DATA589 Project

Varshita Kyal, Shveta Sharma, Ujjwal Upadhyay

30 April, 2023

Introduction

We have 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.

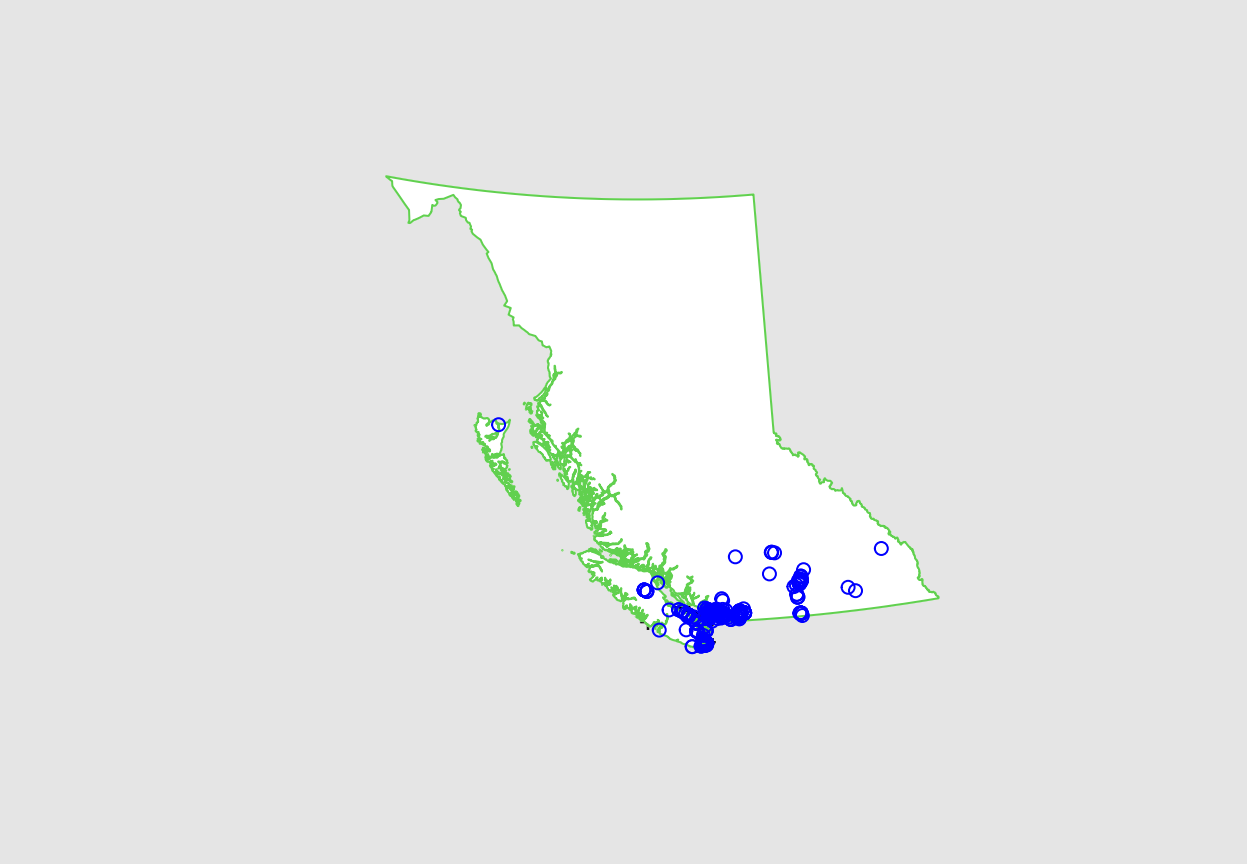

Figure 1: Occurences of Canada Goose in BC

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.

Next we looked into the relationship of the intensity with each covariate.

Results

First Moment Analysis

With some point data in hand, the first summary statistics we have calculated the average number of points per unit area (i.e., our 'expectation', or 'first moment') to check if the occurrence of Canada Goose in BC is homogeneous or not.

We noticed that the average intensity is 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 homogeneous or not by plotting it by dividing the regions in quadrants.

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.

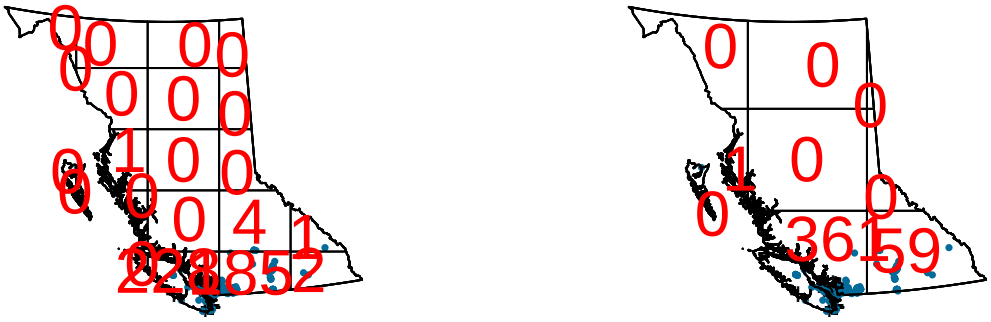

Figure 2: Quadrat counts of Canada Goose occurences, left 5x5, right 3x3

We have conducted a quadrat count of homogeneity with both 3 x 3 and 5 x 5 quadrats. These quadrats are shown in Figure 2, where we can visually tell that the intensity in each quadrats are not the same.

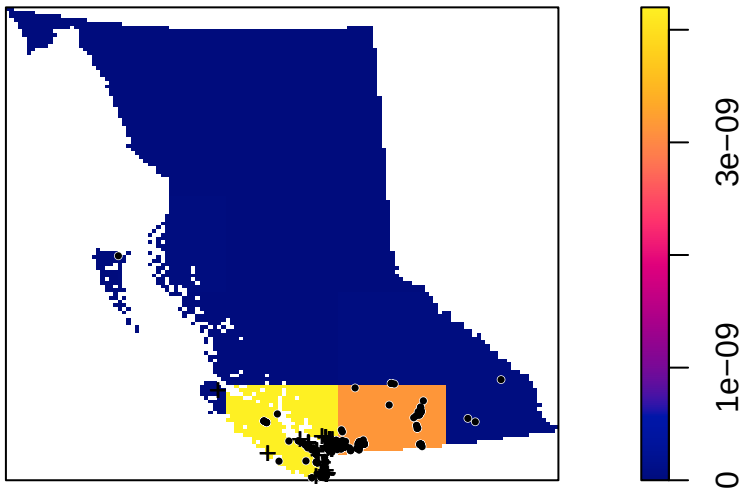

Figure 3: Quadrat counts with intensity of Canada Goose occurences (5x5)

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 = 3048, 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

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

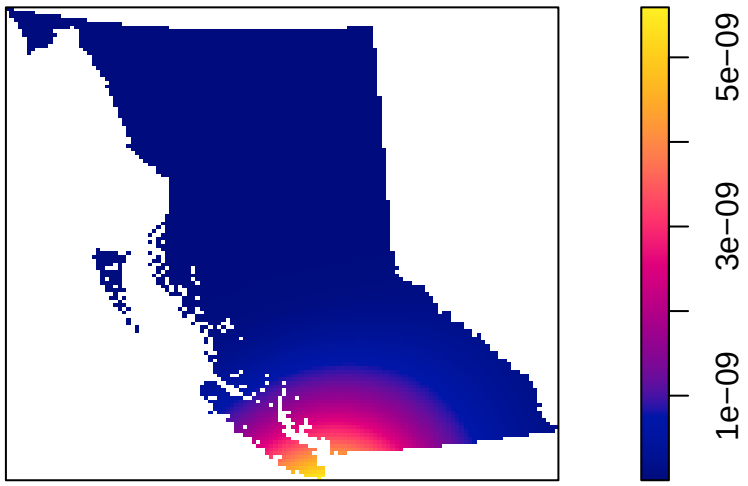

Figure 4: Kernel Estimation of Canada Goose occurences

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 shown in Figure 5.

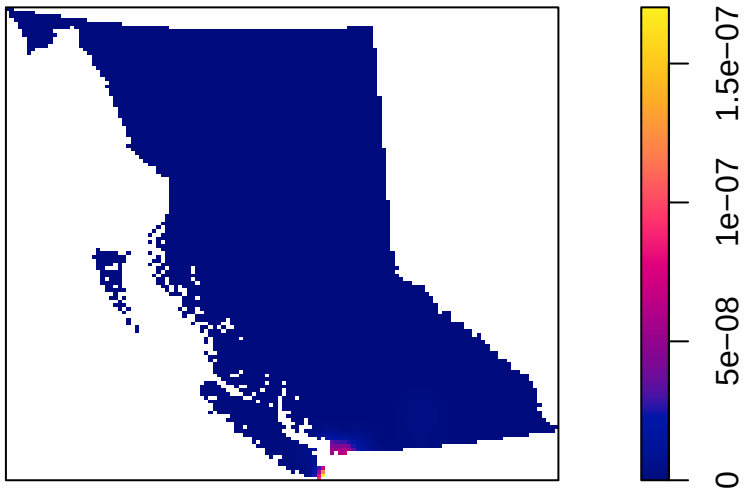

Figure 5: Adaptive kernel estimate of intensity

Hot spot analysis

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

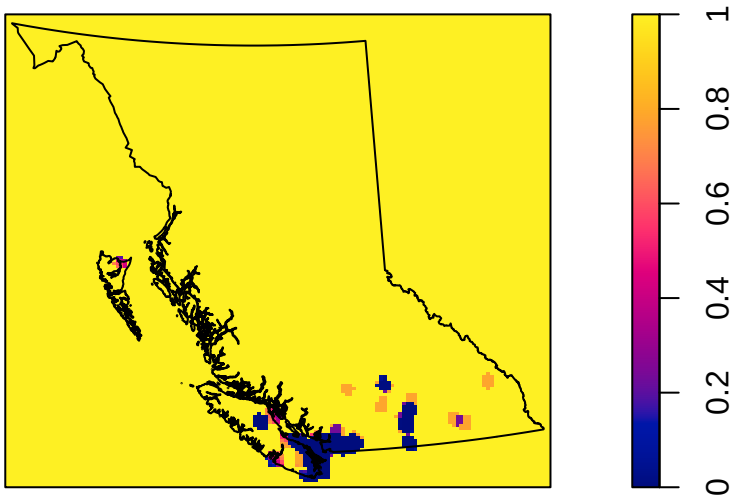

Figure 6: Local p-values

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.

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 homogeneous. Using the `Kinhom` function ensures that we are not assuming the intensity is homogeneous by weighting the data based on $\lambda(u)$

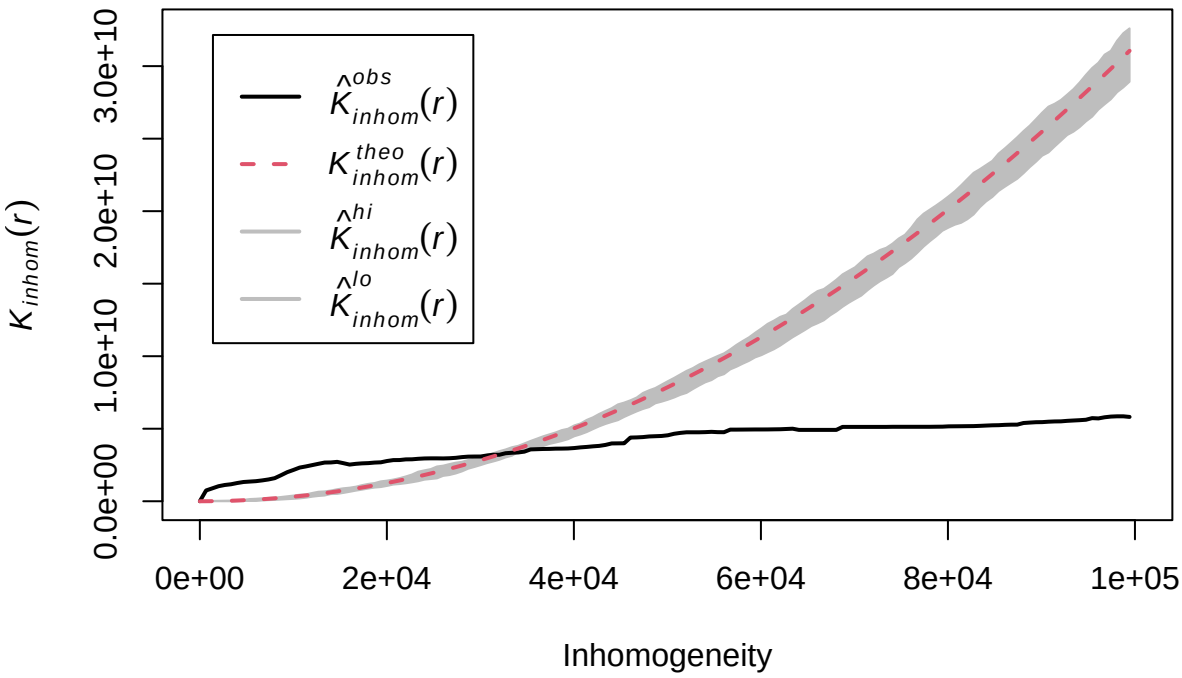

Figure 7: Ripley's K function with border correction assuming inhomogeneity

From figure 7, there is evidence of clustering, as the black line indicating the observed data is separated from the 95% confidence bands of the values expected with no clustering. This suggests that the relationship between points may be due to effects between points rather than relationship with covariates.

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.

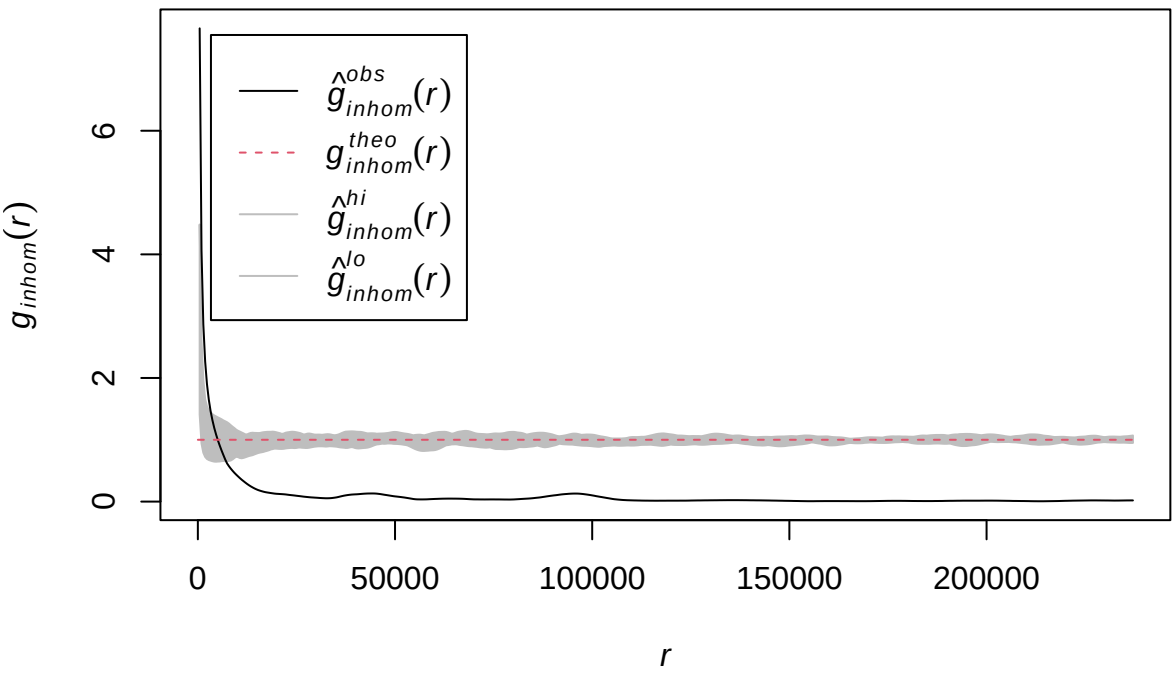

Figure 8: Pair correlation function assuming inhomogeneity

To get a sense of the distances for which clustering occurs, we used the pair correlation function. Figure 8 shows evidence for clustering at distances smaller than around 10 km but after that the observed values are showing avoidance behaviours as the species is just present in the southern part of the region.

Relationships with covariates

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

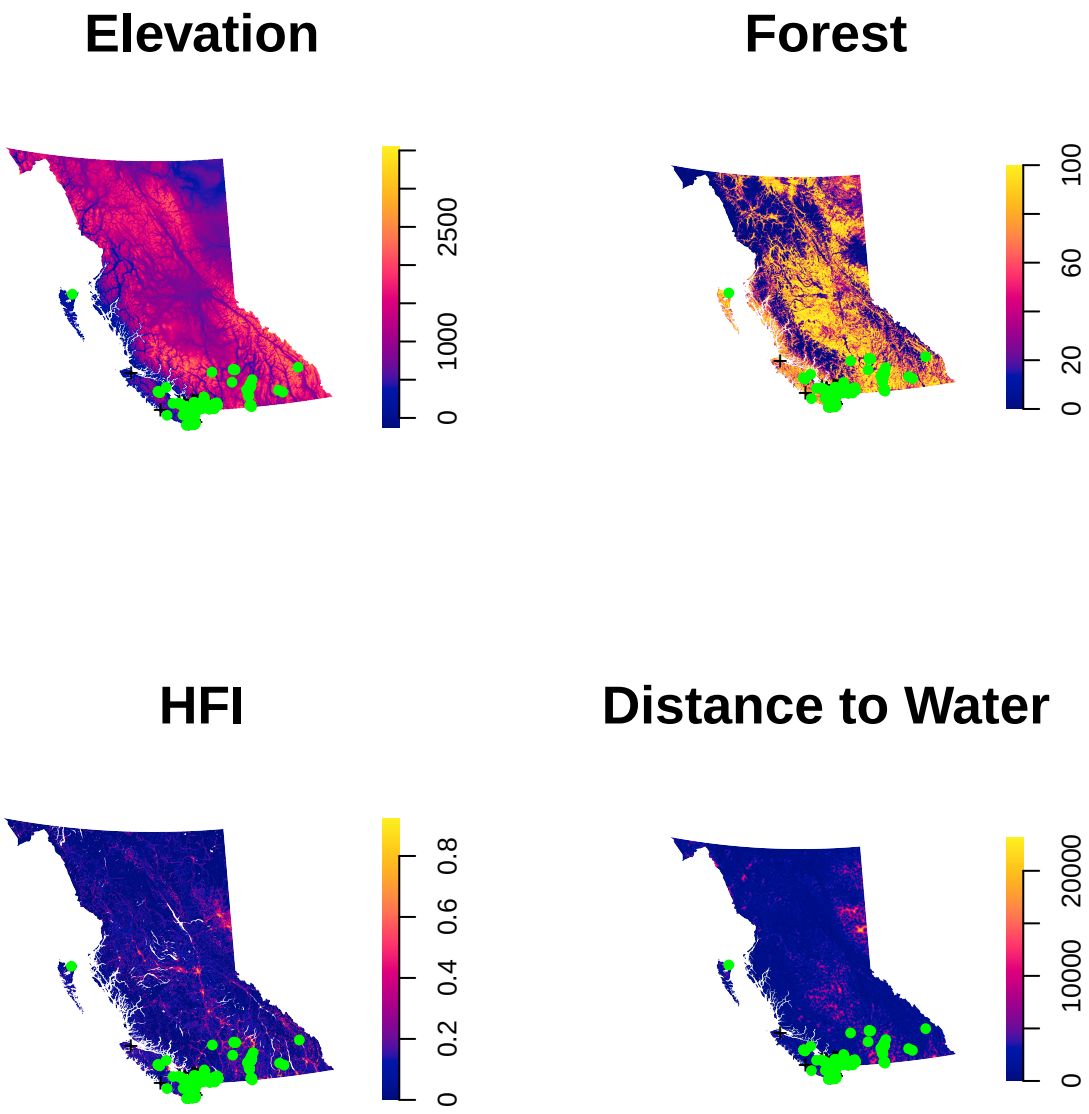

Figure 9: 4 covariates explaining the occurence of Canada Goose in BC

From figure 9, we can notice that Canada Goose are present on the low elevation area and also in highly densed forest area. Moreover, Canada Goose prefers low densed population areas and places where distance to water is in the lower range.

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.

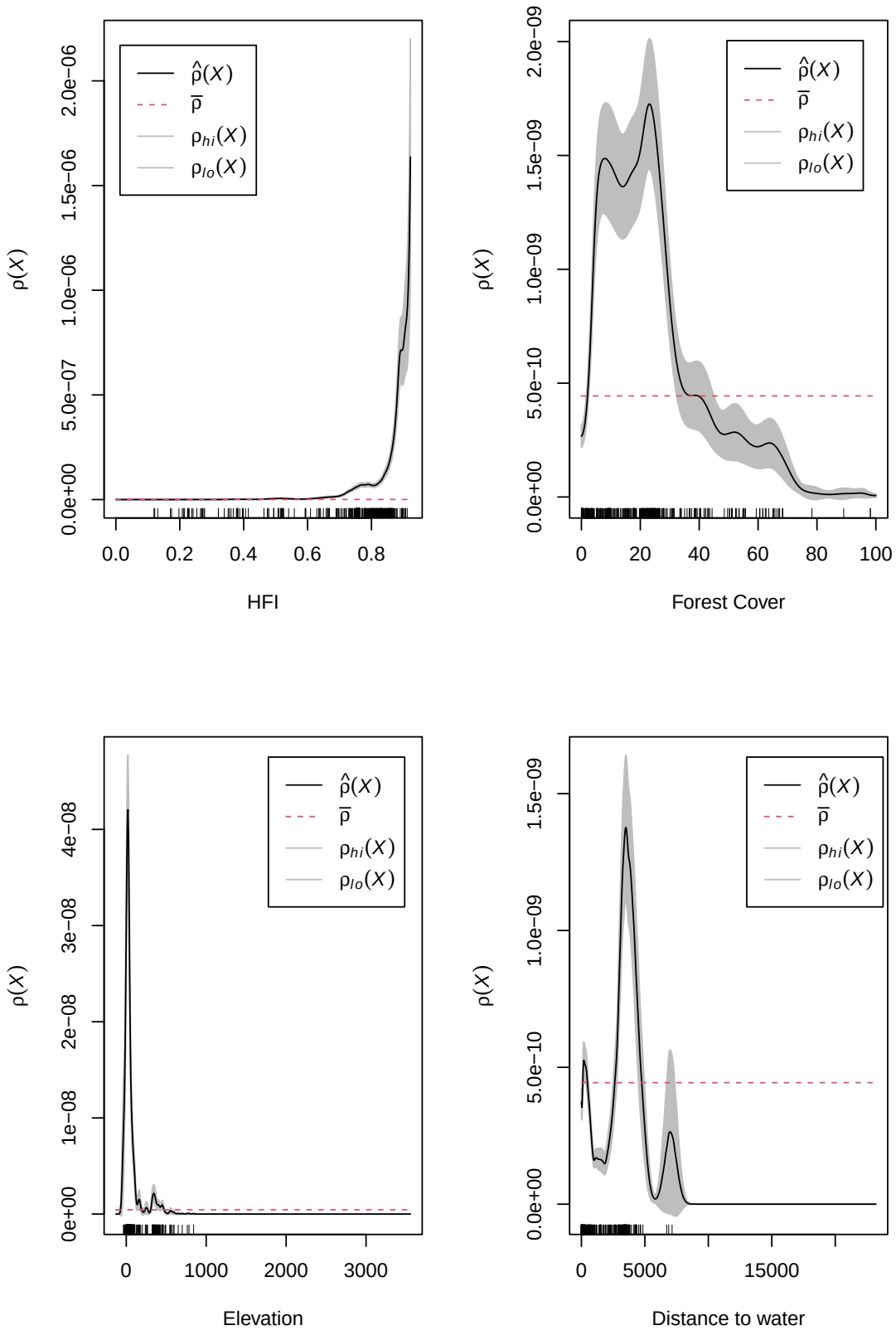

Figure 10: Comparison of Intensity with all the four covariates

From Figure 10, we can be misled by the first HFI figure, as it appears that there is no relationship until an HFI of around 0.8, at which point there seems to be an exponential relationship. This suggests that there is a non-linear relationship between HFI and Canada Goose occurrence, with the highest frequency of observations occurring at high HFIs. This finding is not surprising, as our dataset is crowdsourced and thus gooses are more likely to be spotted by humans in areas with higher HFIs.

As for forest cover, we can see that there is a non-linear relationship between forest cover and the number of observed Canada Goose. There is an increase in observations with intermediate forest cover, but then a decrease beyond that point. To avoid identifiability issues when modeling the data, it is necessary to examine whether there is any correlation between the covariates in the dataset (i.e. collinearity).

In terms of elevation, at first glance it seems like there is no relationship with Canada Goose occurrence. However, upon closer inspection, there appears to be a non-linear relationship, as the graph shows varying results at different elevations without any clear pattern.

Lastly, the figure depicting the relationship between Distance to Water and Canada Goose occurrence shows a non-linear relationship between the two variables.

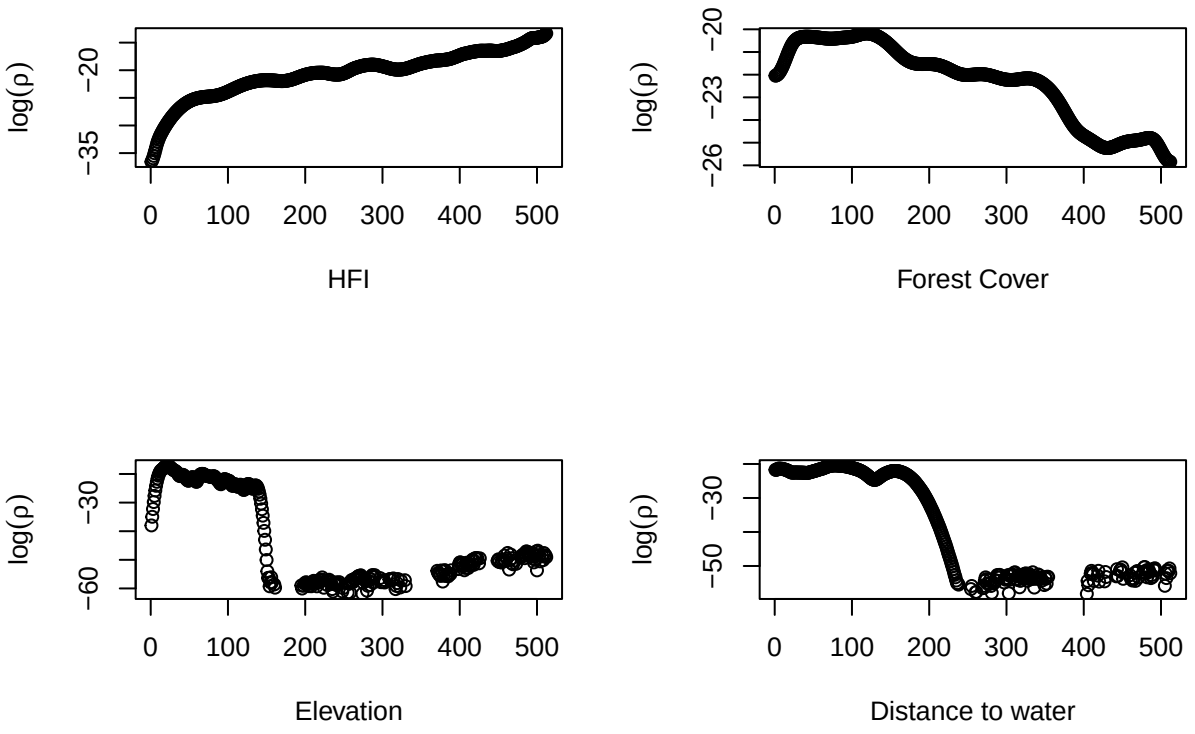

Figure 11: Comparison of logarithmic Intensity with all the four covariates

In Figure 11, we plot the $log(\rho)$, we get a line that could be reasonably interpreted as linear for HFI however the plots for other covariets deosnt seem linear.

Model Fitting

Collinearity

It is possible that HFI, elevation, 'forest cover' and 'distance to water' are correlated, which would cause identifiability issues when modelling the locations of Canada Goose.

```
##             ..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
```

Here, the correlation coefficients are relatively weak, so we can proceed without too much worry.

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 quadratic terms to fit our model as an initial guess of our fitted model.

Based on these initial analysis, a reasonable form for the model ($Model1$) 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)
##          -2.688252e+01          -1.780819e+00           2.199618e-01
##                 Forest            I(Forest^2)                    HFI
##          -1.460935e-02           5.787119e-05           1.277929e+01
##              I(HFI^2)      Dist_Water_scaled I(Dist_Water_scaled^2)
##          -6.835604e+00           5.355972e-02          -8.334819e-02
##
##                             Estimate         S.E.       CI95.lo       CI95.hi
## (Intercept)            -2.688252e+01 5.280534e-01 -2.791749e+01 -2.584756e+01
## Elevation_scaled       -1.780819e+00 5.462052e-01 -2.851361e+00 -7.102765e-01
## I(Elevation_scaled^2)   2.199618e-01 1.793507e-01 -1.315592e-01  5.714828e-01
## Forest                 -1.460935e-02 6.638534e-03 -2.762064e-02 -1.598060e-03
## I(Forest^2)             5.787119e-05 9.520483e-05 -1.287268e-04  2.444692e-04
## HFI                     1.277929e+01 1.436377e+00  9.964045e+00  1.559454e+01
## I(HFI^2)               -6.835604e+00 1.293085e+00 -9.370004e+00 -4.301204e+00
## Dist_Water_scaled       5.355972e-02 6.596289e-02 -7.572517e-02  1.828446e-01
## I(Dist_Water_scaled^2) -8.334819e-02 3.840362e-02 -1.586179e-01 -8.078478e-03
##                         Ztest        Zval
## (Intercept)               *** -50.9087249
## Elevation_scaled           **  -3.2603480
## I(Elevation_scaled^2)           1.2264337
## Forest                     *   -2.2006889
## I(Forest^2)                     0.6078598
```

```
## HFI                      ***   8.8968931
## I(HFI^2)                 ***  -5.2862765
## Dist_Water_scaled              0.8119674
## I(Dist_Water_scaled^2)     *  -2.1703212
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.6% (14 out of 2320) 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 ($Model2$):

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

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Elevation_scaled + Forest + HFI + I(Dist_Water_scaled^2) +
## I(HFI^2)
##
## Fitted trend coefficients:
##           (Intercept)       Elevation_scaled                Forest
##          -27.39640990            -2.51469385           -0.01128053
##                   HFI I(Dist_Water_scaled^2)              I(HFI^2)
##           12.42495228            -0.06608439           -6.54916803
##
##                            Estimate        S.E.      CI95.lo       CI95.hi
## (Intercept)            -27.39640990 0.430020264 -28.23923413 -26.553585672
## Elevation_scaled        -2.51469385 0.156901246  -2.82221464  -2.207173057
## Forest                  -0.01128053 0.003231708  -0.01761456  -0.004946496
## HFI                     12.42495228 1.347737875   9.78343458  15.066469977
## I(Dist_Water_scaled^2)  -0.06608439 0.032877086  -0.13052230  -0.001646487
## I(HFI^2)                -6.54916803 1.231039166  -8.96196046  -4.136375605
##                         Ztest       Zval
## (Intercept)               *** -63.709579
## Elevation_scaled          *** -16.027239
## Forest                    ***  -3.490578
## HFI                       ***   9.219116
## I(Dist_Water_scaled^2)      *  -2.010044
## I(HFI^2)                  ***  -5.320032
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.6% (14 out of 2320) of
## the quadrature points
##
## *** Fitting algorithm for 'glm' did not converge ***
```

In this fitted model, all the predictor variables are statistically significant and therefore we can proceed with this model for now.

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.

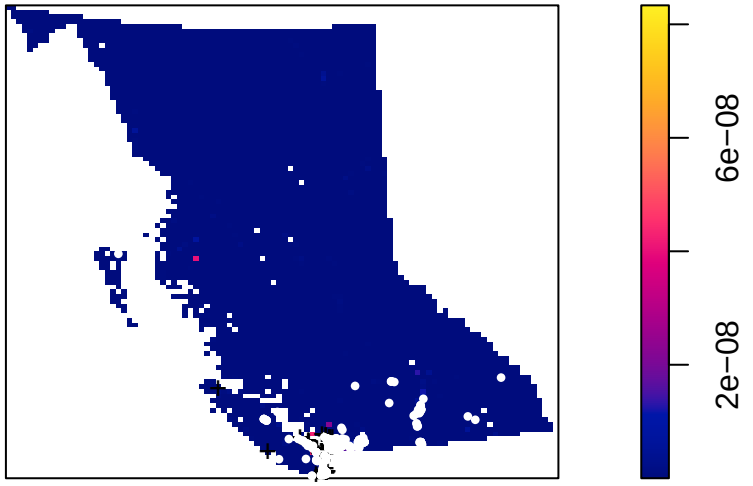

Figure 12: Fitted trend of the model

The predicted values of $\lambda_{BC\_CanadaGoose}$ are a function of all of the fitted covariates. Because the point process occurs over four 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 'distance to water' and 'HFI'.

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

```
## [1] 32.49056
```

With a ΔAIC of ca. 32, 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 = 83.119, df = 3, 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

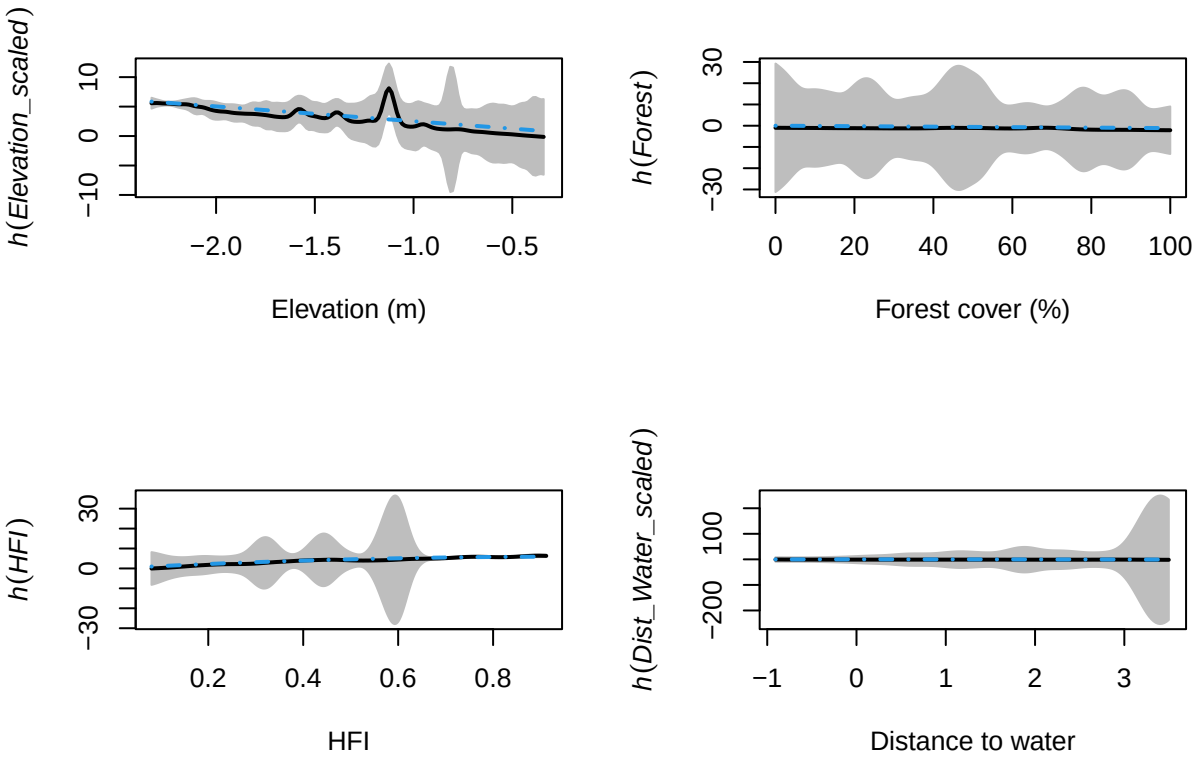

Figure 13: PPP Residuals plots

From Figure 13, 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.

Discussion

Upon examining the occurrence data of Canada goose on a map of British Columbia, it was observed that the species is mostly present in the southern part of the province. After conducting a first moment analysis, it became evident that the data intensity is non-uniform, and a few regions of the province stand out as hotspots. Subsequently, a second moment analysis was conducted using Ripley's K-function, which revealed data clustering, and the pair correlation analysis showed that there was no significant clustering beyond 10km. Thus, it can be concluded that there is little to no clustering in the occurrence of Canada goose.

Furthermore, analysis of the covariates - HFI, forest cover, and elevation - in relation to the Canada goose data showed that these variables have a non-linear relationship. Distance to water, the fourth covariate, was also examined, but initial assessments did not show a promising linear relationship, so a quadratic form of the covariate was used in the further analysis. Additionally, there was no significant correlation observed between the variables, which allowed them to be combined in further modeling.

The first combined model (Model 1) was fitted with linear and quadratic terms for all four covariates. However, just 'distance to water' and 'HFI' was found to be an significant predictor in quadratic terms and other 2 coviartes as linear and the next model (Model 2) was fitted with elevation, forest cover and HFI using linear and distance to water and HFI as quadratic terms. The AIC scores for the three models are presented in the table below.

	Model 1	Model 2
AIC	15423.13	15455.44

As demonstrated earlier, Model 1 has a lower AIC score compared to the other model, and after conducting a quadrat test, it was concluded that none of the models were superior. Therefore, we opted for a parsimonious model and selected Model 1. The residual, partial residual, and covariate effect plots confirmed the adequacy of the model fit. However, the plots also revealed that there is room for improvement, particularly in areas of high and low elevation.

Our research objectives were two-fold. Firstly, we aimed to determine whether Canada Goose occurrence is present in close proximity to human habitation, and the model indicated that the answer is "yes." Secondly, we sought to establish whether Canada Goose avoid forest cover, but we were unable to conclude that there is a significant relationship between the species and this covariate. Nonetheless, we found that there is a significant association with elevation.

During the analysis and modeling process, we encountered several intriguing challenges and insights that we would like to share to assist with future research. Firstly, the Rho plots and standard residuals plot for this data could not be generated due to NAs in the data. Secondly, if higher-order elevation variables are included in Model 2, the model does not converge.

Finally, a model was fitted which concludes that occurences of Canada Goose in BC depends on the following equation which can be improved further in future work.

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

References

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

2. Research topic: https://wildlife-species.canada.ca/bird-status/oiseau-bird-eng.aspx?sY=2019&sL=e&sM=a&sB=CANG