Using Citizen Science Data to Predict Northern Goshawk (*Accipiter gentilis*) Distribution Based on Environmental, Climate, and Geological Data in the State of Oregon

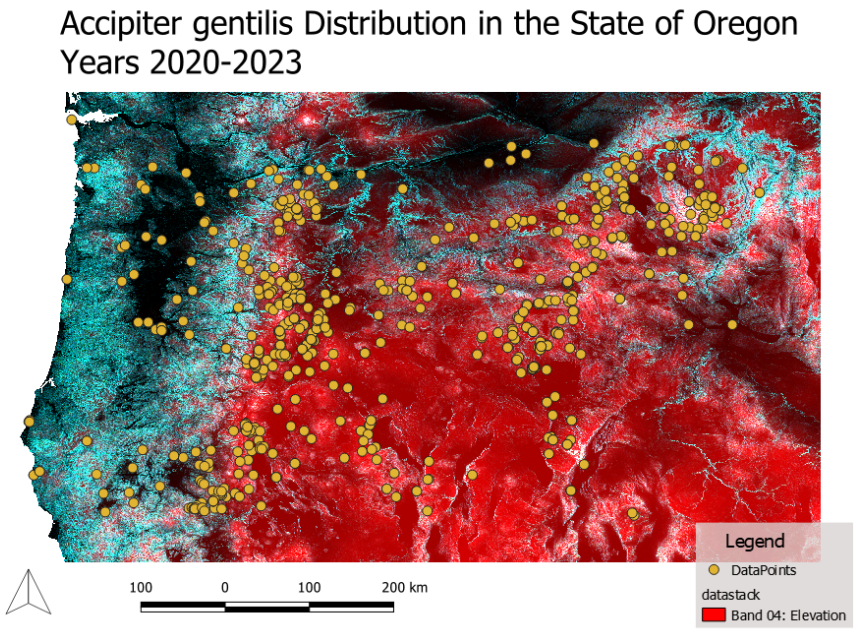


Fig. 1. State of Oregon with E-bird Citizen Presence data for *Accipiter gentilis*

My response variable is the presence and pseudo-absence of (*Accipiter gentilis*) in the state of Oregon based on collected citizen data. The data source is the E-bird database accessible through the GBIF | Global Biodiversity Information Facility data portal. The years of collection for my presence data are 2020-2023 in order to increase the number of observations.

Data Selection and Retrieval:

1. E-bird Data: Bird data is a citizen science initiative that collects birdwatching observations from participants, providing valuable insights into bird distribution, migration patterns, and population trends for research and conservation efforts.

GBIF.org (17 October 2024) GBIF Occurrence Download https://doi.org/10.15468/dl.px3sqc

2. Elevation - The height above sea level, influencing climate and vegetation types.

U.S. Geological Survey. (n.d.). National Elevation Dataset. Retrieved from USGS NED website.

3. Aspect - The compass direction a slope faces, affecting sunlight exposure and microclimates.

Derived from Elevation through functions included in the terra package

4. Roughness - A measure of landscape variability that impacts water runoff and habitat complexity.

Derived from Elevation through functions included in the terra package

5. Slope - The steepness of a surface, influencing erosion, drainage, and vegetation patterns.

Derived from Elevation through functions included in the terra package

6. Precipitation - All forms of water falling from the atmosphere, crucial for ecosystem health and water availability.

PRISM Climate Group. *PRISM Climate Data*. Oregon State University, 2024. Web. https://prism.oregonstate.edu.

7. Mean Temperature of the Dew Point (TDMean) - The average temperature at which air becomes saturated with moisture, indicating humidity levels.

PRISM Climate Group. *PRISM Climate Data*. Oregon State University, 2024. Web. https://prism.oregonstate.edu.

8. Mean Temperature - The average temperature over a period, affecting species distribution and ecosystem processes.

PRISM Climate Group. *PRISM Climate Data*. Oregon State University, 2024. Web. https://prism.oregonstate.edu.

9. Maximum Temperature - The highest recorded temperature in a period.

PRISM Climate Group. *PRISM Climate Data*. Oregon State University, 2024. Web. https://prism.oregonstate.edu.

10. Minimum Temperature - The lowest recorded temperature in a period.

PRISM Climate Group. *PRISM Climate Data*. Oregon State University, 2024. Web. https://prism.oregonstate.edu.

11. Land Cover - The physical material covering the Earth's surface, vital for biodiversity and ecosystem services.

Homer, C. G., et al. *Completion of the 2021 National Land Cover Database for the Conterminous United States—Representing a Decade of Land Cover Change Information*. U.S. Geological Survey, 2021. Web. https://www.mrlc.gov/nlcd-2021.

12. Canopy Cover - The percentage of ground covered by the upper layer of vegetation, affecting light and habitat conditions.

U.S. Forest Service. *National Land Cover Database (NLCD) 2021 Canopy Cover Data*. 2021. Web. https://www.mrlc.gov/nlcd-2021.

13. Population by Census - Data on human populations that inform land use and environmental impact assessments.

U.S. Census Bureau. *2020 Census: Demographic and Housing Characteristics*. 2021. Web. https://www.census.gov/programs-surveys/decennial-census/2020.html.

Data Processing:

E-Bird Data: The occurrence data from E-BIRD was downloaded using a specific occurrence download ID. After importing the data, its structure and summary statistics were examined to understand the dataset better. Relevant columns were selected, focusing on occurrence ID, species, location, and individual counts. The dataset was then filtered to include records from the years 2020 to 2023, ensuring only recent data was analyzed. Finally, the data was converted into a spatial format using the sf package and subsequently transformed into a SpatVector.

Elevation Data: Latitude and longitude ranges for the area of interest were defined to facilitate the downloading of elevation data. The elevation data was retrieved for each coordinate combination, and the results were merged into a single dataset. This elevation data was then visualized through plotting and projected to match the coordinate reference system of the spatial data, ensuring consistency across datasets.

Terrain Analysis Data: Terrain metrics, such as aspect, roughness, and slope, were calculated from the elevation data to enhance the understanding of the landscape. Each of these metrics was saved as a raster file, and visualizations were created to showcase the terrain characteristics of the study area.

Climate Data: Specific climate variables of interest were defined for analysis- precipitation, max temperature, minimum temperature, mean temperature, dew point. A loop was established to process data for each year from 2020 to 2023. During this phase, climate raster data was read and cropped to fit the bounding box of the study area, allowing for a focused analysis. Mean values for each climate variable were calculated to create a comprehensive stacked raster representation.

Land Cover Data: NLCD land cover data for 2021 was loaded and projected to the appropriate coordinate reference system. This data provided valuable insights into the land use and cover types in the region, which could influence ecological analyses.

Canopy Cover Data: was sourced from the National Land Cover Database (NLCD), providing critical insights into tree canopy percentages across the study area. After downloading, the raster was projected to match the coordinate reference system of the elevation data, ensuring compatibility with other spatial datasets.

Census Data Integration: The analysis included demographic data accessed through the census API. The population variable was downloaded and converted into raster format to integrate demographic information with the spatial datasets, facilitating a comprehensive examination of population characteristics in relation to other environmental factors.

Data Preparation for Modeling: All raster datasets were cropped to a defined bounding box to ensure uniformity in spatial extent. Resampling was performed to align the extents and resolutions of different rasters (92.8 meters). Ultimately, all processed raster layers were combined into a final data stack, setting the stage for further analysis.

Point Data Preparation: A response variable was prepared to indicate presence or absence based on individual counts in the E-Bird dataset. To enhance the dataset, pseudo-absence points were generated by sampling randomly within the study area. The raster values at both presence and pseudo-absence points were extracted and combined to create a comprehensive dataset ready for modeling.

Statistical Analysis: The extracted data was assessed for completeness and potential duplicates. Statistical analyses, including cumulative density function analysis and the Kolmogorov-Smirnov test, were conducted to evaluate spatial bias, helping to identify any inconsistencies in the dataset.

Spatial Bias Mitigation: To address potential bias, a grid raster was created to represent stratified sampling zones. Within each grid cell, one random point was selected, ensuring representative

sampling across the study area. This strategy was revisited, and the data was reevaluated with the selected points, allowing for a thorough assessment of bias mitigation.

Final Dataset Creation: The final phase involved extracting raster values for the selected points and compiling these into a cohesive dataset. This dataset, which included presence and pseudo-absence data along with environmental variables, was prepared for subsequent analyses, ensuring a robust foundation for understanding ecological patterns.

Model Creation and Detail:

In the following sections, I will be providing some results from the exploratory data analysis for this model, model optimization and testing results, and visualizations for the final model.

This project is a supervised classification model, with the two defined classes being presence and absence (1 and 0 respectively)

This project is pixel-based as the data are raster layers, and the final visualization is a rasterized prediction matrix.

The final resolution of the model is 92.8 meters.

EDA

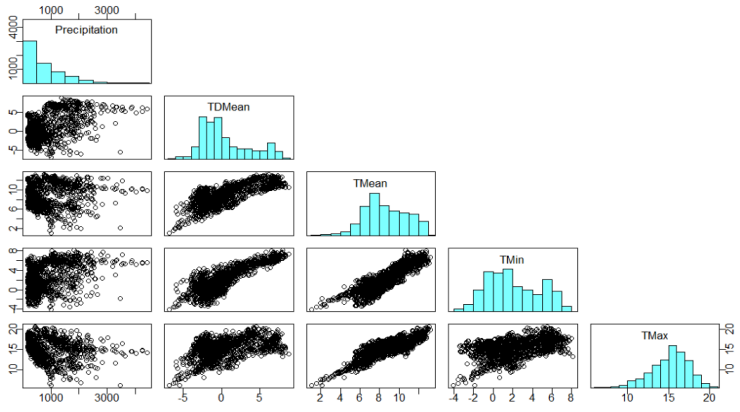


Fig. 2.

Boxplots:

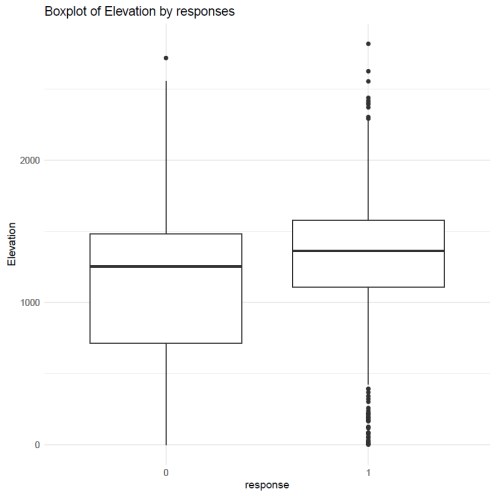


Fig 3. The distribution of elevation across both responses is significantly varied.

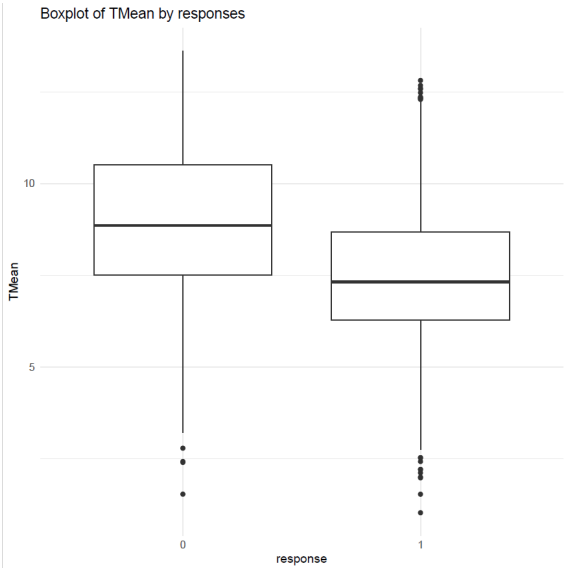


Fig 3.1. The distribution of Mean Temperature readings across both responses is significantly varied.

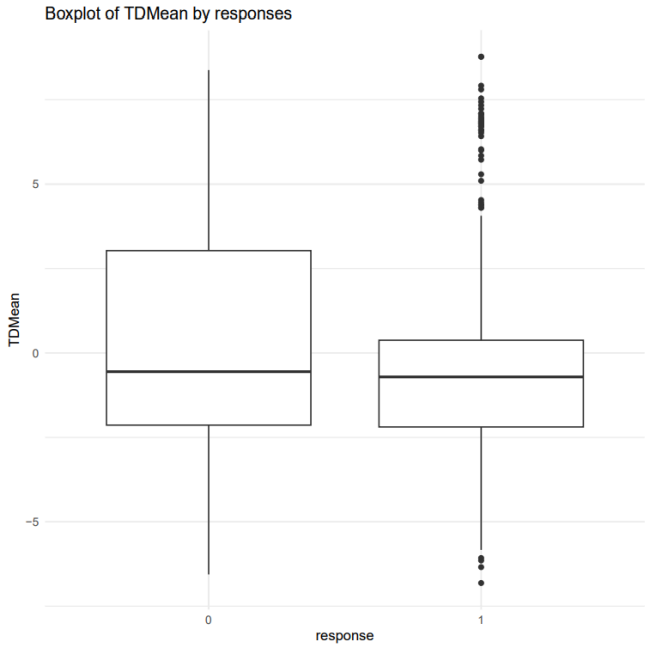
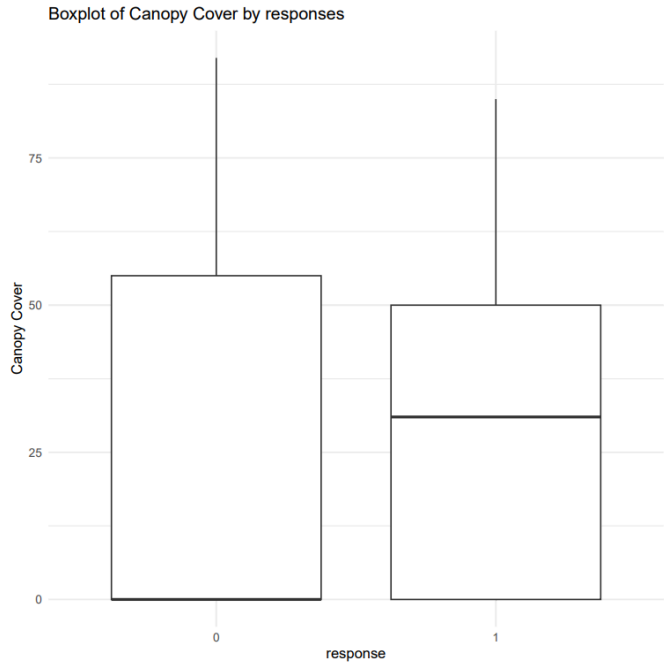


Fig 3.2. The distribution of TDMean across both responses is significantly varied.



3.3. The distribution of Canopy Cover across both responses is significantly varied.

Correlation Matrix:

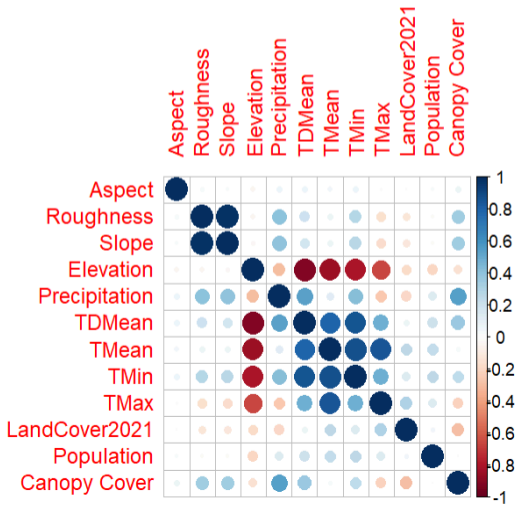


Fig 4. Correlation Matrix

Of significance:

(-) Correlation between Temperature Data and Elevation

(+) Correlation between Precipitation and Canopy Cover

(+) Correlation between Roughness and Slope

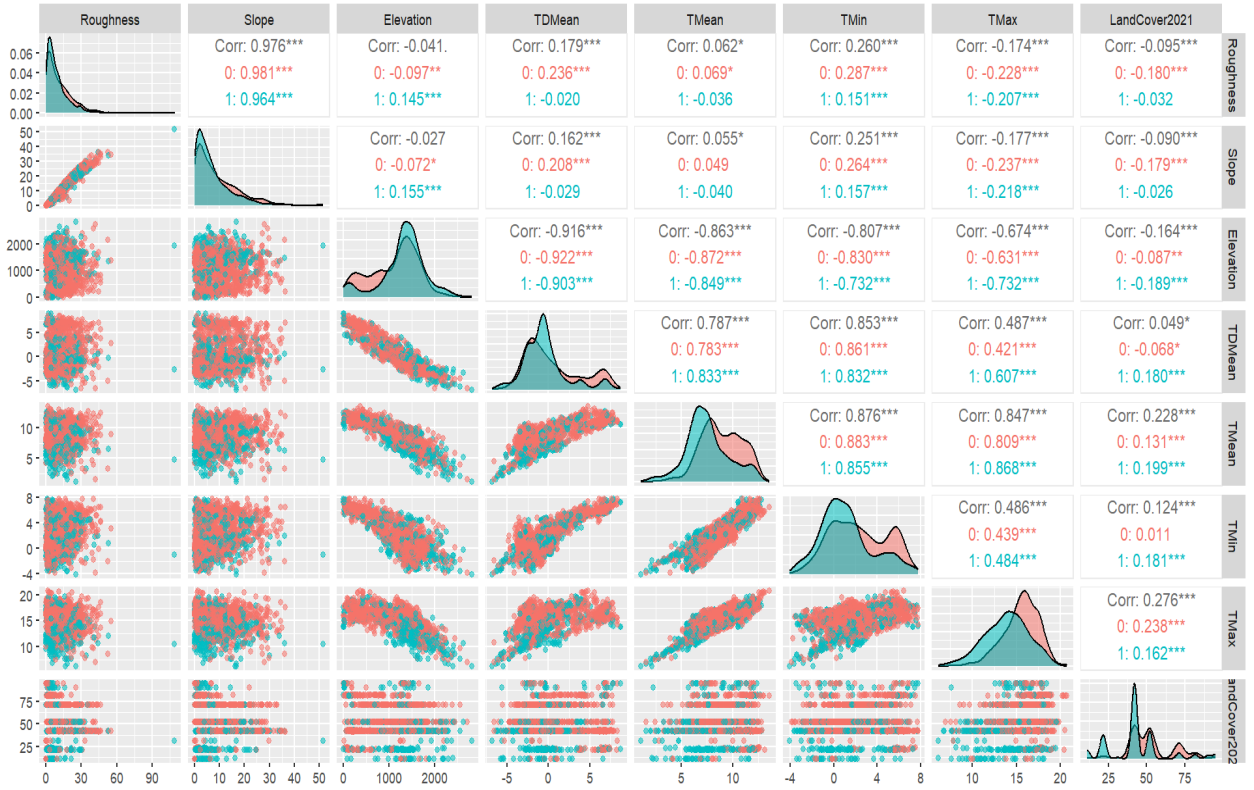


Fig 5. Correlation between response and variables

Summary of Welch Two Sample t-test Results

- **Aspect**:
  - **p-value**: 0.1107
  - **Means**: Group 0: 174.98 | Group 1: 166.30
  - **Significance**: Not significant
- **Roughness**:
  - **p-value**: < 0.0001
  - **Means**: Group 0: 11.41 | Group 1: 9.26
  - **Significance**: Significant
- **Slope**:
  - **p-value**: < 0.0001
  - **Means**: Group 0: 8.64 | Group 1: 6.84
  - **Significance**: Significant
- **Elevation**:
  - **p-value**: < 2.2e-16
  - **Means**: Group 0: 1074.17 | Group 1: 1309.50
  - **Significance**: Significant
- **Precipitation**:
  - **p-value**: 0.9833
  - **Means**: Group 0: 814.00 | Group 1: 813.31
  - **Significance**: Not significant
- **TDMean**:
  - **p-value**: < 0.0001
  - **Means**: Group 0: 0.80 | Group 1: -0.41
  - **Significance**: Significant
- **TMean**:
  - **p-value**: < 2.2e-16
  - **Means**: Group 0: 9.05 | Group 1: 7.51
  - **Significance**: Significant
- **TMin**:
  - **p-value**: < 2.2e-16

    - **Means**: Group 0: 2.49 | Group 1: 1.19
    - **Significance**: Significant
- **TMax**:
    - **p-value**: < 2.2e-16
    - **Means**: Group 0: 15.60 | Group 1: 13.84
    - **Significance**: Significant
- **LandCover2021**:
    - **p-value**: < 2.2e-16
    - **Means**: Group 0: 53.81 | Group 1: 43.43
    - **Significance**: Significant
- **Canopy Cover**:
    - **p-value**: 0.04138
    - **Means**: Group 0: 25.61 | Group 1: 28.61
    - **Significance**: Significant
- **Population**:
    - **p-value**: 0.04498
    - **Means**: Group 0: 3271.99 | Group 1: 3134.37
    - **Significance**: Significant

Based on the exploratory data analysis the following are variables that differ significantly between the presence and absence groups: Roughness, Slope, Elevation, TDMean, TMean, TMin, TMax, LandCover2021, and Canopy Cover show significant differences between the two groups defined by the response variable.

The most redundancy is present in the temperature data. Land cover and canopy cover also have some redundancy, since higher canopy cover indicates a specific value on the land cover raster as well (forest). In the final model, I plan to use most of the data regardless of redundancy, but for further exploratory analysis I will only use TMean and Land Cover, since they are the most comprehensive of the data that possesses some redundancy.

The three variables of most significance are Elevation, Mean Temperature, and Land Cover.

Cluster analysis:

K-means Cluster Analysis Elbow Method

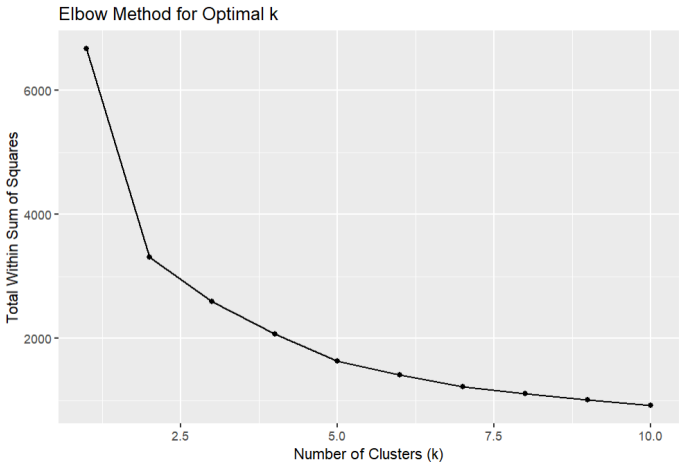


Fig 6. Elbow Method Graph to determine optimal number of clusters- elbow present at 3 and 5. K means analysis was performed for both 3 and 5 clusters.

K-means for 3 clusters:

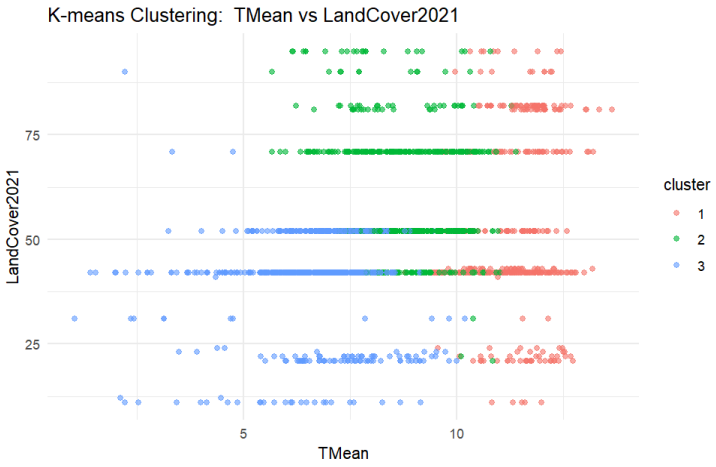


Fig 7. K-means (3 cluster) analysis for TMean vs. LandCover2021

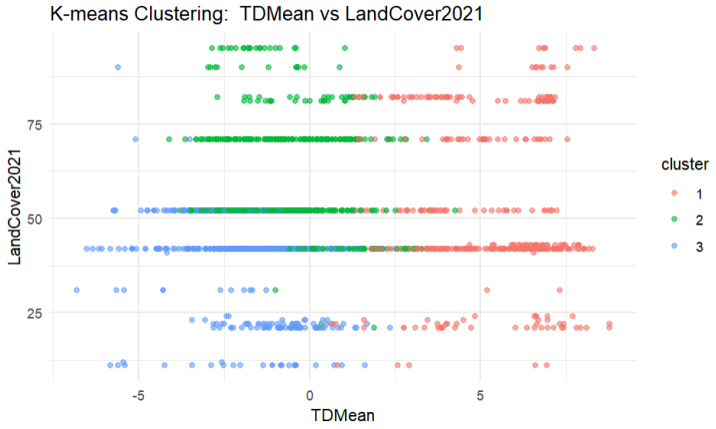


Fig 7.1. K-means (3 cluster) analysis for TDMean vs. LandCover2021

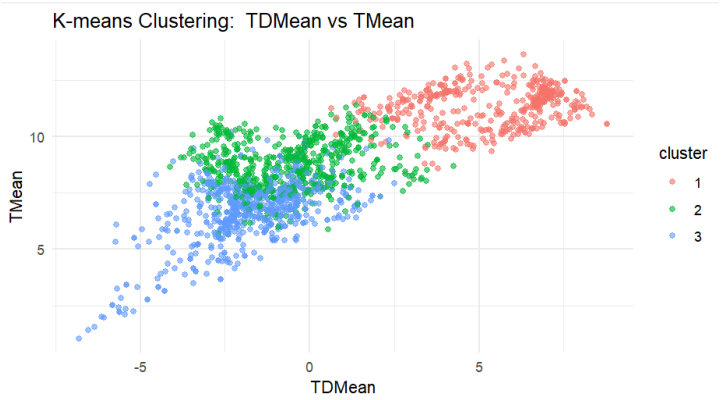


Fig 7.2. K-means (3 cluster) analysis for TDMean vs. TMean

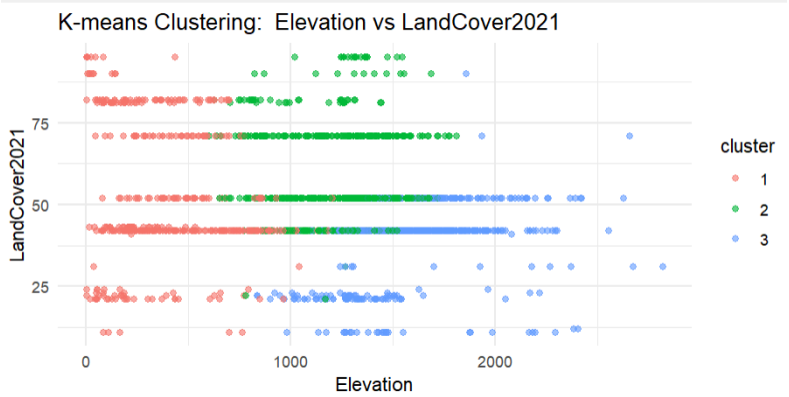


Fig 7.3. K-means (3 cluster) analysis for Elevation vs. LandCover2021

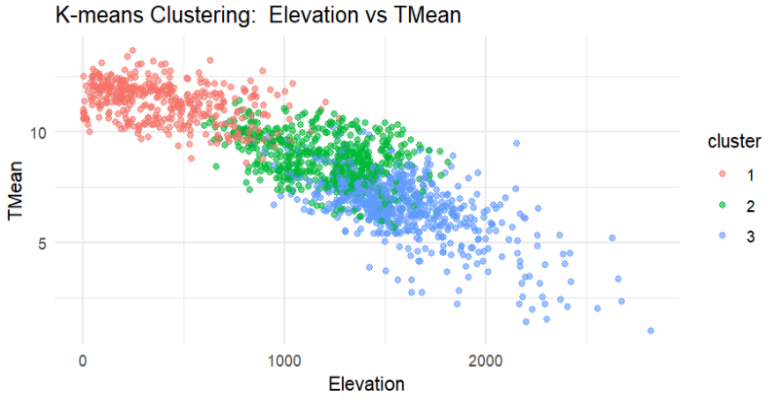


Fig 7.4. K-means (3 cluster) analysis for Elevation vs. TMean

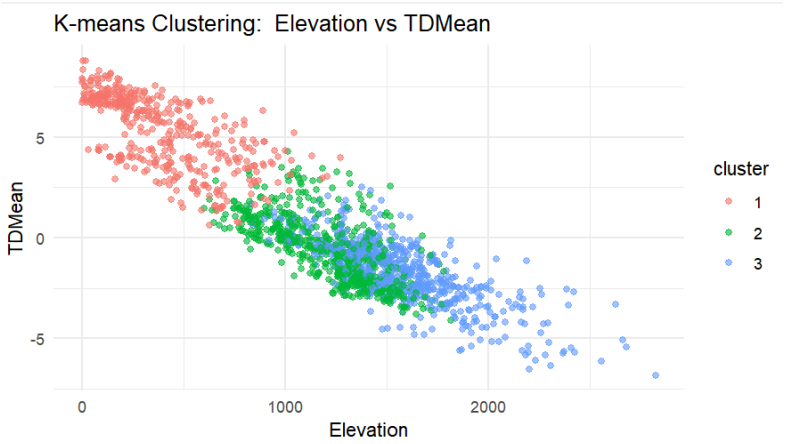


Fig 7.5. K-means (3 cluster) analysis for TMean vs. LandCover2021

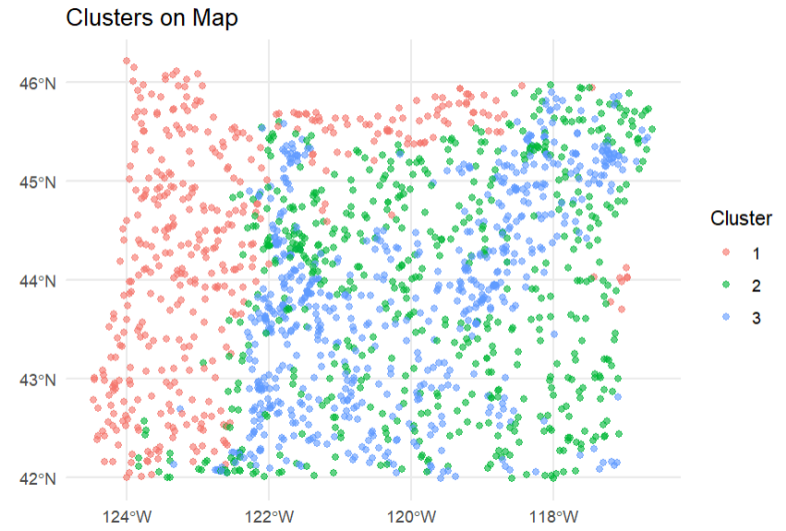


Fig 7.6. Spatial representation of the K-means Cluster analysis for 3 clusters.

```
cluster Elevation TDMean TMean LandCover2021
  <fct>       <dbl>  <dbl> <dbl>         <dbl>
1 1            395.  5.26  11.3           53.0
2 2           1213. -0.561  8.73          59.7
3 3           1560. -1.74   6.69          40.9
```

K-means for 5 clusters:

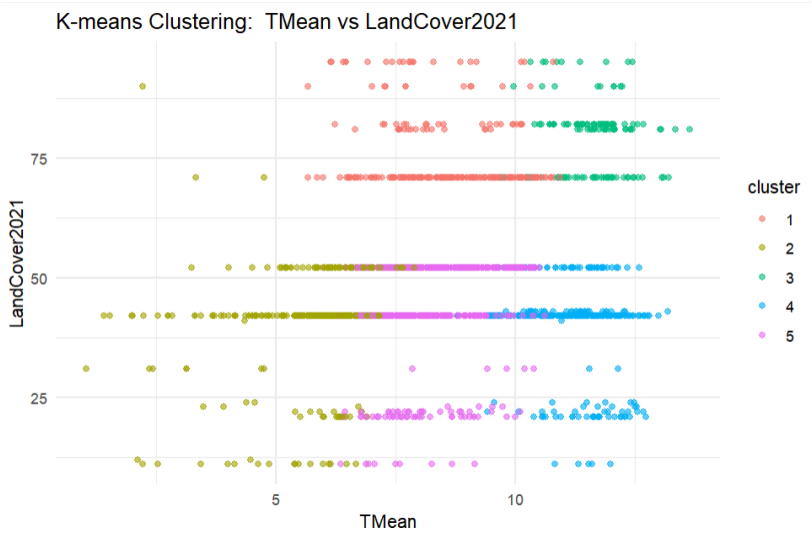


Fig 8. K-means (5 cluster) analysis for TMean vs. LandCover2021

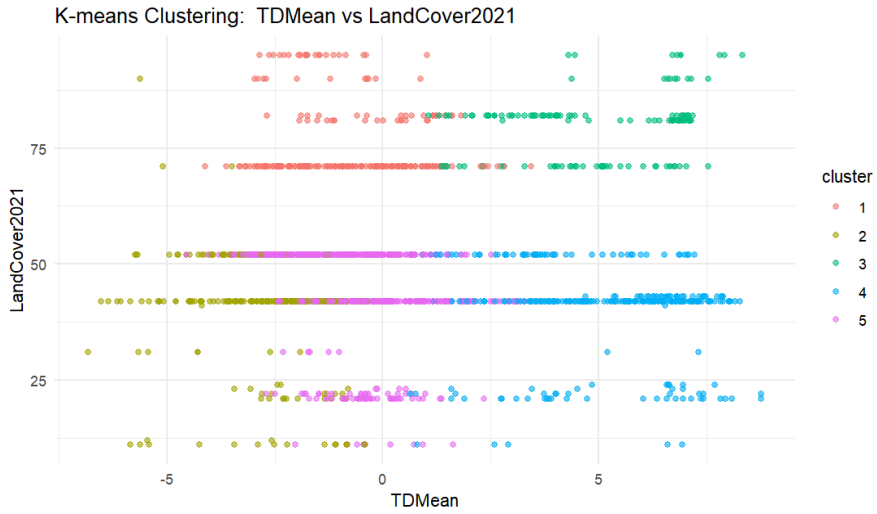


Fig 8.1. K-means (5 cluster) analysis for TDMean vs. LandCover2021

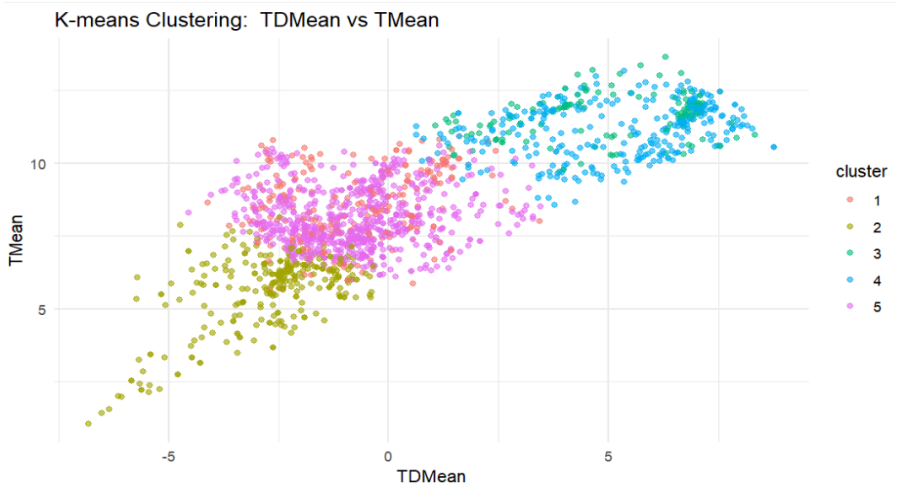


Fig 8.2 K-means (5 cluster) analysis for TDMean vs. TMean

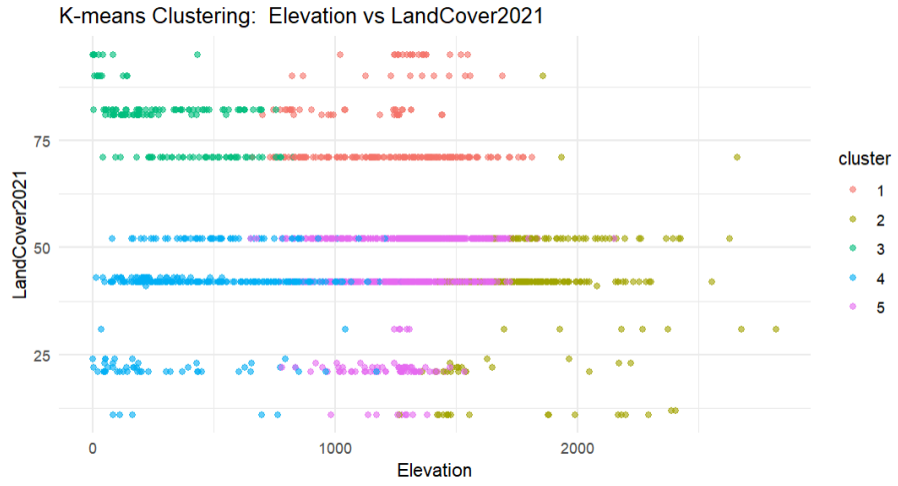


Fig 8.3 K-means (5 cluster) analysis for Elevation vs. LandCover2021

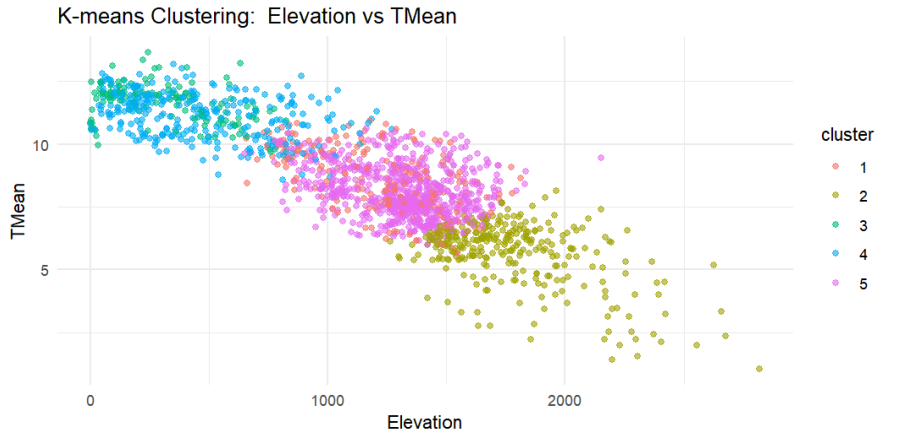


Fig 8.4. K-means (5 cluster) analysis for Elevation vs. TMean

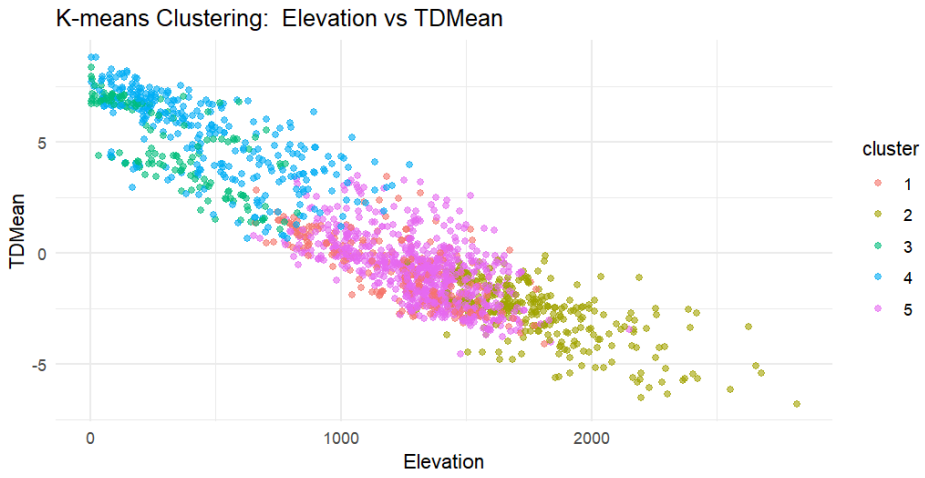


Fig 8.5 K-means (5 cluster) analysis for Elevation vs. TDMean

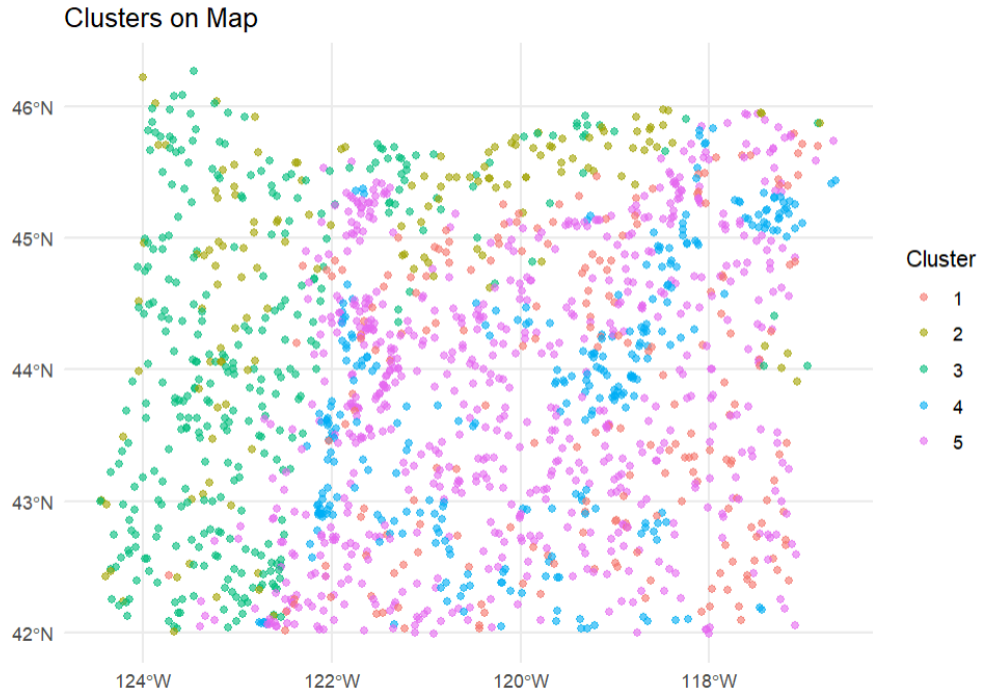


Fig 8.6. Spatial representation of the K-means Cluster analysis for 3 clusters.

	cluster	Elevation	TDMean	TMean	LandCover2021
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	1272.	-1.09	8.45	76.2
2	2	357.	4.41	11.4	78.4
3	3	457.	5.03	11.2	39.3
4	4	1759.	-2.77	5.76	42.4
5	5	1317.	-0.861	8.07	44.7

Density Based Spatial Clustering

minPts- is the minimum number of points required to form a dense region in the dataset. If a region contains at least this number of points within a specified distance (the epsilon distance), it is classified as a dense region.

minPts = 5

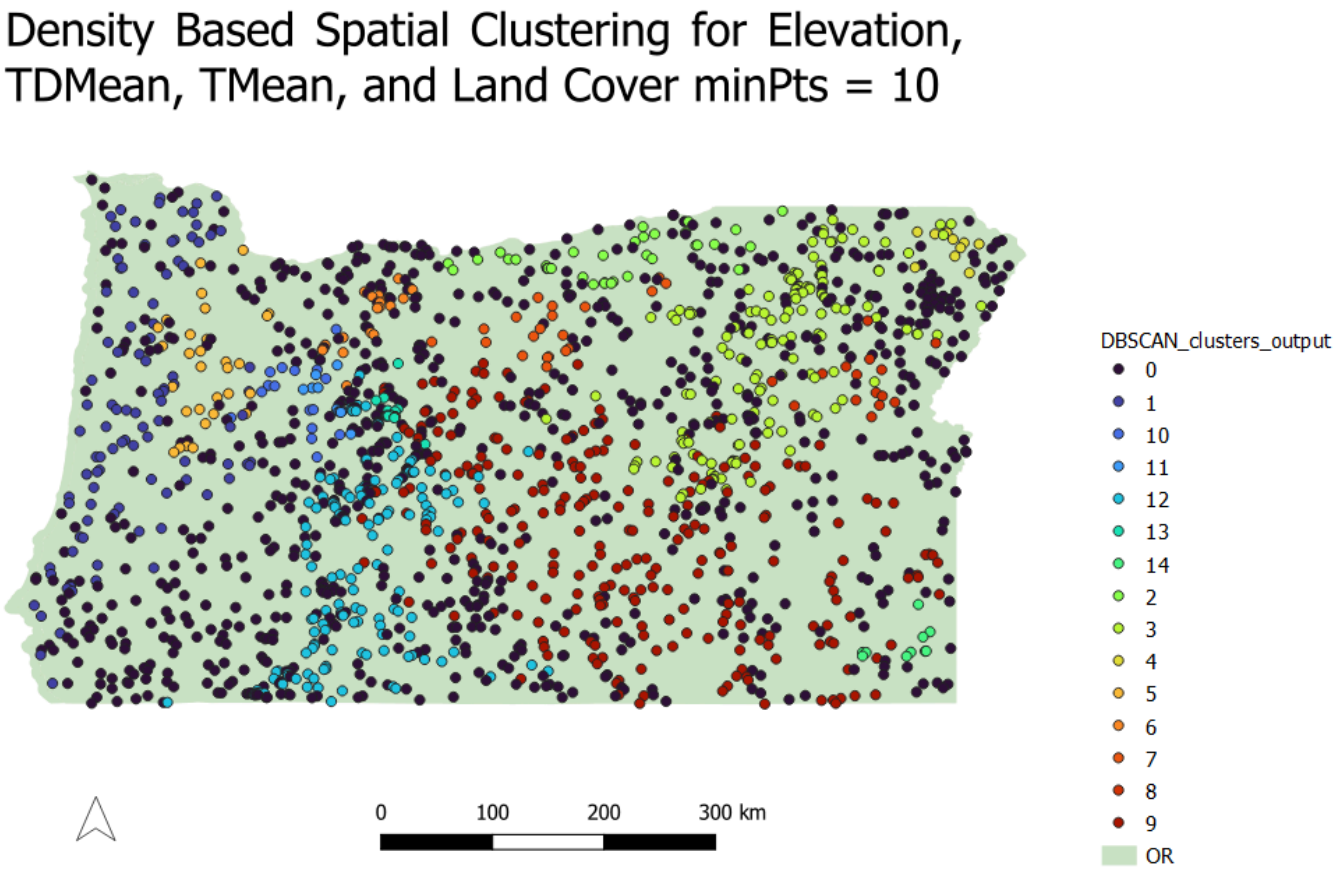


Fig 9. DBSC with minPts = 10

Summary of the Scaled Means in the DBSC:

	cluster	Elevation	TDMean	TMean	LandCover2021
	<fct>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	1140.	0.530	8.64	51.0
2	1	281.	6.74	11.1	42.3
3	2	377.	3.45	11.6	82
4	3	1471.	-1.36	6.92	42
5	4	1331.	-0.492	7.90	71
6	5	109.	6.92	11.9	81.6
7	6	1232.	1.17	7.05	42
8	7	900.	0.538	9.80	71
9	8	1356.	-1.19	7.32	52
10	9	1385.	-2.02	8.25	52
11	10	782.	4.11	9.70	42
12	11	1028.	2.40	8.54	42
13	12	1501.	-0.925	7.04	42
14	13	1004	0.227	8.31	42
15	14	1502	-2.61	9.36	71

Next steps:

Hypothesis: Environmental factors significantly influence the presence of species, with notable differences in mean values of Elevation, Land Cover, Temperature (TDMean, TMean, TMin, TMax), Roughness, and Canopy Cover between presence and absence groups. Specifically, I expect that higher elevations, specific land cover types, and varying temperature conditions correlate positively with species presence.

I am going to use all variables with significant effects on my response variable in my model. I will not be adding any additional variables for the remainder of the project.

Do you have any other steps pending before moving into model training?

One of the remaining steps prior to model training is calibrating my bias mitigation strategies to allow for a larger number of presence points to become part of the final model. This will allow for a more thorough analysis.

Briefly state what are the main difficulties that you have faced in your analysis and your plans to overcome them.

My main difficulties have been in creating processes that involve such large raster data-sets. I have 12 rasters in my main raster stack. Processing this data including seemingly simple actions such as projections, clipping, and resampling take a lot of time. Many times I've had to abort the R sessions, resulting in loss of loaded variables. I had to use larger processing power to complete the cluster analysis, as my rasters include values for each grid cell for the entire state of oregon. The dependencies for the DataExplorer library in R are not compatible with my software as well, so some data exploration actions are hard to complete.

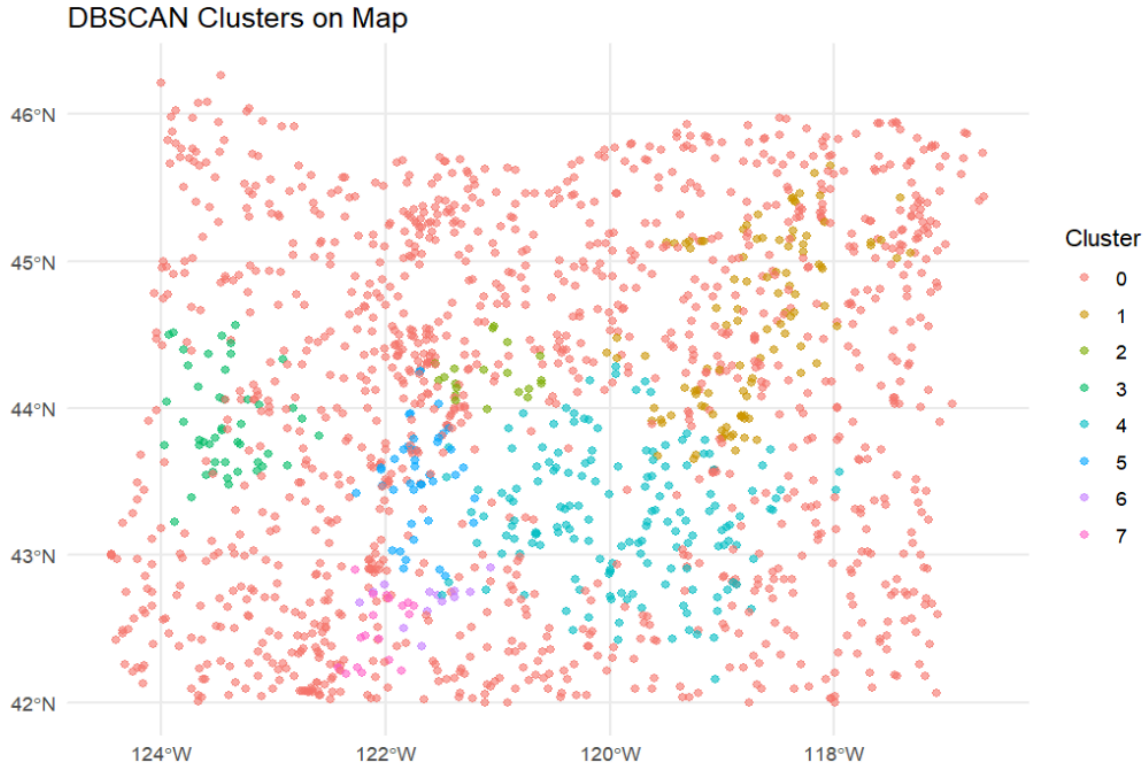


Fig 10. Resampled DBScan

Fig. 11. Northern Goshawk (*Accipiter Gentilis*)

Bias Testing

```
Moran I test under randomisation

data:  SpatData$population_density

weights: lw

Moran I statistic standard deviate = 4.4214e-10, p-value = 0.5

alternative hypothesis: greater

sample estimates:

Moran I statistic       Expectation          Variance

    -8.888889e-04     -8.888889e-04      8.231805e-17
```

```
Asymptotic two-sample Kolmogorov-Smirnov test

data:  SpatData$population_density and population_density_values
D = 0.2495, p-value < 2.2e-16
alternative hypothesis: two-sided
```

There is not much spatial clustering along a single range of population density values according to Moran's statistic.

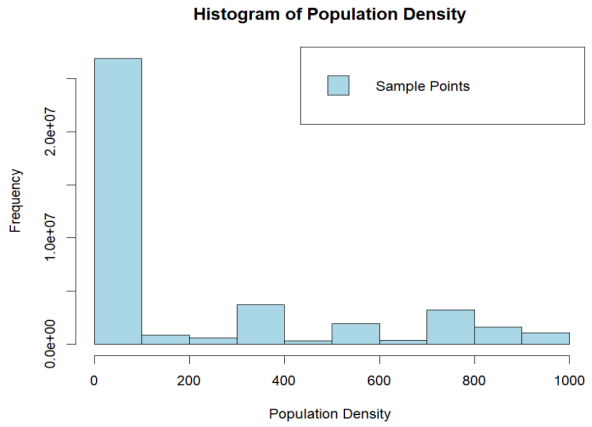


Fig 12.

Choosing a Model:

Random Forest achieved the highest accuracy (84%), the best ROC AUC (92.1%), and the lowest Brier score (0.113).

Accuracy: Correctness of predictions.

ROC AUC: Distinguishing ability between classes.

Brier Score: Reliability of predicted probabilities.

Model	Accuracy	ROC AUC	Brier Score
Random Forest	0.849	0.921	0.113
Neural Network	0.786	0.852	0.188
XGBoost	0.774	0.832	0.210

Fig. 13 Model Testing

I chose random forest and XGBoost, because I don't have enough data for the neural network to come up with reliable results despite model testing.

Model Optimization

Data split: 70-30 split.

Random Forest:

ntree = 854: Number of trees in the forest- best accuracy reducing computation time.

mtry = 3: Number of features randomly considered for splitting at each node to control model complexity and avoid overfitting.

nodesize = 7: Minimum number of samples required in a leaf node, larger values simplify the model.

importance = TRUE: Computes feature importance, allowing insight into which features contribute most to predictions.

XGBoost:

objective = "binary:logistic": Specifies binary classification, where the model predicts probabilities for two classes (0 or 1).

eta = 1.94: Learning rate, which controls the step size during gradient descent. A higher value speeds up learning but may lead to overfitting.

max_depth = 9: Maximum depth of the trees, controlling the complexity of the model. Deeper trees can capture more patterns but may lead to overfitting.

colsample_bytree = 0.782: Fraction of features to sample for each tree, controlling model complexity and preventing overfitting.

gamma = 1.65: Minimum loss reduction required to make a further partition on a leaf node. Larger values prevent overfitting by making splits more conservative.

subsample = 0.782: Fraction of rows used for each tree, reducing overfitting by introducing randomness and preventing the model from memorizing the data.

Results:

Random Forest:

```
Confusion matrix:
     0   1 class.error
0  644 132   0.1701031
1  129 671   0.1612500
```

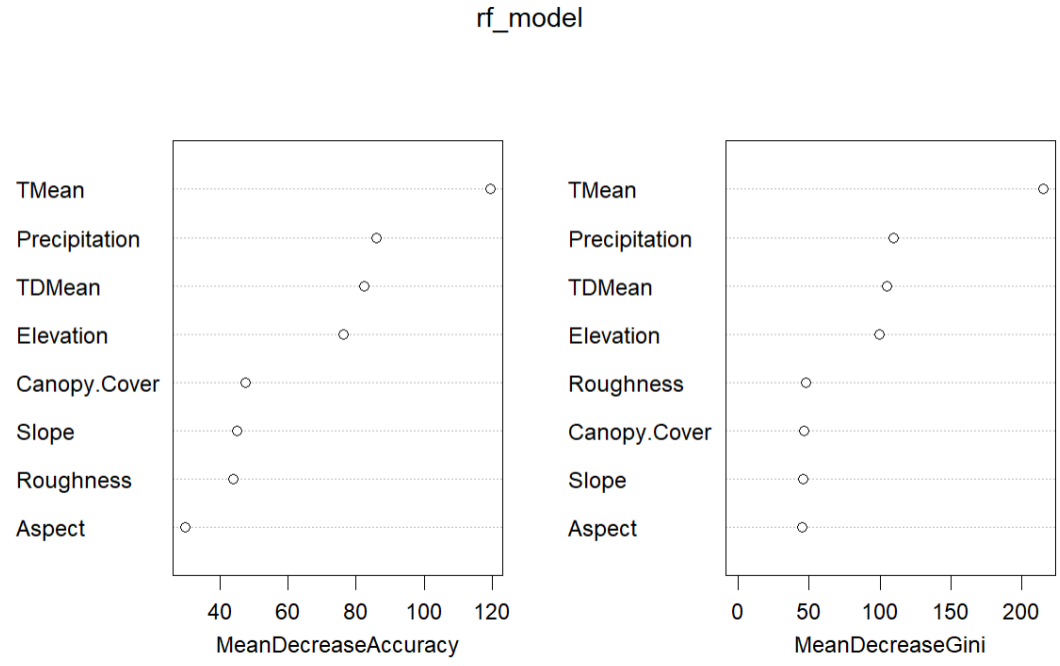


Fig. 14 Feature importance assessment

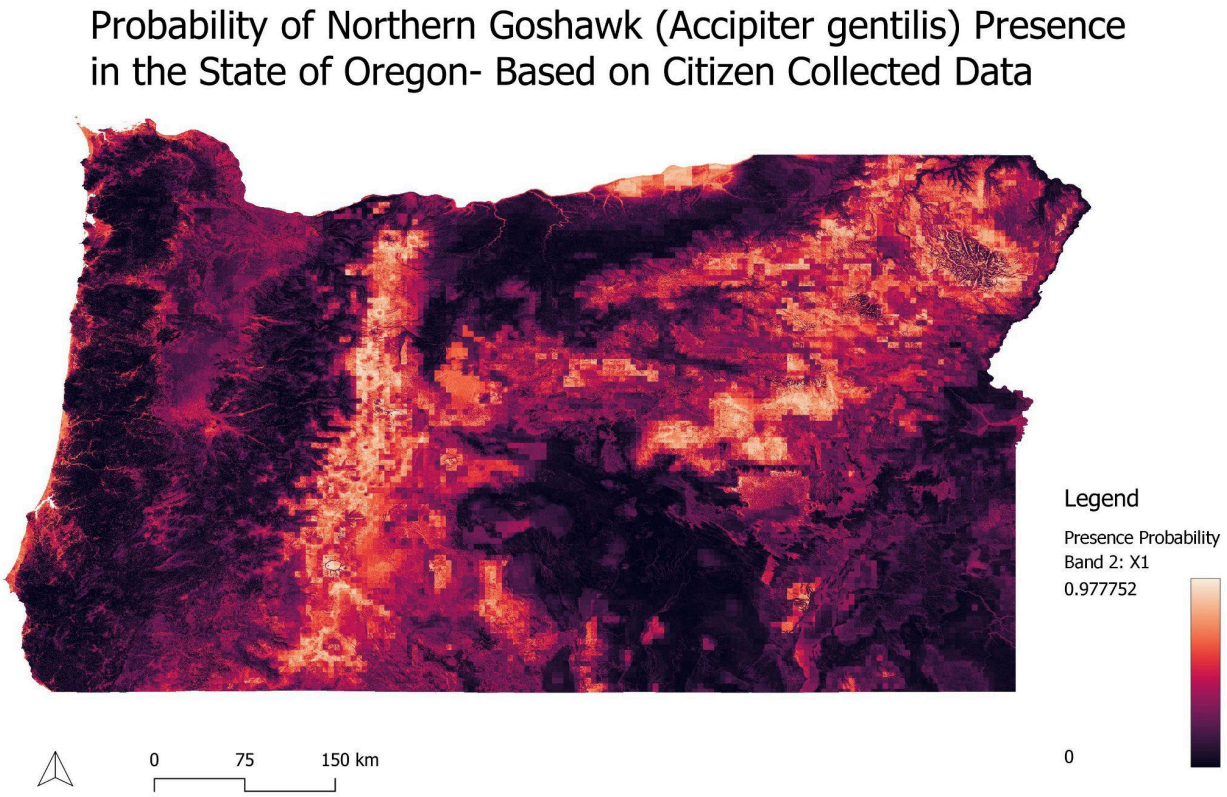


Fig. 15 Final map for the Random Forest Model

XGBoost:

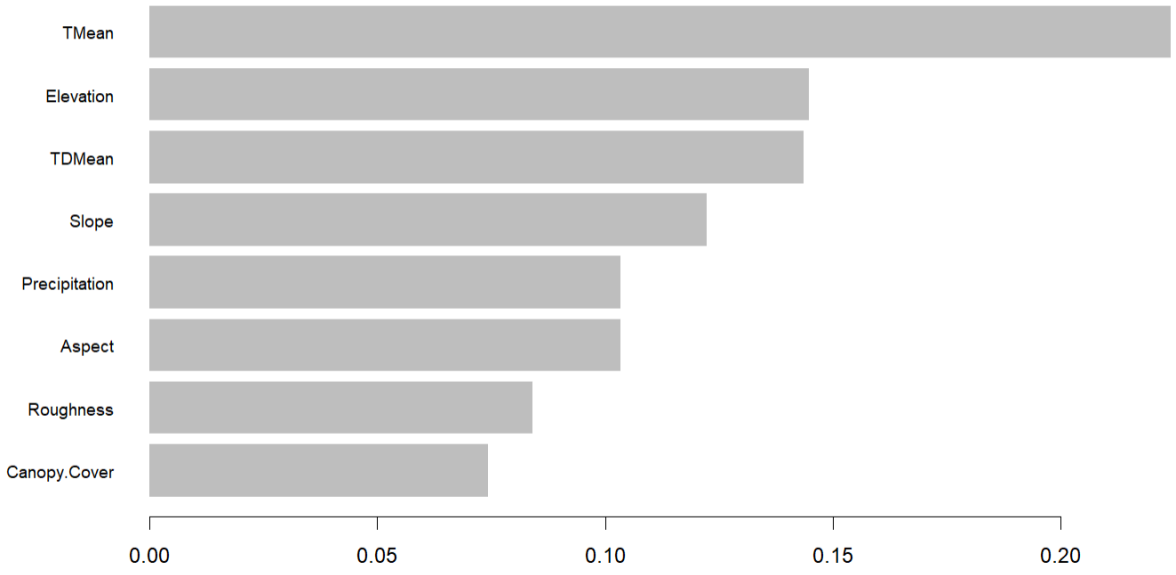


Fig 16. Importance matrix xgboost

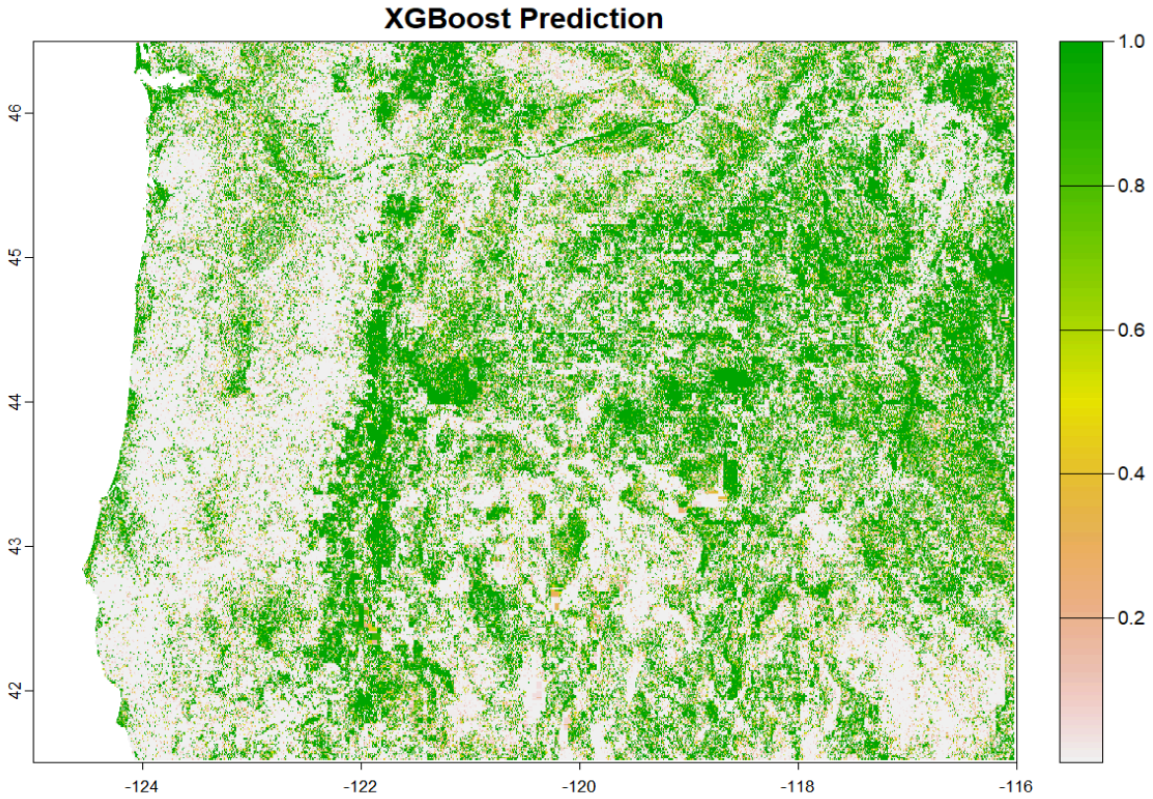


Fig 17. Final XGBoost Probability Prediction

Discussion

Improvements on the model:

There are some collinearities that would be good to address in the model, even though xgboost and random forest are models that both take some measures to reduce the effects of collinearity.

More predictors: vegetation types, more comprehensive land cover and canopy data.

Smaller area with larger # of data points for a different species.

Incorporate bias mitigation on roads for species with larger data counts.

Make distinction between year-round and winter ranges.

Create a server for data storage for larger scale models.

This was a challenging project for the storage space and the virtual memory it required in processing the data.