The Analysis of Reptile and Amphibian Observations in Los Angeles County

Timothy Stegman

December 2020

Contents

1. Introduction

1.1 Project Description

The purpose of this project is to analyze reptile and amphibian observations that have been collected in Los Angeles County through the iNaturalist initiative. iNaturalist is a social networking service that brings citizen scientists, naturalists, and academics together to crowdsource the observation and identification of living organisms across the globe. In this analysis, we will look at the research grade observations that have been collected for the classes Reptilia and Amphibia in Los Angeles County from Oct-2015 through Sep-2020. The Random Forest classification algorithm will be utilized in an effort to predict common name groupings of organisms based upon seasonal, geographic, and weather features.

1.2 The Data Sets

Three data sources were used in this analysis:

1. **GBIF_iNaturalist.csv** - Research grade observations from iNaturalist
2. Files of the form **NOAA_mmmyy_through_mmmyy.csv** - Climate data from the National Oceanic and Atmospheric Administration (NOAA)
3. **common_name_groupings.csv** - Organism groupings

GBIF_iNaturalist.csv:
Most of the research grade iNaturalist observations have been ingested at the Global Biodiversity Information Facility (GBIF), which is an online repository with free and open access to biodiversity data. On 06Nov2020, a query was run at GBIF.org to extract iNaturalist research grade observations for Los Angeles County, for classes Reptilia and Amphibia, and for all time. This raw data has 25,460 observations and 50 variables. Of the 50 variables, only the following 5 were used for this analysis:

eventDate – character variable which captures date and time (PDT)
family – character variable which indicates the scientific family that the organism belongs to
month – variable indicating the month of the observation as an integer from 1-12
decimalLatitude – numeric variable which gives an estimate of the latitude of the observation
decimalLongitude – numeric variable which gives an estimate of the longitude of the observation

NOAA data sets:
Three data sets were pulled from the NOAA's National Centers for Environmental Information (NCEI) website. The NCEI site offers a climate data search tool which allows for the querying of historical weather data by region. For this project, daily weather summaries were queried by station for the time period of 01Jan2015 – 31Oct2020 for Los Angeles County. Due to size limitations, three separate extracts were run from the website, and the following csv files were created:

NOAA_jan15_through_dec16.csv
NOAA_jan17_through_dec18.csv
NOAA_jan19_through_oct20.csv

Each data set consists of 16 variables, and combined they have a total of 150,336 observations. However, for this analysis, only the following 7 variables were kept:

STATION – character variable indicating the weather station
DATE – character variable which captures the date of the weather summary

AWND – numeric variable which captures the average daily wind speed in miles per hour
PRCP – numeric variable which captures daily precipitation in inches
TAVG – integer variable which captures the daily average temperature in °F
TMIN – integer variable which captures the daily minimum temperature in °F
TMAX – integer variable which captures the daily maximum temperature in °F

common_name_groupings.csv:
For the purposes of this exercise, a custom csv file was created which maps the organism families to the following common name groupings:

1. Frogs
2. Lizards
3. Salamanders
4. Snakes
5. Turtles

This file consists of the following 4 character variables:

class – indicates the scientific class the organism belongs to
order – indicates the scientific order the organism belongs to
family – indicates the scientific family the organism belongs to
grouping – indicates the common name grouping the organism belongs to

The last variable, grouping, is what we will attempt to predict using the seasonal, geographic, and climate features found in the iNaturalist and NOAA data sources.

2. Transformations of the Raw Data

2.1 Transformations of the iNaturalist Data

1. After importing the iNaturalist data set, variable names were converted to lower case.
2. The eventdate variable, which stores the date and time of the observations as a character variable, was converted to class POSIXct.
3. Once converted to class POSIXct, the hour and date were extracted from eventdate.
4. The hour and month variables were cast as factors.
5. The data set was trimmed down to include only the variables of interest.
6. Observations were restricted to the time period of 01Oct2015 – 30Sep2020.
7. The final iNaturalist data set was inspected for missing values and was found to be complete.

2.2 Transformations of the NOAA Data

1. After importing the 3 NOAA csv files, the data sets were combined into one data frame.
2. The variable names were converted to lower case.
3. The date variable was converted from character to date class.
4. The variables that were out of scope for the analysis were excluded.
5. The data was inspected for missing values. Many missing values were found, and it appears that certain weather measurements were not captured at particular stations.
6. Since the aim of this project is to utilize daily aggregate weather statistics across Los Angeles County, the decision was made to ignore the missing values, and compute the median values per day for each

of the weather variables. This process eliminated the missing values, and also provided a summary of the weather variables across the county for each day.

2.3 Transformations of the Common Name Groupings Data Set

For the common name groupings data set, the grouping variable was converted to a factor variable.

2.4 Joining the Data Sets

A SQL inner join was used to create one combined data set containing the relevant columns from the iNaturalist, NOAA, and groupings data sets. Specifically, the NOAA data set was joined to the iNaturalist data set on date, and the common name groupings data set was joined to the iNaturalist data set on family.

3. Exploratory Data Analysis

3.1 Basic Summary Statistics

In Figure 1 below, we summarize the final joined data set. If we look at the grouping variable, we can see that the majority of the observations captured were of lizards. In fact, 69.45% of the observations fell within this common name grouping. So, when we look at the Random Forest classification algorithm later, we would like to find a model which has greater accuracy than the majority class classifier, which is the model that simply predicts that every animal observed is a lizard.

```
   family          decimallatitude decimallongitude     month           hour
Length:21179       Min.   :32.81   Min.   :-118.9   4      :5182   11     :2364
Class :character   1st Qu.:34.06   1st Qu.:-118.5   5      :3090   10     :2332
Mode  :character   Median :34.13   Median :-118.3   6      :2510   12     :2172
                   Mean   :34.13   Mean   :-118.3   3      :2145   13     :2016
                   3rd Qu.:34.20   3rd Qu.:-118.1   7      :1779   14     :1856
                   Max.   :34.82   Max.   :-117.6   9      :1700   15     :1680
                                                    (Other):4773   (Other):8759
   event_dt           event_dt_tm                       grouping          awnd
Min.   :2015-10-01   Min.   :2015-10-01 15:29:56   Frogs      : 1841   Min.   : 2.010
1st Qu.:2017-06-15   1st Qu.:2017-06-15 18:44:42   Lizards    :14709   1st Qu.: 5.140
Median :2018-10-05   Median :2018-10-05 10:56:11   Salamanders:  879   Median : 5.820
Mean   :2018-09-03   Mean   :2018-09-04 06:04:54   Snakes     : 2502   Mean   : 6.039
3rd Qu.:2019-10-24   3rd Qu.:2019-10-24 15:46:04   Turtles    : 1248   3rd Qu.: 6.710
Max.   :2020-09-30   Max.   :2020-09-30 14:09:26                       Max.   :15.880

     prcp              tavg            tmin            tmax
Min.   :0.00000   Min.   :36.00   Min.   :33.00   Min.   : 46.00
1st Qu.:0.00000   1st Qu.:59.00   1st Qu.:50.50   1st Qu.: 71.00
Median :0.00000   Median :65.00   Median :56.00   Median : 77.00
Mean   :0.01235   Mean   :65.16   Mean   :55.48   Mean   : 77.68
3rd Qu.:0.00000   3rd Qu.:72.00   3rd Qu.:61.00   3rd Qu.: 85.00
Max.   :2.13500   Max.   :95.00   Max.   :78.00   Max.   :109.50
```

Figure 1: Basic Statistical Summary

3.2 Observations by Grouping and Time Period

Since reptiles and amphibians are ectotherms, the temperature of the surrounding environment is an important factor in determining their behavior. Therefore, we would expect to see patterns in observations at both the monthly and hourly levels. Figure 2 below displays the percentage of total observations by month and by hour for each of the common name groupings.

We can see that for all of the groupings, the observations peak in spring. Also, with the exception of salamanders, the lows occur in winter months. This is likely due to the fact that salamanders are particularly sensitive to ambient moisture levels, and the rainy season occurs in the winter.

Regarding the hourly plots, as expected, most of the observations occur during daylight hours. However, we can see that frogs and snakes, which have some constituent species which are known to be nocturnal, tend to be observed at night more often than the other organisms.

Curiously, there are spikes in the hourly plots corresponding to hour=0 (midnight). This is due to the fact that some of the event dates in the iNaturalist data did not capture the time of the event, and so the time value was defaulted to midnight.

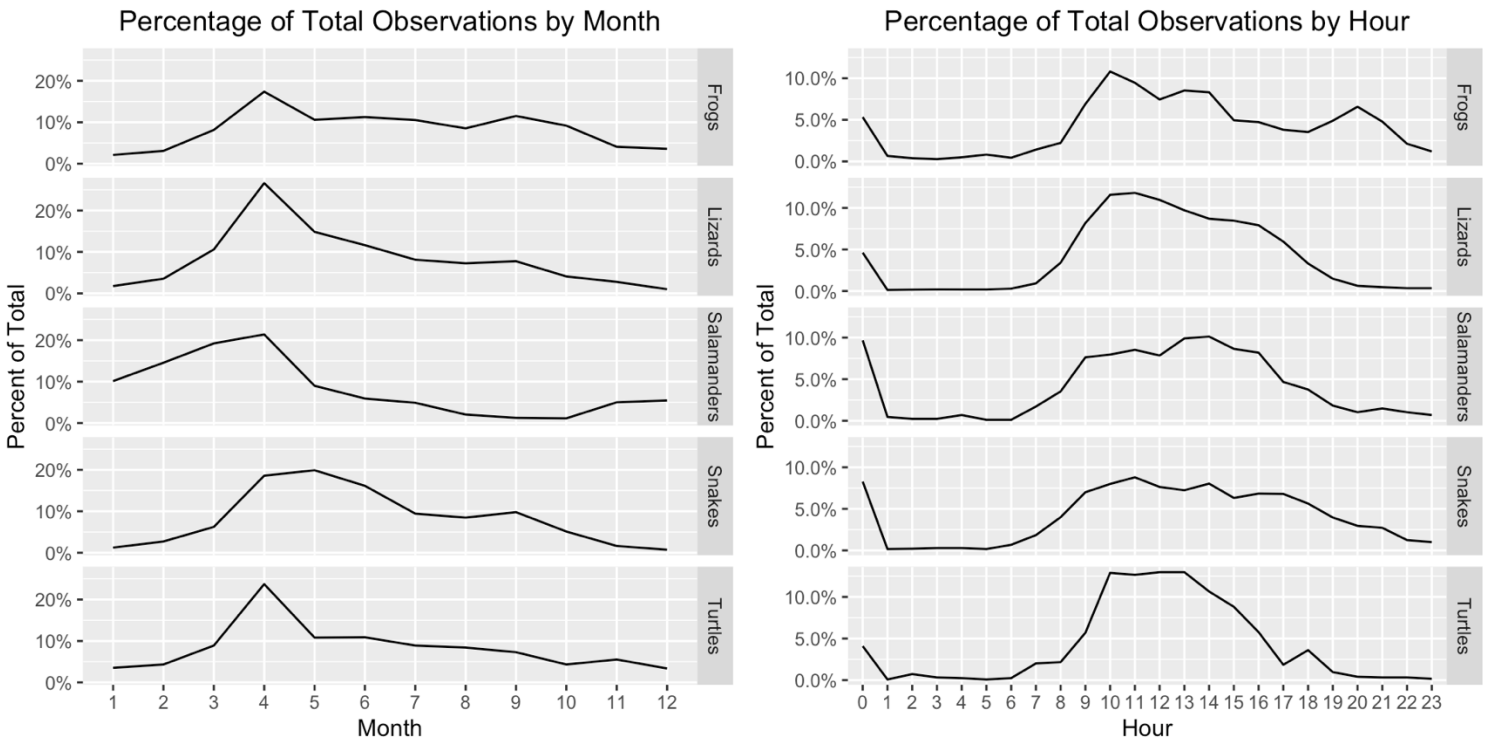

Figure 2: Percentage of observations by month and hour of day

3.3 Geographic Distribution by Grouping

Los Angeles County is home to a wide variety of geographic features. Within the county one can find coastal, montane, riparian, desert, and urban habitats, all of which can be home to reptiles and amphibians.

Figure 3 presents the geographic distribution of observations overlayed onto a map of Los Angeles county. In order to reduce point congestion on the plot, a random sample of 5,000 observations without replacement was used to generate this graph. From these plots we can see that lizards have been widely observed throughout the county, including in highly urbanized areas, while snakes are less frequently found in the most urbanized areas. Very few observations of frogs and salamanders have been logged in the desert areas in the northeast part of the county, which might be expected due to their moisture requirements. Another interesting piece of information revealed by this plot is that turtles were observed primarily in the urbanized region of the county. This is perhaps due to the fact that there are a number of small lakes and ponds scattered throughout the area where captive turtles have been released.

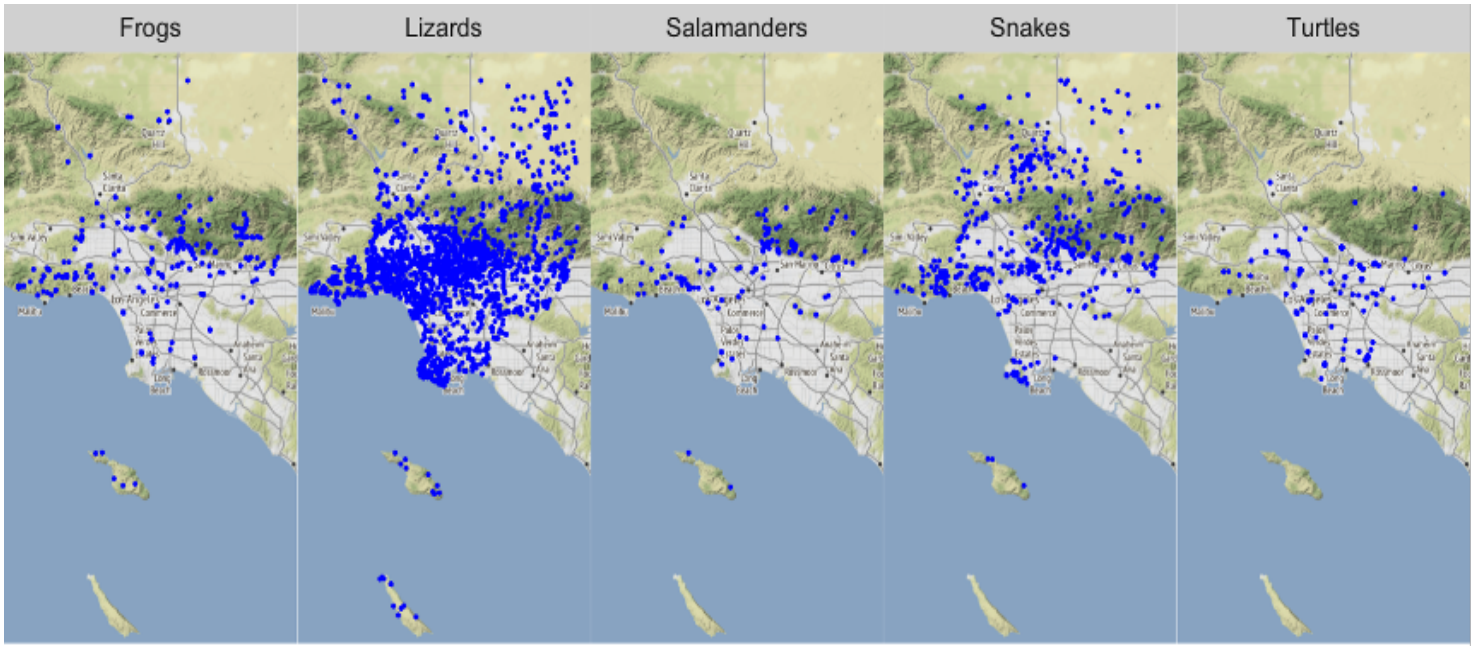

Figure 3: Geographic distribution of observations

3.4 Boxplots of Weather Statistics by Grouping

During the data transformation phase of the project, for each of the 5 weather variables, we computed the median value across all of the weather stations for each day. This was done in order to obtain a general picture of what the weather was like on any particular day in Los Angeles County.

Figure 4 presents boxplots to give us an idea of how these daily weather aggregates are distributed in relation to the observations within the different common name groupings. The distributions are generally symmetric, and the average wind speed and precipitation variables exhibit quite a few outliers. Based upon these plots, it appears that salamanders seem to be observed more often on cooler days when compared to the other organisms. It also appears that the different organism groupings have similar distributions in regard to wind speed, with salamanders exhibiting slightly more dispersion. Since the values of precipitation are so low (the 3rd quartile is zero), it is hard to differentiate between the groupings here.

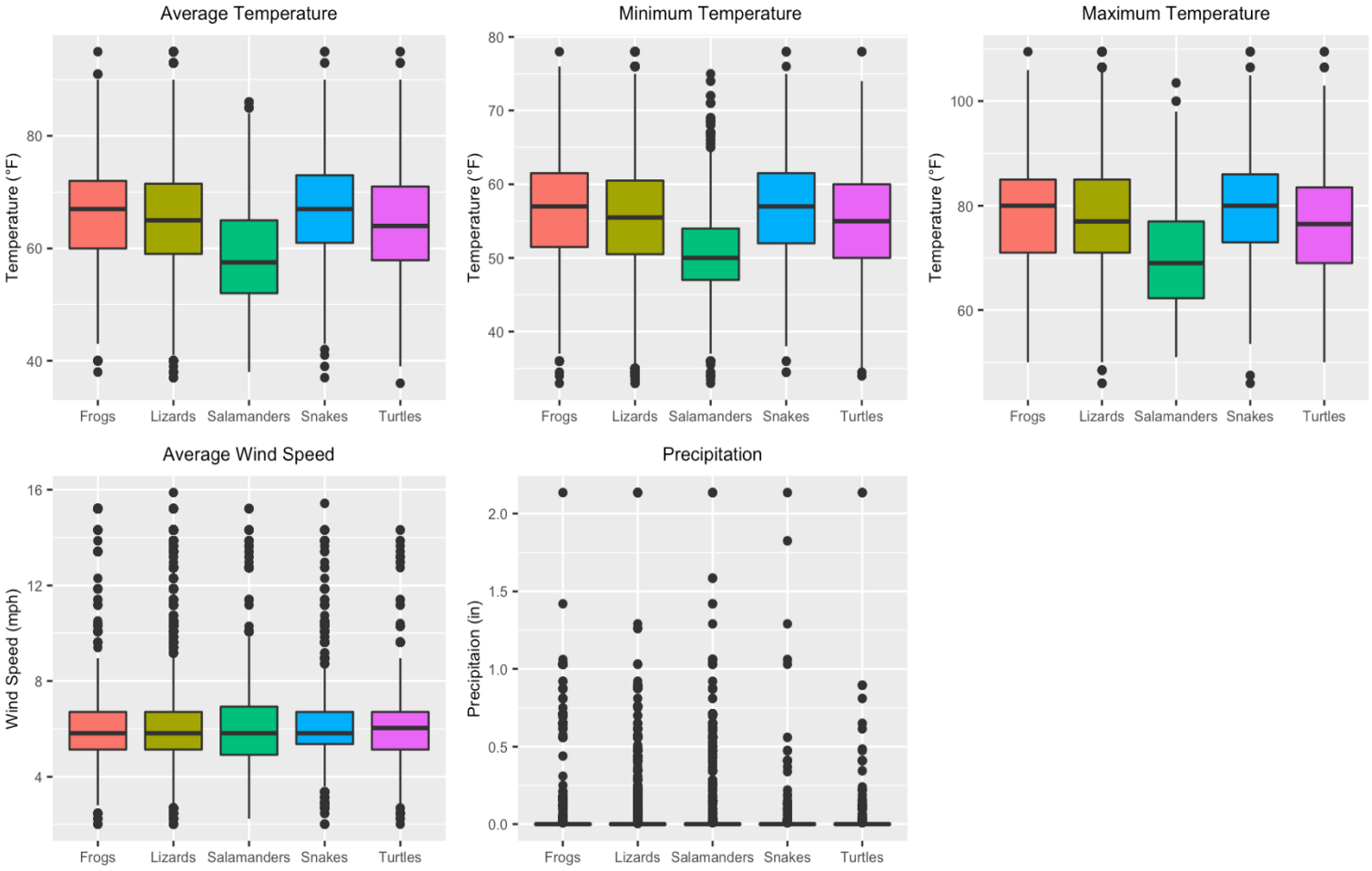

Figure 4: Boxplots of weather variables by grouping

3.5 Scatter Plot Matrix

Figure 5 below displays the scatter plot matrix along with the corresponding LOWESS smoothing curves for seasonal, geographic, and weather variables. A random sample of 1,000 observations without replacement was used here to make the graphs more readable.

We see strong linear correlations between the temperature variables (tavg, tmin, and tmax). The Pearson correlations between these three variables are all over 0.89, with the correlation between tmin and tmax being the weakest. This indicates that there may be some redundancy amongst these variables, and so tavg will be excluded from future modeling efforts, since it is tightly correlated with both tmax and tmin.

We also notice that the relationship between prcp and the temperature variables appears to be non-linear, with higher precipitation values corresponding to colder days. This is expected since the rainy season occurs during the winter in Los Angeles. Also, as expected, we see a non-linear relationship between month and temperature variables with the highs occurring in the summer months.

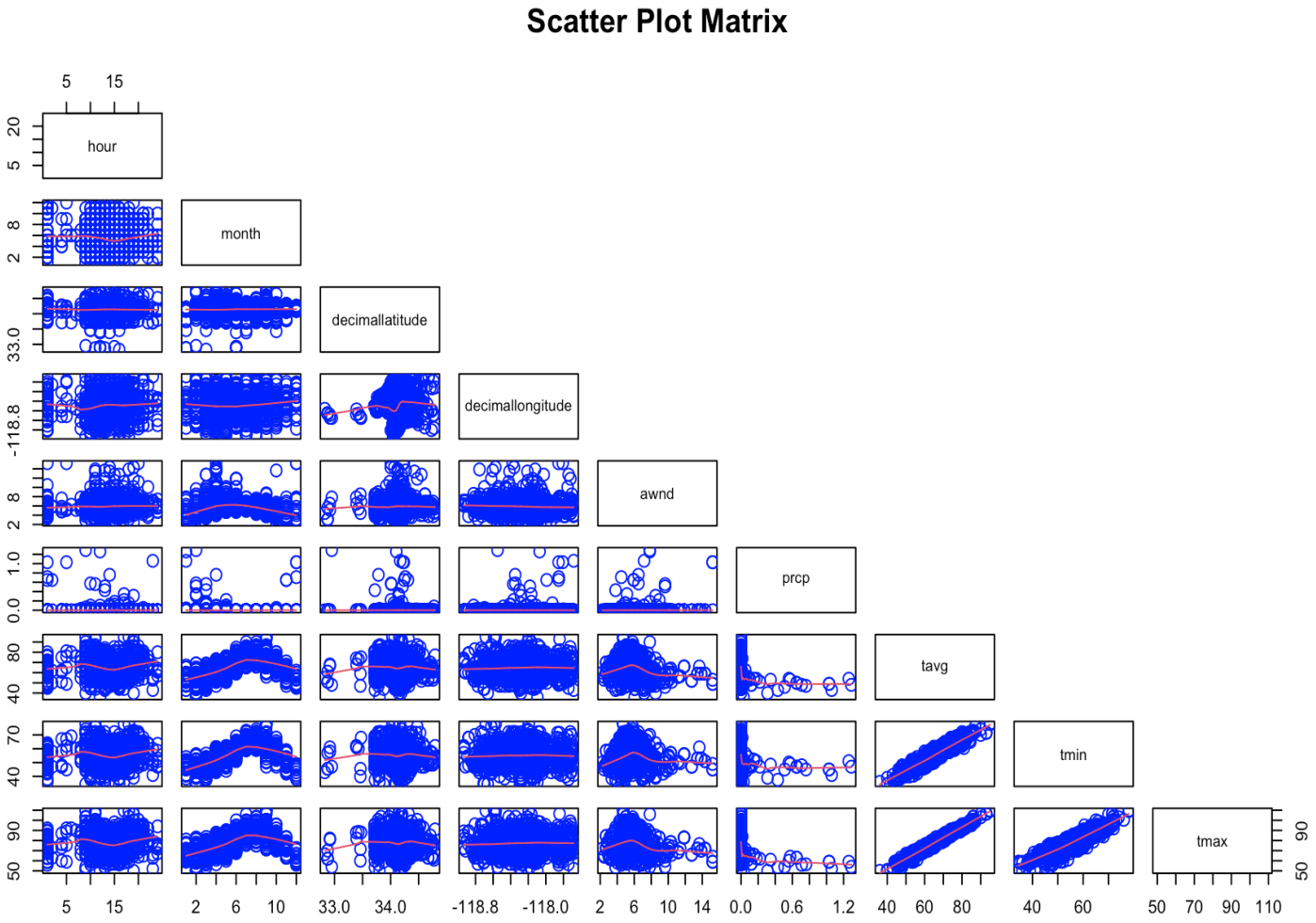

Figure 5: Scatter plot matrix

4. Machine Learning

4.1 Random Forest Classifier

In an attempt to predict the common name groupings, we utilized the Random Forest classification algorithm as implemented in the randomForest function from the R package of the same name. This is a popular method which looks at an ensemble of classification trees which are calculated on random subsets of the data, and where a random subset of features is used at each split in each tree. For this analysis, we used a 60/40 training/test split of the data. Also, when running the algorithm, we considered a variety of values for the number of trees to grow (ntree) in combination with the number of features to be sampled at each split (mtry). The model which produced the lowest test misclassification error rate was deemed to be the best.

To assess the feasibility of running many different combinations of ntree and mtry, the algorithm was first run with ntree=500 and mtry=5. Due to the fact that this took a long time to run, the algorithm was run again where the hour and month variables were converted from factors to numeric. This fit the model in one tenth the time, and with a very similar test misclassification error rate. So, the decision was made to use the numeric versions of these two variables for further modeling.

A total of 30 models were fit where the ntree value ranged from 100 to 1000 and the mtry value ranged from 3 to 5. Figure 6 shows the misclassification error rates calculated on the test data for these 30 models. The error rates seemed fairly stable and were all around 0.23. Also, the models with an mtry value of 5 tended to outperform the models with lower mtry values. Specifically, the model with ntree=900 and mtry=5 produced the lowest test misclassification error rate at 0.2292.

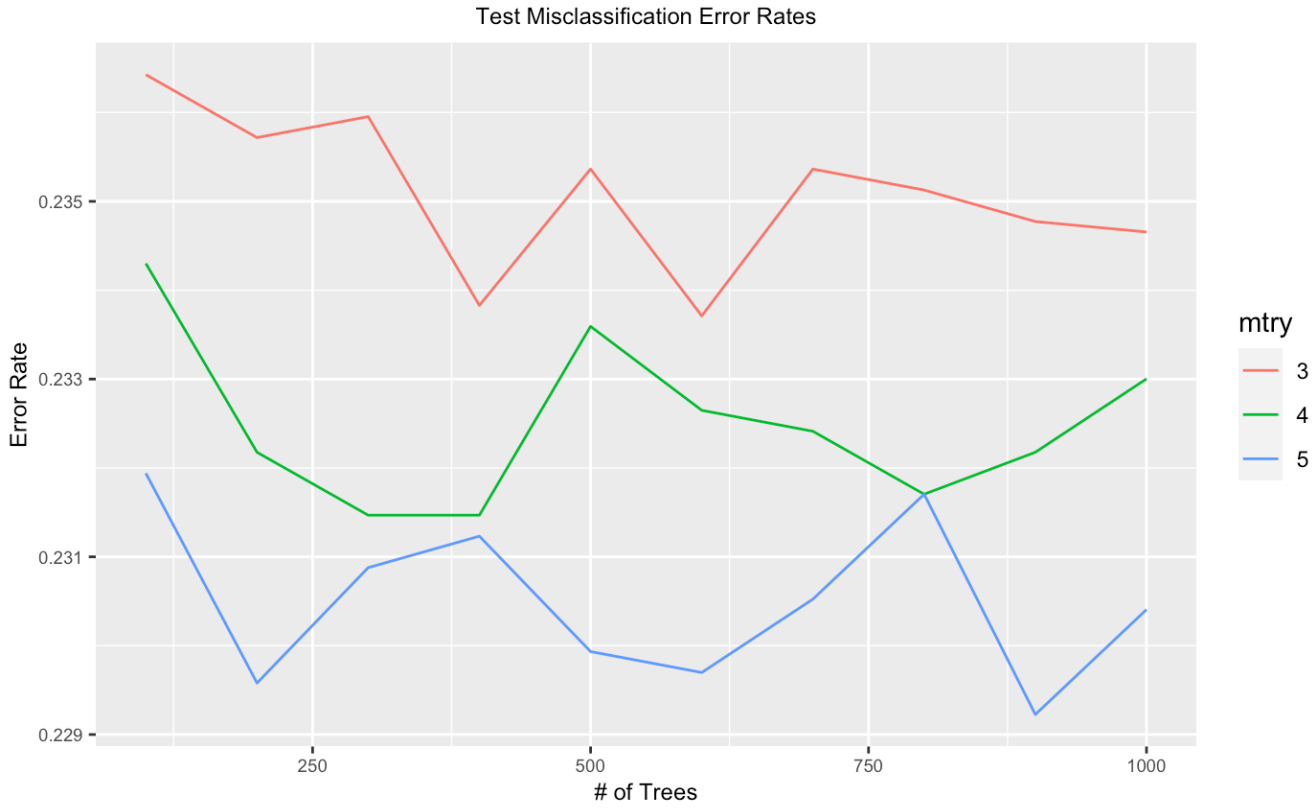

Figure 6: Test misclassification error rates by ntree and mtry

4.2 Variable Importance

Variable importance plots for the best model (with ntree=900 and mtry=5) were calculated on the full data set and are presented in Figure 7. Both the Mean Decrease Accuracy and Mean Decrease Gini plots produced the same ranking and found that the top four most important variables for predicting the common name grouping were decimallatitude, decimallongitude, hour, and tmax. They also found that the least important variable was prcp. In fact, if prcp is exluded from the model with ntree=900 and mtry=5, we get a test misclassification error rate of 0.2299, which is very close to the rate for the model that includes prcp.

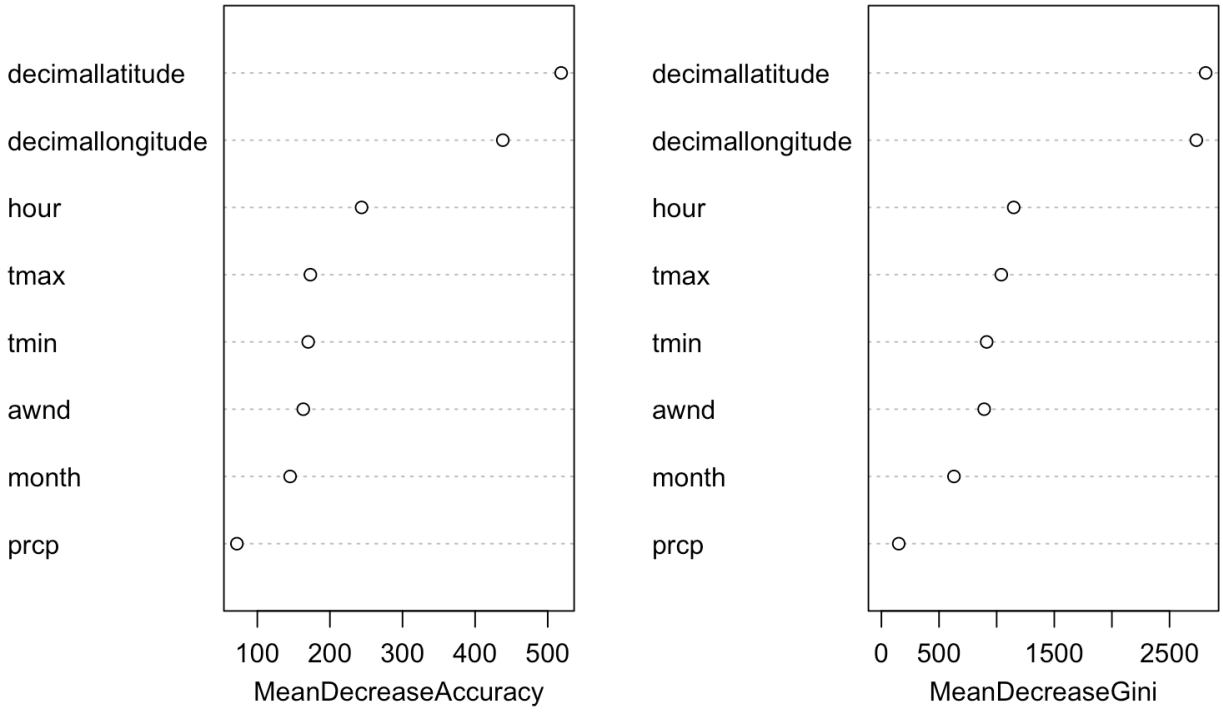

Figure 7: Variable importance plots

5. Conclusion

In this project, we were able to utilize the Random Forest classification algorithm in R to fit a model which predicted the common name grouping variable with a test misclassification error rate of 0.2292. While this model was not terribly accurate, it did beat the majority class classifier which predicts that every organism observed is a lizard and has a test misclassification error rate of 0.3055.

6. References

[1] iNaturalist.org

[2] GBIF.org (06 November 2020) GBIF Occurrence Download https://doi.org/10.15468/dl.vapdjg

[3] Menne, Matthew J., Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, Ron Ray, Russell S. Vose, Byron E.Gleason, and Tamara G. Houston (2012): Global Historical Climatology Network - Daily (GHCN-Daily), Version 3. FIPS:06037. NOAA National Climatic Data Center. doi:10.7289/V5D21VHZ 2020-11-01, 2020-11-04

[4] Gutierrez, D. D. (2015). Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R. United States: Technics Publications.

[5] Stebbins, R. C., McGinnis, S. M. (2012). Field Guide to Amphibians and Reptiles of California: Revised Edition. United States: University of California Press.

Citations for R packages used in this analysis:

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

[7] Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/

[8] Nicholas Tierney, Di Cook, Miles McBain and Colin Fay (2020). naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.6.0. https://CRAN.R-project.org/package=naniar

[9] G. Grothendieck (2017). sqldf: Manipulate R Data Frames Using SQL. R package version 0.4-11. https://CRAN.R-project.org/package=sqldf

[10] Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

[11] D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

[12] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18--22.