Final Project Report: Climate Change and Valley Fever

Data 102: Data, Inference, and Decisions, Spring 2021

by Danny Ticknor, Hannah Chea, Meer Wu, and Bridget Nix

1 Introduction

Every individual's carbon footprint, plastic consumption, and use of air conditioning is increasing the Earth's temperature. The evidence of climate change has been more visible, with an increase in the number and magnitude of natural disasters, the increasing number of organism extinctions, and the onset of diseases such as COVID-19. Valley fever (*coccidioidomycosis*) is an infection that mainly targets the lungs and is caused by the fungus Coccidioides. These fungal spores live in dust and soil in the southwestern United States[7]. Affecting only certain dry areas, valley fever is a specific disease that is often misdiagnosed. Taking this into account, our goal of our final project is to explore how different environmental aspects of climate change affects the number of valley fever cases.

2 Data Overview

We used data from many datasets in our attempt to best estimate the effect of climate change. We do not claim to know the true effect of climate change in the future because it has not occurred yet.

2.1 Valley Fever

The valley_fever dataset[1] consists of monthly case counts of coccidioidomycosis (Valley fever) by date and county among the states of Arizona, California, and Nevada from 2000 to 2015. This public data was developed by Morgan Gorris with support from a Department of Defense (DoD), National Defense Science & Engineering Graduate Fellowship and is based on confirmed cases and diagnoses from each state's health agencies. There are some limitations with the data. People may have been infected in a different county than where they were diagnosed, there is a lag between when a person is infected and when the case is reported, and some valley fever cases have likely been reported as other health conditions, such as asthma.

2.2 Ozone

The Ozone dataset[2] from the Center for Disease Control and Prevention consists of predicted Ozone values as well as their standard deviation. We used data from 2011-2014 and worked only in states where Valley Fever is prevalent (ie. California, Arizona, and Nevada).

2.3 PM

The PM dataset[3] from the Center for Disease Control and Prevention consists of predicted Particulate Matter values as well as their standard deviation. We used data from 2011-2014 and worked only in states where Valley Fever is prevalent (ie. California, Arizona, and Nevada).

2.4 Drought

The drought dataset[4] has 5 categories of drought, from D0 to D4, with specifics given below. We will be utilizing weekly data for counties in our states of interest from the years 2011-2014. Drought categories, from

least severe to most severe, include “No Drought,” “Abnormally Dry,” “Moderate Drought,” “Severe Drought,” “Extreme Drought,” and “Exceptional Drought.”

2.5 Temperature & Precipitation

The temperature dataset[5] is from the National Oceanic and Atmospheric Administration National Centers for Environmental Information and is real-time analysis of monthly temperature and precipitation across the United States. This dataset was originally created to study climate variability and change, and the observations have been adjusted to account for the artificial effects introduced into the climate record.

2.6 Population Data

The population dataset[6] is from the American Community Survey (ACS) and was downloaded from data.census.gov as a csv file. The data are 5-year estimates for the years 2010 to 2014 and are based on a sample of the population. Each row represents an estimate of the total population for a county in one of the following states: Arizona, California, and Nevada.

2.7 Cleaning

We cleaned each of the datasets before merging them together. In order to merge the datasets, we had to remove some of the top rows and convert the top row of variable names into the column names, reformat date strings, and rename columns. None of the datasets had missing values.

The particulate matter and ozone datasets had records for each day, so we took the average of the PM2.5 and ozone concentrations for each month. It is possible that averaging the values might remove some patterns in the data. We also used population estimates for each county to add a column to the valley fever dataset with the number of cases per 100,000 people.

3 EDA

After combining our datasets into one large dataset, we created a heatmap (Figure 1) and a pairplot (Figure 2) to look at all the correlations between all the variables in the dataset.

The number of cases per 100,000 has weak positive correlations with temperature, mean PM2.5 concentration, and mean ozone concentration, although the number of cases actually decreases after a certain threshold of mean ozone concentration. There is a very weak negative correlation with precipitation, but a high number of valley fever cases only occurs when precipitation is low. There is no correlation with any of the drought classifications. Since the percentage of each county covered by drought was not correlated with any other variables, we decided to exclude drought data from the rest of our analysis.

Temperature and ozone concentration have the strongest correlation (r=0.52). Ozone also has a negative association with precipitation (r=-0.3), and the highest cases only occur. The correlation coefficient for ozone and particulate matter is -0.28, although they do not appear to be correlated based on the scatter plot. Temperature is negatively correlated with precipitation (r=-0.28), and the highest and lowest temperatures only occur when precipitation is low.

Figure 3 shows the number of Valley Fever cases over a given year by each state. We compared the number of Valley Fever cases for the years with the highest and lowest number of cases in each state. There were observable differences between the number of cases each month and the distribution of cases over time across different years, which motivated our later analysis to use data separated by months in a given year.

In figure 4, we visualized the relationship between ozone concentration and the number of cases per 100,000 people by state. California and Nevada have positive correlations, but Arizona does not appear to have any association between ozone concentration and valley fever cases. Also, California had the largest range of case numbers (0 to 43.6) and Arizona had the second largest range (0 to 32.2). Nevada had a much smaller range of case numbers than the other two states (0 to 5.9).

Visualizing the mean number of cases per 100,000 people for each month did not reveal any patterns (Figure 5). However, separating the data by state revealed that valley fever cases are highest for the months from late summer through winter for California, cases are highest from late winter through early summer for Arizona, and Nevada has spikes in cases that seem to be more random.

4 Research Questions

4.1 Question 1: Does the increase in ozone concentration cause the rise in Valley Fever cases?

Answering this question can help people avoid getting Valley Fever by living in places with lower ozone concentration or advance scientific research on causes of Valley Fever. This question investigates causal relationships.Since we are not domain experts for Valley Fever and are limited to using observational data, causal inference with an instrumental variable is a good fit to our research question.

4.2 Question 2: Can we predict the number of cases based on environmental factors?

Knowing which environmental factors affect Valley Fever gives scientists a direction to study and improve the currently limited understanding of Valley Fever causes. We want to do prediction using multiple features. Hence, we compare and contrast different models (GLM, nonparametric) to see how we can make the best predictions.

5 Technique 1: Causal Inference

5.1 Methods

Our analysis used an instrumental variable and 2-stage least squares regression to estimate the average treatment effect of ozone levels on the number of Valley Fever cases. We chose ozone concentration (ppb) as the treatment variable, the number of Valley Fever Cases per 100,000 people in each county in Arizona, California, and Nevada as the outcome variable, and precipitation level (inches) as the instrumental variable. During EDA, we identified ozone levels as one of the variables that is most strongly correlated with the number of Valley Fever Cases and precipitation as one of the variables most strongly associated with ozone levels that is also uncorrelated with the number of Valley Fever cases. While the lack of correlation does not guarantee independence, this was the best variable we could find amongst all the variables we examined. Some possible confounders we investigated are temperature, PM2.5 levels, and geographic location. Both temperature and

PM2.5 levels positively correlated with the number of Valley Fever cases. California counties had the highest number of cases per 100,000, followed by Arizona counties then Nevada counties.

Based on our background research on the topic, we suspected that sunshine, wind speed, and humidity may also be confounders. However, we were unable to obtain data for those variables, so we were unable to examine the effect of these environmental factors on the onset of Valley Fever. Since there were a lot of possible confounding variables and we were unable to obtain data for all of them, we used an instrumental variable to control for known and unknown confounders. Using an instrumental variable, we assumed a linear structural model between the treatment and the outcome variable as well as between the treatment and the instrumental variable. We then performed 2-stage least squares regression by first predicting ozone concentration from precipitation level then regressing the number of Valley Fever cases on the predicted ozone concentration.

5.2 Results

We found an estimated average treatment effect of 0.164. Based on our result, increase in ozone concentration does cause a rise in Valley Fever cases. This result is consistent with our hypothesis and the positive correlation we observed between the two variables. Our result is valid under the assumption that the level of precipitation is completely independent of the number of Valley Fever cases. We were only able to investigate uncorrelation between precipitation and Valley Fever case rate, so if the two variables are not actually independent, our estimate may not hold. In addition to the independence assumption, we assume that the relationships between ozone concentration and Valley Fever case rate as well as between ozone and precipitation levels are linear. While we observed a correlation between these variables, there still may exist nonlinear relationships between them. It is also difficult to guarantee that our instrumental variable, precipitation level, has a causal effect on ozone. We investigated for association between precipitation and ozone, but correlation does not imply causation.

5.3 Discussion

We are not very confident that there exists a causal relationship between ozone and the number of Valley Fever cases since our instrumental variable is not perfect. Without the instrumental variable working properly, there are many confounders that could easily affect the outcome. For example, if the effect of temperature or PM2.5 were not eliminated through the instrumental variable, our result would show a stronger ATE between ozone concentration and Valley Fever cases. In our EDA, we observed that ozone concentration did not correlate with the number of Valley Fever cases for counties in Arizona, which affirms our suspicion that the instrumental variable may not be working as it should. Data on sunshine duration, wind speed, and humidity will be helpful in eliminating confounders. Limitations include lack of information to create a *perfect* instrumental variable and lack of access to data for other environmental factors. We observed Simpson's paradox after stratifying by state. Cases in Arizona had a much stronger estimated ATE than cases in the other Nevada or California.

6 Technique 2: GLM and Non-parametric methods

6.1 Methods

Method 1 was GLM with bootstrap, as seen in figures 7,8, and 9. An assumption for bootstrapping is that the sample size acts as a representative sample of the total population. Some assumptions for the GLM include independence of data points, a linear relationship between predictor and response, and the negative binomial being the correct distribution for the residuals. We used a negative binomial model to predict three different scenarios onto three bootstraps. This included the states of California, Arizona, and Nevada from the years 2011-2014. Scenario one predicted the effect of less than 1 inch of precipitation on the number of Valley Fever cases. Scenario two predicted the effect of ozone over 50ppb on the number of Valley Fever cases. Scenario three predicted the effect of temperatures above 85^0 F on the number of Valley Fever cases. The negative binomial was chosen because it's error most closely resembled the error of the bootstrap. We evaluated the GLM's performance based on the correlation coefficient and standard error given by the model. We calculated the beta value of each feature as a 95% confidence interval. It looked like:

```
[corr coef + 2(sd error), corr coef - 2(sd error)]
```

For our nonparametric method, we used Random Forest to predict the number of cases per 100,000 people from environmental factors (particulate matter, ozone, average temperature, and precipitation)'. We split the data into a training and test set, with 30% of the data in the test set. We have over 2,700 rows in the "final" data table, so the training and test sets should both have enough data. We chose features for the model based on correlations we observed from EDA and fit the model to the training data. We chose the best model by minimizing the root mean squared error and finding a scatter plot of the true number of cases per 100,000 people vs. the predicted values where the points are closest to the line y=x as possible.

We initially used a multiple linear regression model on the data using the same features that were used in the random forest model, but the range of predictions for low number of cases was too large, and the model severely overestimated the number of valley fever cases per 100,000 people when true number was less than about 10. We then decided to use a random forest model to limit overfitting.

6.2 Results

The results of the GLM are as follows and can be confirmed through our code. $Beta_{0\_precip}$ is between 0.192 and -0.204, $Beta_{0\_temp}$ is between 0.20 and -0.047, and $Beta_{0\_ozone}$ is between 0.014 and -0.026. In plain English, one additional unit of less than 1 inch of precipitation is correlated with approximately -0.006 Valley Fever cases.. Similarly, one additional unit of 85^0 F is correlated with approximately 0.077 Valley Fever cases. Last but not least, one additional unit of 50 ppm Ozone is negatively correlated with approximately -0.006 Valley Fever cases.These values were calculated as the top of the confidence interval plus the bottom of the confidence interval divided by two.

For the Random Forest model, the root mean square error for the training set was 1.537 and the root mean square error for the test set was 4.2289. Figure 6 shows the relationship between the true values and predicted values. If our model was 100% accurate, the points would be along the red line y=x. Points are fairly close to the line for true values 3 through 20. However, the model underpredicts values for true values above 20, and

for values under 3, the model predicts a large range of values. The random forest is definitely the better performer.

6.3 Discussion

The GLM results are overall not very good. We can see an extremely slight relationship with high temperatures, but little to no relationship for ozone or precipitation. High temperatures are correlated with more cases, while ozone and light rainfall are close to zero. This is likely due to the difficulty of quantifying climate change as a variable. We used ozone, particulate matter, precipitation, and drought data and still barely scratched the surface as to the true estimate of the effect of climate change. The bootstrap errors are not too severe and would likely be helped with a larger dataset and sample size. We also had some limitations with the amount of Valley Fever data available to us. Overall, between 2011 and 2014, cases actually decreased year to year which is definitely not as we or climate scientists would expect. Our model may have been helped by the addition of dust storm or desertification data, but these were difficult to find and process. I am not confident in applying this to future datasets without more data and the opinion of some domain knowledge experts..'

The mean squared log errors for the GLM were extremely high as to be expected.
Temperature performed the best with 23.91, ozone had 74.7, and precipitation was at 85.97. These are absurdly high compared to the mean squared log error of 0.701 for the random forest model. We concluded that the random forest model was the best for predicting Valley Fever cases using environmental features.

7 Conclusion

Our analysis examined the relationships between Valley Fever and environmental factors. In merging data sources for Valley Fever cases and other environmental variables, we were able to examine relationships across all combinations of the variables. However, this forced us to aggregate some of the data from daily observations into monthly observations using mean estimates. Based on our results, ozone concentration causes Valley Fever. Temperature and precipitation are good predictors for Valley Fever cases. For every degree of increase in temperature or every inch of increase in precipitation, Valley Fever cases increase by 0.2. Studies have shown that climate change increases ozone concentration[8]. Our result shows that climate change, in addition to its negative effects on the environment, also has a negative impact on lung health. Climate change needs to be taken seriously. There needs to be stronger climate protection policies in place. There are currently no studies that investigate the relationship between ozone levels and Valley Fever. Similar studies should be done in order to ensure the replicability of our results. Our analysis was also limited to a subset of environmental factors we wanted to explore due to lack of public datasets available. Future studies should explore the effects of other environmental factors such as sunshine duration, wind speed, and humidity on Valley Fever.

8 Academic honesty statement

We give our word that all of the work in this project is our own, and that we have cited all of our sources, including my classmates, to the best of our ability.

9 Appendix

Figure 1: Heat map of correlations between variables, with Pearson standard correlation coefficients included.

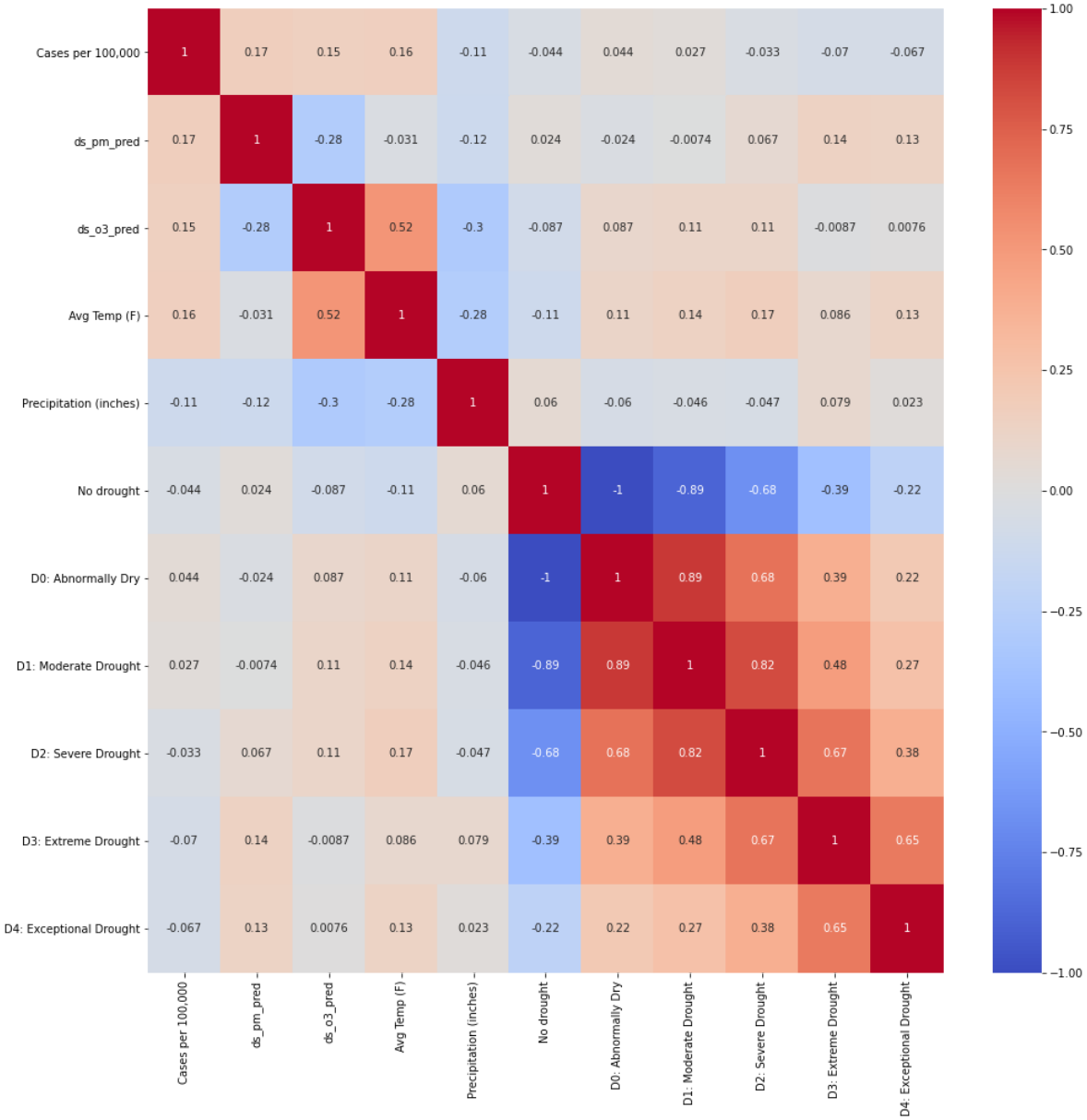


Figure 2: Scatter plots for each pair of variables. Kernel density estimate plots for the variables are shown on the diagonals.

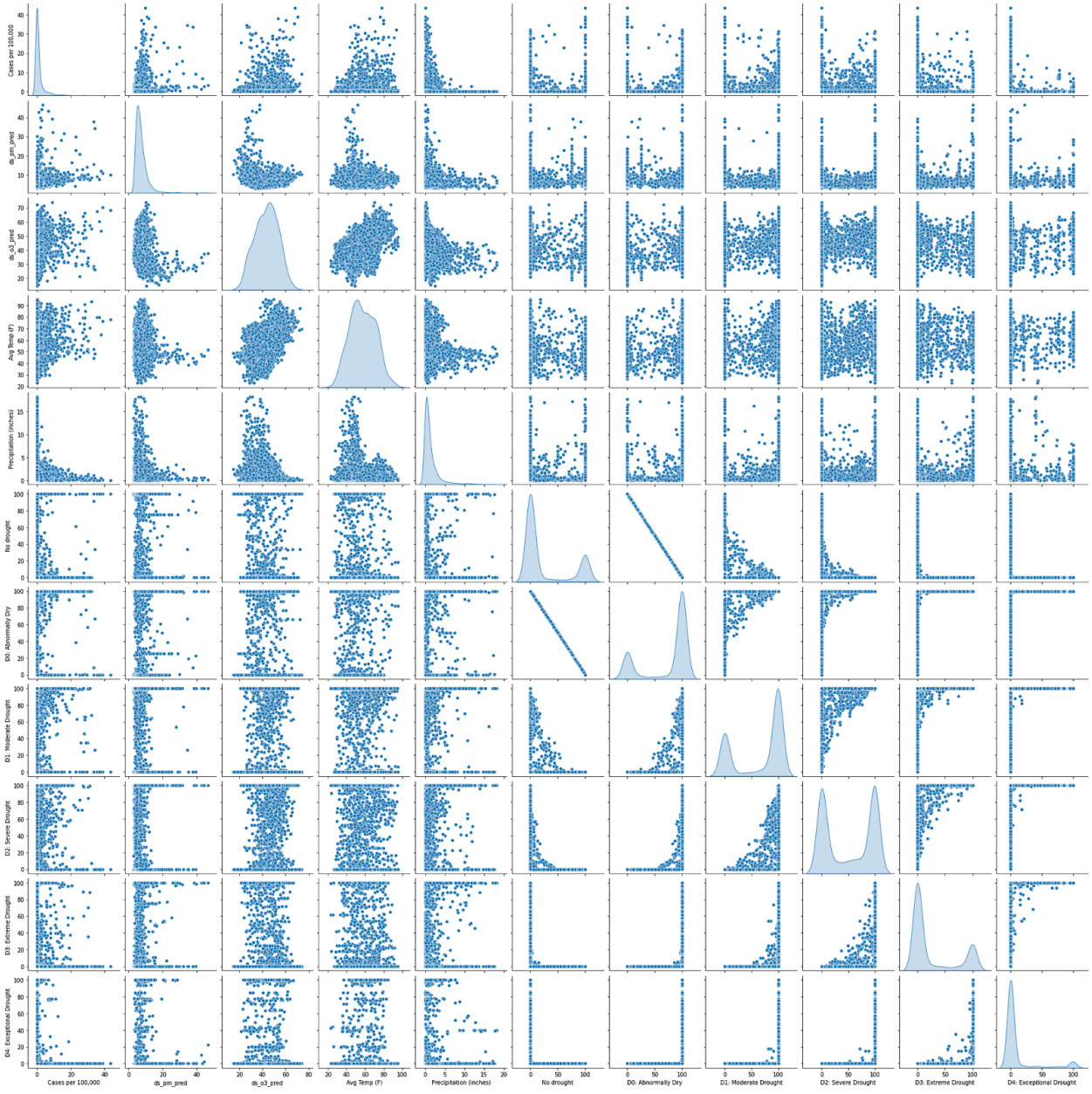


Figure 3: Number of valley fever cases per 100,000 people over time, by state. Data for the years with the lowest and highest number of cases are shown for each state.

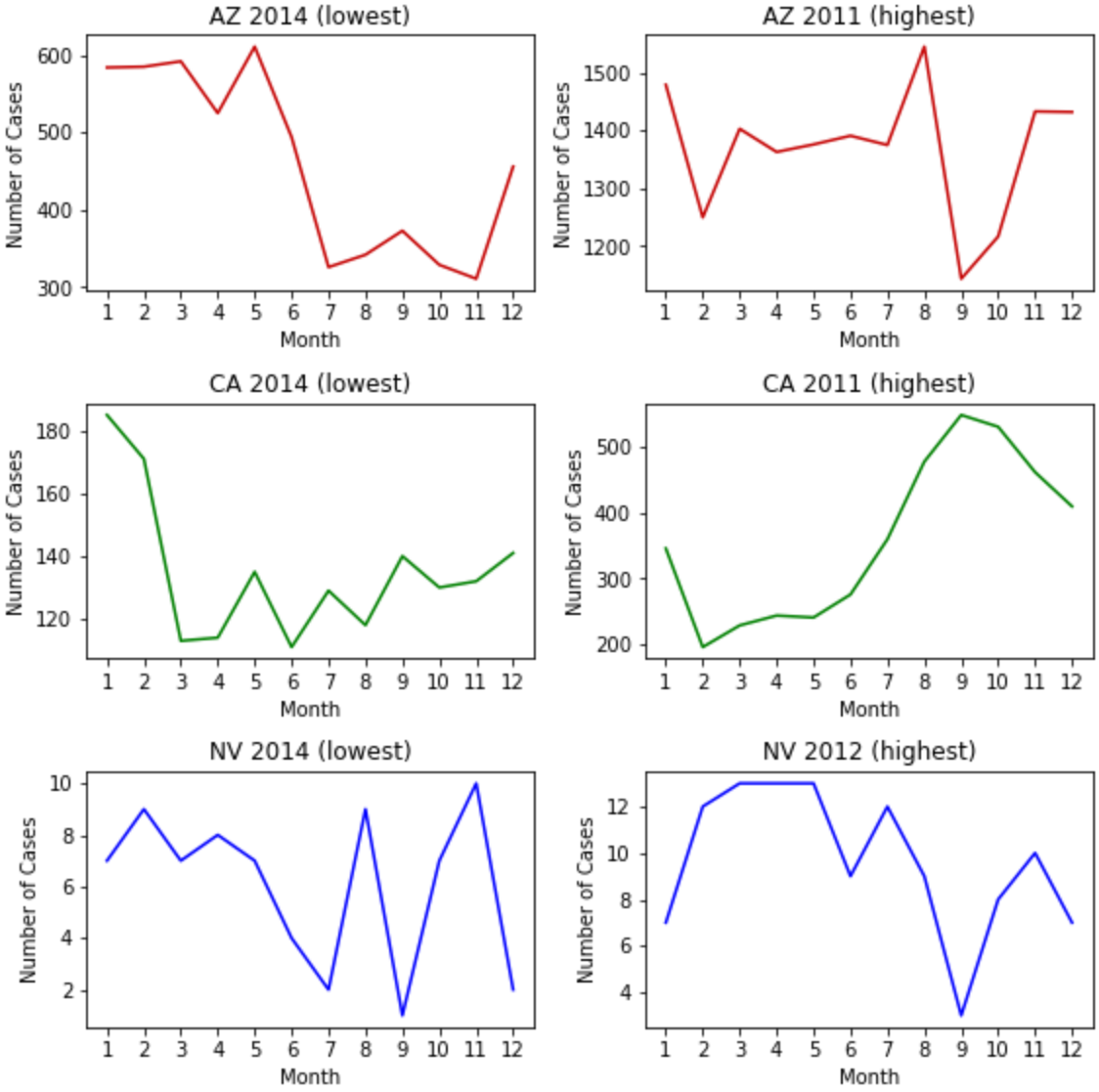


Figure 4: Valley fever cases per 100,000 people vs. ozone concentration, by state. Note that different states have different scales for the y-axis.

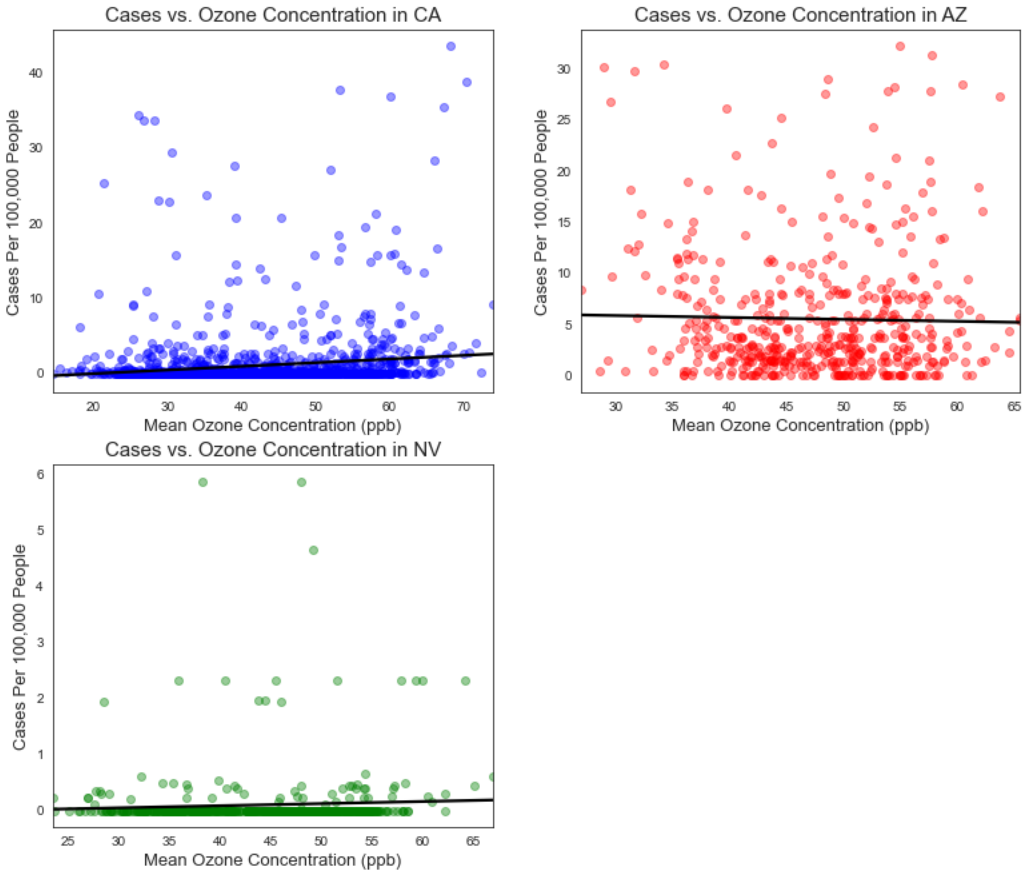


Figure 5: Average number of valley fever cases per 100,000 people for each month, by state.

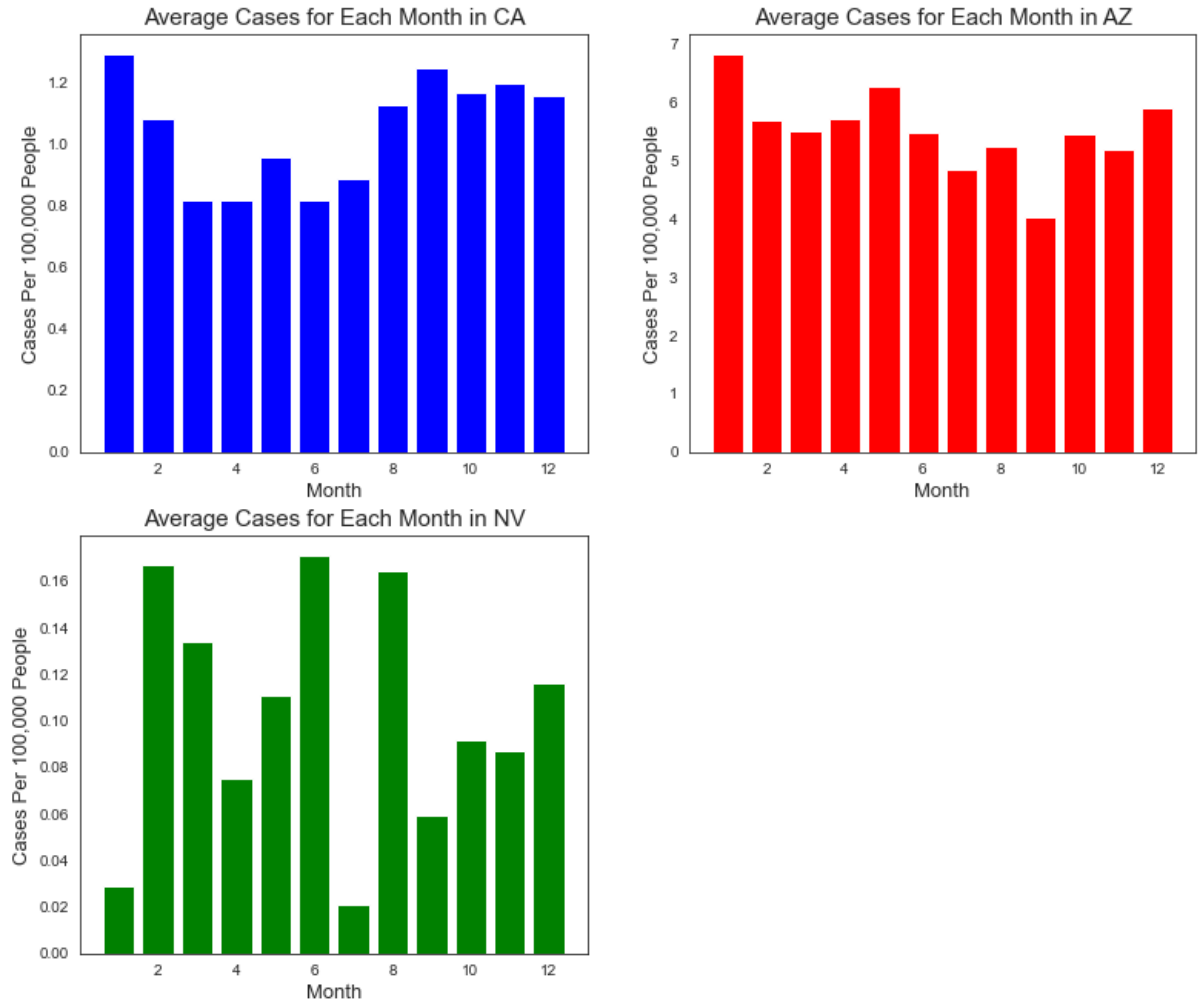


Figure 6: Scatter plot of the true number of valley fever cases per 100,000 people vs. the predicted values from the random forest model.

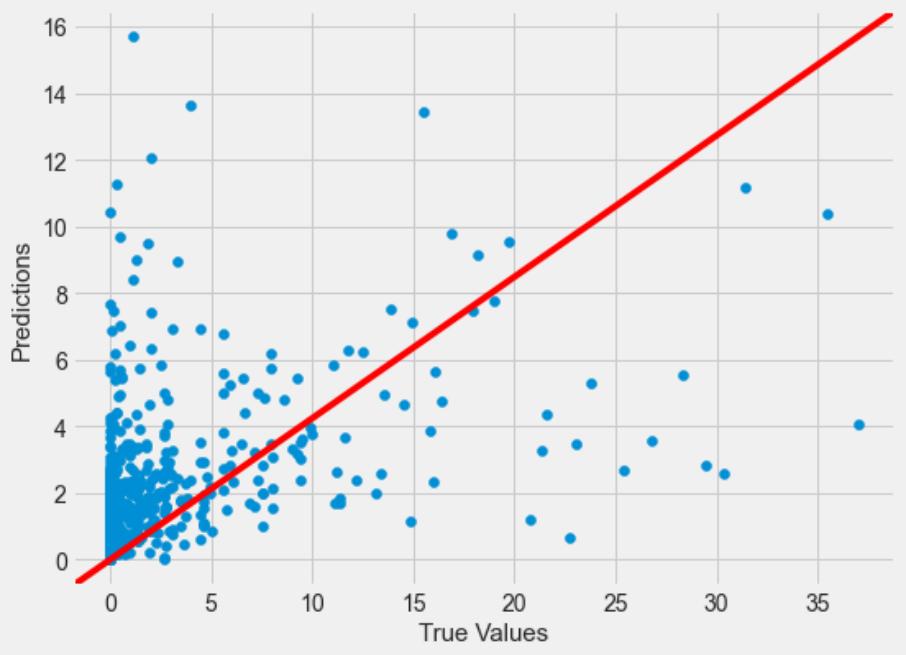
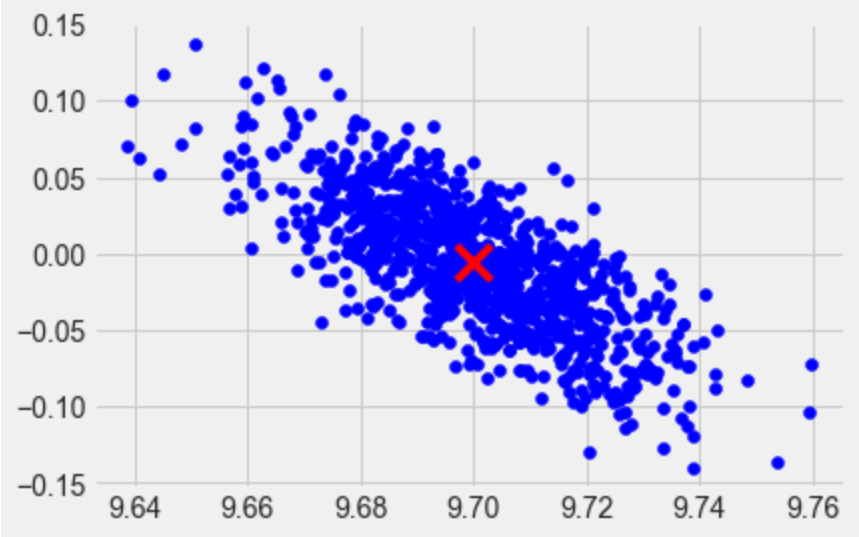
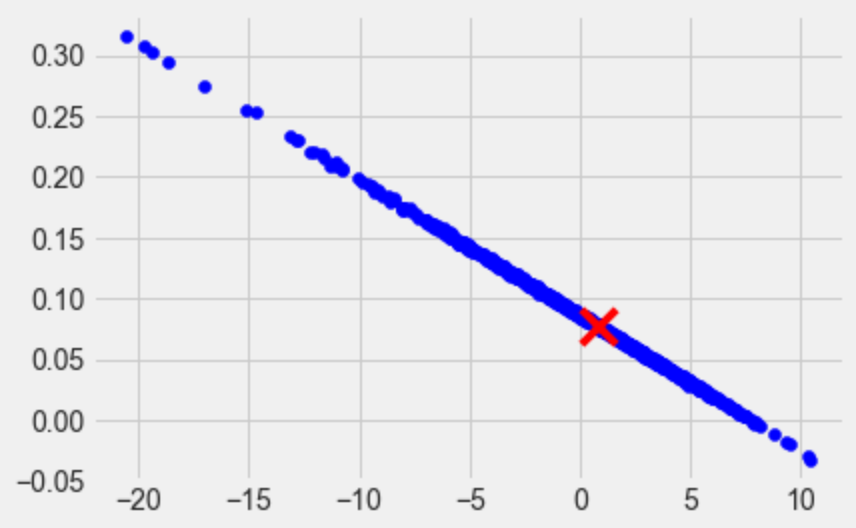


Figure 7: Scatter plot of GLM bootstrap samples. Constants are on the x-axis, and the coefficient for precipitation is on the y-axis. A summary of the results for the GLM is included below the graph.



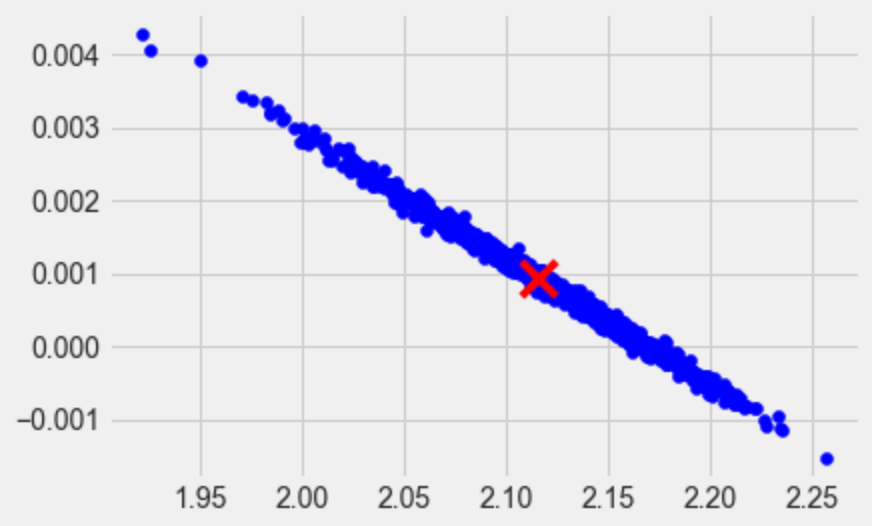
```
Bootstrap std error for const: 0.018
Bootstrap std error for year: 0.043
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 totals   No. Observations:                 1196
Model:                            GLM   Df Residuals:                     1194
Model Family:        NegativeBinomial   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12795.
Date:                Mon, 10 May 2021   Deviance:                       569.23
Time:                        16:16:21   Pearson chi2:                     219.
No. Iterations:                     5
Covariance Type:            nonrobust
===============================================================================================
                              coef    std err          z      P>|z|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                       9.6998      0.043    225.156      0.000       9.615       9.784
Precipitation (inches)     -0.0061      0.099     -0.061      0.951      -0.201       0.189
===============================================================================================
```

Figure 8: Scatter plot of GLM bootstrap samples. Constants are on the x-axis, and the coefficients for average temperature are on the y-axis. A summary of the results for the GLM is included below the graph.



```
Bootstrap std error for const: 4.122
Bootstrap std error for year: 0.046
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 totals   No. Observations:                   41
Model:                            GLM   Df Residuals:                       39
Model Family:        NegativeBinomial   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -353.33
Date:                Mon, 10 May 2021   Deviance:                       53.740
Time:                        16:17:28   Pearson chi2:                     17.3
No. Iterations:                     6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7968      5.506      0.145      0.885      -9.995      11.589
Avg Temp (F)   0.0773      0.062      1.239      0.215      -0.045       0.199
==============================================================================
```

Figure 9: Scatter plot of GLM bootstrap samples. Constants are on the x-axis, and the coefficient for ozone concentration is on the y-axis. A summary of the results for the GLM is included below the graph.



```
Bootstrap std error for const: 0.049
Bootstrap std error for year: 0.001
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 totals   No. Observations:                  550
Model:                            GLM   Df Residuals:                      548
Model Family:        NegativeBinomial   Df Model:                            1
Link Function:                    log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -5489.2
Date:                Mon, 10 May 2021   Deviance:                       266.58
Time:                        16:17:53   Pearson chi2:                     107.
No. Iterations:                     5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          9.3368      0.539     17.322      0.000       8.280      10.393
ds_o3_pred    -0.0064      0.010     -0.664      0.507      -0.025       0.013
==============================================================================
```

$$\textbf{Beta}_{0\_precip} = [0.192, -0.204]$$
$$\textbf{Beta}_{0\_temp} = [0.20, -0.047]$$
$$\textbf{Beta}_{0\_ozone} = [0.014, -0.026]$$

10 Citations

[1] Gorris ME, Cat LA, Mathlock M, Ogunseitan OA, Treseder KK, Randerson JT, Zender CS. (2018). Coccidioidomycosis (valley fever) case data for the southwestern United States. [Data file]. Open Health Data, In Review. Retrieved from https://github.com/valleyfever/valleyfevercasedata

[2] Center for Disease Control and Prevention. (2018). Daily Census Tract-Level Ozone Concentrations, 2011-2014. [Data file]. Retrieved from https://data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-Ozone-Concentrations-2011/372p-dx3h

[3] Center for Disease Control and Prevention. (2018). Daily Census Tract-Level PM2.5 Concentrations, 2011-2014. [Data file]. Retrieved from https://data.cdc.gov/Environmental-Health-Toxicology/Daily-Census-Tract-Level-PM2-5-Concentrations-2011/fcqm-xrf4. 2018.

[4] United States Drought Monitor. (2021). Comprehensive Statistics. [Data file]. Retrieved from https://droughtmonitor.unl.edu/Data/DataDownload/ComprehensiveStatistics.aspx.

[5] NOAA National Centers for Environmental information. (2021). Climate at a Glance: County Mapping. [Data file]. Retrieved from https://www.ncdc.noaa.gov/cag/county/mapping/4/tavg/202103/1/value

[6] U.S. Census Bureau. (2021). Total Population for all counties in CA, AZ, NV. [Data file]. 2010-2014 American Community Survey 5-Year Estimates. Retrieved from https://data.census.gov/cedsci/table?q=total&g=0400000US04.050000,06.050000,32.050000&tid=ACSDT5Y2014.B01003&hidePreview=true

[7] *Valley Fever (Coccidioidomycosis)*. (2020, December 29).Centers for Disease Control and Prevention. Retrieved May 10, 2021, from https://www.cdc.gov/fungal/diseases/coccidioidomycosis/

[8] Sheffield, P. E., Knowlton, K., Carr, J. L., & Kinney, P. L. (2011). Modeling of regional climate change effects on ground-level ozone and childhood asthma. American journal of preventive medicine, 41(3), 251–A3. https://doi.org/10.1016/j.amepre.2011.04.017