DATA589 Project : Spatial Analysis of Ascomycota (Fungi)

Justin Chan, CM Tong (Kenny)

2023-04-29

1. Introduction

1.1 Ascomycota as a phylum of the kingdom Fungi

Ascomycota is a phylum (or category) of the kingdom Fungi. In fact, it is the largest phylum of Fungi. Its members are also commonly known as sac fungi.

Among many other Fungi phylum, Ascomycota is of particularly useful to humans. As it can be used as sources of medicinally important compounds (e.g. antibiotics), and fermentation for various food products such as alcoholic beverages, bread and cheese. One famous example of its members is Penicillium, which has been widely used and utilized in natural environment, in food spoilage, and food and drug production.

The Ascomycota also plays an important role in most land-based ecosystems, because they are efficient decomposers, breaking down organic materials (e.g. dead leaves and animals) and completing the nutrient cycle.

1.2 Data Source

Data is downloaded directly from GBIF with the gbif R package. According to the GBIF classification system, Ascomycota comprises of about 20 Classes as immediate children species. The whole phylum makes up a total of 16 million occurences records across the globe.

The details will be discussed in the later sections.

1.3 Research Questions

Due to the potential medical and various practical values of Ascomycota mentioned above, we would be interested in exploring and investigating its growth distribution, patterns and population density in terms of spatial analysis.

During this project, we hope to answer the following questions :

- Are there any spatial patterns?
- Is the distribution homogeneous and inhomogeneous?
- Are there any significant correlation with the available covariate data such as Elevation, Forest, and Distance to Water?
- Are there evidence to support any occurrence clustering, independence or avoidance? Are these affected by the homogeneity and inhomogeneity assumption of the data?
- What are the possible prediction models? Does complicated model outperform simple model? Is the complicated model worth it?
- Can higher-degree polynomial model further improve the prediction model?

2. Methods and Results

2.1 Dataset Variables and Selection

A total of about 69,000 occurrences records are available in BC, which are consolidated from a total of 109 datasets. However, the following three major parties constitute about 85% of the total occurrences records :

- UBC Herbarium - Licen Collection
- iNaturalist Research-grade Observations
- Assembly and activity of microbial communities in the Pacific temperate rainforest

We tried to load the complete 69K occurrences set into our computer, we found that loading the whole dataset takes a tremendous amount of time for most processes and operation. For example, it took about 4 hours just to load the complete set of data from gbif api! Running operations like scanLRTS(), rhohat(), envelope() took even longer! After consulting the course instructor, we have set the limit to 2000 records when we retrieve the data with the gbif api to make it 'computing-resources affordable'.

```
# install.packages("rgbif")
library(rgbif)
#?rgbif

species <- c("Ascomycota")

# Get number of occurrence records from rgbif
#occ_count(scientificName = species) #16069587 in the dataset

#Filter Ascomycota data to only in BC
asc_count <- occ_count(scientificName = species,
                           hasCoordinate = TRUE,
                           country = "CA",
                           stateProvince = "British Columbia")

asc_bc_data <- occ_data(scientificName = species,
                            hasCoordinate = TRUE,
                            country = "CA",
                            stateProvince = "British Columbia",
                            limit=2000) #* 102 illegal points stored in attr(,"rejects") ***

# class(asc_bc_data) #gbif_data
asc_bc_data <- asc_bc_data$data #gbif_data to data.frame
## head(asc_bc_data) #contain "Ascomycota" data only in BC
```

2.1.1 Data Records

Firstly, screen through some sample records of the dataset to see what information it contains and which attributes would and would not be useful and valuable in our analysis.

A total of 75 data columnns are available. These columns can be broadly divided into the following three main categories :

- Scientific Classification and Taxonomy Details : Scientific Name, Hierarchy of Biological classification and taxonomic ranks, etc

- Point and Location Details : Continent, State/Province, Coordinates, Coordinates Uncertainty Meters, etc
- Data Collection Details : Collector, Date and Timestamp

2.1.2 Data Cleaning

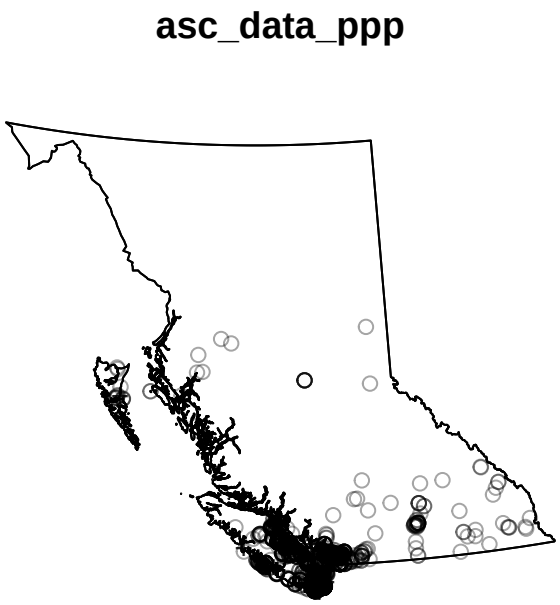
As the dataset contains too many information, including many detailed timestamps, internal key/identifiers information and dataset/records identifiers which should not be valuable in our analysis and complicate our subsequent analysis, we have performed preliminary cleaning procedure and performed attributes selection, in order to only retain those few potentially useful attributes, speeding up our subsequent analysis and making it more focused. The cleaned list iis shown below.

```
# View(data.frame(names(asc_bc_data)))
cleaned_asc_bc <- asc_bc_data[ , c("decimalLongitude", "decimalLatitude", "order", "family", "genus", "s
                                   "year", "month", "day", "eventDate", "occurrenceStatus", "class", "cou
                                   "verbatimEventDate", "collectionCode", "gbifID", "verbatimLocality"
                                   )]
```

2.1.3 Spatial Dataset Object Conversion and Preparation

As the current format of the fungi data is not in a PPP class object, to facilitate subsequent analysis and plots, we would firstly convert it to the PPP, which would make subsequent plotting and analysis library much more accessible.

```
plot(asc_data_ppp)
```



2.2 Statistical Packages

The following packages are used to perform the analysis :

- rgbif : allow searching and retrieving data from GBIF, access various dataset metadata, species names, and occurrences details
- sp, rgdal : process, manipulate and transform the data source longitude and latitude information
- spstat : provides many statistical tests, analysis and plots
- maptools : provides utilities to handle and analyze spatial objects
- Others supporting libraries include kdensity and splines

2.3 Analytical Workflow

The report will start with some basic Exploratory Data Analysis and basic data plottings to have an initial understanding of the overall data points pattern.

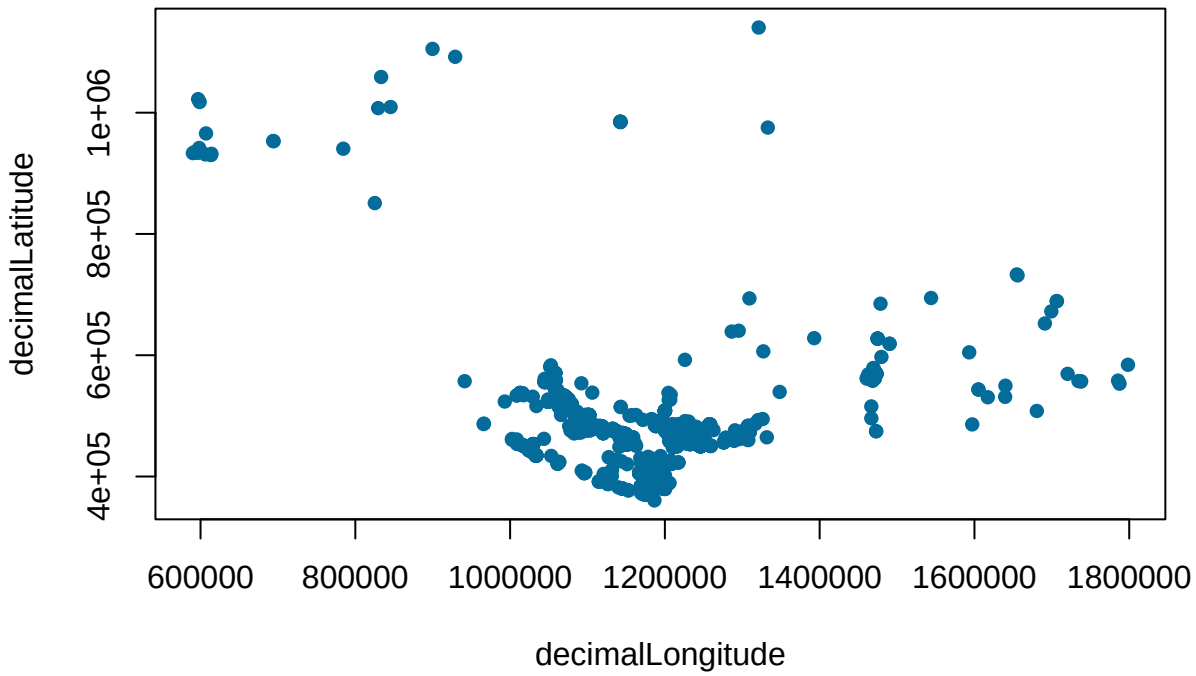
We will then conduct First Moment Analysis including homogeneity analysis and covariates study on some variables. Afterwards, we will perform Second Moment analysis where popular tools such as K-function and Pair Correlation Function will be deployed to investigate any clustering.

The report will then conduct model fitting and selection with both lower and higher polynomial variable terms and evaluate the fitness and its costs with AIC values. The report is concluded by a final model validation process for the resulted model to assess the wellness of the model fitting.

2.3.1 EDA : Initial Coordinates Plotting

First, we have a coordinate plot to observe the general pattern in terms of the coordinates.

```
#Visualise the data
plot(decimalLatitude ~ decimalLongitude,
     pch = 16,
     col = "#046C9A",
     data = cleaned_asc_bc)
```



From the initial coordinate plot, we have identified the following very preliminary observation :

- Points clustering is highly likely as a large group of data points is spotted at the decimal Longitude between 1000K and 1300K and Latitude between 4e+5 and 5e+5.

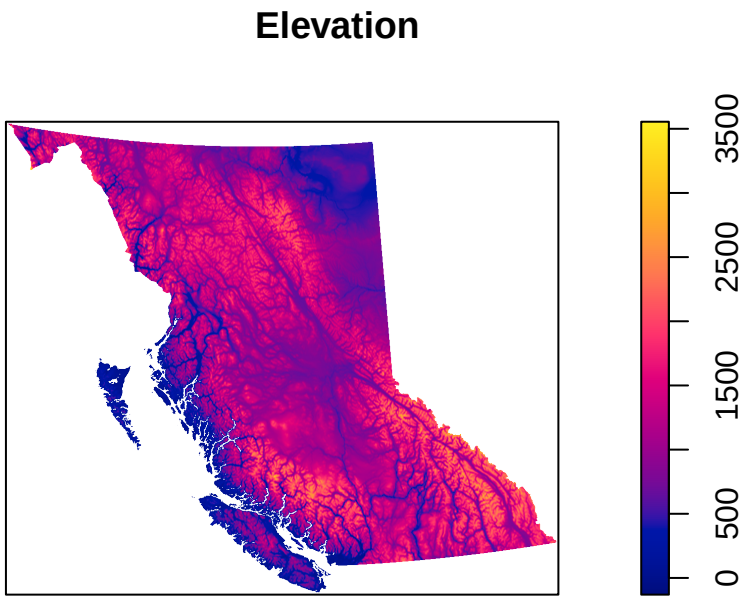
- There is only a single major and significantly large clustering only. Although there are some other scattered points observed near the Longitude range 1500K-1800K, they are simply incomparable to the large clustering in terms of the scale.

2.3.2 EDA : BC Windows and Covariates Data Walkthough

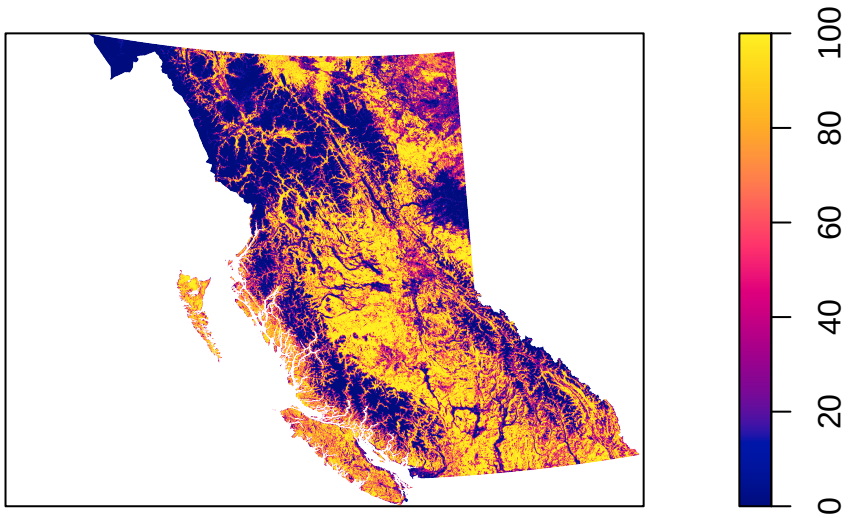
Let's see what the BC_Covariates.Rda file provides us which may help us to identify and choose appropriate fields to be used in the covariates analysis in the later part of this analysis. It should contain the information on the BC province :

- Windows
- Elevation
- Forest
- Dist Water

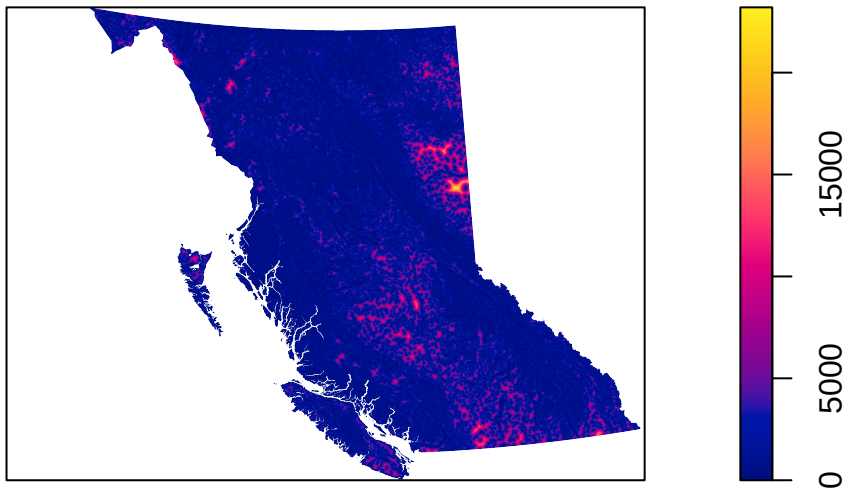
```
#par(mfrow=c(2,2))

#Elevation
plot(BC$Elevation, main = "Elevation")
```



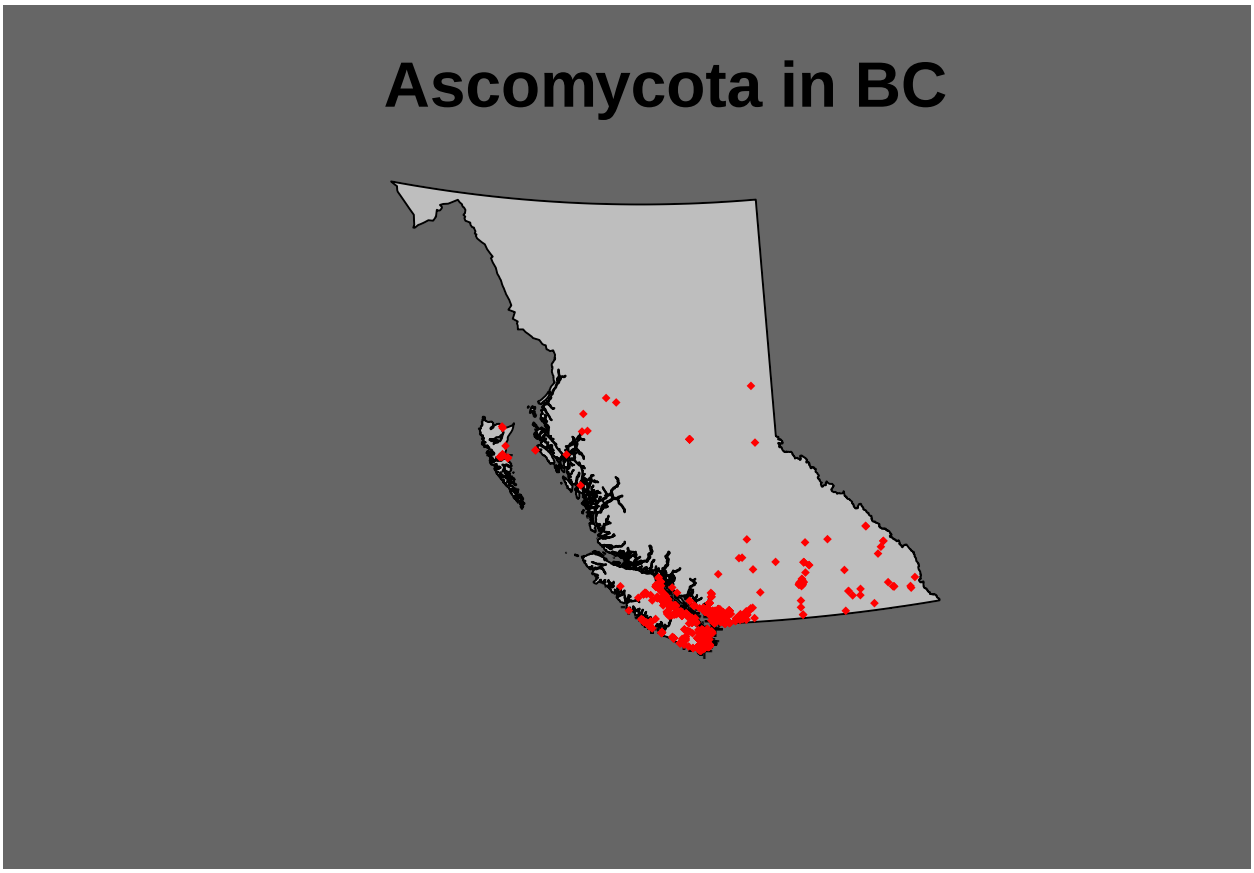

```
#Forest
plot(BC$Forest, main = "Forest")
```




```
#Dist_Water
plot(BC$Dist_Water, main = "Distance to Water")
```




```
plot(asc_data_ppp,
     which.marks = "species", # Which mark to use
     col = "grey", #The colour of the window
     cols = 'red', #The colours of the points
     cex = 0.6,
     pch = 18, # The plotting symbol
     main = "Ascomycota in BC", # The title
     par(bg="grey40", cex.main = 2),
     cex = 0.6,
     legend = T) # Turn off the legend depending on needs
```



Next, We try to pick Elevation to divide its values into 5 levels, and see if there are any patterns for the distribution of the fungi points in the different elevation levels.

```
cut <- cut(BC$Elevation,5,
labels = c("low","low-medium","medium","medium-high","high"))
table(cut[asc_data_ppp]) #most in low elevation
```

```
##
##          low  low-medium      medium medium-high        high
##         1481          67           8           0           0
```

Finding :

- Overwhelming majority of the fungi points are found in the low elevation region.
- There is a trend : the higher the elevation region is, the less occurrences of the fungi is observed.
- The fungi density is dropping substantially when the elevation increases, until it reached the level of zero starting from the meidum-high elevation level.

2.3.3 First Moment Descriptive Statistics

After preliminary EDA and very high level plots, we will study various first moment descriptive statistical measures.

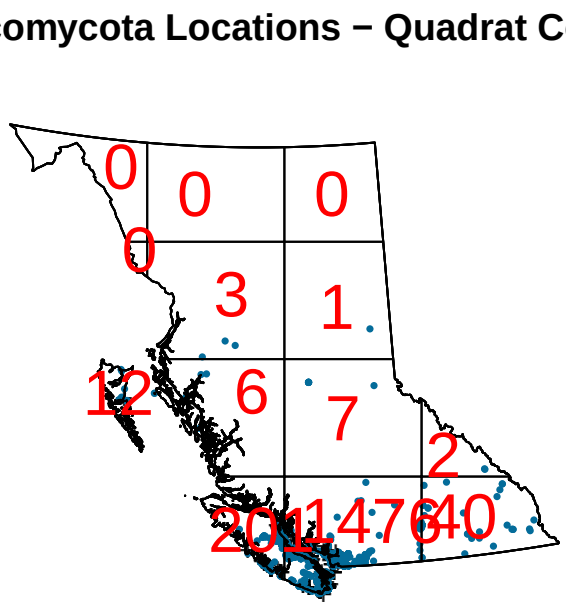
```
intensity(asc_data_ppp)
```

2.3.3.a Intensity

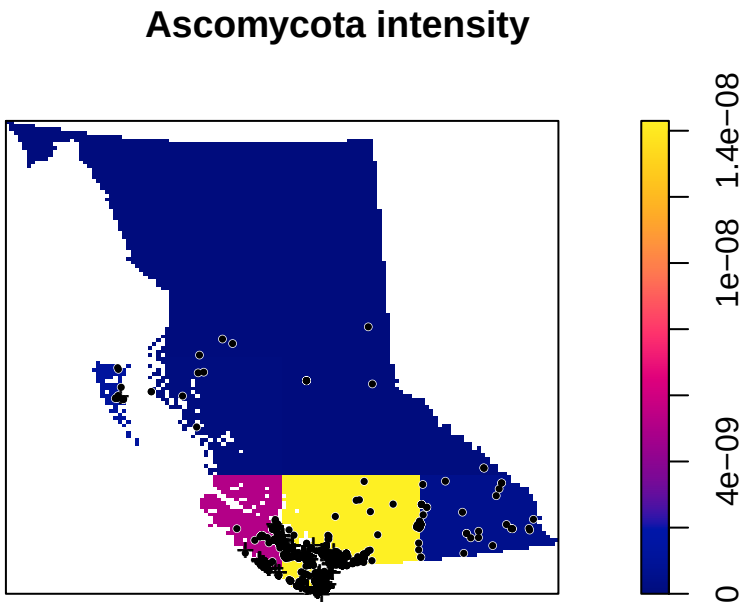
```
## [1] 1.843372e-09
```

2.3.3.b Homogeneity Studies, Quadrat Test and Hotspot Analysis Note that the intensity is a very small number. This is consistent with our plots above. There are not many points in the whole BC windows. Overwhelming regions in BC have (nearly) zero occurrences while the rest are highly clustered in a few regions. Let's verify this with the Quadrat Count Plot.

```
Q <- quadratcount(asc_data_ppp,
                  nx = 4,
                  ny = 4)

#Plot the output
plot(asc_data_ppp,
     pch = 16,
     cex = 0.5,
     cols = "#046C9A",
     main = " Ascomycota Locations - Quadrat Count")

plot(Q, cex = 2, col = "red", add = T)
```



```
plot(intensity(Q, image = T),
     main = "Ascomycota intensity")

plot(asc_data_ppp,
     pch = 16,
     cex = 0.6,
     cols = "white",
     add = T)

plot(asc_data_ppp,
     pch = 16,
     cex = 0.5,
     cols = "black",
     add = T)
```



Next, we perform a Quadrat test of homogeneity

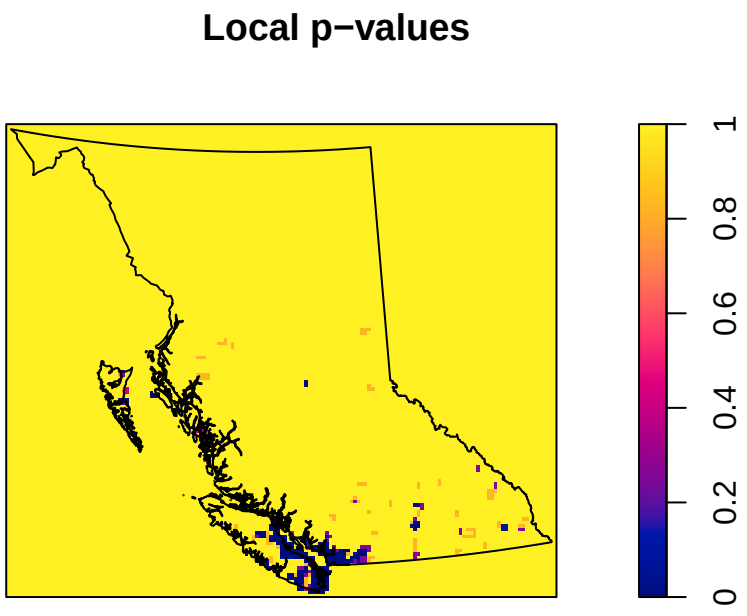
```
#Quadrat test of homogeneity
quadrat.test(Q)
```

```
##
##  Chi-squared test of CSR using quadrat counts
##
## data:
## X2 = 10279, df = 12, p-value < 2.2e-16
```

```
## alternative hypothesis: two.sided
##
## Quadrats: 13 tiles (irregular windows)
```

In addition, We also perform a Likelihood Ratio Test to evaluate the degree of homogeneity as a cross reference.

```
R <- bw.ppl(asc_data_ppp)
LR <- scanLRTS(asc_data_ppp,r=R)

#plot(LR,"Likelihood Ratio Test")
#plot(asc_data_ppp$window,"Likelihood Ratio Test")

pvals <- eval.im(pchisq(LR,
                        df = 1,
                        lower.tail = FALSE))


#Plot the output
plot(pvals, main = "Local p-values")
plot(asc_data_ppp$window,add=T)
```



Once again, these plots have strongly reinforced the spatial inhomogeneity nature of the fungi distribution :

- Overwhelming majority of the regions have zero or near zero points. The quadrat count plot has shown drastic difference between the quadrats.

- From the intensity plot, The points are observed being mostly clustered in the south-eastern region of BC
- Quadrat Test and Likelihood Ratio Test indicates a significant deviation from homogeneity for the points.
- Hot spot analysis shows only a very few prominent hot spots with low p-values which are mainly located in the southern region. Most regions rarely have significant number of occurrences, and therefore are shown in high p-value (i.e. p=1).

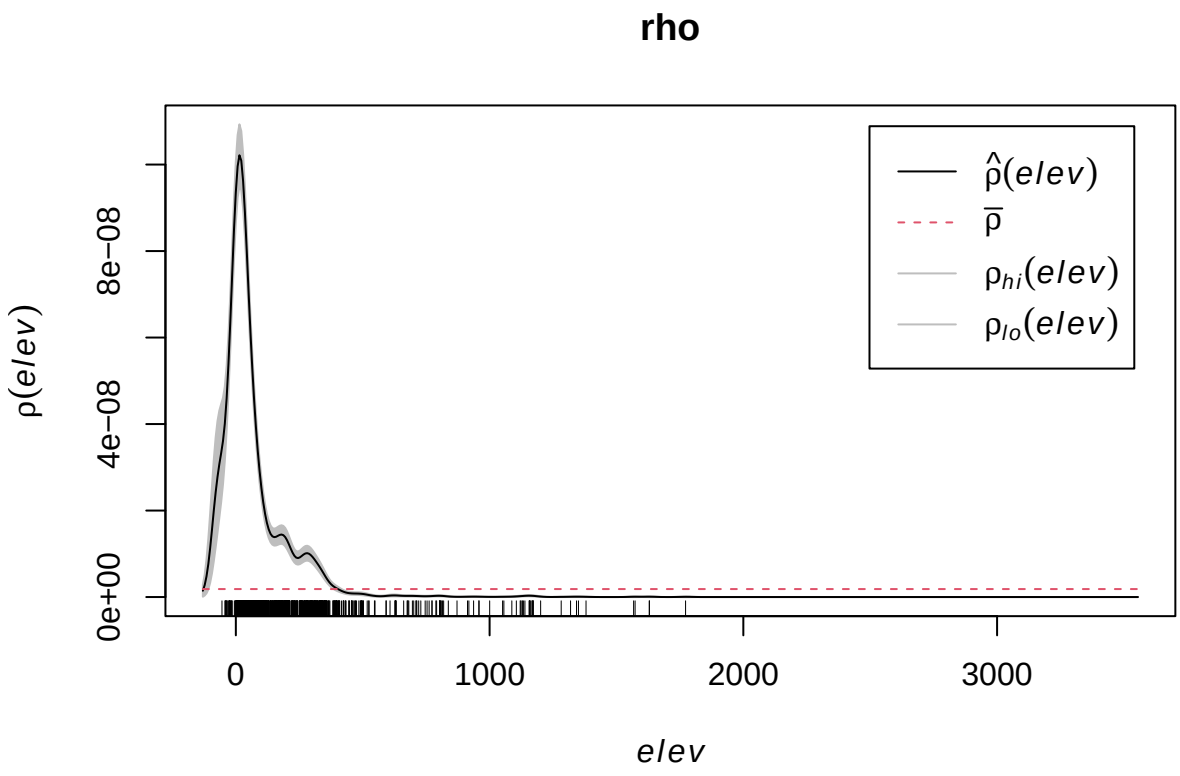
2.3.4 Covariate Study

Now, we will study the covariate behaviour individually, and see whether there are strong support or observable patterns. We will conduct quantile split into 4 sections for each of the covariates.

2.3.4.a Covariate Variable : Elevation Let's start with Elevation.

```
elev <- BC$Elevation
b <- quantile(elev,probs=(0:4)/4,type=2)
Zcut <- cut(elev,breaks=b)
V <- tess(image=Zcut)
quadratcount(asc_data_ppp,tess=V)
```

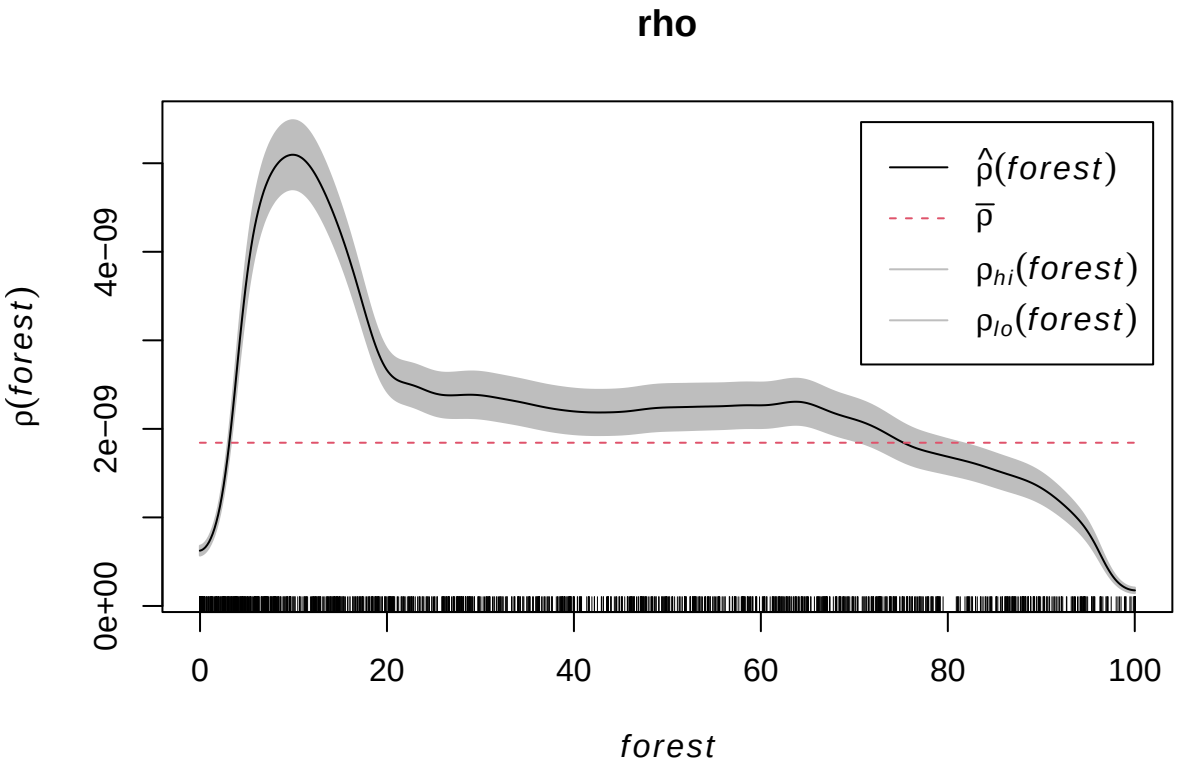
```
## tile
##          (-130,761]       (761,1.1e+03]  (1.1e+03,1.46e+03] (1.46e+03,3.56e+03]
##                1694                  24                  25                   5
```



2.3.4.b Covariate Variable : Forest Then, let's see Forest.

```
forest <- BC$Forest
b <- quantile(forest,probs=(0:4)/4,type=2)
Zcut <- cut(forest,breaks=b)
V <- tess(image=Zcut)
quadratcount(asc_data_ppp,tess=V)
```

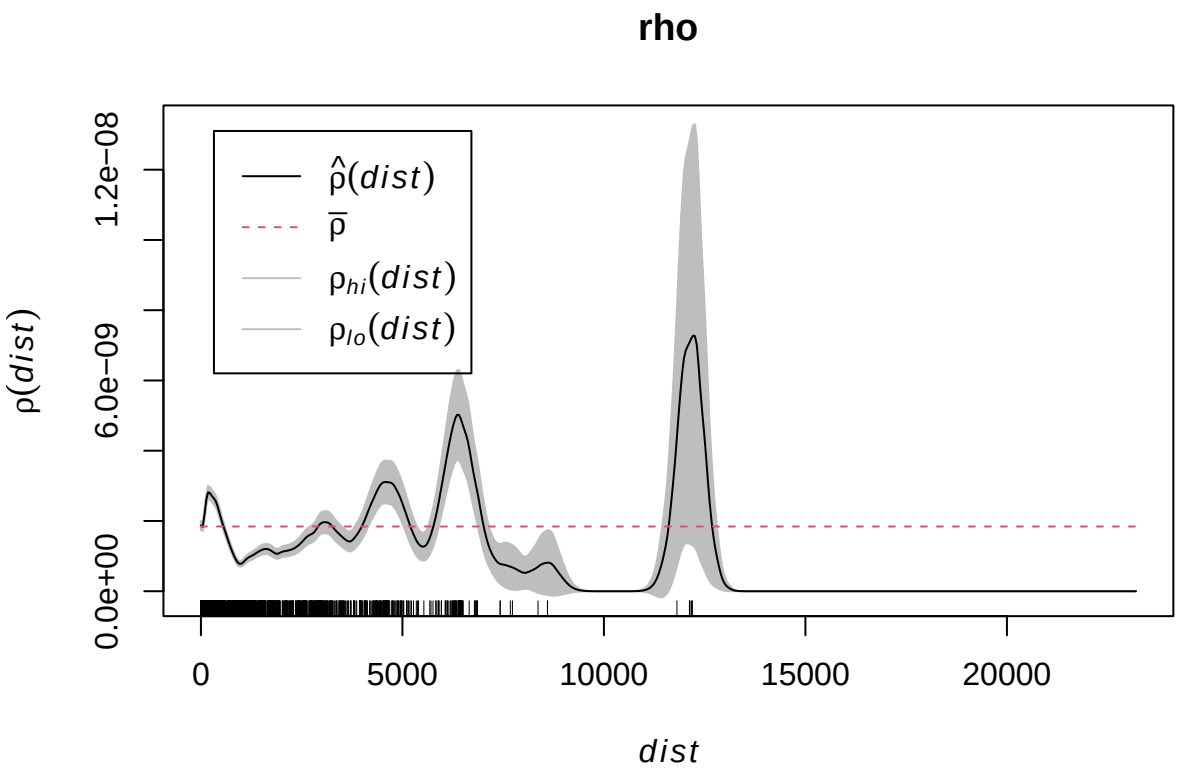
```
## tile
##    (0,11.6] (11.6,50.2] (50.2,86.9]  (86.9,100]
##         582         533         455         177
```



2.3.4.c Covariate Variable : Distance to Water Followed by Distance to Water.

```
dist <- BC$Dist_Water
b <- quantile(dist,probs=(0:4)/4,type=2)
Zcut <- cut(dist,breaks=b)
V <- tess(image=Zcut)
quadratcount(asc_data_ppp,tess=V)
```

```
## tile
##             (0,483]       (483,1.1e+03]  (1.1e+03,2.18e+03] (2.18e+03,2.32e+04]
##                 785                 220                 267                 424
```



2.3.4.d Section Conclusion Observation :

- For Elevation Level :
  - It is obvious that the fungi is overwhelmingly correlated to the low elevation level.
  - The proportion of occurrence that appears in the lowest elevation sector accounts for about 95% of the identified points.
  - A significant clustering is observed at the low elevation.
  - There are basically a very few (if any) occurrences for elevation higher than around 400m.
- For Forest and Distance to Water :
  - It is observed that the fungi is correlated to both Forest and distance .
  - Clustering of observed at the low forest value range and at middle distance to water level.
  - For the highest forest value region, the intensity has been decreasing to near-zero intensity.

2.3.5 Second Moment Descriptives

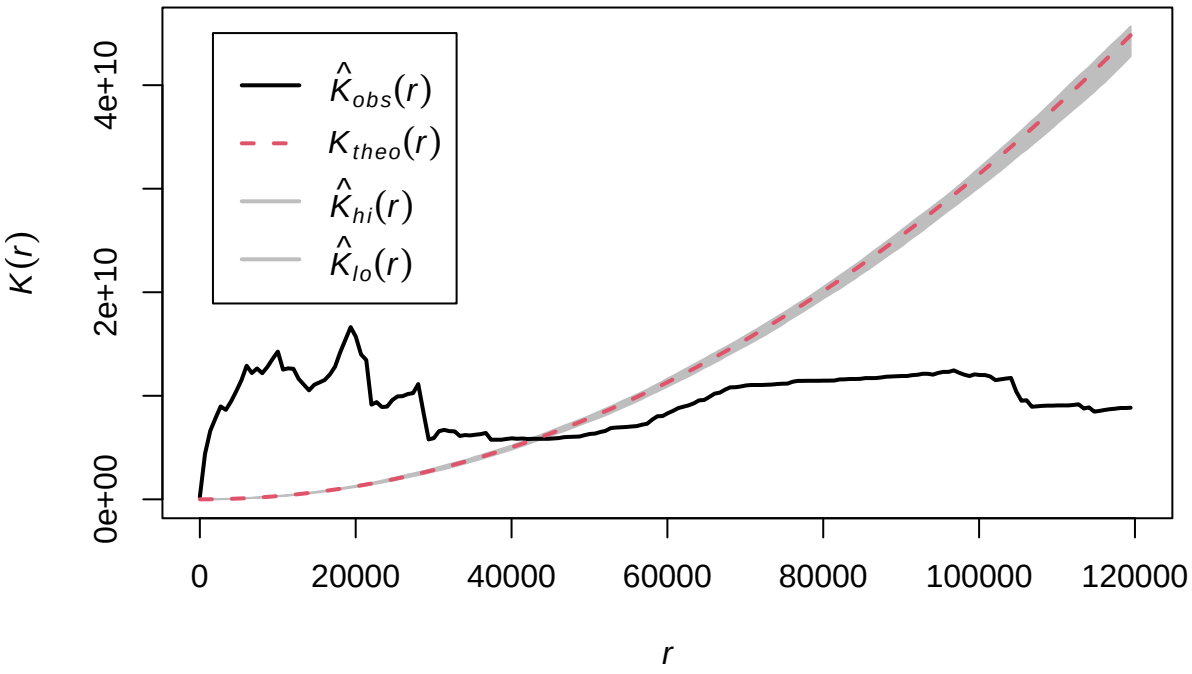
To continue the analysis, we will explore the second moment descriptives to uncover any variances and correlation characteristics of the data.

2.3.5.a Ripley's K-function We firstly use the typical K-function to measure the spatial clustering and point pattern.

```
# Bootstrapped CIs
# rank = 1 means the max and min
# Border correction is to correct for edges around the window
# values will be used for CI
E_asc <- envelope(asc_data_ppp,
                  Kest,
                  correction="border",
                  rank = 1,
                  nsim = 19,
                  fix.n = T)
```

```
## Generating 19 simulations of CSR with fixed number of points  ...
## 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,  19.
##
## Done.
```

```
# visualise the results
plot(E_asc,
     main = "",
     lwd = 2,
     xlim=c(0,120000))
```

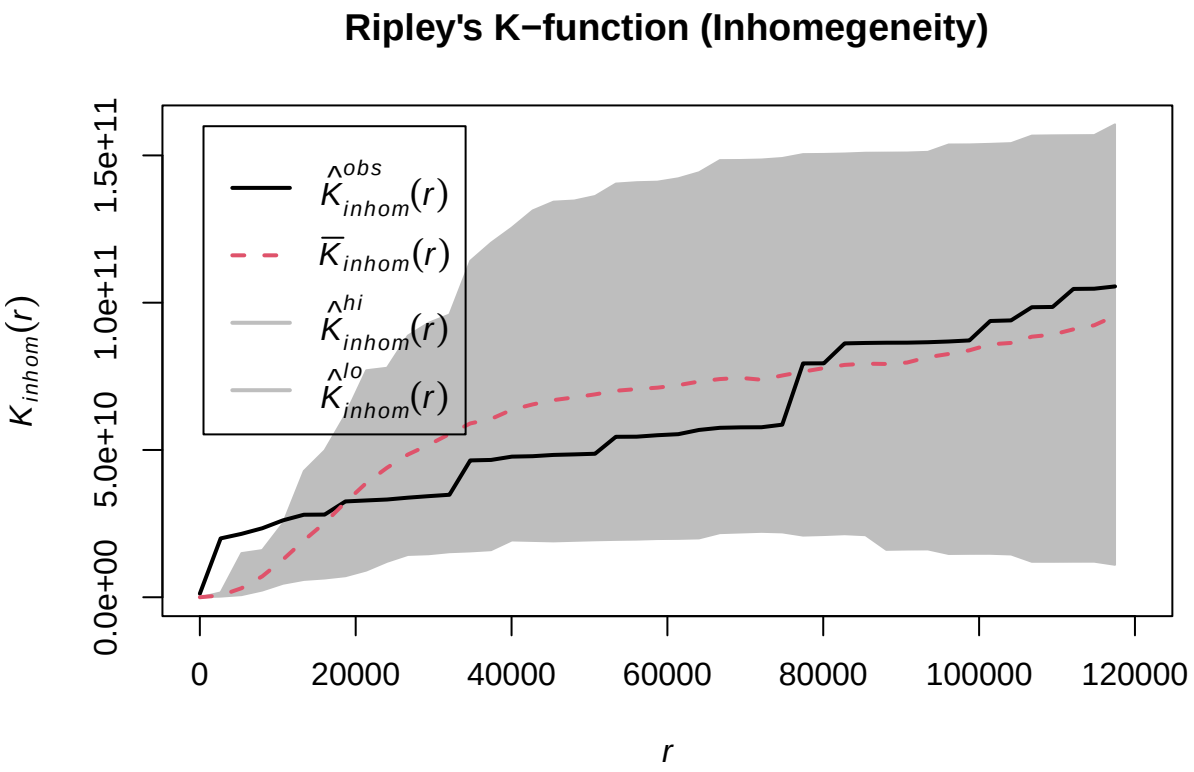


However, note that the above analysis assumed homogeneity. In our case, as previous analysis reveals that there is inhomogeneity in our data, we are going to correct for inhomogeneity and reperform the test for a more accurate result.

```
lambda_asc <- density(asc_data_ppp, bw.ppl)
Kinhom_asc <- Kinhom(asc_data_ppp, lambda_asc)

#Estimate a strictly positive density
lambda_asc_pos <- density(asc_data_ppp,
                          sigma=bw.ppl,
                          positive=TRUE)

#Simulation envelope (with points drawn from the estimated intensity)
E_asc_inhom <- envelope(asc_data_ppp,
                        Kinhom,
                        simulate = expression(rpoispp(lambda_asc_pos)),
                        correction="border",
                        rank = 1,
                        nsim = 19,
                        fix.n = TRUE)
```

```
## Generating 19 simulations by evaluating expression  ...
## 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,  19.
##
## Done.
```

```
# visualise the results
# par(mfrow = c(1,2))
plot(E_asc_inhom,
     main = "Ripley's K-function (Inhomegeneity)",
     lwd = 2,
     xlim=c(0,120000))
```



Finding :

- When assumed homogeneity, the observed data shows a strong clustering at small r value range.
- When corrected for inhomogeneity, except that there is some evidence pointing to clustering at the starting region r (0-10,000), obvious clustering pattern is no longer observed.

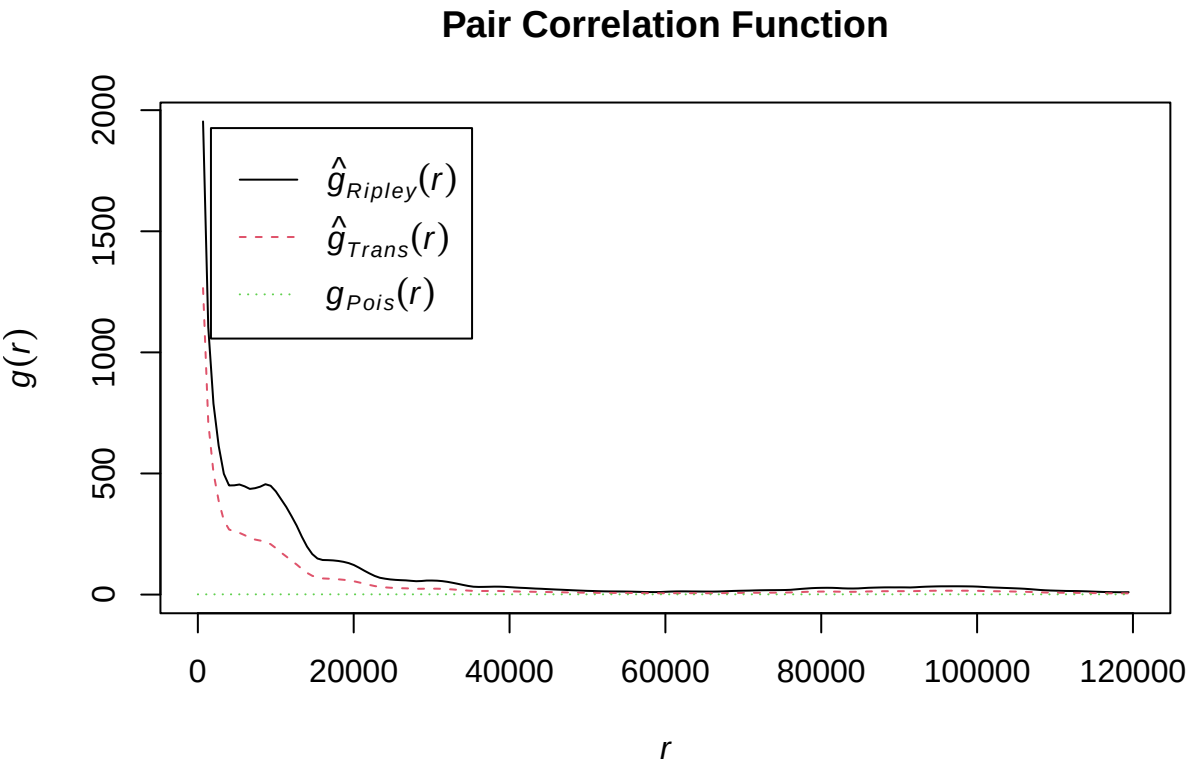
2.3.5.b Pair Correlation Function We are also interested in checking the points relation using pair correlation function.

```
# Estimate the g function
pcf_asc <- pcf(asc_data_ppp)

pcf_asc
```

```
## Function value object (class 'fv')
## for the function r -> g(r)
## ....................................................
##       Math.label        Description
## r     r                 distance argument r
## theo  g[Pois](r)        theoretical Poisson g(r)
## trans hat(g)[Trans](r)  translation-corrected estimate of g(r)
## iso   hat(g)[Ripley](r) isotropic-corrected estimate of g(r)
## ....................................................
## Default plot formula:  .~r
## where "." stands for 'iso', 'trans', 'theo'
```

```
## Recommended range of argument r: [0, 341660]
## Available range of argument r: [0, 341660]
```

```
plot(pcf_asc,xlim=c(0,120000),main="Pair Correlation Function")
```

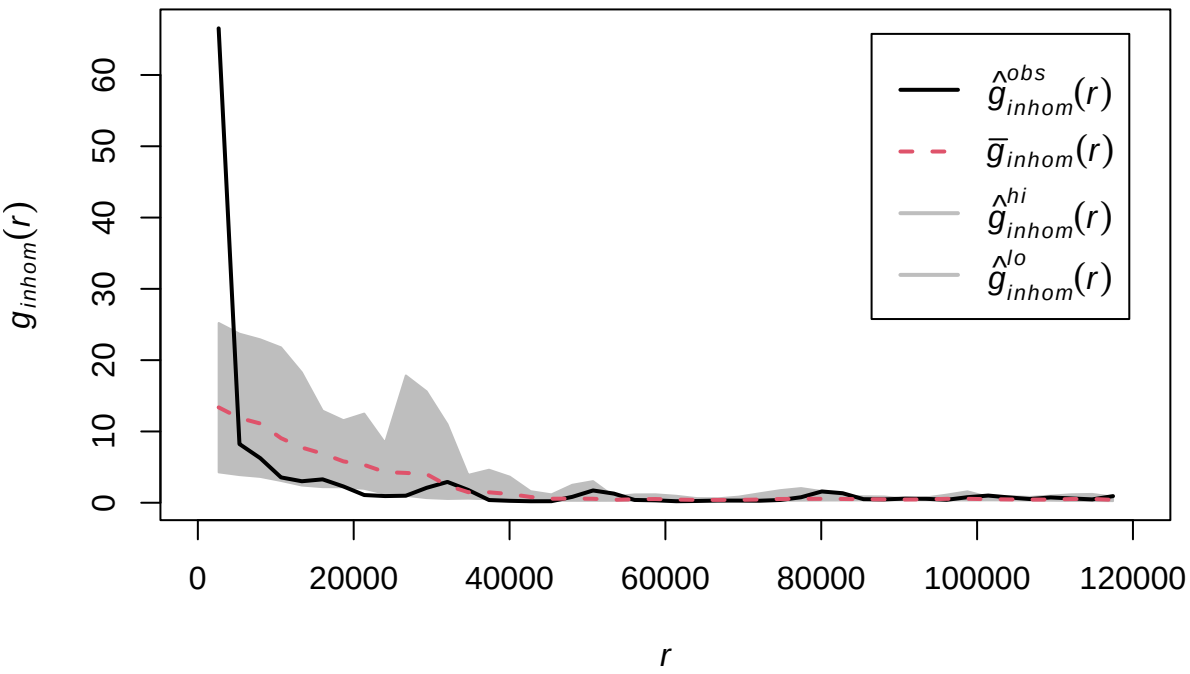


The above estimator also assumes homogeneity. Let's relax this assumption to produce a more accurate analysis.

```
#Simulation envelope (with points drawn from the estimated intensity)
pcf_asc_inhom <- envelope(asc_data_ppp,
                          pcfinhom,
                          simulate = expression(rpoispp(lambda_asc_pos)),
                          rank = 1,
                          nsim = 19)
```

```
## Generating 19 simulations by evaluating expression  ...
## 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,  19.
##
## Done.
```

```
plot(pcf_asc_inhom,
     xlim = c(0,120000),
     main = "",
     lwd = 2)
```



Finding :

- When corrected for homogeneity, the evidence of clustering becomes weaker now.
- Except at the very beginning region of the r-value, the locations of the fungi do not exhibit significant deviation from the theoretical values, therefore lacking strong evidence of clustering or dispersal.

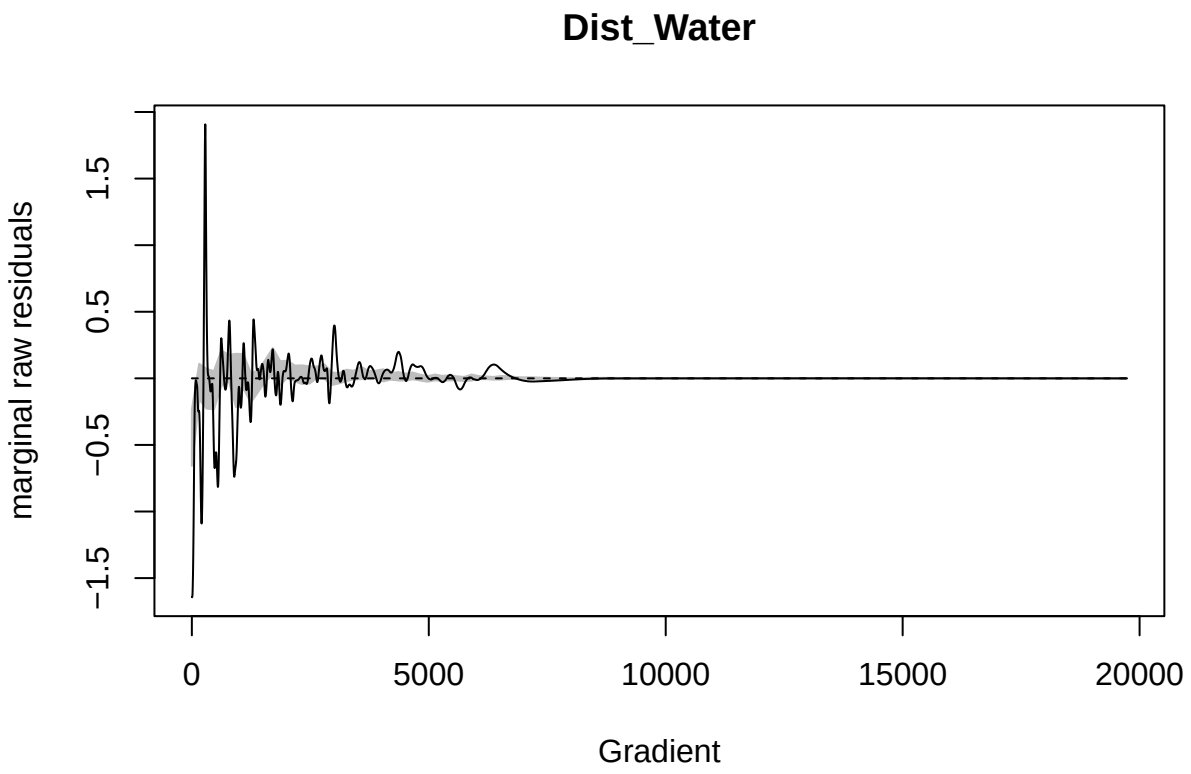
2.3.6 Model Fitting and Selection with AIC

In this section, we will proceed with model fitting and selection, in order to achieve a reasonably effective model, while following the rule of parisomny as much as possible. We will try to start with the simplest linear models. Then we will adjust the model according to the data behaviour and increase the model complexity if situation warrants.

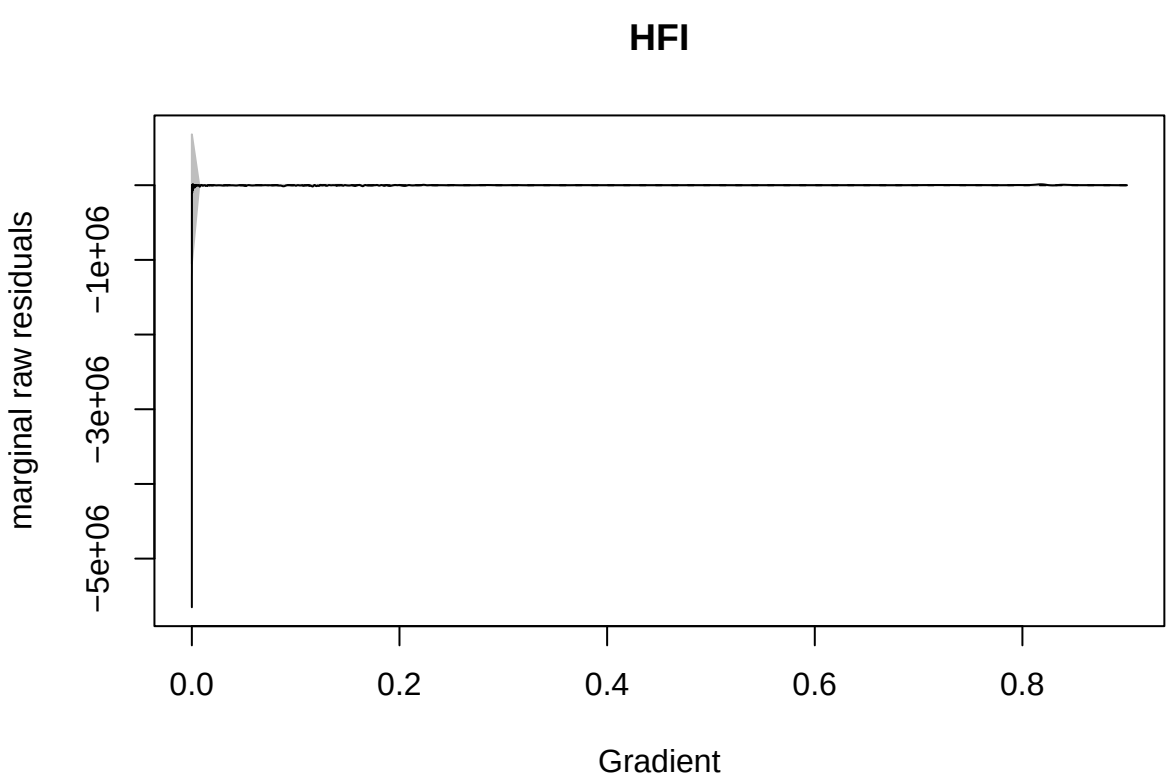
Lurking Variable plot could give us some insight on whether or not we should include the co variate that were missing from our original model. After knowing some potentially useful covariates that could improve our model, Partial Residual Plot can then help us to understand if the added covariates have linear or non-linear relationship with our independent variable, which could allow us to refine our model by using residual as a diagnostic tool. We can then use the Likelihood Ratio Test to confirm whether we can increase the fit of our model by adding more covariates(complexity) to our model.

First we would Build our model using only Elevation as our predictor:

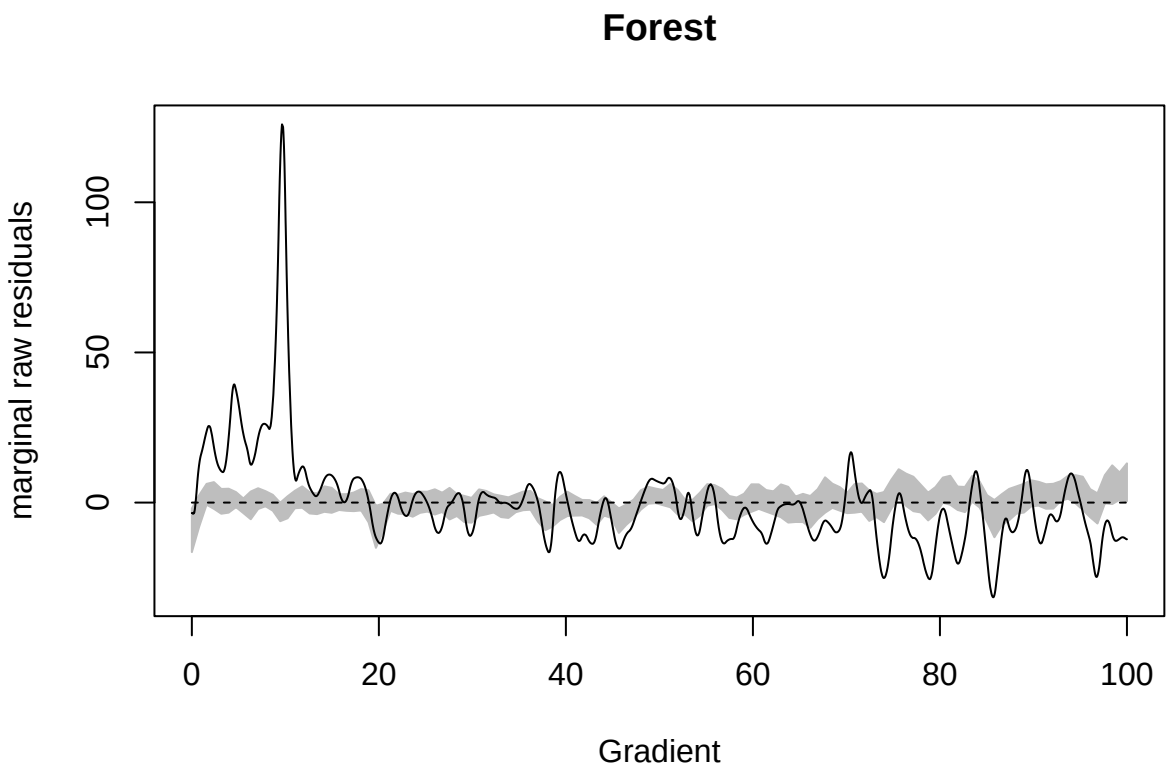
```
## Generating 39 simulated patterns ...1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
## Processing.. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 2
## Done.
```




```
## Generating 39 simulated patterns ...1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
## Processing.. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 2
## Done.
```

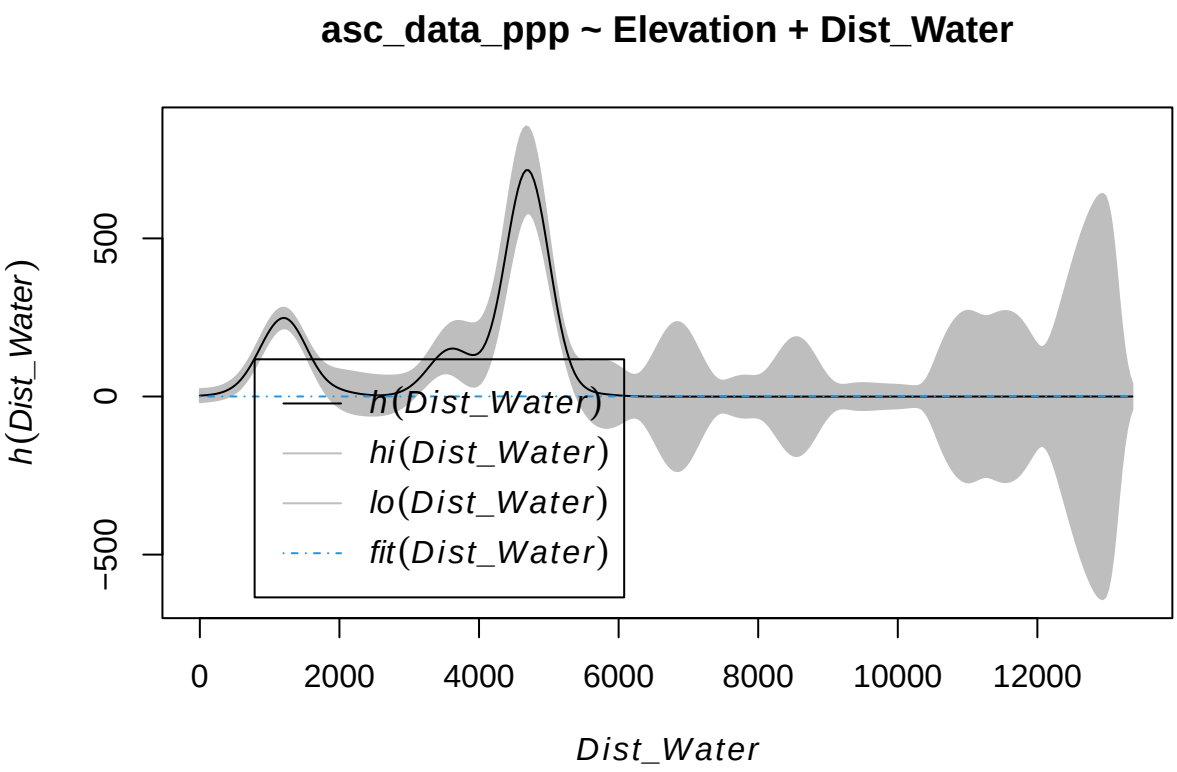



```
## Generating 39 simulated patterns ...1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
## Processing.. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 2
## Done.
```

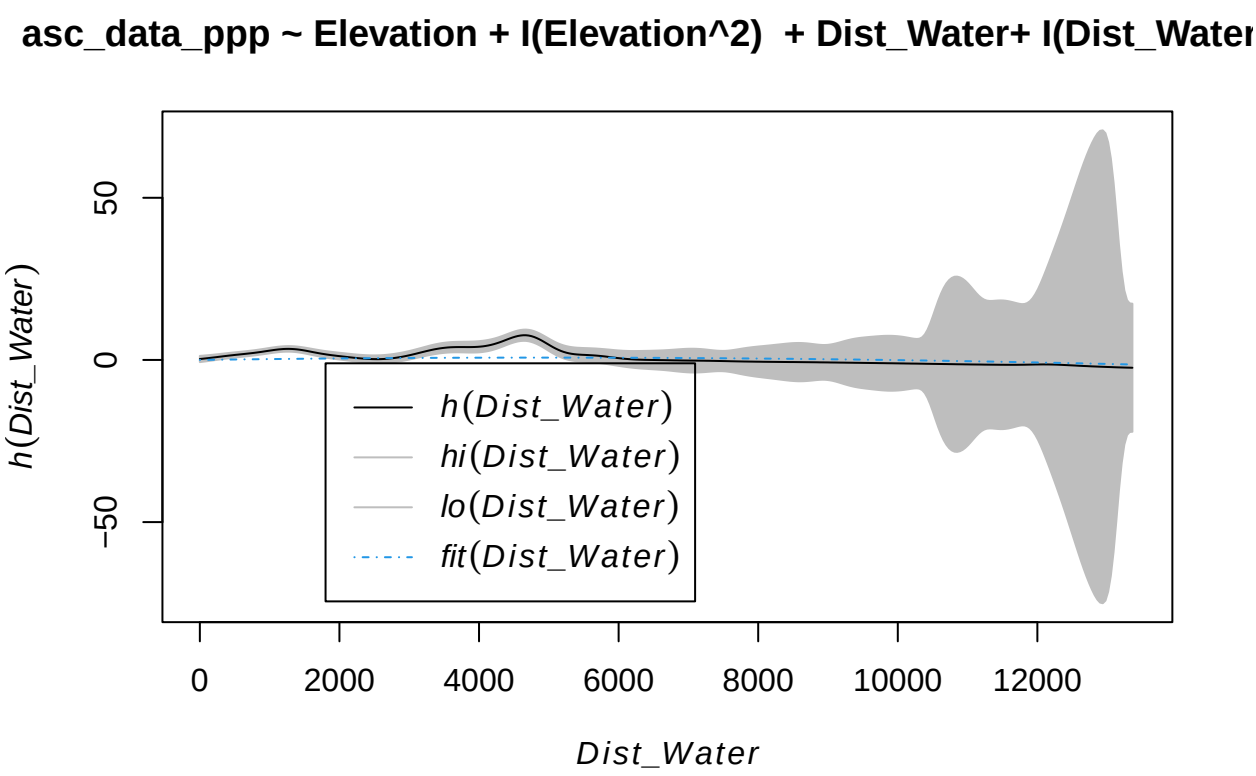


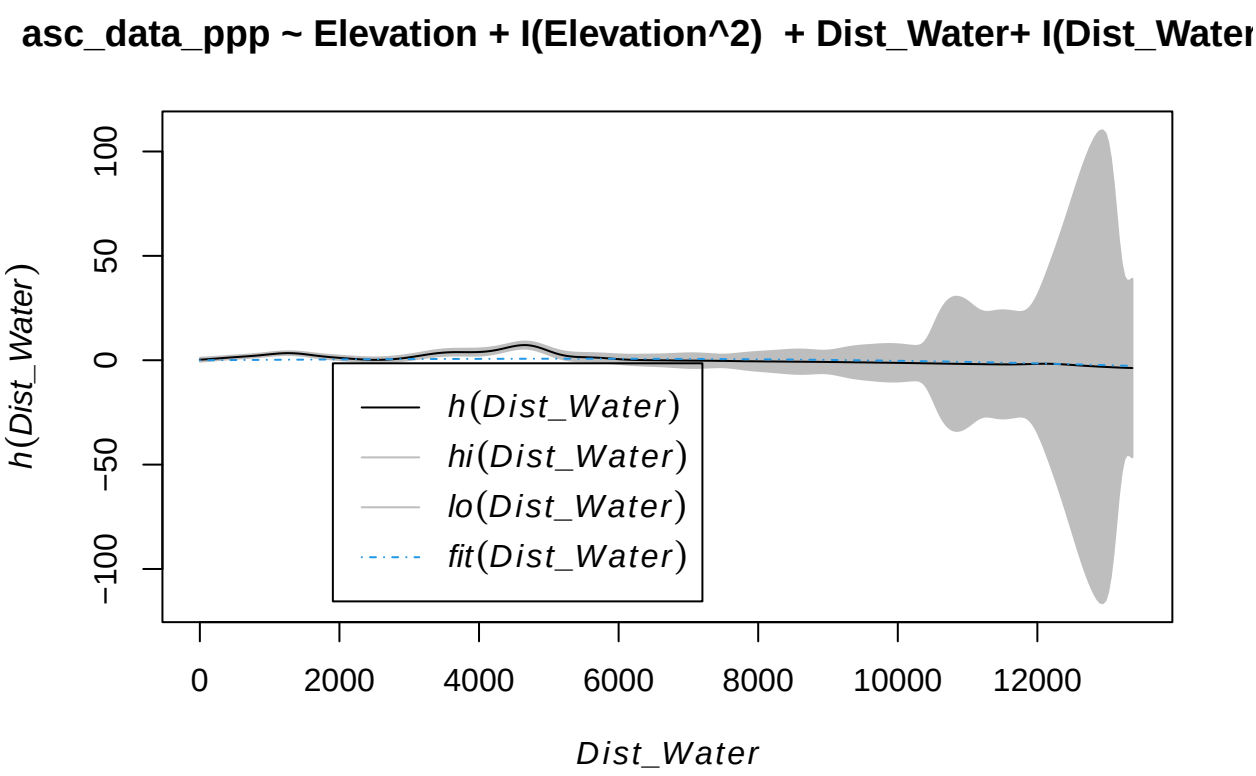
Based on the above Lurking plots, we can see that the all Forest,HFI,Dist_Water have some relation with our predict variable. So we would fit them individually to our model and using Partial Residual to see weather if their relation is linear or Quadratic and then use Likelihood Ratio Test to see if our proposed models (more complex models) are better at prediction than our original model (simpler model).

First, we add Dist_Water into our model and to see if Dist_Water have a linear or non-linear relationship with our trend:



Looks like Dist_Water have a non-linear relationship with our trend. We then can increase the complexity of our covariate by using a high order polynomials to evaluate the degree of fitting:


asc_data_ppp ~ Elevation + I(Elevation^2) + Dist_Water+ I(Dist_Water
h(Dist_Water)
50
0
−50
h(Dist_Water)
hi(Dist_Water)
lo(Dist_Water)
fit(Dist_Water)
0
2000
4000
6000
8000
10000
12000
Dist_Water



Visually, our model fitting have improved after adding higher polynomials for both Elevation and Dist_Water. Now we want to run through the Likelihood ratio test to see if this more complex model has in fact increased the fitting of our data compared to our original model. The null-hypothesis of this test is the simple model(reduced model), if the p-value of the test is lower then our significant value(0.05), it means that our more complex models are a better fit of our data. suggesting we should reject the simple model and favor the more complex model.

```
## Analysis of Deviance Table
##
## Model 1: ~Elevation + Dist_Water      Poisson
## Model 2: ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3)   Poisson
##   Npar Df Deviance  Pr(>Chi)
## 1    3
## 2    5  2   261.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From our Likelihood Ratio Test, we have achieved a low p-value of 2.2e-16, signify extra complexity is desirable. Therefore we will adopt our new model.

Through the same model validation steps, we want to check if adding Forest will improve the fitting of our model. After adding Forest into our model and increasing the complexity of our covariants, we have achieved the following result for our Likelihood Ratio Test result:

```
## Analysis of Deviance Table
##
## Model 1: ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3)   Poisson
```

```
## Model 2: ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3) + Forest + I(Forest^2)    Poisson
##   Npar Df Deviance  Pr(>Chi)
## 1    5
## 2    7  2   594.34 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above Likelihood Ratio Test on the model that have include Forest, we have achieved a low p-value of 2.2e-16, signifying extra complexity is wanted. so we will adopt our new model.

Similarly, the same model validation steps will be taken to assess the use of the HFI variable for our model fitting. After adding HFI into our model, we have achieved the following result from our Likelihood Ratio Test result:

```
## Analysis of Deviance Table
##
## Model 1: ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3) + Forest + I(Forest^2)    Poisson
## Model 2: ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3) + Forest + I(Forest^2) + HFI + I
##   Npar Df Deviance  Pr(>Chi)
## 1   46
## 2   48  2   2129.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above updated Likelihood Ratio Test,, we have now achieved a low p-value of 2.2e-16, implying that extra complexity is desirable. Therefore we will adopt it as our new model.

```
## Error in solve.default(M) :
##   system is computationally singular: reciprocal condition number = 7.05944e-26
## Error in solve.default(M) :
##   system is computationally singular: reciprocal condition number = 7.05944e-26
## Nonstationary Poisson process
## Fitted to point pattern dataset 'asc_data_ppp'
##
## Log intensity:  ~Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3) +
## Forest + I(Forest^2) + HFI + I(HFI^2)
##
## Fitted trend coefficients:
##     (Intercept)       Elevation  I(Elevation^2)      Dist_Water I(Dist_Water^3)
##   -1.942422e+01   -7.137738e-03    1.761325e-06    4.447410e-05   -1.296935e-12
##          Forest     I(Forest^2)             HFI        I(HFI^2)
##   -3.017571e-02    4.184102e-04    1.149646e+01   -6.876242e+00
##
## Standard errors unavailable; variance-covariance matrix is singular
## Problem:
##  Values of the covariates 'Elevation', 'HFI' were NA or undefined at 0.73% (40
## out of 5445) of the quadrature points
##
## *** Fitting algorithm for 'glm' did not converge ***
```

Following the above model fitting and assessment, the following conclusion is drawn :

- Model section suggest us that all of our features are significant to our data

- The model with quadratic terms provides a better fit to the data
- So, a possible model should be [Elevation + I(Elevation^2) + Dist_Water + I(Dist_Water^3) + Forest + I(Forest^2) + HFI + I(HFI^2)]

$$\lambda_{ASC}(u) = e^{-1.94 \ - \ 7.137738e{-}03 \ \text{elevation}(u) \ + \ 1.761325e{-}06 \ \text{elevation}(u)^2 \ - \ 4.447410e{-}05 \ \text{Dist}_{\text{W}}\text{water}(u) \ - \ 1.296935e{-}12 \ \text{Dist}_{\text{W}}\text{water}(u)^3 \ - \ 3.0}$$

2.3.6.a Quadrat Test Let's see the quadrat test result of our previous final derived model.

```
#Run the quadrat test
quadrat.test(fit_simple2, nx = 4, ny = 4)
```

```
##
##  Chi-squared test of fitted Poisson model 'fit_simple2' using quadrat
##  counts
##
## data:  data from fit_simple2
## X2 = 492.1, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 13 tiles (irregular windows)
```

Conclusion :

- This has small p value, suggesting significant deviation from our model's prediction. Room for further improvement is therefore expected, but it does not provide hint for how to achieve any improvement.
- Although the model is not yet perfect, it is progressively having improvement after rounds of variables selection process.

- Considering the fact that we are predicting the locations of one species of fungi in a biodiverse continent based only on available covariates like Elevation and Forest, and have no information on all of the many other factors that would significantly influence fungi growth (e.g. humidity, moisture level, temperature, pH value, oxygen content, etc.), the current model has probably already been the best practically achievable version under various constraints.

3. Discussion and Conclusion

3.1 Data Inhomogeneity

From the various plots and analysis above, the inhomogeneity of the fungi occurrence can be concluded with strong evidence. The Quadrat Test and Hotspot Analysis provide certain significant evidence suggesting the data inhomogeneity. While the occurrences are concentrated in only a few high-density clusters, most regions in BC have no record of the fungi.

3.2 BC Windows Covariates

Relation with covariates has been studied. Elevation, Distance to Water, and Forest values are all appearing to be covariates and effectively affect fungi distribution. Relatively speaking, the rhohat plots of different covariates reveal that the Elevation and Forest demonstrate less fluctuation/variation with the fungi density, while the distance to water suggests a more complicated model and relation with the fungi.

3.3 Correction for Inhomogeneity

For the second moment analysis, both K-function and PCF analysis have revealed much less possibility of clustering after recorrecting for inhomogeneity. Although there are still potential clustering patterns identified at the beginning r-value range, strong evidence of autocorrelation among the fungi can hardly be observed during the rest of the higher r value range.

3.4 Model Fitting and Validation

Poisson Model allows us to define a framework for modeling point process, with the count as Poisson distributed random variable. In order to come up a model that could describe our model well, we first use Lurking Variable plot to gain some insight on whether or not we should include any new covariates on top of the existing model. We also use Partial Residual Plot to assess the possible relationship (say, quadratic or linear) between the covariates and the fungi distribution. This allows us to further refine our model effectively. Lastly, we use Likelihood Ratio Test to confirm whether the new model can actually increase the fit of our model. We then iterate this modeling fitting and validation process till we could find a better yet achievable model for our data.

Our Final Model:

$$\lambda_{ASC}(u) = e^{-1.94 \; - \; 7.137738e{-}03 \text{ elevation}(u) \; + \; 1.761325e{-}06 \text{ elevation}(u)^2 \; - \; 4.447410e{-}05 \text{ Dist}_{\text{W}}\text{ater}(u) \; - \; 1.296935e{-}12 \text{ Dist}_{\text{W}}\text{ater}(u)^3 \; - \; 3.0}$$

Form the above model, we can see that all of our features are significant in our model fitting : Elevation, HFI, Dist_Water and Forest are important explanatory factors of the intensity of Ascomycota. Quadrat.test() shows there is still a significant deviation from our model's prediction, so there is still room for improvement in the future. Such improvement can possible be achieved by adding more covariates and the degrees/complexity of the model. Although the model is not yet perfect, it is progressively having improvement after rounds of variables selection process. Moreover, We understand there are certain limitation when we are formulating our models. We are also aware that we have only a few available covariates at our disposal, which are likely insufficient for our model fitting given the complex biodiversity behaviour in the fungi world. Finally, we recognize there are many other variables (e.g. humidity, moisture level, temperature, pH value, oxygen content, etc.) not being able to be considered due to the limited information.

4. References:

- https://www.britannica.com/science/fungus/Outline-of-classification-of-fungi
- https://en.wikipedia.org/wiki/Penicillium
- https://en.wikipedia.org/wiki/Ascomycota