PROJECT REPORT:

ANALYSIS OF SNOWSHOE HARES IN BONANZA CREEK

By

Weixuan Chen

v. Apr 23, 2023

Journal Style: Biology

1. Introduction

Snowshoe Hares, Lepus americanus, are a 'keystone' prey species in northern boreal forests and experience population fluctuations of 8-11-years. **(Service)**Despite intense responses of both vegetation and predators to changes in hare densities, landscape-scale comparisons of hare populations in Alaska have been limited to qualitative descriptions. Researchers conducted capture-recapture studies of snowshoe hares at 5 locales in the Tanana valley, from Tok in the east to Clear in the west from 1999 to 2002. Snowshoe hare densities were highest in 1999 (=6.36 ha-1, SE=0.63) and declined thereafter. Researchers were unable to detect declines in apparent survival during declining densities in the study populations. **(Ernest, 1974)**Movement distances did not vary temporally and persistence of individuals through declining densities may be associated positively with body condition at the peak. The relationship of hare pellets and hare densities was weak and limits the utility of this methodology for estimating hare densities in Interior Alaska. **(Kielland, 2017) (Hodges, 1999)**

The following paper discusses characteristics of Snowshoe Hares across various sites of the [Bonanza Creek Experimental Forest] **(Fairbanks)** between 1999 to 2011. It is an important prey for larger animals and likewise affects the vegetation of the forest as well as an essential indicator of climate change. **(Centers, 2021) (Willis, 2021)**The species is known to have a population fluctuation of 8-11 years and the original study explores more reasons for observed population decline in detail. The initiative is supported by the Institute of Arctic Biology, University of Alaska Fairbanks. Detailed information can be retrieved from the EDI Portal. **(EDI)**

jjj

2. Materials and Methods

Dataset used for this paper is drawn from EDI Data Portal consisting of variables about snowshoe hares and trappings across the three sites of the experiment site: Bonanza Creek Forest (Bonanza Riparian, Bonanza Mature and Bonanza Black Spruce).

Table 1. Variables and Information of the Dataset.

Table Column Descriptions

	date	time	grid	trap	l_ear	r_ear	sex	age	weight	hindft	notes	b_key	session_id	study
Column Name:	date	time	grid	trap	l_ear	r_ear	sex	age	weight	hindft	notes	b_key	session_id	study
Definition:	Date	Time	name of trapping grid	trap ID	left ear tag	right ear tag	sex of hare captured	age of capture	weight of hare	hind foot	notes on capture	unique animal identifier	unique session idenifier (A session is defined as multiple consecutive trap nights at a given site.)	The study type of this particular trapping session.
Storage Type:	string	string	string	string	string	string	string	string	integer	integer	string	integer	integer	string
Measurement Type:	nominal	nominal	nominal	nominal	nominal	nominal	nominal	nominal	ratio	ratio	nominal	nominal	nominal	nominal

The data is used to analyze statistical characteristics of the hares. The hares' physical characteristics can be used for testing the relationship between the variables of interest. This paper tested the normality of weights and length of hindfeet of hares to check if the data meets the requirements of statistical tests. To examine the relationship between variables, a two-sample t-test is used after the normality examination, and a linear regression of response variable 'weights' regresses on explanatory variable 'height' to examine the relationship further. Finally, machine learning algorithm, specifically the Support Vector Machine, a popular classification algorithm, was performed to investigate if the classification of the habitant site of hares is possible by given features.

jjj

3. Results

3.1 Exploratory Data Analysis

As the data given, the count of snowshoe hares trapped each year can be a strong indicator of how the number of hares fluctuate in the forest year by year. Since the young hares represents the future prosperity of hares, only the hares at age 1 year were selected. The first plot indicates how the number of young hares change over years.

Graph 1. Count of Hares Over Year.

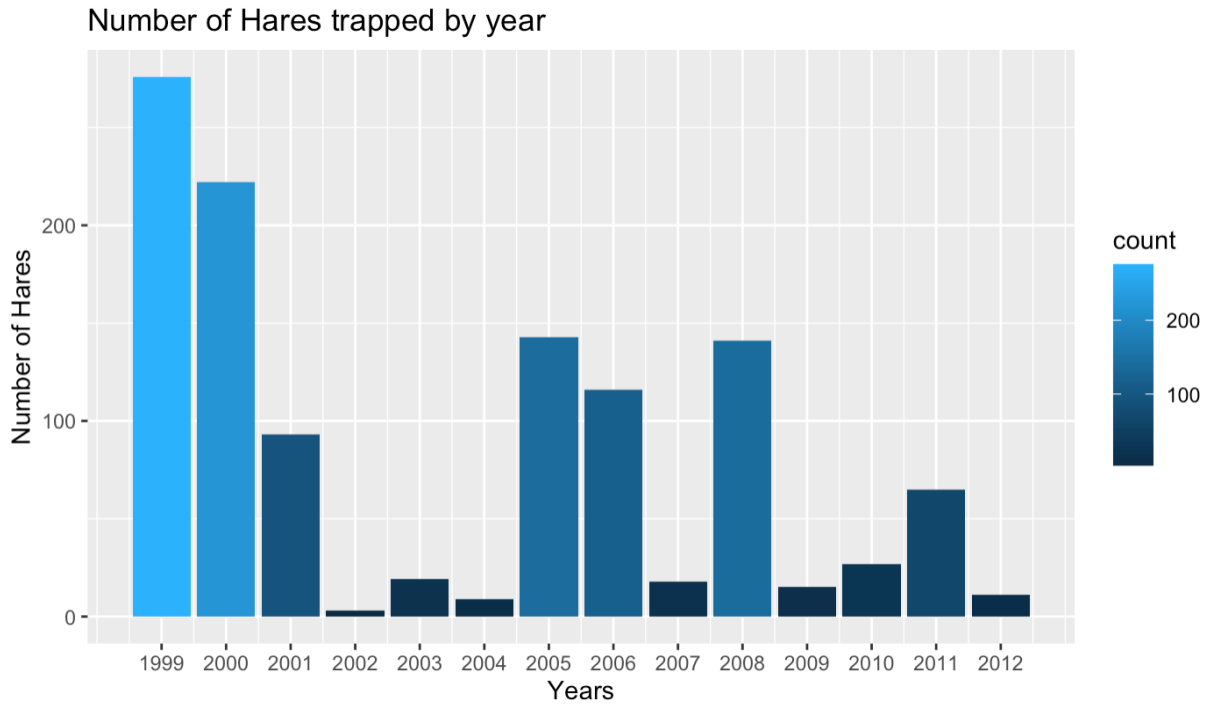


Figure 1: Number of Hares trapped from years 1999 to 2012

Hare population trappings between the years 1999 to 2012 have been depicted in the graph above. As observed, the highest number of trappings, n = 276, has occurred in 1999 and least number of trappings in 2003, n = 3. On an average, the mean number of juvenile hares trapped were 82.71 between 1999 and 2012. There was a sudden drop in the trappings in the year 2002, which fluctuates almost up to year 2008 and gradually never recovered compared to the initial years.

Besides the count by year, the investigation of weights against how sex of snowshoe hares changes with respect to habitant site is also worth studying. The following graph shows the relationship.

Graph 2. Weights vs Sex against grid.

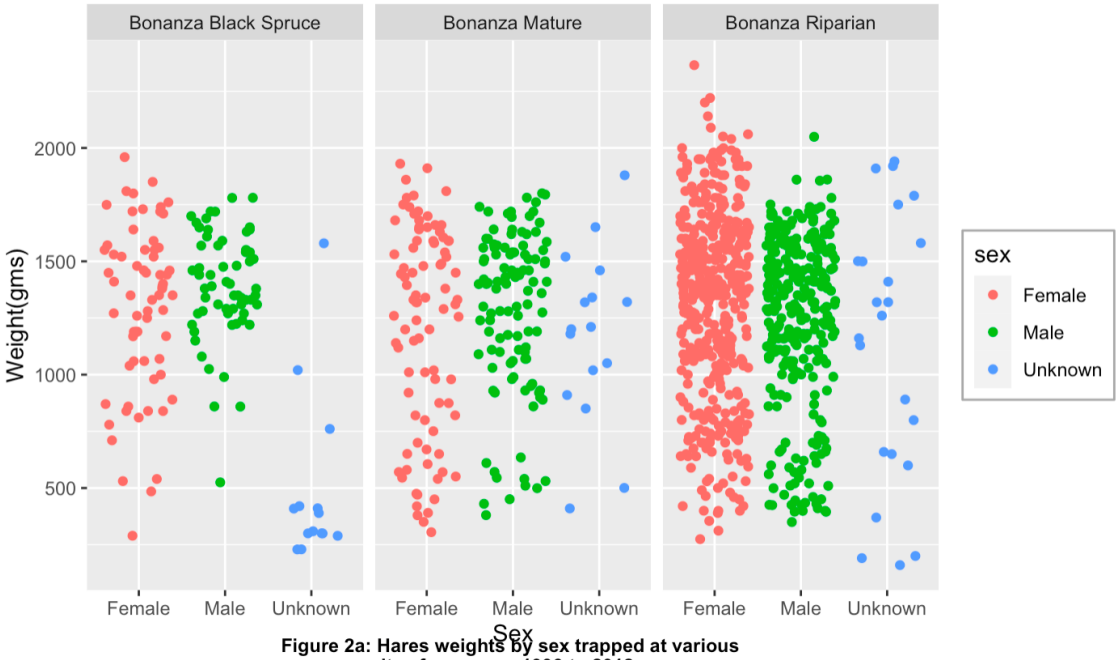


Figure 2a: Hares weights by sex trapped at various sites from years 1999 to 2012

From this plot, the distribution of weights across sex is quite uniform at each grid. How about the distribution of weights for sex at each grid?

Graph 3. Boxplot of Weights vs Sex against grid.

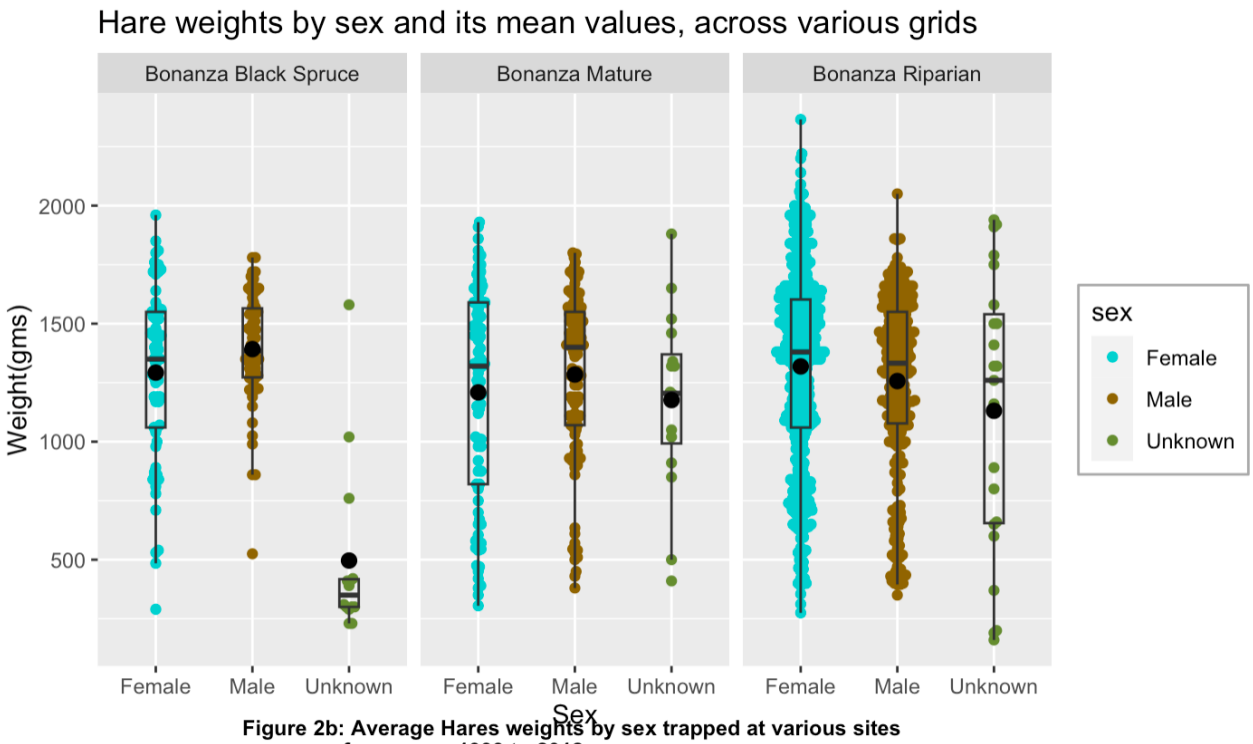


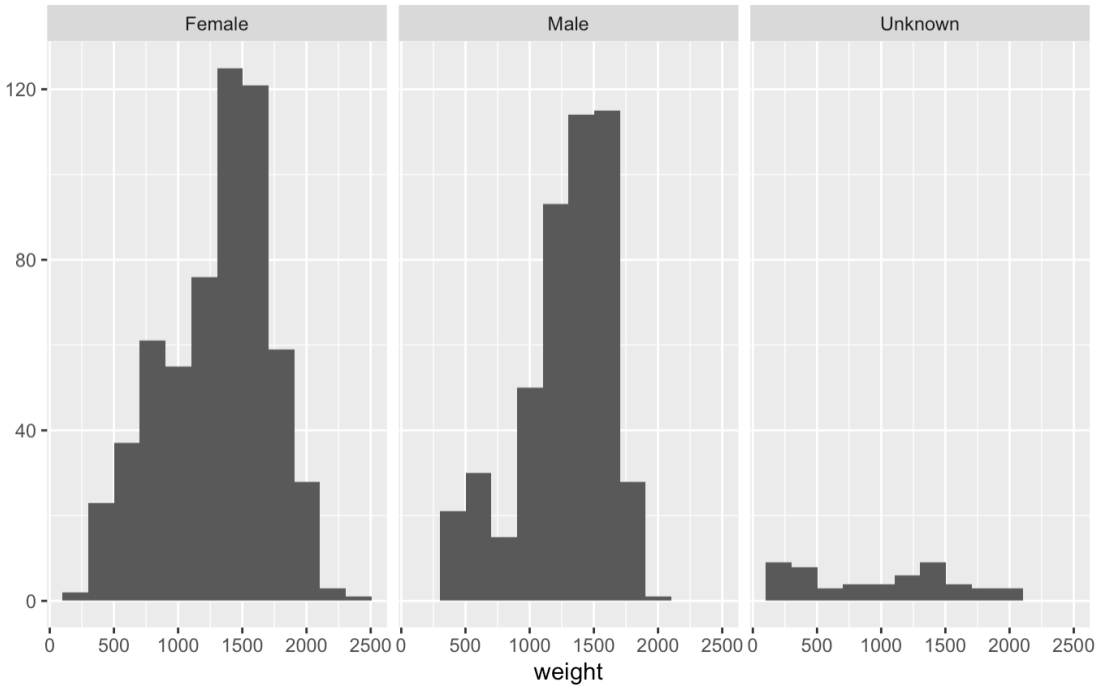
Figure 2b: Average Hares weights by sex trapped at various sites from years 1999 to 2012

On an average it can be observed from Graph 3 that the male hares weigh more than the female hares across the three sites, since the means of male hares are higher than those of female. More hares were trapped at the Bonanza Riparian site compared to the other sites, because of the dense points area covered in the graph.

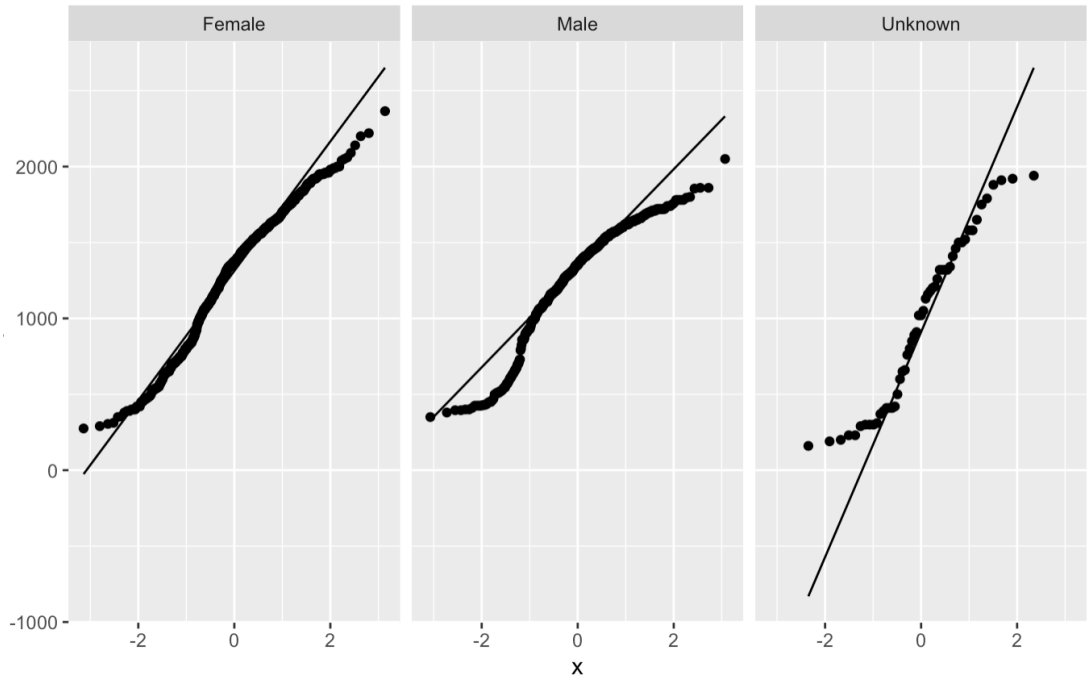
3.2 Statistical Analysis

The previous graph shows evidence of the mean weights of male hare is higher than that of female hare. In this section, a precise statistical test is conducted to test if such relationship is true. Based on the normality assumption of t-test, checking if the data is normally distributed is essential. Therefore, plots of the distribution of male hare weights and female hare weights can be a good start. The following plots show the distribution of weights of male and female hares.

Graph 4. Distribution of male and female hares' weights.



Graph 5. QQ-plot of distribution of male and female hares' weights.



From the graph 4, the distribution of male and female hares' weights looks like almost normal. To further examine the normality, the graph 5, which depicts the QQ-plot, shows the data satisfies the normal assumption, because the data points almost follow the trend for normal quantile.

3.3 Two-sample t test

The variables of interest are weights for male hares and female hares respectively. From the t-test, the mean weights of male hares and female hares are 1281.253 and 1299.682 respectively, and the p-value is greater than 0.05 significant level. Therefore, there is no evidence to show that the mean weight of male hares is difference from the mean weight of female hares.

Table 2. Unpaired t-test for mean weight of male and female hares.

```
        Welch Two Sample t-test

data:  weights_male and weights_female
t = -0.78027, df = 1050.2, p-value = 0.4354
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -64.77496  27.91652
sample estimates:
mean of x mean of y
 1281.253  1299.682
```

3.4 Linear Regression

The hindfeet is also an important characteristic of snowshoe hares. In previous part, the test suggests that there is no difference in mean weight for male hares and female hares statistically. Besides the sex, the relationship between the hindfeet and weight of snowshoe hares is also worh investigating. Therefore, a linear regression with weight as response variable and length of hindfeet as explanatory variable is performed. The following plot and table give the detail statistics:

Graph 6. Linear regression for weights vs length of hindfeet.

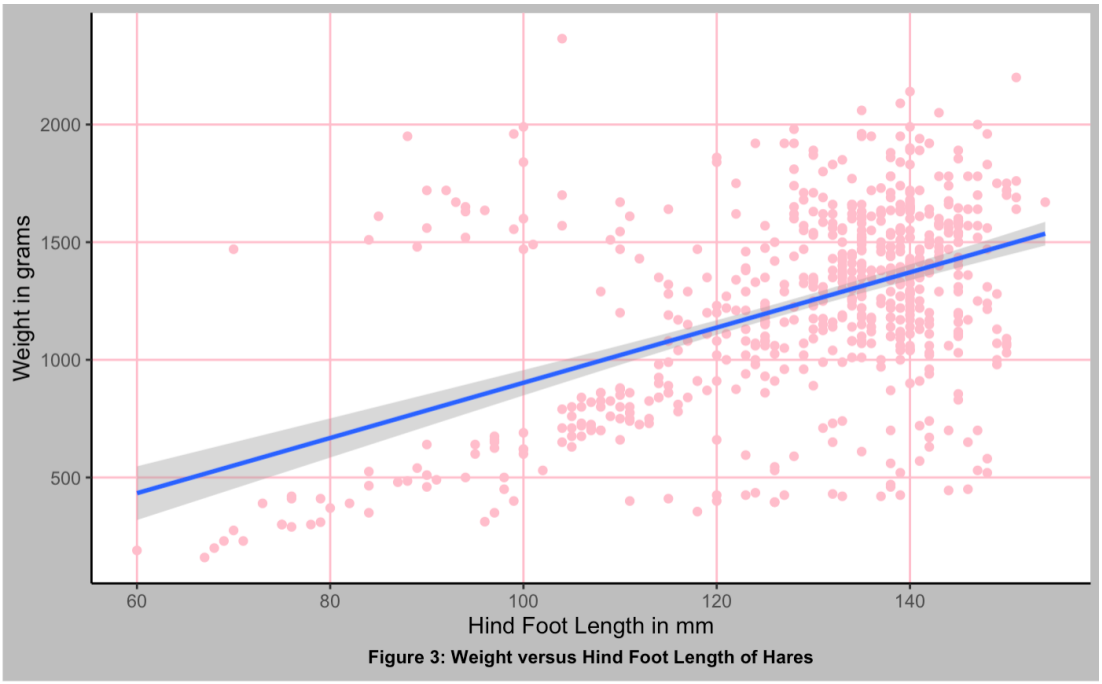


Figure 3: Weight versus Hind Foot Length of Hares

Table 3. Summary of Linear Regression for weights vs hindfeet

```
Call:
lm(formula = weight ~ hindft, data = snow_hare_df)

Coefficients:
(Intercept)       hindft
    -270.62        11.73


Call:
lm(formula = weight ~ hindft, data = snow_hare_df)

Residuals:
    Min      1Q  Median      3Q     Max
-992.18 -235.71   -6.01  213.51 1415.55

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -270.6178   106.7185  -2.536   0.0115 *
hindft        11.7315     0.8242  14.234   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 364.9 on 649 degrees of freedom
  (507 observations deleted due to missingness)
Multiple R-squared:  0.2379,    Adjusted R-squared:  0.2367
F-statistic: 202.6 on 1 and 649 DF,  p-value: < 2.2e-16
```

The simple linear regression indicates that the prediction power of hindfeet to weight of hares is moderate by the p-value of variable weight is < 0.001 (significant) and R-squared value is 0.2379. The slope of hindfeet is 11.7315, which implies one unit increase in hindfoot expects to increase weight by 11.7315.

3.5 Support Vector Machine on Snowshoe hares data

In this section, a classification technique is used to determine which habitant site a hare comes from based on given features. The algorithm chosen is support vector machine with kernel tricks, because it has great ability to deal with linear and non-linear separable hyperplanes. The kernels applied in this section are: linear, polynomial, radial, and sigmoid. The performance metrics is confusion matrix which are presented following:

Table 4. Confusion Matrix for Linear Kernel (left) and Polynomial Kernel (right) of SVM.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 11  0  0
         2  0  1  0
         3  0  0 45

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9373, 1)
    No Information Rate : 0.7895
    P-Value [Acc > NIR] : 1.407e-06

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity             1.000  1.00000   1.0000
Specificity             1.000  1.00000   1.0000
Pos Pred Value          1.000  1.00000   1.0000
Neg Pred Value          1.000  1.00000   1.0000
Prevalence              0.193  0.01754   0.7895
Detection Rate          0.193  0.01754   0.7895
Detection Prevalence    0.193  0.01754   0.7895
Balanced Accuracy       1.000  1.00000   1.0000
Confusion Matrix and Statistics
```

```
          Reference
Prediction  1  2  3
         1 11  0  0
         2  0  1  0
         3  0  0 45

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9373, 1)
    No Information Rate : 0.7895
    P-Value [Acc > NIR] : 1.407e-06

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity             1.000  1.00000   1.0000
Specificity             1.000  1.00000   1.0000
Pos Pred Value          1.000  1.00000   1.0000
Neg Pred Value          1.000  1.00000   1.0000
Prevalence              0.193  0.01754   0.7895
Detection Rate          0.193  0.01754   0.7895
Detection Prevalence    0.193  0.01754   0.7895
Balanced Accuracy       1.000  1.00000   1.0000
Confusion Matrix and Statistics
```

The confusion matrices in Table 4 indicate strong overfitting as the accuracy of both model is 1. Therefore, these two models are not the best model for prediction.

Table 5. Confusion Matrix for Radial Kernel (left) and Sigmoid Kernel (right) of SVM.

```
          Reference
Prediction  1  2  3
         1  0  0  0
         2  0  1  0
         3 11  0 45

Overall Statistics

               Accuracy : 0.807
                 95% CI : (0.6809, 0.8995)
    No Information Rate : 0.7895
    P-Value [Acc > NIR] : 0.4478

                  Kappa : 0.1387

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity             0.000  1.00000  1.00000
Specificity             1.000  1.00000  0.08333
Pos Pred Value            NaN  1.00000  0.80357
Neg Pred Value          0.807  1.00000  1.00000
Prevalence              0.193  0.01754  0.78947
Detection Rate          0.000  0.01754  0.78947
Detection Prevalence    0.000  0.01754  0.98246
Balanced Accuracy       0.500  1.00000  0.54167
Confusion Matrix and Statistics
```

```
          Reference
Prediction  1  2  3
         1  0  0  0
         2  0  0  0
         3 11  1 45

Overall Statistics

               Accuracy : 0.7895
                 95% CI : (0.6611, 0.8862)
    No Information Rate : 0.7895
    P-Value [Acc > NIR] : 0.5765

                  Kappa : 0

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity             0.000  0.00000   1.0000
Specificity             1.000  1.00000   0.0000
Pos Pred Value            NaN      NaN   0.7895
Neg Pred Value          0.807  0.98246      NaN
Prevalence              0.193  0.01754   0.7895
Detection Rate          0.000  0.00000   0.7895
Detection Prevalence    0.000  0.00000   1.0000
Balanced Accuracy       0.500  0.50000   0.5000
```

The confusion matrices in Table 5 show relatively good models where support vector machine model with radial kernel has accuracy 0.807 and with sigmoid kernel has accuracy 0.7895. By comparison, the model with radial kernel has better performance than the other.

4. Conclusions and Discussions

There is an overall decrease in the number of snowshoe hare trappings over the years from 1999-2012 across the various sites. The data also shows that a greater number of trappings were cited at the Bonanza Riparian site compared to the other sites from the exploratory data analysis. Male hares weigh more than female hares, but there is not significant difference in weights, which is interesting to investigate. The machine learning technique, specifically the support vector machine, can classify the sites hares come from, which is meaningful in research level.

References:

1. https://www.ualberta.ca/folio/2020/12/less-winter-snow-could-spell-disaster-for-snowshoe-hares.html(willis,2021)
2. https://www.usgs.gov/news/identifying-effects-climate-and-land-use-change-snowshoe-hare-midwest(CENTERS, 2021)
3. https://www.adfg.alaska.gov/static/home/library/pdfs/wildlife/research_pdfs/74_hare_ernest.pdf(1974)
4. http://www.lter.uaf.edu/data/data-detail/id/1 (FAIRBANKS)
5. https://www.nps.gov/articles/snowshoe-hare.htm(Service)
6. https://www.fs.usda.gov/research/treesearch/50629(1999)
7. Kielland, K., F.S. Chapin, R.W. Ruess, and Bonanza Creek LTER. 2017. Snowshoe hare physical data in Bonanza Creek Experimental Forest: 1999-Present ver 22. Environmental Data Initiative. [Bonanza Creek LTER. Institute of Arctic Biology, University of Alaska Fairbanks](https://doi.org/10.6073/pasta/03dce4856d79b91557d8e6ce2cbcdc14) (EDI)(2017)
8. David Robinson, Alex Hayes and Simon Couch (2021). broom: Convert Statistical Objects into Tidy Tibbles. R package version 0.7.9. https://CRAN.R-project.org/package=broom
9. Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.
10. Kielland, K., F.S. Chapin, R.W. Ruess, and Bonanza Creek LTER. 2017. Snowshoe hare physical data in Bonanza Creek Experimental Forest: 1999-Present ver 22. Environmental Data Initiative. https://doi.org/10.6073/pasta/03dce4856d79b91557d8e6ce2cbcdc14
11. Kirill Müller (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here

12. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

13. Erik Clarke and Scott Sherrill-Mix (2017). ggbeeswarm: Categorical Scatter (Violin Point) Plots. R package version 0.6.0. https://CRAN.R-project.org/package=ggbeeswarm

14. Sam Firke (2021). janitor: Simple Tools for Examining and Cleaning Dirty Data. R package version 2.1.0. https://CRAN.R-project.org/package=janitor

15. David Meyer (2023). e1071: Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier, generalized k-nearest neighbors. https://cran.r-project.org/web/packages/e1071/index.html

16. nl Zhang (2017). CatEncoder: Contains some commonly used categorical variable encoders, such as 'LabelEncoder' and 'OneHotEncoder'. Inspired by the encoders implemented in Python 'sklearn.preprocessing' package. https://cran.r-project.org/web/packages/CatEncoders/CatEncoders.pdf

17. Max Kuhn (2023). caret: Misc functions for training and plotting classification and regression models. https://cran.r-project.org/web/packages/caret/index.html