Data 589 Project: Research on Hornets

Yuyan Peng, Gevin Yang, Alan Zhang

Introduction

The study system involves the species Dolichovespula maculata, commonly known as the bold-faced hornet. This insect belongs to the order Hymenoptera and is part of the Vespidae family. The Global Biodiversity Information Facility (GBIF) database provides extensive data on occurrences of this species, which is crucial for understanding its distribution and ecological role. The dataset can be found at https://www.gbif.org/species/1311815

Here are the questions we would like to explore with the dataset:

- Question 1: To delineate the population distribution of hornets.
- Question 2: To determine the relationship within hornet occurrence data points.
- Question 3: To identify environmental factors influencing the species' altitudinal distribution in British Columbia, examining variables such as forest elevation, water sources, and the Human Footprint Index.
- Question 4: To construct and calibrate a model evaluating the influence of these variables on hornet distribution patterns.

We will employ a model to establish which factors prominently affect the distribution of hornets, guided by prior observations and literature, including the species' water collection behavior during specific periods as noted by Reichholf (2014), which said that water distribution might have an impact on hornets' distribution.

Our investigation represents a thorough ecological evaluation that employs spatial statistical methods to decode the determinants influencing the distribution of the bold-faced hornet. This analysis endeavors to transcend beyond the intuitive understanding of the species' distribution, delving into the quantitative elements that govern their spatial patterns. Furthermore, we aim to uncover the intrinsic relationships among the data points of hornet occurrences, enhancing our comprehension of their ecological interactions and dispersion mechanisms.

Methodology

Data Introduction

The dataset we obtained contains a detailed compilation of biological records, specifically focusing on hornets, with 6913 entries and 50 variables.

The dataset mainly focuses on three key areas regarding the specimens: identification, location, and classification.

- Identification variables: gbifID, occurrenceID, identifiedBy, dateIdentified, typeStatusor, etc.
- Location variables: countryCode, locality, stateProvince, decimalLatitude and decimalLongitude, coordinateUncertaintyInMeters, elevation, etc.
- Classification variables: kingdom, phylum, class, order, family, genus, species, taxonRank, scientificName, etc.

These categories collectively provide a full spectrum of information essential for scientific analysis, ranging from the precise identification of each specimen to its location and detailed taxonomic classification.

In the analysis we do, we mainly focus on the location variables.

Analytical Workflow Based on Questions

Question 1: To delineate the population distribution of hornets.

Plot Distribution:

Plotting the hornet distribution among BC can give us an intuitive understanding of the distribution.

The ggplot2 package is employed to create a scatter plot visualizing the geographic distribution of hornet observations in British Columbia using longitude and latitude coordinates. To achieve a sleek, modern look, theme_minimal() is applied, which strips away background gridlines and styles the axis lines in black, producing a clean and focused scatter plot of hornet locations based on the dataset.

Projection Process:

The projection process is a critical aspect of handling spatial data because it involves converting the three-dimensional surface of the earth to a two-dimensional plane, which is necessary for any kind of map-making and spatial analysis.

The process begins by loading the sp library, which is essential for manipulating spatial data. Following this, a projection string specific to the BC Albers projection, as described in the project description on Canvas, is defined to set up the appropriate coordinate reference system. As the result, coordinates are compiled into a new dataframe named df_bc_albers, which holds the converted spatial data ready for further analysis or visualization.

Question 2: To determine the relationship within hornet occurrence data points.

Intensity Calculation:

From the plot, we can see most hornets tend to gather in lower BC, showing high intensity, especially at the southwest corner. But for the other part, especially for upper BC, there is no evidence of correlation in hornet location from eyes, indicating the distribution is inhomogeneous.

Quadrat Test:

The quadrat test results, along with the visualization, would provide insights into the spatial dynamics of the hornet distribution, informing if the distribution is homogeneous or inhomogeneous.

In the analysis of hornet spatial distribution, the study area is subdivided using the quadratcount function into a 100-quadrat grid to assess point density variations. The subsequent chi-squared test on this grid reveals a significant deviation from Complete Spatial Randomness, suggesting patterns of clustering or regular spacing among the points. This statistical evidence points towards non-random distribution, which is further elucidated through visualizations that map out the density and distribution anomalies within the hornets' habitat.

Ripley's K-function is simply a function of the area of a circle with radius r. Any deviations between the empirical and theoretical K-functions are thus an indication of correlations between points.

Estimating a strictly positive density for hornet points using kernel density estimation with bandwidth from the plug-in method. Generating an envelope of simulated point patterns to test against the null hypothesis of complete spatial randomness, correcting for edge effects. Running 19 simulations to create the envelope ($\alpha = 0.05$), with each simulation having the same number of points as the original data. Visualizing the simulation envelope to identify the range of distances with significant spatial correlation or dispersion. Zooming in on the envelope plot to focus on areas of significant deviation from the expected CSR pattern.

Question 3: To identify environmental factors influencing the species' altitudinal distribution in British Columbia, examining variables such as forest elevation, water sources, and the Human Footprint Index.

Elevation and Hornet Distribution:

How does hornet distribute among BC? If hornets are homogeneously distributed among BC, the median elevation of hornets should align with the median elevation of BC, let's test that.

To analyze the relationship between elevation and hornet distribution in British Columbia, an elevation map is created using a color gradient to represent varying elevations, with hornet observations overlaid to visualize their spatial distribution relative to elevation. Elevation data is further categorized into quantile-based classes, and hornet observation points are superimposed on this classified map for a clearer view of their distribution across different elevation ranges. Additionally, the median elevations for the entire region and specifically at hornet locations are calculated and compared to discern the preferred elevation range of hornets. A Kernel Density Estimation is also performed to explore the density distribution of elevations and hornet occurrences throughout British Columbia, enhancing understanding of hornet environmental preferences.

Forest and Hornet Distribution:

Same as the previous step when we compare the median elevation of BC and hornets while plotting the density distribution of both, we would like to know if the distribution of hornets has anything to do with the forest distribution in BC.

Homan Footprint Index (HFI)and Hornet Distribution:

Same as the previous steps when we tested the elevation and forest distribution to see if they aligned with the hornet distribution, we would lastly like to know if the distribution of hornets has anything to do with the HFI in BC.

Question 4: To construct and calibrate a model evaluating the influence of these variables on hornet distribution patterns.

Estimating ρ (rho), which represents the strength of spatial correlation, is critical for understanding how the density of hornet locations is associated with environmental variables. The parameter rho reflects the spatial dependence inherent in our sample data, measuring the average influence on observations by their neighboring observations. By estimating rho and plotting it, we should be able to grasp the relationship between hornet distribution and the 4 variables (elevation, forest, water distance, and HFI). We will further build our model based on that.

To analyze how environmental factors influence hornet distribution, spatial correlations between hornet locations and key variables—elevation, forest cover, distance to water, and Human Footprint Index (HFI)—are estimated using the rhohat function. Each correlation is then visualized through dedicated plots. This approach provides a clear view of the relationships, showing whether hornets prefer specific elevations, forest densities, proximity to water, or areas with varying human impact. These insights help pinpoint the environmental preferences of hornets, guiding effective conservation and management strategies.

Fit the model:

After the rho estimation, we would find a suitable model to fit our data. The next step is to test the model and refit until all the variables have a good result by the Z-test.

To analyze hornet distribution, a nonstationary Poisson process model is used with elevation, forest cover, and distance to water as covariates, including their squared terms for nonlinear relationships. Following statistical evaluation, ineffective squared terms are removed based on z-test outcomes, and the model is refitted without them. The refined model's predictions are then plotted, allowing for visual inspection to assess how environmental factors influence hornet distribution accurately. This iterative process enhances model reliability and interpretability.

Model selection:

Here, we use AIC and BIC to do model selection. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are being used to compare the two models. These criteria evaluate the models based on their goodness of fit to the data while penalizing for complexity (the number of parameters used).

In the model selection process, we calculate the AIC and BIC values for both the originally fitted model and the refitted model.

Model validation:

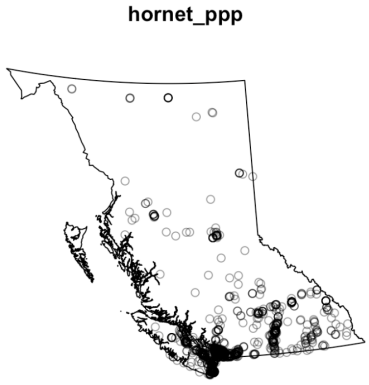
In the model validation process, we would like to know how well our model fits, and if the model does not fit well, how should we improve the model.

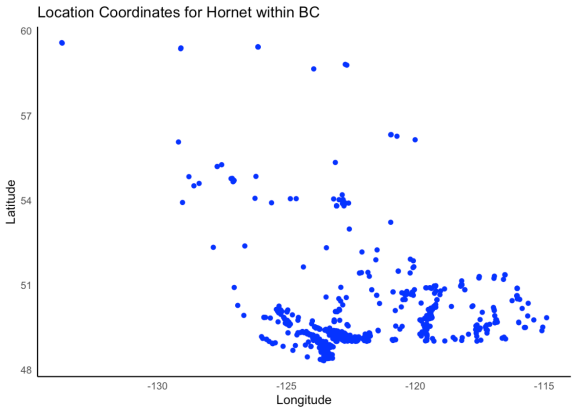
To validate a model predicting hornet distribution, a series of checks are conducted. First, a quadrat test assesses significant deviations from the model's predictions. Residuals for the key covariates—elevation, forest cover, and distance to water—are then plotted to evaluate fit. The model is refined by incorporating splines and re-fitting as a Generalized Additive Model (GAM), with new residuals plotted to check improvements. Finally, the revised model is compared to the original using AIC/BIC metrics to quantify enhancements, ensuring the model's accuracy and robustness in predicting hornet distributions.

Results

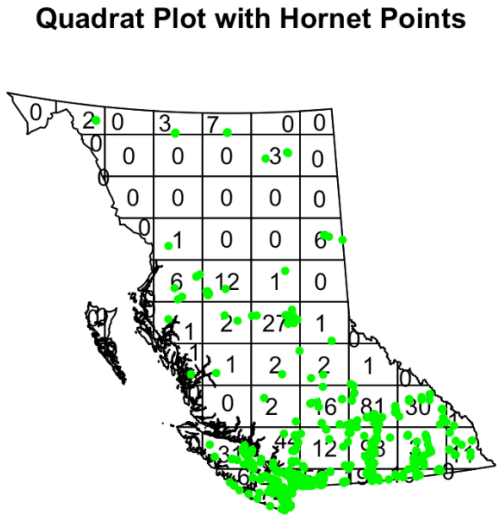
First Moment Descriptive Statistics:

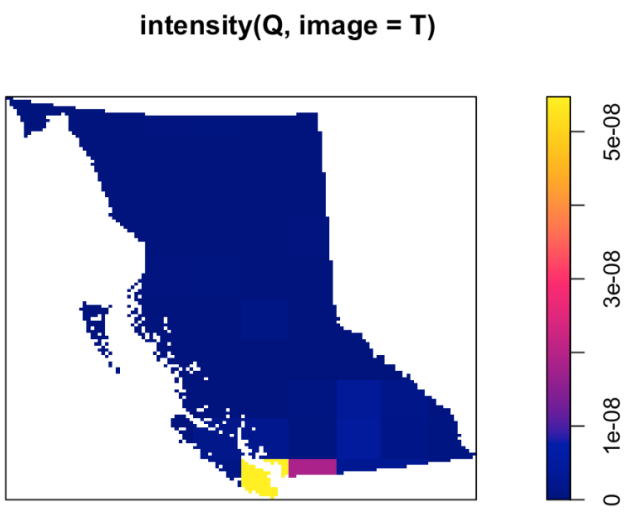
The analysis using spatial plotting and projection processes revealed a clear geographic distribution of Dolichovespula maculata within British Columbia. Using ggplot2 and sp, we visualized the hornet's occurrences, which are predominantly concentrated in the southwest corner of the region. The transformation to BC Albers projection further clarified these patterns.



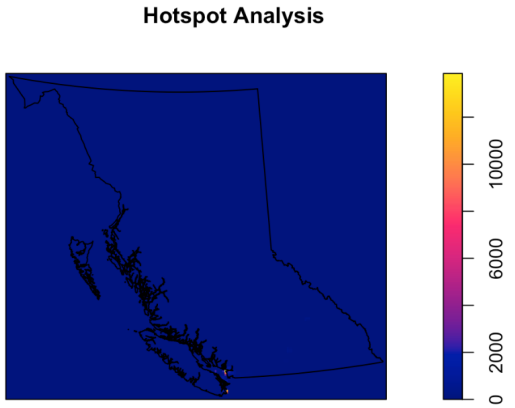


We applied a quadrat test to evaluate the homogeneity of hornet distribution. Dividing the study area into 100 quadrats, we found a significant deviation from Complete Spatial Randomness (CSR), with a chi-squared test yielding a p-value < 0.01. This result indicates that the distribution of hornets is not uniform, suggesting underlying environmental or biological processes influencing their distribution.



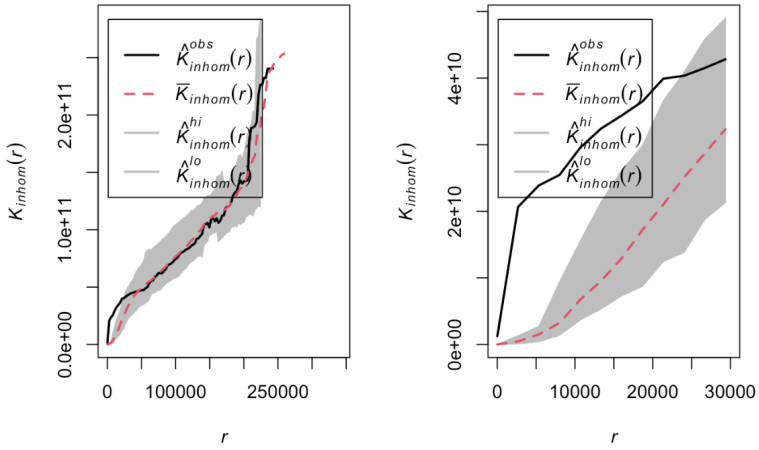


The hotspot analysis plot intensifies this interpretation by assigning a color gradient scale to the density estimations, with warm colors representing higher concentrations of points. This analysis identifies the most significant clusters of hornet activity, quantified by the KDE values on the color scale. The areas with the highest intensities, indicated by red and yellow hues, correspond to locations where conservation efforts or further ecological studies might be most needed or where the hornets have optimal conditions for their survival and proliferation.

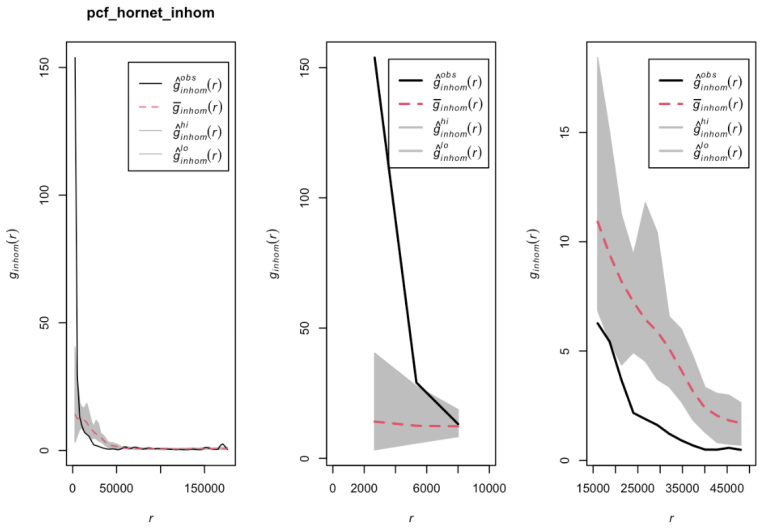


Second Moment Descriptive Statistics:

The Ripley's K-function analysis evaluated spatial correlation between hornet locations to discern patterns beyond what could be attributed to random chance. By comparing the observed K-function with the expected K-function under inhomogeneity, significant insights into spatial patterns were revealed. A strictly positive kernel density was employed to simulate an inhomogeneous Poisson process, which accounts for varying intensity of occurrences across the study area. The simulations generated, under the null hypothesis of spatial randomness with inhomogeneity, produced an envelope capturing the range of expected K-function values. This envelope is essential in identifying distances at which the actual spatial pattern deviates significantly from the model of inhomogeneity. Upon examination of the observed K-function against the simulation envelope, significant clustering of hornet occurrences was detected at smaller spatial scales, up to around 22,000 meters. The observed K-function surpassed the upper confidence bounds of the envelope in this range, indicating a greater level of aggregation among hornets than would be expected under the null hypothesis. Such clustering could suggest underlying ecological processes or behaviors that result in hornets being found in closer proximity to each other than would occur by chance alone. For larger distances beyond 22,000 meters, the observed K-function fell within the bounds of the simulation envelope, suggesting that at these scales, the distribution of hornets does not show significant spatial correlation. This implies that environmental heterogeneity could be the driving factor in the distribution of hornets across the broader landscape of British Columbia, as opposed to localized interactions between individuals.

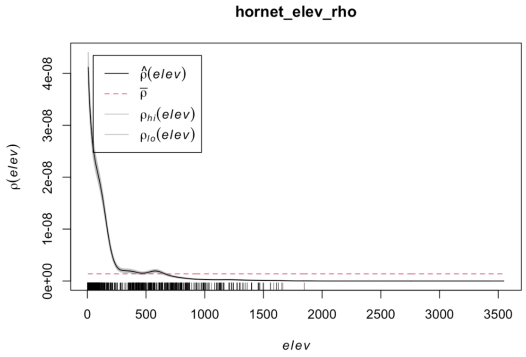


The Pair Correlation Function (PCF) analysis serves as a tool to understand the fine-scale spatial relationship among hornet locations, providing a more nuanced understanding than the cumulative measure offered by Ripley's K-function. Through this analysis, the degree of spatial clustering or dispersion at various distances is assessed. For this study, the PCF was calculated for the hornet point pattern, with the intent to reveal any non-random interactions between points at different scales. The envelope surrounding the PCF represents the 95% confidence interval obtained from 19 simulations of an inhomogeneous Poisson process, serving as a reference for detecting significant deviations from randomness. The results demonstrated notable deviations from the expected values in two key distance ranges. Firstly, there was a higher-than-expected number of pairs of points at distances up to approximately 6,000 meters, indicating a stronger local clustering than would be predicted by the inhomogeneous Poisson process. This observation suggests that hornets tend to form localized aggregations at these scales, possibly due to environmental preferences or social behaviors that prompt them to be near one another. Conversely, in the distance range between 15,000 to 50,000 meters, the observed PCF values fell below the expected range, indicating less clustering than predicted by the model. This lower level of correlation suggests that, at these intermediate scales, the distribution of hornet occurrences does not exhibit the clustering that might be expected given their densities at smaller distances. The PCF analysis thus provides insights into the spatial structure of hornet locations, revealing significant small-scale clustering, while at larger distances, interactions appear to be more random or even dispersed than expected. This dual pattern underscores the complexity of the spatial dynamics governing hornet distributions and may reflect varying influences of ecological processes at different scales.

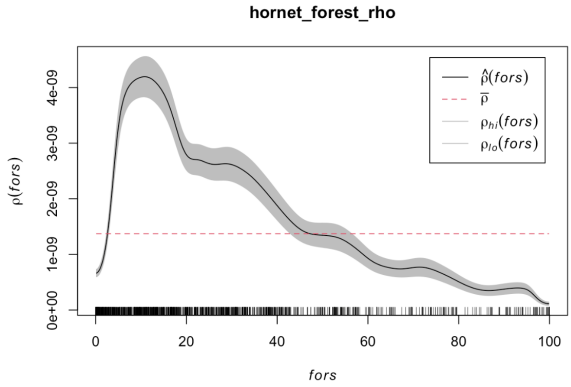


The Poisson Point Process Model Fitting

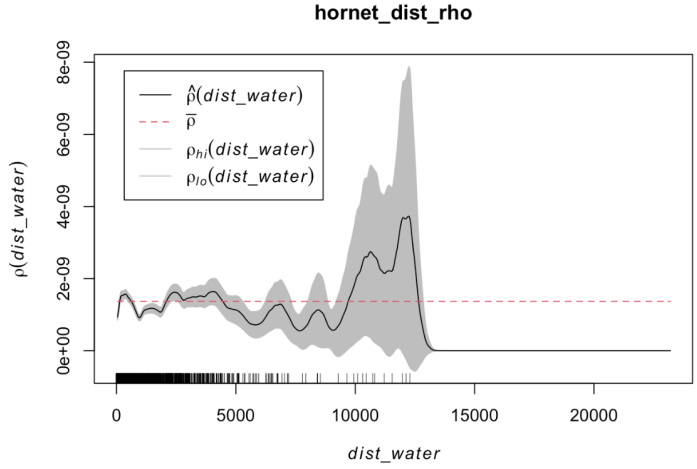
The Poisson point process model fitting section investigates the relationship between the occurrence of hornets and various environmental indicators: elevation, forest density, distance to water sources, and the Human Footprint Index (HFI). The ρ (rho) values, estimated using the rhohat function, quantify the strength and nature of spatial correlation between hornet occurrences and each environmental variable. This involves fitting a spatial model to the data, and ρ represents how each environmental factor influences hornet distribution after accounting for all other factors.



The “hornet_elev_rho” plot examines the correlation between hornet occurrences and elevation. The solid line represents the estimated ρ for elevation, which remains relatively stable across the range of elevations. The dashed red line indicates the overall mean of ρ, providing a reference for comparison. This plot suggests that while elevation may play a role in hornet distribution, its influence does not strongly fluctuate across different elevations.



For forest density, the "hornet_forest_rho" plot reveals a more variable relationship with hornet occurrences. The initial peak in ρ values at lower forest densities suggests a higher correlation in these regions, which could indicate a preference for habitats with certain forest density characteristics. As forest density values increase, ρ decreases, implying that the strength of the correlation diminishes, potentially reflecting an upper threshold of forest density beyond which additional density does not further influence hornet distribution.



The "hornet_dist_rho" plot, focusing on distance to water sources, shows a pronounced peak in the correlation strength at lower distances, followed by a sharp decline. This indicates a strong preference for areas closer to water, which is consistent with the needs of hornets for water in their diet and nest-building activities. As the distance to water increases, the correlation strength significantly drops, underscoring the importance of proximity to water for hornet distribution. After evaluating the influence of squared terms of these indicators, a Z-test was conducted to determine their significance in the model. The tests suggested that the squared terms for forest density and distance to water were not improving the model and were removed in the refined model.

We considered the model shown below first:

$$\lambda(\mu) = e^{\beta_0+\beta_1 ele_u+\beta_2 water_u+\beta_3 HFi_u+\beta_4 forest_u}$$

But we found that variable of forest is not significant with z-value equal to 0.304, which indicating removing it is acceptable. So, the model without this feature is shown like below:

$$\lambda(\mu) = e^{\beta_0+\beta_1 ele_u+\beta_2 water_u+\beta_3 HFi_u}$$

But from the rhohat graphs shown above non-linear relationship should be considered for HFI and elevations:

$$\lambda(\mu) = e^{\beta_0+\beta_1 ele_u+\beta_2 water_u+\beta_3 HFi_u+\beta_4 elev^2+\beta_4 HFI^2}$$

Model Selection

We employed both AIC and BIC criteria to select the optimal model. The model incorporating elevation, forest density, and proximity to water as predictors provided the best fit to the data, balancing complexity, and explanatory power.

```
AIC(fit_2);AIC(fit_1)
```

```
## [1] 47779.17
```

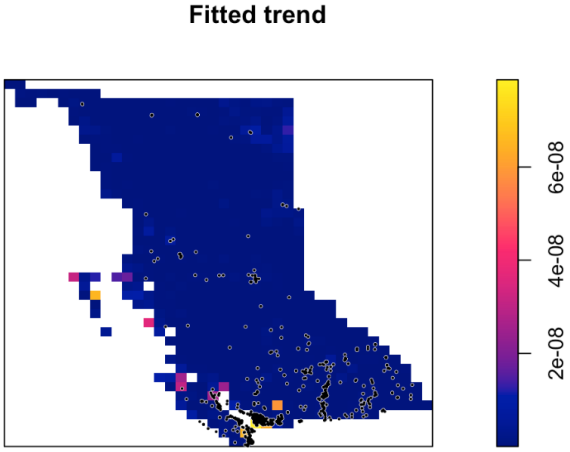
```
## [1] 47850.98
```

```
BIC(fit_2); BIC(fit_1)
```

```
## [1] 47810.2
```

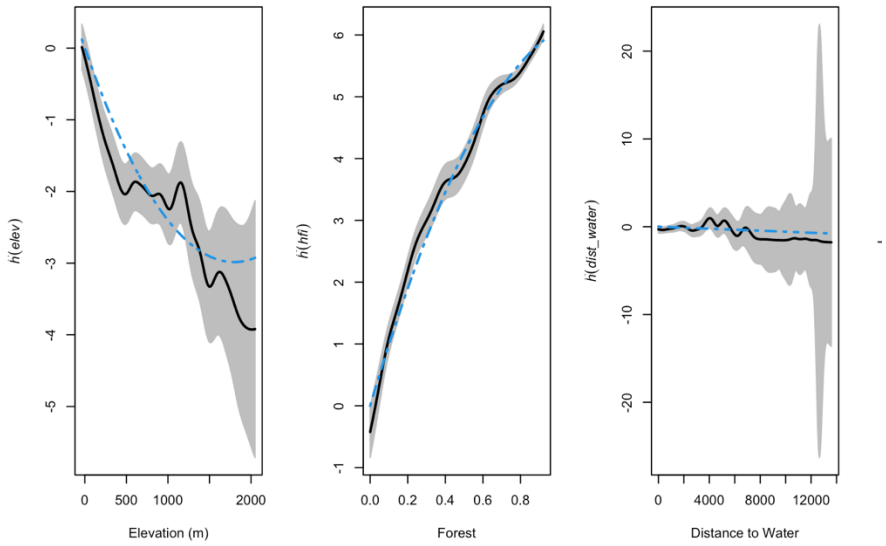
```
## [1] 47871.66
```

A visualization of the fitted model's predictions superimposed on the observed data points allowed for a direct comparison between predicted and actual occurrences, showing how well the model captured the underlying spatial trends of hornet distribution. This fitted trend map is valuable for visualizing the potential "hotspots" for hornet presence based on the environmental factors considered, which could guide field surveys and conservation strategies.



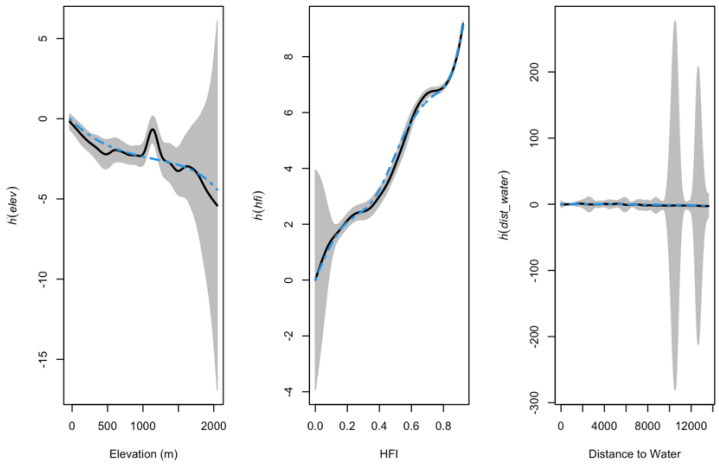
Model Validation

To get a better feel for how we should specify our elevation and water and forest effects, we can use partial residual plots, which show the fitted effect of a covariate alongside the observed effect.



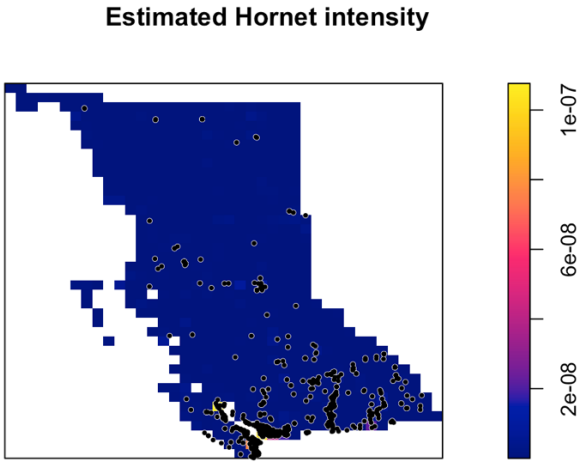
From these figures we can see that the quadratic terms are not capturing the patterns in our data particularly well. As an improvement, we could try adding higher-order polynomials, but polynomials can be unstable. In this situation, it may be worth switching from a linear modelling framework to an additive modelling framework.

After adjusting the parameters inside the model, we specify basis expansions for predictor variables such as elevation (4 basis functions), habitat fragmentation index (7 basis functions), and distance to water (2 basis functions), allowing for capturing potentially non-linear relationships between these variables and the response in spline-based models.



It is not well fitted than our expectation, but better than the previous one. To ensure we are not overfitting, we can again use our model selection techniques (AIC, BIC, ANOVA).

The ANOVA results indicate that Model 2 has significantly better fit to the data than Model 1, as evidenced by the large difference in deviance (516.81) and the highly significant p-value (< 2.2e-16), which is well below any conventional significance level. All lines of evidence point towards these more complex models being a better fit to the data.



Discussion

The spatial distribution of hornets in British Columbia is a complex interplay of various environmental factors, and our study aimed to untangle these relationships using robust statistical modeling. Our investigation into the spatial distribution of hornets in British Columbia provides valuable insights into the interplay between species presence and environmental factors. Through rigorous statistical modeling, we determined that elevation, while a factor, did not exhibit a strong selective preference within the range of hornet occurrences. This implies that conservation efforts focused on elevation might have less impact than previously considered. The analysis of forest density and proximity to water revealed pronounced preferences, with hornets favoring areas with moderate forest coverage and close to water sources. These findings highlight the ecological requirements of hornets for nesting and foraging, suggesting that conservation efforts should prioritize the preservation of these specific habitat features. The relationship with water sources is particularly striking, reinforcing the need for careful management of aquatic and riparian ecosystems to support hornet populations. The Human Footprint Index (HFI) analysis adds a dimension to our understanding of how human activity influences hornet distribution, with a clear tendency for hornets to avoid areas of high human impact. This avoidance behavior points towards the need for habitat protection in less disturbed areas, which could offer the resources and safety required by hornets. Our modeling efforts, particularly the Poisson point process model, quantified the correlations between hornets and these environmental variables, affirming their significance in predicting hornet locations. By refining the model and validating it against observed data, we established a reliable tool for predicting hornet occurrences. This predictive capacity is crucial for effective habitat management and conservation, as it allows for the anticipation of hornet distribution patterns in response to environmental changes. The model's predictive visualizations, illustrating potential "hotspots" for hornet activity, are instrumental for future field research and habitat conservation. By identifying areas with optimal

conditions for hornets, our research provides a foundation for targeted conservation actions, ensuring these ecologically important species continue to thrive in their natural habitats.

Reference

Josef H. Reichholf (2014): Wozu benutzen Hornissen Vespa crabro das Wasser, das sie im Frühsommer eintragen? – Mitteilungen der Zoologischen Gesellschaft Braunau – 11: 285 - 288