Habitat usage of reindeer at a Storliden mountain area in Malå municipality in Sweden

Author: Kanchana Weerasinghe
Business Intelligence (AMI295) (AMI295, VT24)
line 4: e-mail: h23kalwe@du.se

***Abstract*— This study aimed to predict reindeer habitat preferences in Arctic ecosystems, focusing on the impact of human-made structures and natural features. Employing logistic regression, recursive feature elimination (RFE), and a random forest approach, key environmental predictors such as proximity to power lines, roads, elevation, and slope were analyzed. The results indicate that distance to power lines alone does not significantly affect reindeer habitat choice. The predictive model demonstrated an accuracy of 64.43%, high sensitivity (94.05%), and low specificity (18.53%), suggesting an inclination to overpredict non-visit instances.**

Keywords- Pellet, Broad Leaved Forest, Clear Cut Forest, Young Forest, Coniferous Forest, Mine, Power Lines

Introduction

I. INTRODUCTION

Understanding reindeer behavior and habitat usage patterns is crucial for effective wildlife management and conservation efforts. Recent scientific investigations have shed light on the unique visual abilities of reindeer, particularly their capacity to perceive ultraviolet (UV) radiation.

This study endeavors to investigate:

1) Assess whether reindeer exhibit avoidance behavior near power grid lines.
2) Determine if this behavior can be influenced by the physical characteristics of the location, the presence of water sources, the distance to man-made structures, and forest characteristics.
3) Reindeer visits are based on the physical characteristics of the location, the presence of water sources, the distance to man-made structures, and forest characteristics.
4)Predict the next visit based on the plot's characteristics.

.

II. LITERATURE REVIEW

Habitat Usage Patterns of Reindeer

Reindeer (Rangifer tarandus) are integral to Arctic ecosystems, and their habitat preferences and behavioral patterns have been the subject of numerous studies due to their ecological and cultural significance. Understanding these patterns is critical for wildlife management and conservation. Reindeer exhibit complex behaviors in response to various environmental stimuli, including vegetation type, human infrastructure, and seasonal changes.

Reindeer Vision and Ultraviolet Perception

One of the unique aspects of reindeer biology is their ability to perceive ultraviolet (UV) light, a capability that significantly influences their interactions with their environment. According to studies by Hogg et al. (2011) and Tyler et al. (2014), reindeer can detect UV light, which aids in foraging by enhancing the contrast between food sources and the snow-covered ground. This visual ability also affects their response to artificial UV sources, such as those emitted by electric power lines, potentially leading to avoidance behavior.

III. METHOD DESCRIPTION

A. The Dataset

The dataset comprises a comprehensive collection of environmental and spatial variables relevant to habitat usage patterns of reindeer in the Storliden mountain area of Malå municipality, Sweden. Each variable provides insights into different aspects of the landscape and its suitability for reindeer habitation. Here's a brief description of each variable.

Description of variables in the data file, Stor1.txt (online resource)		
Variable Name	**Explanation**	**Vales/Range**
Elevation	Elevation	308 – 504 m amsl
Slope	Slope	0 – 19 degrees
VRM	Ruggedness index	0 – 0.084
kNN	Forest age structure	0-107 year old
Distpow, logDistpow (the logarithm+1)	Distance to power lines	1241 – 5169 m
Distroad, logDistroad (the logarithm +1)	Distance to all roads	0 – 921 m
Distbig, logDistbig (the logarithm +1)	Distance to all big roads	500 - 3354 m
Distgruva, logDistgruva (the logarithm+1)	Distance to the (mine north of the mountain, now closed)	200 - 4215 m
Distmall, logDistsmall	Distance to all small roads (forest roads mainly)	0 - 863 m
SMDBLeav	Broad-leaved forest	Dummy; 0, 1
SMDConi	Coniferous forest	Dummy; 0, 1
SMDClear	Clear cut	Dummy; 0, 1
SMDYoung	Young forest	Dummy; 0, 1
SMDMire	Mires	Dummy; 0, 1
SMDLake	Lake	Dummy; 0, 2
Aspect4Flat	Flat areas	Dummy; 0, 1
Aspect4NW	Northwest slope	Dummy; 0, 1
Aspect4NE	Northeast slope	Dummy; 0, 1
Aspect4SE	Southeast slope	Dummy; 0, 1
Aspect4SW	Southwest slope	Dummy; 0, 1
ID	One ID for each plot	factor
Pellet_2009	Wheteht any pellet was found in 2009	Dummy; 0,1
Pellet_2010	Wheteht any pellet was found in 2010	Dummy; 0,1

Table 1: Description of variables in the data set

B. Data Mining process :

This section provides an overview of the process initiated to achieve the aim, which is illustrated in Figure 1.

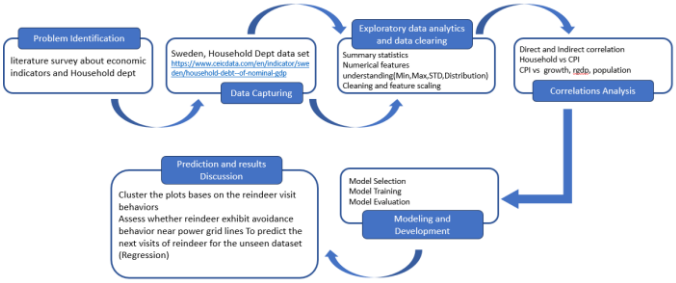


Figure 1: Data Manning process

Exploratory data analysis:

In the exploratory data analysis (EDA) techniques involved examining the numerical features of the dataset, such as pellet counts and forest characteristics, to understand their distributions and relationships. Summary statistics, including measures of central tendency and dispersion, were computed to describe the data's central tendencies and variability.

Additionally, data cleaning procedures were implemented to ensure the quality and integrity of the dataset. This involved identifying and handling missing values, outliers, and inconsistencies that could potentially affect the analysis results. Special attention was paid to maintaining the accuracy of the data while preparing it for further analysis.

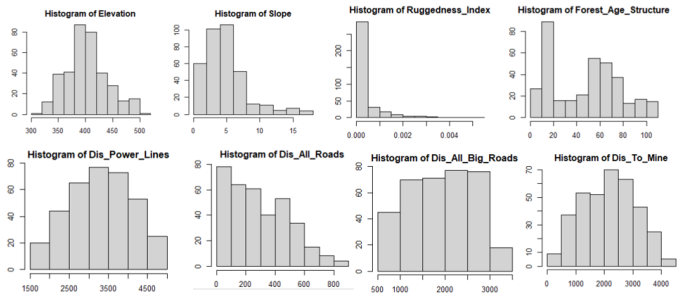


Figure 2: Numeric variables distribution

Feature Scaling:

For the analysis and visualization purpose distance features has been normalized using log scale and Min-Max normalization technique

Feature Extraction:

Three New feature has been introduced called , Visit-categorical , Visit Status- Numerical and Pallet_Combination to make the analysis phase more productive and efficient

Visit – To indicate whether give slot has been visited by the reindeer in 2009 or 2010 or both (no – Not visited , yes- visited)

Visit Status – To indicate whether give slot has been visited by the reindeer in 2009 or 2010 or both (no – 0 visited , yes- 1)

Pallet_Combination- This feature is to consolidate the Pallet_2009 and Pallet_2010. This field consist with categorical data as 2009-Visited ,2010-Visited ,Both ,None

Feature Grouping:

Features has been conceptually categorized to analysis the data in many different aspects

Physical characteristics of the place	Water source	Distance to man made	Forest Characteristics
Elevation Ruggedness index Flat areas Slope Northwest slope Northeast slope Southeast slope Southwest slope	Mires (wet, swampy areas) Lake presence	Distance to power lines Distance to all roads Distance to all big roads Distance to the (mine north of the mountain, now closed) Distance to all small roads (forest roads mainly)	Forest age structure Broad-leaved forest Coniferous forest Clear cut Young forest

Table 2: Feature Grouping

Correlation:

This analysis was conducted to understand the correlation among the variables

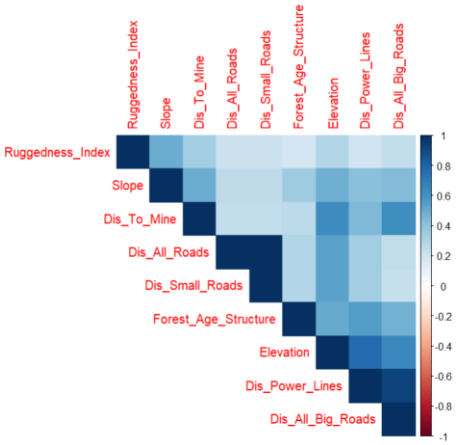


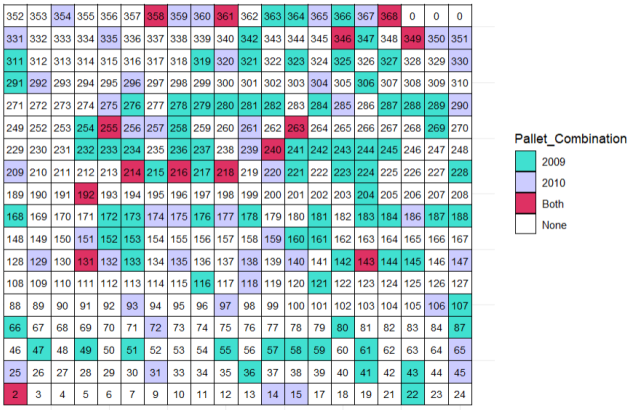
Figure 3: Feature correlation

Based on the above findings, it's evident that a robust correlation exists between Dis_All_Rods and Dis_Small_Rods. This multicollinearity may significantly influence the feature importance analysis within the Random Forest model

IV. Results and Analysis

Reindeer visited slots Analysis:

The purpose of this analysis is to identify and visualize the visited slots regardless of any underlying reasoning factors



Total Slots	357	%
Visited in 2009 Count	79	22.1
Visited in 2010 Count	46	12.8
Visited in both years Count	15	4.20

Figure 4: visited slots

Feature impotency analysis for the visits:

This analysis has been conducted to identify most significant features for the reindeer visit using Random Forest model

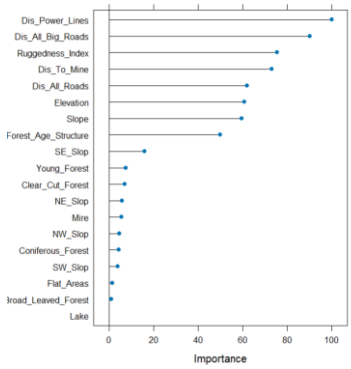


Figure 5: feature importance test results

***Research Quation 1**: Assess whether reindeer exhibit avoidance behavior near power grid lines:*

To understand the statistical significance of the power line impact on the reindeer exhibit, the investigation needs to focus on how well the predictor Dis_Power_Lines can predict reindeer visits and the statistical significance of that outcome. To evaluate this, a statistical model has been developed using random forest, with Dis_Power_Lines as the predictor and Visit as the response variable.

H0- Power line has no significant adverse effect on the reindeer exhibit
Predictor: Distance to power line | Target: Visit
Setting
Original Data set: No Class observations: 217, Yes Class observations: 140
Number Of Trees: 600, K-fold: 10, Number of repeats:100 hyperparameters to tune: 1 – 10
Results

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  14686  8408
       yes  7014  5592

               Accuracy : 0.568
                 95% CI : (0.5629, 0.5732)
    No Information Rate : 0.6078
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0776

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.6768
            Specificity : 0.3994
         Pos Pred Value : 0.6359
         Neg Pred Value : 0.4436
             Prevalence : 0.6078
         Detection Rate : 0.4114
   Detection Prevalence : 0.6469
      Balanced Accuracy : 0.5381

       'Positive' Class : no
```

Figure 6: Model evaluation result of power line distance (predictor) and Visit (target)

- The model has an **accuracy** of 56.8%, meaning it correctly predicts the class of about 56.8% of the instances.
- **Sensitivity** (67.68%) is higher than **specificity** (39.94%), indicating the model is better at correctly identifying "yes" instances compared to "no" instances.
- The **Kappa** value (0.0776) suggests only slight agreement between predicted and actual classifications, indicating the model's performance is not significantly better than random guessing.
- Mcnemar's Test **P-Value** (<2e-16) indicates a significant difference in the errors made by the model compared to another method.
- The **positive predictive value** (63.59%) indicates that when the model predicts "no", it is correct about 63.59% of the time.
- The negative predictive value (44.36%) indicates that when the model predicts "yes", it is correct about 44.36% of the time.

Overall, a p-value of 1 suggests that there **is no statistical evidence to reject the null hypothesis**. In other words, Dis_Power_Lines alone does not have the statistical significance to conclude that the distance of the power line has a significant adverse effect on the reindeer exhibit.

***Research Quation 2**: Assess whether reindeer exhibit can be decided by Physical characteristics of the place, Water source, Distance to man-made, Forest Characteristics*

H0- Above factors no significant adverse effect on the reindeer exhibit
Predictor: Distance to power line, Distance to man-made, Physical characteristics of the place, Forest Characteristics | Target: Visit
Setting
Original Data set: No Class observations: 217, Yes Class observations: 140
Number Of Trees: 600, K-fold: 10, Number of repeats:100 hyperparameters to tune: 1 – 10
Results

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  17898  9004
       yes  3802  4996

               Accuracy : 0.6413
                 95% CI : (0.6363, 0.6463)
    No Information Rate : 0.6078
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1945

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.8248
            Specificity : 0.3569
         Pos Pred Value : 0.6653
         Neg Pred Value : 0.5679
             Prevalence : 0.6078
         Detection Rate : 0.5013
   Detection Prevalence : 0.7536
      Balanced Accuracy : 0.5908

       'Positive' Class : no
```

Figure 7: Model evaluation result of other predictors and Visit (target)

- **Accuracy**: The model correctly classifies about 64.13% of instances, which is an improvement over the NIR (60.78%).
- **Kappa**: At 0.1945, the Kappa statistic suggests only slight to fair agreement between predicted and actual classes, indicating that while the model is better than random guessing
- **Sensitivity**: The model is good at identifying "no" instances, with a high sensitivity of 82.48%.
- **Specificity**: The model struggles with identifying "yes" instances, with a low specificity of 35.69%.
- **Precision**: When the model predicts "no", it is correct 66.53% of the time.
- **Negative Predictive Value**: When the model predicts "yes", it is correct 56.79% of the time.
- **Mcnemar's Test P-Value**: The very low p-value (< 2.2e-16) indicates that the model's errors are significantly different from what would be expected by random chance, affirming that the model has predictive power.
- **Balanced Accuracy**: At 59.08%, the balanced accuracy suggests that the model has moderate performance but is biased towards predicting "no".

***Research Quation 3**: Visit clustering based on Physical characteristics of the place, Water source, Distance to man-made, Forest Characteristics*

Distance to power lines Clustering:

• Distance to power lines – DPL • Distance to all roads - DAR • Distance to all big roads - DBR • Slope -SLP • Ruggedness index - RugI	• Distance to the (mine north of the mountain, now closed)- DMI • Distance to all small roads (forest roads mainly)- DSR • Elevation- Ele

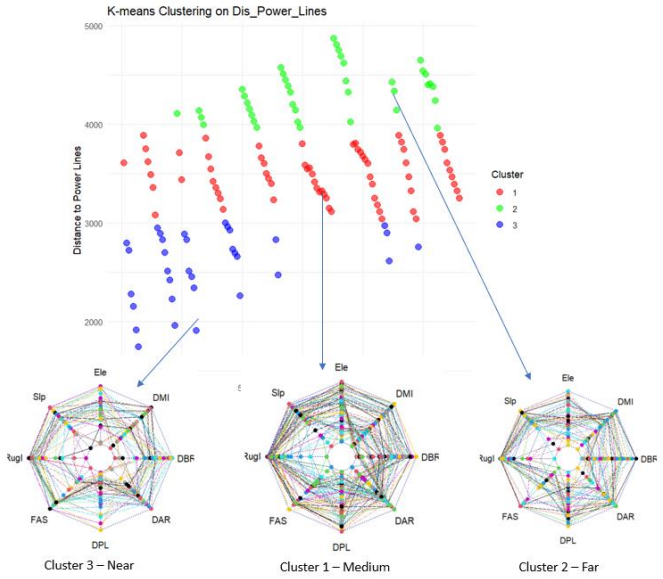


Figure 8: Distance to the power line clustering results and other features mapping

Figure 8 illustrates the clustering based on the distance to power lines and visit data. The radar chart highlights how other behaviors vary with the cluster-specific observations. These observations indicate that there are no significant differences in the other characteristics of the plots across the individual clusters

Forest age structure clustering:

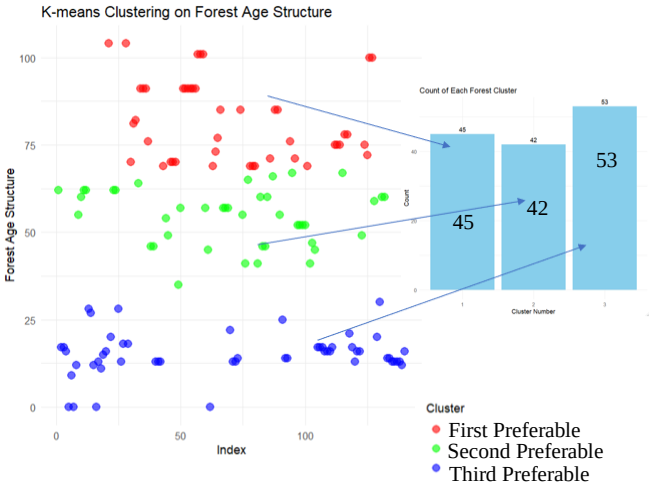


Figure 9: Forest Age structure clustering results and preference

Based on the clustering results for forest age structure, it is evident that forests aged 0 to 25 years are the most preferable forest structure than others.

***Research Quation 4**: Predicting next visit based on the plot's characteristics*

Logistic regression would be a good choice for predicting the next visit due to:

1) Relationships between the predictors and the log-odds of the response variable. When all predictors are numerical, it can efficiently capture these linear relationships.

2) With numerical data, logistic regression coefficients directly represent the change in the log-odds of the outcome for a one-unit change in the predictor. This makes the model easy to interpret and understand, which is valuable for explaining the impact of each variable.

Feature selection:
Feature selection is carried out with the Recursive Feature Elimination (RFE) feature selection method that helps to identify the most important predictors in a dataset by recursively eliminating less important features

Setting
Original Data set: No Class observations: 217, Yes Class observations: 140
K-fold: 10, Number of repeats:5, algorithm: rfe

```
Recursive feature selection

Outer resampling method: Cross-Validated (10 fold, repeated 5 times)

Resampling performance over subset size:

 Variables Accuracy   Kappa AccuracySD KappaSD Selected
         1   0.5456 0.01153    0.07819  0.1649
         2   0.6297 0.17787    0.07015  0.1653
         3   0.6224 0.13634    0.05706  0.1394
         4   0.6314 0.20045    0.07161  0.1492
         5   0.6335 0.20211    0.06742  0.1371
         6   0.6459 0.22913    0.06092  0.1285        *
         7   0.6447 0.22259    0.06615  0.1389
         8   0.6458 0.22460    0.07524  0.1598
         9   0.6414 0.22250    0.07039  0.1496
        10   0.6320 0.19990    0.07923  0.1678
        11   0.6348 0.20696    0.07509  0.1578
        12   0.6320 0.19847    0.07140  0.1527
        13   0.6330 0.19636    0.06927  0.1458
        14   0.6297 0.19076    0.06026  0.1292
        15   0.6441 0.22177    0.06692  0.1398
        19   0.6436 0.22083    0.05887  0.1241

The top 5 variables (out of 6):
   Dis_All_Log_Big_Roads, Dis_Log_Power_Lines, SE_Slop, Elevation, Dis_To_Log_Mine
```

Figure 10: Recursive Feature Elimination results

- The selected subset provides a good trade-off between accuracy and model complexity, with an average accuracy of 64.59% and a Kappa of 0.22913.
- The most important predictors identified are related to the **distance to roads, power lines, slopes, elevation, and distance to mines**.

Logistic regression model for prediction:

Setting
Original Data set: No Class observations: 217, Yes Class observations: 140
K-fold: 10, Number of repeats:100 , family = "binomial"
Selected Features: Dis_Log_Power_Lines, Dis_All_Log_Big_Roads, SE_Slop, Elevation, Dis_To_Log_Mine, Young_Forest
Result

```
Confusion Matrix and Statistics

          Reference
Prediction    no   yes
       no  20366 11417
       yes  1334  2583

               Accuracy : 0.6428
                 95% CI : (0.6378, 0.6478)
    No Information Rate : 0.6078
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.141

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9385
            Specificity : 0.1845
         Pos Pred Value : 0.6408
         Neg Pred Value : 0.6594
             Prevalence : 0.6078
         Detection Rate : 0.5705
   Detection Prevalence : 0.8903
      Balanced Accuracy : 0.5615

       'Positive' Class : no
```

Figure 11: Logistic regression results and preference

- The model's **sensitivity** is high at 93.85%, indicating that it is effective at identifying positive cases.
- However, the **specificity** is low at 18.45%, suggesting that the model struggles to correctly identify negative cases.
- The **accuracy** is slightly better than the **no information rate** (60.78%), but not significantly high, standing at 64.28%.
- The model's performance might be skewed due to **class imbalance**, as indicated by the **prevalence** (60.78%) and **detection prevalence** (89.03%).
- The model's accuracy of 64.28% is slightly better than what would be achieved by chance (the no information rate), which is 60.78%. This difference is statistically significant with a p-value of < 2.2e-16.
- However, despite the statistically significant accuracy, the model's kappa statistic of 0.141 indicates only slight agreement beyond chance.
- Additionally, Mcnemar's test shows a highly significant p-value of < 2.2e-16, indicating a significant difference between the model's predictions and the true values.

V. CONCLUSION

By examining spatial data collected during the spring seasons of 2009 and 2010, the research aimed to determine whether reindeer exhibit avoidance behavior near power grid lines and to identify key environmental factors influencing their habitat preferences.

Impact of Power Grid Lines: Reindeer tend to avoid areas near power grid lines, but the statistical significance is limited when considering only this predictor. The random forest model showed an accuracy of 56.8%, with higher sensitivity than specificity.

Significant Predictors of Reindeer Visits: Recursive feature elimination (RFE) and logistic regression identified key predictors for reindeer habitat preference: distances to various roads, power lines, slope, elevation, and proximity to mines. The logistic regression model had an accuracy of 64.28% and a kappa statistic of 0.141

Model Performance and Class Imbalance: The logistic regression model had high sensitivity (93.85%) but low specificity (18.45%), indicating it is better at identifying reindeer visits but struggles with non-visits. This is due to class imbalance in the dataset.

Cluster Analysis: Clustering based on forest age and proximity to power lines indicated a preference for younger forests (0 to 25 years) among reindeer, but did not show significant differences in other plot characteristics across clusters.

REFERENCES

[1] C. Hogg, M. Neveu, K. A. Stokkan, L. Folkow, P. Cottrill, R. H. Douglas, and G. Jeffery, "Arctic reindeer extend their visual range into the ultraviolet," Journal of Experimental Biology, vol. 214, no. 12, pp. 2014-2019, 2011. [Online]. Available: https://jeb.biologists.org/content/214/12/2014

[2] N. J. C. Tyler, K. A. Stokkan, C. Hogg, C. Nellemann, I. Vistnes, and G. Jeffery, "Ultraviolet vision and avoidance of power lines in birds and mammals," Conservation Biology, vol. 28, no. 3, pp. 630-632, 2014.
[Online].Available: https://conbio.onlinelibrary.wiley.com/doi/abs/10.1111/cobi.12262

[3] E. Reimers, B. Dahle, S. Eftestøl, J. E. Colman, and E. Gaare, "Effects of a power line on migration and range use of wild reindeer," Biological Conservation, vol. 145, no. 1, pp. 149-153, 2012.

[4] C. Nellemann, I. Vistnes, P. Jordhøy, and O. Strand, "Progressive impact of piecemeal infrastructure development on wild reindeer," Biological Conservation, vol. 113, no. 2, pp. 307-317, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0006320702003228