1 Calculating Precipitation Error

We have talked several times about how to measure precipitation error. Below is a step by step description of how I calculate this in the "pointMicro.R" plots:

- Create a binary variable for precipitation. We had previously done this based on the following condition.

$$\texttt{Precip\_Actual} = \begin{cases} 1 & \texttt{PrecipitationIn} > 0 \\ 0 & \text{otherwise} \end{cases}$$

  However, there are cases in the data where there was a rain event that didn't produce enough moisture to be detected in the `PrecipIn` variable. These events however are indicated in the `Event` variable prompting the following updated designation for `Precip_Actual`.

$$\texttt{Precip\_Actual} = \begin{cases} 1 & \texttt{PrecipitationIn} > 0 \text{ OR } \texttt{Event} \text{ CONTAINS "Snow" or "Rain"} \\ 0 & \text{otherwise} \end{cases}$$

- Separate precipitation forecasts by the 10 unique values ($p_f = 0, 10, 20, \cdots, 100$).
- Determine the maximum forecasted prediction (from AM and PM forecasts) for each day. Denote this maximum prediction as `F_PrecipMAX`.
- Group forecasts by location and lag.
- For each set of grouped values, perform the following calculation for each of the 10 unique forecast values.

$$ERR_f = [\texttt{sum(Precip\_Actual[F\_PrecipMAX == p\_f])/sum(F\_PrecipMAX == p\_f)}] - \frac{p_f}{100}$$

  In other words, we compute the actual proportion of days that received rain for each p_f forecast and compare that to the particular p_f value in question. If a forecast were accurate, then we would expect the actual proportion of days to be close to the predicted forecast for those same days.

- Let $n_f = \texttt{sum(F\_PrecipMAX == p\_f)} \quad (f = 0, \cdots 10)$. Calculate the average error as a weighted average of these proportion comparisons i.e.

$$\texttt{ErrALL} = \frac{\sum_{f=0}^{10} ERR_f * n_f}{\sum_f n_f}$$

- Similarly, compute

$$\texttt{absErrALL} = \frac{\sum_{f=0}^{10} |ERR_f| * n_f}{\sum_f n_f}$$

- Note that there are cases where $n_f = 0$ resulting in an undefined value for ERR_f. Given the use of the weighted average, these values are simply ignored (and trivially replaced with 0)

when calculating the precipitation prediction average. Note also that the above method does not provide a daily measure of accuracy as is found with max and min temps. However, I feel this is a more appropriate way to measure predictive accuracy rather than trying to compare two binary variables.

2 Data Preparation for Cluster Analysis

My work on clusters was simply an extension of the analysis and tools created by Jill. The final cluster results followed her methodology nearly exactly with a few exceptions outlined in below:

- Replacement of an outlier precipitation reading (more than 39 inches) at Oklahoma city with the precipitation value from a National Weather Service station (0.8 inches). This value was heavily skewing the scaled standard deviation measurement for this value in the clustering.
- Replace missing values for wind speed and visibility at Baltimore with values from Dover, DE, rather than Wilmington, NC. We originally thought that this was Wilmington, DE, which is why we used it originally.
- Replace cloud cover AND visibility measurements at Austin, NV with value from Las Vegas, NV. The visibility measurements for Austin were actually missing values erroneously recorded as 0.
- Made adjustments to the following variables in the cluster analysis:
  - Cluster analysis included the mean and standard deviation for each measured variable.
  - Latitude and longitude were excluded as I wanted to see if clusters were preserved without this raw information.
  - The `arpt.` variables were also removed as they were (in most cases) a repeat of the location information of the original locations.
  - The forecast error variables were removed and replaced with the actual temperature measurements. My reasoning with this was to see if the cluster analysis could identify the same groups of locations that we have previously identified in our error analysis last semester.
  - The `Mean_` variable measurements were removed for various measurements as they were near perfect linear combinations of their respective `Max_` and `Min_` measurements.
  - Almost all locations would list a seemingly arbitrary value of 10 to indicate `Max_Visibility`. This value was near constant across all days. However, there were a few stations that used a different number than 10 to indicate max visibility. This led to a scaled version of the variable where almost all locations had value of 0 with a few wild outliers. For these reasons, this variable was removed prior to clustering.

3 pointMicro.pdf Maps

The "pointMicroTest.pdf" files in the images folder represent preliminary attempts to visualize the accuracy information across clusters. Each document contains maps of the US highlighting each cluster (an adaptation of Jill's cluster plot code). Then, the mean absolute error for maximum temperature, minimum temperature, and precipitation are plotted in the right three columns respectively. Correlations between each of the accuracy measurements (using Kendall's τ) within each cluster are located beneath each map. In this version, the clusters are ordered by their average longitude, and observations within each cluster are likewise ordered by longitude. I did this because longitude had a fairly strong overall correlation for each accuracy measurement. Specifically, locations farther north had less accurate forecast measurements for all three variables. This can likewise be seen in each of the listed plots. The four images represent different combinations of cluster size (5 or 6) and distance (Euclidean vs Manhattan).

The Euclidean and Manhattan distances provide fairly similar cluster results. The biggest difference being found in the clustering of locations in the midwest. I still need to do more work to determine the similarity of clusters when I eliminate the variables I proposed cutting out in a previous email.