Lake Michigan Influences

Blake Wallace
Capstone Technical Report

May 18, 2019

Data science objectives:

1. Are there any statistically significant differences in temperature closer to the water?
2. Can a predictive model that explains at least 80% of the variance in the precipitation differences be constructed?

Data Sources:

There are four data sources presented here. Three of them were crucial elements during the execution of the current model and comparison analyses presented below. The fourth, Station FSTI2 buoy, was initially used and then discarded for the Botanical Garden data because of a lack of long term weather measurements. However, in the next iterations of this project, when temperature fields are constructed, this buoy will be incorporated into the model construction.

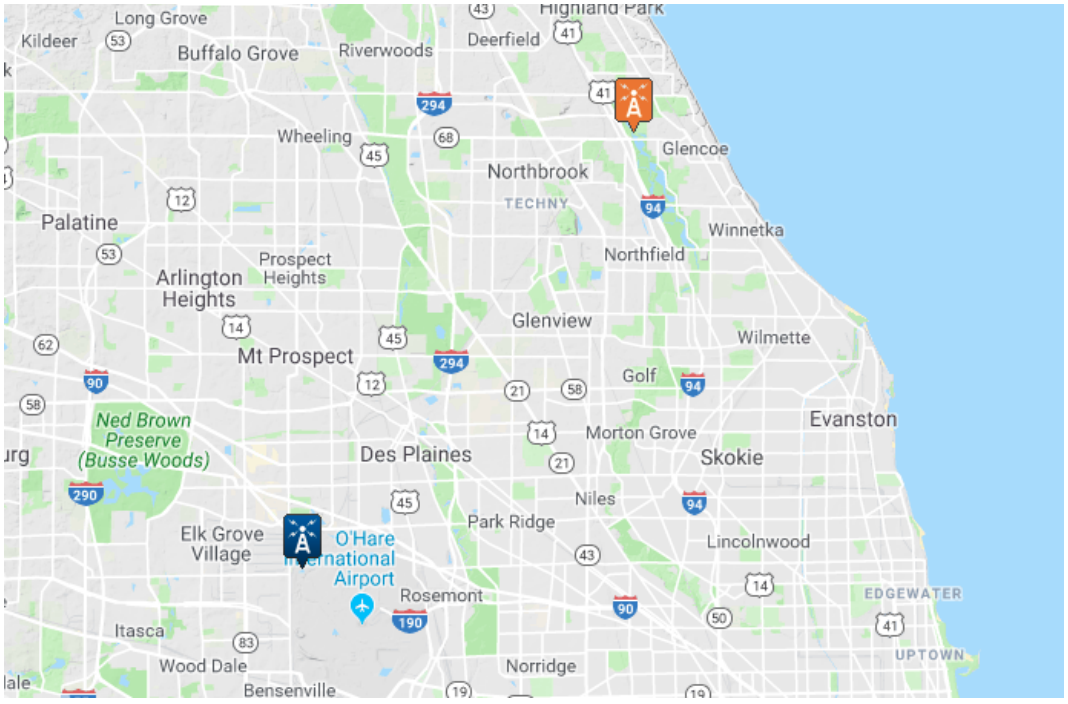


Figure 1: In the top right is the location of the weather tower inside of the Chicago Botanical Gardens, while the bottom left shows the location of the tower in the O'Hare Airport. Photo generated by the GHCND Search engine.

CHICAGO OHARE INTERNATIONAL AIRPORT, IL US
- Source: National Centers for Environmental Information
- GHCN (Global Historical Climatology Network) Daily Documentation
- ID GHCND:USW00094846

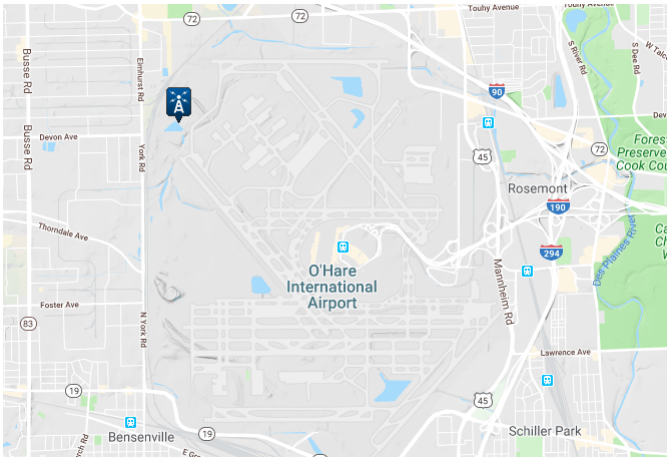


Figure 2: In the top left is the location of the weather tower inside of O'Hare Airport. Photo generated by the GHCND Search engine.

- 41.995 N 87.9336 W
- Airport Information

CHICAGO BOTANIC GARDEN, IL US

- Source: National Centers for Environmental Information
- GHCN (Global Historical Climatology Network) ? Daily Documentation
- ID GHCND:USC00111497
- 42.13987 N 87.78537 W
- Garden Information

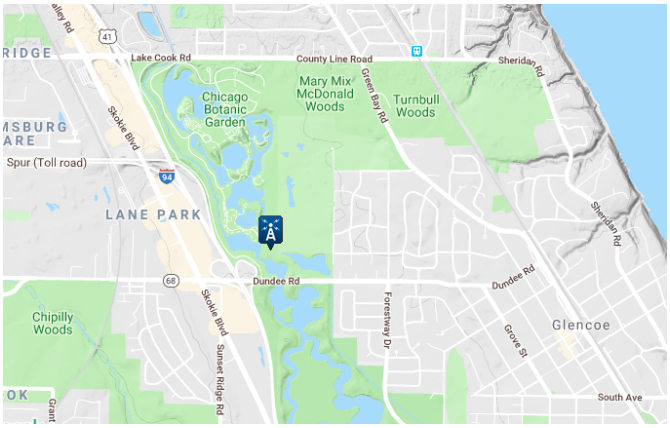


Figure 3: The weather tower at the Chicago Botanical Gardens. Photo generated by the GHCND Search engine.

Lake Michigan

- Source: Great Lakes Statistics: Average Surface Water Temperature from the Great Lakes Surface Environmental Analysis (GLSEA)
- 44.0 -87.0 (44 00' 0.00" N 87 00' 0.00" W)
- Data Set for 2018

Station FSTI2 - Foster Ave., Chicago, IL

- Source: National Data Buoy Center
- Owned and maintained by Chicago Park District
- 41.976 N 87.648 W (4158'35" N 8738'51" W)

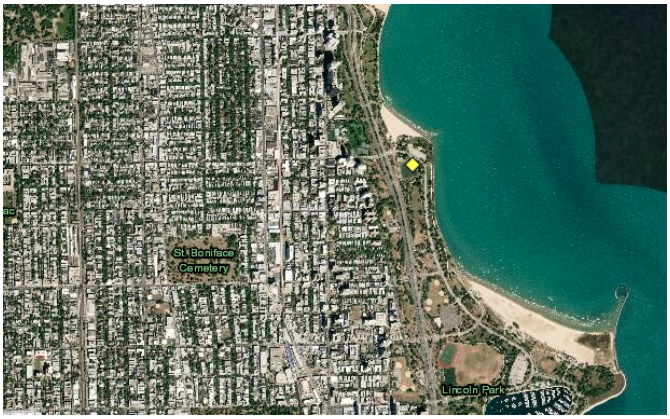


Figure 4: In the top right, on the yellow diamond, is the location where the FSTI2 buoy measurements are taken. Photo generated by the NOAA's National Data Buoy Center.

Data

There are 20 variables listed here. Some are used for partitioning the data and graphing. Others were pertinent to the modeling process and comparison analysis.

There were two methods of collection used to gather the data for this project. First, the data for average daily temperatures of Lake Michigan was scraped from the website for the Great Lakes Statistics: Average Surface Water Temperature from the Great Lakes Surface Environmental Analysis (GLSEA). This work is contained in the notebook here. Second, the GHCN database offers multiple methods for downloading data. The data for the Chicago Botanical Gardens and for O'Hare Airport was downloaded directly as comma separated files, and then merged in the notebook Merging Data.

Links

Data Dictionary
Bouy Data Dictionary
O'Hare Airport Data Dictionary
Botanical Garden Data Dictionary

"The five core values are:"

ohare_prcp - Precipitation (PRCP) (inches)
ohare_snfall - Snowfall (SNOW) (inches)
ohare_sndpth - Snow depth (SNWD) (inches)
ohare_maxtmp - Maximum temperature (TMAX) (Fahrenheit)
ohare_mintmp - Minimum temperature (TMIN) (Fahrenheit)

Other Features

lake-temp - Average Daily Surface Water Temperature for Lake Michigan (Fahrenheit)
garden_prcp - Precipitation (PRCP) (inches)
garden_maxtmp - Maximum temperature (TMAX) (Fahrenheit)
garden_mintmp - Minimum temperature (TMIN) (Fahrenheit)
garden_tobs - Temperature at time of observation (TOBS) (Fahrenheit)
ohare_wspd - Average daily wind speed (AWND) (miles per hour)
ohare_atmp - Average Temperature (TAVG) (Fahrenheit)
ohare_w2dir - Direction of fastest 2-minute wind (WDF2) (the direction the wind is coming from in degrees clockwise from true N)
ohare_w2spd - Fastest 2-minute wind speed (WSF2) (miles per hour)

Feature Engineering

target - absolute difference between the precipitation measurements at Ohare and the garden (ohare_prcp - garden_prcp)

garden_didrain - categorical, 1 for yes, 0 for no

ohare_didrain - categorical, 1 for yes, 0 for no

garden_medtmp - Median daily temperature at the Garden/ midpoint between the max and min temperatures ((garden_maxtmp + garden_mintmp)/2)

ohare_medtmp - Median daily temperature at ohare/ midpoint between the max and min temperatures ((ohare_maxtmp + ohare_mintmp)/2)

tmpdiff - difference between the median temperatures at ohare and the garden (ohare_medtmp - garden_medtmp)

Data Cleaning/Data Manipulation/EDA:

For this project there are two primary sets that are explored. The first pertains to the merging of the FSTI2 buoy data with the Lake Michigan and O'Hare data. This set spans the four year period starting January 01, 2015 and ending December 31, 2018. It contains 1461 observations, none of which are null. A notable fact pertaining to this dataset, figure 5 shows that wind speeds near Lake Michigan are negatively correlated with lake temperatures, having a Pearson correlation of -0.47. This linear correlation is stronger than any present between the wind direction at the FSTI2 buoy and any other considered variables measuring temperature. At first glance it would seem that this number captures short term fluctuations in the lake's temperature, represented by the squiggly curve seen in Figure 8. However, the size of Lake Michigan makes it is hard to imagine the windspeed at a single spot is indicative of the average temperatures of the entire lake. Instead, we believe it is more reasonable to assume that there are, on average, higher wind speeds during months when there is lower average daily water temperatures.

The second set that was considered included the merging of the Chicago Botanical Garden data with the O'Hare data spanning the years 1995 to 2018. There

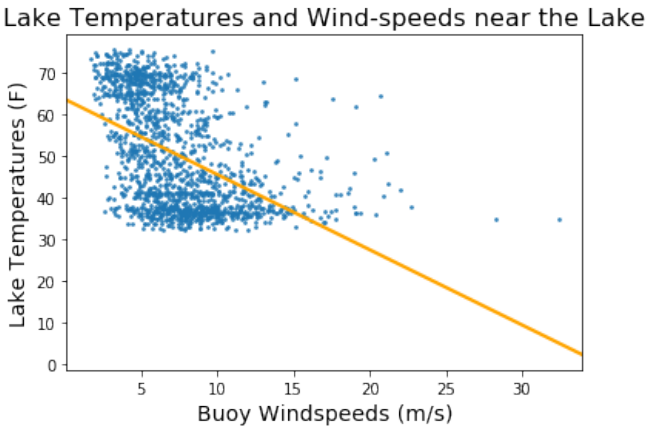


Figure 5: The linear relationship between the FSTI2 buoy wind-speeds and the Lake Michigan water temperatures between January 2015 and December 2018.

Analysis of the data

Tests and Evaluation:

In this section we will perform several two-tailed hypothesis tests, all of the form:

$$H_0 : \text{the average temperatures are the same,}$$
$$H_A : \text{the average temperatures are different.}$$

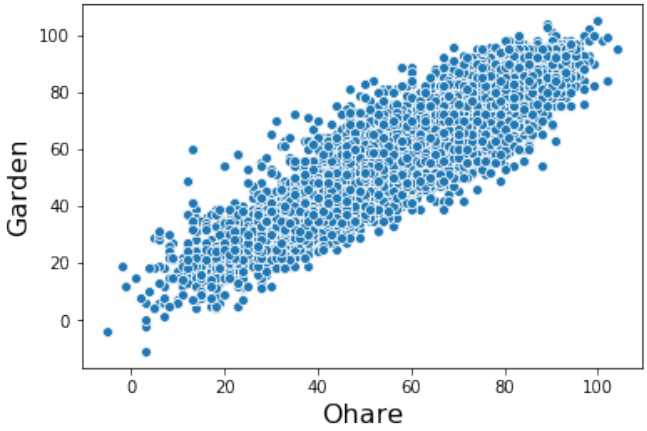
Table 1 shows the results from five different tests. The first row considers the dataset containing all of the weather measurements from January, 1995 through December, 2018. The second row represents when there is no rain at either location, the third row when there is rain at both locations, and the last two indicate when there is rain at only the airport, or only the gardens, respectively. As can be seen, with a very small p-value, there is a significant

Data	Quantity of Data	t-score	p-value	Significance	Gardens Avg (F)	Ohare Avg (F)
All Data	7923	0.5876	0.5568	None	59.24	59.43
No Rain	4022	3.285	0.0010	Yes	58.99	60.57
Both Rain	1648	-2.629	0.0086	Yes	59.48	57.7
ohareRain	1193	-1.9557	0.0506	None	59.06	57.43
gardensRain	1060	0.0904	0.9280	None	59.99	60.07

Table 1: Statistical Tests with Results

difference between the average maximum temperatures when it rains at both locations and when it is not raining at either location. This indicates that, given that there is no precipitation at either location, there is less than 0.1% chance of rejecting the Null Hypothesis incorrectly. Since this number is so small, there is evidence to reject the null hypothesis, indicating that there is most likely a difference in the average maximum temperatures between the two locations.

Figure 6: **Temperature/ Full Data (F)**



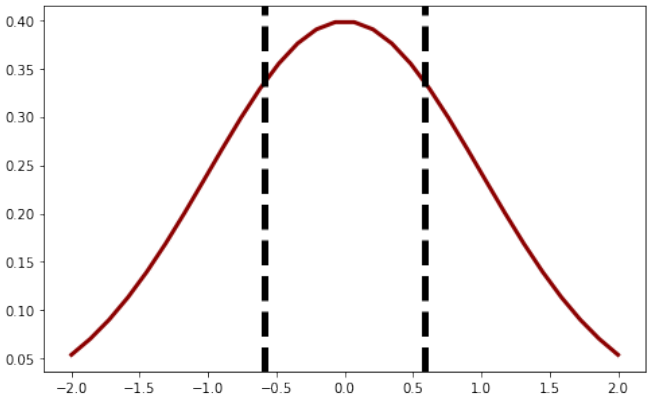


Figure 7: t-statistic for the full dataset

Models and Evaluation:

The second data science objective in this project is, "Can a predictive model that explains at least 80% of the variance in the precipitation differences be constructed?" To this date, the answer is no. Table 2 shows several of the models built with the larger data set containing measurements between 1995 and 2018 for the Chicago Botanical Gardens and the O'Hare Airport. To date the best constructed model was a random forest, which explained 11% of the variance on unseen data. There are other models that were constructed, which can be found in the jupyter notebook Modeling the Garden-Ohare-Lake Michigan data (1995-2018). Some of these other models include regularized linear models and feed forward neural networks. But, they exhibited such low scores they were not even included in the table below.

Table 2: Predictive Models with their scores

Model	Training score*	Testing score*	Training MSE**	Testing MSE**	Cross Validation
Linear no poly	0.0825	0.1052	0.0933	0.0683	0.0785
Linear gs	0.1222	0.1329	0.0893	0.0662	0.0984
Decision Tree	0.1139	0.0691	0.0901	0.0711	0.0429
Decision Tree gs	0.0937	0.0584	0.0922	0.0719	0.0450
Random Forest	0.8614	0.0517	0.0134	0.0724	0.0554
Random Forest	0.8711	0.1078	0.0131	0.0681	0.0770
Random Forest gs	0.8658	0.0905	0.0136	0.0694	0.0651
Random Forest	0.8677	0.0682	0.0135	0.0711	0.0660
Random Forest	0.8704	0.1153	0.0132	0.0676	0.0767
Random Forest	0.8080	0.0957	0.0195	0.0690	0.0787
Random Forest ada	0.9547	0.0549	0.0331	0.0722	0.0331
Random Forest ada	0.9445	0.0525	0.0056	0.0723	0.0283
Random Forest bag	0.6735	0.1130	0.0332	0.0677	0.0928
Random Forest bag	0.6705	0.1239	0.0335	0.0669	0.0943

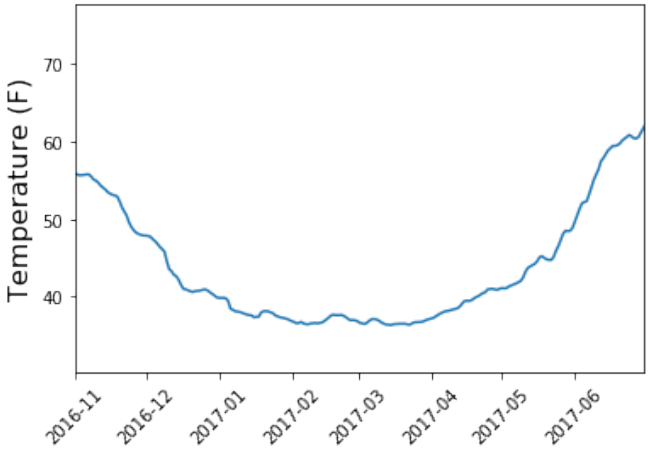
***** The score refers to the Coefficient of Determination.
****** MSE - Mean Squared Error
gs denotes a Grid Search was performed.
ada denotes an Ada Boost model was performed.

Future Iterations:

The presence of the statistically significant differences between the average maximum daily temperatures closer to the water and further away from the water is striking, and this must be the next line of investigation. It is likely that there are seasons of the year when more rain falls, traditionally in the Spring, when the Lake Michigan water temperature is usually close to it's lowest point. It is believed that, since the temperature difference when there is no rain has a lower temperature at the Botanical Gardens, that this must occur more often when the lake is at its lowest temperature and climbing. As depicted in Figure 8, the lowest point of the lake temperatures occurs around the end of winter, beginning of spring. This will be the next line of investigation.

Regarding modeling, the next step will be to investigate temperature fields around the Chicagoland area. To do this, it is necessary to consider data from many other locations. While the problem of incomplete data will be present, increasing the volume, and time-correlating it will mean that a less lengthy time period will be necessary to capture the temperature trends. The predictive element is present when extrapolating temperature in the holes between the specific locations. The spatio-temporal nature of this problem means a more sophisticated approach must be taken to predict temperatures in the gaps between locations where temperature measurements are made.

Figure 8: **Lake Michigan Daily Temps, Nov 2016 - June 2017**



Resources: