Assignment 02 - Exploratory Data Analysis

Dhun Sheth

08 April, 2024 13:30:13

Complete the following exercises within the lab period and submit to Canvas before leaving. In addition to the points detailed below, 5 points are assigned to the quality of the annotation and to the 'cleanliness' of the code and resulting pdf document.

```
library(spatstat) # loading spatstat library
```

Exercise 1 – 2 points

We will again be working with the BC Parks dataset, which contains information on the locations of Provincial Parks in British Columbia. The parks belong to 5 different regions. There is also information on elevation (in m) contained within the dataset.

- Import the BC park locations dataset and convert the data to a `ppp` object (being sure to include information on regions as marks). – 0.5 point(s)
- Plot the resulting `ppp` object. The marks need to be visually distinct. – 0.5 point(s)
- Inspect the spatial distribution of parks. Do you expect the process to be homogeneous? Justify why you came to this expectation. – 1 point(s)

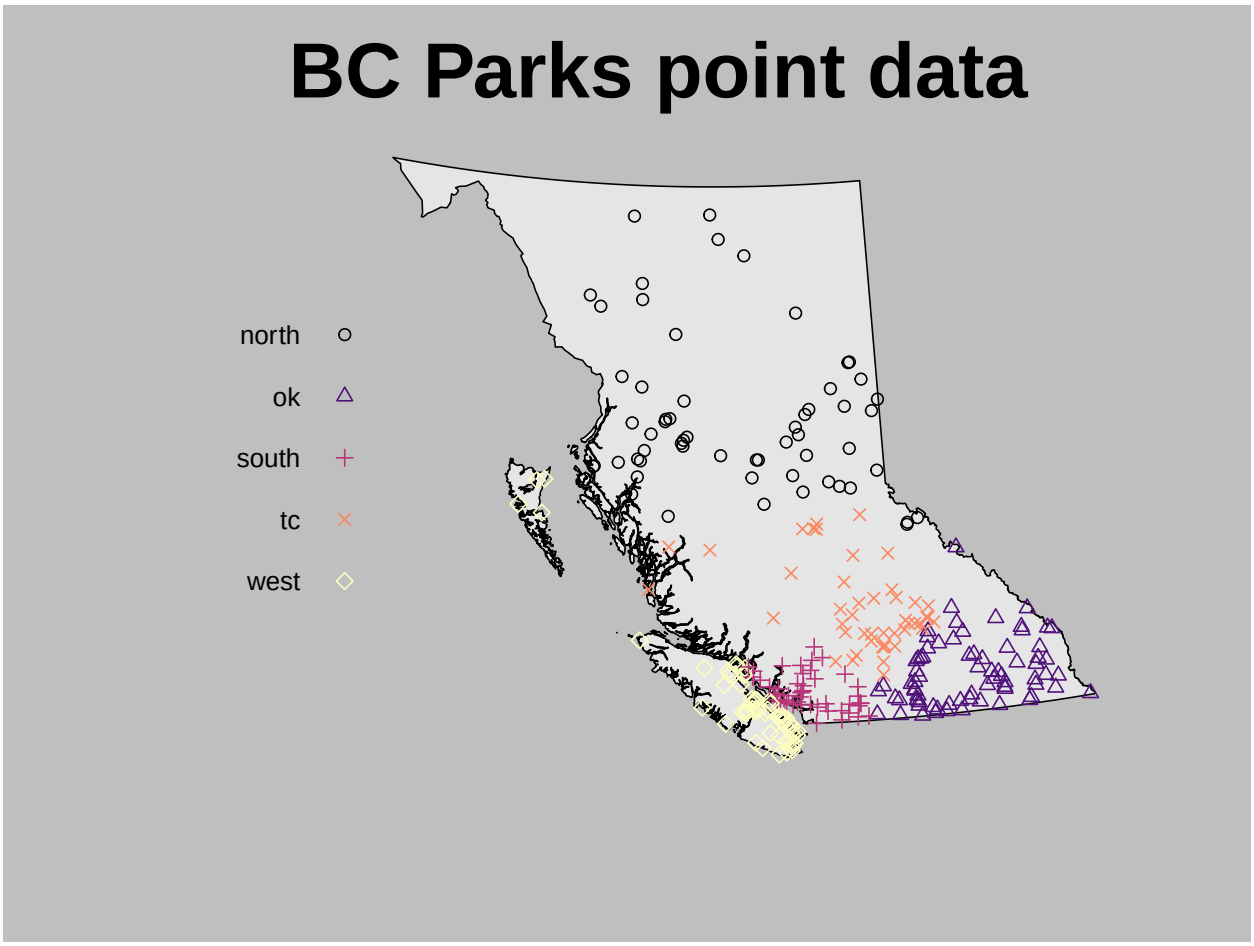
Note: You will need to load the `maptools` or `sp` packages and make use of the `as.owin()` function.

Loading Data

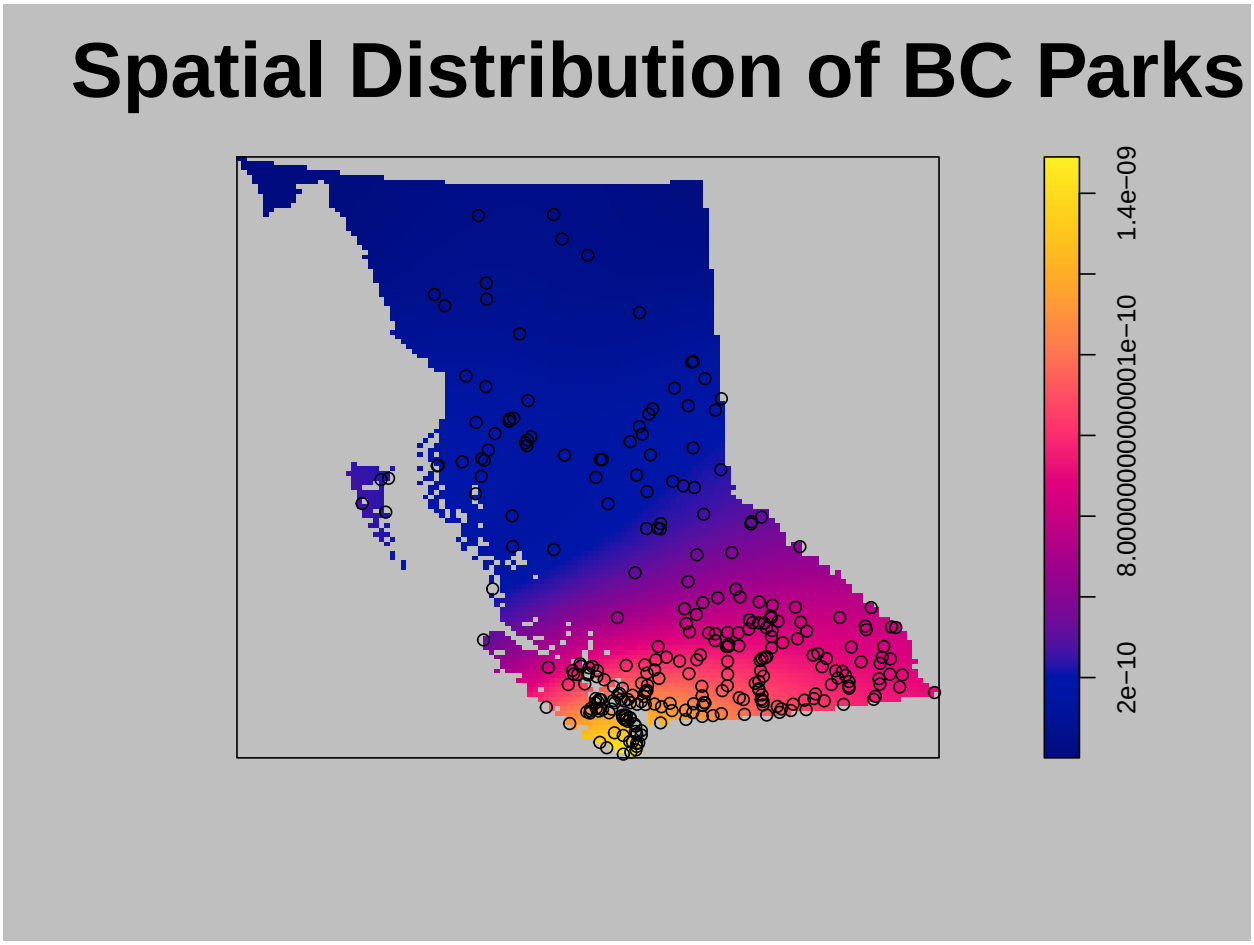
```
load('BC_Parks.Rda')
library(viridis)
library(colorspace)
library(sf)
library(spatstat.geom)
```

```
# Window
parks_win <- as.owin(DATA$Window)

# Converting data to ppp object
parks_ppp <- ppp(x = DATA$Parks$X, # X coordinates
                 y = DATA$Parks$Y, # Y coordinates
                 window = parks_win) # Observation window
```

```r
#Visualise the dataset
marks(parks_ppp) <- data.frame(Group = DATA$Parks$Region,
                               Park_Name = DATA$Parks$Park)

col_pal <- magma(length(unique(DATA$Parks$Region)))

plot(parks_ppp,
     which.marks = "Group",
     col = "grey90",
     cols = col_pal,
      par(bg="grey75", cex.main = 3),
     main = "BC Parks point data",
     legend=T)
```




```r
# density estimation
lambda_hat <- density(parks_ppp)
plot(lambda_hat, main='Spatial Distribution of BC Parks')
points(parks_ppp)
```



Based on the above spatial distribution, the data is not randomly distributed across the window but rather its concentrated in the bottom left of BC, suggesting the process is not homogeneous.

Exercise 2 – 2 points

- Under an assumption of homogeneity, what is the intensity of parks/km^2 in BC? – 1 point(s)
- Is the estimated intensity trustworthy? Why/why not? – 1 point(s)

Hint: see `?rescale`

```
I <- intensity(parks_ppp)
print(I)
```

```
## [1] 3.04768e-10
```

```
units <- unitname(parks_ppp)
print(units)
```

```
## unit / units
```

```r
# rescaling to Km
win_km <- rescale(Window(parks_ppp), 1000, "km")

# Intensity in Parks/km^2
I_km <- npoints(parks_ppp)/area(win_km)
print(I_km)
```

```
## [1] 0.000304768
```

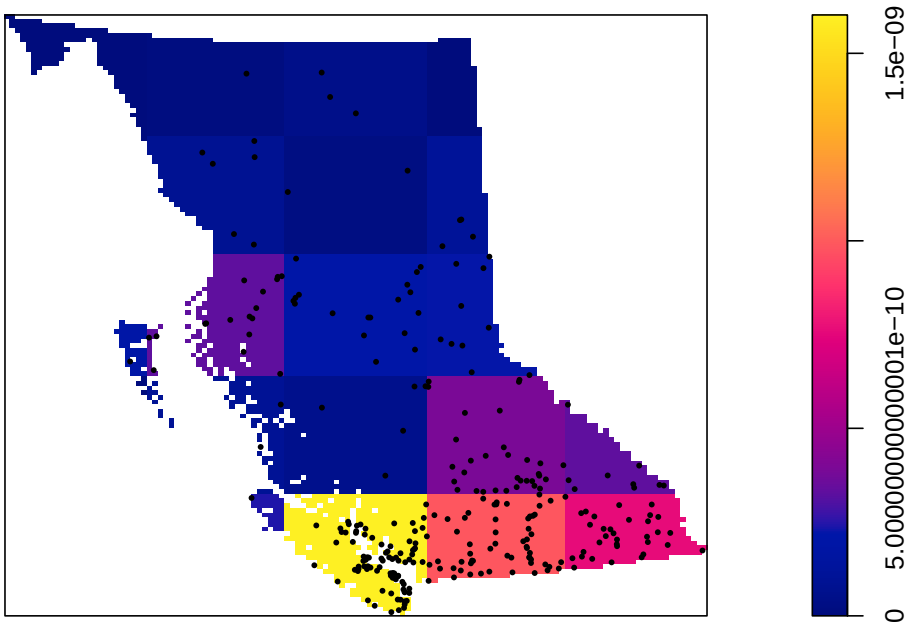
The intensity is calculated under the assumption of homogeniety which as determined earlier was not accurate, thus, the above calculated intensity should not be trusted.

Exercise 3 – 2 points

- Use a quadrat test to determine whether the assumption of homogeneity is met. – 1 point(s) Note: Be sure to set the number of quadrats appropriately, to avoid issues with the quadrat test.
- Visualise both the quadrats and estimated intensity, being sure to include the points in each figure. – 1 point(s)
- Is the estimated intensity from exercise 2 trustworthy, and why? – 1 point(s)

```r
# Split into a 5 by 5 quadrat and count points
Q <- quadratcount(parks_ppp,
                  nx = 5,
                  ny = 5)

plot(intensity(Q, image = T),
     main = "BC Parks intensity")

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

```
## Plotting the first column of marks
```




```
quadrat.test(Q)
```

```
## Warning: Some expected counts are small; chi^2 approximation may be inaccurate

##
##  Chi-squared test of CSR using quadrat counts
##
## data:
## X2 = 532.63, df = 20, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 21 tiles (irregular windows)
```

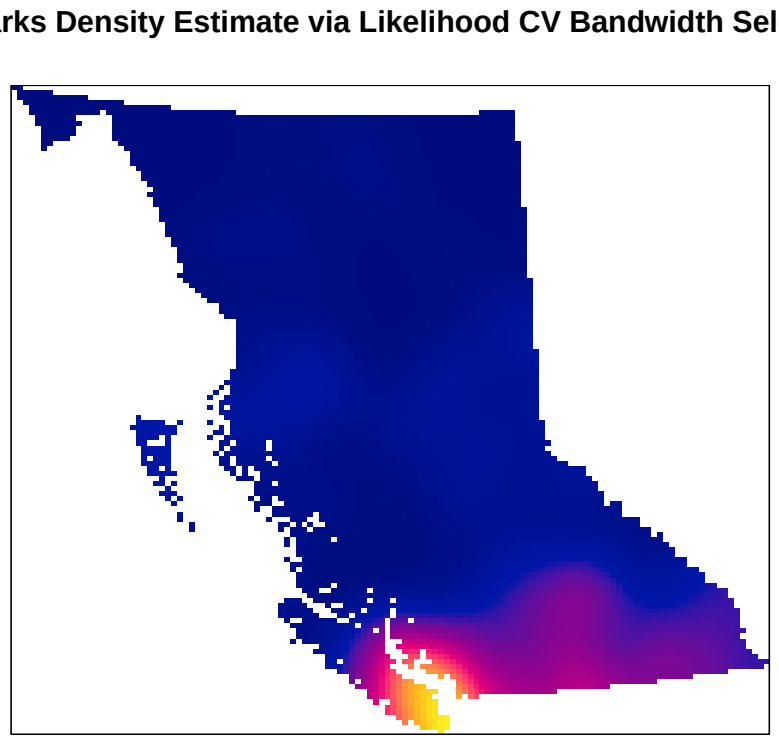
Based on the above quadrat test, the small p-value suggests the null hypothesis should be rejected, ie. there is significant evidence against the assumption of homogeneity, therefore, the intensity calculated in Exercise 2 should not be trusted because it was calculated under the assumption of homogeneity which is false. In addition, looking at the plot of quadrat and estimated intensity in each quadrat, it also shows the intensity is not homogeneous.

Exercise 4 – 4 points

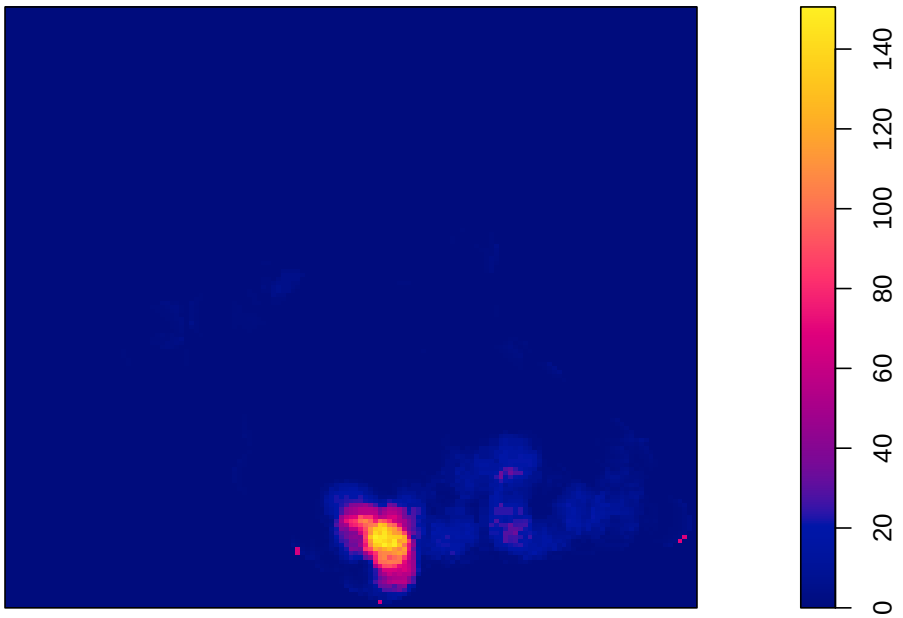
- Estimate the intensity using kernel estimation with likelihood cross validation bandwidth selection. – 1 point(s)

- Perform hotspot analysis to identify locations of elevated intensity. – 1 point(s)
- Visualise the output (be sure to include the window). – 1 point(s)
- Based on the estimated intensity and hotspot analysis, where would choose to go if you were planning a vacation to tour different provincial parks. – 1 point(s)

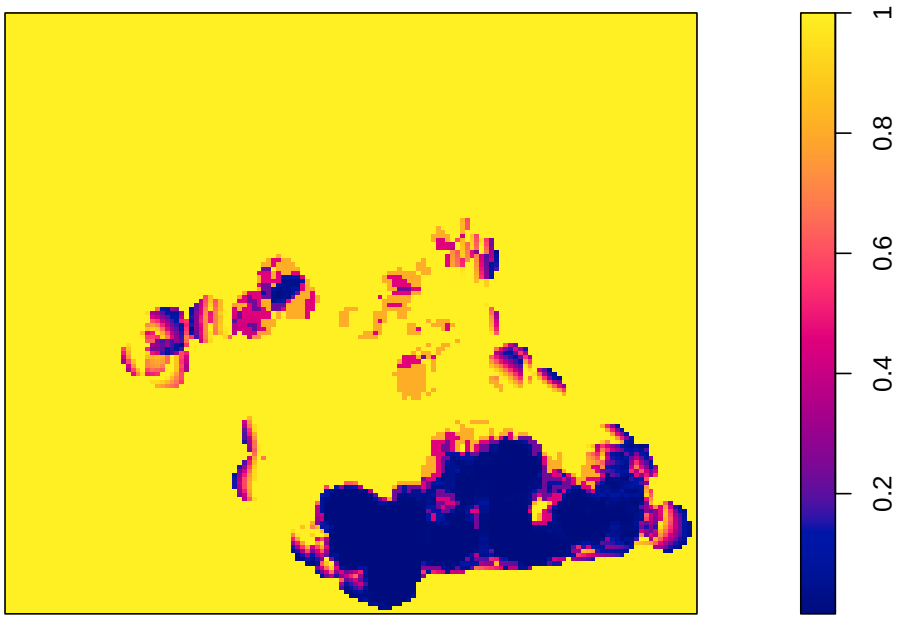
```
kde <- density(parks_ppp, sigma = bw.ppl)

# Likelihood Cross Validation Bandwidth Selection
plot(density(parks_ppp, sigma = bw.ppl),
     ribbon = F,
     main = "BC Parks Density Estimate via Likelihood CV Bandwidth Selection")
```




```
# Hotspot analysis
# Estimate R
R <- bw.ppl(parks_ppp)

# Calculate test statistic
LR <- scanLRTS(parks_ppp, r = R)

# Plot the output
plot(LR)
```




```
# Compute local p-values
pvals <- eval.im(pchisq(LR,
                        df = 1,
                        lower.tail = FALSE))

# Plot the output
plot(pvals, main = "Local p-values")
```



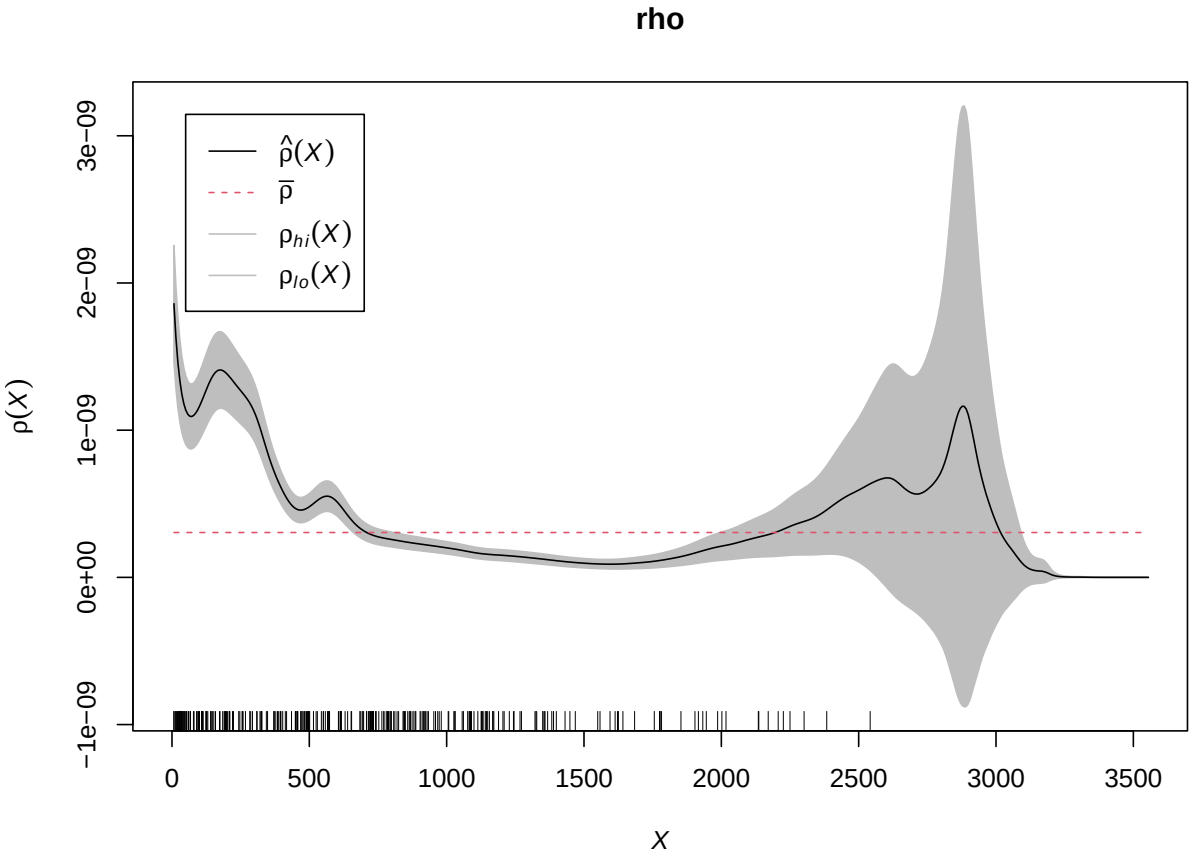
If I was planning a vacation to tour different provincial parks, I would visit most south-west tip of BC because there is a large concentration (ie. lots of hotspots) of BC parks in that area versus the northern parts of BC.

Exercise 5 – 3 points

- Estimate ρ for the locations of parks as a function of elevation. – 1 point(s)
- Plot ρ vs. elevation. Be sure that the x-axis for elevation does not go below 0. – 1 point(s)
- Do you think that there is a relationship between elevation and park intensity? Use the results/data to support your statements. – 0.5 point(s)
- Would you be more or less likely to find a park at 1500m compared to the average intensity of parks across B.C.? Why? – 0.5 point(s)

Note: Estimating rho can be slow ($\sim$ 1-2 min). Be sure to leave enough time for the document to knit.

```
# Estimate Rho
rho <- rhohat(parks_ppp, DATA$Elevation)

plot(rho, xlim = c(0, max(rho$X)))
```



Based on the plot of rho vs. elevation, there does seem to be some non-linear relationship between elevation and park intensity.

Because rho is below the average intensity of parks across BC (ie. red doted line), I would expect it to be less likely to find a park at 1500m compared to the average across BC.

Exercise 6 – 5 points

- Using Ripley's K-function, test for a significant (i.e., $\alpha = 0.05$) correlation between park locations. – 4 point(s)
- Is there any evidence of correlations in park locations? Why? – 1 point(s)

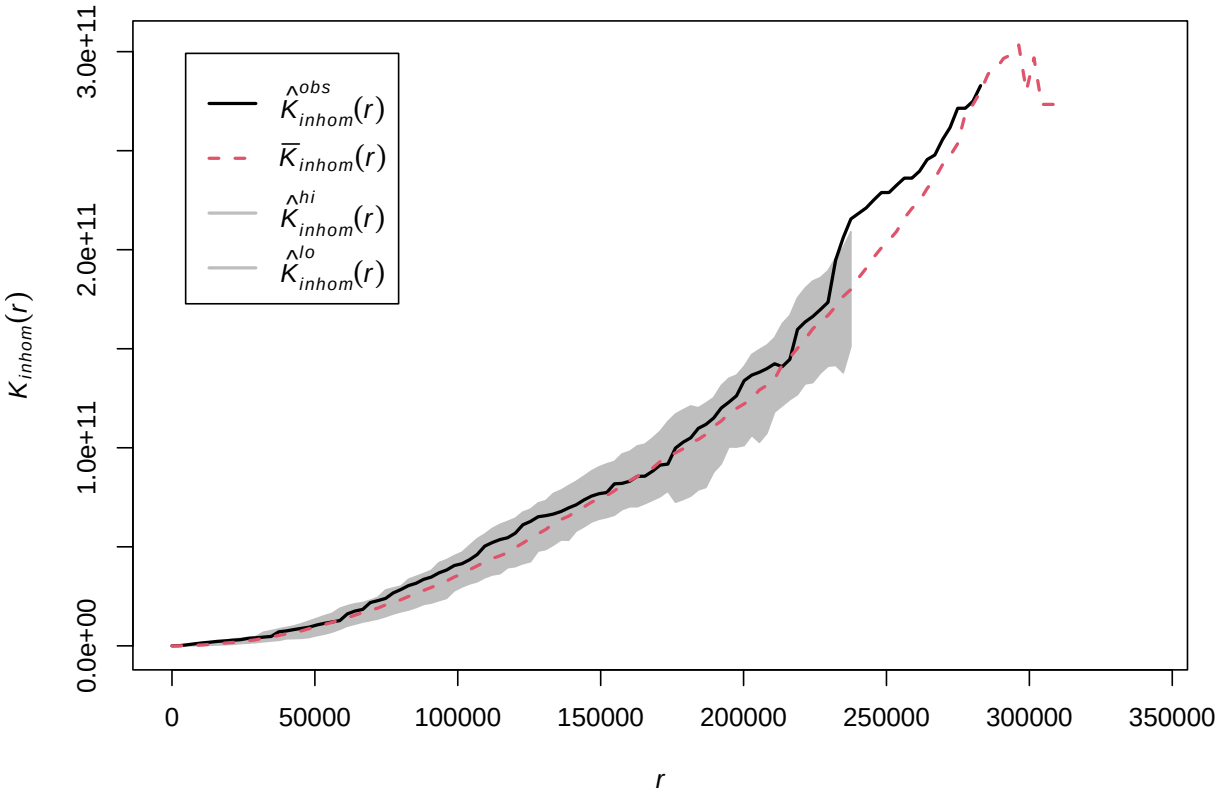
Notes: Use border corrections (i.e., `correction="border"`) and be sure the estimators assumptions are being met.

```
# Estimate the empirical k-function
#Estimate intensity
lambda_parks <- density(parks_ppp, bw.ppl)

Kinhom_parks <- Kinhom(parks_ppp, lambda_parks)

#Estimate a strictly positive density
lambda_parks_pos <- density(parks_ppp,
```

```
                          sigma=bw.ppl,
                          positive=TRUE)

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

```
## Warning in envelope.ppp(parks_ppp, Kinhom, simulate =
## expression(rpoispp(lambda_parks_pos)), : fix.n and fix.marks were ignored,
## because 'simulate' was given

## 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(E_parks_inhom,
     main = "",
     lwd = 2)
```



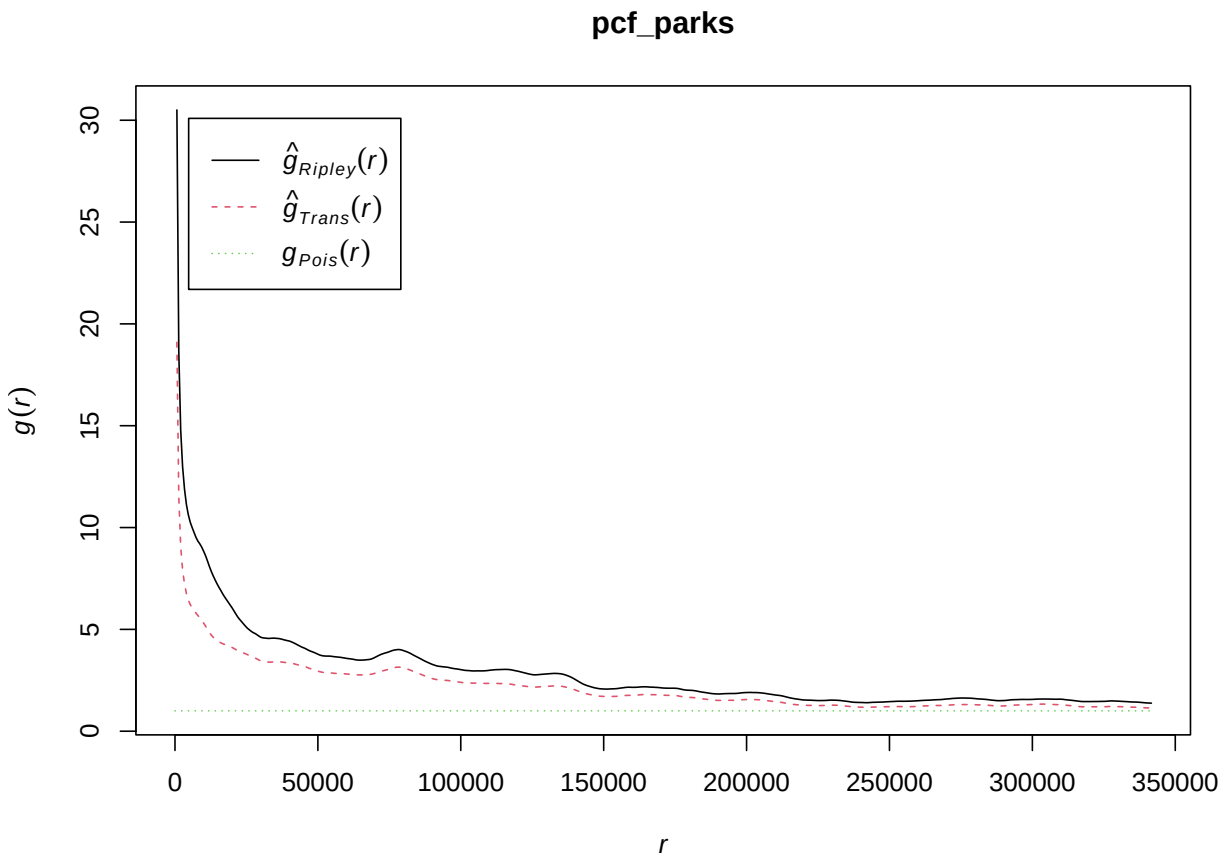
Under the empirical k-function estimation, this assumes estimators are homogeneous, however, as we have established previously, this is not true. Relaxing this assumption using Kinhom() and correcting for inhomogeneity (simulating 19 times which is equivalent to an alpha=0.05), we see the deviations are less meaningful. This suggests there is no correlation between park locations, but rather the correlation likely exists due to the relationship with covariates.

Exercise 7 – 3 points

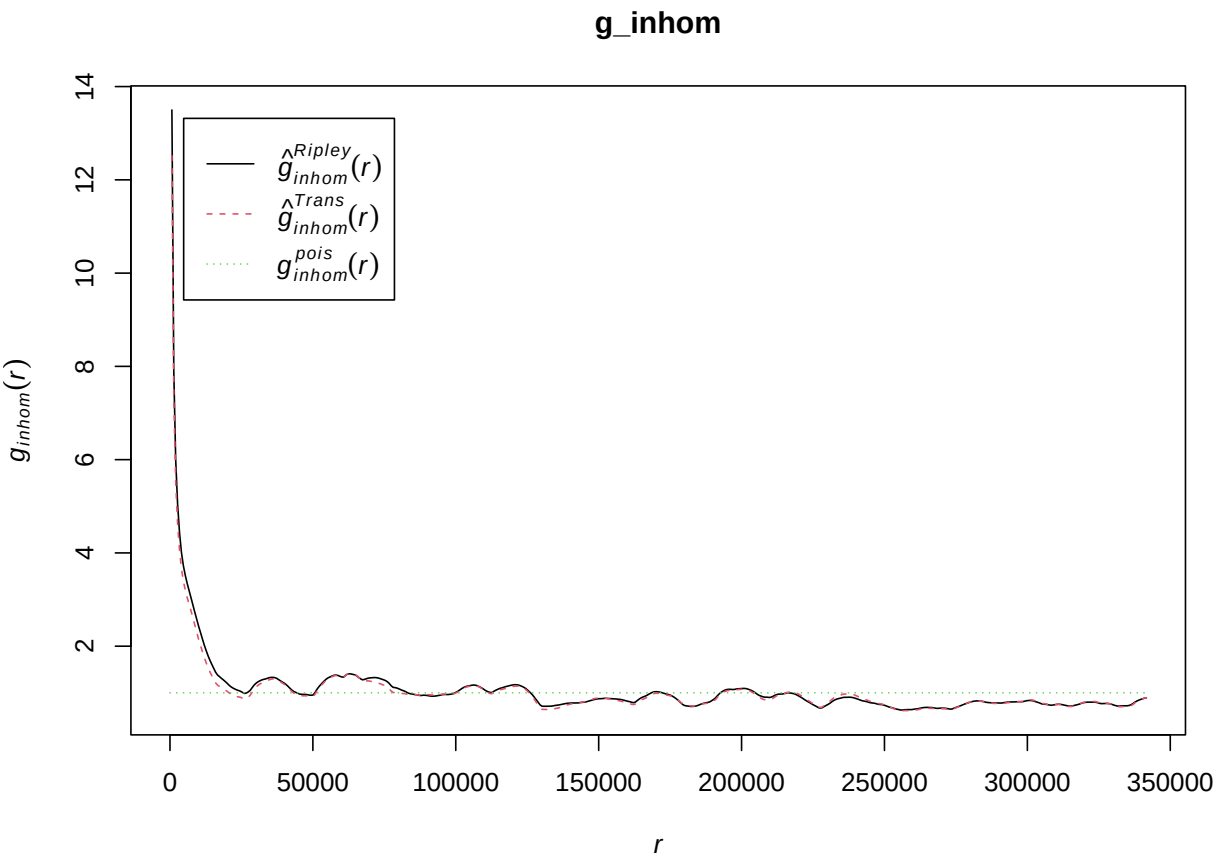
- Using simulation envelopes, estimate both the homogeneous and inhomogeneous pair correlation functions. – 1.5 point(s)
- Visualise the results. – 0.5 point(s)
- Are the estimates comparable? Which of these would you use to draw conclusions about the clustering of provincial parks? – 0.5 point(s)
- Are parks in BC clustered? – 0.5 point(s)

Note: These steps can be slow (∼ 1-2 min). Be sure to leave enough time for the document to knit.

```
pcf_parks <- pcf(parks_ppp)
plot(pcf_parks)
```




```
g_inhom <- pcfinhom(parks_ppp)
plot(g_inhom)
```



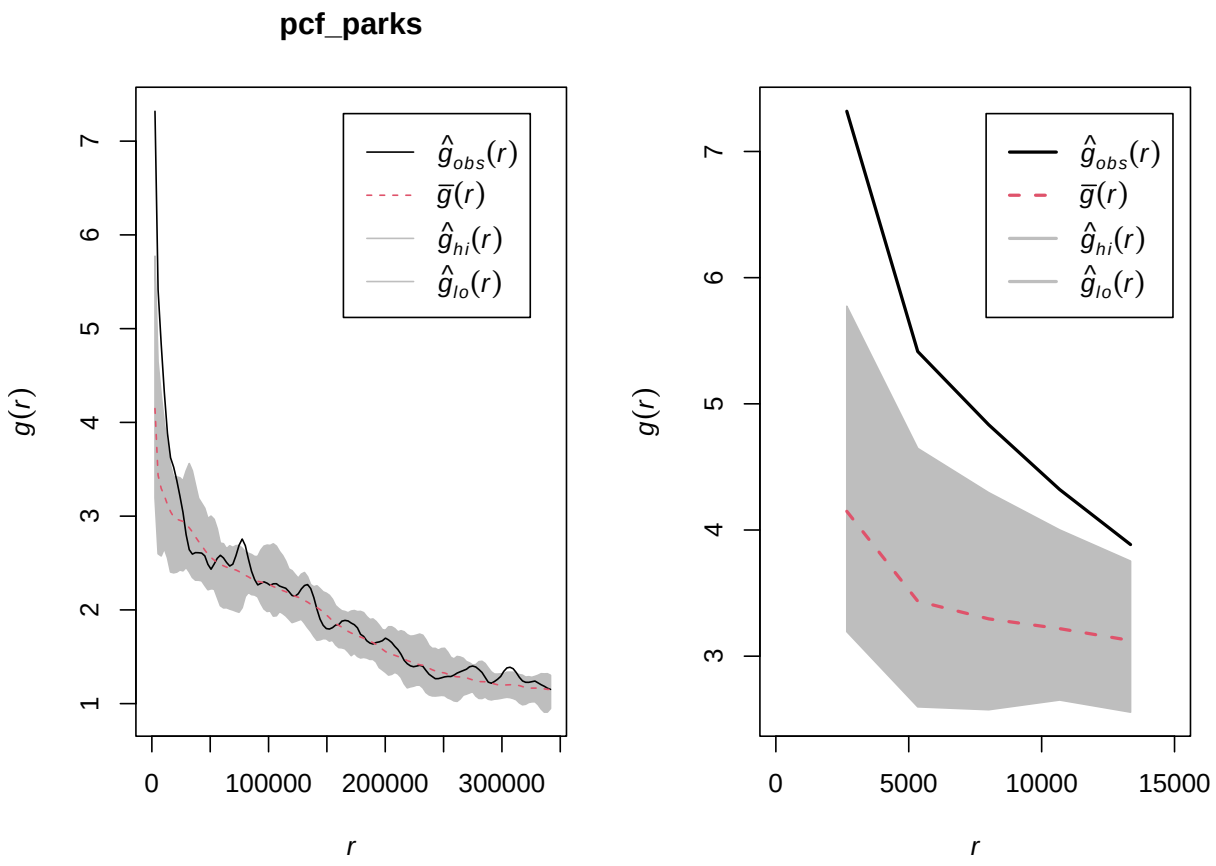

```
pcf_parks <- envelope(parks_ppp,
                          pcf,
                          simulate = expression(rpoispp(lambda_parks_pos)),
                          rank = 1,
                          nsim = 19)
```

```
## 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.
```

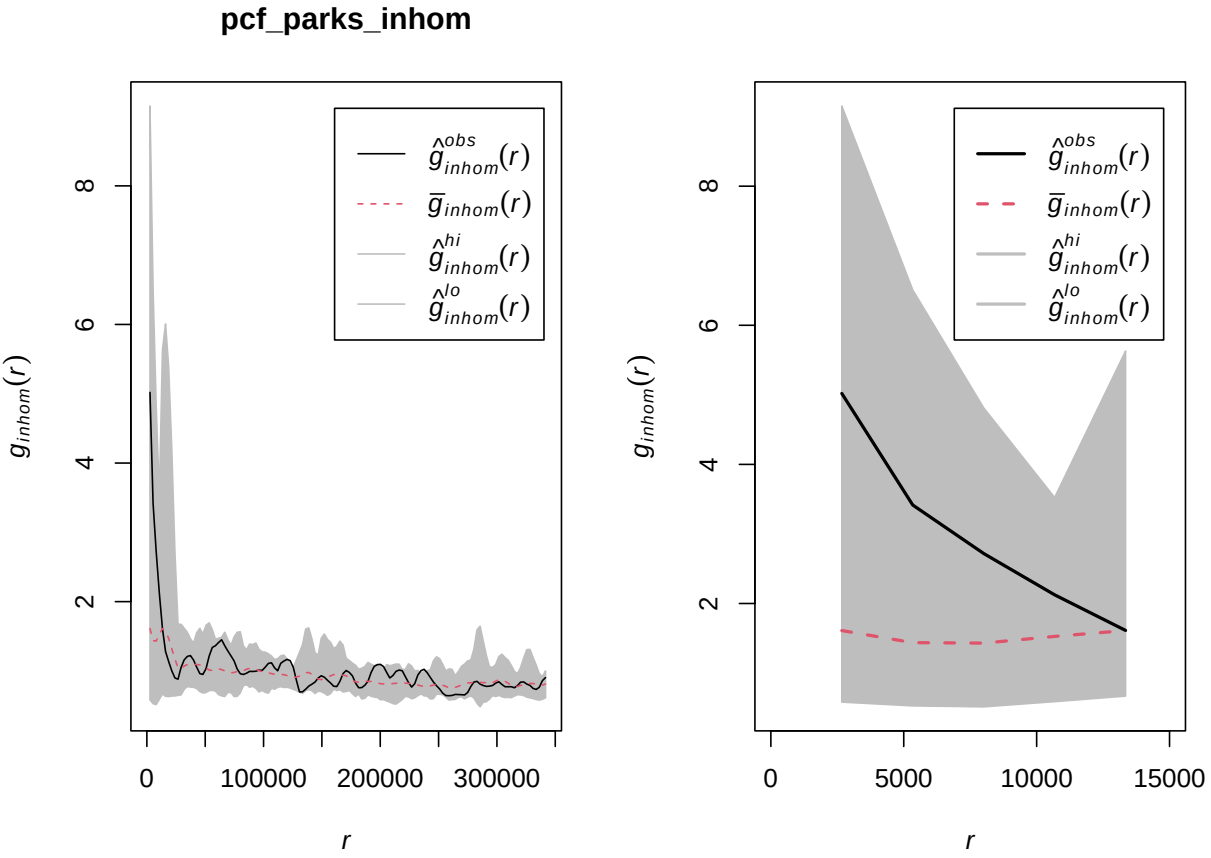
```
pcf_parks_inhom <- envelope(parks_ppp,
                          pcfinhom,
                          simulate = expression(rpoispp(lambda_parks_pos)),
                          rank = 1,
                          nsim = 19)
```

```
## 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.
```

```r
par(mfrow = c(1,2))
plot(pcf_parks)
# Zoom in
plot(pcf_parks,
      xlim = c(0,15000),
      main = "",
      lwd = 2)
```




```r
par(mfrow = c(1,2))
plot(pcf_parks_inhom)
# Zoom in
plot(pcf_parks_inhom,
      xlim = c(0,15000),
      main = "",
      lwd = 2)
```



The results are not comparable because the empirical deviations appear weaker for the inhomogeneous case compared to the homogeneous case, and persist for only 90,000 meters.

By adding simulations, we can determine if the clustering results are significant or not. Again in the homogeneous case, the clustering seems to be significant for r ~15,000 meters, whereas when corrected for inhomogeniety, the results are within the error bands, indicating the clustering being observed is not significant.

I would draw conclusions from the pair correlation corrected for inhomogeneity (because we know the homogeneity assumption has been violated) with simulation envelopes because it shows significance of clustering if it exists, and in this case it was not.

I would conclude, the parks in BC are not clustered, based on the above arguments.

Exercise 8 – 3 points

- Based on these descriptive statistics, what have you learned about the spatial distribution of parks in BC?

Based on the above descriptive statistics, I have learned that parks in BC do not have a homogeneous spatial distribution and are concentrated more in the South.

I have learned that BC park locations are not correlated with each other but rather some covariates, and the concentration (intensity) of park locations have some non-linear relationship with elevation.

I have also learned that parks in BC are not clustered.