589Project

Nowshaba Durrani, Ricky Heinrich, Viji Rajagopalan

2023-04-09

Introduction

For our Data 589 project, we have selected Red Fox (Scientific Name - Vulpes Vulpes) to do the analysis. In the GBIF database they have approximately, 610,958+ georeferences records for this species around the world, however for this project we have selected to do the analysis of the occurance of Red Fox in BC only. So with the above function we have fetched the information for British Columbia only in 127 columns and 242 number of entries.

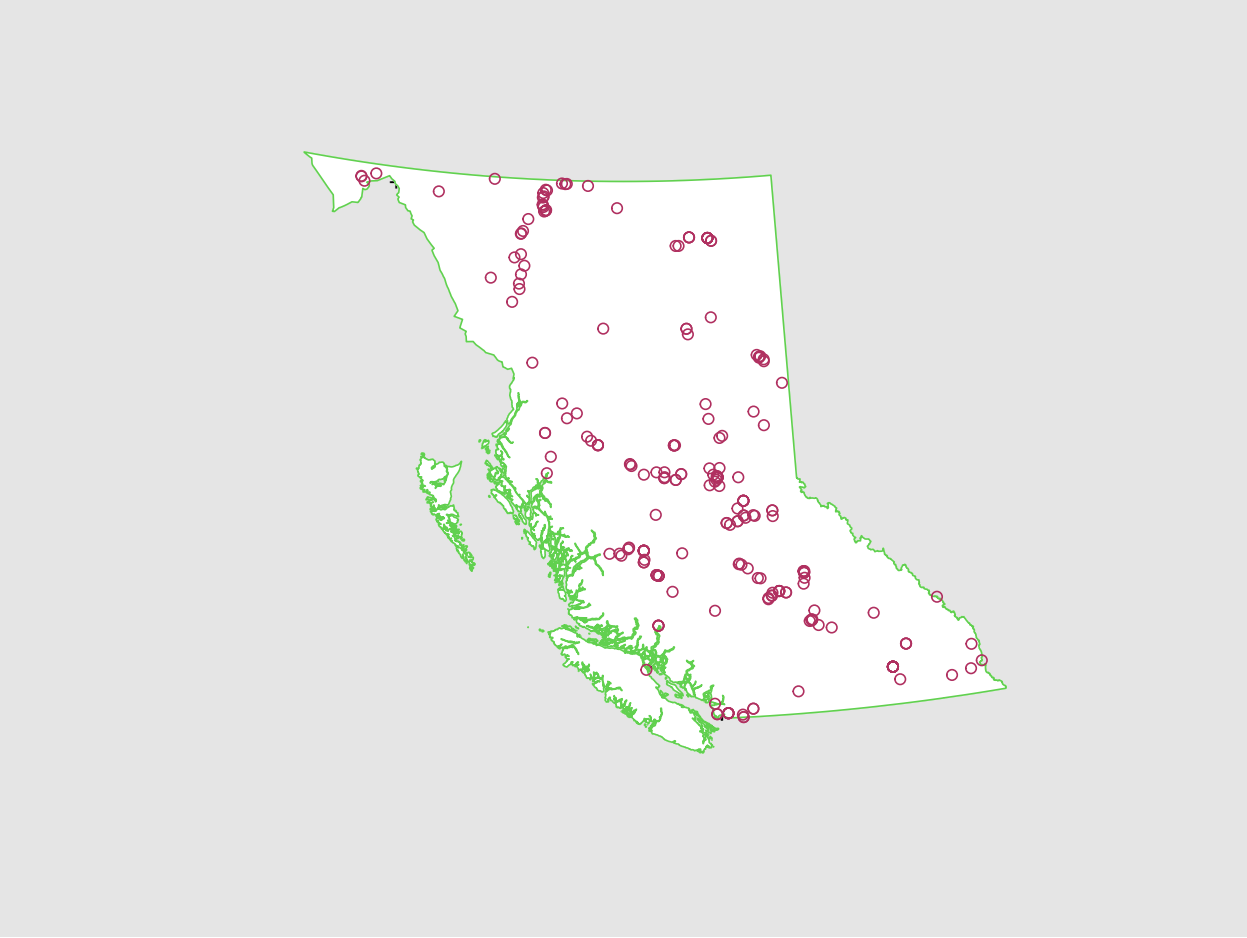


Figure 1: Occurence of Red Foxes in BC

Here we have plotted all the occurrences of Red Fox in the BC region and we can see that the species are scattered in the region specially in the upper and middle part of the province. Now we will be exploring what

is contributing to the occurrences of the species in the specific places based on various factors like elevation, close to water bodies, forests, human habitats, etc.

Methods:

Briefly describe the data and what variables are included. Provide a detailed description of the analytical workflow that was applied to the data, citing any relevant literature and statistical packages employed. There should be enough information that anyone can reproduce the workflow if they had access to the data. Length: As long as necessary.

The data comes from the Global Biodiversity Information Facility (GBIF) databases. We used the package `rgbif` to access the 'Vulpes Vulpes' data from R directly, sorting by instances occuring in BC. We've extracted the longitude and latitude data from this, and converted it appropriately using the `sp` package.

Our second source of data contained the BC Window object, as well as possible covariate data: elevation, forest cover, Human Footprint Index (HFI) and distance to water.

We used the package `spatstat` to build a `ppp` object with the converted coordinates of the Red Fox locations from the GBIF data and the window from our second data source.

To conduct first moment analysis, we used functions from the aforementioned `spatstat` package. We did a quadrat test as well as hotspot analysis to gain insight into the homogeneity assumption of the point process.

For second moment analysis, we looked into Ripley's K-function and pair correlation function using functions from `spatstat`. This provides us with insight into possible clustering tendencies of the point process.

Next we looked into the relationship with covariates.

First Moment Analysis

```
## [1] 2.509854e-10
```

Per the summary, Average intensity 5.063089e-10 points per square unit which is 0.0000000005063089 per square unit and this does not explain the observance of Vulpes Vulpes in a meaningful way.

Quadratcount:

5 by 5 and 10 by 10 - Both convey different view points on the intensity of the observance. According to plot 1, most of the Vulpes Vulpes are spotted in the South West areas around Vancouver.

The 10X10 figure shows the intensity is high in the coastal areas with higher density in the South West region.

Quadrat counting suggests varying intensity and to confirm that the variation is not due to chance alone, we conduct an objective test for spatial (in)homogeneity. We do a Chi-square test to validate if the deviations are significant.

```
##
##  Chi-squared test of CSR using quadrat counts
##
## data:
## X2 = 352.69, df = 63, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 64 tiles (irregular windows)
```

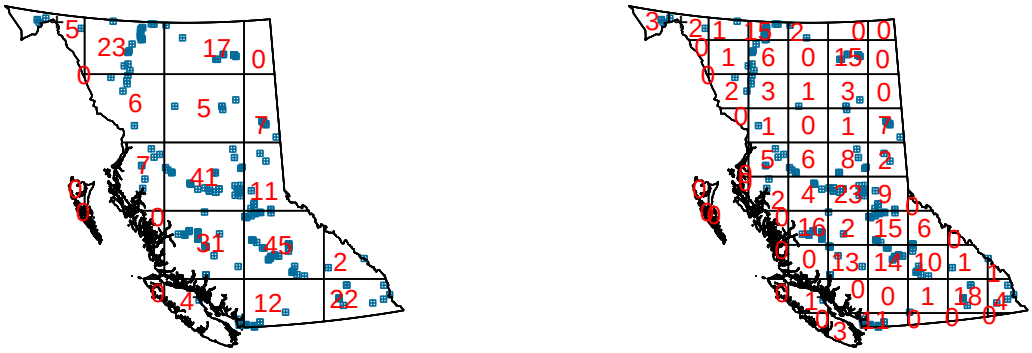
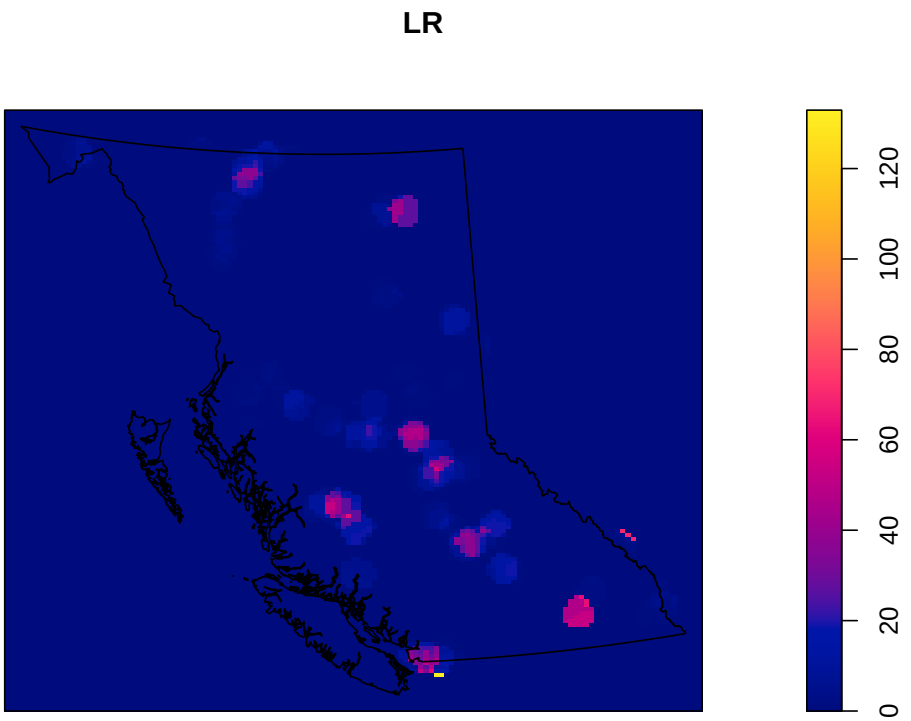


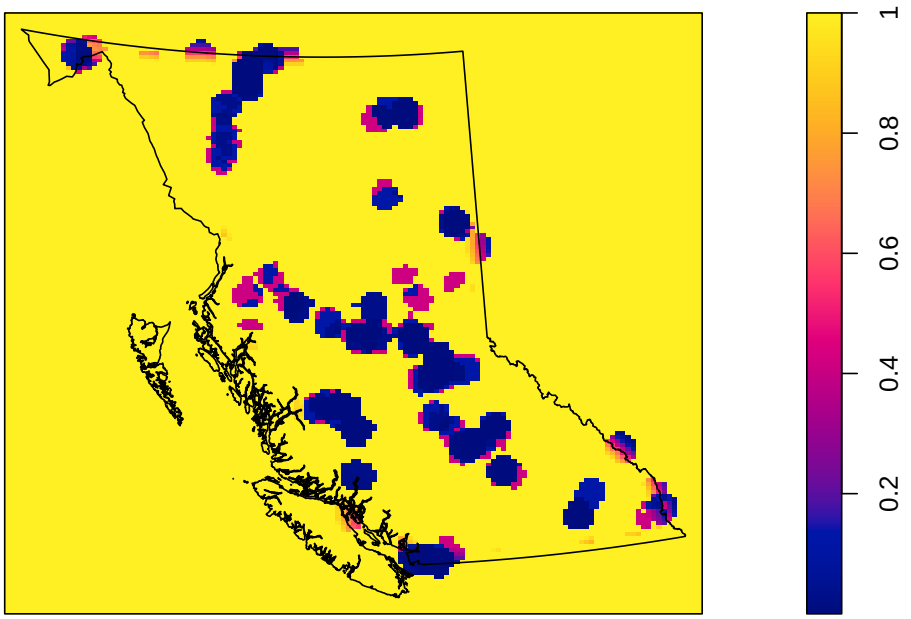
Figure 2: Quadrat counts of Red Fox occurences, left 5x5, right 10x10

The null hypothesis of the test suggests homogeneity in the process and as the p-value is very small, the null hypothesis is rejected and its confirmed there is significant deviation from homogeneity.

Hot spot analysis:

As the next step, we analyze for any hot spots in the south west coastal areas of BC.





Question: Do we need p -value intensity analysis? Also, is it possible to add the window for better observation window boundary (shape of BC)?

2nd Moment Analysis

Ripley's K function

Ripley's K-function provides information on whether there are significant deviations from independence between points.

Assuming homogeneity in the point process, we see that the actual occurence lines appear clustered, as the black line does not overlap with the expected red line confidence intervals (significance level of 0.05).

In Figure 3 we see that the effect appears significant. It is suspicious that the confidence bands increase a lot, (what's the explanation for that?)

However, we know from first moment analysis that the intensity does not seem homogenous. Using the Kinhom function ensures that we are not assuming the intensity is homogenous. From Figure 4, it seems like from the smaller numbers, ie smaller distances between points, there is 'evidence' of clustering, whereas there are funny things going on as distances increase. The deviations are still meaninful in the 'smaller' distances, suggesting that the relationship between points may be due to effects between points rather than relationship with covariates.

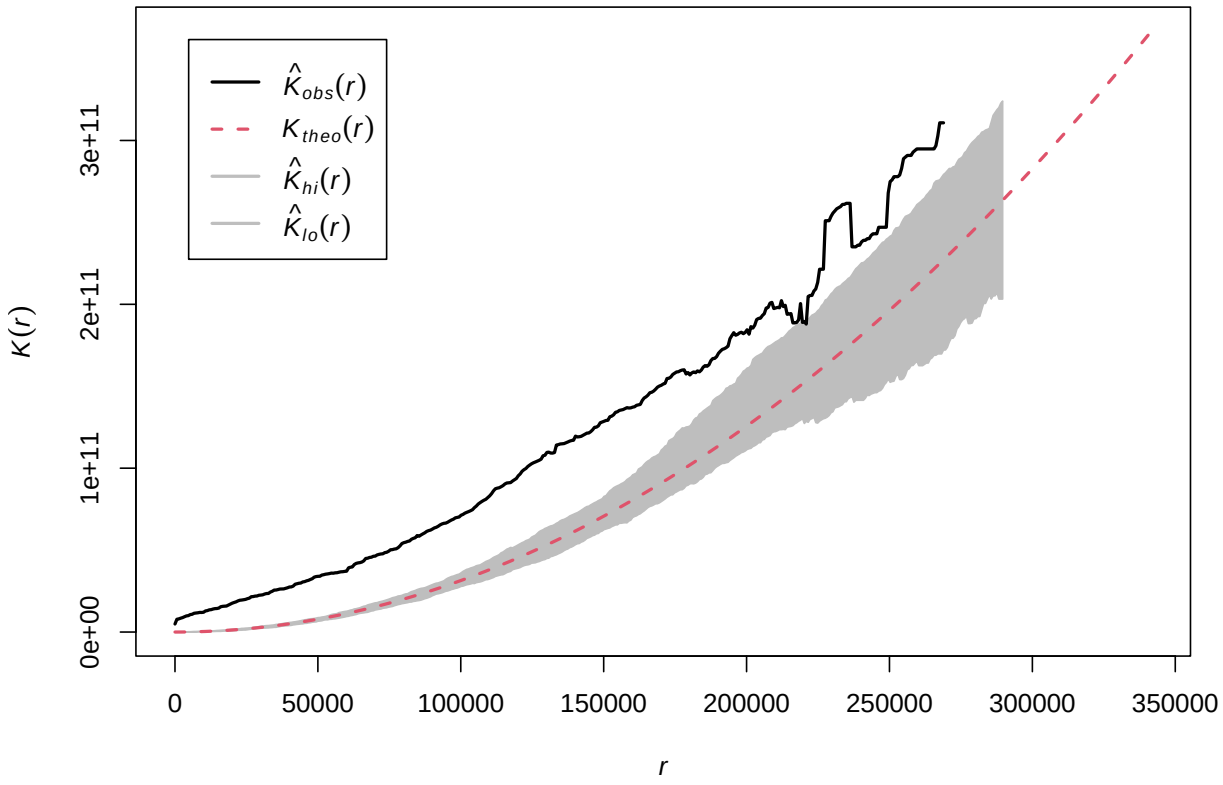


Figure 3: Ripley's K function with border correction assuming homogeneity

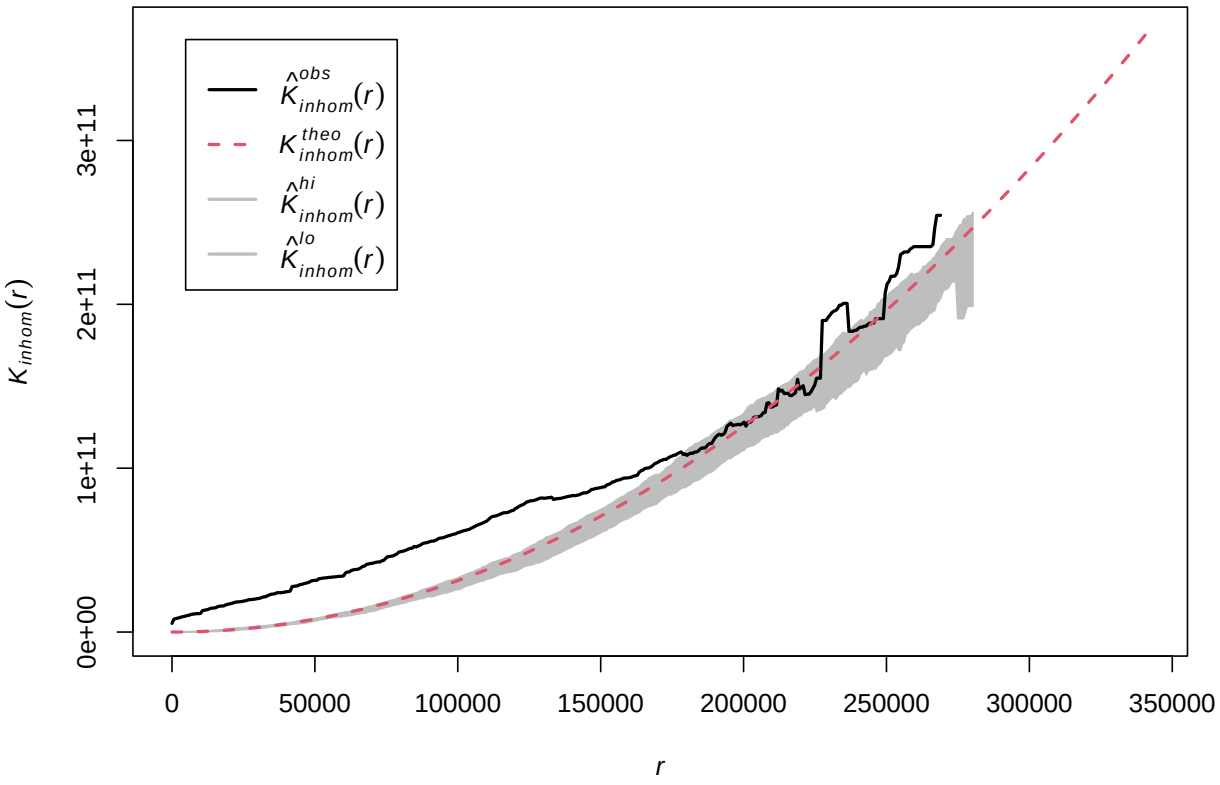
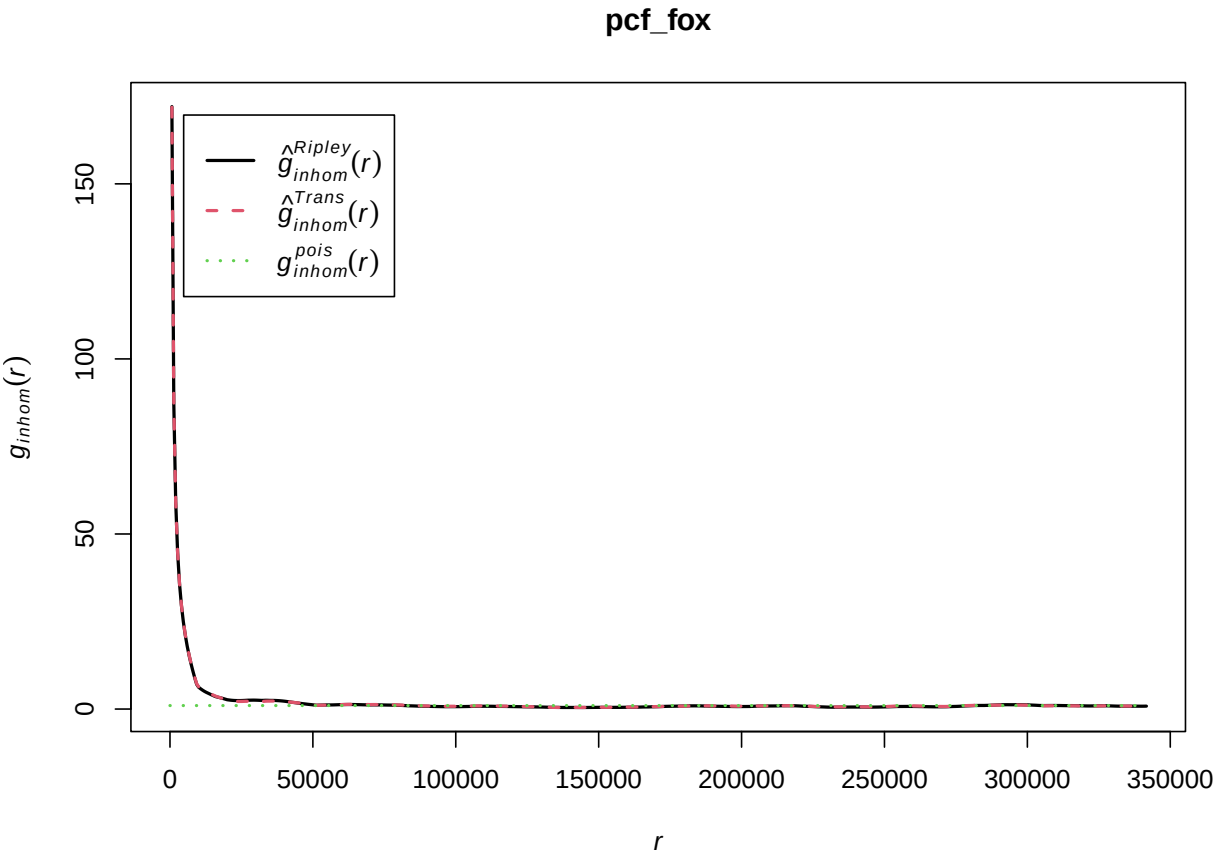
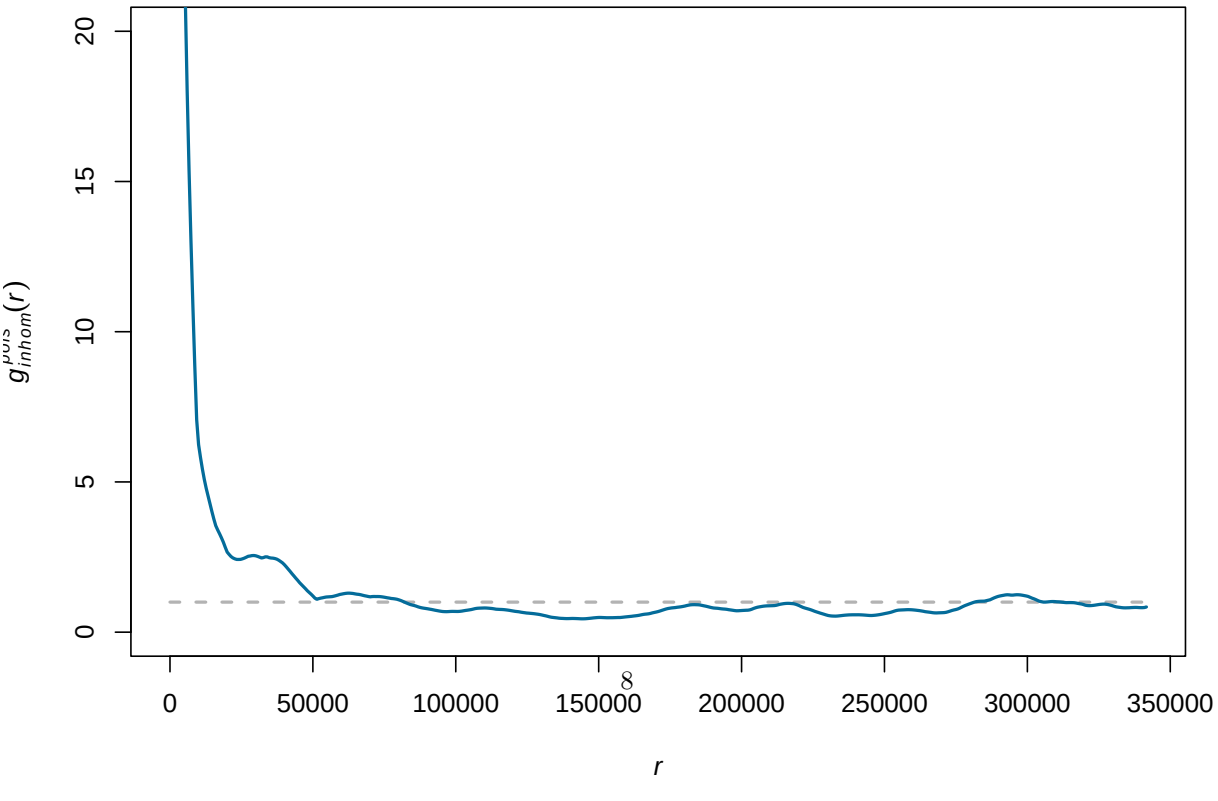


Figure 4: Ripley's K function with border correction assuming inhomogeneity

Pair correlation function





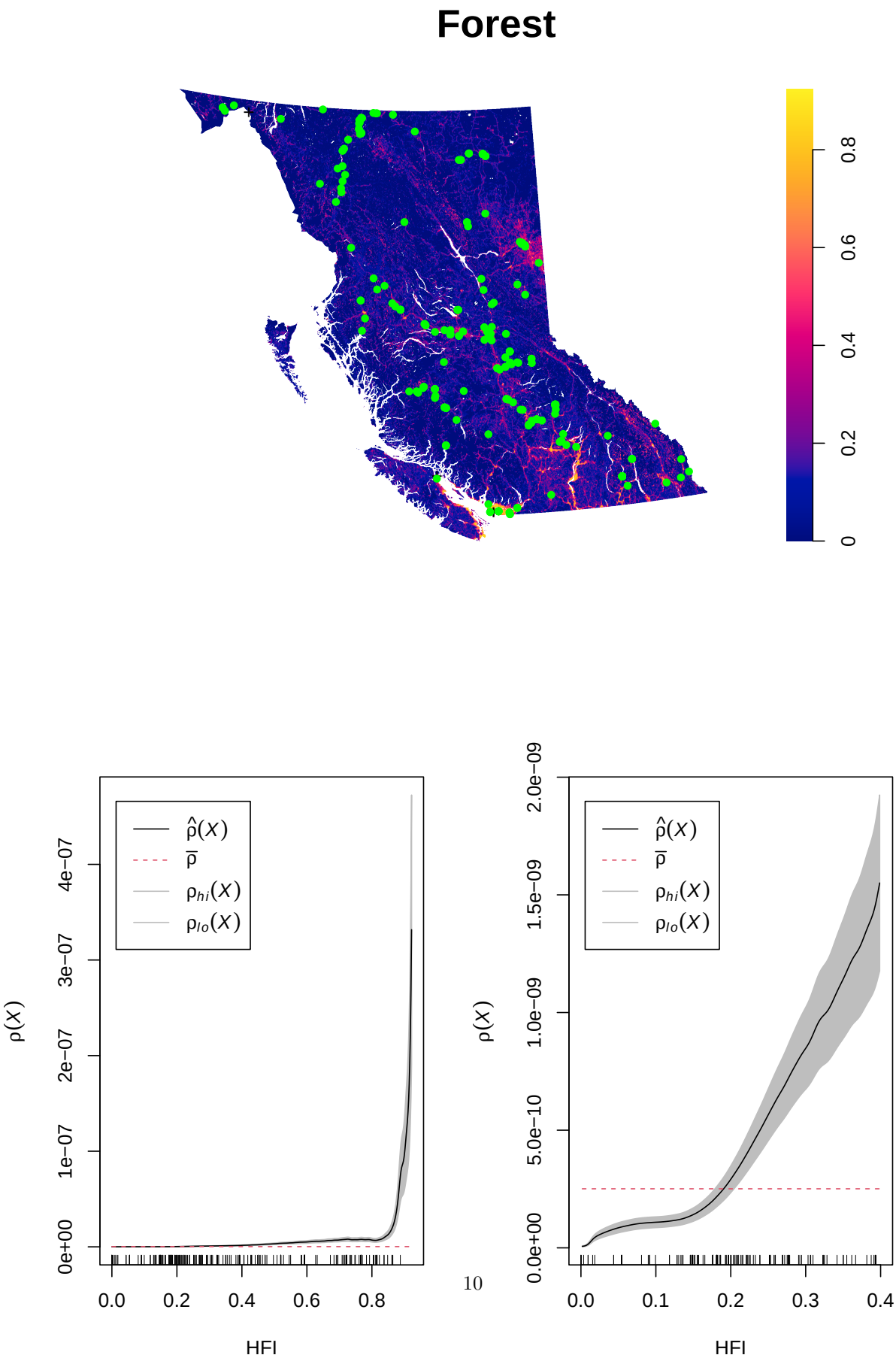
We observe that there seem to be evidence for clustering at smaller than 50 000, but after that it rides the y = 1 line slightly under.

We need to validate clustering effects after modeling the data with covariates.

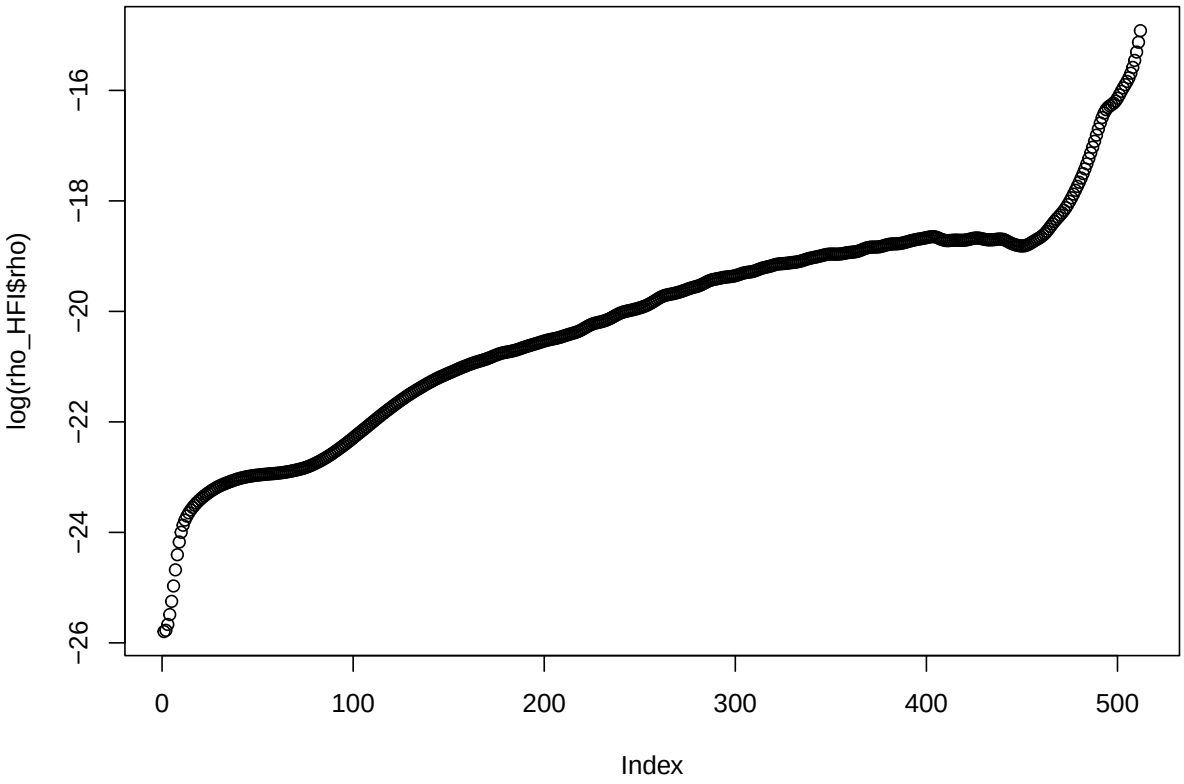
Relationship with Covariates

Our data includes 4 covariates we can explore: the elevation, the forest cover, the human footprint inventory (HFI), and the distance to water. Given our research questions, we will start with investigating the HFI and the forest cover.

HFI


Forest
0
0.2
0.4
0.6
0.8
ρ(X)
0e+00
1e−07
2e−07
3e−07
4e−07
0.0
0.2
0.4
0.6
0.8
HFI
$\hat{\rho}(X)$
$\bar{\rho}$
$\rho_{hi}(X)$
$\rho_{lo}(X)$
ρ(X)
0.0e+00
5.0e−10
1.0e−09
1.5e−09
2.0e−09
0.0
0.1
0.2
0.3
0.4
HFI
$\hat{\rho}(X)$
$\bar{\rho}$
$\rho_{hi}(X)$
$\rho_{lo}(X)$

In the first figure, we could be fooled into thinking that there is no relationship up to around HFI = 0.4, until which it seems like an exponentional relationship. However, zooming in from HFI 0 to 0.4, we see that the confidence bands don't intersect at all with the red line, which is the expected value given no relationship. This relationship appears non-linear and possibly exponential, where the greatest intensity of observed red foxes occurs at high HFIs. This relationship was expected, as our dataset is not exhaustive but rather is crowdsourced, and naturally foxes are more likely to be noticed by humans in spaces with higher HFIs.

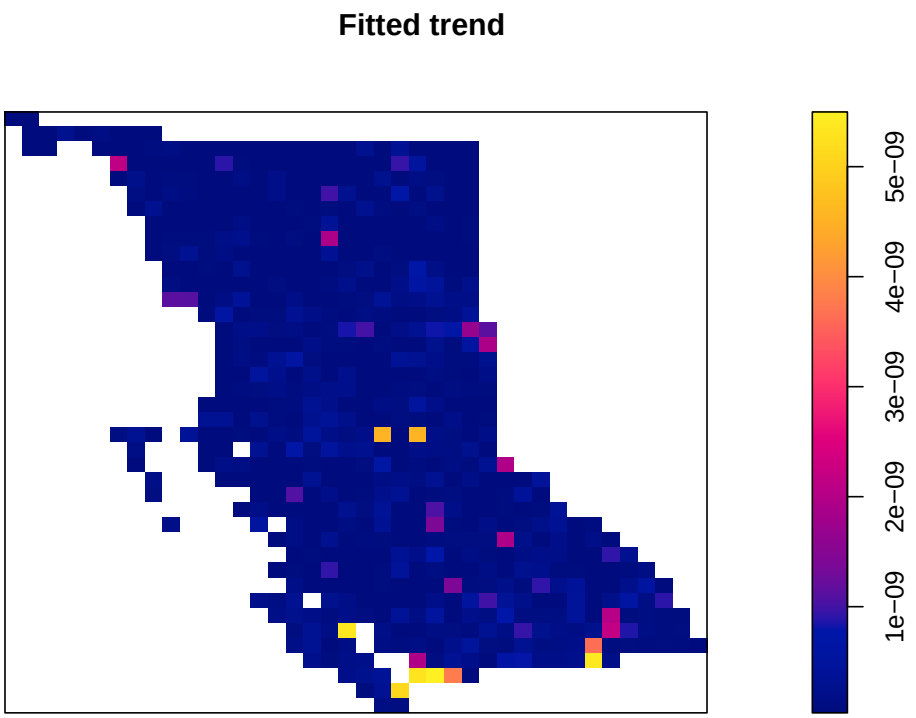


If we plot the log of the rho, we get a line that could be reasonably interpreted as linear.

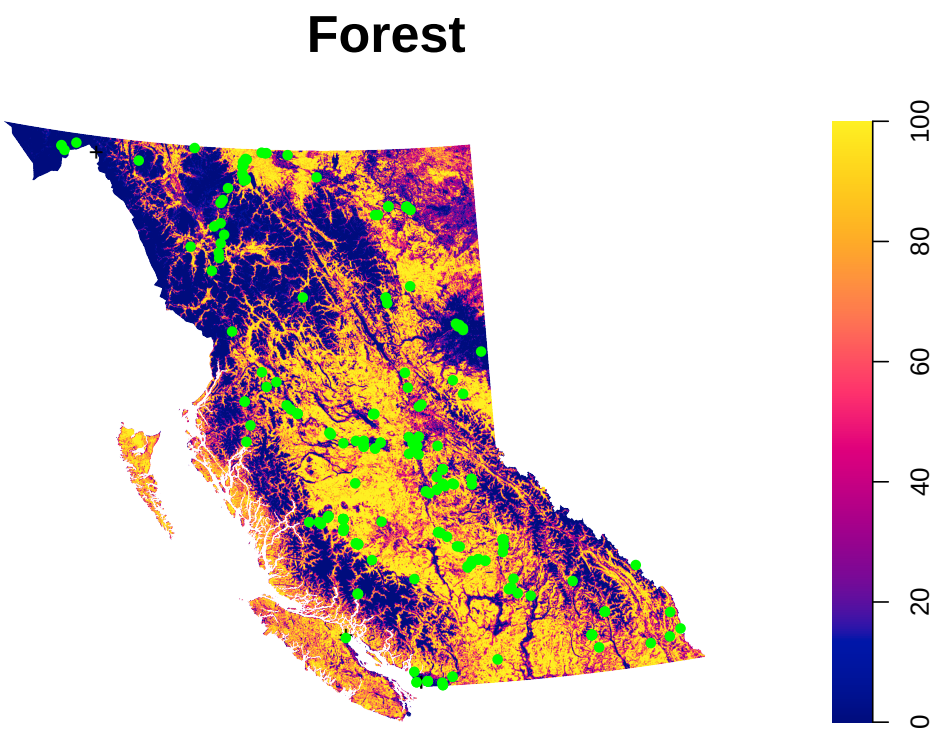
simple model

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~HFI
##
## Fitted trend coefficients:
## (Intercept)         HFI
##   -23.31421     5.98177
##
##               Estimate      S.E.    CI95.lo    CI95.hi Ztest       Zval
## (Intercept) -23.31421 0.1058336 -23.521645 -23.106785   *** -220.29132
## HFI           5.98177 0.2148471   5.560677   6.402862   ***   27.84199
## Problem:
```

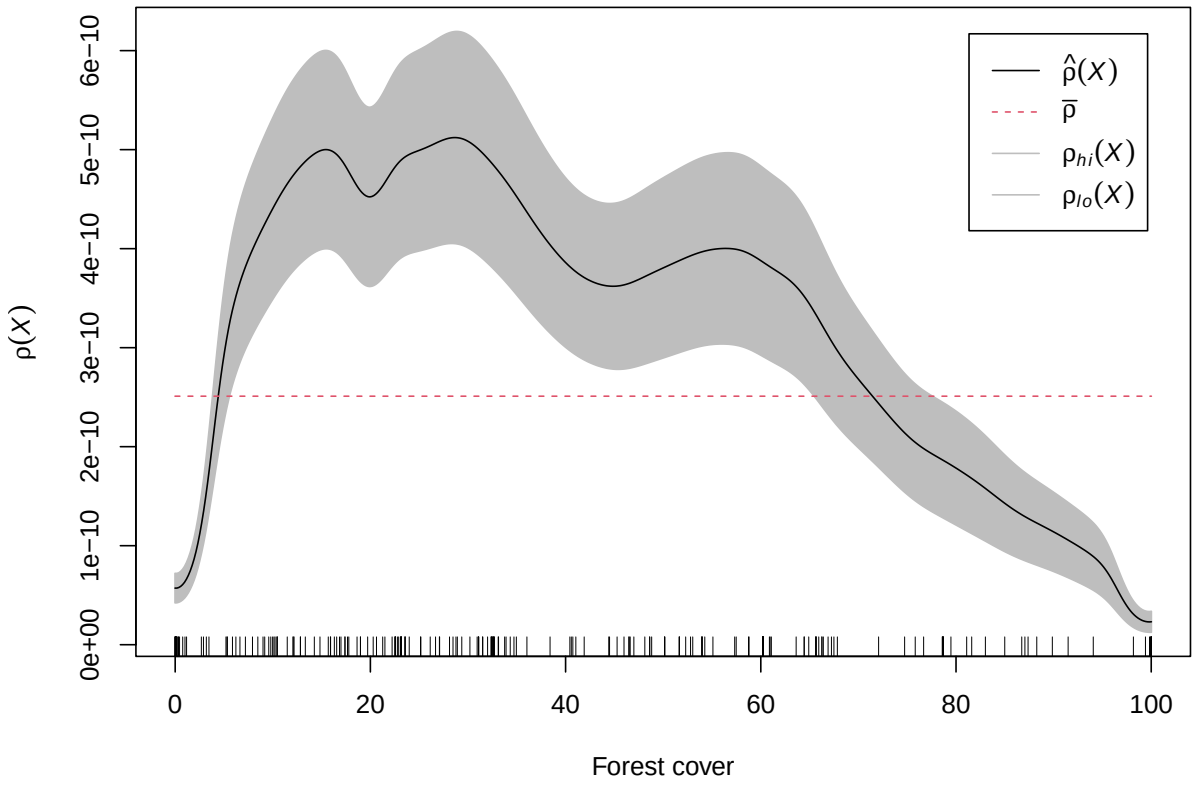
```
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points

## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~HFI + exp(HFI)
##
## Fitted trend coefficients:
## (Intercept)         HFI    exp(HFI)
##   -13.71641    22.17067   -10.35893
##
##              Estimate     S.E.   CI95.lo    CI95.hi Ztest      Zval
## (Intercept) -13.71641 1.404309 -16.46881 -10.964015   *** -9.767372
## HFI          22.17067 2.382441  17.50117  26.840165   ***  9.305861
## exp(HFI)    -10.35893 1.522442 -13.34286  -7.374997   *** -6.804153
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points

## [1] 10468.83

## [1] 10420.78
```



Forest Cover



1. To add : Add other covariates together - side by side



There seems to be non-linear relationship between forest cover and number of red lions observed. The observance increase with increase in forest cover at intermediate coverage and then it decreases.

Next is to observe if there is any correlation between the covariates (collinearity in the covariates dataset). This is necessary to avoid any identifiability issues in modeling the data.

2. Fix issues with collinearity

3. Finalize and summarize about other two variables - elevation and water.

Model Fitting

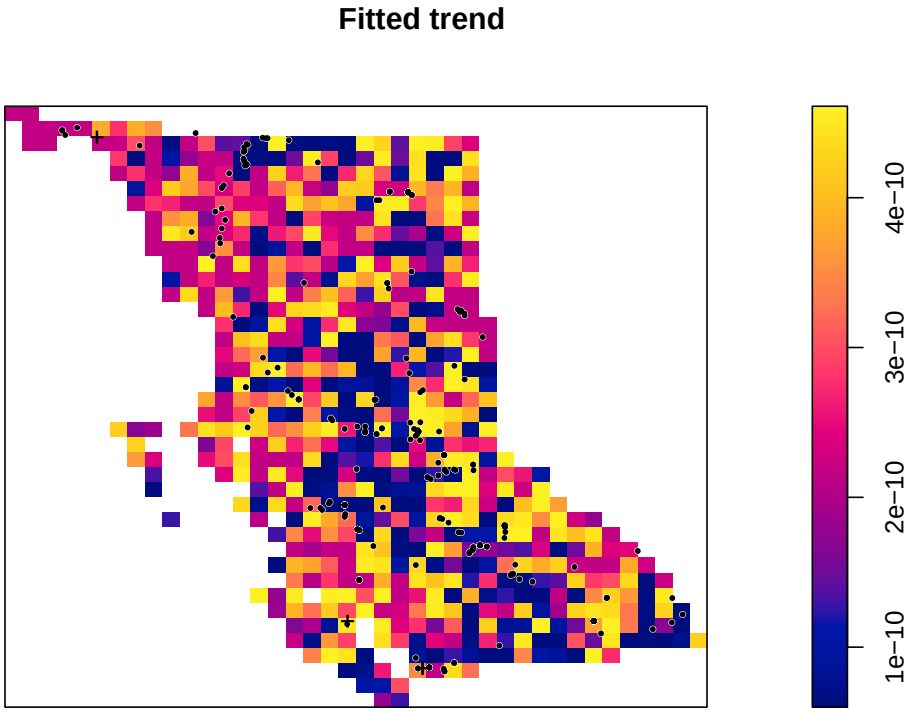
Try different models - linear and quadratic, do necessary anova and other tests for model selection and validation

Model with Forest

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Forest + I(Forest^2)
```

```
##
## Fitted trend coefficients:
##   (Intercept)        Forest   I(Forest^2)
## -2.225373e+01  3.996522e-02 -5.280288e-04
##
##                   Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept) -2.225373e+01 1.330469e-01 -2.251450e+01 -2.199297e+01   ***
## Forest       3.996522e-02 7.091557e-03  2.606603e-02  5.386442e-02   ***
## I(Forest^2) -5.280288e-04 7.699618e-05 -6.789385e-04 -3.771191e-04   ***
##                     Zval
## (Intercept) -167.262353
## Forest         5.635606
## I(Forest^2)   -6.857857
```

All variables look significant.



Does not look like a good fit for all points and also not easy to interpret.

Analyze - linear+quad, AIC

Validate if we need GAM - 5. linear vs gam - table: evaluate, AIC, visualize

Results: Length: Describe your statistical findings. Tables and figures should be used throughout. Length: As long as necessary.

Findings from EDA

model: Tablulate observations -4. table to summarize models - linear+quad, AIC

model: GAM - Table if necessary

Discussion: Provide a brief summary of your findings. Length: ca. 1 page.

References: Include references to all necessary literature.

1. Data: GBIF.org (09 April 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.p6tsaa
2. Research topics: https://cwf-fcf.org/en/resources/encyclopedias/fauna/mammals/red-fox.html