Andrea Bruckner
Predict 413: Final

Forecasting Monthly Rainfall

Introduction

Being able to forecast rainfall benefits nearly everyone. Farmers and gardeners use these forecasts to decide what to grow and how often they need to water their crops; business executives use weather forecasts to decide whether they should walk to the office or take a cab; party planners use forecasts to determine if they should plan the party indoors or outdoors, or maybe if they should rent a tent for an event.

Not knowing the amount of rainfall can present many problems. In terms of the environment, too little rainfall can result in crop failure or forest fires, while too much rainfall can lead to flooding or landslides. Earlier this month, Louisiana experienced prolonged periods of heavy rainfall that led to severe flooding in much of the state. Many areas experienced more than 20 inches of rainfall within a day, while at least one area had more than 30 inches of rainfall in a day[i]. Some reports indicate that meteorologists correctly forecasted the flood risk but failed to accurately predict the magnitude of the flooding[ii]. Being able to predict natural disasters, such as the recent flooding in Louisiana, and convey the forecasts to people who may be affected can help save lives.

In this report, I will be building several models to forecast the amount of monthly precipitation for the provided data set. These models could potentially be used for warning people about severe floods, agriculture planning, urban rainwater collection scheduling, and various other instances where monitoring rainfall is important. Although humans cannot control the weather, they can attempt to predict it and act accordingly.

Data and Literature Review

The data set for this analysis comes from the Daily Global Historical Climatology Network, a database of climate data collected from land-based weather stations around the world. It is available for download via the National Oceanic and Atmospheric Administration's official website. The data we are examining contains 24,806 rows of consecutive daily weather data spanning from September 1, 1946 to July 31, 2014. The data set includes a date variable plus three predictor variables: PRCP (precipitation, inches), ACMH (average cloudiness, percent), and AWND (average wind speed, mph). After conducting some exploratory data analysis, I decided to focus on building models using just the precipitation variable since it was the only variable with complete data for each day between September 1, 1946 and July 31, 2014. The cloudiness variable was missing 13,698 records, and the wind variable was missing 13,636 records. This means that they were each missing approximately 55 percent of their data. Although ACMH and AWND had far too many missing values to impute, I originally thought that maybe I could still use these variables in my models by eliminating all the records where either of these variables had a missing value. I even corrected the outlier in the wind variable from 3263.7 mph to the average wind speed of 7.7 mph before realizing I would not be able to use these two variables at all. After eliminating all the records where ACMH or

AWND had missing values, I had 4169 records, or approximately 17 percent of the original data set. This small data set represented only data between January 1984 and May 1995, which would not be helpful in predicting the 2014 to 2016 rainfall required.

After looking over the data, I aggregated the daily data into monthly data because it seemed like it would be easier to model without sacrificing the integrity of the data. I also divided the data set into two sets: a training set containing about 75 percent of the data (September 1946 to June 1997), and a testing set with the remaining 25 percent of the data (July 1997 to July 2014).

Prior to building my models, I completed some exploratory data analysis to familiarize myself with the data. In **Figure 1**, we can that there appears to be some seasonality to the amount of rainfall, although the amount of rainfall varies greatly over the years.

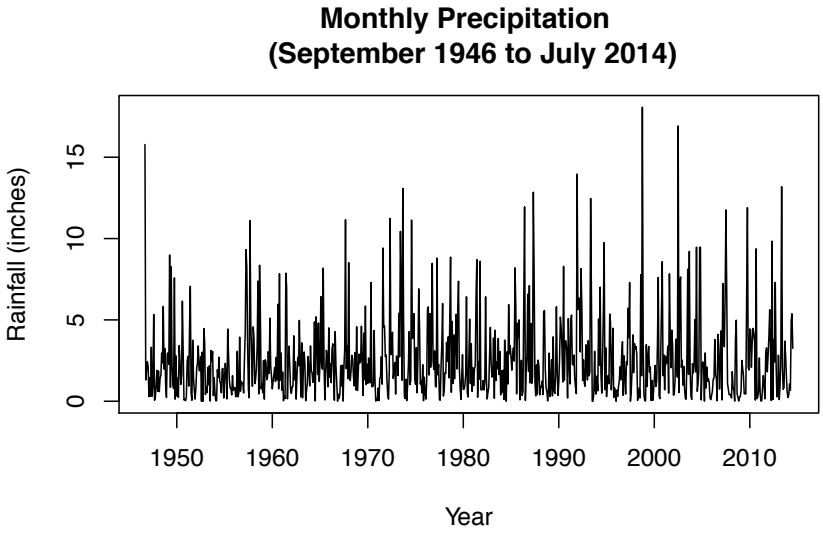

Figure 1: Plot of Monthly Precipitation Data

I then created month and seasonal deviation plots to further explore the data. The plots in **Figure 2** do not show a clear pattern, such as a bell-curve, but it appears that the spring and the fall generally see the highest amount of rainfall.

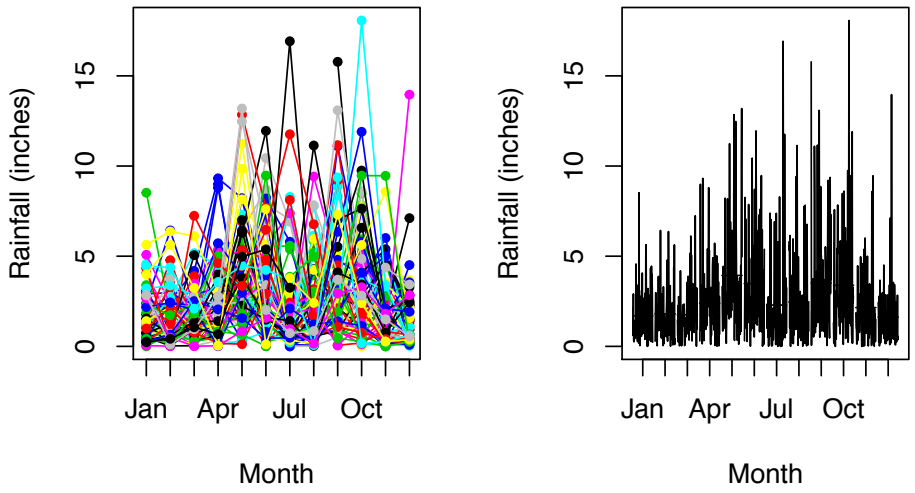

Figure 2: Month and Seasonal Deviation Plots

The plots I generated during my exploratory data analysis did not provide too many insights into how I might model the data, so I decided to research how others have forecasted rainfall data. The first article I looked at came from the *Journal of Hydrology* and was about how researchers used multiple regression and neural network models to forecast three years of spring rainfall values for Victoria, Australia[iii]. They discovered that the neural network models were more accurate in forecasting long-term rainfall than the multiple regression models. They used a training and test set while building their models, and evaluated the models using mean square error (MSE), mean absolute error (MAE), and the Pearson correlation coefficient.

I found another article from the same journal where researchers modeled rainfall and runoff to predict flooding in urban environments[iv]. Although the modeling technique they used—two-dimensional dynamic and diffusive wave modeling—is beyond the scope of what we learned in this class, I thought it was very interesting to read about this other modeling technique. In addition to mathematical models, the researchers also constructed architectural models of urban building configurations to see how building placement affected rainfall and runoff, and they also took into account variations in topography. This article definitely gave me ideas of how it might be beneficial to pair mathematical models with physical models when trying to solve spatial issues.

Another article I looked at also used neural networks to forecast rainfall. Researchers in Argentina sought to predict the monthly rainfall in an arid region of the country in order to improve agricultural water resources management. Some interesting advice the researchers shared is as follows:

> When a rainfall series is being analyzed, it is important to make use of the simplest possible models. Specifically, the number of unknown parameters must be kept at a minimum. For forecasting problems, Bayesian analysis generates point and interval forecasts by combining all the information and sources of uncertainty into a predictive distribution for the future values[v].

Their advice to keep the models simple is something I tried to remember when creating my own models.

In addition to the three articles I just discussed, I looked at another article where the researchers used neural networks to forecast rainfall. The researchers in this study looked at four years of hourly data collected from 75 rain gauge stations throughout Bangkok, Thailand, in order to improve flood management[vi]. They found that while rainfall is the most important variable in predicting future rainfall, the wet bulb temperature, or "the lowest temperature that can be reached by evaporating water into the air," is a very important variable to consider when predicting rainfall[vii]. While the data set I examined did not contain information on the wet bulb temperature, this article was still very helpful. It explained the importance of reducing the size of the neural network if possible since a smaller network requires less training time.

The final article I looked at also discussed using neural networks to forecast rainfall data, but the most interesting lessons I learned from it were about the data collection and homogenization techniques used for a lot of weather data. This article discussed how weather data is often homogenized in order to account for weather station relocation, replacement of recording equipment, and other factors that might affect the data[viii]. The article also discussed that lighthouses have provided the most consistent time series data in Australia because they are located far from, and therefore not affected by, urban heat islands. These are all interesting things I will remember when working with weather data in the future.

Overall, the studies and articles I reviewed helped me form ideas on which kinds of models to create (neural network seems to be the most popular method), as well as reinforce how important predicting rainfall is.

Forecasting Models

For this analysis, I created several models but will focus on just four in this report: a multiple linear regression model, a Holt-Winters additive model, a neural network model, and a seasonal auto ARIMA model. I chose to build a multiple linear regression model and a Holt-Winters additive model because I am comfortable with these modeling techniques and think it is easier to compare more recently learned modeling techniques against already familiar techniques. That being said, I chose a neural network model and a seasonal auto ARIMA model because I am not yet comfortable with these methods and want to get better at creating and understanding them.

I formulated each model using the training and test sets in order to get an idea which model performed best, but I made sure to include all the aggregated data in each model's formula prior to deploying the models for Kaggle. The reason I deployed each model using the full set of aggregated data rather than the training data was because the training set contained data only up until June 1997. I did not think that any model could accurately predict the rainfall for the target months, August 2014 to July 2016, if the data set it used was missing the 17 most recent years of data.

In **Figure 3**, we can see the forecasts for each model in blue against the actual rainfall data in red. As the plots show, none of these models did a really great job predicting the rainfall that was in the test set.

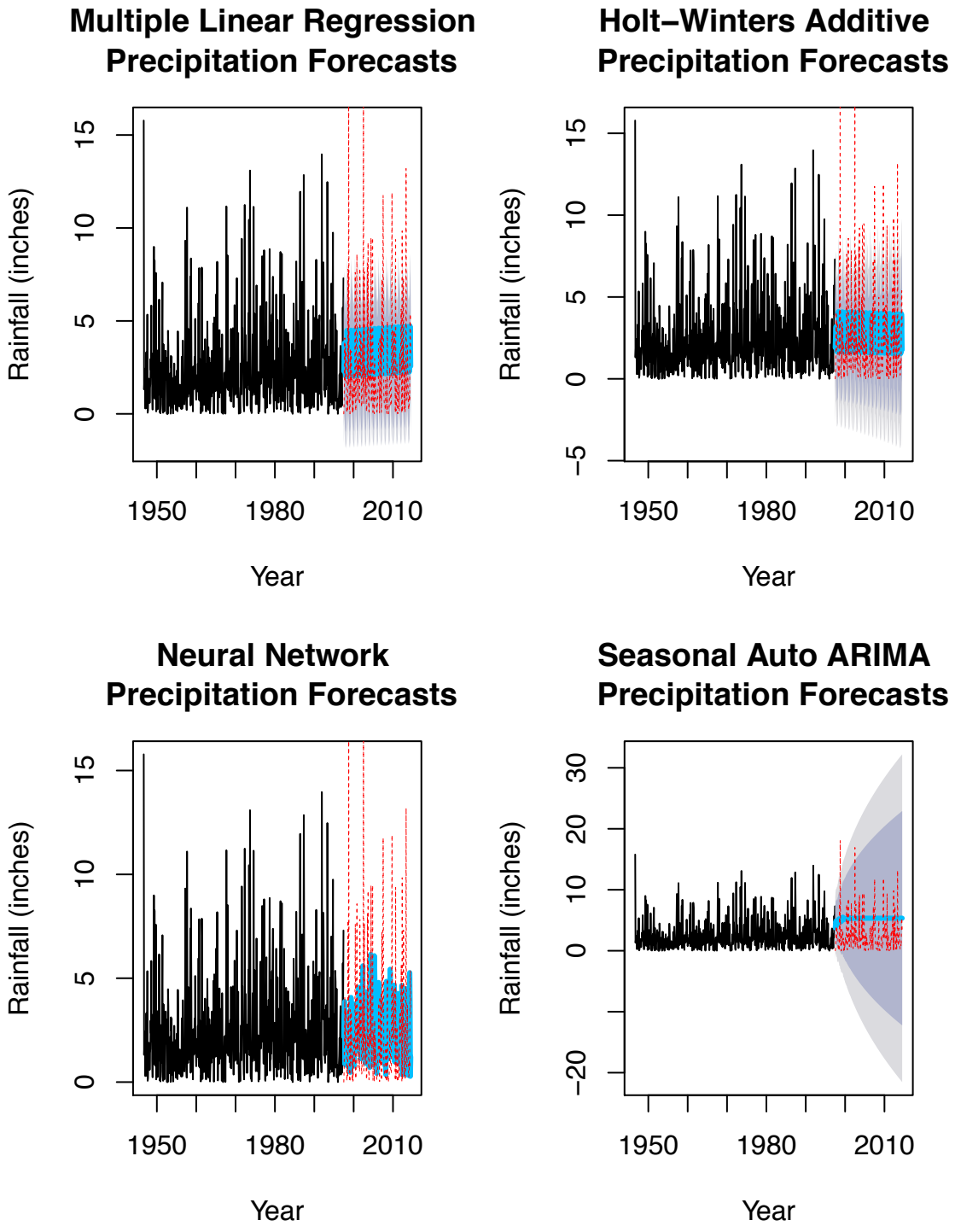

Figure 3: Predicted v. Actual Plots for Train/Test Models

By only looking at the plots, we can infer that the neural network model is probably most accurate while the seasonal auto ARIMA model is least accurate. The forecast fluctuations for the neural network model more closely mimic the actual data while the forecasts for the seasonal auto ARIMA model appear to be a nearly straight line. None of the models, however, seem to really capture the fluctuations in the data. The large confidence intervals shown in **Figure 3** also suggest that while these models are able to make predictions, they are not very good at predicting the actual values.

```
Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 1.2162853  0.3625939   3.354    0.000846 ***
trend       0.0012870  0.0005324   2.417    0.015933 *
season2     0.2932228  0.4584886   0.640    0.522716
season3     0.0064456  0.4584896   0.014    0.988788
season4     0.9969233  0.4584911   2.174    0.030070 *
season5     2.4705382  0.4584933   5.388 0.000000102 ***
season6     2.0133689  0.4584961   4.391 0.000013333 ***
season7     0.2951765  0.4607751   0.641    0.522023
season8     0.9134895  0.4607754   1.983    0.047881 *
season9     1.8065206  0.4584933   3.940 0.000091070 ***
season10    1.4148414  0.4584911   3.086    0.002124 **
season11    0.5294368  0.4584896   1.155    0.248658
season12    0.1234439  0.4584886   0.269    0.787837
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.315 on 597 degrees of freedom
Multiple R-squared:  0.1194,    Adjusted R-squared:  0.1017
F-statistic: 6.745 on 12 and 597 DF,  p-value: 0.00000000001882
```

Figure 4: Multiple Linear Regression Results

For the multiple linear regression model, I forecasted the amount of rainfall using the trend and seasonal components of the data. In **Figure 4**, we can see that the adjusted R-squared value is very low: 0.1017. This value means that only 10.17 percent of the variation in the monthly rainfall is accounted for by this model.

Looking at accuracy metrics and residuals will help us determine which of these models is best at forecasting rainfall since the plots are not enough to make these judgments.

Method	Data Set	ME	RMSE	MAE	MASE	ACF1
Multiple Linear Regression	Train	-2.68035E-16	2.290449	1.655367	0.7063969	0.04010777
	Test	-0.4270609	2.94144	2.182045	0.9311471	0.06127568
Holt-Winters Additive	Train	-0.01251963	2.321727	1.697192	0.7242449	0.02335104
	Test	0.1486201	2.906965	2.01599	0.8602863	0.05985043
Neural Network	Train	0.000955847	0.2564389	0.1829876	0.07808656	0.06307684
	Test	-0.024105368	3.1397222	2.2570505	0.96315424	0.04182
Seasonal Auto ARIMA	Train	-0.02662744	2.567683	1.874101	0.7997376	0.01135866
	Test	-2.6587013	3.98977	3.597793	1.5352911	0.10217627

Table 1: Accuracy Metrics

The accuracy metrics in **Table 1** provide very mixed results. The dark green indicates which modeling method performed best on the test set, while the light green indicates which modeling method performed best on the training set. Interestingly the neural network performed best in training but not in testing. The table shows that the Holt-Winters additive model performed best on the test set. These mixed accuracy metrics do not instill confidence in any of the models.

Because the metrics were very mixed, I decided that looking at residual plots might help clarify which model is best. The plots in **Figures 5, 6, 7,** and **8**, however, do not definitively steer us to one model over the others. The residuals for all models appear random, have a mean zero, and generally do not indicate any cause for concern.

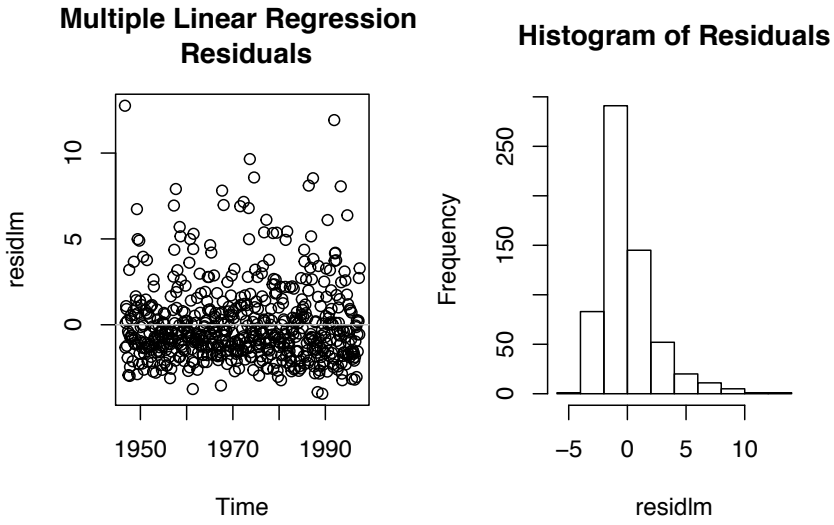

Figure 5: Multiple Linear Regression Model Residuals

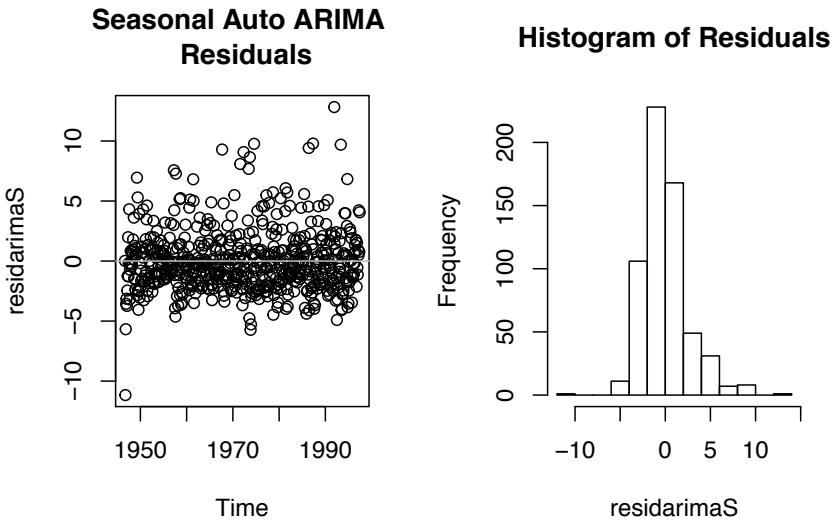

Figure 6: Seasonal Auto ARIMA Model Residuals

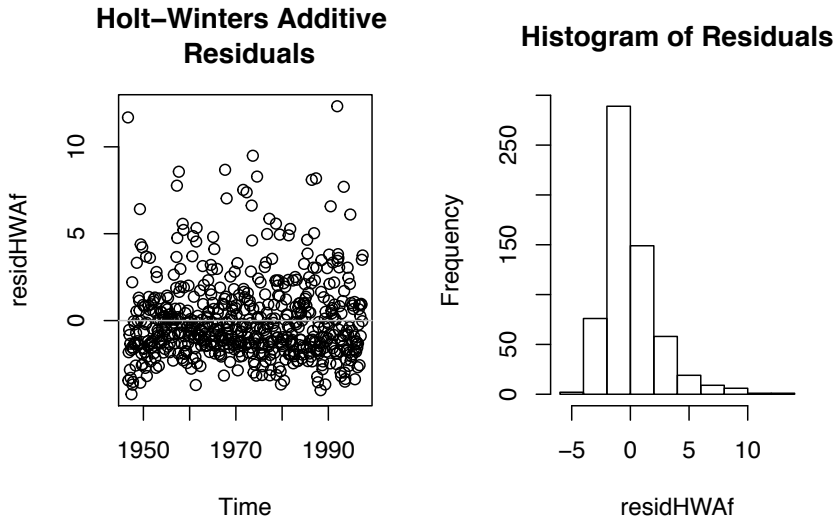

Figure 7: Holt-Winters Additive Model Residuals

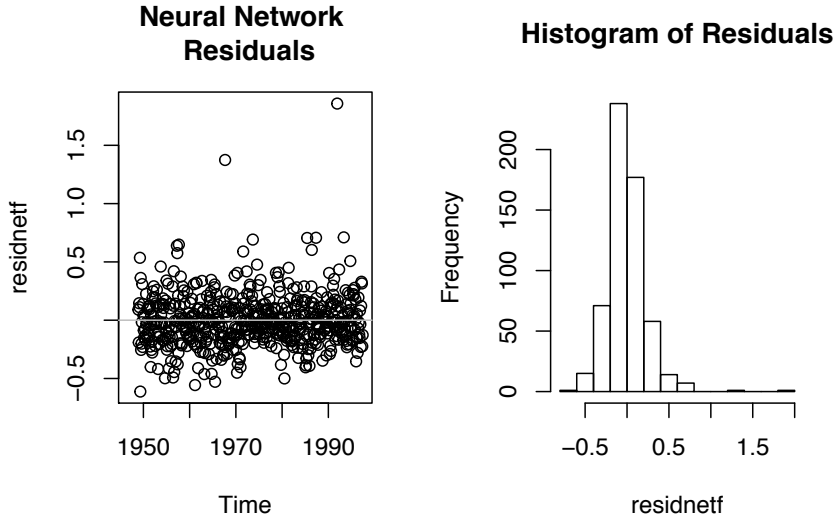

Figure 8: Neural Network Model Residuals

None of the plots, metrics, or residuals indicated a standout winner. After modifying the code so that the models used the entire aggregated data set rather than the training set, I submitted each model to Kaggle. I was surprised that the multiple linear regression model performed the best.

Model	RMSE
Multiple Linear Regression	2.41164
Holt-Winters Additive	2.45272
Neural Network	2.47538
Seasonal Auto ARIMA	2.71574

Table 2: Kaggle Results

As **Table 2** shows, the multiple linear regression model had the lowest RMSE, while the seasonal auto ARIMA model had the highest. The mixed messages the plots, accuracy metrics, and Kaggle results suggest that predicting rainfall is not easy and that I perhaps need to try more modeling methods or techniques.

Limitations, Lessons Learned, and Looking Forward

While working with this data, I encountered several limitations. The data set provided had ample data, but it only allowed us to work with one variable: precipitation. The other variables provided, cloudiness and wind speed, were completely useless because they were missing more than half of their data. Although we have created many time series models using just one variable throughout this course, having more variables with complete or nearly complete data would have opened up more modeling possibilities. Precipitation is tied to so many other factors that it seems wise to consider other variables when creating models. Not knowing where the rainfall was recorded (was it in Chicago? Kansas? Ecuador?) also seemed to place a limit on this project. Although we do not need to know where the rainfall was recorded in order to create models, knowing this would help place what we are predicting in context.

I think all of the models I created have one main limitation: they do not accurately and consistently predict rainfall. As I discussed throughout this report, none of the models was a standout winner. The plots, accuracy metrics, and Kaggle results all pointed at different models. This inconsistency would make me reluctant to deploy any of these models in a real-world situation.

In spite of the inconsistent and poorly performing models, I did learn a lot from this project. I learned how to subset, impute, and aggregate data. I also figured out how to verify that the data I had contained consecutive dates. Figuring out how to do these things enabled me to create models with greater confidence and ease.

Although I did not end up using the wind and cloud variables—or the data set I created containing only complete records for each precipitation, cloudiness, and wind speed—I learned a lot from preparing the data.

I also learned a lot from formulating my models. I decided early on that aggregating the data would be the cleanest way to build models. Why work with 24,000 + records when I could work with a fraction of that? It seemed most logical to aggregate at the beginning since the data had to be aggregated eventually.

Another thing I learned is that although a test/train set was not required, using one demonstrated good practice and could probably help me narrow down which models to focus on. I learned though that I needed to use the entire aggregated data set while formulating the models I would submit to Kaggle. As I mentioned earlier, the training set stopped 17 years before the dates we had to predict. Using the training set for the Kaggle submissions would likely have led to very poor models.

Looking forward, I would really like to figure out how to create a zero-inflated model and a conditional random forest model. I struggled for hours trying to create these models and kept getting error after error. I ultimately commented out this code and focused instead on the models I already created. I hope to figure out these modeling techniques in the near future though because they have a reputation for being very effective.

Conclusions

The models I created did not do a great job at predicting the rainfall for August 2014 to July 2016. As I mentioned earlier, the plots, accuracy metrics, and Kaggle results did not clearly favor one model over the others. The mixed messages my models were sending indicates that I need to spend more time creating new models or fine-tuning the ones I already have. Perhaps I need to isolate and cap outliers to create stronger forecasts as well. Although I spent a lot of time on this project, there are still so many techniques I can try that might produce a more reliable model.

References

[i] Yan, H. (2016, August 22). *Louisiana's mammoth flooding: By the numbers*. Retrieved from

[ii] Samenow, J. (2016, August 18). *'They didn't warn you': Louisiana disaster reveals deep challenges in flood communication*. Retrieved from https://www.washingtonpost.com/news/capital-weather-gang/wp/2016/08/18/they-didnt-warn-you-louisiana-disaster-reveals-deep-challenges-in-flood-communication/

[iii] Mekanik, F., Imteaz, M.A., Gato-Trinidad, S., & Elmahdi, A. (2013, October 30). Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. *Journal of Hydrology*. Retrieved from http://www.sciencedirect.com.turing.library.northwestern.edu/science/article/pii/S0022169413006227

[iv] Cea, L., Garrido, M., & Puertas, J. (2010, March 1). Experimental validation of two-dimensional depth-averaged models for forecasting rainfall–runoff from precipitation data in urban areas. *Journal of Hydrology*. Retrieved from http://www.sciencedirect.com.turing.library.northwestern.edu/science/article/pii/S0022169409007987

[v] Rodriguez Rivero, C., & Pucheta, J. A. (2014). Forecasting Rainfall Time Series with stochastic output approximated by neural networks Bayesian approach. *International Journal of Advanced Computer Science and Applications*. Retrieved from http://thesai.org/Downloads/Volume5No6/Paper_23-Forecasting_Rainfall_Time_Series.pdf

[vi] Hung, N. Q., Babel, M. S., Weesakul, S., & Tripathi, N. K. (2009, August 7). An artificial neural network model for rainfall forecasting in Bangkok, Thailand. *Hydrology and Earth System Sciences*. Retrieved from http://hydrol-earth-syst-sci.net/13/1413/2009/hess-13-1413-2009.pdf

[vii] Vermont State Colleges. (n.d). Wet Bulb Temperature. Retrieved from http://apollo.lsc.vsc.edu/classes/met130/notes/chapter4/wet_bulb.html

[viii] Marohasy, J., & Abbot, J. (2015, December 1). Assessing the quality of eight different maximum temperature time series as inputs when using artificial neural networks to forecast monthly rainfall at Cape Otway, Australia. *Atmospheric Research*. Retrieved from http://www.sciencedirect.com.turing.library.northwestern.edu/science/article/pii/S0169809515002124