DATA589 Project

Varshita Kyal, Shveta Sharma, Ujjwal Upadhyay

25 April, 2023

Introduction

We selected **Canada Goose** species dataset in British Columbia region for the spatial analysis. In GBIF database, this species has approximately 17326212 occurrences. However, we have filtered the data set based on BC, Canada only. When filtered the dataset, we fetched that Canada Goose species in BC has 500 rows and 77 columns of entries.

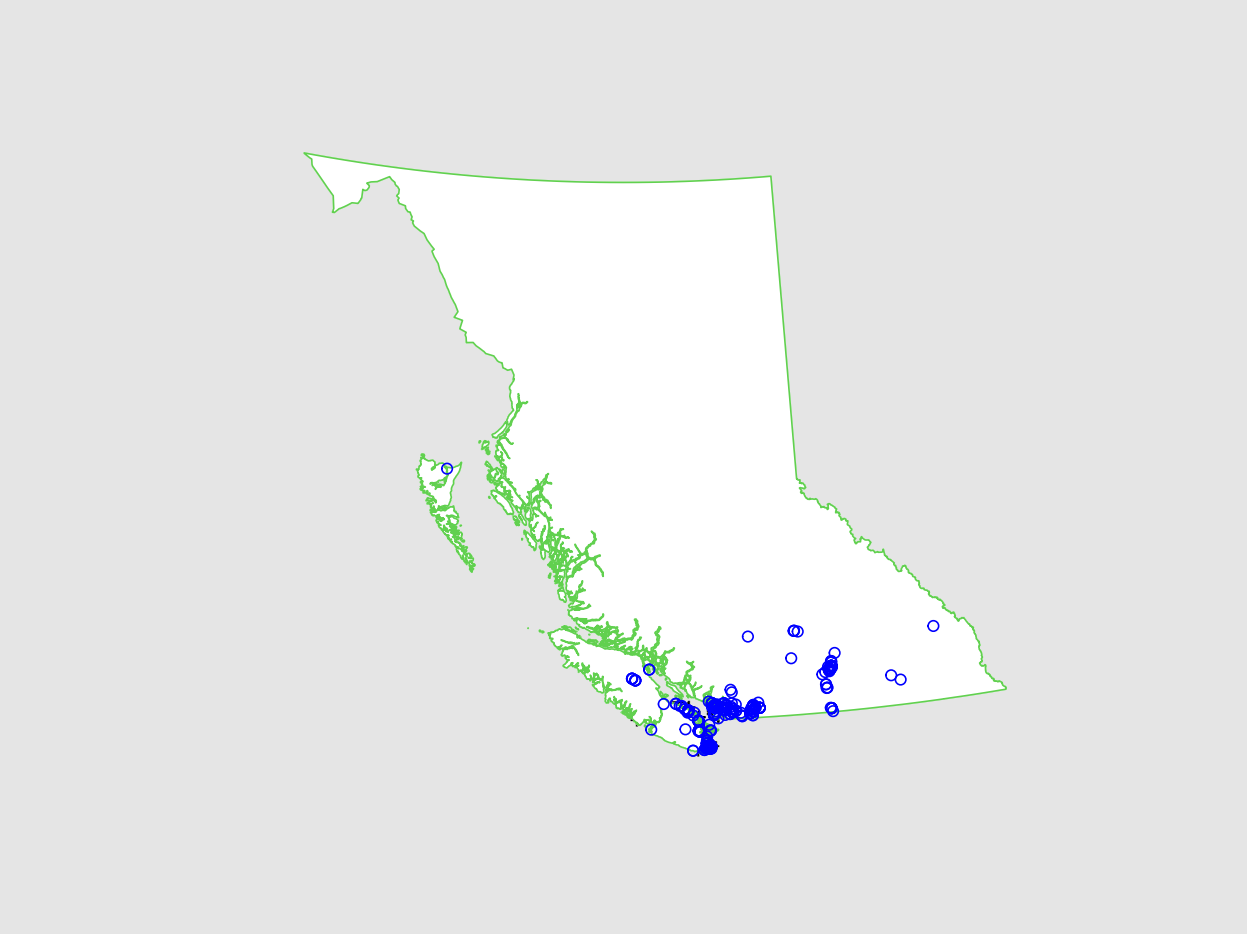

Figure 1: Occurences of Canada Goose in BC

Here we have plotted all the occurrences of Canada Goose in the BC region and we can see that the species are mostly present in the south and south-west region 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

The described analysis involves examining Canada Goose data in British Columbia, using various R packages such as `rgbif`, `sp`, and `spatstat`. The data is obtained from the Global Biodiversity Information Facility (GBIF) databases and latitude and longitude data is extracted using `rgbif`, and then converted to a usable format with `sp`. Covariate data including elevation, forest cover, HFI, and distance to water is also obtained to create a ppp object using `spatstat`.

To analyze the data, first moment analysis is performed using the quadrat test and hotspot analysis from `spatstat`, which provide information about the homogeneity of the Canada Goose point process. Second moment analysis is performed using the Ripley's K-function and pair correlation function, which can reveal clustering tendencies in the data. Overall, the analysis is a comprehensive spatial exploration of the Canada Goose data in BC, using various techniques and R packages.

First Moment Analysis

With some point data in hand, the first summary statistics we want to calculate is the average number of points per unit area (i.e., our 'expectation', or 'first moment').

```
## [1] 4.450245e-10
```

The average intensity 4.397517e-10 points (Canada Goose) per square unit and this does not explain the observance of Canada Goose in a meaningful way.

Therefore, we can check if the dataset is inhomogeneous or not.

Quadrat counting for checking inhomogeneity

When λ is spatially varying, $\lambda(u)$ can be estimated nonparametrically by dividing the window into sub-regions (i.e., quadrats) and using our simple points/area estimator.

Clearly, the assumption of homogeneity is not appropriate for this dataset as the Canada Goose tends to be clustered in certain areas, whereas others have none at all. Quadrat counting suggests a spatially varying, inhomogeneous $\lambda(u)$, but point processes are stochastic and some variation is expected by chance alone.

We can therefore test for significant deviations from complete spatial randomness (CSR) using a χ^2 test

```
##
##  Chi-squared test of CSR using quadrat counts
##
## data:  parks_ppp
## X2 = 3083.7, df = 20, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 21 tiles (irregular windows)
```

The small p-value suggests that there is a significant deviation from homogeneity. The p-value doesn't provide any information on the cause of inhomogeneity, however, and significant deviations can be due to the processes truly being inhomogenous, but also due to a lack of independence between points.

Kernel estimation

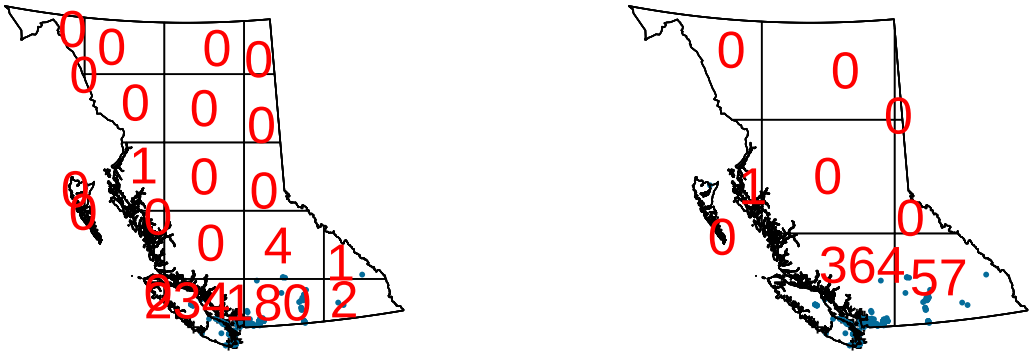

Figure 2: Quadrat counts of Canada Goose occurences, left 5x5, right 3x3

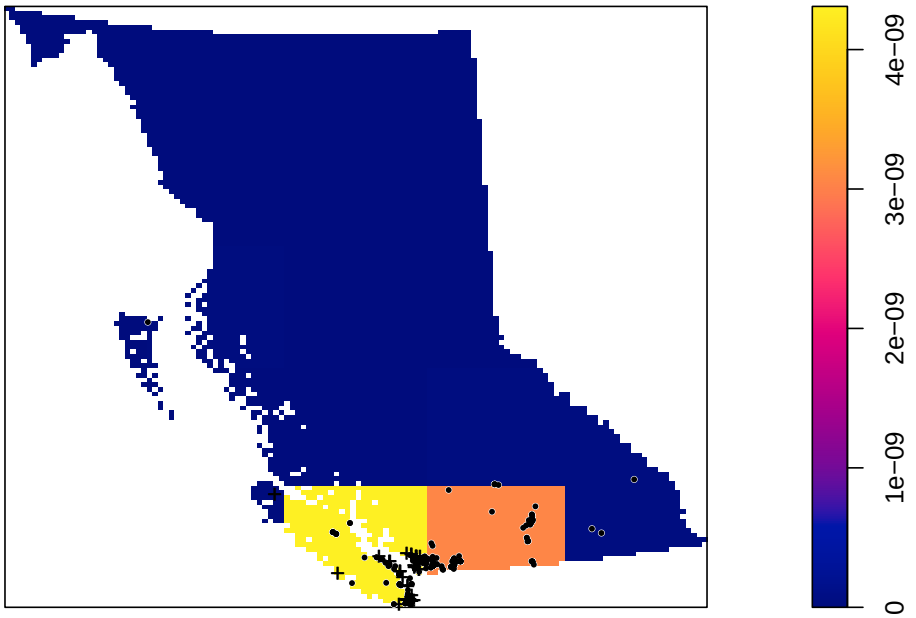

Figure 3: Quadrat counts with intensity of Canada Goose occurences (5x5)

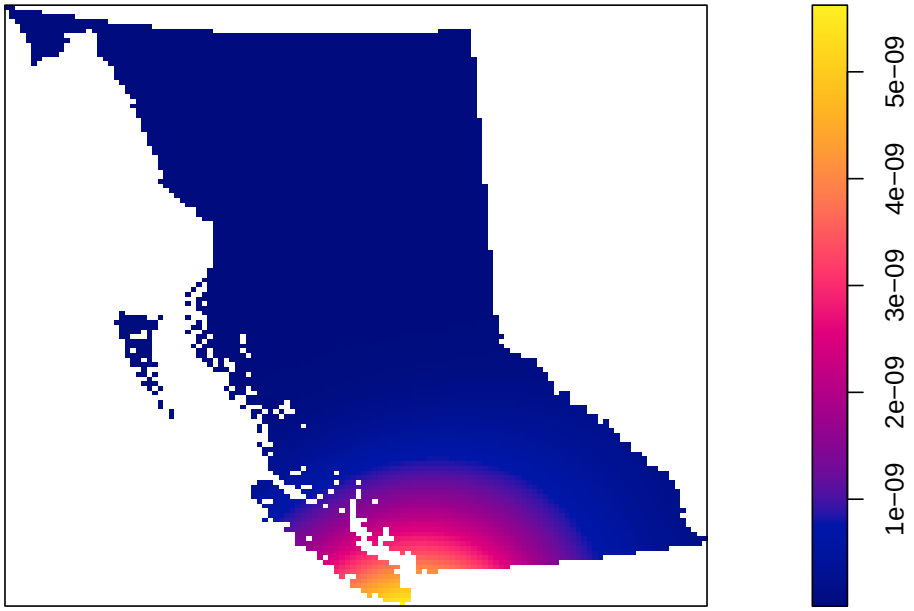

Figure 4: Kernel Estimation of Canada Goose occurences

A spatially varying, $\lambda(u)$ can also be estimated non-parametrically by kernel estimation.

The Figure 4 above shows comparable results as compared to Figure 2 but with finer-scale resolution.

Kernel estimation uses a single bandwidth across the whole dataset, but this can be relaxed by using adaptive smoothing via the `adaptive.density()` function.

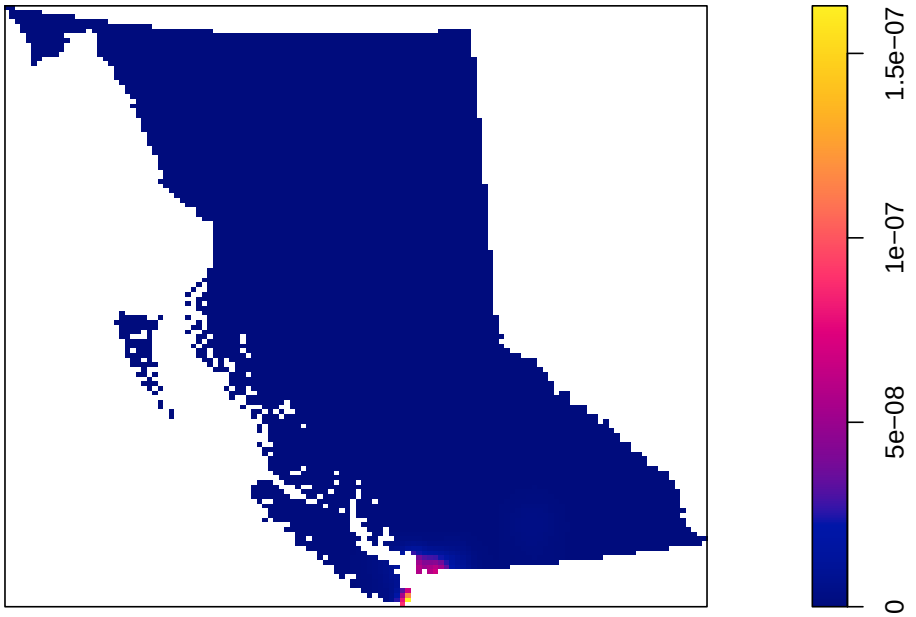

Figure 5: Adaptive kernel estimate of intensity

Hot spot analysis

If the intensity is inhomogeneous, we often want to identify areas of elevated intensity (i.e., hotspots).

From the p-values of Figure 6, we can see that southern region of BC has more occurences of Canada Goose which is again comparable with Figure 2 results.

Relationships with covariates

Our data includes 4 covariates we can explore: elevation, forest cover, human footprint inventory (HFI), and distance to water.

We are usually interested in determining whether the intensity depends on a covariate(s). Testing for relationships with covariates, we are assuming that λ is a function of Z, such that

$$\lambda(u) = \rho(Z(u))$$

A non-parametric estimate of ρ can be obtained via kernel estimation, available via the `rhohat()` function.

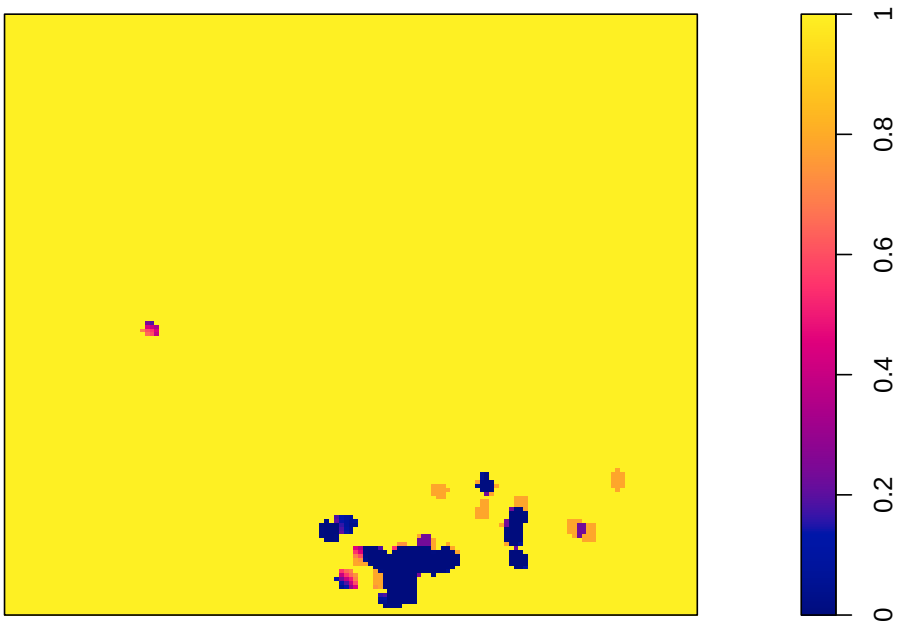

Figure 6: Local p-values

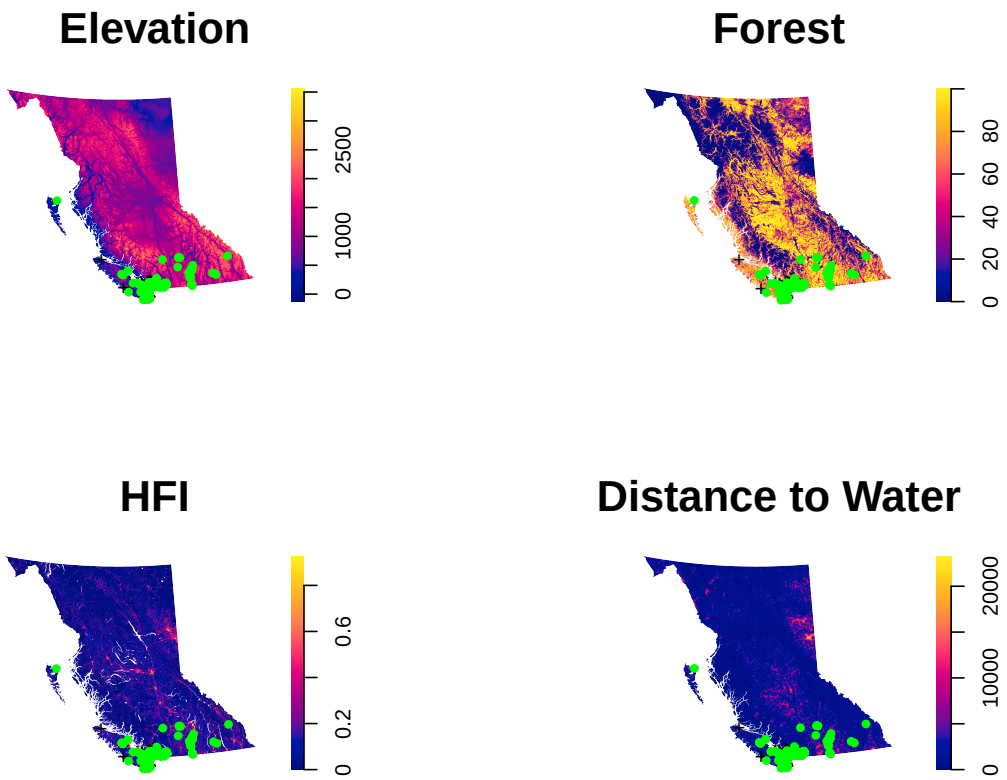

Figure 7: 4 covariates explaining the occurence of Canada Goose in BC

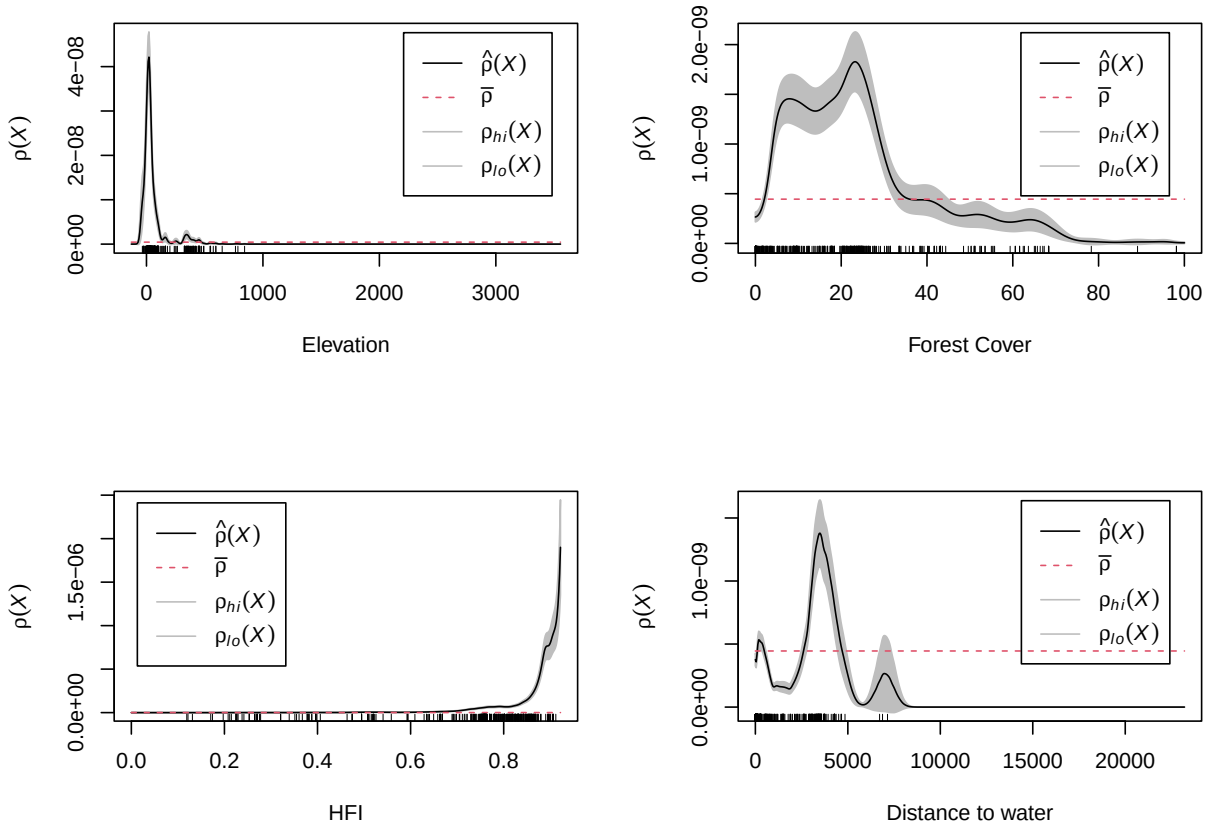

Figure 8: Comparison of Intensity with all the four covariates

Figure 8 suggest that there is likely to be a relationship between the number of Canada Goose species wrt elevation, forest cover, human footprint index and distance to water. These relationships appear to be non-linear with the highest intensity of species occurring at intermediate elevation, forest cover, human footprint index and distance to water.

Second Moment Descriptives

Morisita's index

As Morisita's index assumes homogeneity and it is already evident that our dataset is inhomogeneous. Therefore, we can skip this method for now. However, Morisita's index serve as a useful visual diagnostic tool when derivation assumed homogeneity.

Ripley's K-function

Morisita's index describes correlations based on the rate at which pairs of points are found 'close' together, but if we're interested in the spacing (or distance) between points.

Ripley's K-function provides information on whether there are significant deviations from independence between points and also assumes homogeneity. 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 by weighting the data based on $\lambda(u)$ (the `pcfinhom` function)

```
## 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.
```

```
## 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.
```

Pair Correlation Function

Ripley's K-function provides information on whether their are significant deviations from independence between points, but provides limited information on the behaviour of the process.

The estimator of the pair correlation function also assumes homogeneity. Here again, we can relax this assumption via the `pcfinhom()` function.

Again, when corrected for inhomogeneity, the empirical deviations appear weaker than in the homogeneous case.

PPP Pre-analysis

Collinearity

```
##             ..1         ..2         ..3         ..4
## ..1  1.00000000 -0.26225376 -0.26625626 -0.03493453
## ..2 -0.26225376  1.00000000  0.06618592  0.04818598
## ..3 -0.26625626  0.06618592  1.00000000  0.13246899
## ..4 -0.03493453  0.04818598  0.13246899  1.00000000
```

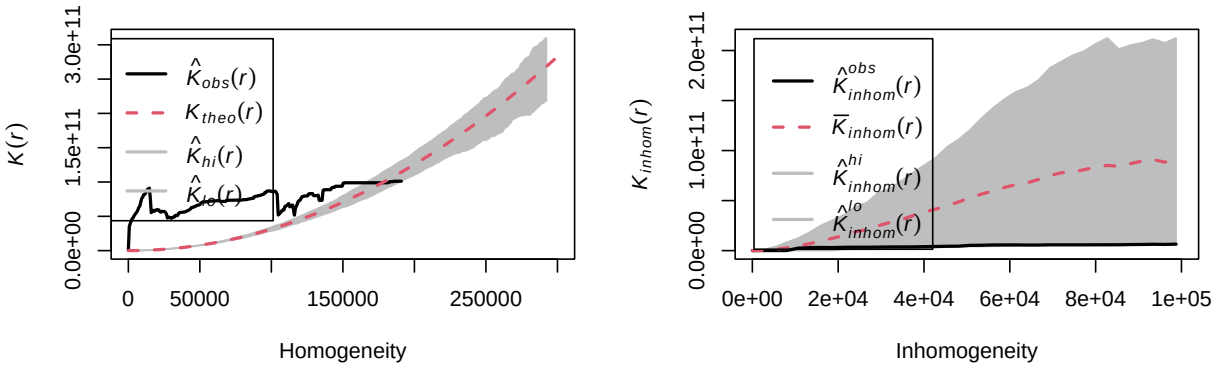

Figure 9: Ripley's K function with border correction assuming homogeneity and inhomogeneity

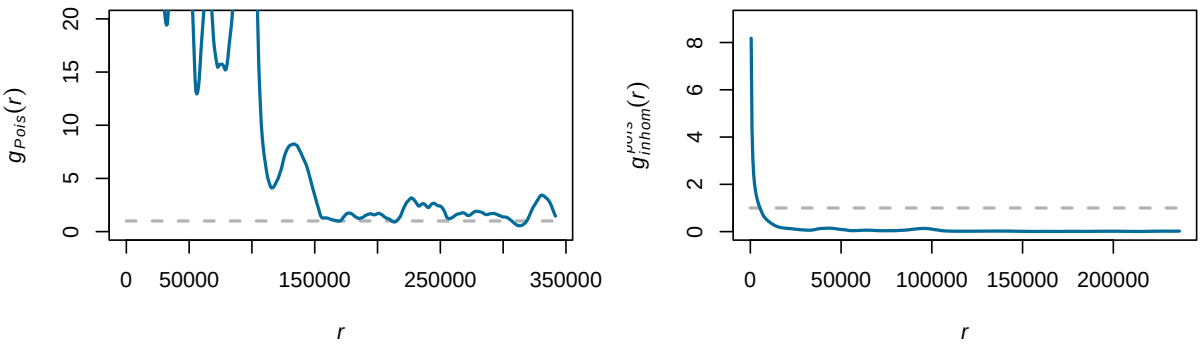

Figure 10: Pair correlation function assuming homogeneity and inhomogeneity

Model Fitting

Now, we already know that our data is inhomogeneous, therefore modelling an inhomogeneous Poisson point process means specifying the form of the model in terms of

$$\lambda(u) = e^{\alpha+\beta_1 Z_1(u)+\beta_2 Z_2(u)+...+\beta_i Z_i(u)}$$

The correlation coefficients are relatively weak, so we can proceed with Poisson model with interaction terms to fit our model as an initial guess.

Based on these initial analysis, a reasonable form for the model might be

$$\lambda_{BC\_Parks}(u) = e^{\beta_0+\beta_1[elevation(u)+forestcover(u)+hfi(u)+dist\_water(u)]+\beta_2[elevation(u)^2+forestcover(u)^2+hfi(u)^2+dist\_water(u)^2]}$$

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Elevation_scaled + I(Elevation_scaled^2) + Forest +
## I(Forest^2) + HFI + I(HFI^2) + Dist_Water_scaled + I(Dist_Water_scaled^2)
##
## Fitted trend coefficients:
##           (Intercept)       Elevation_scaled  I(Elevation_scaled^2)
##         -2.677919e+01          -1.824749e+00           2.297263e-01
##                Forest            I(Forest^2)                    HFI
##         -1.448102e-02           3.796748e-05           1.234093e+01
##              I(HFI^2)      Dist_Water_scaled I(Dist_Water_scaled^2)
##         -6.664687e+00           4.411578e-02          -8.401530e-02
##
##                            Estimate         S.E.       CI95.lo       CI95.hi
## (Intercept)            -2.677919e+01 5.276265e-01 -2.781332e+01 -2.574506e+01
## Elevation_scaled       -1.824749e+00 5.573567e-01 -2.917148e+00 -7.323496e-01
## I(Elevation_scaled^2)   2.297263e-01 1.824459e-01 -1.278611e-01  5.873136e-01
## Forest                 -1.448102e-02 6.721227e-03 -2.765439e-02 -1.307660e-03
## I(Forest^2)             3.796748e-05 9.659756e-05 -1.513603e-04  2.272952e-04
## HFI                     1.234093e+01 1.406587e+00  9.584066e+00  1.509779e+01
## I(HFI^2)               -6.664687e+00 1.276708e+00 -9.166988e+00 -4.162385e+00
## Dist_Water_scaled       4.411578e-02 6.620229e-02 -8.563832e-02  1.738699e-01
## I(Dist_Water_scaled^2) -8.401530e-02 3.872927e-02 -1.599233e-01 -8.107329e-03
##                         Ztest        Zval
## (Intercept)               *** -50.7540703
## Elevation_scaled           **  -3.2739333
## I(Elevation_scaled^2)           1.2591476
## Forest                     *  -2.1545208
## I(Forest^2)                     0.3930480
## HFI                       ***   8.7736668
## I(HFI^2)                  ***  -5.2202126
## Dist_Water_scaled               0.6663785
## I(Dist_Water_scaled^2)     *  -2.1692974
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.6% (14 out of 2321) of
## the quadrature points
##
## *** Fitting algorithm for 'glm' did not converge ***
```

Considering coefficients which are statistically significant, and suggest that $\lambda_{BC\_CanadaGoose}$ can be estimated as:

$$\lambda_{BC\_Parks}(u) = e^{\beta_0 + \beta_1[elevation(u) + forestcover(u) + hfi(u)] + \beta_2 hfi(u)^2}$$

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Elevation_scaled + Forest + HFI + I(HFI^2)
##
## Fitted trend coefficients:
##      (Intercept) Elevation_scaled           Forest              HFI
##     -27.44452373      -2.61890733      -0.01144439      11.82082034
##         I(HFI^2)
##      -6.15049026
##
##                       Estimate        S.E.      CI95.lo       CI95.hi Ztest
## (Intercept)      -27.44452373 0.423327465 -28.27423031 -26.614817142   ***
## Elevation_scaled  -2.61890733 0.160255136  -2.93300162  -2.304813035   ***
## Forest            -0.01144439 0.003209137  -0.01773418  -0.005154593   ***
## HFI               11.82082034 1.315201667   9.24307244  14.398568241   ***
## I(HFI^2)          -6.15049026 1.207864258  -8.51786071  -3.783119817   ***
##                       Zval
## (Intercept)      -64.830482
## Elevation_scaled -16.342112
## Forest            -3.566188
## HFI                8.987839
## I(HFI^2)          -5.092038
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.6% (14 out of 2321) of
## the quadrature points
##
## *** Fitting algorithm for 'glm' did not converge ***
```

Model visualisation

Seeing the summary output is useful, but perhaps not the easiest way to interpret the fitted model, and certainly not one of the more effective ways of communicating the results to broader audiences. Visualisations help us here.

The predicted values of $\lambda_{BC\_CanadaGoose}$ are a function of all of the fitted covariates. Because the point process occurs over two dimensions, it can be difficult to understand how the individual coefficients in-and-of-themselves influence $\lambda_{BC\_CanadaGoose}$.

Model Selection

The quadratic term on gradient is significant, but the figure of $\rho(x)$ vs elevation, dist_water and HFI may be reasonably approximated by a straight line. To ensure we're not overfitting, we can use the `AIC()` function to calculate the AIC value of the fitted model, and compare it to a reduced model without a quadratic effect on gradient.

$$\lambda_{BC\_Parks\_reduced}(u) = e^{\beta_0 + \beta_1[elevation(u) + forestcover(u) + hfi(u) + dist\_water(u)]}$$

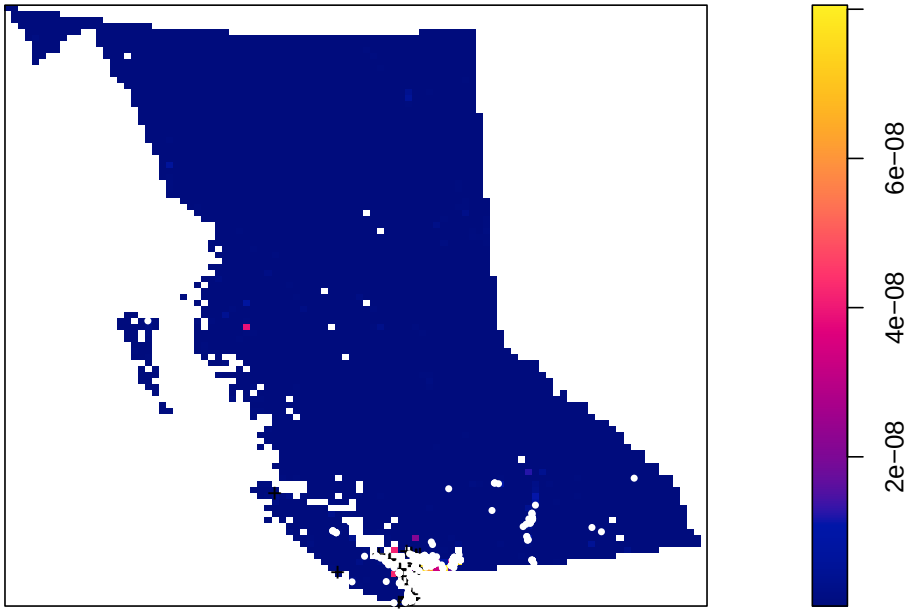

Figure 11: Fitted trend of the model

```
## [1] 27.69607
```

With a ΔAIC of ca. 27, the extra complexity is well supported by the data.

Model Validation

Model selection can tell us which models from a pool of candidates have the best support given our observations, but it doesn't tell us anything about how well our model does at predicting the occurrence of . When we fit a model to some data we are always assuming that the model has been correctly specified. In addition, when we use software to fit a model to some data it will always estimate some coefficients even if the model is a poor fit to the data. It is therefore critical to evaluate a model's behaviour to ensure that it is a reasonable fit to the data.

Quadrat counting

```
##
##  Chi-squared test of fitted Poisson model 'fit' using quadrat counts
##
## data:  data from fit
## X2 = 90.18, df = 4, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 9 tiles (irregular windows)
```

The small p value tells us that there's a significant deviation from our model's predictions. While this is useful for suggesting that our model has room for improvement, it provides us with no direction on how to do so (e.g., missing parameters, model mispecification (e.g., polynomial vs. linear), a lack of independence, non-stationarity, etc. . .).

PPP Residuals

From these figures we can see that the fitted model covariate terms are capturing the patterns in our data particularly well. Therefore, we can conclude our fitted model for the Canada Goose is working good.

Results

Discussion

References

1. GBIF.org (25 April 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.qs6zmf

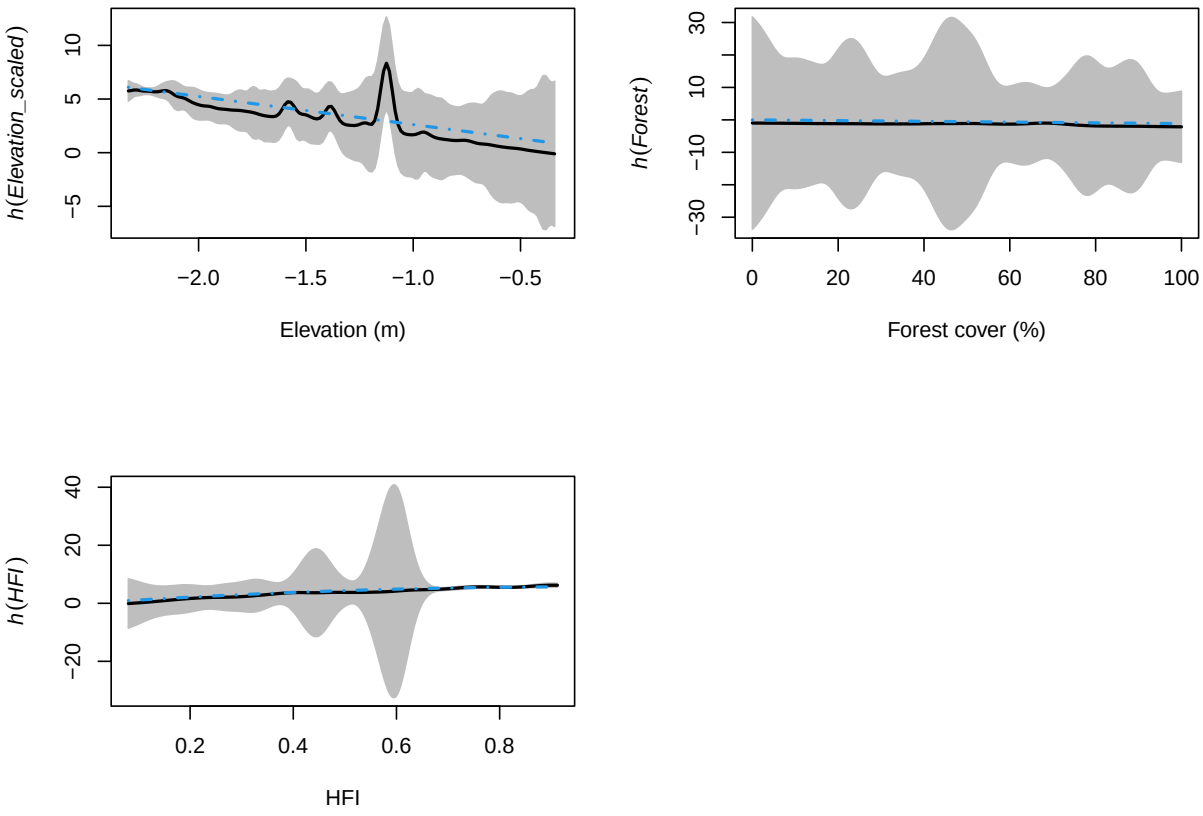

Figure 12: PPP Residuals plots