Capstone One: Statistical Analysis Report

Interactive documentation of this analysis can be found on Github in this jupyter notebook.

Because the project aims to forecast temperature, obscuration, and humidity, statistical analysis focuses on time series forecasting with three different prediction techniques: ARMA modeling, exponential smoothing, and generalized additive modeling. Models fit daily data for the three variables and then predict future values.

Some preliminary data processing takes place before modeling. Exploratory visualization previously revealed that the hours of 10 AM to 4 PM contained the best sky clarity and temperature for each day; for this reason, analysis considers only hours in this range across the dataset's complete years (2010-2018). Before investigating any of the models, the hourly humidity data needs to be resampled to occur at a daily interval; the resample method applies a mean to each day's 10 AM to 4 PM data to arrive at this figure. The temperature and obscuration data already occur at a daily interval.

Analysis continues from preliminary processing to ARMA process modeling. As ARMA modeling requires time series data to be time stationary, the first step of the analysis investigates the data's stationarity. A Dickey-Fuller test provides a p-value in response to the null hypothesis that the time series is not stationary; the data set's temperature, obscuration, and humidity time series all yield low p-values in the test. This indicates that all three series are stationary and can be fruitfully approached using ARMA modeling.

To get a sense of the proper order of ARMA models, the next stage of analysis examines the three variables' autocorrelation and partial autocorrelation plots. The minimized Bayesian Information Criterion (BIC) determines model orders, after fitting and assessing all possible ARMA models up to a maximum AR order of 4 and a maximum MA order of 2. Lastly, plots compare each variable's ARMA model forecast for the month of January, 2019 against recorded data for that month. Additional techniques that may add more model accuracy may include simple differencing or seasonal differencing in advance of modeling.

Next, exponential smoothing provides alternative forecasting models. Naive STL decomposition provides an initial view of the data's relevant linear and cyclic components. After this, Holt-Winters triple exponential smoothing models each variable's series. For each variable in the data set, plots compare several possible models with changes to the following two variables: (1) damped vs. undamped additive trend modeling, and (2) additive vs. multiplicative seasonal modeling. Damped additive trend and multiplicative seasonality performed best for temperature, while damped additive trend with additive seasonality performed best for humidity data; obscuration model results were ambiguous between damped and undamped additive trend with additive seasonality. An average root mean square error metric helps quantify model assessment. These plots compare forecast 2018 data against observed 2018 data.

The final technique examined utilizes Facebook's Generalized Additive Model, Prophet, to model and predict temperature, obscuration, and humidity values. This technique captures a trend with hierarchical seasonality through a combination of Fourier harmonic analysis with Bayesian smoothing on a piecewise linear model. Prophet's temperature and obscuration

forecasts yielded lower error than Holt-Winters exponential smoothing, while Holt-Winters smoothing narrowly outperforms Prophet’s humidity predictions.