Pollution and Temperature

Team 23: Aaditya, Albert, and Cindy
Boston Regional Data Open

February 9, 2019

Abstract

In this paper, we built a predictive model of land temperature at certain cities using pollution. In broad terms, we set out to quantify the ”greenhouse gas” effect. After performing some significant preprocessing, we experimented with using the different chemical concentrations as inputs to our one hidden-layer neural network model. We found that we achieved the best results when we filtered out superfluous pollutants that were irrelevant, because it made the learning task easier for our model. By combining some hand-crafted features (such as our spacetime-based, sinusoidal temperature adjustment) with the learned features characteristic of the neural network, we were able to get a predictive model with a mean squared error of just 0.14 on our test set. Thus, our modeling indicates that the gaseous pollutant concentration does have a correlation to relative temperature increases, lending evidence to what is commonly known as the ”greenhouse effect”.

1 Introduction

Interested in understanding how different the emissions of various human activities affect the environment, we attempted to understand how various gaseous pollutants affect local land temperature, utilizing datasets 3 and 6 provided to us for the Data Open. We found that all gases were present in a variable concentration over time, as seen in Figure 1 for Challam, WA, one of the 72 counties accounted for in the data. This variability implied that we had a plethora of data to utilize in modelling the correlation between gaseous concentrations and local land temperature.

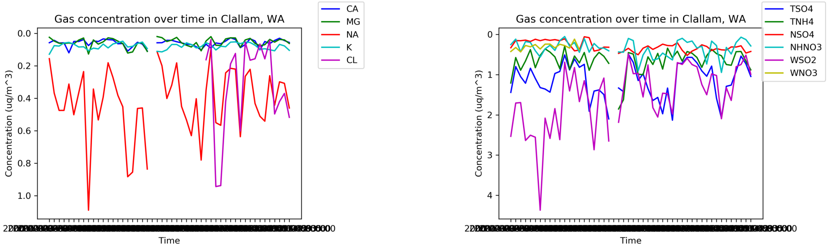

Figure 1: Gas concentration over time

2 Data Processing

In order to take note of these trends, we used Pandas to scrape the data sets provided and looked at the total power output in MW in various power plants. This data set included the city/state of interest as well as their latitudinal and longitudinal coordinates. We used this information to orient another data set, the `land_temperatures_by_city` dataset which included the city, but didn't have the state names. Cities of the same name in multiple states were checked for their relative distances to datapoint location to determine the correct state. Then, we used the `land_temperatures_by_city` dataset to determine the closest city of each observation in the `filter_pack_concentrations_weekly` dataset. The distance between cities was determined using the Haversine formula for calculating distance between two points on a sphere given latitude and longitude coordinates.

Next, in preparation for training the model, we created a few handcrafted features for temperature rescaling. For each observation in the filter concentrations dataset, we found the average temperature of all land temperature recordings which were from the same state and within the time frame of the filter recording. Filter pack concentration recordings which had no land temperature recordings in the given time frame were removed. In order to account for the variations in temperature caused by the change of seasons, we found the minimum and maximum at each city, and created sine function that matched the fluctuations from seasonal variations in temperature. We set the winter solstice, December 31st, as the minimum of the sine curve and the summer solstice, June 21st, as the maximum, and then adjusted temperatures accordingly to normalize them across different locations.

We used this processed data to develop a machine learning model to predict future energy trends in these areas of interest.

3 Modeling

To go from levels of gaseous pollutants to temperature, we realized two potential approaches. Using carefully handcrafted features that factor in the interdependence of the sources and then training a linear regression model was one option. However, instead of spending the large amount of time it would take to make these features, we decided to let the model ”learn” these features from the data. We did this by using a one-hidden layer neural network. We used ReLU activation for the hidden layer and linear activation for our outputs. The inputs were some subset of the gaseous pollutants and the output for the model was the normalized temperature. We trained this model on training pairs of gaseous pollutant information and temperature recordings from matching time and locations. We set aside 20% of our training dataset to use as a validation set to compare different models and evaluate our model's final performance. The final training set size was 8448 training pairs.

After experimenting with different learning rates for training, different dimensions of the hidden layer, and different model inputs. For example, in some runs, we didn't include CA, MG, NA, K, CL concentrations as we thought these would have little effect on the temperature but only ”confuse” the network. In the end, we found that higher hidden dimensions led to a lower validation error, as well as manually removing some chemical concentrations. Specifically, using a hidden dimension of size 20 and using solely the TSO4, TNH4, NSO4, NHNO3, WSO2, and WNO3 concentrations as input, our model achieved a test error of just 0.14 on the normalized temperature readings.

In the future, we hope to experiment with more model architectures and training hyperparameters to establish a stronger correlation. Furthermore, factoring other gaseous pollutants such as CO2 concentration, CH4 concentration, and H2O concentration, could yield better results. However, this information was not provided and unfortunately we didn't have time to look for these values.

4 Conclusion

We set out to understand the relationship between human pollution and the environment, with our first step being to model the correlation between chemicals picked up by Teflon and Nylon filters and local land temperature. We utilized the latitude and longitude to pair the relevant filters and temperature measures for each county, and trained a feed-forward neural network to predict local land temperature from pollut Interested in understanding how different the emissions of various human activities affect the environment, we attempted to understand how various gaseous pollutants affect local land temperature, utilizing datasets 3 and 6 provided to us for the Data Open. We found that all gases were present in a variable concentration over time, as seen in Figure 1 for Challam, WA, one of the 72 counties accounted for in the data. This variability implied that we had a plethora of data to utilize in modelling the correlation between gaseous concentrations and local land temperature.