Quantifying Local Energy Demand through Pollution Analysis

Cole Smith
Courant Institute of Mathematical Sciences
New York University
New York, USA
css@nyu.edu

Andrii Dobroshynskyi
Courant Institute of Mathematical Sciences
New York University
New York, USA
andrii.d@nyu.edu

***Abstract*—In this paper, we explore the process of quantifying and predicting energy demand given measurable emissions of airborne pollutants throughout the United States, from 2014-2019. Airborne pollutants often take the form of criteria gasses like Ozone, NOx, Carbon Monoxide, Sulfur Dioxide, and toxic compounds. These gasses are created by a mixture of atmospheric conditions, and fossil fuel emissions. Using daily emission measurements throughout the United States from the EPA, and spot prices for various Oil and Gas products from the EIA, we define a mapping between these two universes. Principally, we form a basis of analysis for the ecological signals behind Oil and Gas prices on a daily scale by defining a quantifiable demand score, based upon the Time-Lagged Cross Correlation (TLCC) between pollution levels, and commodity spot prices. For this, we provide a scalable, pure-Spark implementation of our TLCC algorithm. We find that these gasses exhibit a seasonal pattern, and a predominantly negative correlation to market oil prices for both the criteria gasses, and toxic compounds.**

***Index Terms*—air quality, oil and gas, spark, time series**

I. Introduction

Spot prices of petroleum, natural gas, and energy are dependent on a variety of influencing factors, and major factors are likely geopolitical in nature. However there does exist a regularity in these markets such that an increase in demand in the United States (which consumes about 25% of the world's oil and gas) often prompts a spike in prices by the Organization of the Petroleum Exporting Countries (OPEC).

It is also known that Ozone and NOx gasses are produced by the combustion of fossil fuels, those being oil, natural gas, and coal. By exploring daily pollution data, we may derive "hot spots" of high energy demand, relative to locations, and demand spikes, relative to time.

We approach the problem of quantifying demand by constructing multiple time-series of pollution data and spot prices, and then analyzing the moving trends between both to measure these "hot spots".

II. Motivation

Air pollution is often viewed as an implicit cost of fossil fuel use. Quantifying the dollar-impact of a given unit of energy on the the environment is typically difficult to compute in practice due to the various ways in which fossil fuel is used. Moreover, the focus of fossil fuel emissions is typically Carbon Dioxide, while other criteria, like Ozone for instance, are less-explored due to their more subtle role in ecological and climate change. However, it is worth exploring the impact of other gasses and particulates not only on our atmosphere, but on our communities. Ozone and particulate matter poses a health risk to sensitive groups, and rising temperatures will only increase the amount of "bad ozone" (Ozone that occurs near sea-level) for these groups.

Our goal for this project is to identify instances of increased fossil fuel usage through a relationship between their emissions and demand for energy, so that future work may be conducted in predicting, and assessing the underlying causes for such usage. Our hope is that this information may be helpful in deciding viable communities for renewable energy projects.

III. Related Work

A. Modelling oil spot price distributions for prediction

There have been multiple works attempting to predict the price of commodities such as petroleum with machine learning methods. Yu et al. [2] have presented a well-performing framework for timeseries prediction when dealing with spikes in the data. The paper specifically tested on crude oil price prediction and was able to get good generalization. The approach decomposed the time series into components that sufficiently explain the spiked movement and then fit a neural network on the extracted components to select the best generalizable ones. Yousefi et al. [1] did a similar approach with wavelets but with a focus on identifying market inefficiencies. Yang et al. [7] and Weron & Misiorek [9] showed promising results on electricity usage prediction, with [7] also using wavelet transforms for feature selection. A common thread was particularly promising results when accounting for non-stationarity and averaging out multiple models to better generalize to irregular movements, i.e. spikes that were the focus of paper [2]. Finally, Che and Wang [3] presented a particularly well-performing framework for short-term prediction of electricity with auto-regressive integrated moving averages, showing performance besting neural network-based methods in capturing non-linearities. Overall, accounting for spikes and utilizing moving averages showed great results in quantifying relationships and local patterns in oil price movements or spot prices for resources in similar

domain. In this work we also choose to focus on moving average techniques due to the modeling and experimental flexibility of re-computing simple moving averages (SMA's).

B. Learning from air pollution data

In a study by Jonathan Wilkey, et al.,the authors are able to accurately predict the emissions and air pollution of oil and gas production itself in Utah from 2015 to 2019. They do so by measuring Volatile Organic Compounds (VOCs), greenhouse gasses, and methane, against drilling activity, and validating with Monte-Carlo methods. Their results come within 5% of the actual predicted activity for drilling and hydraulic fracturing. Their approach exhibits the feasibility of our approach, although we will consider data from throughout the United States, rather than constraining ourselves to only air quality information around gas production. [6]

Mohamed and Ahmed in [5] show that Oil prices are predictable using economic signals alone. This is a natural approach, and one that the authors had success with. They express that the price of oil could be a successful indicator of GNP. The authors then predict daily oil prices using Gene Expression Programming (GEP) and Neural Networks. Our approach would like to quantify this information at a local level. It is critical to note that this paper considers predicting oil prices on the national level – As in, the spot prices which are seen within markets.

Different areas of the United States demand energy differently. From our paper survey, this has appeared as an ongoing research problem, and one that is rather difficult to quantify. While our approaches will use different data, their modeling approach is also successful, and the authors found that the GEP model outperforms existing Neural Network solutions to the same problem.

[8] outlines the predicted energy demand, and the geographical indicators that occur with a (predicted) increase in oil drilling and fracking. Moreover, it introduces the models that are used to predict the probability of drilling at certain sites, but does not go into details on those models. This is helpful for the purpose of potentially identifying drilling locations spatially as they do in the article. We would then use this information to assess the pollution indicators around these zones. In this way, we may evaluate the ecological impact on not only fossil fuel usage, but also its extraction and refinement, two factors which invariantly relate to its price.

IV. Datasets

As touched upon in earlier sections, we primarily use two datasets for analyzing energy demand - a fine-grain measure of air data that allows us to quantify pollution and various spot price data for oil that we use to measure demand.

A. AirData

Fields: Criteria Gasses (Ozone, NO2, Carbon Monoxide, Sulfur Dioxide), Toxic Compounds (NOy, NO, NOx, NMOC)

Metadata: Size: ∼350GB, Collection: Batch with support for real-time, Time Scale: Hourly from 1985-2019

Source: https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw

Schema: See Table II

Description: The dataset gives us detailed features for air quality based on location. There are many features (columns) we are using, but some examples are: longitude, latitude, Ozone levels, NOx levels. We have an ETL pipeline in place that transforms multiple measures from this dataset into normalized quantities independent of units.

Meta-analysis: The most important group from the AirData data sets is the Criteria Gasses. We begin by assessing the trends in air pollution by considering all spatial data points together, grouped by day of year, over the past five years to visualize the seasonality of pollution. In particular, we notice that most criteria gasses are seasonally correlated except for Ozone, which exhibits a negative correlation with the other criteria gasses.

See figure 2

Lastly, we assess the L2 Norm of all normalized criteria gasses, grouped by day of the year, and averaged over the past 5 years. We see that the signal itself appears to maximally correlate with the discrete Ozone signal. Thus, the majority of total pollution magnitude for a given day of the year can be explained by the Ozone level on that day.

See figure 3

B. Petroleum Data

Fields: Spot Price (USD), Date (format: YYYYmmDD)

Metadata: Size: ∼50MB, Collection: Batch

Source: https://www.eia.gov/opendata/qb.php?category=241335&sdid=PET.RWTC.D

Schema: See Table III

Description: The dataset is a daily indicator for the spot price of crude oil, in other words the market price that somebody is willing to pay for the resource. Available in daily granularity, we use this data to quantify some indirect demand for the resource. In turn, when aligning the spot price for oil with daily air quality, we can set up an analytic to find relationships in the two time series that we call `x_features` and `y_features`. The ETL pipeline in place for PetroleumData transforms the price column into a normalized quantity and computes extra columns based on the date column of the spot price.

V. Description Of Analytic

We store both data sets in HDFS, on the Dumbo cluster provided by NYU. Our main analytic application is built using SBT version 11.0 with Spark version 2.11, and Scala version 2.11.8. Our analysis is dependent upon the Spark-Core, Spark-SQL and MLLib libraries, as well as Apache Math for our statistical functions.

A. Time-Lagged Cross Correlation

We provide a ground-up implementation in pure Spark and Scala of a Time-Lagged Cross Correlation (TLCC) framework using Spark's `MLLib` package. This library assumes an X set of columns, a Y set of columns, and a list of integers representing the number of days, forwards or backwards, for which to slide each column in X.

Every column in X is shifted by each lag value, and the resulting lagged set of X is compared to the non-shifted columns in Y in an all-play-all fashion by computing the Cartesian product between the two sets. For signal correlation, we use Pearson's Correlation statistic.

We also provide a `TLCCIngress` class to easily verify the input DataFrames, and simplify the implementation of the `TLCC` class itself.

B. TLCC Performance

In testing, our implementation can complete 8 cross correlations with signal lengths of 90 days in one minute and 10 seconds. It takes 45 seconds to load AirData and EIA data from HDFS, and only **25 seconds** to compute the full cross correlations.

Computing the auto-correlation of the AirData gasses dataset (See table II) from 2014-2015 takes **2 minutes, 14 seconds**. Of this time, 1 minute and 25 seconds is spent normalizing a previously prepared DataFrame using the `bdad.etl.airdata.AirDataset` class. Thus, computing **64 cross correlations** with a signal length of **1825 days** takes only **45 seconds**.

All values were collected using the `time` output on NYU's Dumbo cluster using 30 executors.

C. TLCC Results

We found that the Criteria Gasses (Ozone, NO2, Sulfur Dioxide, and Carbon Monoxide) were generally more correlated with Oil spot-prices at 0 and 30-day lags then Toxic Compounds (NOy, NO, NOx, NMOC) were at the same lag values.

CO	-0.81	**NOy**	-0.32
SO2	-0.64	**NOx (30 Day Lag)**	-0.47
Ozone	0.81	**NO**	-0.31
NO2	-0.40	**NMOC (30 Day lag)**	-0.53

TABLE I
BEST CORRELATION TO PETROLEUM SPOT PRICE FOR 2019 (LEFT: CRITERIA CLASSES, RIGHT: TOXIC COMPOUNDS)

CO has the highest correlation to price, although the values of CO had the lowest variance of all the other signals, and auto-correlated with itself at a 30-day lag. Ozone had the highest correlation with price signals.

These results are inline with what we are expecting – As prices of petroleum increases, we expect quantity demanded to subsequently decrease, and therefore exhibit a negative relationship with the resulting pollution. In [4], Cavallo shows that oil prices often exhibit a similar trend due to production saturation by OPEC. This speculation however cannot be proven with our results alone. Future work needs to be conducted to assess the causal nature of our features.

VI. APPLICATION DESIGN

Our code base is provided in two main Scala packages. Namely we implement the `bdad.etl` and `bdad.model` packages, which provide functionality for loading cleaned data sets from the EPA and EIA, and performing analysis on those data sets respectively.

A. Design Diagram

See Figure 1

The overall process of our application starts by downloading the relevant information in batches. This is done so using a series of Python scripts, and then the files are placed into HDFS. The raw files are then read from HDFS by the classes in `bdad.etl`. The data from both the EIA and EPA need to be within the same format for use in our models. The transformation and normalization is handled along with data ingress in their respective ETL classes, described in detail below.

The cleaned data sets are then fed to `TLCCIngress` objects, and delivered to the `TLCC` class to compare the required correlations. After interpreting the correlation data, we form a scoring relationship from the best found lag, and more correlated air data features.

Lastly, we compute a Heatmap to be displayed by our UI, which reflects the assumed energy demand in local areas throughout the United States.

B. Extract, Transform, Load Design

1) AirDataset: The `AirDataset` class provides an intuitive interface to the otherwise unwieldy EPA AirData air quality data sets. In particular, we provide functionality for selecting the required date range and criteria to create specific *Scenarios* for analysis. The recorded measurements are spatial to a given latitude and longitude, the precision of which can be set by the class to average the pollution for either larger or smaller areas. For the purposes of this paper, we consider the Criteria Gasses and Toxic Compounds features. This class will first select only the required files from HDFS, pivot the selected criteria into feature columns, and normalize those selected features into a single `Vector` column for use in Modelling.

2) PetroleumDataset: Similar to the functionality of AirData, the `PetroleumDataset` class provides an interface for interacting with the data on crude oil spot price from the EIA (Energy Information Administration) data repository. A basic key functionality that it supports is selecting data based on a time range, which is most often yearly in granularity in order to align with the AirData. The main features of the class work with the the Price, tied to a particular date. The interface provided by `PetroleumDataset` helps with processing the Price column and generating new columns from the timestamp,

such as `year` and `dayofyear` for the purposes of joining when computing correlation analytics.

3) Scenarios: The Scenarios class provides specific *implementations* of the above ETL classes.

C. Model Design

1) TLCC: The `TLCC` class is the root of our analysis of the relationship between pollution and energy demand. *See Section V.*

2) CostRepresentation: The `CostRepresentation` class is the actuation of our TLCC results, and constitutes a representation of the dollar-value-impact of energy demand in a given area. *See Section VII.*

3) Heatmap: The `Heatmap` model class generates a spatial detail of the AirData information for use in our UI component. The magnitude of pollution for a given latitude and longitude is computed by taking the L2 norm of the criteria gasses. In this way, the Heatmap shown by the UI is closely related to the Cost Representation, but considers all measures of pollution, not only those correlated with petroleum spot-prices.

VII. Actuation, Remediation

Given the results found from our TLCC analysis (See section V-C), we now deduce a scoring measure for any given region with the current petroleum spot price(s) and past air quality data. We first define a measure of overall pollution as the L2 Norm of each (normalized) criteria gas moving averages for a given latitude/longitude pair, on a given day.

Before computing the L2 norm, the air quality moving averages are computed by taking the Simple Moving Average (SMA) over n periods, where n is the best lag found from the TLCC analysis. We assume this final value to be the *magnitude of increasing pollution*.

This is then divided by the petroleum spot price for the current day to get the dollar-impact of pollution. If there are multiple petroleum prices considered, they are again the L2 Norm of the normalized price signals. We consider this the *magnitude of current demand*.

Local Pollution Scoring Measure

Let $X :=$ Criteria Gas signals by Time and Location
Let $X_{\text{SMA}_n} :=$ SMA of X over n periods
Let $Y :=$ Petroleum Price signals by Time (for each Location)

For each row i in X and Y,

$$score := \frac{\|X_{\text{SMA}_n i}\|}{\|Y_i\|}$$

We consider this final score to model the magnitude of an increase in environmental damage per unit of current demand.

Scoring Interpretation

Using this score for a region, we can deduce the trend of the overall scoring signal for a given area of interest by computing the historical scores for that area. Moreover, one can compare regions by energy demand using this measure by comparing their scores. Notice that this is the same as comparing the regions by the pollution values themselves, since the numerator of the score is the same for all regions.

VIII. Analysis

A. User Interface (UI)

As part of our application, we've included a lightweight web application built to visualize the data outputted from our `Heatmap` model analytic generating spatial detail of the AirData information per longitude and latitude, as well as from our `CostRepresentation` model that generates local pollution scores.

The UI includes tools for controlling the rendering parameters of the two heatmaps, and the geo-map layers are fully interactive maps where one can zoom in/out for finer-level inspection of the heatmap patterns.

For an example UI, please see Figure 4

B. Visualization Analysis

In figure 4, we notice that our scoring heatmap appears similar to our pollution-only heatmap, but places more focus on major metropolitan areas. This is likely the case for a few reasons. The pollution heatmap is averaged over all of of 2019, while our score in this instance only considers our best found lag time of 10 days. Moreover, division by price will allow areas with less overall pollution to be represented by the heatmap. This allows one to consider more subtle movements in pollution scoring over 10-day periods.

IX. Conclusion

In this work, we have constructed a series of processes for quantifying localized energy demand from measurable air data across the United States. We have shown a dominant inverse correlation between market spot prices of oil and gaseous compound emission metrics. We have implemented a TLCC model as the foundation of our signal generation and quantified the relationship between time-lagged air data pollution levels and oil spot prices via a cost-score given per latitude/longitude of a location data point. Using the best performing lag value for the TLCC model found experimentally, our cost-score generation model, via the subsequently generated metrics, was able to pick up major metropolitan areas, placing more weight accordingly on the areas with higher values of quantified air pollution.

X. Future Work

Future work will need to be conducted on the feasibility of our scoring measure for analysis on a per-capita basis. We deliver gross scores for a given latitude/longitude pair, but as we see in our cost-score heatmap, this gross value will always be higher for high population-density areas.

We also consider looking at the cost-score distributions over time as future work. We think it may be useful to study the patterns of a cost-score given to a latitude/longitude pair changing over time per specific time lags and to measure consistency of pollution quantified by our metric across regions.

Acknowledgment

We would like to thank the NYU High Performance Computing Group for providing us access to their cluster, and for their continued support in maintaining a suitable environment for our application. We would like to especially thank Dr. Suzanne McIntosh for her guidance, support, and direction in developing this project.

References

[1] Shahriar Yousefi, Ilona Weinreich, and Dominik Reinarz. "Wavelet-based prediction of oil prices". In: *Chaos, Solitons & Fractals* 25.2 (2005), pp. 265–275.

[2] Lean Yu et al. "Oil price forecasting with an EMD-based multiscale neural network learning paradigm". In: *International conference on computational science*. Springer. 2007, pp. 925–932.

[3] Jinxing Che and Jianzhou Wang. "Short-term electricity prices forecasting based on support vector regression and auto-regressive integrated moving average modeling". In: *Energy Conversion and Management* 51.10 (2010), pp. 1911–1917.

[4] Alfred Cavallo. "Elephant in the room: How OPEC sets oil prices and limits carbon emissions". In: *Bulletin of the Atomic Scientists* 69.4 (July 2013), pp. 18–29. DOI: 10.1177/0096340213493583. URL: https://doi.org/10.1177/0096340213493583.

[5] Mohamed M. Mostafa and Ahmed A. El-Masry. "Oil price forecasting using gene expression programming and artificial neural networks". In: *Economic Modelling* 54 (2016), pp. 40–53. ISSN: 0264-9993. DOI: https://doi.org/10.1016/j.econmod.2015.12.014. URL: http://www.sciencedirect.com/science/article/pii/S0264999315004101.

[6] Jonathan Wilkey et al. "Predicting emissions from oil and gas operations in the Uinta Basin, Utah". In: *Journal of the Air & Waste Management Association* 66.5 (2016). PMID: 27064908, pp. 528–545. DOI: 10.1080/10962247.2016.1153529. eprint: https://doi.org/10.1080/10962247.2016.1153529. URL: https://doi.org/10.1080/10962247.2016.1153529.

[7] Zhang Yang, Li Ce, and Li Lian. "Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods". In: *Applied Energy* 190 (2017), pp. 291–305.

[8] Brad D. Wolaver et al. "An Improved Approach for Forecasting Ecological Impacts from Future Drilling in Unconventional Shale Oil and Gas Plays". In: *Environmental Management* 62.2 (Apr. 2018), pp. 323–333. DOI: 10.1007/s00267-018-1042-5. URL: https://doi.org/10.1007/s00267-018-1042-5.

[9] Rafal Weron, Adam Misiorek, et al. "Forecasting spot electricity prices with time series models". In:

XI. Figures

year(dateGMT)	IntegerType	Year of sample collection [2014-2019]
dayofyear(dateGMT)	IntegerType	Day of year of sample collection [1-365]
Reactive oxides of nitrogen (NOy)	DoubleType	Normalized Sample Measurement [-1,1]
Nitric oxide (NO)		
Oxides of nitrogen (NOx)		
Total NMOC (non-methane organic compound)		
Ozone		
Nitrogen Dioxide (NO2)		
Carbon Monoxide		
Sulfur Dioxide		

TABLE II
FORMATTED EPA AIRDATA SCHEMA (CRITERIA GASSES AND TOXIC COMPOUNDS)

year(dateGMT)	IntegerType	Year of spot price collection [2014-2019]
dayofyear(dateGMT)	IntegerType	Day of year of spot price collection [1-365]
price	DoubleType	Normalized Price [-1,1]

TABLE III
FORMATTED EIA PETROLEUMDATA SCHEMA (DAILY CRUDE OIL SPOT PRICE)

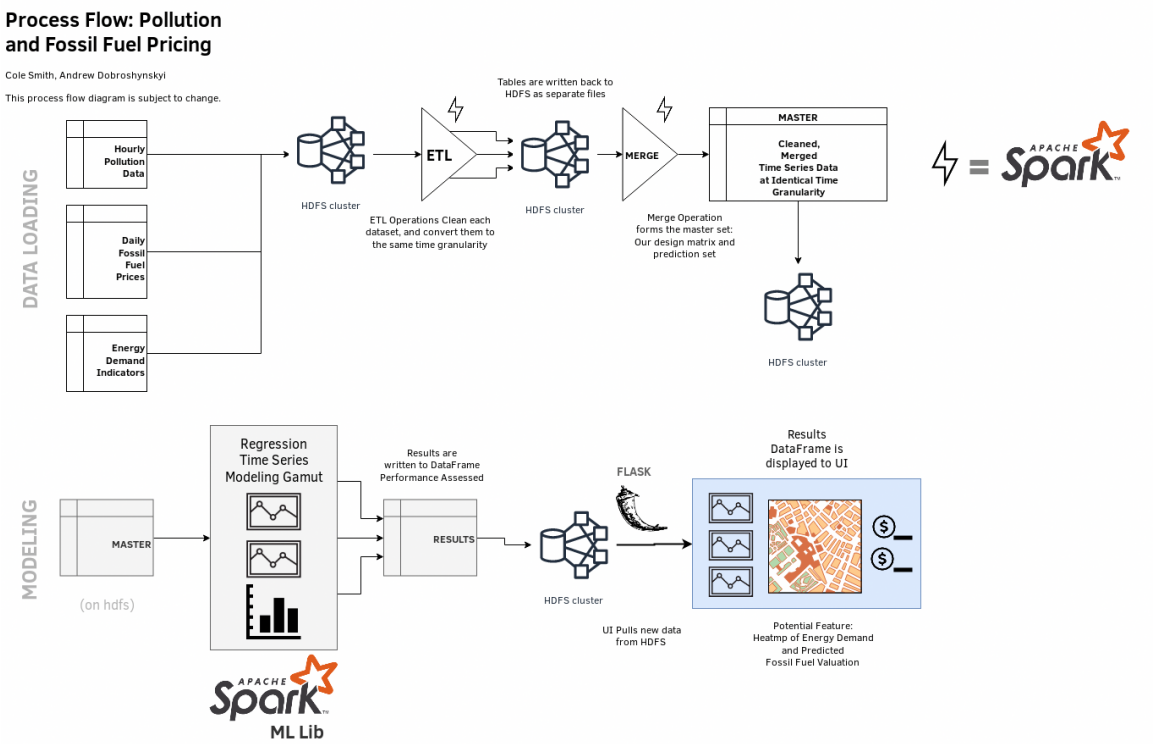

Fig. 1. Application Design Architecture: ETL and Model

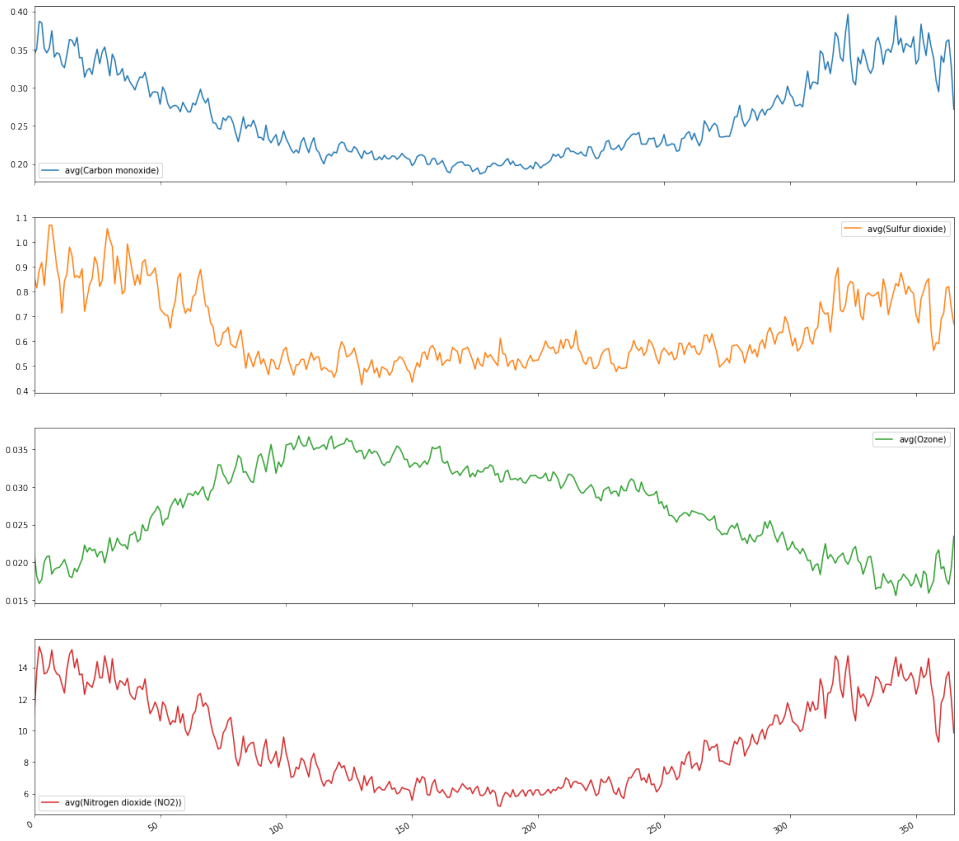

Fig. 2. Average Pollution Levels for Day of Year Averaged over 5 years

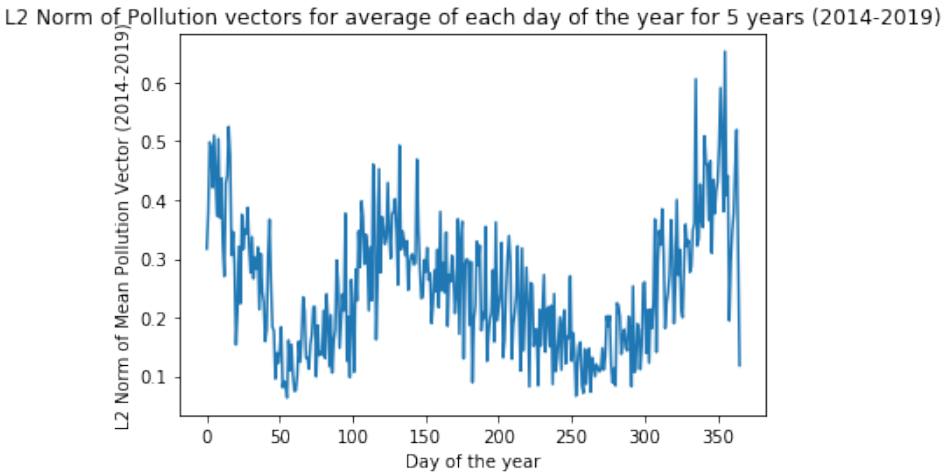

Fig. 3. Average Pollution Levels for Day of Year Averaged over 5 years, L2 Norm

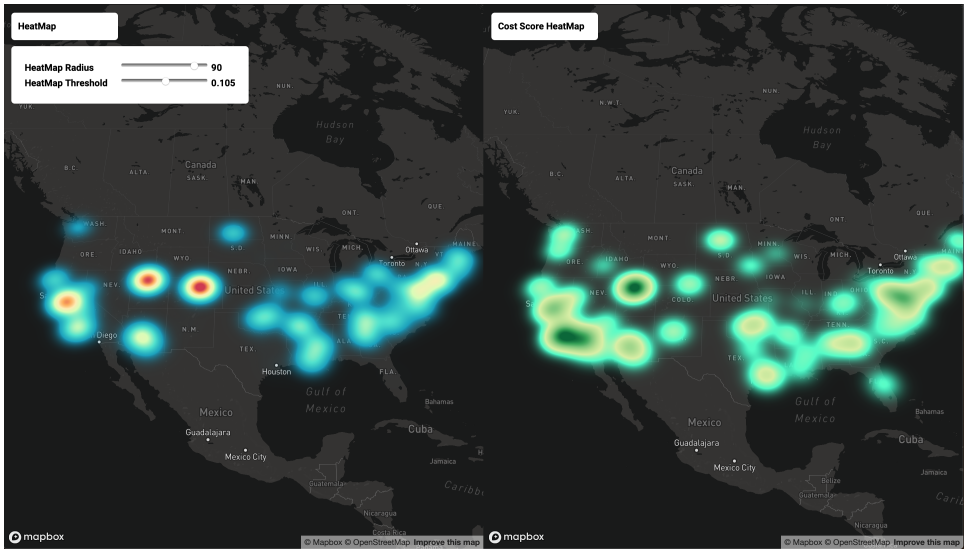

Fig. 4. UI For HeatMap Viz of Geo-Spatial Pollution Data (L2 Norm)