Proposal: Predict Air Quality in China

Brenton Arnaboldi and Zach Zhang

Intro

Air pollution is a pressing issue in China. While the country's rapid economic growth in the last 30 years has improved overall quality of life, it has also led to unsafe levels of industrial pollution. According to a World Bank study in 2007, only 1% of China's 560 million city dwellers breathe air considered safe by the European Union.[1] Poor air quality is a major health hazard. One 2015 study estimated that outdoor air pollution contributes to the deaths of 1.6 million people in China each year.[2] Given the severity of this issue, a model to predict air pollution might be very useful. With an effective predictive model, local governments would be able to issue public warnings further in advance, giving the citizenry more time to prepare for high pollution.

Data

Over the past few years, China has collected hourly real-time pollutant data from 1497 stations across the country. The metrics measured include Air Quality Index (AQI), carbon monoxide, ozone, nitrogen dioxide, sulfur dioxide, particulate matter < 2.5 microns and particulate matter < 10 microns. For this project, we will seek to predict AQI, as AQI is the most common benchmark used to evaluate overall air quality. Specifically, the data range from 5/13/2014 to 12/31/2016. Theoretically, this would mean that each station has (964 days) * (24 hours) = 23,136 observations. However, there are many missing values. We are not exactly sure yet how we will deal with missing values.

In addition to the raw readings, we plan to incorporate different weather features. Weather features such as temperature, wind speed, and precipitation will all impact AQI. We plan on scraping historical weather data from Weather Underground. In a live setting, one could use the weather forecast that they provide to make predictions.

Experimental Design

If we are to build an autoregressive model to predict AQI at hour "h" in Beijing, but want to make our model useful from a real-world perspective, we should consider ignoring AQI values for all hours between "h" and "h-24" in Beijing. A useful model would be one that can make AQI predictions far into the future — a model that predicts AQI simply based on the previous hour isn't very useful.

Feature space will include 1) autoregressive AQI values 2) AQI values from other stations, and 3) weather. For #2, how many other stations should we consider? Will 1497 other stations

[1] "As China Roars, Pollution Reaches Deadly Extremes" (New York Times, August 26 2007).

[2] "Air Pollution in China: Mapping of Concentrations and Sources" (Robert A. Rohde, Richard A. Muller). Available http://berkeleyearth.org/wp-content/uploads/2015/08/China-Air-Quality-Paper-July-2015.pdf

(multiplied by “x” timestamps) result in overfitting? For #3, are we considering weather only for the target city, or also weather from other stations?

Models --

We will begin our modeling process with a univariate time series model for a given station. The model will likely incorporate autoregressive, seasonal, and trend components.

However, this approach fails to model to interconnectivity of the different stations (if station A and B are close then changes in A are likely to be reflected in B). We can model these connections using a latent variable model that jointly predicts all stations. In addition to this we believe that it is important to incorporate weather data to make accurate predictions on a longer time horizon.