IJCAI ECAI
VIENNA 22

Deciphering Environmental Air Pollution with Large Scale City Data

Mayukh Bhattacharyya*, Sayan Nag* and Udita Ghosh

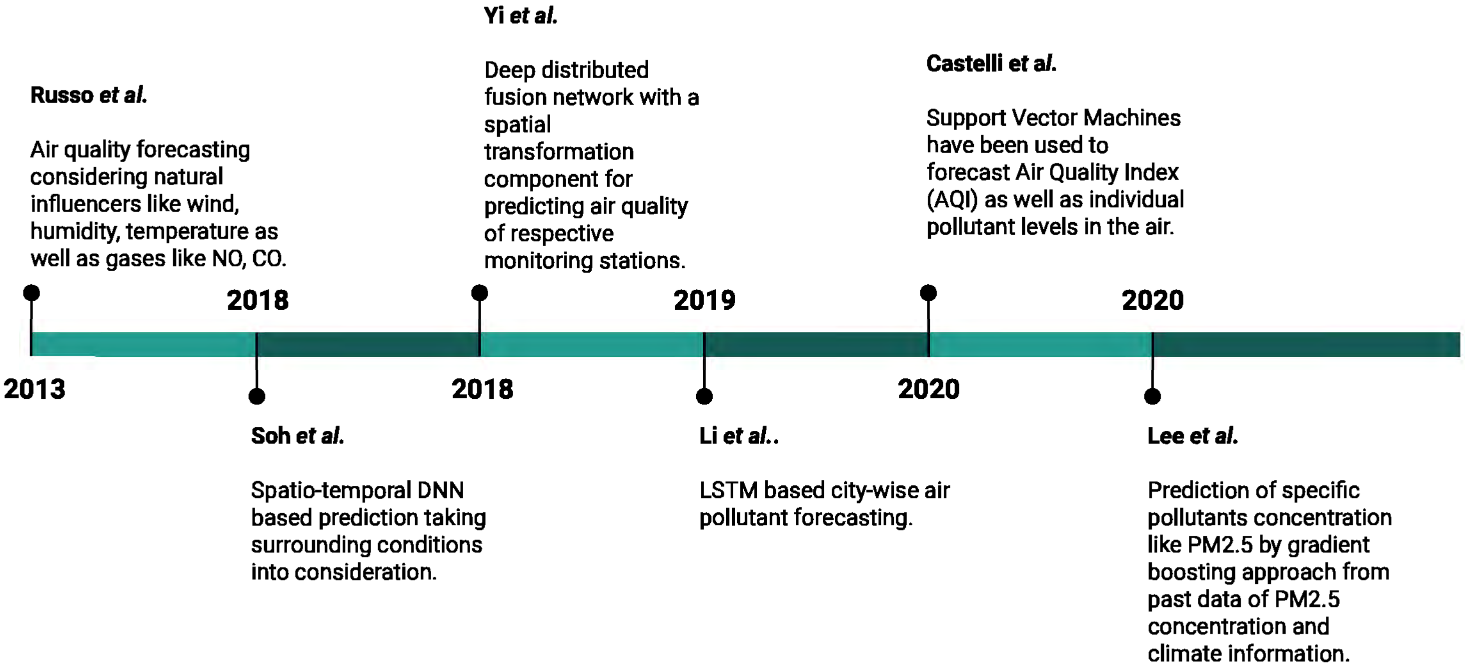
Introduction

Air pollution holds severe and immediate impacts on the climate and ecosystem of the planet leading to Global Warming.

PM2.5 and NO2 the two most common air pollutants are well known to inflict irreversible respiratory disease.

Air pollution affects rainfall and regional weather patterns.

Related Works



2013

Russo *et al.*

Air quality forecasting considering natural influencers like wind, humidity, temperature as well as gases like NO, CO.

2018

Soh *et al.*

Spatio-temporal DNN based prediction taking surrounding conditions into consideration.

2018

Yi *et al.*

Deep distributed fusion network with a spatial transformation component for predicting air quality of respective monitoring stations.

2019

Li *et al.*.

LSTM based city-wise air pollutant forecasting.

2020

Castelli *et al.*

Support Vector Machines have been used to forecast Air Quality Index (AQI) as well as individual pollutant levels in the air.

2020

Lee *et al.*

Prediction of specific pollutants concentration like PM2.5 by gradient boosting approach from past data of PM2.5 concentration and climate information.

Drawbacks

- Most of the previous works are concentrated on a single region which makes the models not universal.
- They do not consider the influences of causal agents of pollution like automobile and industry emissions.
- Lack of large scale dataset hinders a larger exploration or a forecasting study.
- Absence of multivariate time-series based approaches.

Contributions

Dataset

- Large-scale dataset encompassing pollutants, meteorology, traffic and power plant emissions.
- Spatio-temporal in nature. Spans over 2 years and over more than 50 cities in the United States.
- First dataset to include the effect of power plant emissions with others.

Methods

- Linear time-complexity Transformer with a non-linear re-weighting attention mechanism enforcing strict locality between neighboring tokens.
- First use of Transformers for multivariate pollutant forecasting.
- Hybrid loss function with soft-DTW for increased robustness for pollutant forecasting task.

Analysis

- Explainable Bayesian modeling to capture the relative importances of the different factors in influencing the pollutant levels.
- Study on the degree of dependency of the pollutants on previous days values reflecting the duration of retention of pollutants in the atmosphere.

Dataset

Air Pollutants

- PM2.5
- PM10
- NO2
- CO
- SO2
- O3

Power Plant Emissions

$$I_{pp_{c,t}} = \sum_{p} G_p / r_{cp}^2$$

Radial influence from point of generation.

Meteorological Factors

- Pressure
- Humidity
- Temperature
- Dew
- Wind Speed
- Wind Gust

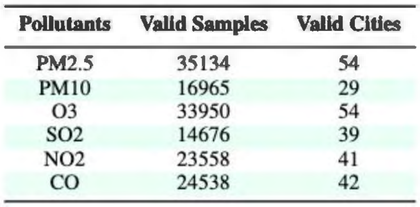
Traffic Emissions

Cumulative trip distance per day per city in million miles.

Dataset: Details

Spatio-temporal Data: 54 cities, 731 days.

Large Feature Set: 9 causal agents (6 natural, 3 artificial), 6 pollutants.



Pollutants	Valid Samples	Valid Cities
PM2.5	35134	54
PM10	16965	29
O3	33950	54
SO2	14676	39
NO2	23558	41
CO	24538	42

Table 1: Dataset Statistics. A city is considered valid here if it has at least 2 months data of the pollutant levels.

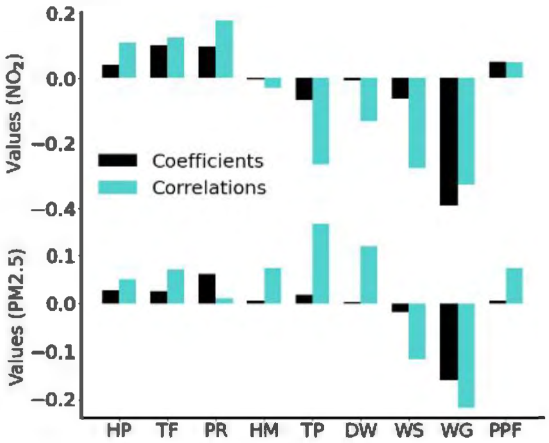


Figure 3: Weights W_is from different inputs in BR model alongside associated correlations of these inputs with NO_2 and PM2.5 levels. X-axis represents (left to right): Population at Home, Traffic, Pressure, Humidity, Temperature, Dew, Wind Gust, Wind Speed and Power Plant Feature.

Method

cosSquareFormer

A linear operation with decomposable non-linear cosine-square based re-weighting mechanism instead of a standard non-linear softmax operation - Alleviates quadratic time and space complexity!

$$s(\tilde{Q}_i.\tilde{K}_j) = \tilde{Q}_i\tilde{K}_j^T cos^2\left(\pi\frac{i-j}{2M}\right) = \frac{1}{2}\left[\tilde{Q}_i\tilde{K}_j^T + \tilde{Q}_i\tilde{K}_j^T cos\left(\pi\frac{i-j}{M}\right)\right]$$

This re-weighting mechanism weights the neighbouring tokens more (compared to cosine) with respect to the far-away ones.

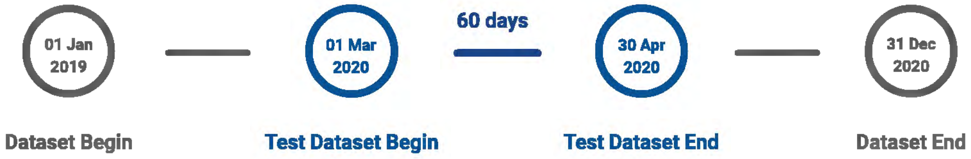
Experiments

- ## Non-Sequential Models:

  Estimation problem: Estimating a pollutant value based on the day's features.
  Ordinary Least Squares, Bayesian Regression, Gradient Boosting Machines.

- ## Sequential Models:

  Forecasting problem: Predicting pollutant levels based on features as well as the pollutant levels of the previous n days.
  LSTM, Transformer, cosFormer, ***cosSquareFormer***.



01 Jan 2019 — Dataset Begin

01 Mar 2020 — Test Dataset Begin

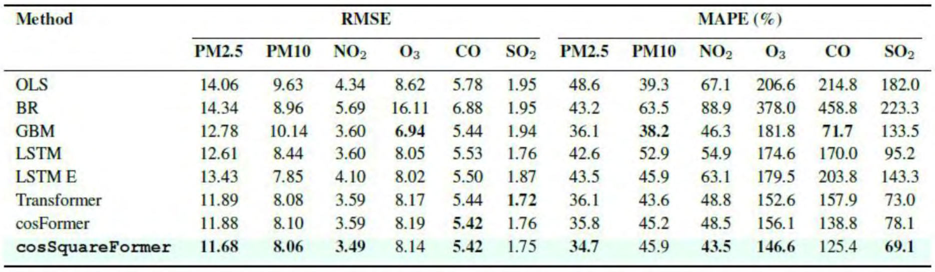
60 days

30 Apr 2020 — Test Dataset End

31 Dec 2020 — Dataset End

Results: Model Performance

cosSquareFormer gives best results in 4/6 pollutants for both RMSE and MAPE metrics.



Method	RMSE						MAPE (%)					
	PM2.5	PM10	NO_2	O_3	CO	SO_2	PM2.5	PM10	NO_2	O_3	CO	SO_2
OLS	14.06	9.63	4.34	8.62	5.78	1.95	48.6	39.3	67.1	206.6	214.8	182.0
BR	14.34	8.96	5.69	16.11	6.88	1.95	43.2	63.5	88.9	378.0	458.8	223.3
GBM	12.78	10.14	3.60	**6.94**	5.44	1.94	36.1	**38.2**	46.3	181.8	**71.7**	133.5
LSTM	12.61	8.44	3.60	8.05	5.53	1.76	42.6	52.9	54.9	174.6	170.0	95.2
LSTM E	13.43	7.85	4.10	8.02	5.50	1.87	43.5	45.9	63.1	179.5	203.8	143.3
Transformer	11.89	8.08	3.59	8.17	5.44	**1.72**	36.1	43.6	48.8	152.6	157.9	73.0
cosFormer	11.88	8.10	3.59	8.19	**5.42**	1.76	35.8	45.2	48.5	156.1	138.8	78.1
cosSquareFormer	**11.68**	**8.06**	**3.49**	8.14	**5.42**	1.75	**34.7**	45.9	**43.5**	**146.6**	125.4	**69.1**

Table 3: Performance of predictions from different models for all 6 pollutants. LSTM E and Attention LSTM E are trained on explicit information of weekday and month whereas the explicit information have been excluded whilen training the remaining models. The sequence length (number of past days) for all the LSTM and Transformer (including variants) is 7 days.

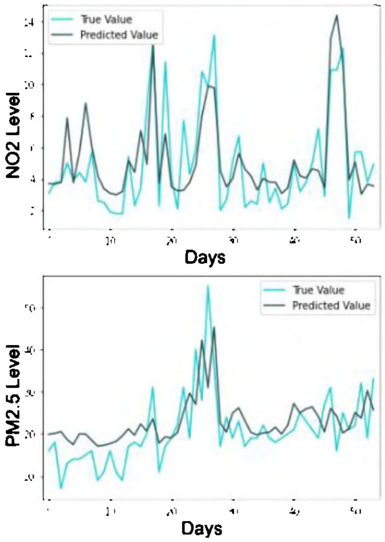
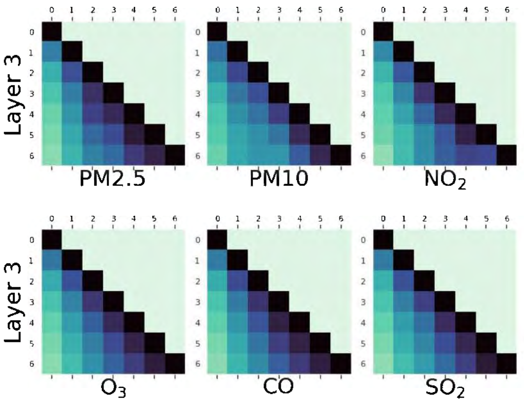


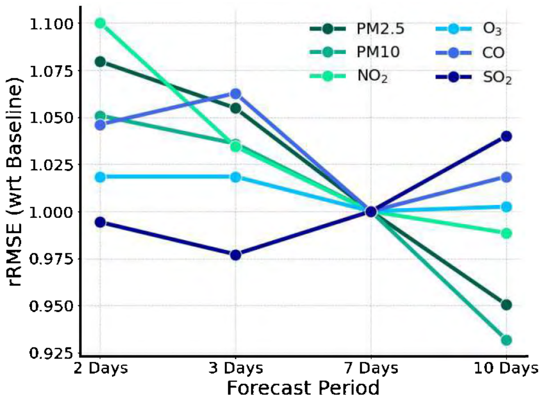
Figure 6: General fit of the proposed model on the test set for the city of Las Vegas.

Results: Sequential Analysis



Varying degree of depending on previous days' information for different pollutants.

→ Denotes level of retention in the atmosphere.



PM2.5, PM10, NO2: Estimation performance improves with longer sequence information.

O3, CO, SO2: Estimation performance stays unchanged/deteriorates with longer sequence information.

Conclusions and Future Work

- An early initiative to tackle the problem of air pollution through our dataset and methodology to establish a baseline for the community to build on.

- Variety of new unexplored factors influencing the air pollution levels captured together in our dataset.

- Improvement and extension of the dataset with more data considering other emission sources and larger spatio-temporal analysis.

Thank You

Dataset and Code are available at:

https://github.com/mayukh18/DEAP