Exploratory Data Analysis for Waquoit Bay Dune Data

By Salvador Balkus - December 10th, 2018

Abstract:

Professor Davis' Exploratory Data Analysis class has analyzed data from sand dunes on Waquoit Bay, Cape Cod, Massachusetts in an effort to explain what factors control the distribution of plant species on the dune. This exploratory data analysis used R and Python to clean the data, generate descriptive statistics, visualize the data, and train a linear support vector machine to try and examine if the growth of different plant species could be modeled mathematically. The study found that maximum wind speed and salt spray seemed to be the most predictive factors - most plants preferred to grow in areas with lower amounts of these variables. In addition, the linear support vector machine confirmed that plants' specific tolerance of one environmental variable depended on other environmental variables to which it was exposed. Therefore, to identify regions in which plant can and cannot tolerate, a machine learning model must be utilized.

Introduction:

Ecological data on sand dunes has been gathered from Waquoit Bay, Cape Cod, Massachusetts by Dr. Rajaniemi, a biology professor here at the University of Massachusetts, Dartmouth, and has been presented to Professor Davis' Exploratory Data Analysis class to be analyzed. The data was collected by dividing the dune into transects and taking measurements of plant species, environmental variables, and microbes at 10-meter intervals from the start of the dune. This analysis focused on three questions posed by the client:

> What factors control the distribution of the plant species in the dunes?
> Do the smaller shrubs ameliorate stressful conditions for the smaller plants?
> What specific conditions are they able to tolerate?

In order to answer these questions, exploratory data analysis was used to examine the data set, with the goal of discovering which conditions were most critical to plant growth on the dune. In addition, visual and mathematical models were constructed to aid in the prediction of plant growth on the dune. These exploratory methods create insight into what conditions are optimal for plant growth.

Methods:

The data in this analysis came exclusively from the Excel spreadsheet of Waquoit Bay dune data provided to the class by Dr. Rajaniemi. Specifically, this analysis focused on two parts of the data: first, the regular transect measurements, which recorded in a pivot table the location and amount of cover for each occurrence of plant species found on the dune; and second, the regular environmental data, which included many variables such as wind speeds, soil moistures, light, temperature, and soil salinity.

This exploratory data analysis was primarily performed in R using the RStudio integrated development environment and several packages from Hadley Wickham's "tidyverse" package collection, including "ggplot2" and "dplyr." In addition, Python was used in Jupyter Notebook to utilize the Linear Support Vector Machine from the "sci-kit learn" Python library, which is used to train and test machine learning models.

To begin the analysis, the data was first cleaned in R. A function was created to transform the species pivot tables into a data frame by reorganizing the species name, plant cover, distance, transect, and groups into individual columns. Then, this data frame was merged with the environment data frame, so that each plant record contained associated environment information for its location. This merge included only environment variables that were not year-specific. For the parts of the analysis that required no missing environmental values, the variables were interpolated by taking measurements from points of the same distance away from shore. This interpolation method was used because it is reasonable to assume that similar distances from the shore would have similar environmental data.

After the data cleaning, the data frame was grouped by species, and the count, median distance from shore, and average of common environmental variables were calculated for the top 8 species that occurred the most often. This part of the analysis used only the data where a plant was actually recorded to have grown. In addition, bubble plots were produced, modeling both the plant locations on the dune by species and how certain variables affected the amount of the transect that the plant covered.

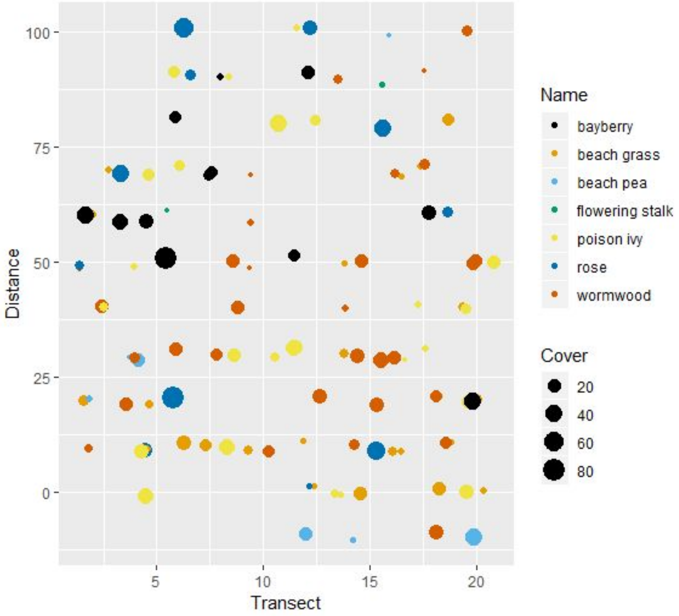
Afterward, the data was pared down to only the species data from 2006, and the additional environmental variables from 2006 were added to the data frame to include more data in the analysis. Logistic regression was used to analyze which variables were most significant in determining whether a plant would grow at a certain location on the dune. For this, a data frame was used with "cover" expressed as a binary value rather than a percentage of the transect. This was to focus the analysis on whether or not a plant would grow in a certain area. The logistic regression was performed on the entire data frame, as well as subsets of specific species, in order to analyze if the significant variables changed when only one species was modeled.

Lastly, machine learning was used to examine whether or not points where species grew appeared in a cluster that was easily distinguishable from the points where plants did not grow. The accuracy of the algorithm determines how predictable the plant growth is, and reveals how many variables ultimately determine where the plant will grow. A linear support vector machine was used for this, which is a machine learning algorithm that draws a straight line to divide the points into two groups. One group contained points where plants grew, and the other contained points where plants did not grow. This algorithm was trained using 70% of the data and its accuracy was tested using 30%. The training and testing sequence was performed 1000 times, and an average of the accuracies was taken to determine the accuracy of the algorithm.

Results:

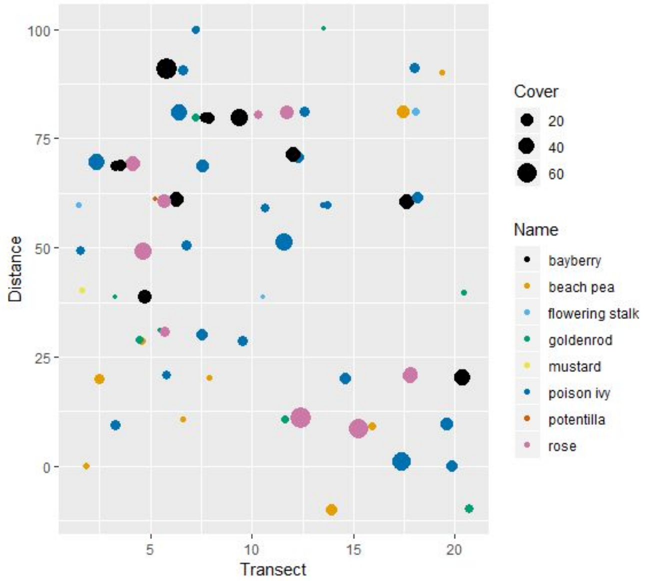
The initial mapping of the top 8 plant species, while an interesting visualization, demonstrates that there are few obvious patterns to the distribution of the most numerous species on the dune. There are clearly spots that are not suitable for plant growth (the empty areas where no plants grow), and certain plants prefer certain areas - for instance, bayberry prefers to be away from the shore.

Figure 1: A jittered bubble plot mapping the locations of the eight most common species and their cover size for all years.



To determine whether large plants shielded small plants from harsh conditions, a similar bubble plot pictured in Figure 2 was produced using eight species with at least five occurrences - four large (rose, bayberry, poison ivy, and potentilla) and four small (beach pea, flowering stalk, goldenrod, and mustard). From Figure 2, it appears that large plants do not shield smaller plants from harsh conditions, as a considerable number of small plants grow closer to the shore, with no large plants in front of them. However, like the previous plot, it is worth noting that there exist large swathes of the dune where none of the common species grow at all.

Figure 2: A jittered bubble plot mapping the locations of the four large species and four small species that occurred in at least five instances for all years.



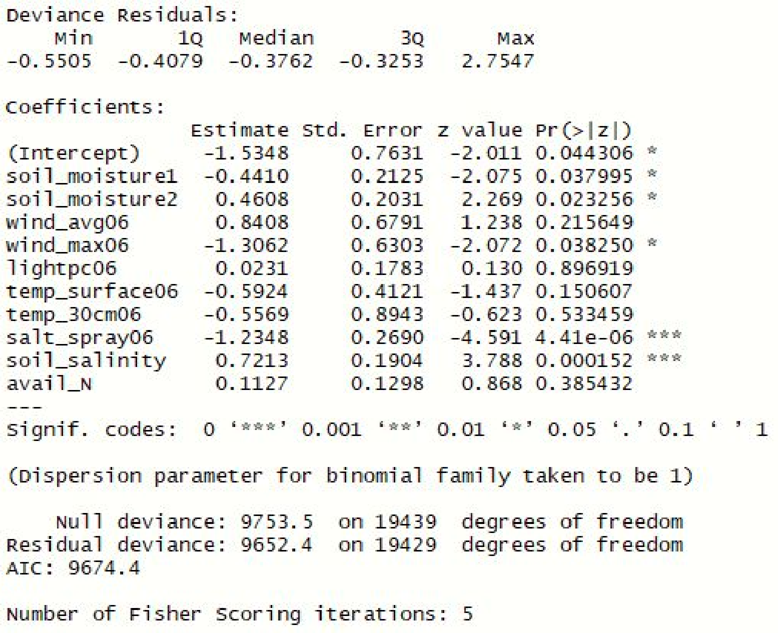
The results of the grouped environmental variable aggregations are pictured in Figure 3. These values help to demonstrate some of the preferred conditions for the more common species for the variables that were not year-specific. For instance, bayberry's low value for mean salt spray gives an indication that it has a much lower salt spray tolerance than the other plants.

Figure 3: Aggregate values for top 8 plant species and their most significant environment variables in the locations where they occurred.

Name	Counts	Median Distance	Mean Soil Moisture 1	Mean Soil Moisture 2	Mean Max Wind Speed	Mean Salt Spray	Mean Soil Salinity
beach grass	60	35	1.985	1.92	4.245417	50.1165	10.6635
wormwood	58	45	1.717241	1.825862	4.96569	50.69017	7.810345
poison ivy	54	50	1.866667	1.961111	3.286111	43.08481	9.256667
rose	23	40	2.065217	2.034783	3.221739	20.64261	9.291304
bayberry	22	69.5	2.822727	2.668182	1.728636	8.766818	10.20545
beach pea	20	9	1.405	1.25	3.9515	60.398	13.305
goldenrod	19	21	1.984211	2.026316	3.927895	56.04368	10.71158
flowering stalk	15	69	2.613333	2.866667	4.990333	56.78533	7.588

In order to test the significance of these values and the variables specific to 2006, logistic regression was used with scaled variables for the 2006 species data. Variables were scaled by dividing by the maximum value of the variable. A logistic regression was calculated for all of the data points, as seen in Figure 4. This R output indicates that, in general, plants tend to prefer areas with minimal salt spray, which explains why more plants prefer living farther from the shore, away from stressful conditions. Salt spray was found to be both statistically significant and very predictive compared to other variables. Surprisingly, in general, plants actually preferred growing in areas with high amounts of soil salinity, as this variable was found to be significant and somewhat predictive. The different soil moistures were also found to be significant, and maximum wind speed was also significant and wound up the most predictive variable of all - plants very much prefer growing in areas with low maximum wind speeds.

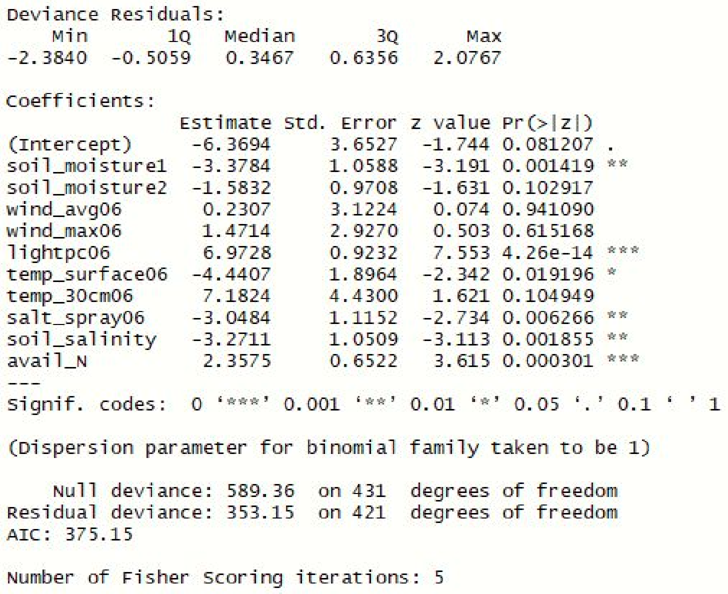
Figure 4: Logistic regression output from R for all plant species.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5505  -0.4079  -0.3762  -0.3253   2.7547

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.5348     0.7631  -2.011 0.044306 *
soil_moisture1  -0.4410     0.2125  -2.075 0.037995 *
soil_moisture2   0.4608     0.2031   2.269 0.023256 *
wind_avg06       0.8408     0.6791   1.238 0.215649
wind_max06      -1.3062     0.6303  -2.072 0.038250 *
lightpc06        0.0231     0.1783   0.130 0.896919
temp_surface06  -0.5924     0.4121  -1.437 0.150607
temp_30cm06     -0.5569     0.8943  -0.623 0.533459
salt_spray06    -1.2348     0.2690  -4.591 4.41e-06 ***
soil_salinity    0.7213     0.1904   3.788 0.000152 ***
avail_N          0.1127     0.1298   0.868 0.385432
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9753.5  on 19439  degrees of freedom
Residual deviance: 9652.4  on 19429  degrees of freedom
AIC: 9674.4

Number of Fisher Scoring iterations: 5
```

This logistic regression was also performed for several individual species, most of which had somewhat similar findings to those of the regression with all of the species. However, some, like wormwood in Figure 5, had additional significant variables. Salt spray and soil moisture 1 still contributed negatively to their growth; however, wormwood, in particular, responded negatively towards soil salinity and positively towards availability of nitrogen and light availability, both variables which were very statistically significant. It also responded negatively towards surface temperature.

Figure 5: Logistic regression output from R for wormwood



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3840  -0.5059   0.3467   0.6356   2.0767

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -6.3694     3.6527  -1.744 0.081207 .
soil_moisture1  -3.3784     1.0588  -3.191 0.001419 **
soil_moisture2  -1.5832     0.9708  -1.631 0.102917
wind_avg06       0.2307     3.1224   0.074 0.941090
wind_max06       1.4714     2.9270   0.503 0.615168
lightpc06        6.9728     0.9232   7.553 4.26e-14 ***
temp_surface06  -4.4407     1.8964  -2.342 0.019196 *
temp_30cm06      7.1824     4.4300   1.621 0.104949
salt_spray06    -3.0484     1.1152  -2.734 0.006266 **
soil_salinity   -3.2711     1.0509  -3.113 0.001855 **
avail_N          2.3575     0.6522   3.615 0.000301 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 589.36  on 431  degrees of freedom
Residual deviance: 353.15  on 421  degrees of freedom
AIC: 375.15

Number of Fisher Scoring iterations: 5
```

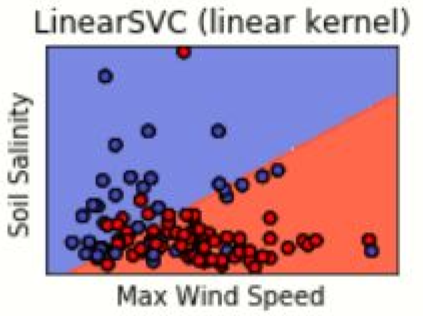
Finally, to examine the predictability of the growth of each plant species, a linear support vector machine was trained on the data for each species of plant. However, the algorithm was only able to successfully identify a region of vegetated points for four of the species. Figure 6 shows these species and the accuracy of the linear SVM at dividing out the vegetated points for each.

Figure 6: Species whose growth was able to be modeled by linear SVM, and associated accuracy

Name of Species	**Accuracy of Linear SVM**
Bayberry	~95%
Wormwood	~80%
Poison Ivy	~69%
Beach Grass	~65%

The linear SVM was trained on the same variables as the logistic regression. However, to help visualize how the algorithm actually divided out the points, a scatterplot for wormwood is pictured in Figure 7 with just two variables. The line generated divides the graph into a section of “vegetated” (red) points and non-vegetated (blue) points.

Figure 7: Demonstration of linear SVM for wormwood with two variables. The red region predicts where on the graph all of the vegetated (red) values should be located, and blue region predicted the non-vegetated (blue) values.



Discussion:

As indicated by the logistic regression, the factors controlling plant growth vary between species - for instance, some plants prefer saline soil while others avoid it. However, few plants can tolerate high levels of salt spray, which is why more plants tend to grow away from shore. It is also apparent that most species tend to do better in areas with lower maximum windspeed. Plant species tend to not be affected by average winds, but even the hardiest common species, wormwood and flowering stalk, tend not to be able to tolerate wind speeds above 5 mph. Thus, maximum wind speed and salt spray are the most predictive factors for plant growth on the dune.

From the visualizations, it appeared that larger plants did not ameliorate conditions for smaller plants behind them. However, the plot did show bare areas of the transects, which are likely areas of high maximum wind speed that made it difficult for any plant to grow there. If additional analysis could be performed on how maximum wind speed affected plant species, a mathematical model could be constructed that might explain how wind affects growth. For instance, it could be possible that wind carries plant seeds over the areas with high maximum wind into areas with low maximum wind, where they settle into the soil and grow. A model may be able to be used to figure out if the wind moves plant seeds between areas of low wind over time.

The question of what specific conditions a plant can tolerate cannot easily be answered with a simple numerical value. This is because a plant’s tolerance of one variable might change depending on its other conditions. For example, a plant might be able to tolerate higher wind speeds if it grows in an area with more light. Therefore, some sort of model must be created to define what conditions a plant can withstand. Machine learning is able to define the complex boundaries where plants can and cannot grow in very high dimensions. In this analysis, a line was drawn through eleven dimensions to try and define where each type of species could grow. Four plants were found to be predictive enough to be modeled using linear SVM: bayberry, wormwood, poison ivy, and beach grass. These species have more predictable relationships between conditions, and can, therefore, be modeled more easily.

The machine learning model that was used seeks to define the relationship between all eleven variables, instead of just the relationship between two, as was shown in Figure 8. This cannot be performed using a simple graph, as the line runs through eleven-dimensional space and there are millions of mathematical comparisons that must be considered in the calculation. Therefore, machine learning is used to examine how the relationships between environmental variables affect where plants can grow, as this process is too complex for humans.

The other species that were not able to be predicted either lacked enough data for the program to create a meaningful model or the relationship between the environmental variables was more complex and not as easily modeled. Therefore, it would be pertinent to continue improving the machine learning model, as this would help define the boundaries of where plants can and cannot grow in a more accurate manner, and thus answer the question of what specific conditions each plant can tolerate.

Conclusion:

This exploratory data analysis has sought to uncover insights that help to answer some of the questions posed by the researcher, Dr. Rajaniemi. In order to prove the statistical validity of these insights, a confirmatory data analysis could be completed to thoroughly test the findings of this report. In addition, actual experiments on each type of plant could be conducted in a lab to test the validity of these findings as well.

It is also worth noting that there existed a few limitations in this exploratory data analysis. Firstly, for most of the plant species, there was not enough data to draw any real conclusions. The insights from this report mostly come from the most common species on the dune. In addition, this report only examines if plant growth could be modeled at all using machine learning algorithms - it does not attempt to build the most predictive model. Therefore, it is recommended to continue improving the machine learning model, using methods such as principal component analysis and cross-validation to create a model that more accurately defines where plants can and cannot grow.