The Effects of Winter Weather and Land Use on Deer Populations in Wisconsin

Taylor McHone and Payton Bartow

University of Wisconsin-Madison

Introduction

Predicting the dynamics of the Wisconsin deer population is vital to setting hunting thresholds and predicting the effects of deer on other life. The Wisconsin Department of Natural Resources (WDNR) has used a metric called the Winter Severity Index (WSI) to predict the effect of winter on the Wisconsin deer population since the late 1900s. The current Winter Severity Index only takes into account the snow depth and temperature of each day during a winter, and the data used to compute the current WSI must be collected with a delay. In an effort to improve the predictive power of the WSI and automation of the data pipeline, we investigated the use of publicly available weather and land use data to predict deer population dynamics. This resulted in the creation of a new, regional dataset designed to be useful for machine learning.

Section 1 details the various datasets used in our investigation, including weather, land use, and deer population datasets. In Section 2, we looked at the correlation between the WSI and the FDR and in Section 3, we incorporated land use into this model. In Section 4 we clustered the state into regions to be used in the final dataset, which is described in Section 5.

1. Data

The daily weather data from the Global Historical Climatology Network (GHCN) includes observations of minimum and maximum temperatures, snow depth and snowfall. Measurements for this dataset were taken from 461 stations located throughout Wisconsin for the years 1997-2020. Much of this data was missing, including 32% of the snow depth data and 43% of the minimum temperature data.

In addition to weather data, land use data from the Multi-Resolution Land Characteristics Consortium (MRLC) was used. The land use data contains land use statistics for the entirety of the US for the years 2001, 2004, 2006, 2008, 2011, 2013, and 2016 in the form of raster files. The land use data was prepared by geocoding the coordinates of each county, and then taking a square image from the original rasterfile that had a center at the coordinates of the respective county and a side length of 1000 pixels. Each county and year of the weather and fawn-to-doe ratio data was then matched with the nearest year of the land use data. The fawn-to-doe ratio (FDR) was used as a metric to measure the dynamics of the deer population. The FDR measured the ratio of the number of fawns to the number of does after a given winter.

Figure 1 shows a plot of the fawn-to-doe ratio for each county from 2017 to 2020. The FDR plots use a white to red scale to plot the magnitude of the FDR and and the blue coloring represents a missing value. There is no clear pattern seen throughout the yearly FDR plots. Therefore, there is a fair amount of noise within the FDR data since the measurements are only samples, and some of the sample sizes are quite small. Of 1921 observations within the data, 24.36% had a fawn sample size of 100 or greater and 25.86% had a doe sample size greater than 100. Additionally, 2.75% of the data contained a doe value of 0; 1.4% of observations also had an FDR of 0, which indicates the the fawns value was 0. Due to the noise within the data and the amount of missing values, we chose to filter the FDR data for quality in our original and experimental models.

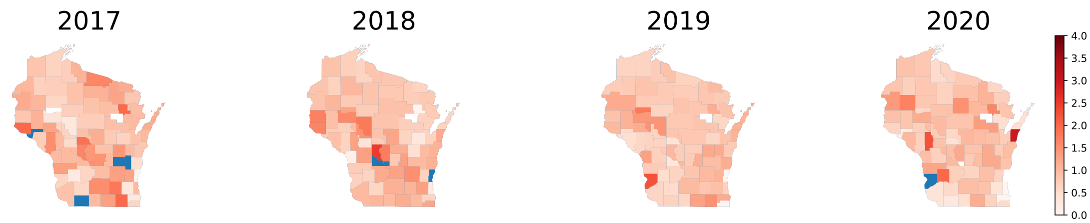

Figure 1: Yearly FDR plots. The red coloration represents the magnitude of the FDR, while a blue color represents a missing value

2. Original Model

The Winter Severity Index (WSI) is defined as the sum of the number of days with temperatures below zero degrees Fahrenheit and the number of days where snow depth is greater than 18 inches. Using GHCN data, we calculated the WSI for each county and compared it to the WDNR's WSI for the years 2014–2017. To avoid including stations with a significant amount of missing data, we chose one station to represent each county for each year based on the amount of missing snow depth data. For our calculations, a day below zero was defined as a day with the minimum temperature below zero. Figure 2 compares the two WSI's for 2014 with each dot representing the value for a county. The blue dots represent the WSI from the WDNR, the orange dots represent calculated WSI and the black line is a line of best fit. While the two statistics are not the same, it appears that they have similar values and follow similar trends. This is confirmed by a root mean squared error of 22 days for 2014-2017.

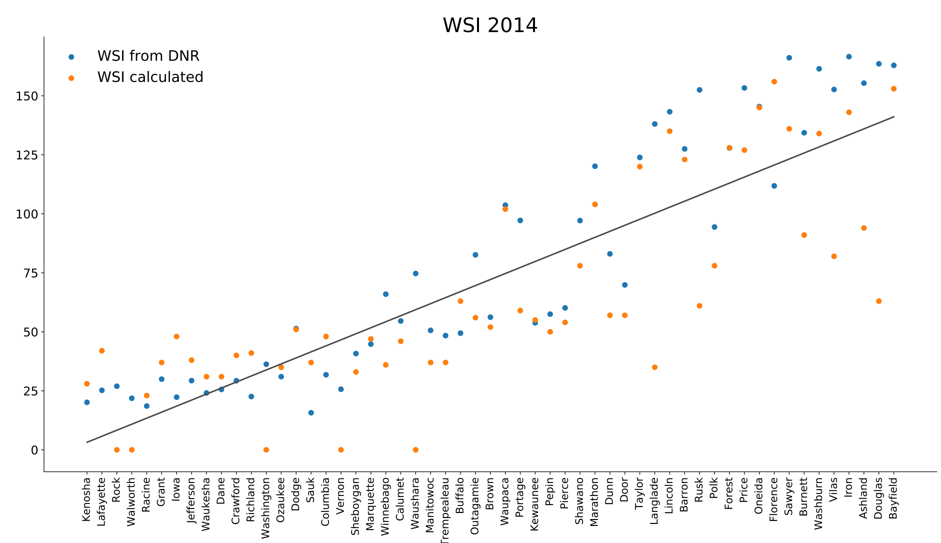

Figure 2: WSI for 2014

To avoid using data with limited or incorrect observations, we placed restrictions on the dataset so that only years with an observed number of does greater than 100 and with a fawn-to-doe ratio less than two and not equal to zero were considered. We used linear regression with a 75%, 25% train, test split to predict the fawn-to-doe ratio from the WSI and found that this explained 1.77% of the variance for the training data and 3.15% for the testing data, which is less than expected.

3. Additional Modeling

In an effort to improve the predictive power of our regression model we investigated the effect of including land use and location statistics in our model. Figure 3 shows a comparison of various models, which used different combinations of land use, location, and weather data.

Figure 3: Model Comparisons

The data used in each iteration underwent a 75%, 25% train, test split. The new model was first trained and scored using only land use data for each county's image sample. This model explained 10% of the variance in the training data and 10.8% of the variance in the testing data. After training the model solely on the land use data the calculated WSI metrics were added in. Using the WSI metrics and land use data, the model explained 12% and 12.7% of the variance in the training and testing data respectively. We also added the latitude to the WSI and land use data, which explained 12.4% of the variance in training data and 12.7% of the variance in testing data. We then divided the original 1000 by 1000 land use image samples into 100 by 100 images to compute percentiles and to compute the land use category which was the maximum value in each division. We tried adding the 10^{th}, 25^{th}, 50^{th}, 75^{th},and 95^{th} percentile of each land use category and the number of times a land use category was a maximum in a division to the model, which accounted for 14.7% of the variance in training data and 8.7% of the variance in testing data. The best model was the model which used the Land Use, WSI, and Latitude, explaining 12.7% of the variance. Since the land use data's R squared Score was far better than the original model - 3.15% vs. 10.8% - we decided to try clustering the data using land use rather than weather.

4. Clustering

To reduce the amount of missing or less reliable data included in the models, we split the state into regions so that we could consider data from several counties at once. Having a greater amount of reliable data may make regional data preferable to county data for machine learning.

We used k-means clustering to split Wisconsin into 7 regions based on location and land use in each county. This clustering consists of northern, central, western and eastern regions, a region containing Waukesha, Brown and Dane counties and a region containing Milwaukee County. Figure 4 shows a map of Wisconsin with each color representing a different region.

To determine the optimal number of regions we looked at inertia, which gives a measurement of the similarity between data points in a region. Inertia is plotted in Figure 5. The plot shows that clustering using 7 regions groups counties that are significantly more similar than clustering using fewer regions. Along with the inertia plot, we were looking to group urban and rural areas together and have most regions contain multiple counties. Groupings with greater than 7 regions tended to be more sporadic with respect to location and contained fewer counties per region.

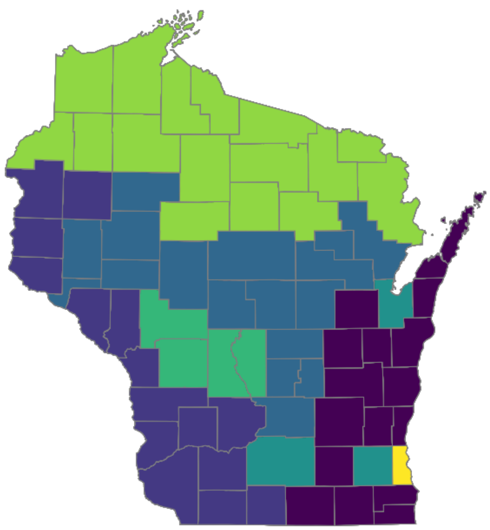

Figure 4: Wisconsin split into regions based on land use and location

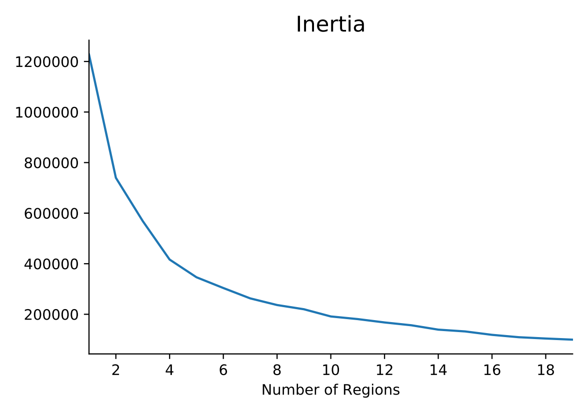

Figure 5: Inertia Plot

5. Final Dataset

The rows of the final dataset are organized by region and the columns consist of yearly weather data, land use data, fawn and doe counts, the fawn-to-doe ratio and overall deer population. Weather data was taken from the GHCN dataset and generalized from station to county by taking the median observation for stations within a county for each day. From this we calculated yearly statistics for the final dataset, which include the number of days with snow depth and snowfall greater than a certain amount and minimum and maximum temperatures below a certain point. The fawn-to-doe ratio was computed by summing the number of fawns and number of does in each region and then taking their ratio. Moreover, the population was calculated by summing the deer population for each county, in each respective region. A dataset with land use data organized by county and year was used to compute land use statistics for the final dataset. The county dataset has columns with averages, percentiles, and the number of times a land use category achieves a maximum value in a 100 by 100 division of the original 1000 by 1000 pixel sample. The columns without a decimal next to the use code are averages and the columns with a "max" next to their name are a maximum. For example, "Open Water" is an average and "Open Water_.5" is a percentile, while "Open Water_max" is a maximum. Average and Percentile columns were computed by averaging and max values were calculated by summation over each respective year and county within a region. In a future study one could look into computing land use statistics from an image containing the entirety of the clustered region. An image of the compressed table is included in Table 1, arbitrary temperature thresholds, snow depth thresholds, and land use statistics were picked since there are too many columns included within the dataset to include them all in this picture. Additionally, regression models were not built for this dataset due to time constraints, a future study could look into implementing machine learning models using different combinations of snow and temperature thresholds and land use data.

In addition to the final dataset, a dataset containing the counties within each cluster was created, and a dataset containing the computed WSI, Land Use, and Population data, organized by county and year, was produced as well.

Table 1: Compressed final dataframe imported as an image

	Region	Year	tmax<15F	snwd>=10in	Cultivated Crops_0.5	Fawns	Does	FDR	Population
0	0	2003	10.0	14.0	1.896291	752.0	666.0	1.129129	85867.0
1	0	2004	10.0	6.0	1.896291	716.0	673.0	1.063893	95984.0
2	0	2005	4.0	5.0	1.836879	581.0	476.0	1.220588	86101.0
3	0	2006	2.0	4.0	1.697104	497.0	473.0	1.050740	102050.0
4	0	2007	7.0	3.0	1.836879	652.0	541.0	1.205176	94404.0

Conclusion

In conclusion, we found that land use explained far more of the variance in the fawn-to-doe ratio than the Winter Severity Index. In turn we have provided a clustered dataset which includes weather data clustered by land use regions to aid in future efforts to improve the Winter Severity Index. Additional weather and geographic variables could also be included in a future study to improve the model.

References

[1] MLRC United States Land Use Data Website,
`https://www.mrlc.gov/data/nlcd-land-cover-conus-all-years`

[2] NOAA GHCN Weather Data Website,
`https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn`

[3] WDNR FDR Dataset.

Addendum

The addendum contains supporting information for the report. Figure 6 contains plots of the fawn-to-doe ratio for each year from 1997-2020, it can be read in the same fashion as Figure 1. Figure 7 contains plots illustrating the WSI for 2014-2017.

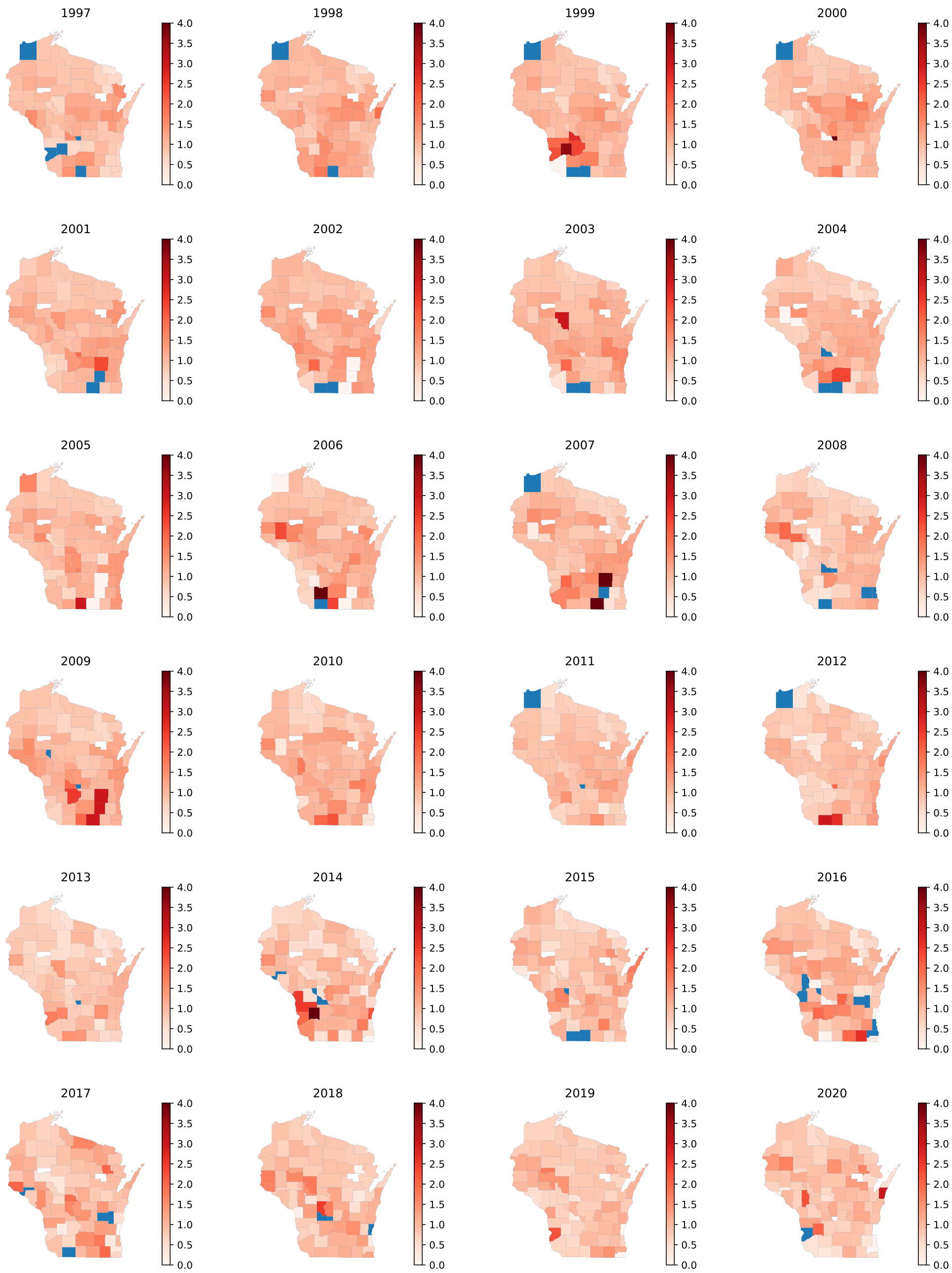

Figure 6: Yearly FDR plots. The red coloration represents the magnitude of the FDR, while a blue color represents a missing value

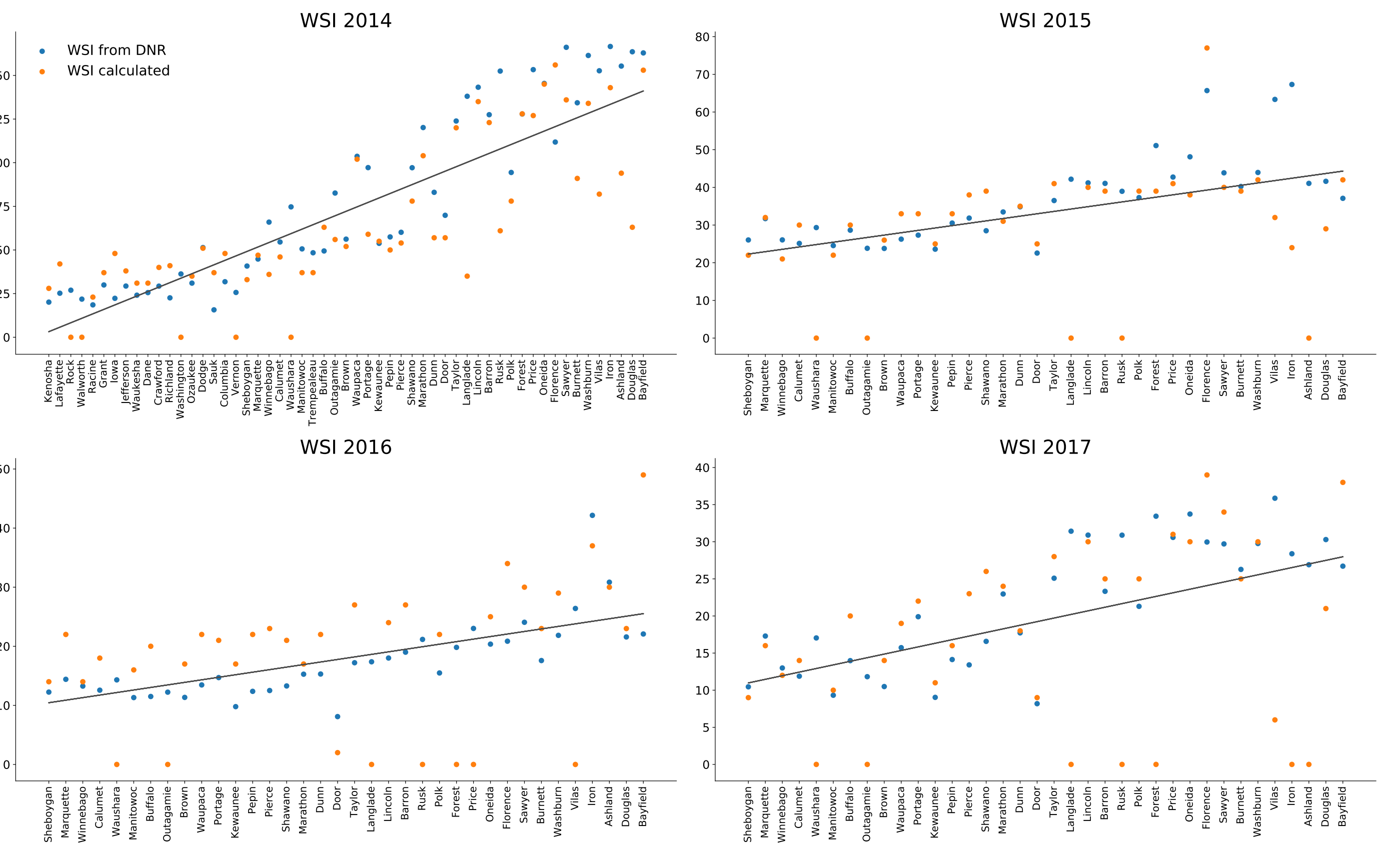

Figure 7: WSI from 2014 to 2017