Honey Bees (Apis mellifera Linnaeus) Spatial Analysis

Alyssa Foote, Noman Mohammad, and Vimaljeet Singh

30 April, 2023

Figure 1: A Western Honey Bee collecting nectar

1 Introduction

Spatial statistics have increasingly become an essential tool in understanding species distribution and habitat preferences, as well as in informing conservation and management decisions. In this report, we will analyze the occurrence data of the western honey bee (Apis mellifera Linnaeus) in the province of British Columbia (BC), Canada.

The western honey bee, a critical pollinator species, plays a vital role in agriculture and ecosystem services. Understanding the distribution patterns and potential drivers of honey bee occurrence can provide valuable insights for the conservation of this species and the ecosystems they support.

The honeybee is a vital pollinator that plays a significant role in both agricultural production and the maintenance of biodiversity in natural ecosystems. These insects contribute to the pollination of a wide range of crops, which directly affects human food supply, as well as sustaining the health and productivity of various plant species within their habitats (Klein et al., 2007).

British Columbia (BC), a province within Canada, encompasses diverse ecosystems that support a rich array of plant and animal species. Honeybees are vital contributors to the pollination processes within these ecosystems and are essential for the productivity of the region's agricultural sector. Understanding the spatial distribution of honeybee populations in British Columbia is crucial for developing effective conservation and management strategies to protect these invaluable pollinators.

In this study, we aim to explore the spatial distribution of honeybee populations (Apis mellifera) in British Columbia using occurrence data obtained from the Global Biodiversity Information Facility (GBIF) database. Our analysis will focus on describing the spatial distribution of Apis mellifera in the region, identifying any clustering or dispersion patterns in the occurrence data, and assessing the relationships between honeybee occurrence and various covariates, such as elevation, forest cover, Human Footprint Index (HFI), and distance to water sources. To achieve our objectives, we will construct a Poisson process model to analyze the spatial distribution of Apis mellifera occurrences in British Columbia. We will explore various statistical techniques, including creating rho plots, and examining the first and second moment statistics to gain insights into the clustering or dispersion patterns in the data. Furthermore, we will evaluate the relationship between honeybee occurrence and covariates such as elevation, forest cover, HFI, and distance to water sources, using spatial regression analysis. This will enable us to identify significant predictors that influence the spatial distribution of honeybee populations in the region. Additionally, we will explore collinearity between covariates, and use model validation techniques such as chi-squared tests and AIC scores. Overall, our approach will involve a combination of statistical modeling and exploratory data analysis techniques to gain a deeper understanding of the factors that influence honeybee occurrence in British Columbia.

1.1 Data collection and Description

This dataset contains occurrence records of the honey bee species Apis mellifera Linnaeus in the province of British Columbia, Canada. The data was collected using the 'rgbif' package in R, which allows for the retrieval of biodiversity data from the Global Biodiversity Information Facility (GBIF) database. It is important to note that the dataset may be subject to potential biases due to the reliance on self-reporting by the individuals or institutions responsible for collecting the data. This means that the accuracy and completeness of the information provided may vary depending on the quality of the reporting.

The first step involved querying the number of records available for Apis mellifera Linnaeus in British Columbia. The complete dataset was then downloaded by specifying the scientific name, coordinates, country, and stateProvince parameters. After obtaining the data, it was saved as a CSV file.

The dataset contains various variables, including decimalLongitude, decimalLatitude, country, stateProvince, occurrenceStatus, coordinateUncertaintyInMeters, taxonID, catalogNumber, institutionCode, eventTime, verbatimEventDate, collectionCode, gbifID, verbatimLocality, class, isInCluster, year, month, day, eventDate, modified, and lastInterpreted. These variables provide information on the location of each occurrence, the institution responsible for collecting the data, and other details about the occurrence event.

To clean the dataset, records indicating absence or zero-abundance were removed. Furthermore, the dataset's coordinates were adjusted to match the covariate data's format. A SpatialPointsDataFrame object was created using the cleaned dataset, and the WGS84 CRS (EPSG:4326) was assigned to it. The coordinates were then transformed to match the CRS of the BC_win object, which has the Albers Equal Area Conic projection, and the transformed coordinates were added to the original cleaned dataframe.

Finally, the cleaned and transformed data was saved as a new CSV file.

In addition to the occurrence data, covariate data for the province of British Columbia is available. The covariate dataset includes variables such as Elevation, Forest, HFI (Human Footprint Index), and Dist_water (Distance to water). These variables provide information about the environmental conditions of each occurrence site, which can be useful for understanding the factors influencing the distribution and abundance of Apis mellifera Linnaeus in the province.

1.2 Exploratory Analysis

The spatstats package has multiple functions to help us do exploratory data analysis. We will start with finding the number of points using the function npoints(). We then find the intensity both homogeneous and inhomogeneous using the functions intensity() and quadratcount() respectively. We then do a test of homogeneity to check if the intensity is homogeneous or not using quadrat.test(). After determining that the intensity is homogeneous, we estimate it using kernel density estimation as well with the function density(). We then move on to find the areas of high intensity with hotspot analysis where we do a scan test and plot the pvalues to see where the areas of high intensity are. Next we check the relation of intensity with covariate using rhohat() function and check for any evidence of clustering using the inhomogeneous K-function and confirm the findings of K-function using pair correlation function pcf_inhom(). We use the plot from rhohat() which shows the relation between intensity and covariates to specify the initial model and perform model selection and validation.

1.3 Model Development

The spatstats package, additionally has numerous functions that allow for easy predictive modeling on spatial data. For this study, before any models were built scaling of the elevation and distance to water (dist_water) covariates was required. This is because the scales for these variables have a wider range than the scales for the human footprint index, 0 to 10, and the forest coverage, 0 to 100 percent. Furthermore, as with all modeling building, it was also important to check for collinearity among the covariates. This was done using the package's cor.im function. Then numerous models, in the form of linear, quadratic, and generalize additive (gam), were built using the ppm function. Additionally, a b-spline function was used from the splines package for the gam models. Each of these models were evaluated by reviewing the partial residuals, AIC scores, likelihood ratio test, and diagnostics via the parres, AIC, anova, and diagnose ppm functions. Finally, all the models were tested on their ability to predict the intensity of honey bees in British Columbia using a fitted trend plot and the quadrat.test function.

2 Results

2.1 Exploratory Analysis

2.1.1 Number of data points

```
## [1] 2575
```

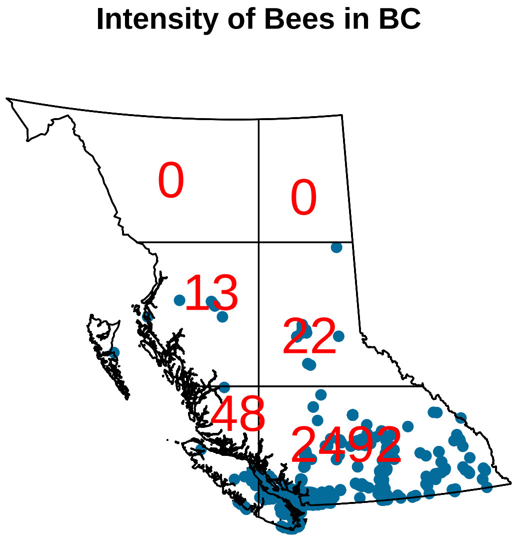
2.1.2 Intensity

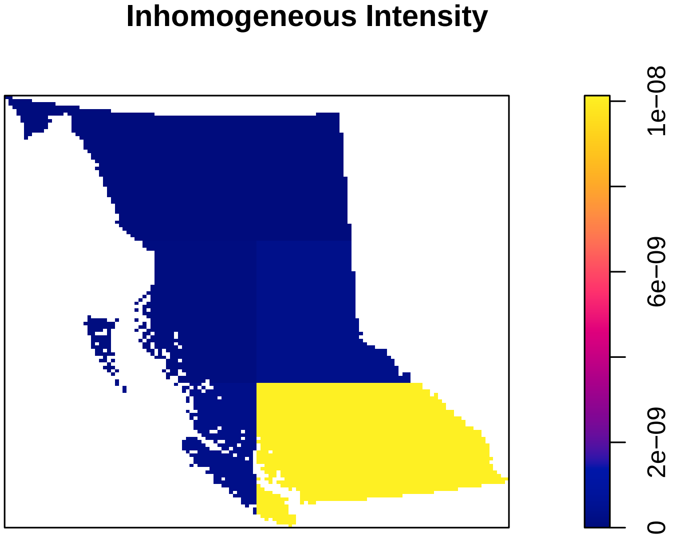
Assuming intensity is homogeneous, we use the in-built function to find the value of intensity.

```
## [1] 2.715493e-09
```

2.1.3 Quadrat counting

In most real scenarios, λ is likely to be spatially varying. When λ is spatially varying, the intensity at any location u is $\lambda(u)$ and it can be estimated non-parametrically by dividing the window into sub-regions called quadrats. Let us do quadrat counting to find the inhomogeneous intensity:





Clearly, the assumption of homogeneity is inappropriate for this dataset as the bees tend to be clustered in certain areas of the study site, whereas others have no bees at all. Quadrat counting suggests a spatially varying, inhomogeneous $\lambda(u)$, but point processes are stochastic and some variation is expected by chance alone. Our eyes are also a poor judge of homogeneity, so we should ideally test for spatial homogeneity objectively. Under a null hypothesis that the intensity is homogeneous, we carry out the test for significant deviations from complete spatial randomness which is the quadrat test.

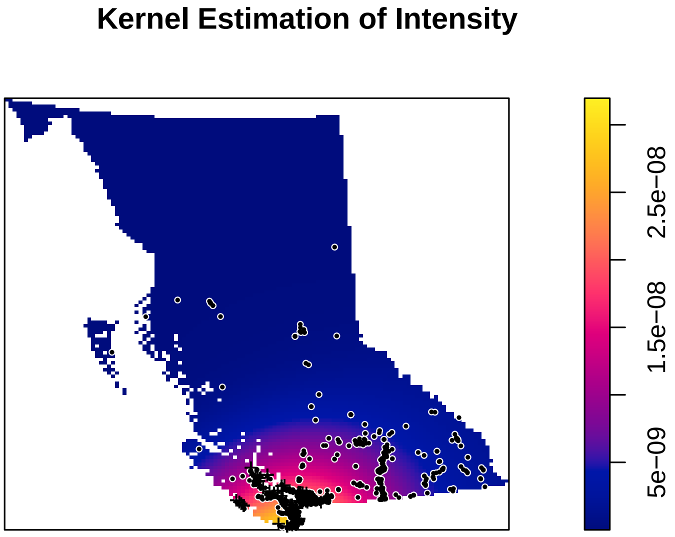
2.1.4 Test of homogeneity

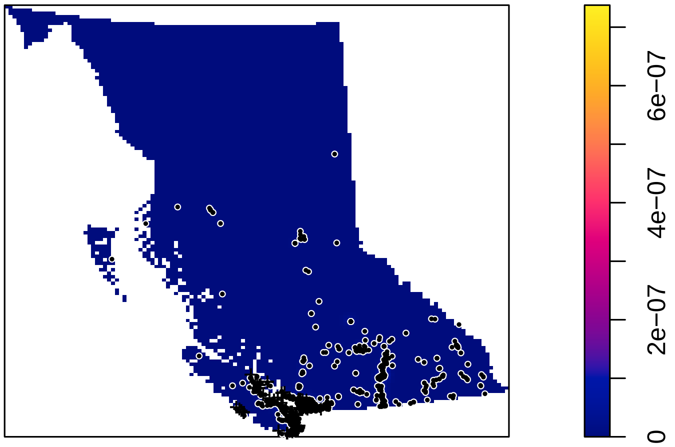
```
##
##  Chi-squared test of CSR using quadrat counts
##
## data:
## X2 = 6736.9, df = 5, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 6 tiles (irregular windows)
```

The small p-value (< 2.2e-16) indicates that the observed pattern of points is significantly non-random with very strong evidence against the null hypothesis of complete spatial randomness (CSR) at any significance level. Hence, there is a significant deviation from homogeneity. The p-value doesn't provide any information on the cause of inhomogeneity, however, significant deviations can be due to the processes truly being inhomogeneous, but also due to a lack of independence between points.

2.1.5 Kernel Estimation

A spatially varying, $\lambda(u)$ can also be estimated non-parametrically by kernel estimation. Kernels estimate $\lambda(u)$ by placing kernels' on each data point (often bi-variate Gaussian) and optimizing the bandwidth' (i.e., the standard deviation of the kernel).



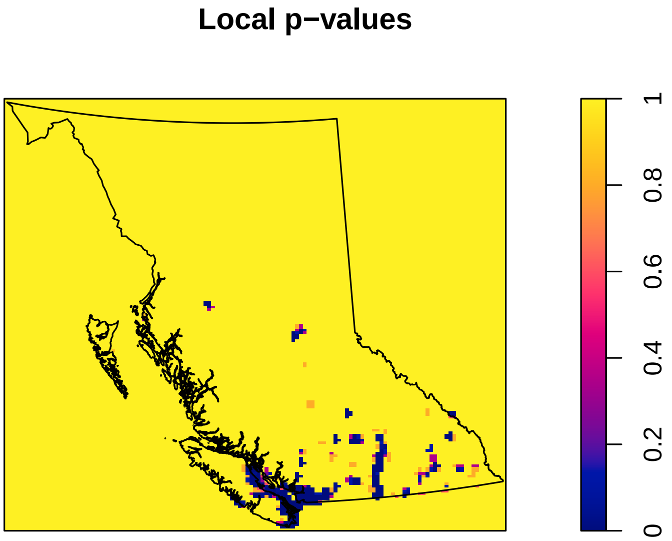


Using the plots above, we can see that southern region of BC has high intensity and there is mostly no change in the northern and central BC area. This is consistent with the findings of the quadrat count graph as well.

2.1.6 Hotspot analysis

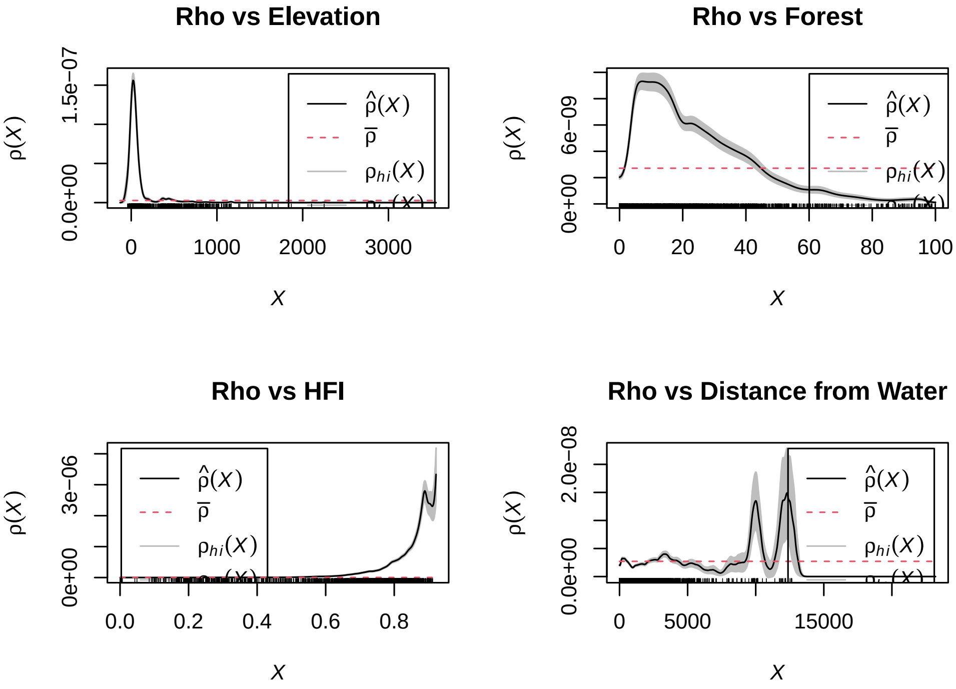
Even though we can see from the plot above that the intensity is inhomogeneous and higher in the south, we would still like to identify areas where the intensity is elevated. This is what we will do below using 'scan test'.

As the intensity is inhomogeneous, at each location u, we can draw a circle of radius 'r' (we call this my_R in the code below) and then count the number of points in and out of the circle. This 'r' has been found through kernel bandwidth optimization. Also, under an assumption that the process is Poisson distributed, we can calculate a likelihood ratio test statistic for the number of points inside vs. outside the circle. The null distribution is χ^2 with 1 degree of freedom allowing us to calculate p-values at each location u.



From the p-value plot, we can see that the hot spot areas are more densely situated towards the south of BC.

2.1.7 Check relation with covariate (Elevation)

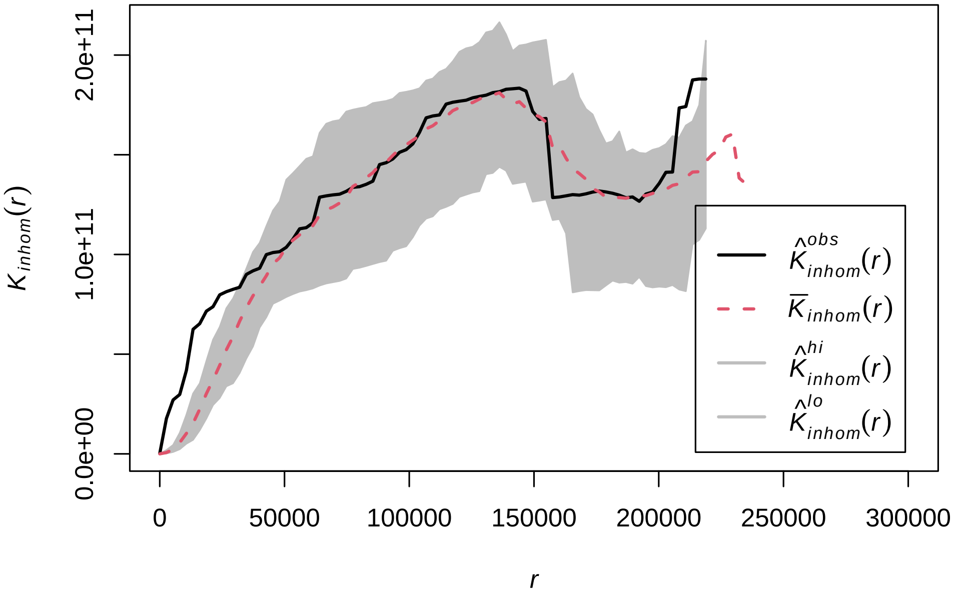


The intensity seems to have a non-linear relation with elevation, forest, human footprint index and distance from water. We can use these relations to be a good starting point to specify an initial model and start selecting and validating more models based on this initial model and how it performs.

2.1.8 Using Ripley's K-function to test for a significant ($\alpha = 0.05$) correlation between locations

As the intensity is inhomogeneous, we will be using 'Kinhom' function to plot the K-function.

```
## 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.
```

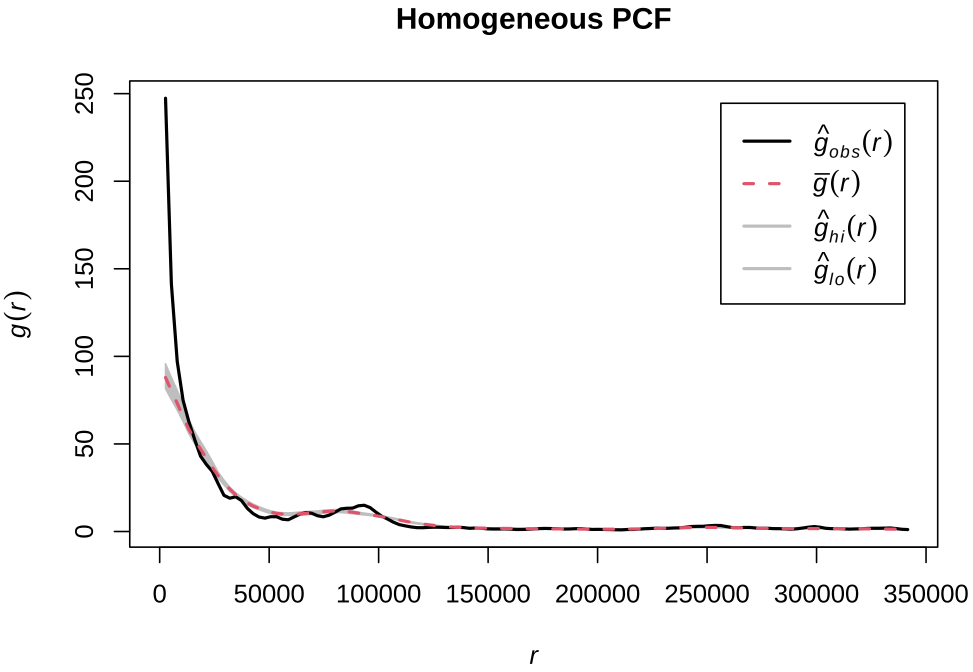


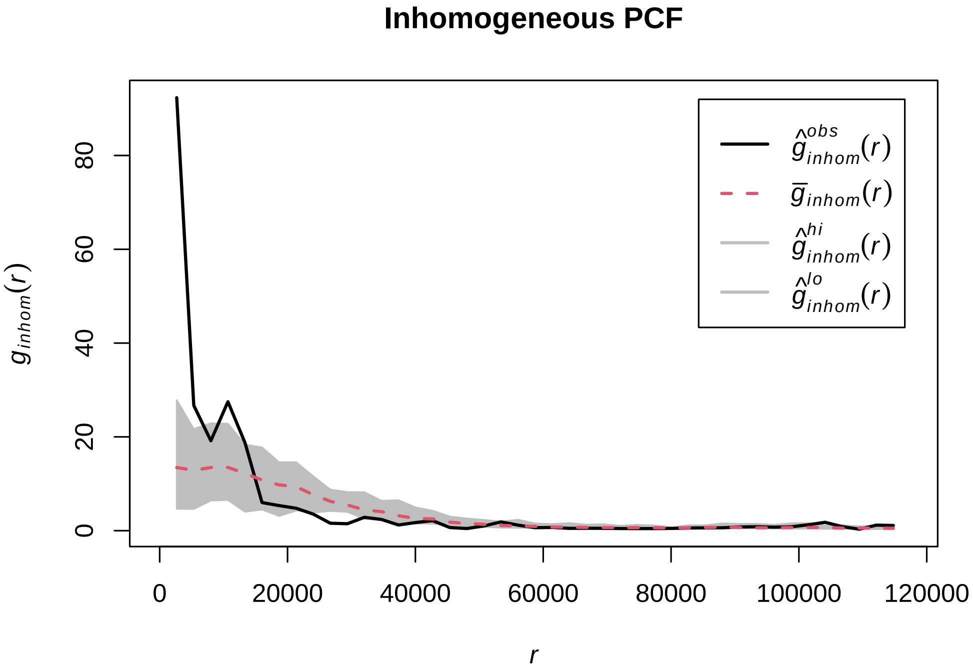
The shading is centered around the estimated K-function and the width of the shaded region reflects the precision of the estimate. The confidence band is around the observed values mostly for the entirety of the plot and the expected values denoted by the red line are inside the band too. Hence, there is no significant evidence of clustering among bees.

2.1.9 Estimating homogeneous and inhomogeneous pair correlation functions, using simulation envelopes.

```
## 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.
```

```
## 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.
```





g(r) is the probability of observing a pair of points of the process separated by a distance r, divided by the corresponding probability for a Poisson process. Pair correlations function contains contributions only from inter-point distances equal to r (and not <= r). We will use the inhomogeneous pair correlation function as our point process is inhomogeneous. As g(r)> 1 indicates that this inter-point distance is more frequent than expected for a completely random pattern, we can say that there is significant evidence of clustering at very low values of r for our point process.

2.2 Model Selection

As describe in the Model Development section, above, before developing any predictive models, it is important to review the collinearity among the covariates. Table 1 and Figure 2, show that although there is some correlation between all covariates, the strongest is between elevation and HFI with a negative correlation of -0.266. Since this value is less than the typical threshold of ±0.4, the model building proceeded without taking these correlations into consideration.

Table 1: Covariate Correlation Matrix

	elev	forest	HFI	dist_water
elev	1.0000000	-0.2620472	-0.2662563	-0.0342639
forest	-0.2620472	1.0000000	0.0661554	0.0482544
HFI	-0.2662563	0.0661554	1.0000000	0.1322941
dist_water	-0.0342639	0.0482544	0.1322941	1.0000000

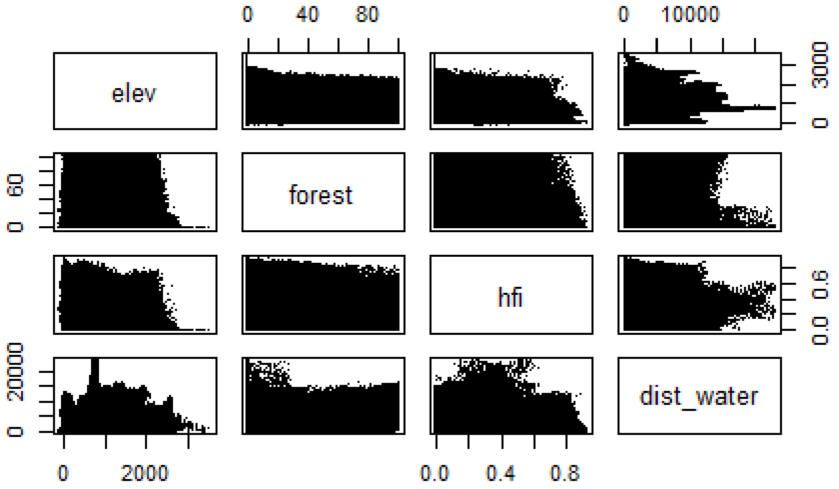
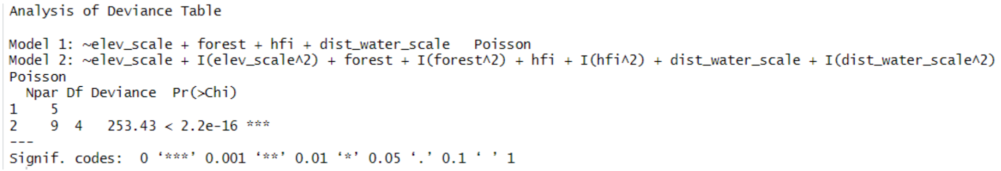


Figure 2: Covariate Correlation Matrix

Based on the results describe in the Exploratory section, the first model taken into consideration was a full quadratic model. This was because of the evidence that the relationship between honey bee intensity and the covariates was at least not linear. Further proven by comparing the AIC values - 8.7043583×10^4 for the linear model versus 8.6798149×10^4 for the full quadratic model - and running a likelihood ratio test that rejected the linear model in favour of this full quadratic model, seen in Figure 3. However, based on the predicted intensity plot in Figure 4, it is clear that this model is not very good. There are many areas of the province that are predicted to have high intensity even though there are few bees located there, which are represented by the black dots. The poor model fit is further exemplified in the quadrat test for deviance as seen in Figure 5. The tiny p-value of 0.001 means that the suitability of this model for predicting the intensity of honey bees should be rejected. Additionally the model summary, in Figure 6, indicates that only the quadratic term of the dist_water covariate is significant.



```
Analysis of Deviance Table

Model 1: ~elev_scale + forest + hfi + dist_water_scale   Poisson
Model 2: ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) + dist_water_scale + I(dist_water_scale^2)
Poisson
  Npar Df Deviance  Pr(>Chi)    
1    5                          
2    9  4   253.43 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Likelihood Ratio Test - Full Quadratic Model

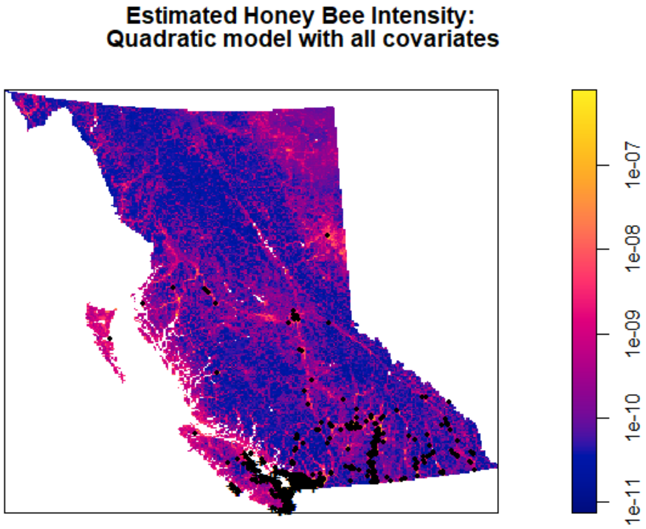
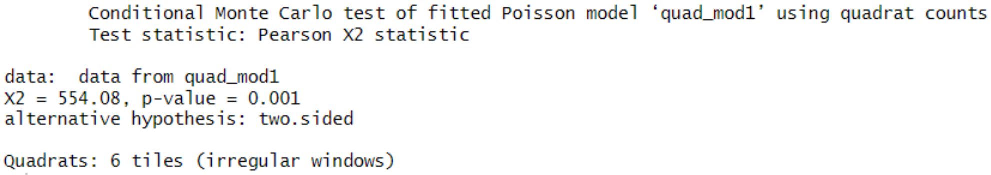
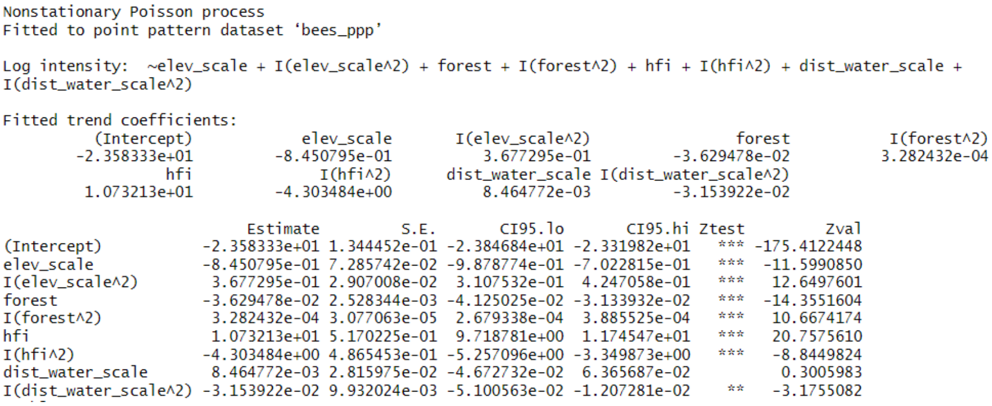

Estimated Honey Bee Intensity:
Quadratic model with all covariates
1e-07
1e-08
1e-09
1e-10
1e-11

Figure 4: Predicted Honey Bee Intensity - Full Quadratic Model



```
        Conditional Monte Carlo test of fitted Poisson model ‘quad_mod1’ using quadrat counts
        Test statistic: Pearson X2 statistic

data:  data from quad_mod1
X2 = 554.08, p-value = 0.001
alternative hypothesis: two.sided

Quadrats: 6 tiles (irregular windows)
```

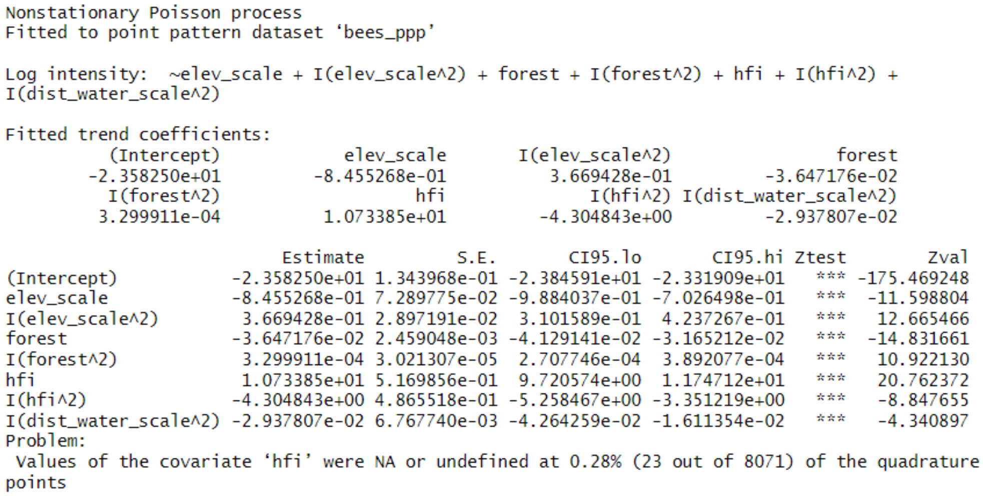
Figure 5: Predicted Deviance - Full Quadratic Model



```
Nonstationary Poisson process
Fitted to point pattern dataset ‘bees_ppp’

Log intensity:  ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) + dist_water_scale +
I(dist_water_scale^2)

Fitted trend coefficients:
          (Intercept)            elev_scale       I(elev_scale^2)                forest           I(forest^2)
        -2.358333e+01         -8.450795e-01          3.677295e-01         -3.629478e-02          3.282432e-04
                  hfi              I(hfi^2)      dist_water_scale I(dist_water_scale^2)
         1.073213e+01         -4.303484e+00          8.464772e-03         -3.153922e-02

                           Estimate         S.E.      CI95.lo      CI95.hi Ztest         Zval
(Intercept)           -2.358333e+01 1.344452e-01 -2.384684e+01 -2.331982e+01   *** -175.4122448
elev_scale            -8.450795e-01 7.285742e-02 -9.878774e-01 -7.022815e-01   ***  -11.5990850
I(elev_scale^2)        3.677295e-01 2.907008e-02  3.107532e-01  4.247058e-01   ***   12.6497601
forest                -3.629478e-02 2.528344e-03 -4.125025e-02 -3.133932e-02   ***  -14.3551604
I(forest^2)            3.282432e-04 3.077063e-05  2.679338e-04  3.885525e-04   ***   10.6674174
hfi                    1.073213e+01 5.170225e-01  9.718781e+00  1.174547e+01   ***   20.7575610
I(hfi^2)              -4.303484e+00 4.865453e-01 -5.257096e+00 -3.349873e+00   ***   -8.8449824
dist_water_scale       8.464772e-03 2.815975e-02 -4.672732e-02  6.365687e-02          0.3005983
I(dist_water_scale^2) -3.153922e-02 9.932024e-03 -5.100563e-02 -1.207281e-02    **   -3.1755082
```

Figure 6: Full Quadratic Model Summary

Therefore, the next model examined was another quadratic model but without the insignificant ‘dist_water’ term. While all the predictors were seen as significant in this model, as seen in Figure 7, the predicted intensity plot did not change, Figure 8. This model was also only a small improvement in AIC from the full quadratic model, 8.6796239×10^4 from 8.6798149×10^4. However, the likelihood ratio test, as seen in Figure 9, does not support the reduced complexity since the p-value of $0.7636 > 0.05$. Meaning that the full quadratic model cannot be rejected in favour of the quadratic model without the ‘dist_water term’. Furthermore, both models had nearly identical diagnostic plots, seen in Figure 10, that indicate both models do not predicate the y coordinates adequately. That is for all parts of the province, the sum of residuals on the y coordinate fall outside the acceptable confidence band indicated by the dotted line. For the x coordinate, only the western and eastern edges of the province had acceptable residuals. No further diagnostics for either of these models were reviewed as a result.



```
Nonstationary Poisson process
Fitted to point pattern dataset ‘bees_ppp’

Log intensity:  ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) +
I(dist_water_scale^2)

Fitted trend coefficients:
          (Intercept)            elev_scale       I(elev_scale^2)                forest
        -2.358250e+01         -8.455268e-01          3.669428e-01         -3.647176e-02
          I(forest^2)                   hfi             I(hfi^2) I(dist_water_scale^2)
         3.299911e-04          1.073385e+01         -4.304843e+00         -2.937807e-02

                           Estimate         S.E.       CI95.lo       CI95.hi Ztest        Zval
(Intercept)           -2.358250e+01 1.343968e-01 -2.384591e+01 -2.331909e+01   *** -175.469248
elev_scale            -8.455268e-01 7.289775e-02 -9.884037e-01 -7.026498e-01   ***  -11.598804
I(elev_scale^2)        3.669428e-01 2.897191e-02  3.101589e-01  4.237267e-01   ***   12.665466
forest                -3.647176e-02 2.459048e-03 -4.129141e-02 -3.165212e-02   ***  -14.831661
I(forest^2)            3.299911e-04 3.021307e-05  2.707746e-04  3.892077e-04   ***   10.922130
hfi                    1.073385e+01 5.169856e-01  9.720574e+00  1.174712e+01   ***   20.762372
I(hfi^2)              -4.304843e+00 4.865518e-01 -5.258467e+00 -3.351219e+00   ***   -8.847655
I(dist_water_scale^2) -2.937807e-02 6.767740e-03 -4.264259e-02 -1.611354e-02   ***   -4.340897
Problem:
 Values of the covariate ‘hfi’ were NA or undefined at 0.28% (23 out of 8071) of the quadrature
points
```

Figure 7: Partial Quadratic Model Summary

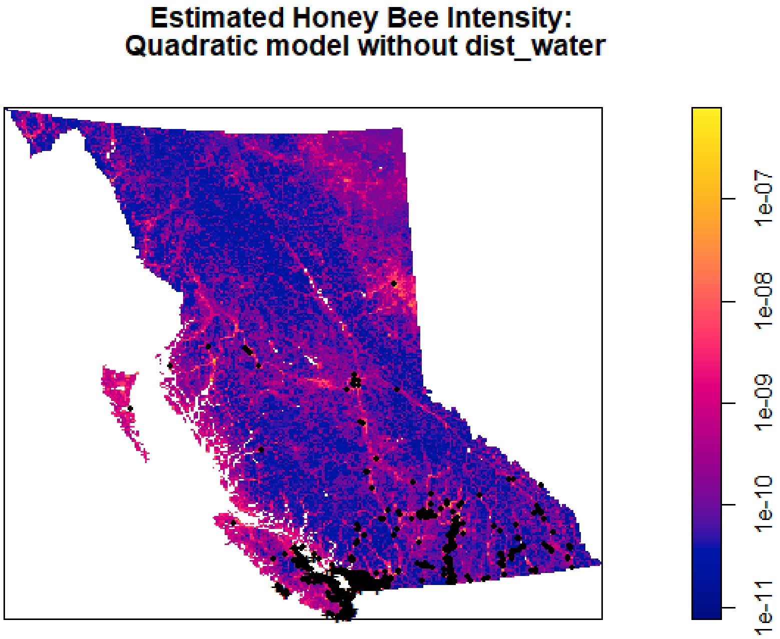
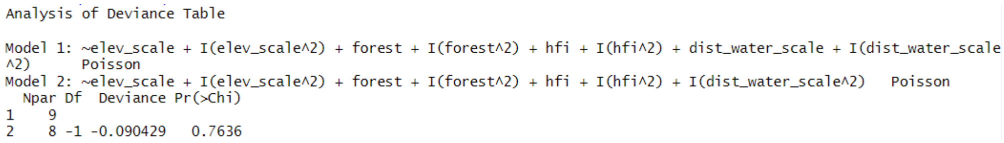

Estimated Honey Bee Intensity:
Quadratic model without dist_water
1e-07
1e-08
1e-09
1e-10
1e-11

Figure 8: Predicted Honey Bee Intensity - Partial Quadratic Model



```
Analysis of Deviance Table

Model 1: ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) + dist_water_scale + I(dist_water_scale
^2)      Poisson
Model 2: ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) + I(dist_water_scale^2)    Poisson
  Npar Df  Deviance Pr(>Chi)
1    9
2    8 -1 -0.090429   0.7636
```

Figure 9: Likelihood Ratio Test - Partial Quadratic Model

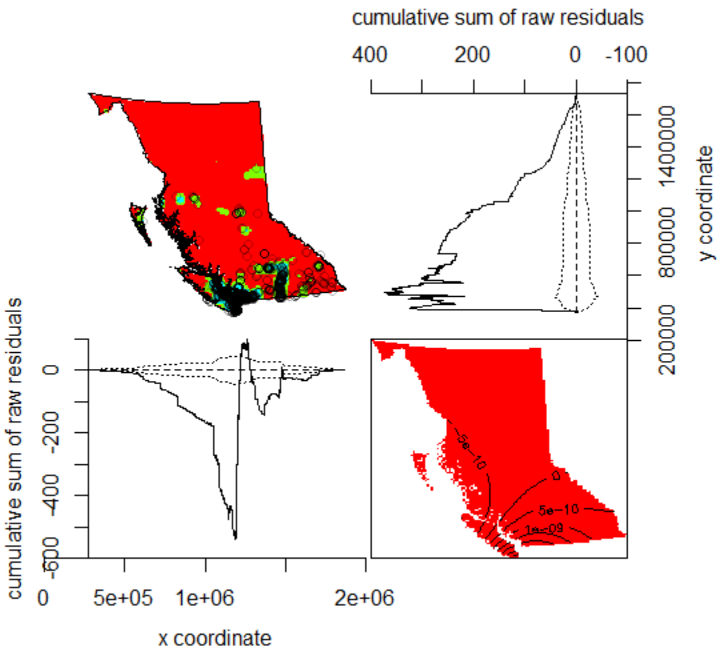
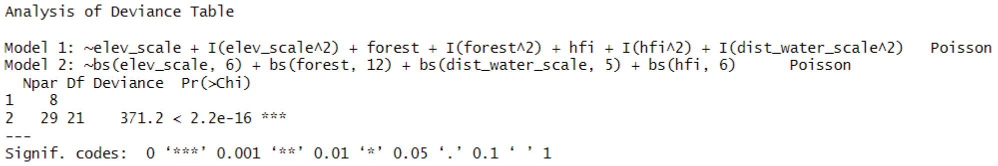


Figure 10: Diagnostics - Quadratic Models

Thus, another type of model, such as a generalized additive model (GAM), seemed appropriate. The first GAM model trialed was a full model with 6 knots on the scaled elevation, 12 knots on the forest coverage, 5 knots on the scaled distance to water, and 6 knots on the HFI. As compared to the quadratic models, this more complex model was supported by the AIC comparison, 8.6469035×10^4 versus 8.6796239×10^4, and the likelihood ratio test, as seen in Figure 11. The tiny p-value means that the quadratic model can be rejected in favour of the current gam model. A review of the partial residuals plot in Figure 12, shows that elevation, distance to water, forest coverage, and human footprint index are all well represented overall. However, as seen in a visualization of the predicted intensity, Figure 13, there are still a lot of low populated areas that have been predicted to have high intensity. The test for deviation between predicted and observed gives evidence, with a p-value of 0.001, to reject the current game model, as seen in Figure 14. The diagnostic plot, seen in Figure 15, shows further evidence that while there was an improvement in the residuals when compared to the quadratic models, there was still room for improvement.



```
Analysis of Deviance Table

Model 1: ~elev_scale + I(elev_scale^2) + forest + I(forest^2) + hfi + I(hfi^2) + I(dist_water_scale^2)   Poisson
Model 2: ~bs(elev_scale, 6) + bs(forest, 12) + bs(dist_water_scale, 5) + bs(hfi, 6)      Poisson
  Npar Df Deviance  Pr(>Chi)
1    8
2   29 21    371.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 11: Likelihood Ratio Test - GAM Model

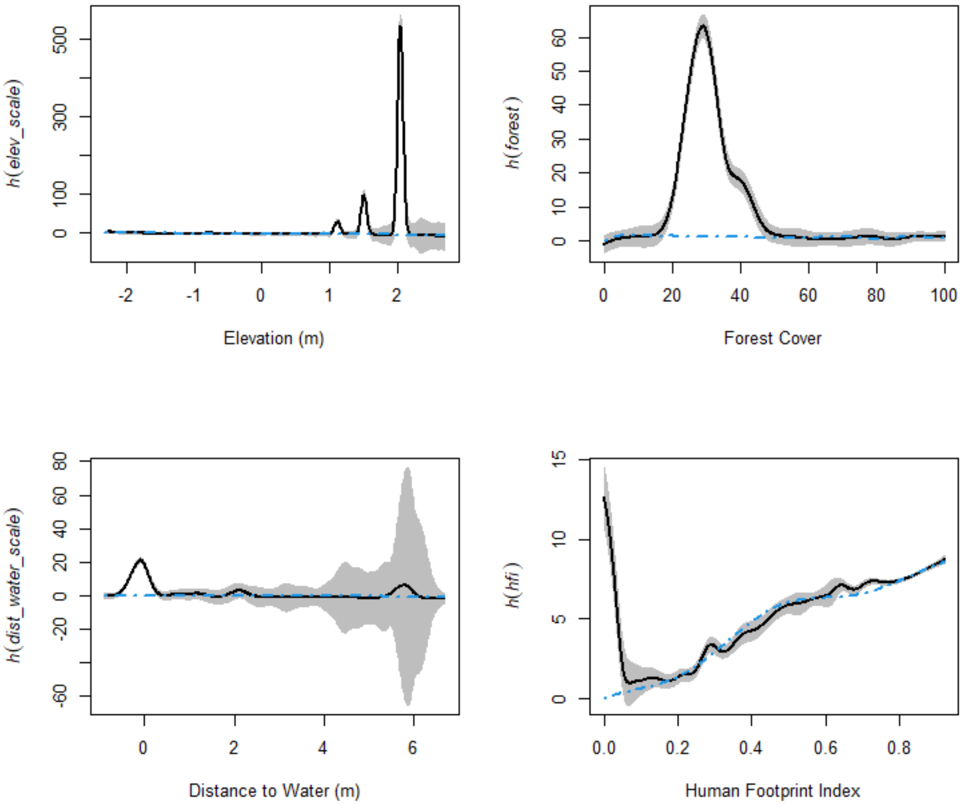


Figure 12: Partial Residuals - GAM Model

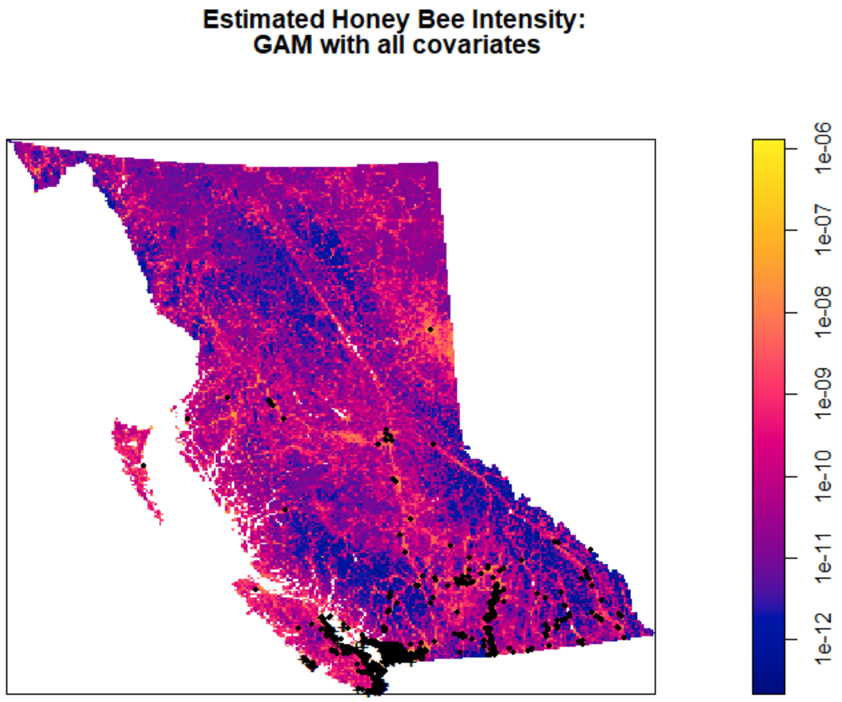
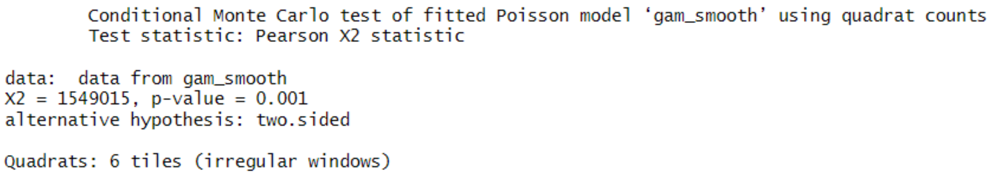


Figure 13: Predicted Honey Bee Intensity - GAM Model



```
        Conditional Monte Carlo test of fitted Poisson model ‘gam_smooth’ using quadrat counts
        Test statistic: Pearson X2 statistic

data:  data from gam_smooth
X2 = 1549015, p-value = 0.001
alternative hypothesis: two.sided

Quadrats: 6 tiles (irregular windows)
```

Figure 14: Predicted Deviance - GAM Model

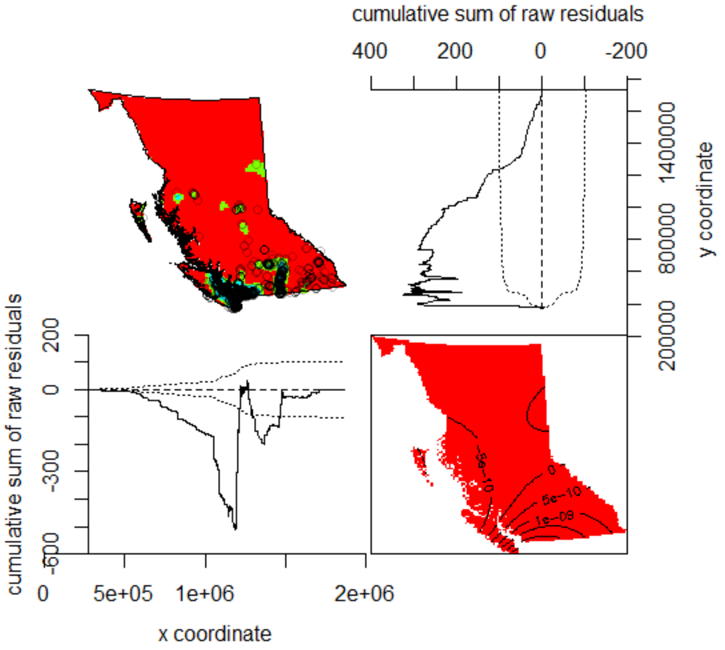
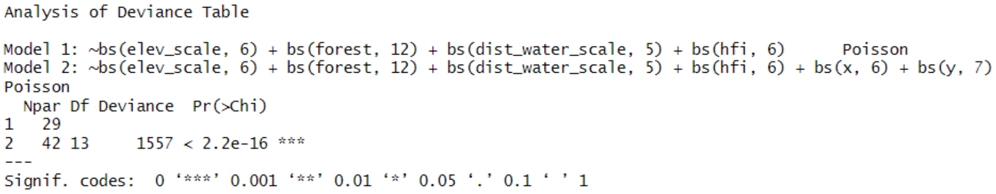


Figure 15: Diagnostics - GAM Model

Since no other covariates were available for consideration, the last gam model was built with the same covariates as the previous model with the addition of the x and y coordinates as proxy covariates. These new predictors had 6 knots on the x coordinate and 7 knots on the y coordinate. When compared to the previous gam model, this new model did better with an AIC of 8.4938034×10^4, as compared to 8.6469035×10^4, and is supported as per the small p-value of $< 2.2e^{-16}$ seen in the likelihood ratio results in Figure 16. The partial residuals plot, in Figure 17, also indicate that all variables are better represented. Although this model should be rejected as per the quadrat test results seen in Figure 18, the resulting p-value of 0.039 is the highest of the 4 models reviewed. Meaning that the deviation between the predicted and observed values was the lowest in this model. Visually, there are some further improvements in the predicted intensity plot, as seen in Figure 19. However, there are still low observation areas that are predicted to have low intensity. This is likely due to the significant amount of residuals in the lower parts of the province as seen in the diagnostics plot in Figure 20.



```
Analysis of Deviance Table

Model 1: ~bs(elev_scale, 6) + bs(forest, 12) + bs(dist_water_scale, 5) + bs(hfi, 6)      Poisson
Model 2: ~bs(elev_scale, 6) + bs(forest, 12) + bs(dist_water_scale, 5) + bs(hfi, 6) + bs(x, 6) + bs(y, 7)
Poisson
  Npar Df Deviance  Pr(>Chi)
1   29
2   42 13     1557 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 16: Likelihood Ratio Test - GAM Model with X-Y

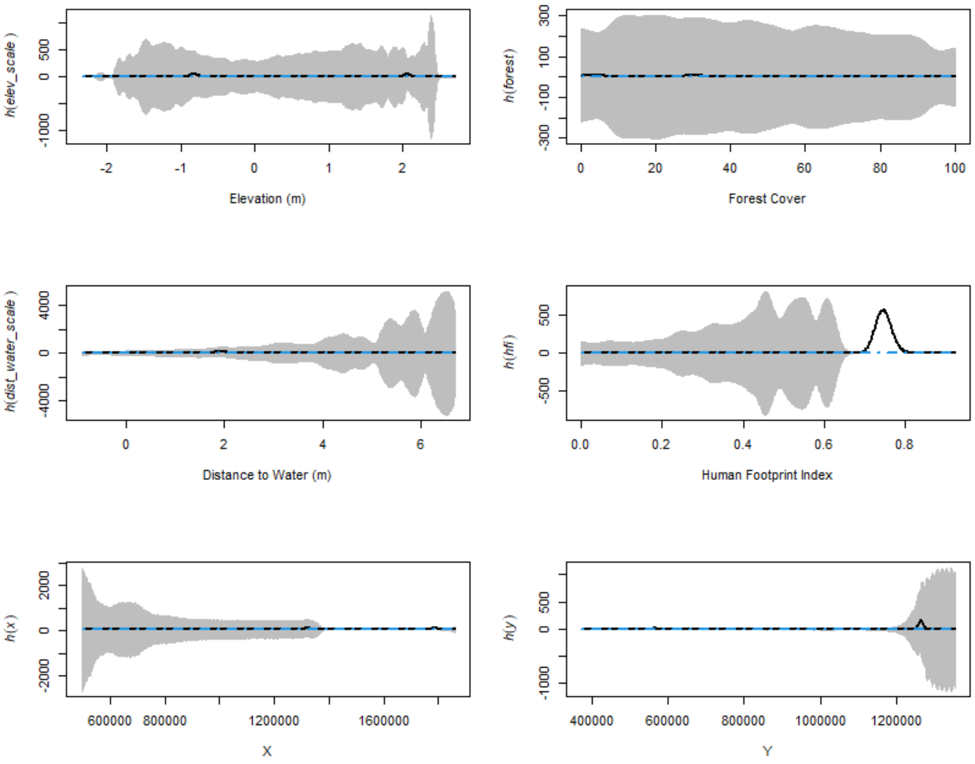
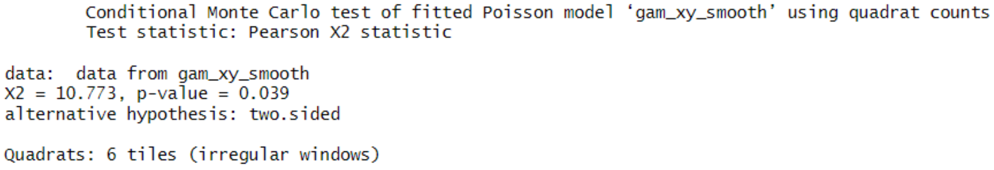


Figure 17: Partial Residuals - GAM Model with X-Y



```
        Conditional Monte Carlo test of fitted Poisson model ‘gam_xy_smooth’ using quadrat counts
        Test statistic: Pearson X2 statistic

data:  data from gam_xy_smooth
X2 = 10.773, p-value = 0.039
alternative hypothesis: two.sided

Quadrats: 6 tiles (irregular windows)
```

Figure 18: Predicted Deviance - GAM Model with X-Y

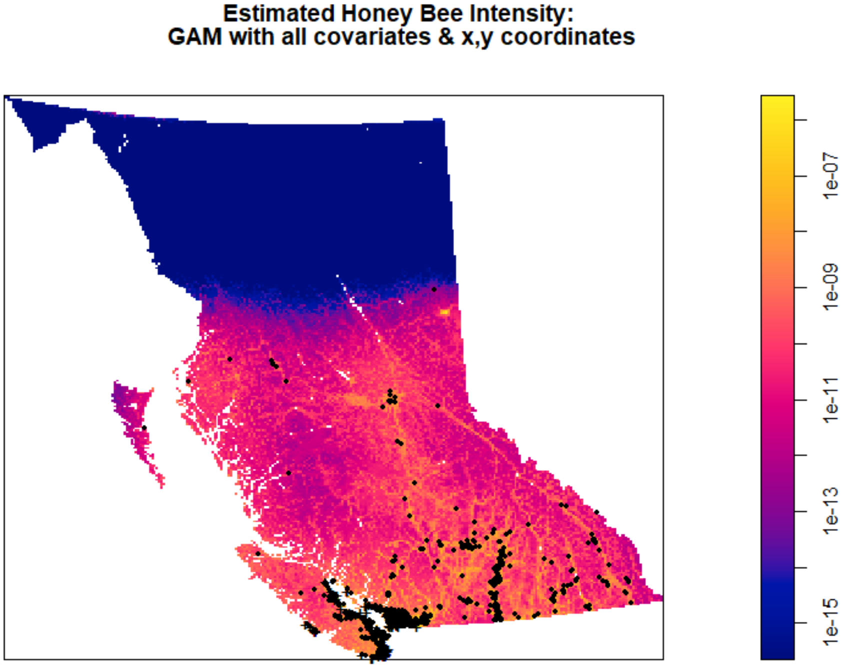

Estimated Honey Bee Intensity:
GAM with all covariates & x,y coordinates
1e-07
1e-09
1e-11
1e-13
1e-15

Figure 19: Predicted Honey Bee Intensity - GAM Model with X-Y

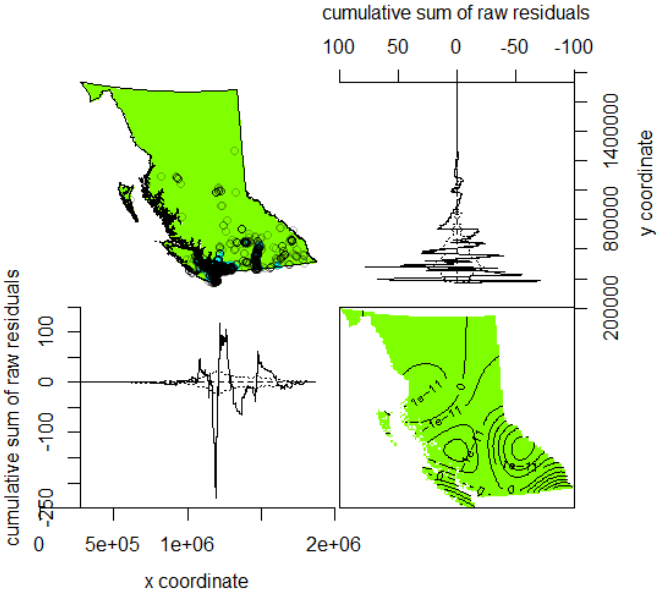


Figure 20: Diagnostics - GAM Model with X-Y

3 Discussion

The number of data points in the dataset is 2614. Test of homogeneity was done to see if the intensity is homogeneous and the intensity was found to be inhomogeneous which was also checked using quadrat count and validated using kernel estimation. On plotting the Kernel estimation of intensity and overlaying the points on it, it was seen that the points are aggregated more towards the south region of BC which was also confirmed by the hotspot analysis. Next, we proceeded to check the relation of intensity with the different covariates and the were all observed to be non linearly related with intensity. This gave us a good starting point for initial model specification. Inhomogeneous K-function (Kinhom) was used to see if there is evidence of clustering. It was found that only at very low values of r (<10000 units) there was evidence of clustering. This finding was confirmed by the inhomogeneous pair correlation function envelopes done with 19 iterations which equals a significance level of 0.05. After reviewing the linear and quadratic models, as well as two different GAM models, it was found that the GAM model with x-y coordinates performed the best. Although there were still low-population areas predicted to have high intensity, this model had the lowest deviation between predicted and observed values, as evidenced by the quadrat test results. The addition of the x and y coordinates as proxy covariates, as well as the use of generalized additive models, were likely the factors that contributed to the overall improved performance of this model. In conclusion, after comparing four models, including linear, quadratic, and two different GAM models, it was found that the generalized additive model with x-y coordinates as proxy covariates performed the best. This model had the lowest deviation between predicted and observed values and was the most accurate in predicting honey bee intensity in British Columbia. However, there were still low-population areas predicted to have high intensity, indicating that further improvements can be made. Therefore, additional covariates are required to accurately estimate honey bee intensity in BC, and future studies should explore the inclusion of other factors to improve the model's performance.

4 References - All

Klein, A. M., Vaissiere, B. E., Cane, J. H., Steffan-Dewenter, I., Cunningham, S. A., Kremen, C., & Tscharntke, T. (2007). Importance of pollinators in changing landscapes for world crops. Proceedings of the Royal Society B: Biological Sciences, 274(1608), 303–313. https://doi.org/10.1098/rspb.2006.3721