation of $\boldsymbol{N}$ in **c)** to our eyes.

e)

Finally, we generate 500 realizations of $\boldsymbol{N}$ and 500 realizations of $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$, which we use to compute estimates of the *a priori* and *a posteriori* expected values, which are displayed in Figure 7a. We also create estimates of the *a priori* and *a posteriori* standard deviations, which are displayed in Figure 7b.

The display shows that the estimated mean and standard deviation of $\boldsymbol{N}$ are approximately constant over the grid, which is anything but unexpected since we simulate from a homogeneous Poisson point process. The estimates related to $\boldsymbol{N}|\boldsymbol{M} = \boldsymbol{m}$, however, show that the mean is mostly constant, but has some 'spikes', which indicates that the intensity of points is not homogeneous.

The estimated standard deviation of $\boldsymbol{N}|\boldsymbol{M}$ clearly varies across the grid. If we compare to Figure 4a, we see that it is inversely proportional to the observation probabilities. This makes sense, at least for the standard deviations: a higher observation probability means less uncertainty and consequently less variation.

Problem 3: Clustered event spatial variables

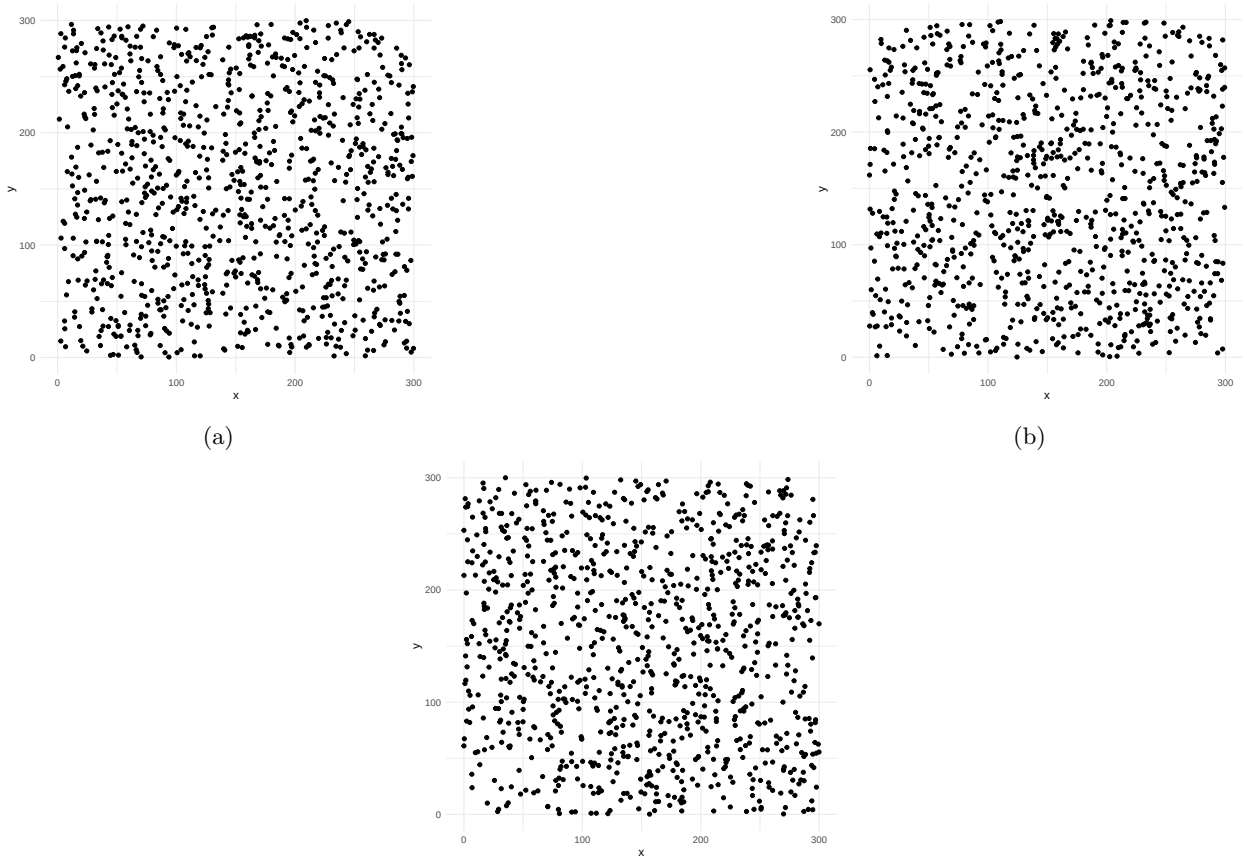
The model

A Neyman-Scott (parent-daughter) process N on $\mathbb{R}^2$ consist of underlying parent points, Y, used to generate daughter points, X_C. That is, each parent point is distributed according to a homogeneous Poisson point process with intensity λ on the given window $W \subseteq \mathbb{R}^2$. The number of daughter points assigned to each parent, C, is distributed with probability mass function $p_C(c) = P(C = c)$, and each daughter is placed by a distance, R_c, and a direction $\hat{r}_c$ away from its parent. Then the location of a daughter point is given by

$$\boldsymbol{X}_C = \boldsymbol{Y} + R_c \hat{r}_c,$$

where $\boldsymbol{X}_C = [X_1, ..., X_C]$ and $\boldsymbol{Y} = [Y_1, ..., Y_C]$.

If the domain is bounded, that is, $W \subset \mathbb{R}^2$, when simulating realization from the model, two problems arise. One is the case when a parent point is outside our domain and has daughters inside. The other is the opposite case, i.e., a daughter point is outside the domain while its parent is inside. Extending the window by a sufficient amount such that $W \subset \tilde{W}$, parents outside the original window with daughters inside will be included. Parents inside the domain may also generate daughters outside the domain, which must be excluded.

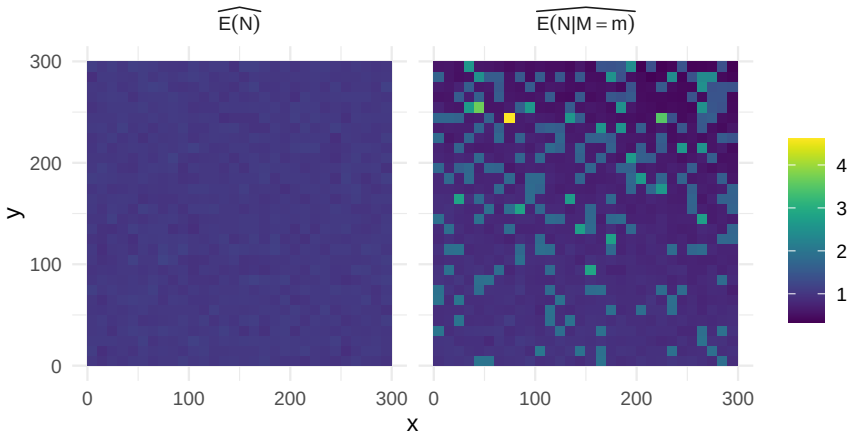


(a)

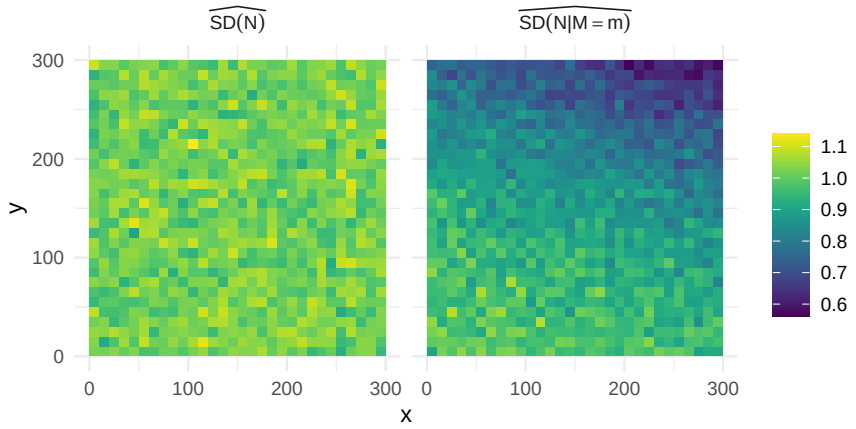
(b)

(c)

Figure 6: Three realizations, (a), (b) and (c) of the inhomogeneous Poisson point process which arises from $\boldsymbol{N}|\boldsymbol{M}$, with estimated $\lambda = \hat{\Lambda}_2$.



(a)



(b)

Figure 7: (a) Estimated *a priori* (left) and *a posteriori* (right) expected values. (b) Estimated *a priori* (left) and *a posteriori* (right) standard deviations.

Model fitting and simulations

Now we will make an empirical fit to the redwood tree dataset displayed in Figure 2. We know that the locations of the daughter points are normally distributed, i.e., $X \sim \mathcal{N}_2(y, \sigma^2\mathbf{I})$, where $Y = y$ is a realization of the parent point location and $\sigma^2\mathbf{I}_2 = \text{Var}(R_c\hat{r}_c)$. We suggest that the total number of clusters $\lambda_n = 6$ where two of the points originate from a parent located outside our window. Each cluster correspond to a parent point. The total number of parents in a domain is given by $\lambda_N = \lambda\nu(W)$, where $\nu(W) = 1$ is the area of the window. Thus, the guestimated intensity $\tilde{\lambda} = 6$. Since the observed number of points per cluster are Poisson distributed we also need a rate parameter, which we suggest to be the mean number of daughters per cluster. Observed number of daughters per cluster were $C_g = \{5, 9, 8, 13, 19, 6\}$ with mean $\hat{\lambda}_C = 10$. The resulting 90% prediction interval of this guestimate displayed in Figure 8a show that the empirical L-function coincide with the prediction interval for distances lower than approximately $x \approx 0.17$. For greater distances, the prediction interval show too high probability for the occurrence of a daughter before it again encapsulates the empirical L-function for distances greater than $x \approx 0.4$. Also, the prediction interval width is small for short distances, but becomes larger as the distance increases.

For the next iteration we used a built in function, `kppm`, in R to generate values for our parameters. This method yielded the estimated parent intensity $\lambda_n = 23.5$ and daughter intensity per parent $\lambda_C \approx 2.63$ with variance $\sigma_c^2 \approx 0.0022$. We test the model by running 100 simulations drawn from the Neyman-Scott process using these parameters. The prediction intervals of the simulations are shown in Figure 8b along with the empirical L-function. We see that the prediction interval has decreased, and it coincide with the empirical L-function. For this iteration we did not take into account that parents outside the considered window may generate daughters inside. Thus, for the next iteration we extend the window by a distance d in all directions such that $W \subset \tilde{W} = [0 - d, 1 + d] \times [-1 - d, 0 + d]$. For the extended window we get a parent intensity $\lambda_N = \lambda_n\nu(\tilde{W}) = 23.5 \cdot 2.25 = 52.875$. In Figure 8c we see the prediction interval of 100 simulations from this method having extended the window by $d = 0.5$ and using the same parameters that we got from the previous iteration. We see that by extending the window we have further reduced the prediction interval from the previous iteration.

In Figure 9 we see the redwood tree data set along with three different realizations from the guestimated Neyman-Scott model. Realization 1 and two seem to have less clustering than that of the real dataset, while realization three bares more ressemblance to the real data. Because of the diversities between realization 1, 2 and 3 it is difficult to make model improvements by adjusting the three parameters λ_n, λ_C and σ_c.

Problem 4: Repulsive point processes

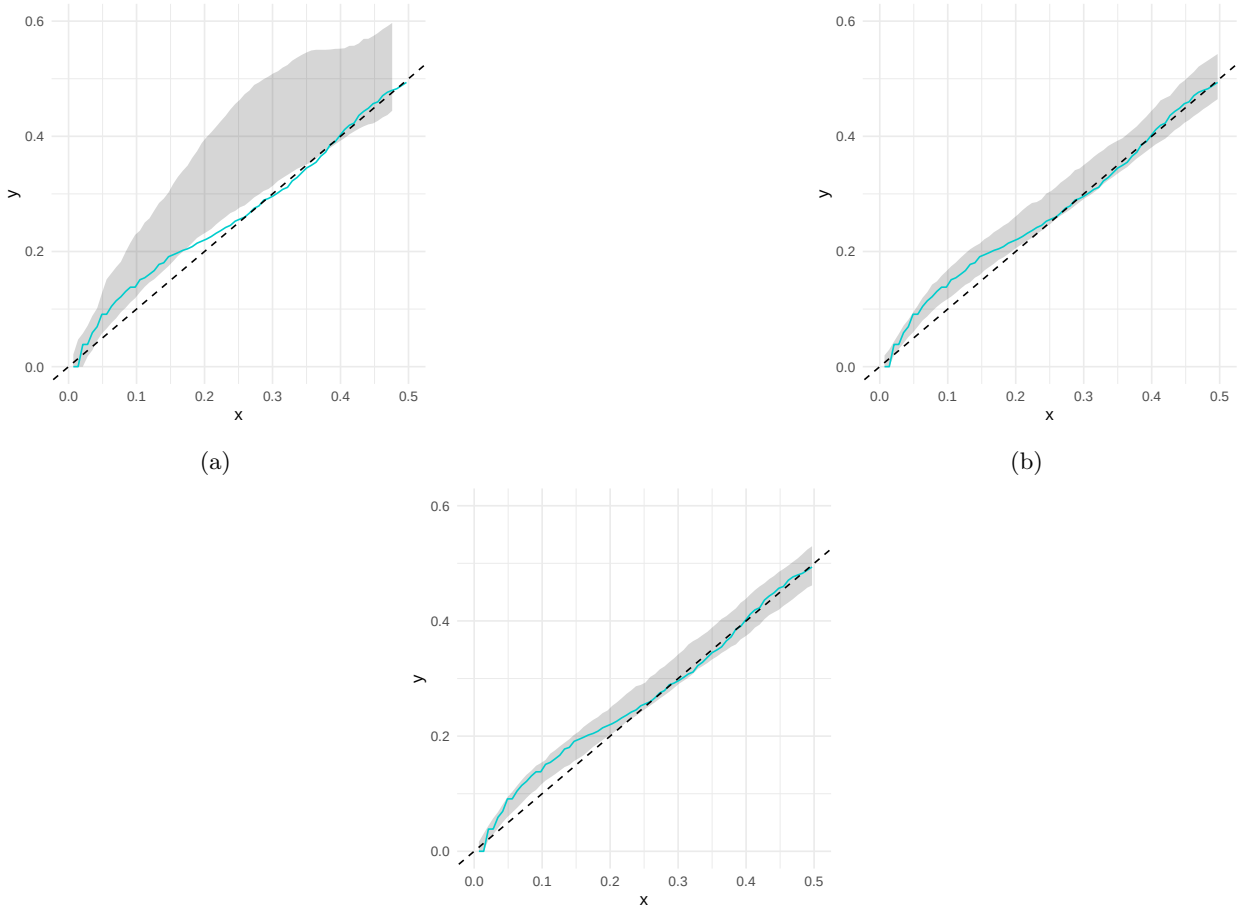
In this section we will take another look at the biological cell data set listed in Problem 1. We plan to model the point pattern using a Strauss process with a fixed number of points, n, and pair potential function

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

The full set of model parameters can be collected in the vector $\boldsymbol{\theta} = (n, \beta, r_0, W)^{\mathrm{T}}$. The parameter n is a the fixed number of points, i.e. we will not consider the situation with a random number of points. The parameter β gives the potential between a pair of points located at, say, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq r_0$. If $\beta > 0$, points will 'tend' to be farther away than a distance of r_0 form each other (i.e. repulsive behavior), since they want to minimize the potential function. W is the observation window and can be specified by the coordinates of its 4 corners.

Since we will use a bounded observation window, $W \subset \mathbb{R}^2$, we may run into some border problems/effects. One such effect is that the intensity of points will be higher along the borders, and at its highest in the corners of W, since there is less area for other potential points to interact with and be repelled by. This effect has the consequence that there will be a 'strip' with lower point intensity within the interaction range of the edge.

We will now guess parameter values based on the cell data set, which is displayed in Figure 3a. The parameter n is given as the number of points in the data set, and W is also given. We take the first guess



(a)

(b)

(c)

Figure 8: 90% prediction intervals for 100 simulated realizations from three different model fits along with the empirical L-function.

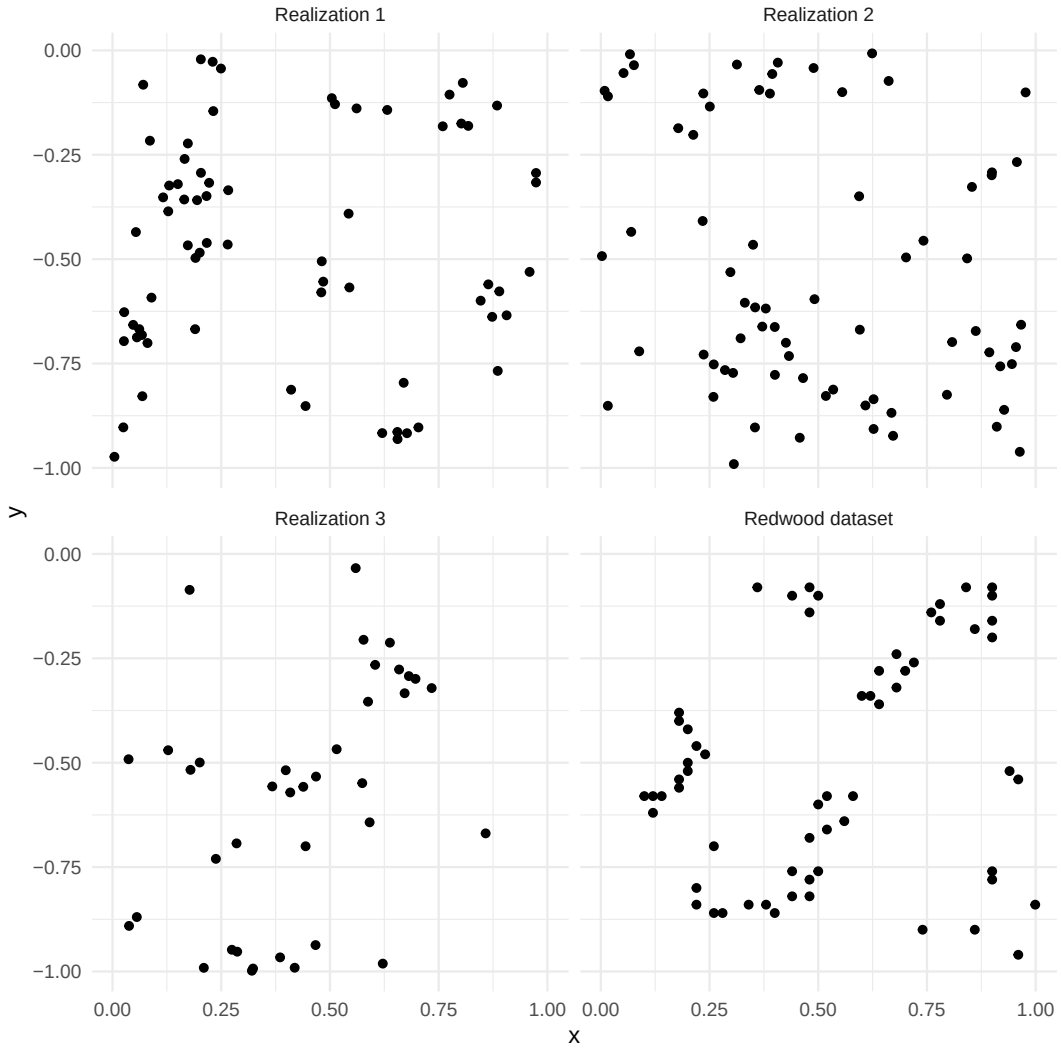


Figure 9: The point pattern of the redwood tree dataset (bottom right) and three different realizations of the Neyman-Scott process with parent intensity $\lambda_n = 23.5$, daughter intensity per parent $\lambda_C \approx 2.63$ with variance $\sigma_c^2 \approx 0.0022$ on the extended window with $d = 0.5$.

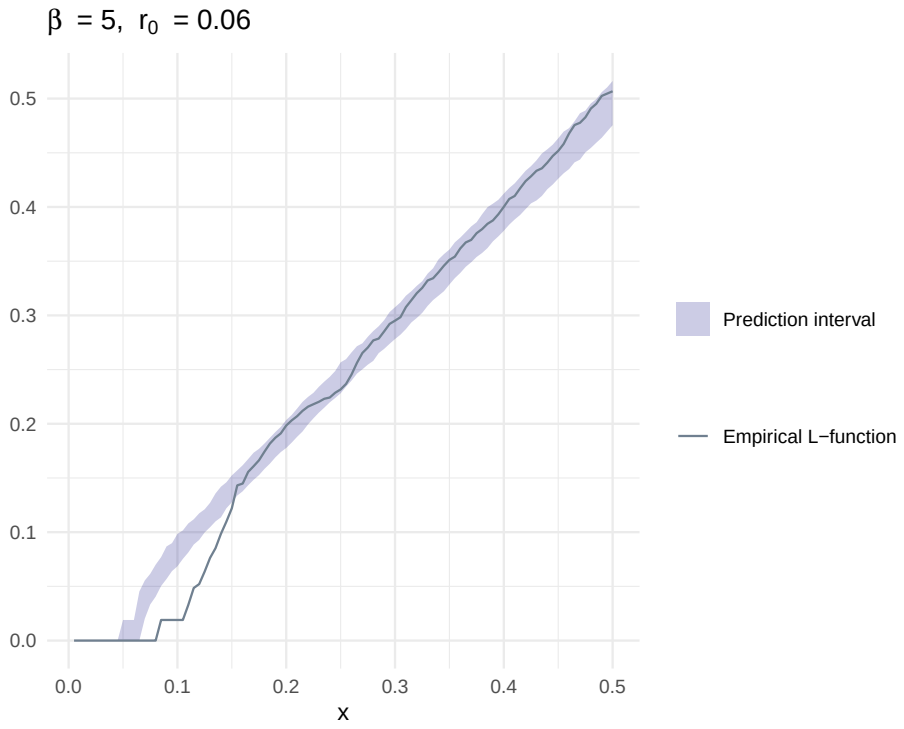


Figure 10: The empirical L-function of the biological cells data set and 90 % prediction intervals for the L-function of a Strauss process with $\beta = 5$ and $r_0 = 0.06$ arising from 100 realization.

of the inhibition distance to be $r_0 = 0.06$, which is approximately the half of the average distance between all points. We set $\beta = 5$ as our initial guess.

We use the function `Strauss` from the `R` package `spatial` to simulate the point process. Figure 10 displays the 90% prediction interval of the empirical L-function resulting from 100 realizations of the Strauss process, along with the empirical L-function of the cells data set.

We iterate our guess, and after some different tries, we end up with $\beta = 6$ and $r_0 = 0.1$. Figure 11 displays the resulting 90% prediction interval of the empirical L-function resulting from 100 realizations of the Strauss process, along with the empirical L-function of the cells data set. We observe that the fit is still not perfect.

We plot the point pattern from the cells data set together with 3 realization from the Strauss process with our last 'guestimates', shown in Figure 12. There are clearly some areas in the middle with fewer points in the Strauss process compared to the cells data set. We also observe that the corners nearly always contain points for all realizations, in contrast to the cells data set, where there is no points in the corners. This can be related to the effects of a bounded observation window discussed earlier. To make the model fit the data better, one could perhaps introduce a clustering effect in addition to the repulsion, so the points will not tend to be farther away from each other than a specified distance. This could perhaps mitigate the gap-formations in the middle which we observe for the Strauss process. Alternatively, one could maybe add a repulsion effect in the corners.

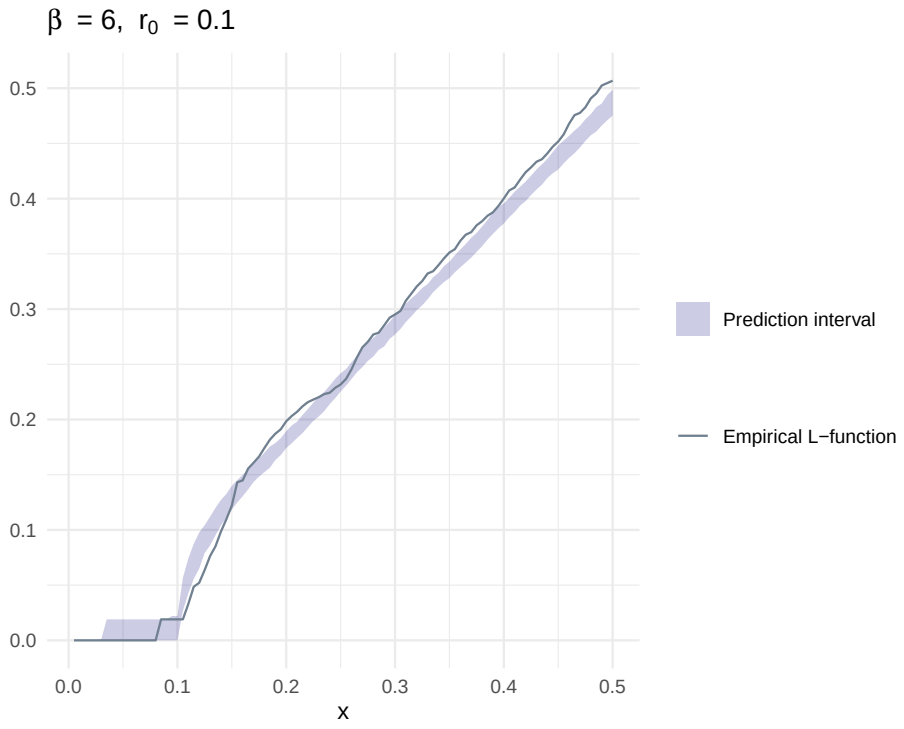


Figure 11: The empirical L-function of the biological cells data set and 90 % prediction intervals for the L-function of a Strauss process with $\beta = 6$ and $r_0 = 0.1$ arising from 100 realization.

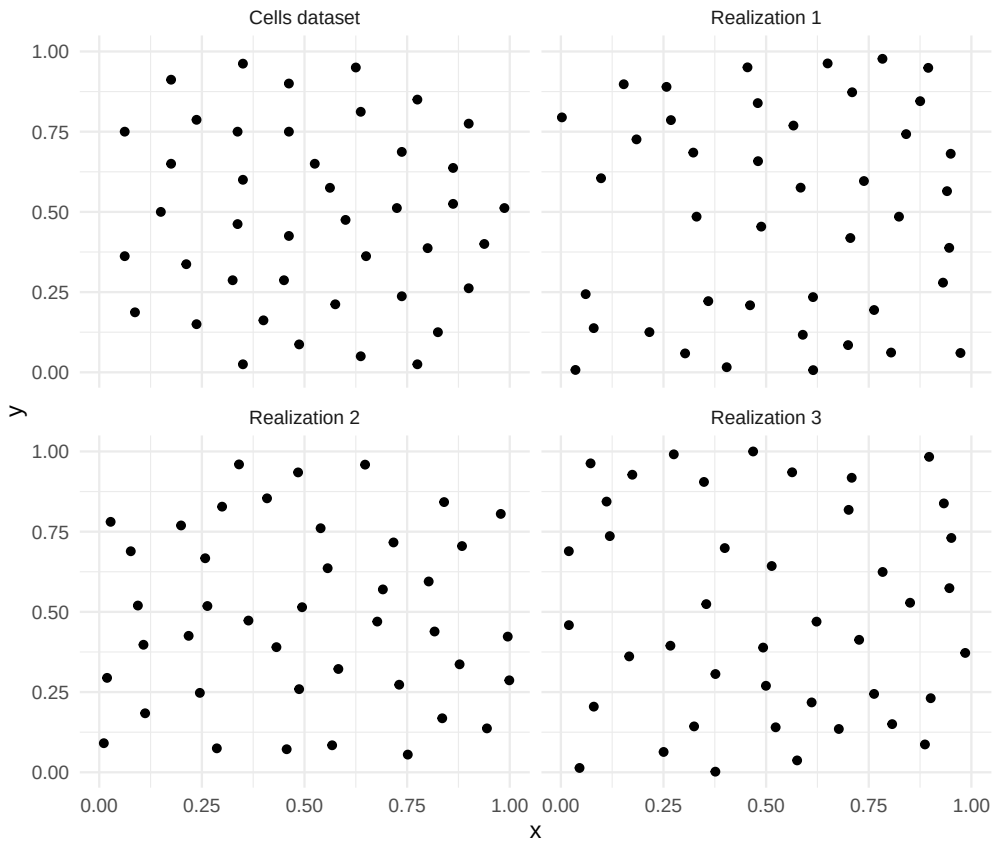


Figure 12: The point pattern of the biological cells data set (upper left) and three realizations of the Strauss process with $\beta = 6$ and $r_0 = 0.1$.