Rodrigo Arriaza
Arvin Rastegar
Johanna Weiss

Project Spatial Point Patterns

Spatial Epidemiology
January 11, 2022

Question 1

1. The data consists of the boundary coordinates of an area of study and the coordinates of bird nest observations. In Figure 1 the observations are plotted, using the area coordinates as the outline of the area. It can be seen that the outline of the study location is a lengthy area, with a lot of observed nests in the southern part and fewer observations in the middle and northern part.

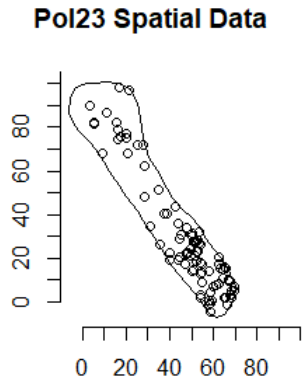


Figure 1: Plotted nest observations

2. The data contains 94 nest observations within the ranges of -6.40 and 69.97 units for the x-coordinate and -7.06 and 100.87 units for the y-coordinate. The observations are bounded by a polygon with 47 vertices. The overall area of study has a size of 2480.82 square units, which means that the intensity is 0.0379 observations per square unit. A Gaussian approximation of the density with $\sigma = 1$ and $\sigma = 5$ of the observations can be seen in Figure 2.

3. The data used again consist of location coordinates of an area of study and coordinates of observed nests in that area. The observations Additionally, there is information on the ‘“time to nest” for each of the observations. The values of “time to nest” ranges from 1 to 26 with a mean of 12.46. The observations with respect to their “time to nest” are plotted in Figure 3. It can be seen that all nests were exclusively observed in the northern part of the study area, with clusters in the western and eastern borders. However, it is not obvious if there is a relation between the time to nest and the location.

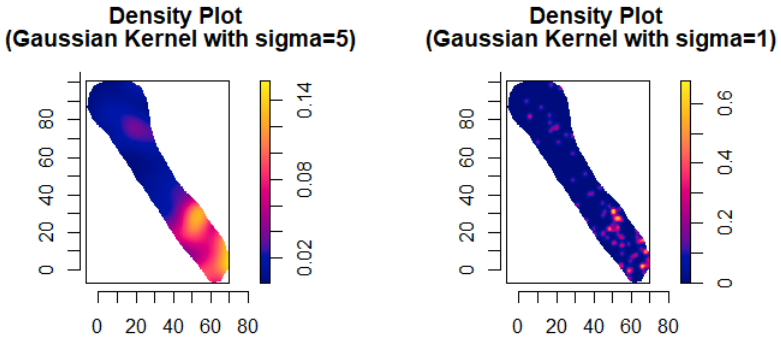


Figure 2: Plotted nest observations

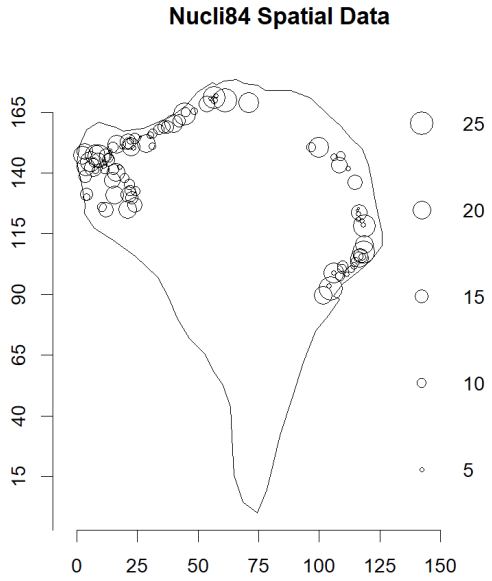


Figure 3: Plotted nest observations

4. The data contains 104 observations which are bounded by a polygon with 60 vertices. The data is shifted by the number of the smallest coordinate in both directions, after which they range between 0 and 125.79 in the x-direction and 0, 178.38 in the y-direction. In total, the area of study has a size of 11231.3 square units and the average intensity of nests is 0.0092 observations per square unit. A Gaussian approximation with $\sigma = 10$ and $\sigma = 5$ of the density of observations can be seen in Figure 4. It can clearly be seen that the observations cluster along the western and eastern border, with

no observations in the center or the southern part of the area of study.

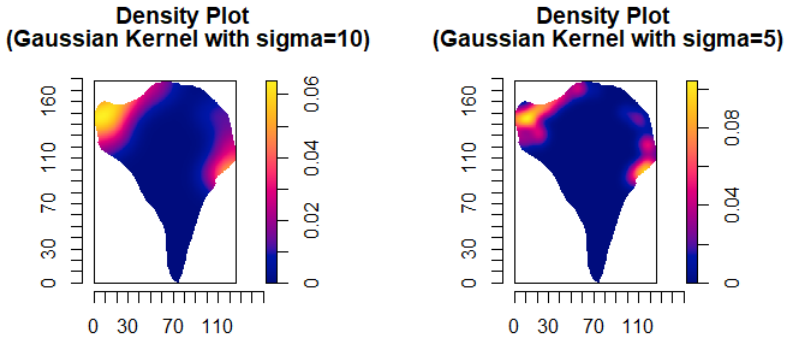


Figure 4: Plotted nest observations

Question 2

1. The data from Question 1.1 (seen in Figure 1) is used and it is assumed that the data follows a homogeneous poisson process. With this assumption a point estimate for the intensity of observations per square units can be done, which is the average observations per square unit, namely 0.0379. The corresponding confidence interval for the estimate is [0.0310, 0.0464].

2. To assess if the spatial is completely spatial random the quadrat test is performed. Dividing the area of study in quadrats and counting the observations gives already an indication that the data is not randomly distributed, as can be seen in Figure 5. Performing the Chi-squared quadrat test with 3 degrees of freedom and a test statistic of 54.32, gives a p-value of <0.001. The corresponding hypothesis are:

   H_0: the process is completely spatial random.
   H_1: the process is not spatially random

   Therefore, one can reject the H_0 hypothesis and conclude that the spatial process is not completely spatially random. The test is repeated with 2000 Monte Carlo simulations of the data to consider a larger amount of observations, which gives similar statistics and leads to the same conclusion.

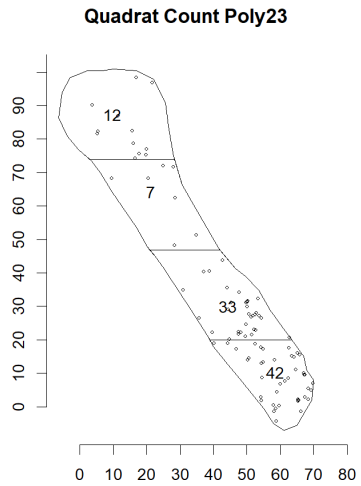


Figure 5: Number of Observations in Quadrats

3. Fitting an inhomogeneous poisson model to the data, the parameters and coefficients in Table 1 are obtained. The plot of the fitted trend and estimated standard error or the model can be see in Figure 6.

parameter	estimate	standard error	lower bound (0.95-CI)	upper bound (0.95-CI)
x	0.0616	0.0157	0.0308	0.0924
y	0.0100	0.0101	-0.0097	0.0298

Table 1

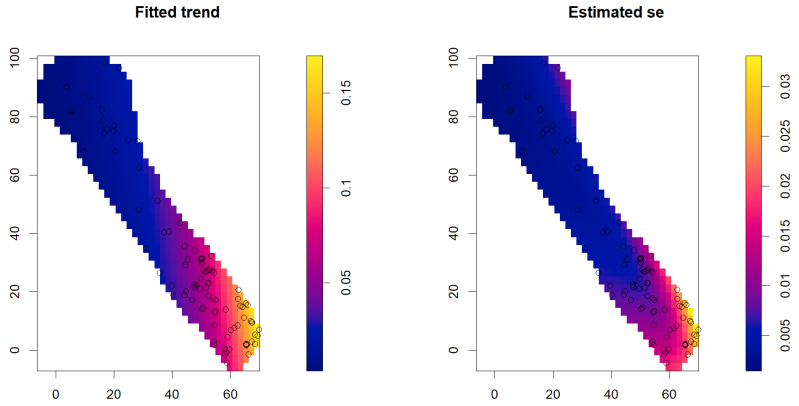


Figure 6: Fitted inhomogeneous poisson model

4. In order to test the goodness of fit of the inhomogeneous model, the Spatial Kolmogorov-Smirnov test is performed for both dimensions. The following hypotheses are used:

   H_0: the variable follows a nonhomogeneous Poisson process
   H_A: the process does not follow a nonhomogeneous Poisson process

   For both dimensions the H_0 hypothesis cannot be rejected with a significance level of α=0.05, as can be seen in Table 2, so the data might follow a nonhomogeneous Poisson process. The observed data on comparison with the expected data of a nonhomogeneous Poisson process for both variables can be found in Figure 7.

dimension	test statistic D	p-value
x	0.1019	0.265
y	0.074869	0.6402

Table 2

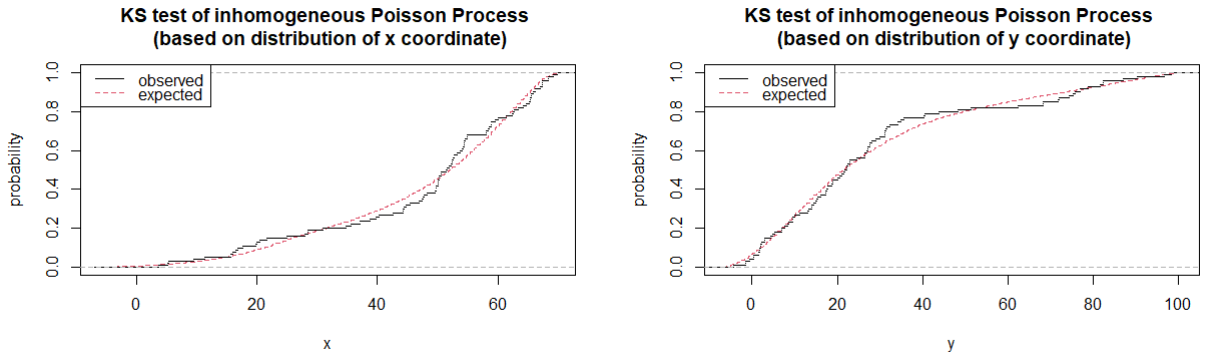


Figure 7: Fitted inhomogeneous poisson model

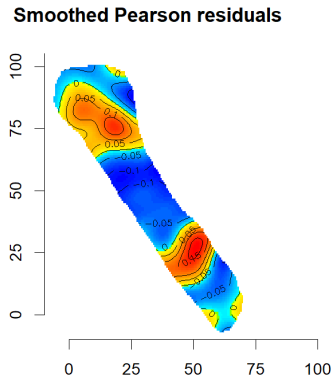


Figure 8: residuals of predicted model

5. As a last measure to check the validity of the inhomogeneous Poisson model, the residuals obtained from the model are looked at. The pearson residuals of the model can be seen in Figure 8. It can be seen that in the top region and the bottom region of the study area the residuals are positive, whereas in the middle region they are lower. Additionally, the lurking plot for both dimensions (Figure 9 is looked at. The lurking plot is helpful to analyze the dependence between the residuals and covariates.

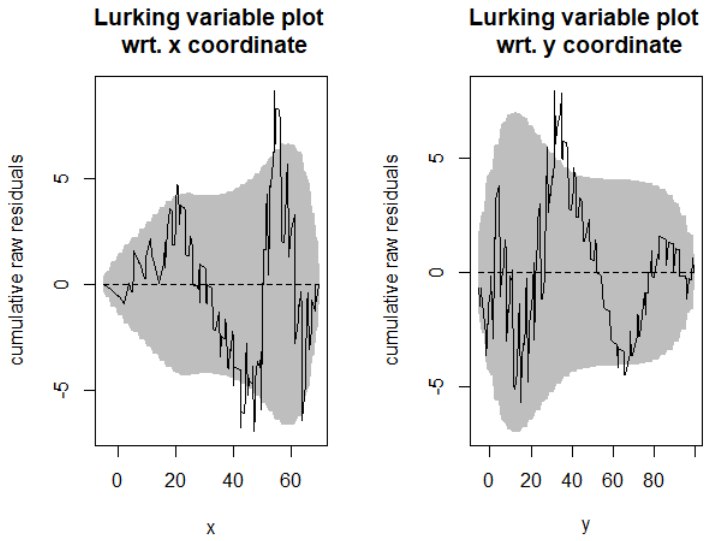


Figure 9: Lurking plots for x and y dimension

Question 3

1. Using the nest data from "nucli 23" we explore the pattern of interaction. In order to visualize the patterns we use three methods.

   Firstly, empty space distances, the distance from a fixed location to the nearest point in a point pattern, in r it is computed by the Fest command.

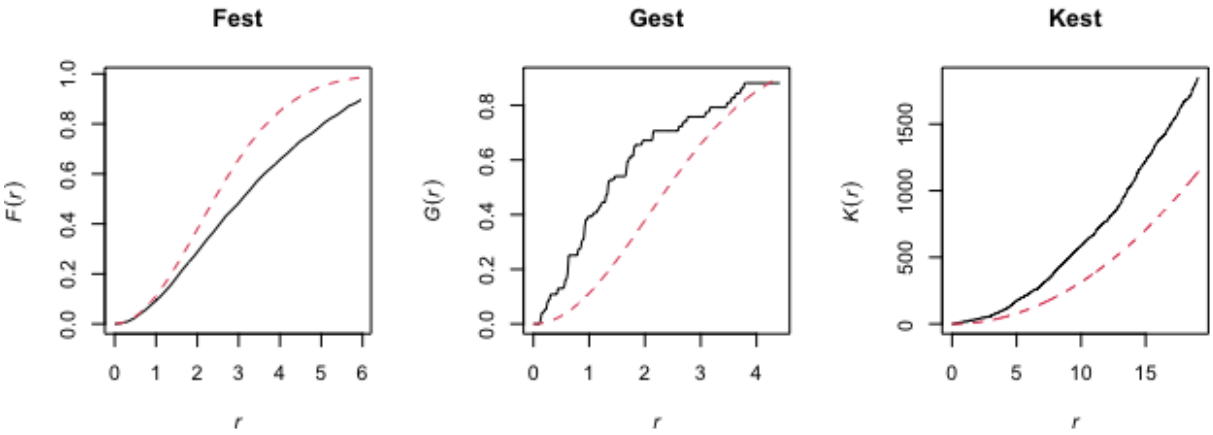
   Because we specify the correction = "best" inside the function, we only keep the best choice of edge corrections in this function. In this case the Kaplan-Meier estimation correction (km) is the best. We see that for the best edge corrected estimate, $\hat{F}_{km}(r) < F_{pois}(r)$, which suggests a clustered pattern.

   Secondly, we observe the nearest neighbor distances.

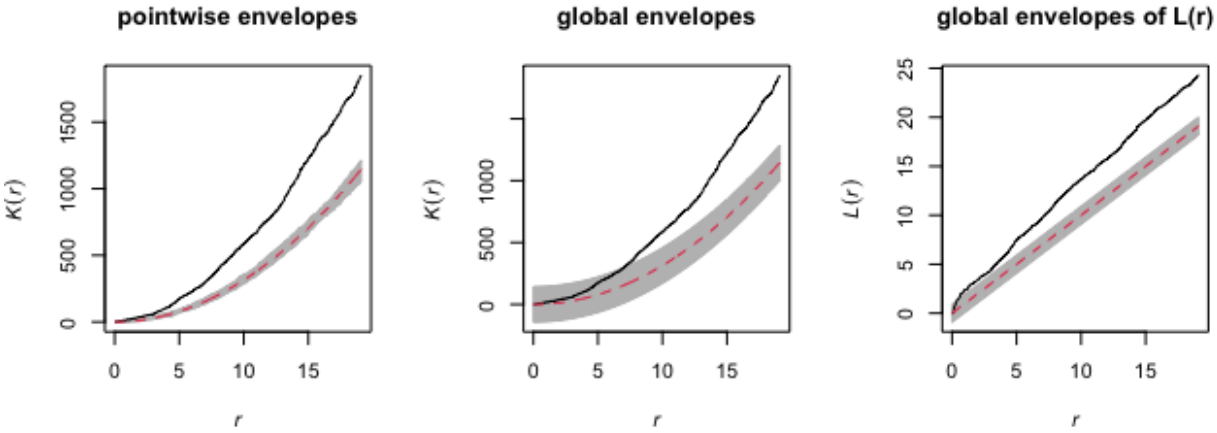
   The interpretation of $\hat{G}(r)$ is the opposite of $\hat{F}(r)$. Values $\hat{G}(r) > G_{pois}(r)$ suggest that nearest-neighbor distances in the point pattern are shorter than for a Poisson process, suggesting a clustered pattern, while values $\hat{G}(r) < G_pois(r)$ suggest a regular pattern(random). The best correction here is again the Kaplan-Meier correction. Again, by default, the graph includes three estimates which we keep only the best one. For a clustered pattern, observed locations should be closer to each other than expected under CSR. The best estimate is $\hat{G}_{km}(r) > G_{pois}(r)$, thus suggesting a clustered pattern.

Project Spatial Point Patterns

Finally, we plot the pairwise distances. We compare the estimate $\hat{K}(r)$ with the Poisson K function. Values $\hat{K}(r) > \pi r^2$ suggest clustering, while $\hat{K}(r) < \pi r^2$ suggests a regular pattern. Values of $\hat{K}_{iso}(r) > K_{pois}(r)$, by keeping the best correction again, suggest clustering.



Simulation Envelopes: In the previous sections, we examined several distance measures to judge whether a point pattern data set is completely random. The general procedure was to compare the observed point pattern with a completely random pattern, the Poisson point process. We generate M independent simulations of CSR inside the study map W. We compute the estimated K functions for each of these realizations,$\hat{K}_{(j)}(r)$ for j=1,2,3,...M and we obtain the point-wise upper and lower envelopes of these simulated curves to see the difference of our actual data to these theoretical points.



Here we have envelopes corresponding to critical values of 5%. This indicates the observations significantly differ from random or the theoretical K for the Poisson point process.

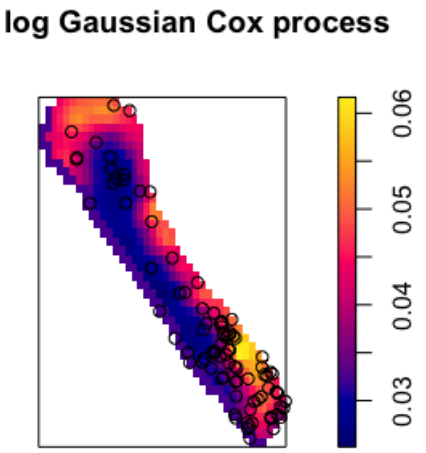
2. A log-Gaussian Cox process is fitted to our data.

```
---------- COX MODEL ----------
Model: log-Gaussian Cox process

        Covariance model: exponential
Fitted covariance parameters:
       var      scale
 0.9081747 12.5101730
```

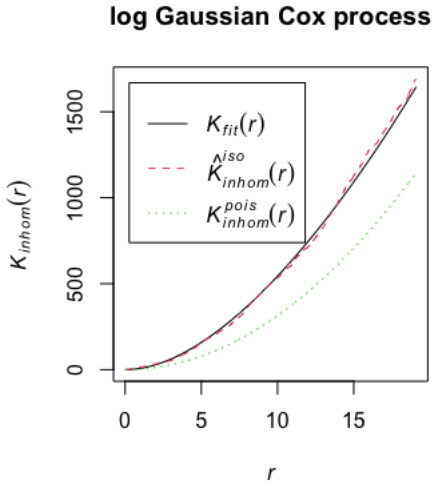
σ^2 is the variance of the log of the driving random intensity(spatial correlation). In some sense, this measures the strength of clustering.
Alpha is the correlation scale in the exponential covariance, $C(r) = \sigma^2 * \exp(-r/\alpha)$ of the driving random intensity. In some sense this measures the scale of clustering.
We can see that a cox process with the correlation range of 12.51 is physically meaningful, hence we have a strong sign of clustering. The intensity of clustering is shown

in the plot below.

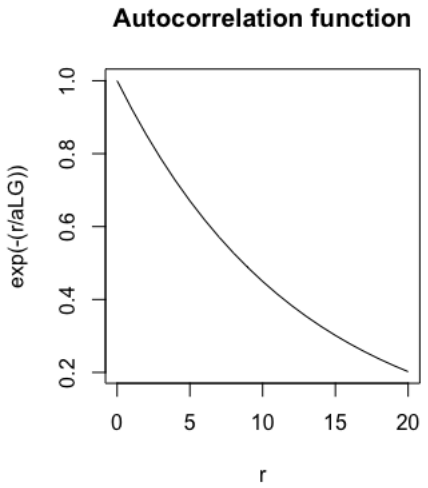


We see that the intensity is higher in the north and southeast of the plot, around 0.06, wherein the center area and southeast is nearly half and around 0.03.

Then we take a look at the edge corrections that are used to correct bias in the estimation of $\hat{K}_{inhom}(r)$.



We see that the K_{fit} is very near $\hat{K}^{iso}$ and very far away from the Poisson one which indicates clustering of the data.



In this graph, we can see the auto-correlation function which makes us understand the range of different auto-correlation in this model. At the radius of about 10, we observe the autocorrelation around 0.5, and after 20, it is almost 0.2.

3. In the last part we fit a Gibbs process model to our data, the area interaction model.

```
------ Interaction: ------

Interaction: Area-interaction process
Disc radius:     3.5
Fitted interaction parameter eta:         9.27025

Relevant coefficients:
Interaction
    2.22681
```

The printout for the area-interaction model uses the "scale-free" parameter eta defined by $\eta = \gamma^{\pi r^2}$. Values of η greater than 1 suggest clustering. Hence, we infer that we have clustered data here.

Plotting a fitted model generates a series of images and contour plots of the fitted first order term $\exp(\hat{\eta} \cdot S(u))$, the fitted conditional intensity $\lambda_{\hat{\theta}}(u, x)$ evaluated for the data pattern x. For Poisson models, the two plots are equivalent and give the fitted intensity function.

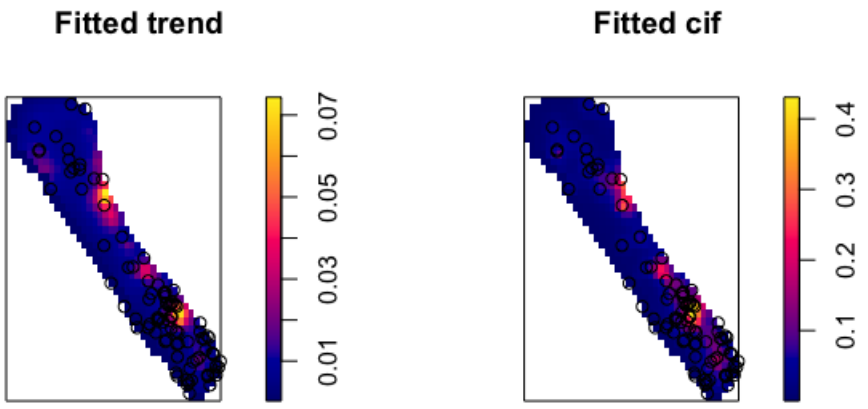


Figure 10: The plots for the Area-interaction model

We observe that the trend is stronger in the light sections of the plot of the fitted trend, and fitted conditional intensity almost shows similar behavior to the fitted trend plot.

In order to validate the models, for a fitted Gibbs process we need to simulate the model to create the critical envelopes. We do this for 95% critical values in here.
We observe here that the model is reasonably good and for most parts in the critical bands.

Residuals for Gibbs processes definition: The total residual in a region $B \subset R^2$ is defined as $R(B) = n(x \cap B) - \int_B \hat{\lambda}(u, x)du$ where again $n(x \cap B)$ is the observed number of points in the region B, and $\lambda(u, x)$ is the conditional intensity of the fitted model, evaluated for the data point pattern x. If the fitted model is correct, the residuals have a mean of zero. This definition is similar to the definition of residuals for Poisson processes except that the intensity $\hat{\lambda}(u)$ of the fitted Poisson process has been replaced by the conditional intensity $\hat{\lambda}(u, x)$ of the fitted Gibbs process evaluated for the data point pattern x.

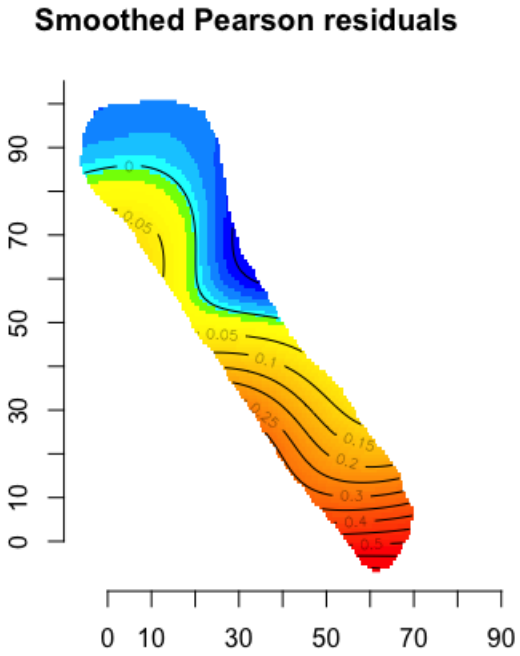


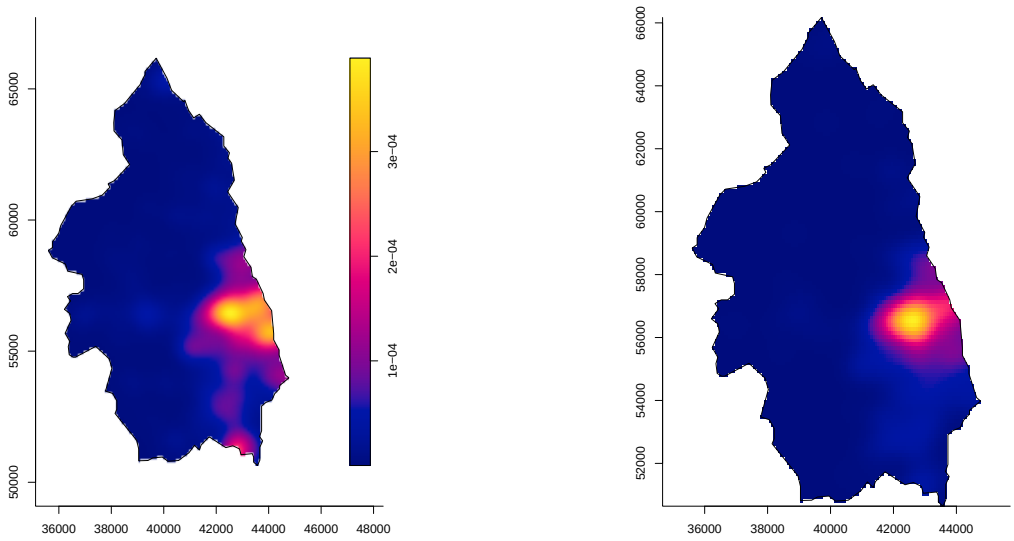
Figure 11: Pearson Residual

When we diagnose the model with Pearson residuals, we observe that residuals are low, especially in the northern area, and the model is a good fit.

Question 4

Assessment of risk variability

To calculate the spatial variation of the relative risk it is needed to analyze the ratio between the intensities of cases and controls. Therefore the data was converted into a ppp object using the polygon data to add a polygonal window and adding as factor the marks column from the pbc dataframe.



(a) Control intensity

(b) Cases Intensity

Figure 12: Control and cases intensity

As seen in Figure 12, it is noticeable that the majority of the cases are concentrated on the mid east region, but this is explainable by the fact that also the majority of the population resides in that area.

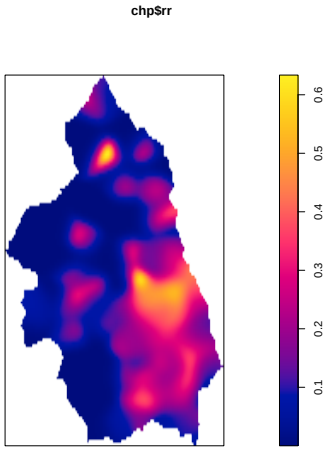


Figure 13: Relative Risk

Consequently when plotting the relative risk in Figure 13, it can be noticed that there are other regions with higher relative risk. For example in the northern part of the region, we see a high relative risk that was not noticeable in Figure 12b, which means that in this part there is a low number of cases but it is actually very high considering the small population.

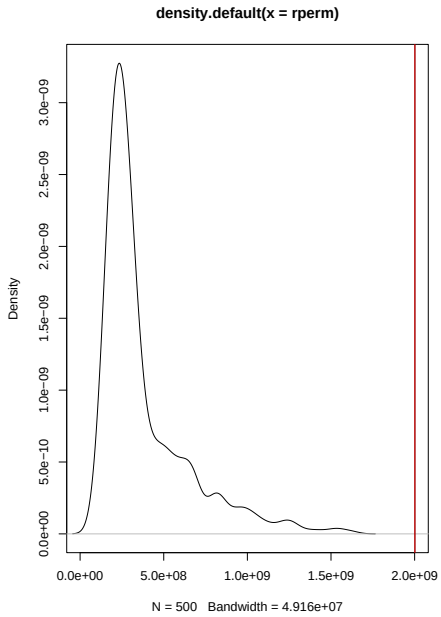
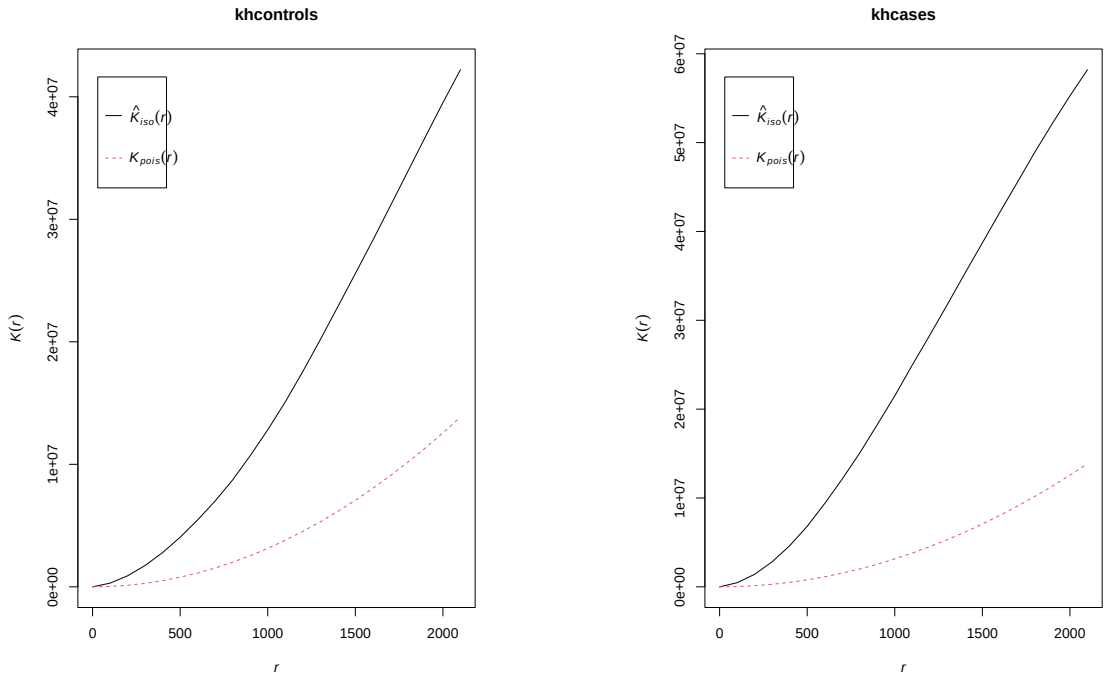


Figure 14: Permutation values vs rho ratio

Finally a permutation test is performed to analyze the spatial variation of relative risk, running 500 simulations and a p-value of 0.0280 is obtained. Thereby the null hypothesis of equal space distribution can be rejected with a significance level of 95%. Therefore we can conclude that the distribution of cases and controls is not equal among the entire region, i.e. there is a spatial variation. This can be seen graphically in Figure 14.

Assessment of interaction

The K function is defined so that the expected number of additional random points equals K(r), where r is a distance of a typical random point of the stationary process point X. The estimate of K is compared to the true value of K for a completely random poisson point process. If these two differ, the difference between the curves indicates spatial clustering.



(a) K(r) Control

(b) K(r) cases

Figure 15: Ripley's reduced second moment function K(r) for control and cases point patterns

As seen in 15, for larger values of r, the distance between the curves, both for the control and the cases, increases greatly. This would indicate that this spatial point process has spatial clustering since for all the interval $\hat{K} > K_{poisson}$, both for control and cases.

To further test the cases interaction a permutation test is performed using the following statistical test:

$$D = \int_A \frac{D(s)}{\sqrt{var(D(s)}}$$

where

$$D(s) = K_{cases} - K_{control}$$

As can be seen graphically in 16, the difference between the K-functions of cases and control is outside the interval, meaning again that the process has spatial clustering.

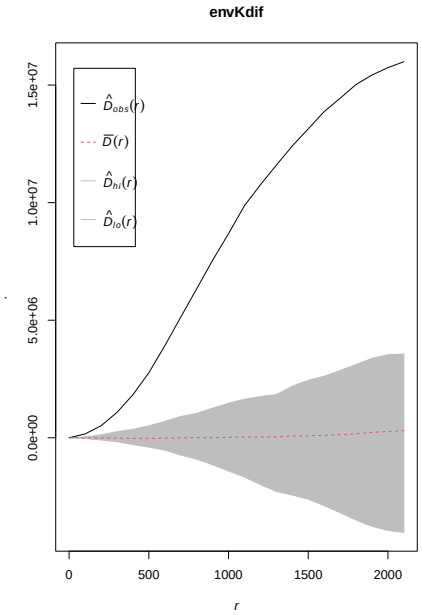


Figure 16: Envelopes for k-function difference vs r

Analytically the permutation yields to a p-value of 0.05, as seen in 17, by which we can conclude that there is spatial clustering, meaning that the occurrence of a case is not random and that the presence of a case increases the probability of other cases appearing nearby.

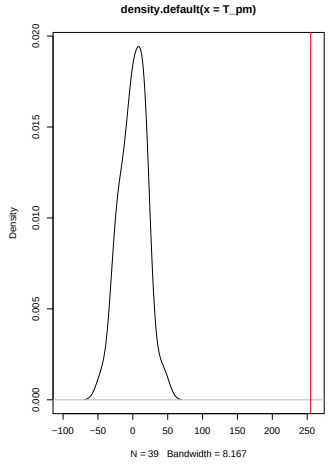


Figure 17: Monte Carlo test for spatial clustering