Red Foxes in BC: an Analysis of the Spatial Point Process

Nowshaba Durrani, Ricky Heinrich, Viji Rajagopalan

2023-04-30

Introduction

Red Foxes (Scientific Name - Vulpes Vulpes) are a widespread animal in Canada, and are present in British Columbia. In the Global Biodiversity Information Facility (GBIF) database, there are approximately 640,000+ geo-references records for this species around the world, and 242 number of entries are classified as from British Columbia. According to the Canadian Wildlife Federation (CWF), farmers used to regard red foxes as pests, as they would steal chickens. This might suggest that they hang around farms and human activity to take advantage of food sources. As well, the CWF says that they "prefer open fields", which suggests that they are more likely to be found in areas with less forest cover. We are interested in finding out if the occurrence of red fox data agrees with these claims, and if there are more factors that affect their presence. Treating the occurrence of red foxes as a point process, we will investigate its characteristics and attempt to model its intensity.

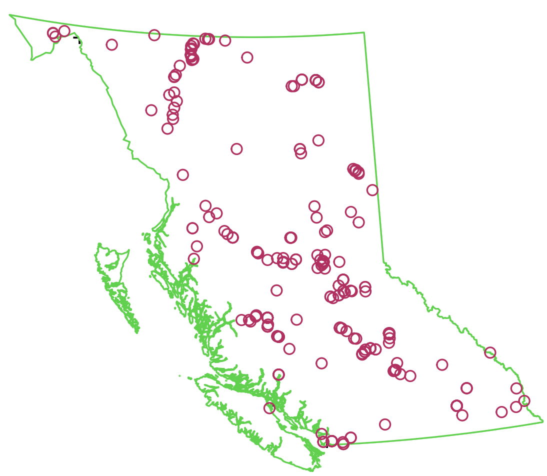


Figure 1: Occurence of Red Foxes in BC

Methods

The data comes from the GBIF databases. We used the package `rgbif` to access the 'Vulpes Vulpes' data from R directly, sorting by instances occurring in BC. We've extracted the longitude and latitude data from this, and converted it appropriately using the `sp` package.

Our second source of data contained the BC Window object, as well as possible covariate data: `Elevation`, `Forest Cover`,`HFI` and `Distance to Water`.

We mainly used the `spatstat` package functions to conduct our analyses. Other supporting libraries for spatial data used are `maptools`, `sp` and `sf`. First we used `ppp` function to build a point pattern process object with the converted coordinates of the Red Fox locations from the GBIF data and the window from our second data source.

To conduct first moment analysis, we used functions from the aforementioned `spatstat` package. We did a quadrat test as well as hotspot analysis to gain insight into the homogeneity assumption of the point process. For second moment analysis, we looked into Ripley's K-function and pair correlation function using functions from `spatstat`. This provides us with insight into possible clustering tendencies of the point process.

Next we looked into the relationship of the intensity with each covariate. For smoothing estimate of the 4 covariates, `rhohat` function was used which helped us to see the relationship between red fox occurrence and each of the covariates. We used correlation function to study the correlation or collinearity of the covariates. Simple models with each of the covariates is first fit and based on the model results and the relationship of covariates with red fox data, we proceeded to fit models with a combination of covariates.

Fitted models were then visualized using `plot` function. We investigated patters in the residuals using `diagnose` function and the partial residuals with each of the covariate is studied using `parres` function to understand the fit of each covariate in the model.

The `splines` package was used to build a GAMs model using functions `ppm` from spatstat package along with `bs` function that generates a basis matrix for representing the family of piecewise polynomials. The purpose was to build a non-parametric model for the data. This model was also diagnosed using techniques mentioned above.

We then assessed the models with `quadrat.test` and `anova` functions to understand any significant deviates from given occurrences and compare and contrast the models. Using `effectfun()`, we also looked at the influence of individual covariates. This computes the intensity of a fitted point process model as a function of one of its covariates. Using AIC as model metric and considering model parsimony and simplicity, we selected the final model for the red fox data.

Exploratory Analysis

First Moment Analysis

We start with investigating whether the occurrence of red foxes in BC seems homogeneous, as it will inform our steps to define the intensity. We have conducted a quadrat test of homogeneity with both 5 x 5 and 10 x 10 quadrats. These quadrats are shown in Figure 2, where we can visually tell that the intensity in each quadrats are not the same. The quadrat test for both the 5 x 5 and the 10 x 10 quadrats provide a p-value of 2.2e-16, confirming that the varied intensities are not due to chance alone, but rather due to an inhomogeneous point process.

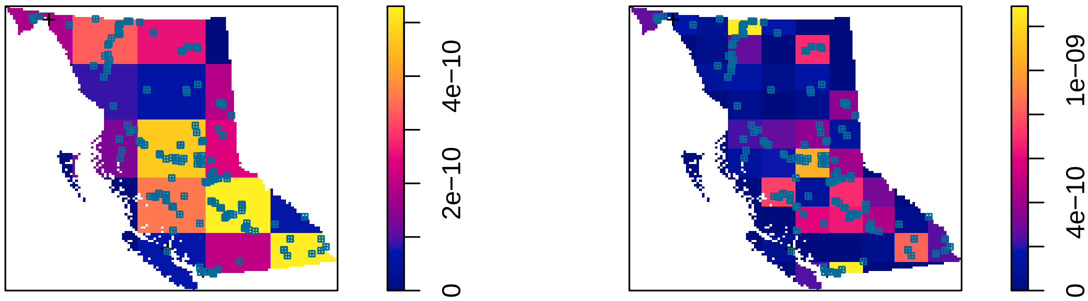


Figure 2: Intensity of Quadrat counts of Red Fox occurences, left 5x5, right 10x10

As the next step, we investigate for any hot spots in the occurrences of red foxes. In Figure 3, we can see that hotspots appear scattered and of moderately high density. It seems like these occurrences are more inland.

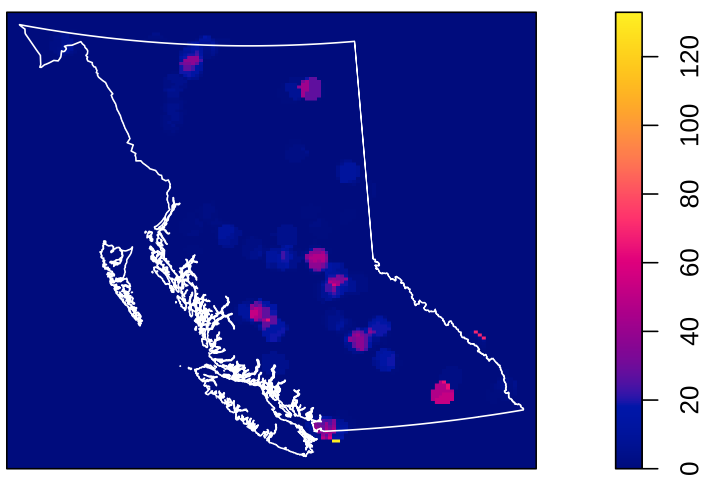


Figure 3: Hotspot of Red Foxes

2nd Moment Analysis

Ripley's K-function provides information on whether there are significant deviations from independence between points. Taking into account that the intensity of red fox occurrences appear inhomogeneous, we can see in Figure 4 that there is some evidence of clustering up to a certain distance, as the black line, indicating the observed data, is separate from the 95% confidence bands of the values expected with no clustering.

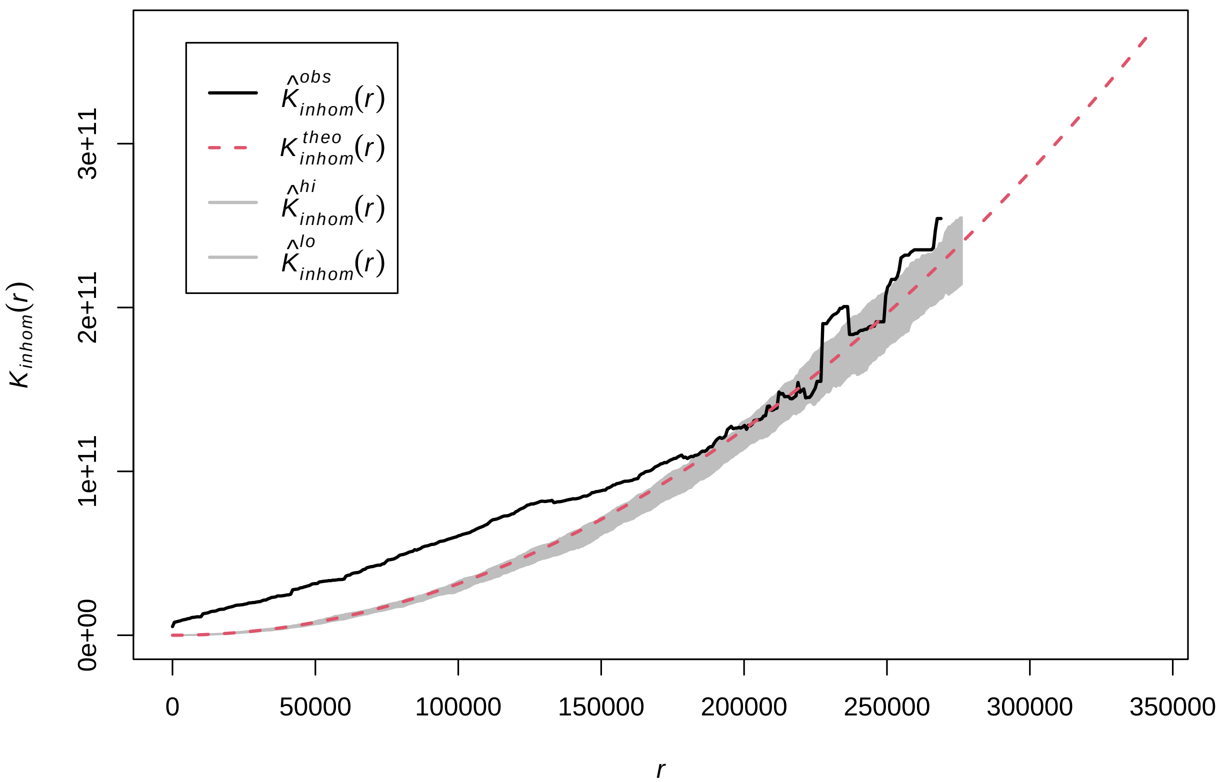


Figure 4: Ripley's K function with border correction assuming inhomogeneity

To get a sense of the distances for which clustering occurs, we used the pair correlation function.

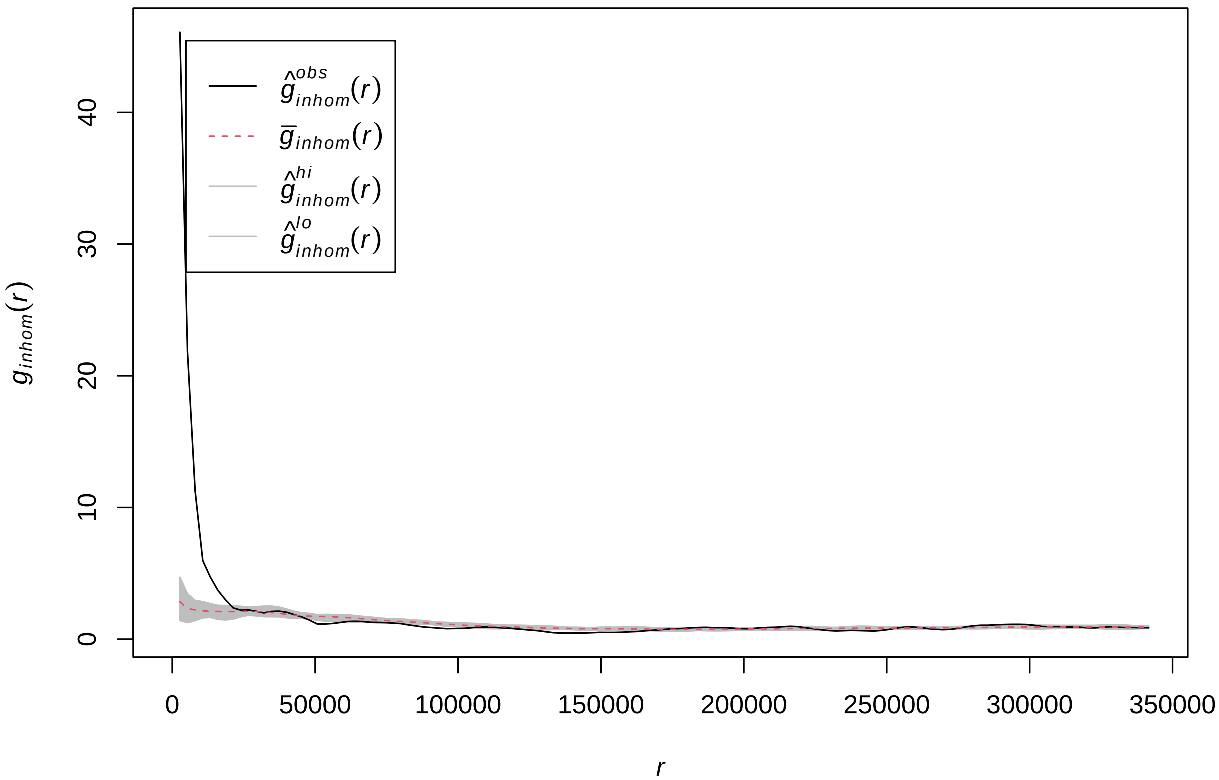


Figure 5: Pair correlation function assuming inhomogeneity

Figure 5 shows evidence for clustering at distances smaller than around 23 000m, or 23km but after that the observed values are not significantly different that those expected from a random spatial process.

Relationship with Covariates

Our data includes 4 covariates which we are exploring: the `Elevation`, `Forest Cover`, human footprint inventory (`HFI`), and `Distance to Water`. Given our research questions, we expect `HFI` and `Forest Cover` to have a relationship with red fox occurrences, however we also investigate the other two covariates.

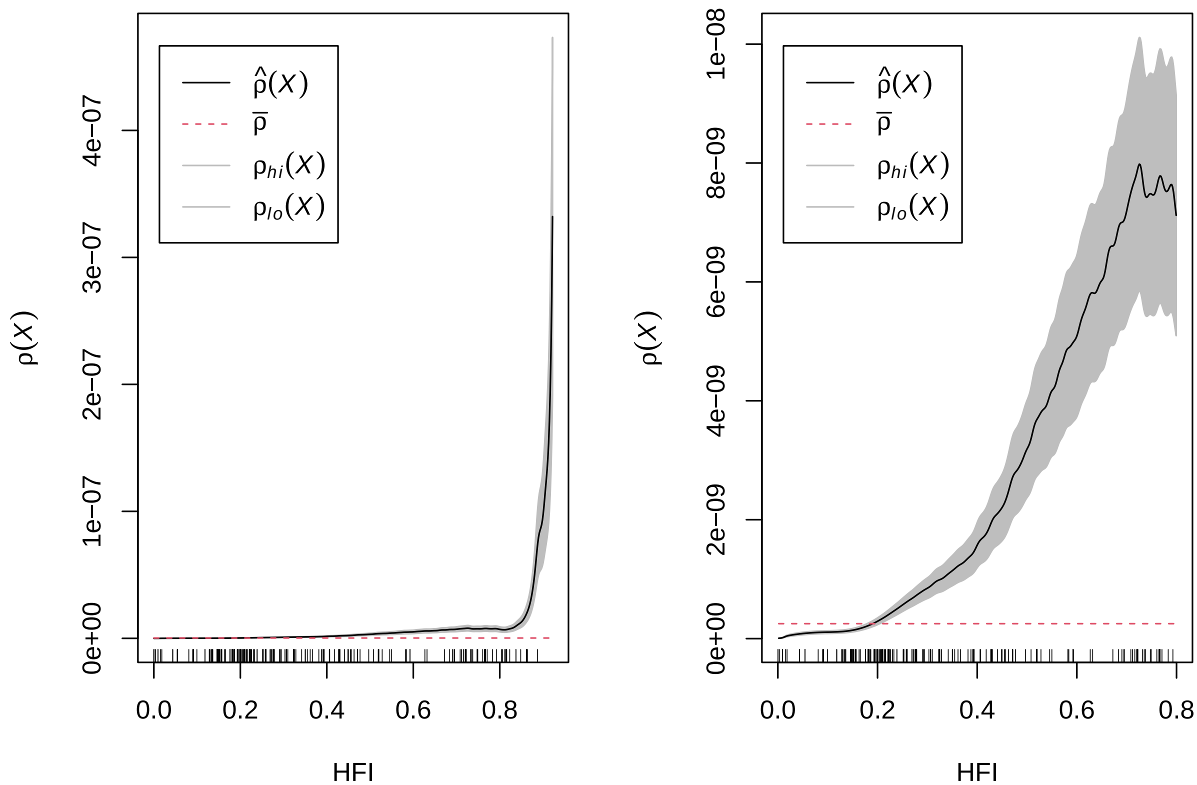


Figure 6: Effect of ‘HFI‘ on intensity of red foxes

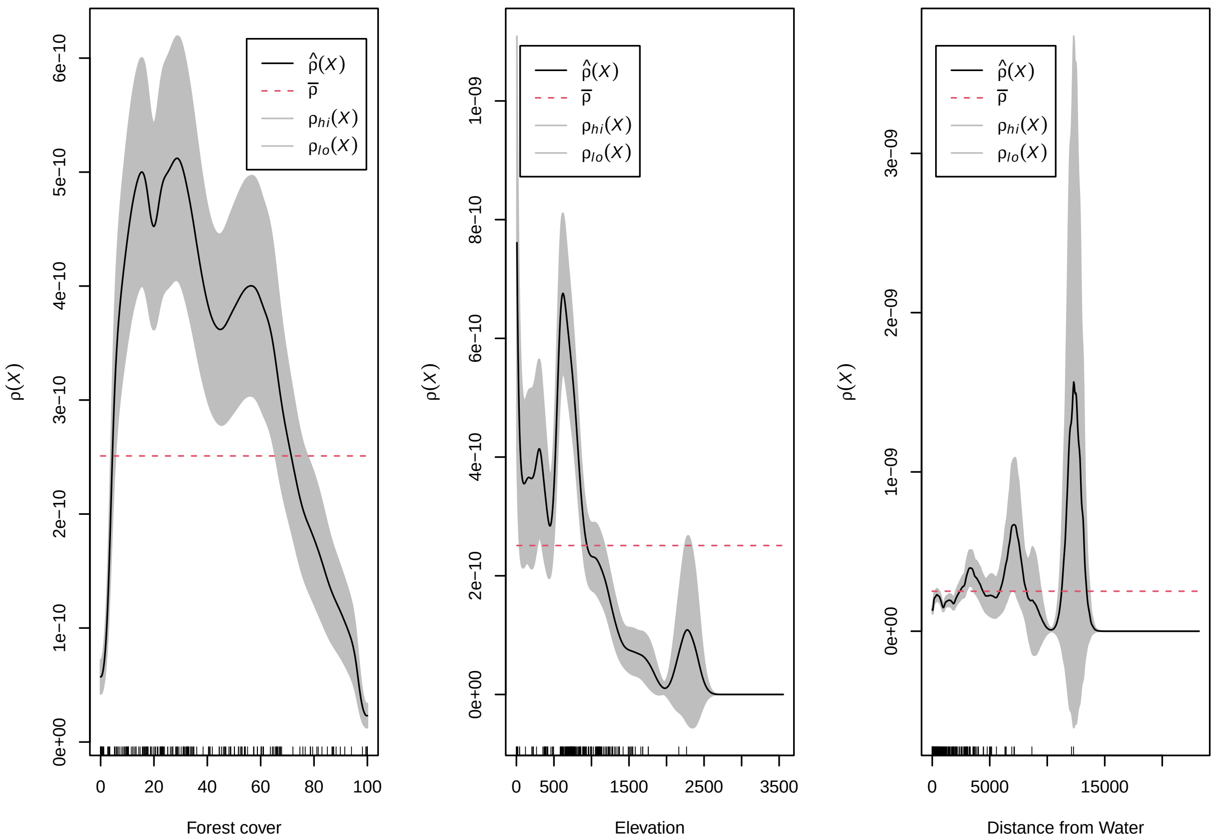


Figure 7: Effect of ‘Forest Cover‘, ‘Elevation‘, and ‘Distance to Water‘ on intensity of red foxes

In the left plot of Figure 6, we could be fooled into thinking that there is no relationship between `HFI` and intensity of red foxes up to around `HFI` = 0.4, until which it seems like an exponential relationship. However, zooming in from `HFI` 0 to 0.8, we see that the confidence bands don’t intersect at all with the red line, which is the expected value given no relationship. This relationship appears non-linear and possibly exponential, where the greatest intensity of observed red foxes occurs at high `HFI`s. This relationship was expected, as our dataset is not exhaustive but rather is crowdsourced, and naturally foxes are more likely to be observed by humans in spaces with higher `HFI`s.

In Figure 7, we see that there seems to be non-linear relationship between `Forest Cover` and number of red foxes observed. The observance increases with increase in `Forest Cover` at intermediate coverage and then it decreases. We also see that there is non-linear relationship with `Elevation`. The relationship appears to be non-linear as the graph is showing different results for different `Elevation` and we cannot see any type of specific pattern from the same. In case of `Distance to Water`, we don’t observe a significant deviation in observed red foxes than expected by chance, indicating that it is not a useful covariate to model.

Fit models for the covariets

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~HFI
```

```
##
## Fitted trend coefficients:
## (Intercept)         HFI
##   -23.31421     5.98177
##
##              Estimate      S.E.    CI95.lo    CI95.hi Ztest       Zval
## (Intercept) -23.31421 0.1058336 -23.521645 -23.106785   *** -220.29132
## HFI           5.98177 0.2148471   5.560677   6.402862   ***   27.84199
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points
```

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~HFI + exp(HFI)
##
## Fitted trend coefficients:
## (Intercept)         HFI    exp(HFI)
##   -13.71641    22.17067   -10.35893
##
##              Estimate     S.E.   CI95.lo    CI95.hi Ztest      Zval
## (Intercept) -13.71641 1.404309 -16.46881 -10.964015   *** -9.767372
## HFI          22.17067 2.382441  17.50117  26.840165   ***  9.305861
## exp(HFI)    -10.35893 1.522442 -13.34286  -7.374997   *** -6.804153
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points
```

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Forest + I(Forest^2)
##
## Fitted trend coefficients:
##   (Intercept)        Forest   I(Forest^2)
## -2.225373e+01  3.996522e-02 -5.280288e-04
##
##                  Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept) -2.225373e+01 1.330469e-01 -2.251450e+01 -2.199297e+01   ***
## Forest       3.996522e-02 7.091557e-03  2.606603e-02  5.386442e-02   ***
## I(Forest^2) -5.280288e-04 7.699618e-05 -6.789385e-04 -3.771191e-04   ***
##                   Zval
## (Intercept) -167.262353
## Forest         5.635606
## I(Forest^2)   -6.857857
```

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Elevation
##
## Fitted trend coefficients:
```

```
##   (Intercept)     Elevation
## -20.775522387  -0.001401046
##
##                  Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept) -20.775522387 0.1298745141 -21.030071757 -20.520973016   ***
## Elevation    -0.001401046 0.0001421902  -0.001679734  -0.001122358   ***
##                    Zval
## (Intercept) -159.966122
## Elevation     -9.853321
```

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Dist_Water
##
## Fitted trend coefficients:
##   (Intercept)    Dist_Water
## -2.211986e+01  1.071969e-05
##
##                  Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept) -2.211986e+01 8.745074e-02 -2.229126e+01 -2.194846e+01   ***
## Dist_Water   1.071969e-05 3.419673e-05 -5.630466e-05  7.774404e-05
##                     Zval
## (Intercept) -252.9408319
## Dist_Water     0.3134712
```

We have fitted 6 models and we came to observe that `HFI`, exp(HFI), `Forest Cover`, I(Forest^2) and `Elevation` seems to be highly significant however `Distance to Water` seems to be in-significant for the occurrence of red foxes in the BC area. When we check the AIC values and it is seen that `HFI` and `HFI`(exp) has lower values so we can consider these models to be a better fit.

```
##           Model      AIC
## 1           HFI 10468.83
## 2      HFI(Exp) 10420.78
## 3  Forest Cover 10931.26
## 4     Elevation 10896.99
## 5 Dist to Water 11000.30
```

```
##             ..1         ..2         ..3         ..4
## ..1  1.00000000  0.06616335 -0.26217406  0.04822162
## ..2  0.06616335  1.00000000 -0.26625709  0.13249159
## ..3 -0.26217406 -0.26625709  1.00000000 -0.03497584
## ..4  0.04822162  0.13249159 -0.03497584  1.00000000
```

We see that no covariate is strongly correlated with another, and so we can move on without taking so into account and treating the covariates as independent.

Model Fitting

From our assessment on individual covariates, we see that there is a non-linear relationship between the covariates and the red fox point data. Based on this knowledge, we move forward to fit the first base model with covariates that are showing strong trends with red fox data and also of research interest. For this purpose, the selected covariates are 1. `Elevation` 2. `HFI` and 3. `Forest Cover`.

We built the model with the linear and quadratic terms of covariates `Elevation`, `HFI` and `Forest Cover`. We call this model as `Model 1`.

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~Forest + I(Forest^2) + HFI + I(HFI^2) + Elevation +
## I(Elevation^2)
##
## Fitted trend coefficients:
##    (Intercept)         Forest    I(Forest^2)            HFI       I(HFI^2)
##  -2.413811e+01   6.206401e-03  -1.160232e-04   1.185642e+01  -7.371997e+00
##      Elevation I(Elevation^2)
##   1.364005e-03  -9.678328e-07
##
##                     Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept)    -2.413811e+01 3.220178e-01 -2.476925e+01 -2.350696e+01   ***
## Forest          6.206401e-03 7.444014e-03 -8.383599e-03  2.079640e-02
## I(Forest^2)    -1.160232e-04 7.767304e-05 -2.682596e-04  3.621315e-05
## HFI             1.185642e+01 1.088713e+00  9.722583e+00  1.399026e+01   ***
## I(HFI^2)       -7.371997e+00 1.232831e+00 -9.788301e+00 -4.955694e+00   ***
## Elevation       1.364005e-03 5.226588e-04  3.396126e-04  2.388397e-03    **
## I(Elevation^2) -9.678328e-07 2.819525e-07 -1.520449e-06 -4.152160e-07   ***
##                      Zval
## (Intercept)    -74.9589241
## Forest           0.8337439
## I(Forest^2)     -1.4937385
## HFI             10.8903097
## I(HFI^2)        -5.9797320
## Elevation        2.6097427
## I(Elevation^2)  -3.4326094
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points
```

The results indicate that covariates `HFI` and `Elevation` show strong significance and explain occurrence of red fox data in both linear and quadratic terms. Though `Forest Cover` showed non-linear relationship in second moment analysis, in this combined model with other covariates, its not significant.

As our next step, we drop the covariate `Forest Cover` from the model and build the next model with `HFI` and `Elevation` with linear and quadratic terms. We call this as `Model 2`.

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~HFI + I(HFI^2) + Elevation + I(Elevation^2)
##
## Fitted trend coefficients:
##    (Intercept)            HFI       I(HFI^2)      Elevation I(Elevation^2)
##  -2.426748e+01   1.242099e+01  -7.734861e+00   1.255468e-03  -9.320293e-07
##
```

```
##                     Estimate         S.E.       CI95.lo       CI95.hi Ztest
## (Intercept)    -2.426748e+01 2.983606e-01 -2.485226e+01 -2.368270e+01   ***
## HFI             1.242099e+01 1.073719e+00  1.031654e+01  1.452544e+01   ***
## I(HFI^2)       -7.734861e+00 1.222941e+00 -1.013178e+01 -5.337941e+00   ***
## Elevation       1.255468e-03 5.201189e-04  2.360535e-04  2.274882e-03     *
## I(Elevation^2) -9.320293e-07 2.837978e-07 -1.488263e-06 -3.757958e-07    **
##                     Zval
## (Intercept)    -81.336078
## HFI             11.568195
## I(HFI^2)        -6.324804
## Elevation        2.413809
## I(Elevation^2)  -3.284131
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points
##
## *** Fitting algorithm for 'glm' did not converge ***
```

Based on the model results, both `HFI` and `Elevation` show strong significance and explain occurrence of red fox data in both linear and quadratic terms. For understanding the fit of the model, we plot the model and red fox locations together by overlaying the observed data on the predicted values from the model.

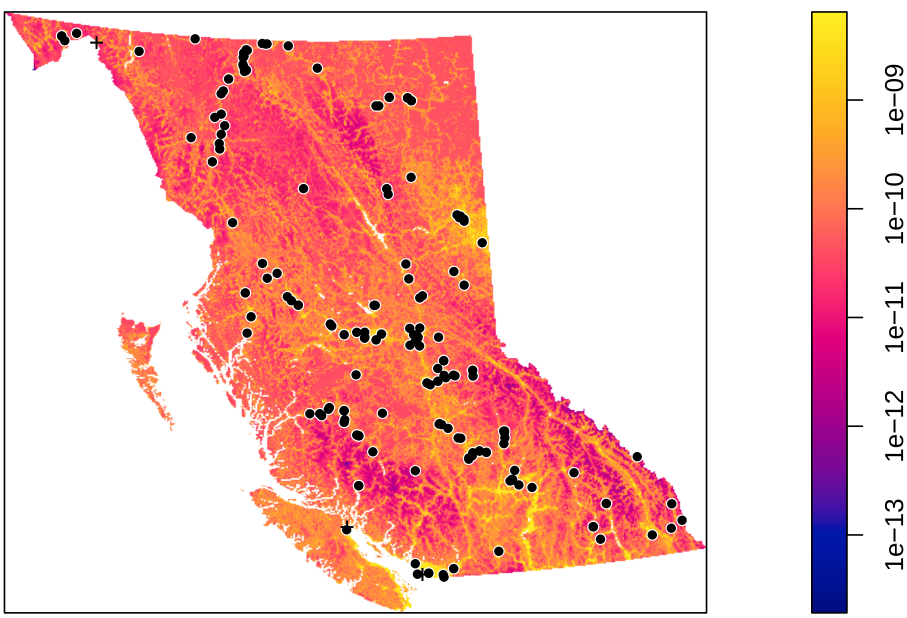


Figure 8: Red fox occurences vs predicted model fit

We see in Figure 8 that the red fox locations are captured well with yellow color in the background which is an indicator of high intensity area as predicted by the model. As the model seem to capture the red fox locations reasonably well, we then proceed with other diagnostics to validate the model and compare with other models to select the best.

First we use the `diagnose` function to validate `Model 2` based on residuals. To the top right of Figure 9, we can see the residual plot of the cumulative sum of raw residuals against y coordinates. The residuals are showing a good fit in the intermediate y coordinates range and the negative high residuals for high and low coordinates indicates the model predictions are higher than actuals. The the bottom left is the residual plot of the cumulative sum of raw residuals against x coordinates. The model overall has a good fit as seen in the plot with residuals mostly within the dotted band and the prediction is low when compared to actual in the higher x coordinate zones.

Overall, the model is providing a good fit in the intermediate x and y coordinate areas and has tendency to deviate in the high and low coordinate areas of BC.

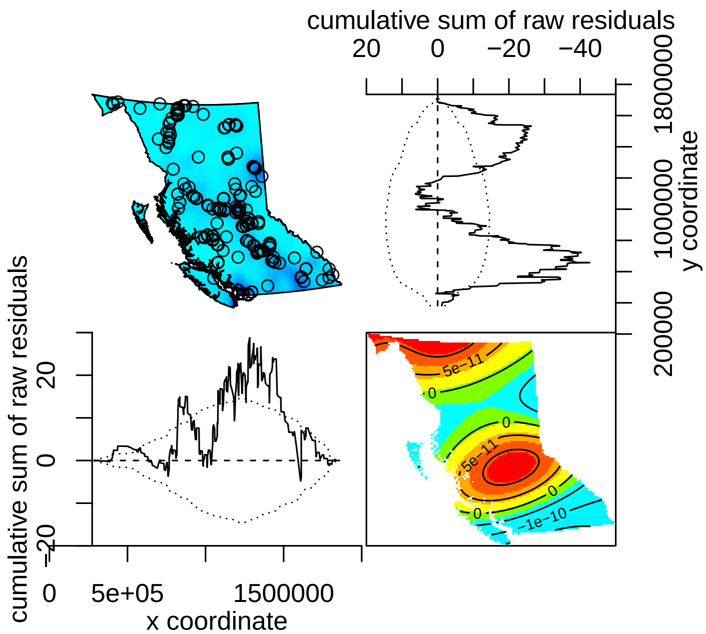


Figure 9: Plotted residuals of Model 2

```
## Model diagnostics (raw residuals)
## Diagnostics available:
##  four-panel plot
##  mark plot
##  smoothed residual field
##  x cumulative residuals
```

```
##  y cumulative residuals
##  sum of all residuals
## sum of raw residuals in entire window = -5.478e-06
## area of entire window = 9.483e+11
## quadrature area = 9.39e+11
## range of smoothed field =  [-1.561e-10, 1.393e-10]
```

```
##
##  Chi-squared test of fitted Poisson model 'fit_red2' using quadrat
##  counts
##
## data:  data from fit_red2
## X2 = 152.84, df = 16, p-value < 2.2e-16
## alternative hypothesis: two.sided
##
## Quadrats: 21 tiles (irregular windows)
```

Additionally, we do a quadrat test on the model which uses a chi-squared test. We get a small p-value indicating the model has significant deviations from the observed data. So we can conclude that this model is useful one however it is not a perfect fit.

We move on to assess this model deeply against a few other models and also investigate further to decide if this is the best choice among the models evaluated.

Model Selection and Validation

With a model that fits the data identified, we proceed to do a thorough validation of this model and also compare with a few other models to select the best one for our data.

First, we start with a simple test evaluating the AIC score of `Model 1` and `Model 2`. The AIC scores are 10401.62 and 10403.57 respectively. We can see that there is not a huge difference in terms of this score between the models. A likelihood ratio test also suggests that there is no evidence of significant difference in performance between the two models.

As we are interested in a parsimonious model, we prefer `Model 2` out of the two options, as it has only two covariates: `Elevation` and `HFI`.

```
## Analysis of Deviance Table
##
## Model 1: ~HFI + I(HFI^2) + Elevation + I(Elevation^2)     Poisson
## Model 2: ~Forest + I(Forest^2) + HFI + I(HFI^2) + Elevation + I(Elevation^2)      Poisson
##   Npar Df Deviance Pr(>Chi)
## 1    5
## 2    7  2   5.9418  0.05126 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Next, next we look at the partial residuals of `Model 2` for each of the covariates, showing the fitted effect of a covariate alongside the observed effect.

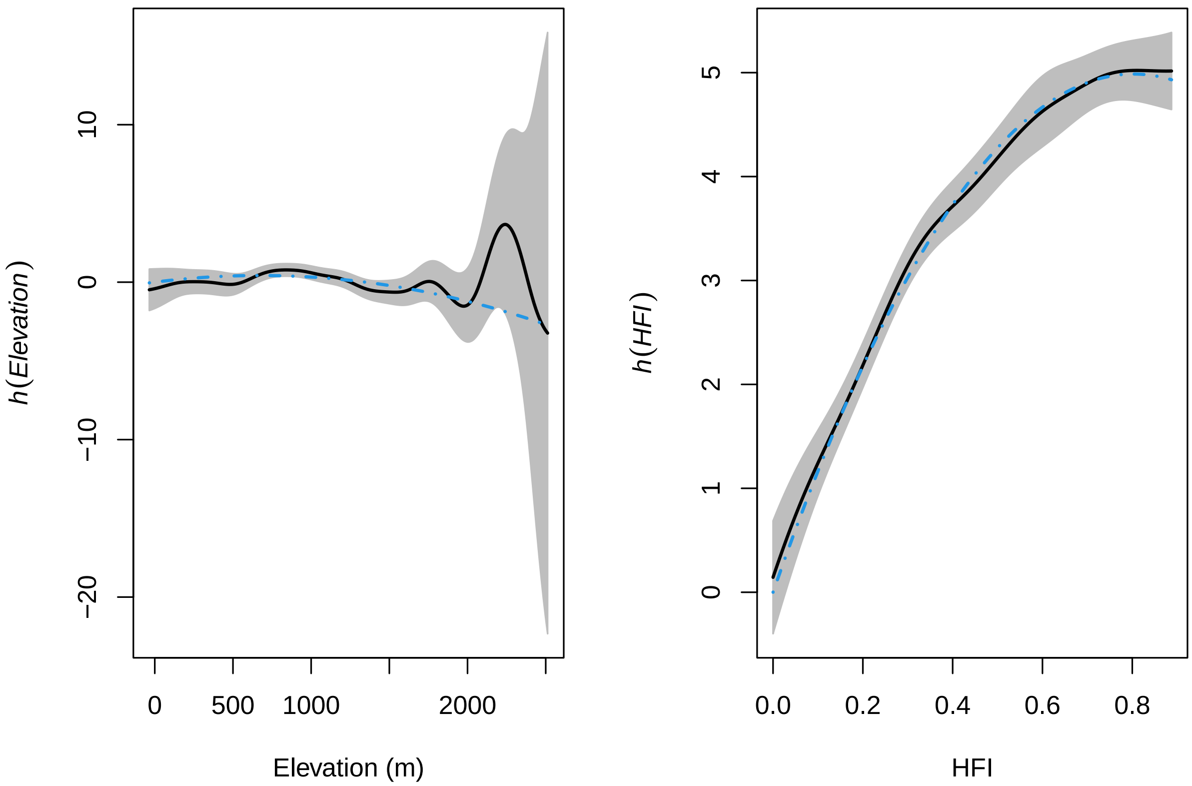


Figure 10: Partial residuals for Model 2 covariates

Based on the plot for `HFI`, on the right side of Figure 10, we see that the model is capturing the patterns in the data really well. Looking at the plot for `Elevation` on the left side, the model is capturing the patterns in the data very well except for higher `Elevation`.

We compute the intensity of a fitted point process model as a function of one of its covariates to look at the influence of the individual covariates.

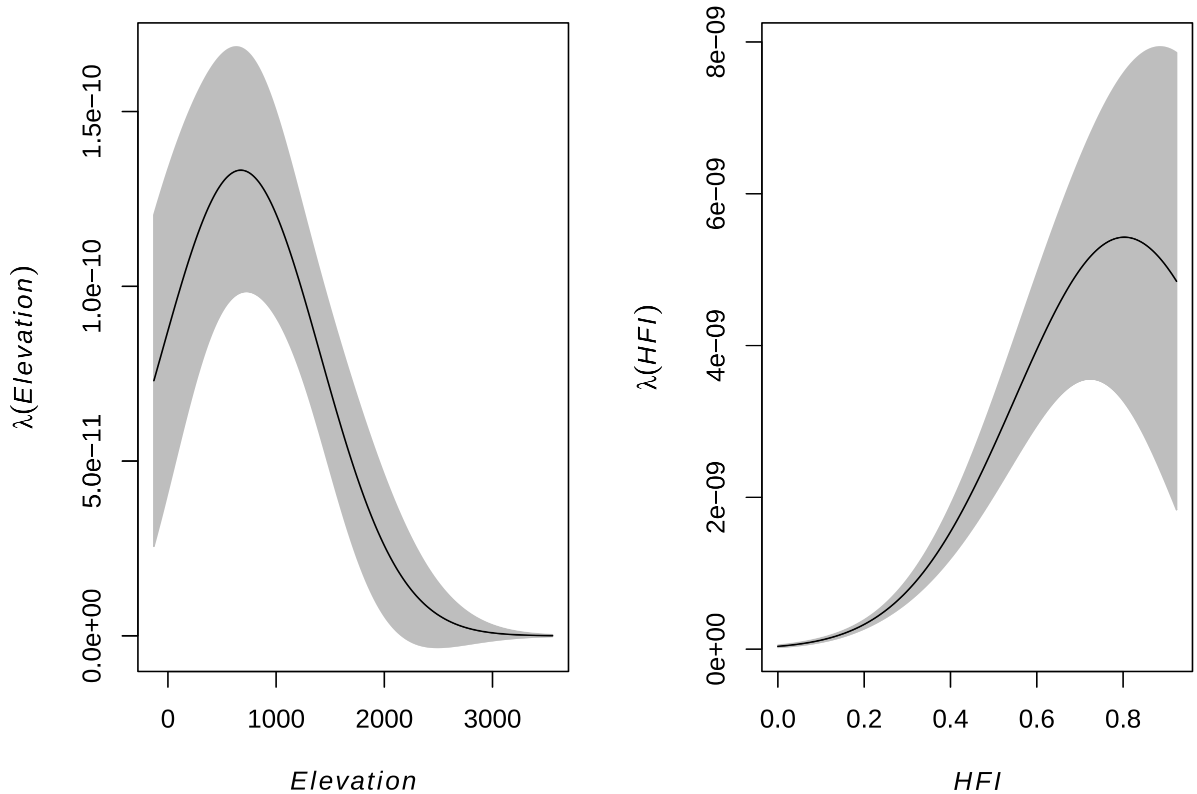


Figure 11: Right: Elevation effect at mean HFI, Left: HFI effect at mean elevation

From Figure 11, we can see that intensity of the model can be described well as a function of `Elevation` and `HFI`.

Though the `Model 2` evaluation so far is very promising, we see that the intensity or occurrence of red fox at higher `Elevation` is not captured well. We try to improve it by adding a higher order polynomial for `Elevation` but it results in convergence error. So, we compare with a non parametric alternative using an additive modelling framework (GAMs) as it allows more flexibility. We call this as our `Model 3`.

```
## Nonstationary Poisson process
## Fitted to point pattern dataset 'parks_ppp'
##
## Log intensity:  ~bs(Elevation, 12) + bs(HFI, 5)
##
## Fitted trend coefficients:
##         (Intercept)  bs(Elevation, 12)1  bs(Elevation, 12)2  bs(Elevation, 12)3
##        -23.06011643         -2.50238970         -0.03994102         -1.39988574
##  bs(Elevation, 12)4  bs(Elevation, 12)5  bs(Elevation, 12)6  bs(Elevation, 12)7
##          0.22518356         -1.66175832         -0.54356329         -0.57428415
##  bs(Elevation, 12)8  bs(Elevation, 12)9 bs(Elevation, 12)10 bs(Elevation, 12)11
##         -1.89370408         -1.12420362         -4.85151633          6.54909789
## bs(Elevation, 12)12         bs(HFI, 5)1         bs(HFI, 5)2         bs(HFI, 5)3
```

```
##        -29.55917527          0.33782061          0.39850473          4.19857566
##         bs(HFI, 5)4         bs(HFI, 5)5
##          4.73799410          4.86649751
##
## For standard errors, type coef(summary(x))
## Problem:
##  Values of the covariate 'HFI' were NA or undefined at 0.56% (12 out of 2137)
## of the quadrature points
```

For a quick assessment, we compare the AIC score of both models and do a quadrat test to validate if one model is superior to the other. The resulting AIC score for the GAMs model is 10408.13 which is higher than the `Model 2`. The quadrat test has a p-value greater than a significance value of 0.05 which tells us that any one model is not superior to the other.

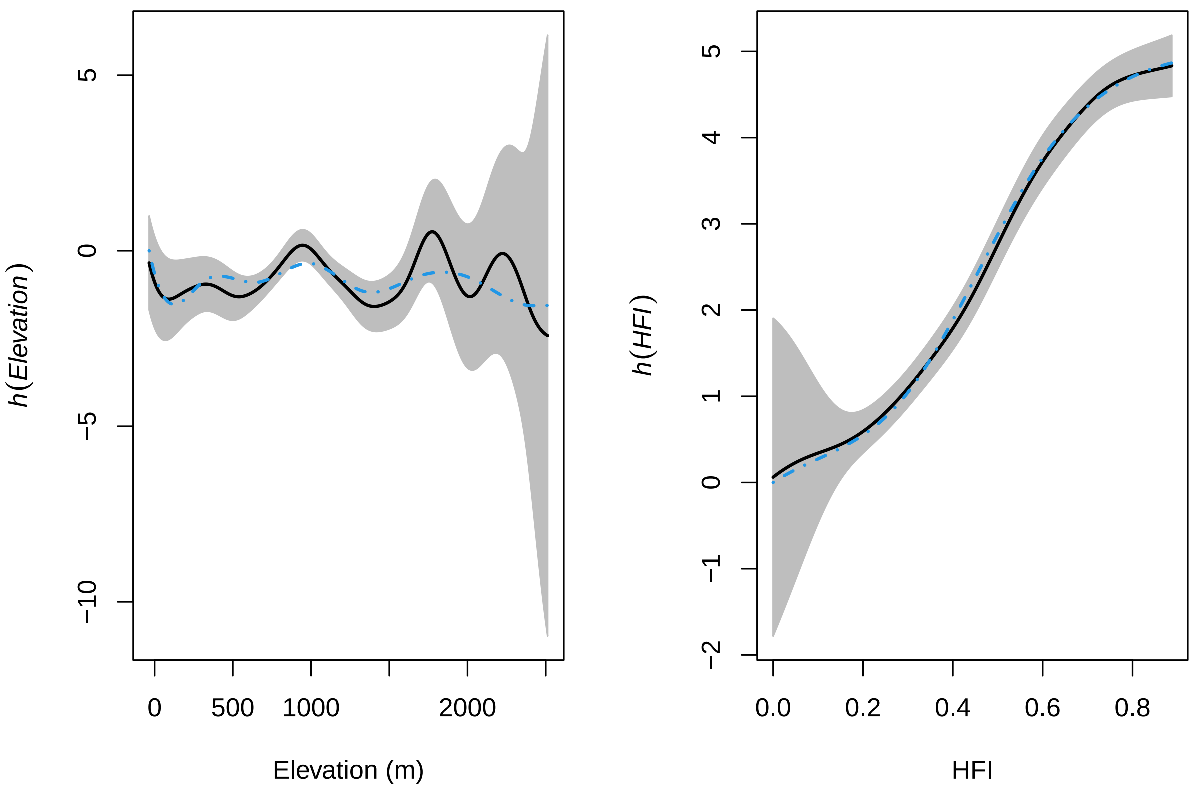


Figure 12: Partial residuals for Model 3 covariates

Additionally, the partial residuals of `Model 3` shown in Figure 12 don't look particularly well. Relatively, `Model 2` has captured the underlying data better and even in the higher `Elevation` area, the output from `Model 3` is still not convincing.

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

```
## Model 2: ~bs(Elevation, 12) + bs(HFI, 5)      Poisson
##   Npar Df Deviance Pr(>Chi)
## 1    5
## 2   18 13   21.436  0.06473 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the LRT test p-value is above a significance level of 0.05, we can conclude that the additional complexity brought on by `Model 3` is not warranted. We conclude `Model 2` as the winner to describe the red fox intensity in British Columbia.

Discussion:

First look at the occurrence data of red fox on BC map showed that the species is spread across the province and look clustered in some places. First moments analysis clearly showed that the intensity of data is inhomogeneous and there are a few hot spots in the province. Second moment analysis using Ripley's K-function showed there is some clustering of the data but the pair correlation which focuses on contribution of all inter-point distances confirmed that there is no significant clustering in data past 23km. Based on this, we conclude there is very little or no clustering in the occurrences of red fox.

Analysis of covariates `HFI`, `Forest Cover`, `Elevation` with this data shows the variables have a non-linear relationship with the data. The fourth covariate available to us is `Distance to Water` and this does not show promising relationship in our initial assessment plots. Individual models with all the covariates is created and studied. `Distance to Water` is not significant and so is excluded from further analysis. We also discovered that there is no significant correlation between the variables and this allows us to combine them in further modeling the given red fox data.

First combined model (`Model 1`) with `Forest Cover`, `Elevation` and `HFI` is fitted with linear and quadratic terms and we discovered that `Forest Cover` is not a significant predictor. So, we removed `Forest Cover` and fitted the next model (`Model 2`) with `Elevation` and `HFI` including linear and quadratic terms. For a good comparison, Model 3 is also fit which is a GAMs model. AIC scores from the three models are tabulated below.

	3 covariates (Model 1)	**2 covariates (Model 2)**	**GAM w 2 covariates (Model 3)**
AIC	10401.62	10403.57	10408.13

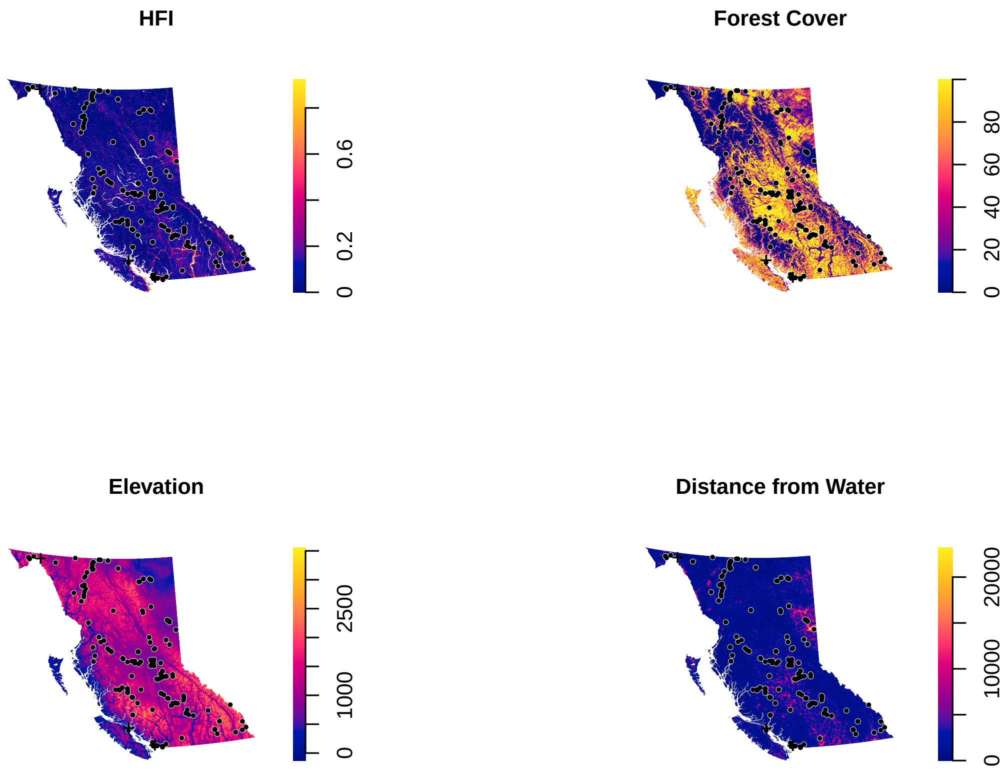
As seen above, `Model 2` has comparatively a lower AIC score and a quadrat test between all three models showed that not any one of the models is superior. We decide to select a parsimonious model and `Model 2` is selected. The residual, partial residual and covariate effect plots support the goodness of fit for this model. They also reveal that there are opportunities for improvements with this model especially at high and low `Elevation` points.

Going back to our research questions, we have found that in this dataset, red foxes are more likely to be observed in areas of higher human activity, as we saw that `HFI` was a significant covariate in all of our models. We have not found that forest cover is significant in modeling the intensity of red foxes in our final model, but the individual effect of forest cover seemed to be significant, suggesting that red foxes were observed more often in places with a forest cover between 10-60%. We found there is significant relationship with `Elevation`, and that there is no evidence of a significant relationship with distance to water.

There are some interesting challenges and insights gathered during the analysis and modeling process and we would like to share to help with future research. The Rho plots and standard residuals plot for this data error out due to NAs in the data. The `Model 2` which is the selected model does not converge if higher order of `Elevation` variables are included. GAMs model is used as a comparison here and only limited tuning is performed in the interest of time. An in depth tuning in our opinion could help build a second model for this data.

Appendix

Plotting red fox occurances with covariates



References: Include references to all necessary literature.

1. Data: https://www.gbif.org/species/5219243, accessed through `rgbif` package
2. Research topics: Canadian Wildlife Federation, https://cwf-fcf.org/en/resources/encyclopedias/fauna/mammals/red-fox.html