Reindeer exhibit avoidance behavior near power grid lines at a Storliden mountain area in Malå municipality in Sweden

Author: Kanchana Weerasinghe
Business Intelligence (AMI295) - Data Science
Dalarna Univerity
e-mail: h23kalwe@du.se

***Abstract*— This study aimed to predict reindeer exhibit avoidance behavior near power grid lines, focusing on the impact of power grid line, 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 showed 64.43% accuracy, with high sensitivity (94.05%) and low specificity (18.53%), indicating a tendency to overpredict instances where visits don't occur.**

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

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 [1].
This research is crucial due to the planned expansion of nuclear generating capacity by 2035, necessitating an assessment of wildlife impact [3]. The study aims to:

1) Assess whether reindeer avoid power grid lines.

2) Determine if avoidance is influenced by location characteristics, water sources, proximity to man-made structures, and forest characteristics.

3) Analyze and predict reindeer visit patterns based on these factors.

II. LITERATURE REVIEW

A. *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.

B. *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 and Tyler [1], 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 [2] 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

Table 1: Dataset

	Feature	*Description*
1	Elevation [float]	Elevation (m)
2	Slope [float]	Slope (degrees)
3	VRM [float]	Ruggedness index
4	kNN [float]	Forest age structure
5	Distpow [float]	Dis power lines (m)
6	Distroad [float]	Dis all roads (m)
7	Distbig [float]	Dis big roads (m)
8	Distgruva [float]	Dis mine (m)
9	Distmall [float]	Dis small roads (m)
10	SMDBLeav [bit]	Broad-leaved forest
11	SMDConi [bit]	Coniferous forest
12	SMDClear [bit]	Clear cut
13	SMDYoung [bit]	Young forest
14	Mires [bit]	Mires
15	SMDLake [bit]	Lake
16	ID [int]	ID for each plot
17	Pellet_2009 [int]	Pellet found 2009
18	Pellet_2010 [int]	Pellet found 2010

different aspects of the landscape and its suitability for reindeer habitation. Here's a brief description of each variable.

B. Data Mining Methods :

1) ***Research Question 1 :*** *Reindeer avoid areas near power grid lines (Clasification):*

A statistical model was developed using the random forest algorithm due to its ability to combine predictions from multiple decision trees, handle complexity and variability, and provide insights into the importance of different variables. The predictor variable was the distance to power lines (Dis_Power_Lines), and the response variable was reindeer visits (Visit). The random forest model was chosen for its robustness and capacity to manage complex interactions between variables. To evaluate the null hypothesis $H0$: Power lines have no significant adverse effect on reindeer behavior, the model's statistical significance was assessed, focusing primarily on the p-value and confusion matrix (sensitivity, specificity, precision).

2) ***Research Question 2:*** *Influence of Physical Characteristics, Water Sources, Man-Made Structures, and Forest Characteristics on Reindeer Visits (Clasification):*

The same technique was employed with additional predictors to cover each aspect. This comprehensive model helped identify the most significant features influencing reindeer habitat selection.

3) ***Research Question 3:*** *Clustering Reindeer Visits Based on Various Features (Clustering):*

K-means clustering was employed to analyze reindeer visits based on the physical characteristics of the place, water sources, distance to man-made structures, and forest characteristics after scaling the features with log. The data were grouped based on similarities in these features, providing insights into the patterns and preferences in reindeer habitat usage.

4) ***Research Question 4:*** *To predict future reindeer visits based on plot characteristics (Regeression):*

Logistic regression was used due to its effectiveness in modeling binary outcomes. With numerical data, the change in the log-odds of the outcome for a one-unit change in the predictor was directly represented by logistic regression coefficients. To enhance the model's performance, feature selection was carried out using the Recursive Feature Elimination (RFE) method. Less important features were iteratively removed by RFE, ensuring that the most relevant predictors were used in the final model. The model was evaluated using the p-value and confusion matrix, including sensitivity, specificity, and precision.

C. 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.

D. Feature Scaling:

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

E. Feature Extraction:

Three New feature has been introduced:

Visit – For the random forest algorithm, the response variable "Visit" indicates whether a given slot was visited by reindeer in 2009, 2010, or both years. The classes are defined as follows: "no" (not visited) and "yes" (visited).

Visit Status – To indicate whether a given slot was visited by reindeer in 2009, 2010, or both, the response variable for logistic regression is defined as follows: "no" (0 - not visited) and "yes" (1 - visited)

Pallet_Combination- This feature consolidates the Pallet_2009 and Pallet_2010 fields. It contains categorical data for specific conditions filtering: "2009-Visited", "2010-Visited", "Both", and "None"

F. Feature Grouping:

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

Table 2: Feature Grouping

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

G. *Correlation:*

This analysis was conducted to understand the correlation among the variables

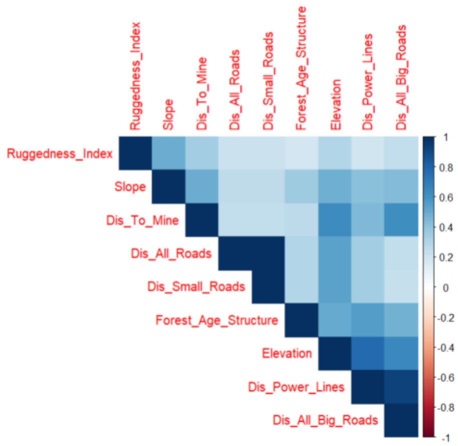

Figure 1: Feature correlation

Based on the above Figure 1, 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

A. *Reindeer visited slots Analysis:*

To provide a clearer understanding of reindeer visits during 2009-2010, the following visualization is based on the Pallet_Combination feature discussed in the feature extraction section..

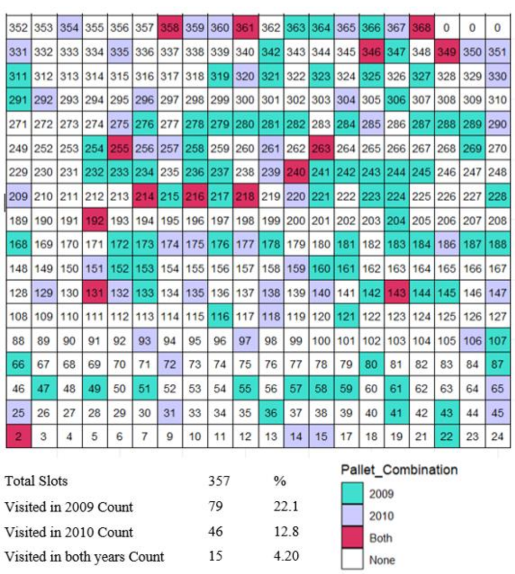

Figure 2: visited slots

B. *Feature impotency analysis for the visits:*

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

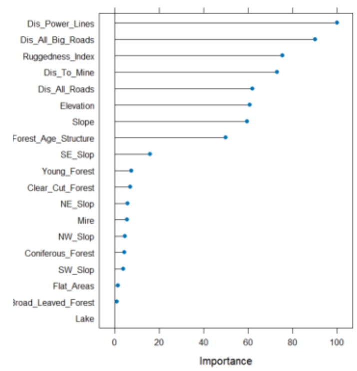

Figure 3: feature importance test results

C. ***Research Question 1****: Assess whether reindeer exhibit avoidance behavior near power grid lines (Clasifictiation):*

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
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

Table 3: Model evaluation result of power line distance (predictor) and Visit (target)

Metric	*Value*
Confusion Matrix	Prediction no yes no 14686 8408 yes 7014 5592
no	14686
yes	7014
Statistics	
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

The model's accuracy is 56.8%, with higher sensitivity (67.68%) compared to specificity (39.94%), indicating it is better at correctly identifying "yes" instances than "no" instances.
The Kappa value (0.0776) suggests only slight agreement between predicted and actual classifications, showing the model's performance is not significantly better than random guessing.
Overall, a p-value of 1 indicates no statistical evidence to reject the null hypothesis, suggesting that the distance to power lines alone does not significantly affect reindeer behavior.

D. ***Research Question 2****: Assess whether reindeer exhibit can be decided by Physical characteristics of the place, Water source, Distance to man-made, Forest Characteristics (Clasifictiation):*

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

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

Table 4: Model evaluation result of other predictors and Visit (target)

Metric	*Value*
Confusion Matrix	Reference Prediction no yes no 17898 9004 yes 3802 4996
no	17898
yes	3802
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

The model correctly classifies 64.13% of instances, improving over the NIR of 60.78%. With a Kappa statistic of 0.1945, it shows only slight to fair agreement between predicted and actual classes, indicating it is better than random guessing.

The model has high sensitivity (82.48%) for identifying "no" instances but low specificity (35.69%) for identifying "yes" instances.

Precision is 66.53% for "no" predictions, while the negative predictive value is 56.79% for "yes" predictions.

Mcnemar's Test P-Value (<2.2e-16) suggests significant predictive power. The balanced accuracy of 59.08% indicates moderate performance but a bias towards predicting "no"

E. ***Research Question 3-1****: Visit clustering based on Distiance to the Power Line (Clustering):*

K-means clustering was employed to analyze reindeer visits based on the physical characteristics of the place, water sources, distance to man-made structures, and forest characteristics after scaling the features with log

Dis power lines (m) -DPL
Dis all roads (m) - DAR
Dis all big roads (m)- DBR
Slope -SLP
Ruggedness index - RugI
Dis to mine (m)- DMI
Dis all small roads (m) - DSR
Elevation- Ele

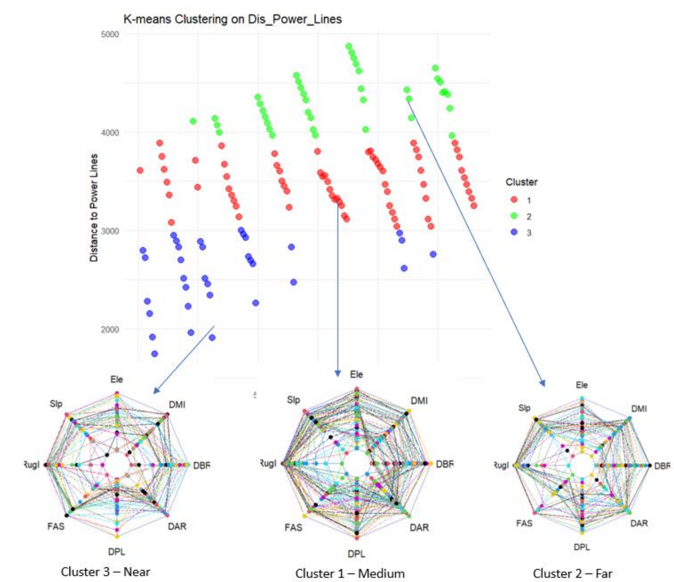

Figure 6: Distance to the power line clustering results and other features mapping

Figure 6 depicts clustering analysis utilizing both distance to power lines and visit data. The radar chart showcases the variations in other behaviors across different clusters. Specifically, it visually indicates any distinctive patterns related to Physical characteristics, Water source, Distance to man-made structures, and Forest Characteristics. However, the graphs reveal no discernible patterns for individual cluster observations based on distance to power lines.

F. ***Research Question 3-2***: *Visit clustering based on Forest Age Stucture (Clustering):*

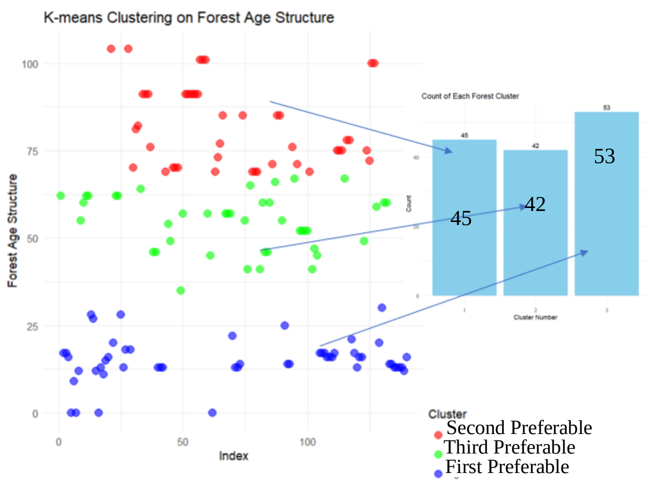

Figure 7: Forest Age structure clustering results and preference

Based on the clustering results (Figure 7) for forest age structure, it is evident that forests aged 0 to 25 years are the most preferable forest structure than others.

G. ***Research Question 4***: *Predicting next visit based on the plot's characteristics (Regression):*

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

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

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.

1) 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

Data set: No Class observations: 217, Yes Class observations: 140

K-fold: 10, Number of repeats:5, algorithm: rfe

Table 5: Recursive Feature Elimination results

variable	*Accuracy*	*Kappa*
1	0.5456	0.01153
2	0.6297	0.17787
3	0.6224	0.13634
4	0.6314	0.20045
5	0.6335	0.20211
power lines	0.6459	0.22913
distance to mines	0.6447	0.22259
distance to roads	0.6458	0.22460
elevation	0.6414	0.22250
10	0.6320	0.19990
11	0.6348	0.20606
12	0.6320	0.19847
13	0.6330	0.19636
14	0.6297	0.19076
slopes	0.6441	0.22177
19	0.6336	0.22083

The selected subset provides a good trade-off between accuracy and model complexity, with an average accuracy of 64.4% and a Kappa of 0.22.

The most important predictors identified are related to the distance to roads, power lines, slopes, elevation, and distance to mines.

2) Logistic regression model for prediction:

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

Table 6: Model evaluation Logistic regression results and preference

Metric	*Value*
Confusion Matrix	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

The model exhibits high sensitivity (93.85%) in identifying positive cases but low specificity (18.45%) for negative cases.

Its accuracy of 64.28% slightly surpasses the no information rate (60.78%).

Class imbalance may skew performance, evident from the prevalence (60.78%) and detection prevalence (89.03%).

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.

1) Impact of Power Grid Lines:

According to the Research Question 1 and Research Question 3-1 results, Reindeer typically avoid areas close to power grid lines, but the statistical significance is somewhat limited when only considering this factor. This suggests that there isn't enough statistical evidence to conclude a direct relationship between the distance to power lines and reindeer natural behaviors based on the dataset. However, this study alone isn't sufficient to definitively determine the extent of human activity's impact, directly or indirectly, on wildlife ecosystems. A more comprehensive study would be necessary to draw concrete conclusions.

2) Impact of all other characteristics:

The findings from Research Question 2 and Research Question 3-2 reveal statistically significant evidence indicating a relationship between reindeer visits and various factors such as Physical characteristics, Water source, Distance to man-made structures, and Forest Characteristics. This suggests that these characteristics directly influence reindeer visits.

3) Significant Predictors of Reindeer Visits:

The results of Research Question 4, utilizing Recursive Feature Elimination (RFE) and logistic regression, pinpointed essential predictors for reindeer habitat preference. These predictors include distances to different roads, power lines, slope, elevation, and proximity to mines. The logistic regression model achieved an accuracy of 64.28% with a kappa statistic of 0.141, suggesting that future reindeer visits can be predicted from this statistically significant model with 64.28% accuracy.

4) Model Performance and Class Imbalance:

The logistic regression model exhibited high sensitivity (93.85%) but low specificity (18.45%), suggesting its proficiency in recognizing reindeer visits but challenges in distinguishing non-visits. This imbalance in classes within the dataset can be mitigated by employing techniques such as Synthetic data creation and advanced methods like Ensemble models.

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] Dataset: https://learn.du.se/bbcswebdav/pid-1005768-dt-content-rid-21790910_1/xid-21790910_1

[3] World Nuclear Association https://www.world-nuclear-news.org/Articles/Roadmap-launched-for-expansion-of-nuclear-energy-i