Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Special Track on AI for Good

Deciphering Environmental Air Pollution with Large Scale City Data

Mayukh Bhattacharyya[*1] , **Sayan Nag**[*2] and **Udita Ghosh**[3]
[1]Stony Brook University
[2]University of Toronto
[3]Zendrive Inc.
mayukh.bhattacharyya@stonybrook.edu, sayan.nag@mail.utoronto.ca, uditag@zendrive.com

Abstract

Air pollution poses a serious threat to sustainable environmental conditions in the 21st century. Its importance in determining the health and living standards in urban settings is only expected to increase with time. Various factors ranging from artificial emissions to natural phenomena are known to be primary causal agents or influencers behind rising air pollution levels. However, the lack of large scale data involving the major artificial and natural factors has hindered the research on the causes and relations governing the variability of the different air pollutants. Through this work, we introduce a large scale city-wise dataset for exploring the relationships among these agents over a long period of time. We also introduce a transformer based model - cosSquareFormer, for the problem of pollutant level estimation and forecasting. Our model outperforms most of the benchmark models for this task. We also analyze and explore the dataset through our model and other methodologies to bring out important inferences which enable us to understand the dynamics of the causal agents at a deeper level. Through our paper, we seek to provide a great set of foundations for further research into this domain that will demand critical attention of ours in the near future.

1 Introduction

Advancement of civilization has led to a lot of interferences on earth generated by humans. While a lot of those pose a threat to the balance of ecosystem and overall climate of the planet, air pollution happens to hold severe and immediate impacts on us humans. PM2.5 and NO_2 the two most common air pollutants are well known to inflict irreversible respiratory disease[Shi *et al.*, 2016]. Besides asthma attacks and cardiovascular issues, it has been observed to cause or exacerbate cancers, diabetes [Bowe *et al.*, 2018] and also influence mortality in infants [Abdo *et al.*, 2019]. The major sources of pollutants like PM2.5, NO_2, O_3 etc are emissions from automobiles, power plants and other heavy industries[2] . The close proximity of industrial zones around highly populated metropolitan areas combine all the sources of the pollutants to create a very poor living conditions in quite a few cities in the world. Governments of multiple cities have tried methodologies ranging from artificial rains by cloud-seeding, partial traffic ban to giant air purifiers. All these efforts showcase the rising importance of the issue with every passing year. A year long stretch of lockdown and work-from-home systems has suddenly proved how the absence of public human activities has an improving effect on the pollution levels in big metropolitan cities. Traffic, a big factor behind urban air pollution was completely absent in the initial 1-2 months. This led to significant drop in different pollutant levels as demonstrated in different studies such as [Menut *et al.*, 2020]. These developments have garnered a lot of attention over finding viable solutions that was inconceivable on a large scale earlier due to lack of data with such variability. We feel this is an exciting opportunity to hasten the research in this domain further. In this regard we present our dataset and methodology which we believe will aid in our common mission.

The primary contributions of the paper are three-fold:

1. We have introduced a large-scale curated spatio-temporal dataset[1] encompassing daily levels of different pollutants and the major casual agents (both artificial and natural) for a duration over 2 years spanning over more than 50 cities in the United States. As per our knowledge, this is the largest dataset in terms of it's coverage of the number of cities. Also, alongside traffic emissions, we adopted a novel method to quantify the critical impact of emissions from power plants, which is a first for studies on urban air pollution till date.
2. We have proposed a non-linear re-weighting attention mechanism for transformers which weights neighboring tokens more with respect to the far-away ones enforcing a strict locality constraint. Although, similar weighting schemes have been very recently adopted for language modeling tasks but this is the first time such a method is being used for time-series analysis. Furthermore, to the best of our knowledge this is the first attempt to use transformers for multi-variate pollutant forecasting tasks. In addition, we have considered a novel hybrid

[*] denotes equal contribution

[1] Dataset and Code: https://github.com/mayukh18/DEAP

[2] NO_2:https://www.epa.gov/{no2-pollution, pm-pollution}

loss function combining Mean Squared Error and Dynamic Time Warping resulting in more robust similarity computation between two temporal sequences.

3. We have also presented a holistic analysis of the dataset that we are introducing. With bayesian modeling, we have captured the relative importances of the different factors in influencing the pollutant levels. We have also analysed the dependency of the pollutants on previous days values thus reflecting the duration of retention of pollutants in the atmosphere. Additionally, we have presented different inferences drawn from the data which brings out actionable information that can used on a wider scale.

2 Related Work

Study of factors leading to air pollution has gained momentum in recent time, although it is a persistent problem for long. Although the studies have taken different problem statements but they have focused mainly around PM2.5 as the central theme. Though previous works existed for air quality forecasting, one of the first works to consider natural influencers like wind, humidity, temperature as well as gases like NO, CO was done by [Russo *et al.*, 2013]. Traditional machine learning methods like Support Vector Machines have been used to forecast Air Quality Index (AQI) as well as individual pollutant levels in the air [Castelli *et al.*, 2020]. Prediction of specific pollutants concentration like PM2.5 by gradient boosting approach from past data of PM2.5 concentration and climate information is presented in [Lee *et al.*, 2020]. However, with the advent of RNNs [Rumelhart *et al.*, 1985; Sherstinsky, 2020], most of the recent works have been using LSTMs [Hochreiter and Schmidhuber, 1997] for air pollution estimation [Qadeer *et al.*, 2020]. [Li *et al.*, 2019] and [Al-Janabi *et al.*, 2020] utilised LSTM based systems to predict the concentration of the air pollutants. A Spatio-temporal DNN has been presented in [Soh *et al.*, 2018] which takes surrounding conditions into consideration while predicting. Leveraging the information from air quality monitoring stations and other factors like city's points of interests, road networks and meteorological data, a combination of feed-forward and recurrent networks has been used to model static and sequential data. [Cheng *et al.*, 2018]. Considering air quality data, meteorology data and weather forecast data, a deep distributed fusion network has been used with a spatial transformation component for predicting air quality of respective monitoring stations [Yi *et al.*, 2018]. Using multi-level attention networks with spatial-temporal and meteorological data, air and water quality prediction have been done in [Liang *et al.*, 2018].

However, the drawbacks of most of these works is either they are concentrated on a single region which makes the models not universal, or that they do not consider the influences of causal agents of pollution like automobile and industry emissions. Although few studies like [Wang *et al.*, 2018] evaluates the effectiveness of several thermal power plant control measures on the air quality, a larger exploration or forecasting study is not available due to lack of large scale data.

3 Dataset

We present a large dataset for modeling the variation of air pollution at the daily level over multiple cities involving data from most of the influencing agents both natural and man-made. The dataset is the largest as per our knowledge in regards to the number of locations and days involved.

Overall, the dataset contains a total of 35,596 unique sample points spanning 54 cities and 24 months with each sample point representing a unique (date, city) combination. The data is collected and curated from multiple sources. Hence, some cities and some dates do not have the values for all the pollutants and features. The different aspects of the dataset are given in Table 1. The sources of the features and the data processing involved are described in the respective sections below:

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.

- **Air Pollutants** The daily data of different pollutant species at a city level was obtained from Air Quality Open Data Platform[3]. The min, max and median values of the pollutant on a day are provided. The concentration values are normalized in US EPA standard. There are six air pollutants in our dataset namely NO_2, PM2.5, PM10, SO_2, O_3, and CO. The violin plots in Figure 1 illustrates the monthly distribution and variation of the aforementioned air pollutants.
- **Meteorological Factors** Meteorological factors like humidity, windspeed, temperature and pressure have an impact on the concentration of pollutants in the atmosphere. The concentration values of these meteorological factors are also obtained from the Air Quality Open Data Platform[3] like above. They serve as input features for our models. The units of the features are provided in the dataset. Figure 2 depicts the correlations among these meteorological factors and the respective pollutants.
- **Traffic** The corresponding daily traffic data is collected from[4] provided by Maryland Transport Institute [Zhang *et al.*, 2020]. The traffic data follows almost the same spatio-temporal granularity as the air pollutant data, apart from one aspect. The traffic data is provided at a county level, not at the city level. But since we are dealing with mainly major metropolitan areas, we have taken the liberty to consider the traffic of the city same as the county it lies within. The trip-based data from [Zhang *et al.*, 2020] is processed to collate all the trips in a day to

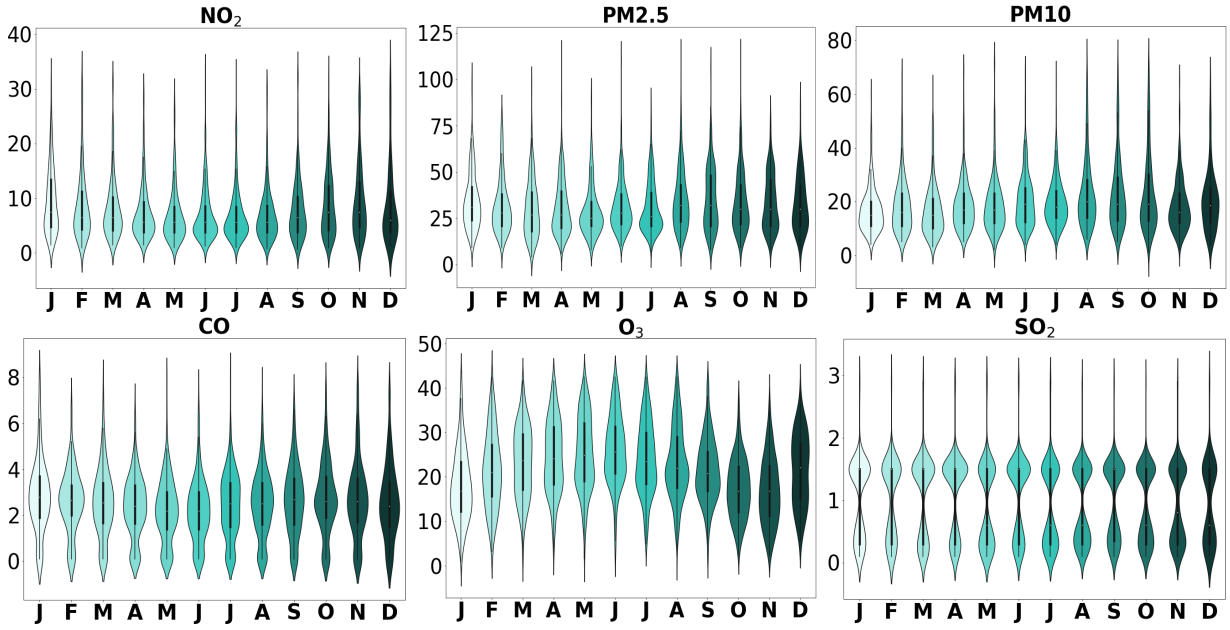

Figure 1: Distribution of NO_2, PM2.5, PM10, CO, O_3, and SO_2 over a period of 12 months (x-axis). Each of the pollutants demonstrate a different distribution both over the period of 12 months and also among the pollutants on each month.

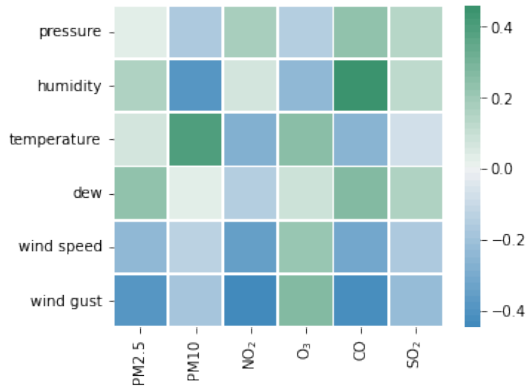

Figure 2: Correlation of Meteorological Factors with Pollutant levels

calculate the "million miles" of travel in the day, which we treat as the measure of traffic in that city for that day.

- **Power Plant Emission** The data around the generation patterns of power plants could only be obtained at a monthly level from US EIA Website[5]. Considering the production patterns of power plants don't change much at the daily level, we made a pragmatic approximation of averaging the monthly value to the daily level. There are 11,833 power plants we have considered in the dataset. It should however be noted that we have only selected generation data of generators running on fuel types - Coal, Oil, Gas and Biomass, since these are the major ones most frequently held responsible for air pollution.

While we do provide the power-plants data in it's granular raw form, we needed a single feature representing the effects of the power plants for a certain (city,date) pair. For that purpose, we design an intuitive metric to form the feature .

$$I_{pp_{c,t}} = \sum_{p} G_p / r_{cp}^2, \quad for\ r_{cp} < R_{limit} \tag{1}$$

where I_{pp} is the feature obtained from power-plants for a city c on a date t. G_p is the average daily generating capacity for the plant for that month and r_{cp} is the linear distance between the power-plant and the centre of the city. We have taken R_{limit} as 30 km.

4 Proposed Method

The general form of transformer is given as: $T(x) = F(A(x) + x)$ where T is the transformer block with an input sequence x, F is the feed-forward network and the self-attention function is given by A which has a quadratic space and time complexity depending on the length of the input sequence (N). Attention function has three learn-able linear matrices namely Query (Q), Key (K) and Value (V) [Vaswani *et al.*, 2017] which when combined together using a dot-product attention with softmax normalization gives the final attention output given as:

$$O_i = \sum_j \frac{e^{Q_i K_j^T}}{\sum_j e^{Q_i K_j^T}} V_j, \ \forall i, j \in N \tag{2}$$

[3]https://aqicn.org/data-platform/covid19

[4]https://data.bts.gov/Research-and-Statistics/Trips-by-Distance/w96p-f2qv

[5]https://www.eia.gov/electricity/data/eia923/

Features	25%	50%	75%
Pollutants (X)			
1. PM2.5	21	28	39
2. PM10	10	15	21
3. NO2	3.8	6.4	10.2
4. O3	14.5	20.8	27.2
5. SO2	0.3	1.1	1.5
6. CO	1.8	2.5	3.6
Traffic Distance(I_T)	19.54	31.32	49.44
Power Plant Emission(I_{pp})	0.40	1.69	6.95

Table 2: Feature Distributions. Since the data is spread over 2 years in 54 cities, we get a good distribution of the feature values.

The quadratic time and space complexity poses a computational challenge, especially for long sequences. Few solutions, including linearization of self-attention were proposed in this regard [Choromanski *et al.*, 2020; Katharopoulos *et al.*, 2020]. Along the same lines, a recently proposed work [Qin *et al.*, 2021], used a linear operation with decomposable non-linear cosine-based re-weighting mechanism instead of a standard non-linear softmax operation. They noticed that non-linear re-weighting introduced by softmax attention results in a stable training process and addition of a decomposable cos-based re-weighting scheme can introduce recency bias to the attention matrix. As a consequence, locality is enforced.

Inspired by the above idea, in order to enforce stricter locality constraint, we proposed a decomposable cosine-square re-weighting mechanism which weights the neighbouring tokens more (compared to cosine) with respect to the far-away ones (see Appendix for Algorithm). This cosine-square reweighting can also be loosely considered as a linear combination of cosFormer [Qin *et al.*, 2021] and Linear Transformer [Katharopoulos *et al.*, 2020]. The similarity function between Q and K with cosine-square re-weighting is defined as:

$$\begin{aligned} s(\tilde{Q}_i, \tilde{K}_j) &= \tilde{Q}_i \tilde{K}_j^T cos^2\left(\pi \frac{i-j}{2M}\right) \\ &= \frac{1}{2}\left[\tilde{Q}_i \tilde{K}_j^T + \tilde{Q}_i \tilde{K}_j^T cos\left(\pi \frac{i-j}{M}\right)\right] \end{aligned} \tag{3}$$

Using Ptolemy's theorem and decomposing the above expression further leads to:

$$\begin{aligned} s(\tilde{Q}_i, \tilde{K}_j) = \frac{1}{2}\Bigg[\tilde{Q}_i \tilde{K}_j^T + \\ \left(\tilde{Q}_i cos\left(\pi \frac{i}{M}\right)\right)\left(\tilde{K}_j cos\left(\pi \frac{j}{M}\right)\right)^T + \\ \left(\tilde{Q}_i sin\left(\pi \frac{i}{M}\right)\right)\left(\tilde{K}_j sin\left(\pi \frac{j}{M}\right)\right)^T\Bigg] \end{aligned} \tag{4}$$

where $i, j = 1, ..., N$, $M \geq N$, $\tilde{Q} = f(Q)$, $\tilde{K} = f(K)$, $f = ReLU$ and output is given as:

$$O_i = \frac{\sum_{j=1}^{N} s(\tilde{Q}_i, \tilde{K}_j) V_j}{\sum_{j=1}^{N} s(\tilde{Q}_i, \tilde{K}_j)} \tag{5}$$

4.1 Loss Function

For training purposes we have used a novel hybrid loss function composing a weighted combination of MSE Loss and a soft-Dynamic Time Warping (sDTW) loss [Cuturi and Blondel, 2017]. For forecasting purposes, it has been noticed in [Cuturi and Blondel, 2017] that sDTW loss performs superior to standard Euclidean loss because of the former's robustness to similarity computation between two temporal sequences. In our work we have seen combining these two losses together gives better performances and training stability as compared to their individual counterparts. Hence, our proposed loss function between ground-truth (y) and predicted ($\bar{y}$) time-series is given as:

$$L(y, \bar{y}) = MSE(y, \bar{y}) + \lambda \, sDTW(y, \bar{y}) \tag{6}$$

We have considered $\lambda = 0.5$ for the experiments.

5 Experiments

In this section, we approach the problem of estimating pollutant levels based on information about the causal and influencing factors.

We have evaluated our sequential proposed method with both non-sequential and sequential methodologies as baselines. For non-sequential models, the problem statement is that of estimation: trying to estimate a pollutant value based on the day's features. However, for sequential models, we incorporated the forecasting problem. This means we are trying to predict a certain day's pollutant levels based on features of that day along with the pollutant level of the previous day.

In total, we compared our method with 9 baselines. Among the estimator models, we use Ordinary Least Squares (OLS), Bayesian Regression (BR), Gradient Boosting Machines (GBM). Among the sequential forecasting models, we use LSTM, Attention LSTM, Transformer and cosFormer.

5.1 Test Data

Since we have both sequential and non-sequential models, we needed an train-evaluation split of the whole dataset which would have let us evaluate sequential models with the same ease as non-sequential traditional models. Usually for time-series data, there is the prevalent norm of selecting a later portion as the evaluation dataset. However, we realized that in doing so we would be restricting the evaluation to a particular season with not much daily variations caused by the features. Since we have both the year 2019 and 2020 in the dataset, we constituted the evaluation or test dataset by taking a continuous 60-day segment for each city starting from the first week of March 2020. Since this time period marked the onset of the COVID lockdowns, we would get a much better variability in terms of the features and pollutants. Considering some values in the test set might be missing due to reasons we discussed before, the evaluation performance is calculated only on the available and valid test data samples.

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.

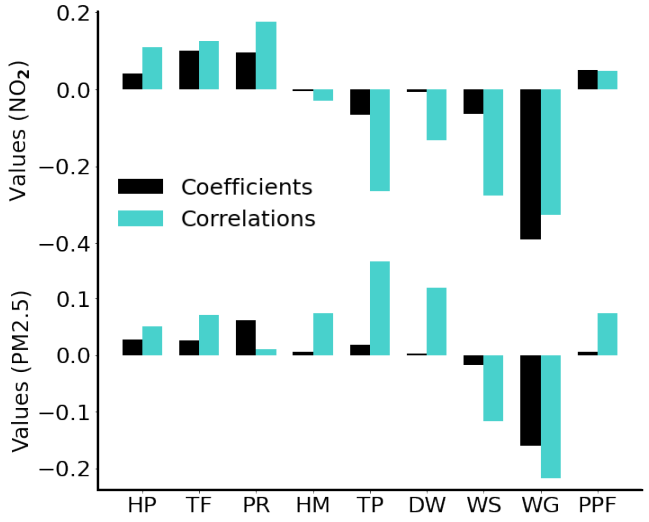

Figure 3: Weights W_is from different inputs in BR model alongside associated correlations of these inputs with NO_2 and PM2.5 levels. X-axis represents (left to right): Population at Home, Traffic, Pressure, Humidity, Temperature, Dew, Wind Gust, Wind Speed and Power Plant Feature.

5.2 Metrics

We evaluate and compare all our methods with 2 metrics: Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) [Botchkarev, 2018]. A combination of these two will give us a holistic picture of the performance of the models being evaluated.

The results from our method as well as other sequential and non-sequential models are presented in Table 3.

6 Analysis

In this section, we explore the findings from our experiments in a little more depth to infer conclusions about the various interrelationships of the features.

Table 3 provides a good idea about the general fit and importance of the models in terms of estimating and forecasting pollutant levels. As we can see our proposed **`cosSquareFormer`** as well as Gradient Boosting Machines (GBM) do well in terms of performance. As we can see in Figure 6 our proposed model does a great job in following sudden daily fluctuations in the pollutant levels.

The good performance of GBM for certain pollutants does raise a question of whether features of past days influence the pollutant levels of future. In order to explore the questions related to the sequential nature of pollutants, we designed an ablation study with multiple sequence lengths with the same experimental setup to maintain parity for modeling. The results given in Figure 5 show that pollutants like PM2.5, PM10 and NO_2 have a better performance with longer sequence lengths, whereas the others either degrade or show a flat trend. Thus it can be assumed that the daily concentration of some pollutants indeed have a good dependence on past concentrations whereas some others are mostly independent of it. This can also be analysed at depth from the attention maps of **`cosSquareFormer`** in Figure 4. The values on the last rows(row 6) denote the dependency of pollutant levels on a particular day on the features of the previous 6 days.

We also wanted to model the uncertainty in the data through Bayesian Inference. Figure 3 shows the weights (means, μ) obtained as a result of Bayesian Inference for NO_2 and PM2.5 alongside correlation values computed corresponding to the pollutants. This plot not only gives us an idea of the importance of each factor and the extent of the influences of each input feature on affecting the pollutant levels, but also demonstrates a parity that exists between the weights and the corresponding correlation values. The complex nature of Transformers in computing the weights for each feature made it difficult to extract the importance of features from it which we would have shed more light on the significance of features.

The visualizations shown in Figure 7 provide some information about each city's conformity with the universal model. It shows us the cities which have pollutant levels which were much higher than that estimated by our model. It provides us the leads to explore the context and reason behind each such outlier city. An analysis on this basis will provide researchers to identify problematic cases in a meaningful way instead of

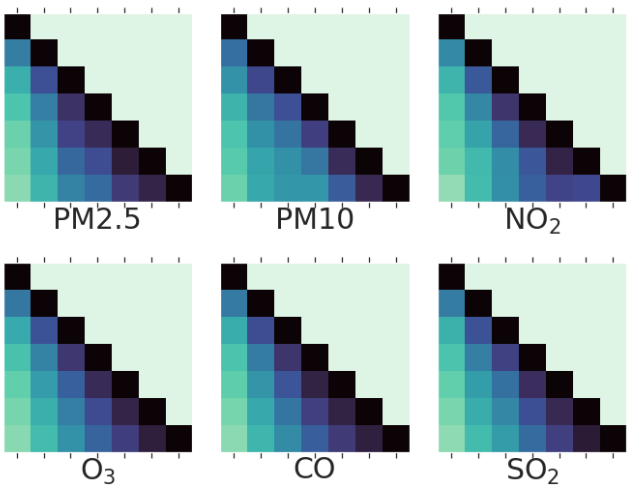

Figure 4: Attention Matrices (mean of all the heads) considering a 7-day forecast period for the third layer of the proposed model for the respective air pollutants.

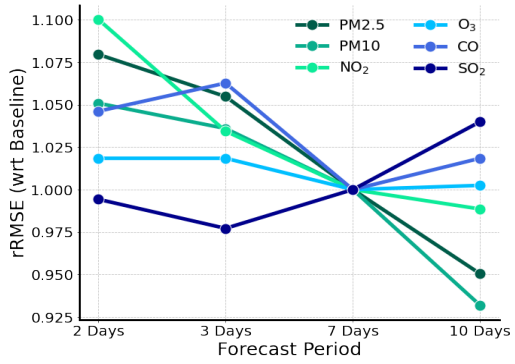

Figure 5: Relative RMSE (rRMSE = RMSE/$\text{RMSE}_{Baseline}$) scores for different sequence lengths with respect to the 7-day baseline for all the pollutants with **`cosSquareFormer`**.

just flagging cities with high pollutant levels.

7 Conclusions and Future Works

Air pollution will lead to be one of the crucial issues of the society in the years to come. An early initiative to tackle the problem may make a big difference in the future. Through our dataset and methodology we have intended to establish a foundation for the community to build on. Our dataset captures a variety of factors influencing the air pollution levels. In this study, we have illustrated the impact of such factors on the air quality indices using extensive studies exploring the various relationships governing the pollutant levels.

Our intention is to improve and extend the dataset with more data considering other emission sources. We believe there are also scopes of further studies like spatio-temporal analysis and other explorations on this dataset itself that may uncover valuable inferences which may progress our understanding of this domain further.

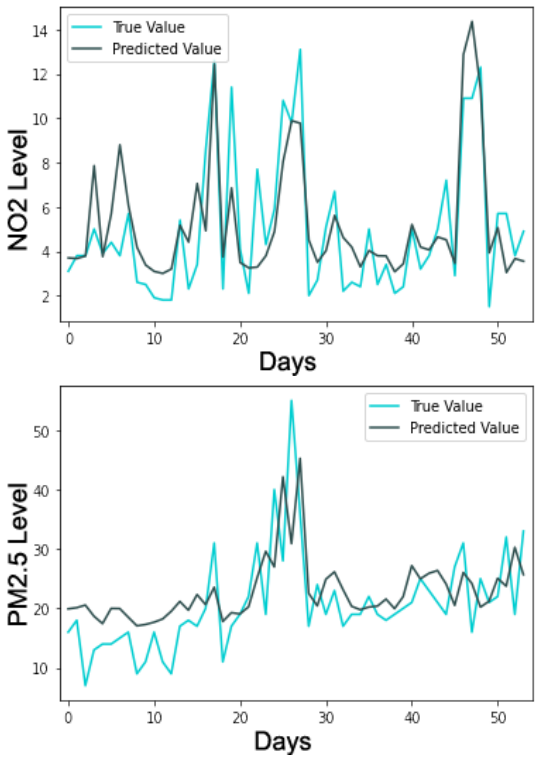

Figure 6: General fit of the proposed model on the test set for the city of Las Vegas.

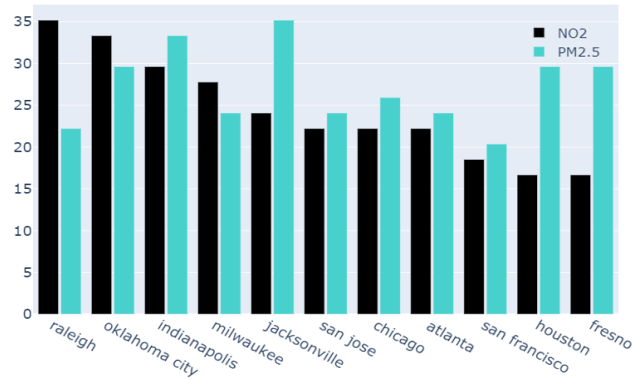

Figure 7: Cities with pollutant levels significantly higher than that predicted by the universal model. The y-axis denotes the % of samples of the city with original pollutant value higher than 125% of the predicted value.

References

[Abdo *et al.*, 2019] Mona Abdo, Isabella Ward, Katelyn O'Dell, Bonne Ford, Jeffrey R Pierce, Emily V Fischer, and James L. Crooks. Impact of wildfire smoke on adverse pregnancy outcomes in colorado, 2007–2015. *International journal of environmental research and public health*, 2019.

[Al-Janabi *et al.*, 2020] Samaher Al-Janabi, Mustafa Mohammad, and Ali Al-Sultan. A new method for predic-

tion of air pollution based on intelligent computation. *Soft Computing*, 24, 01 2020.

[Botchkarev, 2018] Alexei Botchkarev. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*, 2018.

[Bowe *et al.*, 2018] B Bowe, Y Xie, T Li, Y Yan, H Xian, and Z. Al-Aly. The 2016 global and national burden of diabetes mellitus attributable to pm2.5 air pollution. *Lancet Planet Health.*, 2018.

[Castelli *et al.*, 2020] Mauro Castelli, Fabiana Clemente, Aleš Popovič, Sara Silva, and Leonardo Vanneschi. A machine learning approach to predict air quality in california. *Complexity*, 2020:1–23, 08 2020.

[Cheng *et al.*, 2018] Weiyu Cheng, Yanyan Shen, Yanmin Zhu, and Linpeng Huang. A neural attention model for urban air quality inference: Learning the weights of monitoring stations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[Choromanski *et al.*, 2020] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

[Cuturi and Blondel, 2017] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997.

[Katharopoulos *et al.*, 2020] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[Lee *et al.*, 2020] Mike Lee, Larry Lin, Chih-Yuan Chen, Yu Tsao, Ting-Hsuan Yao, Min-Han Fei, and Shih-Hau Fang. Forecasting air quality in taiwan by using machine learning. *Scientific Reports*, 10, 03 2020.

[Li *et al.*, 2019] Hongmin Li, Jianzhou Wang, Ranran Li, and Haiyan Lu. Novel analysis–forecast system based on multi-objective optimization for air quality index. *Journal of Cleaner Production*, 208:1365–1383, 2019.

[Liang *et al.*, 2018] Yuxuan Liang, Songyu Ke, Junbo Zhang, Xiuwen Yi, and Yu Zheng. Geoman: Multi-level attention networks for geo-sensory time series prediction. In *IJCAI*, volume 2018, pages 3428–3434, 2018.

[Menut *et al.*, 2020] Laurent Menut, Bertrand Bessagnet, Guillaume Siour, Sylvain Mailler, Romain Pennel, and Arineh Cholakian. Impact of lockdown measures to combat covid-19 on air quality over western europe. *Science of The Total Environment*, 741:140426, 2020.

[Qadeer *et al.*, 2020] Khaula Qadeer, Wajih Ur Rehman, Ahmad Muqeem Sheri, Inyoung Park, Hong Kook Kim, and Moongu Jeon. A long short-term memory (lstm) network for hourly estimation of pm2. 5 concentration in two cities of south korea. *Applied Sciences*, 10(11):3984, 2020.

[Qin *et al.*, 2021] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *ICLR*, 2021.

[Rumelhart *et al.*, 1985] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, 1985.

[Russo *et al.*, 2013] Ana Russo, Frank Raischel, and Pedro G. Lind. Air quality prediction using optimal neural networks with stochastic variables. *Atmospheric Environment*, 79:822–830, 2013.

[Sherstinsky, 2020] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[Shi *et al.*, 2016] L Shi, A Zanobetti, I Kloog, BA Coull, P Koutrakis, SJ Melly, and JD. Schwartz. Low-concentration pm2.5 and mortality: Estimating acute and chronic effects in a population-based study. *Environ Health Perspect*, 2016.

[Soh *et al.*, 2018] Ping-Wei Soh, Jia-Wei Chang, and Jen-Wei Huang. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. *IEEE Access*, 6:38186–38199, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Wang *et al.*, 2018] Liqiang Wang, Pengfei Liu, Shaocai Yu, Khalid Mehmood Sipra, Zhen Liu, Chang Shucheng, Weiping Liu, Daniel Rosenfeld, Richard Flagan, and John Seinfeld. Predicted impact of thermal power generation emission control measures in the beijing-tianjin-hebei region on air pollution over beijing, china. *Scientific Reports*, 8, 01 2018.

[Yi *et al.*, 2018] Xiuwen Yi, Junbo Zhang, Zhaoyuan Wang, Tianrui Li, and Yu Zheng. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 965–973, 2018.

[Zhang *et al.*, 2020] L Zhang, S Ghader, M Pack, A Darzi, C Xiong, M Yang, Q Sun, A Kabiri, and S. Hu. An interactive covid-19 mobility impact and social distancing analysis platform. 2020.