Springboard Data Science Career Track

Capstone Project 1

Projecting Monthly Rainfall Totals for North Carolina for the Next 50 Years

William Fetzner

Table of Contents

Introduction:

One question that has been asked for generations is, "How much will it rain tomorrow?" For centuries humans have asked this question an attempted to predict when the rain would come and how much rain would fall. For a millennium, there was no accurate way of predicting rainfall amounts for more than a few days in advance. However, now with waves of new technology and systematically recording weather statistics, humans have been able to extend the ability to make predictions about future weather predictions with about 70% accuracy, and up to a week to 10 days in advance. This is important on a daily basis because a person going to work in the morning needs to know whether to put on rain boots and a rain jacket in the morning, or whether to bring their sunglasses. There is more that we can do than just predict exact weather conditions for a given day, but we can predict rainfall amounts in an area several years in advance. This has extremely important implications especially for the agricultural industry. In agriculture, rainfall predictions can mean the difference between whether a farmer plants one type of crop for a single growing season or another depending upon the water requirements of each. In addition, if rainfall in an area is trending down, then a farmer may decide to take his business elsewhere in order to cash in on a crop that is more productive in wetter environments.

In addition, government agencies must know how to plan for varying levels of precipitation. Plans must be put in place to create reservoirs that will be available for the water consumption needs of the population. Conversely, an increase in precipitation levels may lead to flash flood events resulting in damage to infrastructure. Governments must plan for these events and try to mitigate the effects of increased precipitation with laws and regulations set in place early.

The following project develops a 50-year projection of monthly total rainfall. The model uses data from the past 39 years from January 1, 1980 to April 30, 2019, to then predict the next 50 years of monthly rainfall totals for 112 different locations in North Carolina.

Data Acquisition and Cleaning:

The data for this analysis was acquired from NOWData NOAA which is an online database provided by NOAA (National Oceanic and Atmospheric Administration) and the National Weather Service for the free use of climate data by U.S. citizens. The following link: https://sercc.com/nowdatamap contains links to all the different regions in the United States. Within each region is a list of all of the different weather collection stations. Monthly totals for rainfall data were collected from all 112 North Carolina locations and from 124 external locations from Virginia, Tennessee, Georgia, and South Carolina. Data was collected into a single Microsoft Excel file where each sheet within the file was a different location. Each sheet had the years since collection started for that location as rows and each month as a column. The final column summed the total rainfall for the year, and after the 2019 row, there were a minimum and maximum row for each month and the average for the month over the entire collection period.

Several steps were taken to clean the data and place it into a workable format. First, instead of having each location as its own sheet, it was placed into a single column and the single year row was extended to 12 to accommodate each month. The summary data that was included from the raw data was also removed. There was a severe issue with missing data from the dataset. This was due to the recording of rainfall data in each location did not begin at the same time statewide. Therefore, many of the older and larger cities had rainfall totals from the late-1800s while many younger or smaller cities did not begin collection until 20 years ago. In addition to the lack of early data, there was also many cities that did not have any data for the past 5 years or had large gaps in the dataset. Thus, there were 3 different tactics used to deal with missing data: only the data since 1980 was used and then two different filling schemes were employed to fill in the missing data.

The first missing data filling schema used was to take the average rainfall total from the previous month, the next month and the rainfall amount the year before. This had several limitations including that the datapoint had to have data in one of the three previously mentioned months. Thus, it could not be the last month in the dataset, the first month in the dataset, or be before January 1981. The second schema employed found locations within 85 kilometers (50 miles) of the location with the missing data and averaged all the rainfall amounts together from each of the surrounding locations. The second schema was able to fill in all the remaining missing datapoints.

To find the locations that were within 85 kilometers the approximate latitude and longitude coordinates were found for each location and placed into a csv file that was then imported into the Jupyter notebook. Distances between each latitude and longitude coordinate were found using the haversine formula. From there the data frame was filtered into just including the target location and all locations within 85 kilometers.

The choice to use locations within 85 kilometers was determined by calculating the distance which rainfall data from the surrounding locations varied less than a single standard deviation away from the target location.

While still using the distance data frame, the next step was to find those locations outside of North Carolina that might be used as exogenous variables inside the prediction model. These variables are external locations outside of North Carolina that are close enough to target locations in North Carolina to possibly effect the performance of a model predictions. These external locations were placed into a dictionary with the target locations as the key and the external locations as values. The number of external locations for a single target location ranged from 1-5. There were 46 target locations that had external locations that were within 50 kilometers of the target location. The original plan was to include all external locations within

85 kilometers of the target location, yet due to time constraints, only locations within 50 kilometers were able to be used in this analysis.

Exploratory Data Analysis:

Before building a model, the rainfall data was first explored through several visual graphs, seasonal decomposition, and correlation. Figure 1 shows the monthly average total rainfall for Raleigh, NC. The error bars are standard error of the mean and depicts conformity for each month across the years. The highest amounts of rainfall come through the months of July through September; however, there is a relatively small difference of only an inch of rainfall between the highest month of July and the lowest of December.

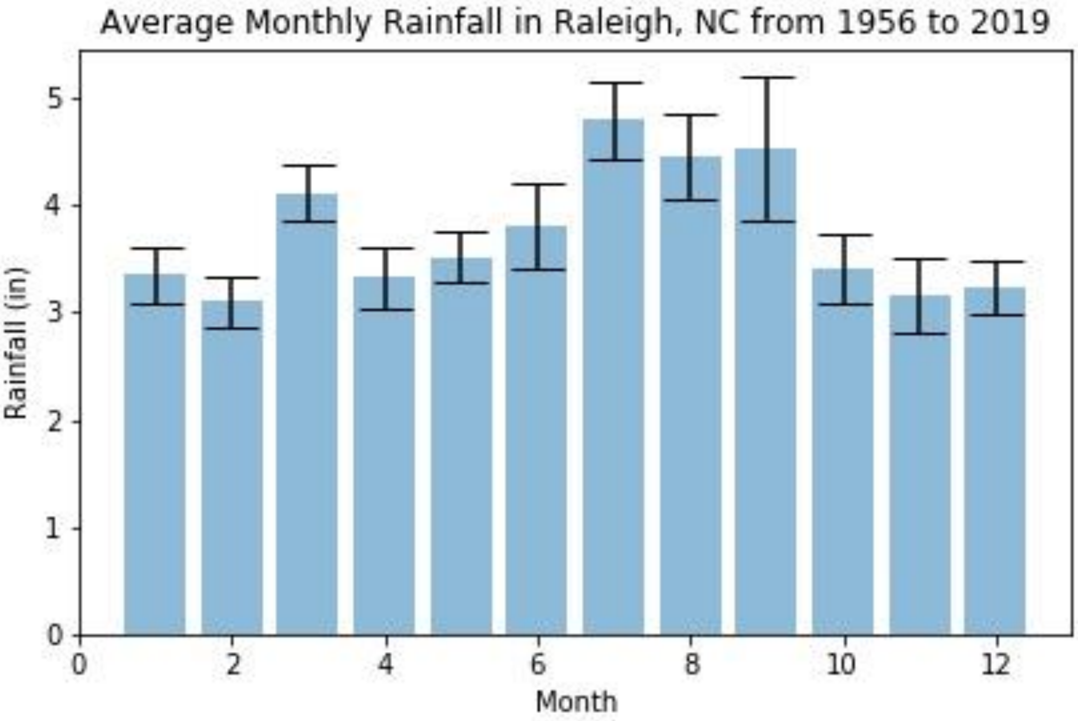

Figure 1: Average monthly rainfall for Raleigh, NC. Error bars show the standard error of the mean.

Figure 2 shows the average rainfall for each year from 1980 to 2019, then compares these between Raleigh, NC and Greensboro, NC which is only about 80 miles away. There are some differences in the average yearly rainfall, but the variation from 2010-2019 is different between the cities and the high peak for Greensboro in 2004.

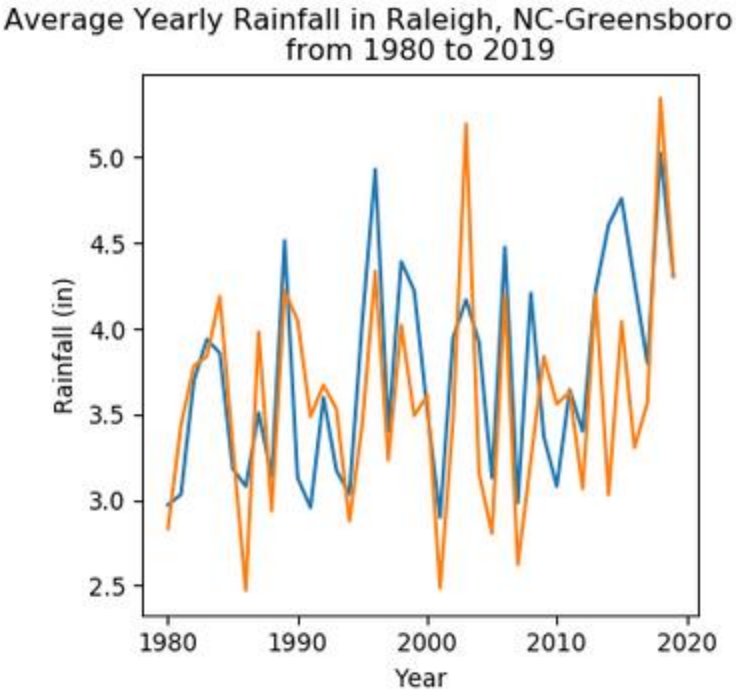

Figure 2: Yearly rainfall for Raleigh and Greensboro, NC from 1980-2019. Notice the similarities and differences in the average rainfall between the two relatively close cities.

Figure 3 shows the results of the seasonal decomposition analysis of the rainfall data. This particularly shows the yearly cycle of the rainfall amounts occurring every year. The seasonal decomposition results also yielded information about the trend of rainfall over the entire sample. It was found that this trend was not significantly greater than 0. Thus, there was no trend in rainfall either increasing or decreasing during the sample.

Autocorrelation is comparing the correlation between the current month for a single location and every previous month before it in that same location. Figure 4 shows an autocorrelation frequency plot for Raleigh, NC. The high peak at the first is the correlation with the value to itself, then there are peaks at 12-month intervals that decrease in correlation behind that. This once again shows the 12-month seasonality of rainfall. Autocorrelations were performed for all locations in three ways. First, the autocorrelation was run just comparing the current value to itself than every previous value. Then a correlation was run for the location can comparing it to the lag 1 values of that same location. Therefore, the correlation compares how well the months align when they are compared to the previous month. The last autocorrelation was done with the lag 12 values which is a similar concept to the lag 1 values only done twelve months behind instead of just a single month.

Lag 1 and lag 12 Correlations were then run between target North Carolina locations and their corresponding exogenous locations that were mentioned previously. These 46 target locations that had corresponding exogenous locations were correlated with both the previous

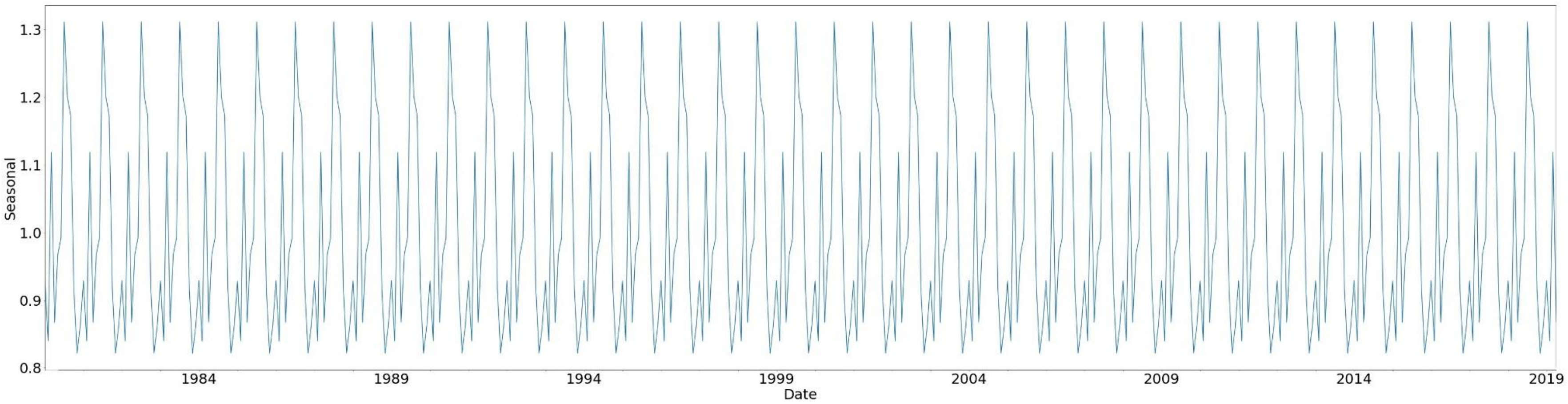

Figure 3: Seasonal Decomposition results for rainfall data. Notice the 12-month cycle of seasonality of rainfall.

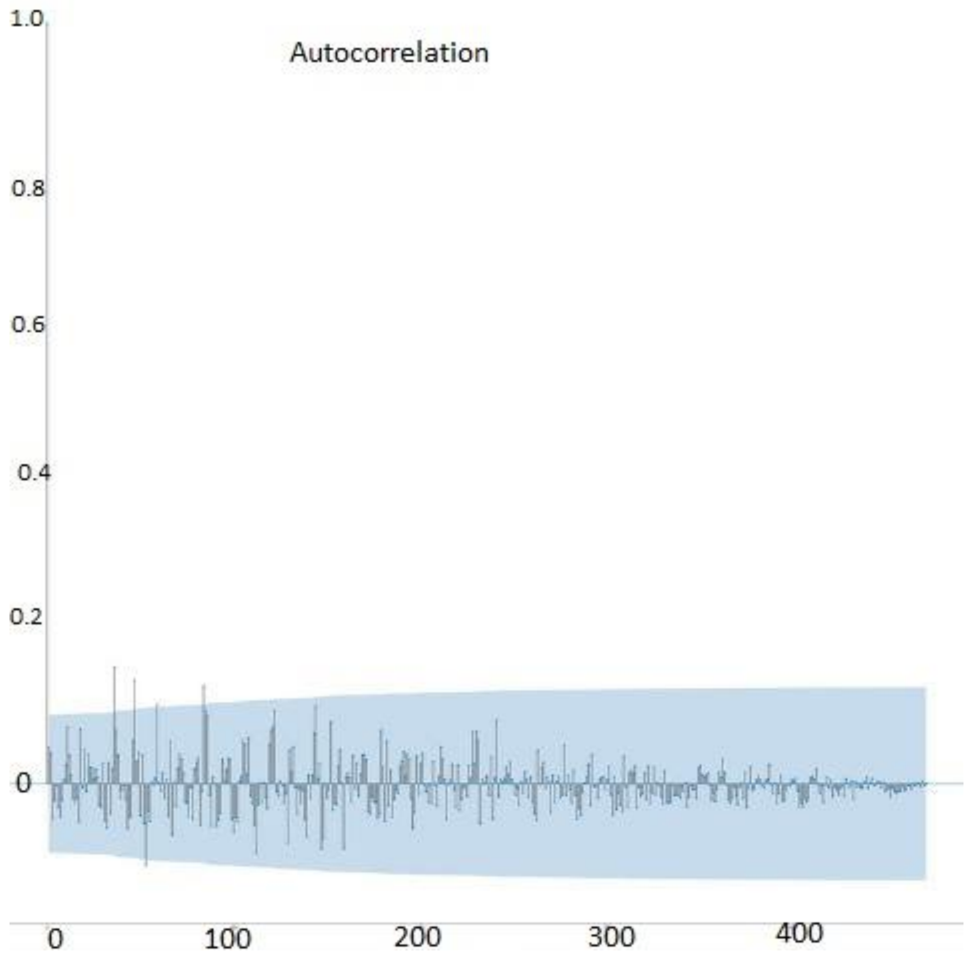

Figure 4: Autocorrelation frequency graph.

month and the month twelve months before the current month.

Machine Learning:

The model used to create the 50-year rainfall predictions was the Seasonal autoregressive integrated moving average (SARIMA). This model can be explained by separating the name into its several parts. The autoregressive part of the model is the target variable being regressed against its own lagged (previous) values. However, this model is integrated so it does not just take the target value and regress it against the previous value but regresses the difference between the target value and the target variables previous values. The moving average model that is created within the SARIMA model indicates a linear combination of the errors when the regression is formed in the autoregression. The seasonal part of the model considers the periodicity of the data and accounts for it in the model. The model is usually denoted as SARIMA (p,d,q)(P,D,Q,m). The parameter p is order or number of time lags of the autoregressive model. The parameter d refers to the degree of differencing or the number of times the data has had past values subtracted. The parameter q refers to the order of the moving-average model. P, D, Q refer to the same parameters only this time in terms of the seasonal component. Lastly, m refers to the number of periods in a season. This model does a great job for time-series analysis data that has a seasonality component to it. Thus, the SARIMA model was used to first be trained on the data then to generate predictions.

The first step was to determine the hyperparameters of the model. The hyperparameters that created the best model was p = 4, d = 0, q=3, P=3, D=0, Q=4, m=12. Both d and D were determined by there not being a trend found in the data during the seasonal decomposition. The m was determined by the seasonality of 12-month periods found several times in the exploratory data analysis, while the other four parameters were tuned through dividing the data into a training (80% of dataset), validation (10%) and testing set (10%). The purpose of this split of 80, 10, and 10 was so that we could tune the model to the best parameters while keeping a portion of the data (testing data) unknown to the model. Therefore, in order to find the best parameters, the training data set was fit to the model then the model predicted a single step forward. The first value in the validation set that the model just predicted was then added to the training set and the model was fit to the training set with the additional datapoint. This continued until all the values of the validation test set were predicted. Then the performance of the model was evaluated by finding the mean absolute error between the real values in the validation dataset and those predicted by all the models ran. The model with the smallest mean absolute error had the hyperparameters listed above.

After finding the hyperparameters, a similar process was completed for finding out if any of the 46 target locations with possible exogenous locations were truly needed to be exogenous locations to benefit the performance of the model and strengthen the predictions. A similar process to splitting the data into training (80%) and testing data (20%), finding the one-step predictions, and comparing the performance of models with the mean absolute error was completed to find the exogenous locations. Out of the 46 target locations with possible exogenous variables only 8 target locations (Whiteville, NC, Casar, NC, Forest City, NC, Gastonia, NC, Lake Lure, NC, Elizabethtown, NC, Mount Holly, NC, Grandfather Mtn, NC) had a better model when including the data from the exogenous locations.

The last step was to use the models to predict the rainfall monthly totals for the next 50 years. The models generated a point prediction for the following month and 50% confidence intervals for the following 50 years.

Conclusion:

The SARIMA model predicted monthly rainfall confidence intervals for the next 50 years. The North Carolina heat map (Figure 5) shows the rainfall that is predicted for years to come. Although there is variation in the predictions, this data can be useful for the agricultural industry in North Carolina by knowing to not to expect much change in rainfall totals over the next 50 years and how much to expect. This may allow farmers to make the decision to remain in North Carolina and continue to expect the rainfall that has been characteristic of this region for the past 40 years.

There are several aspects to weather that were not included in this model that could influence the predictions. One potential influence may be climate change that is affecting rainfall in drastic ways. Many places are seeing stark contrasts between severe flooding and severe droughts. These factors were not included into the SARIMA model but may be included in future models. In addition, the dataset could have been made stronger by not using only recorded data but using climate data for the past centuries and seeing the trend for the past century since the possible beginning of climate change to the previous centuries. This may be able to incorporate the climate change factor into the rainfall predictions for the next 50 years in North Carolina. Therefore, these findings should be interpreted only while keeping in mind the limitations of the variables included in the model.

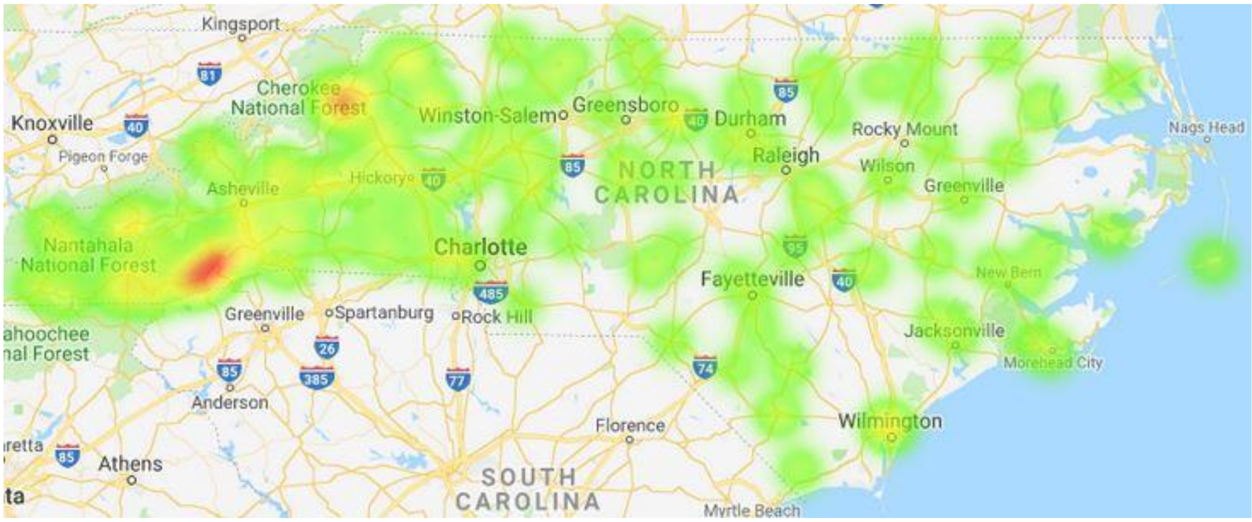

Figure 5: North Carolina Heatmap of the first month's rainfall prediction