# Prediction of Atmospheric Particulate Matter ($PM_{2.5}$) Over Beijing, China using Machine Learning Approaches

P. S. Brahmanandam
Dept. of Physics, Shri Vishnu Engineering College for Women (A), Vishnupur
Bhimavaram- 534202, India

***Abstract*— Air quality is, in general, represented by the annual mean concentration of $PM_{10}$ (particulate matter (PM) pollution particles of size ~ 10 microns, or micrometers or µm) and $PM_{2.5}$ (pollution particles of size ~ 2.5 µm). In this study, we employ popular machine learning (ML) approaches such as LSTM (Long-short term memory) and ARIMA (Auto regressive integrated moving average) on time-series of pollution data (during 01 Jan 2011 and 31 December 2014) to predict $PM_{2.5}$ over Beijing, China. The particulate matter data were, initially, trained using LSTM approach and, later, a comparison with test data showed a one-to-one correspondence between them. Further, ARIMA ML predicted data has also shown such a similar trend, which implies that both LSTM and ARIMA approaches may be efficient and reliable in predicting the temporal dynamic behavior of pollutants.**

Keywords— Machine learning approaches, atmospheric pollutants, temporal dynamics, LSTM, ARIMA

## I. INTRODUCTION

Broadly, air pollution is the presence of particles in solid, liquid and gaseous particles suspended in the air that may cause serious ill- effects to humans and other living beings, and often cause serious damage to climate [1]. Further, particulate matter (PM), aka particle pollution, a complex mixture of solid and liquid droplets, could damage humans' health if their size is as small as 2.5 micrometer or even pretty smaller up to ~1 micro meter. Based on their size, composition, and origin, PM particles are classified as coarse (fine) particles that are having sizes between ~2.5 and ~10 micrometer (between ~2.5 and ~1.0 micrometer). As far as the origins of them are concerned, the so-called coarse particles generally emit from roadways and dusty industries, while fine particles emit from smoke and haze.

It is obviously true that several serious efforts are being taken by various governmental and non-governmental agencies around the world to minimize (because complete removal of them is almost impossible) their consequent effects on humans and other living beings. Still, the widespread of them in the environment is increasing many folds primarily due to ever increasing urbanization (thereby lots of new constructions will come that eventually produce air pollution), biomass burning, agricultural residuals, heavy industries' emissions and vehicular emissions. To minimize pollution effects, we have left with an option that to forecast their trends (up to 0200 to 2400 hours or more) so that awareness among the public may be created. Several research workers have kept their ardent efforts to forecast them for a long time with [2] conventional forecast methods (particularly statistical methods) that would need lots of computational facilities and a plethora of memory to read, store and carry out the analysis. However, recent advances in computer languages (such as Python and R) and cloud computing facilities (Google Colab) have allowed us to forecast pollution trends with much ease and relatively in lesser time. In this research work, we have, therefore, attempted to forecast $PM_{2.5}$ over Beijing, China using ML approaches including LSTM and ARIMA.

### Database and methodology

Beijing, China's capital, has witnessed the highest levels of air pollution ever since the rapid economic growth has started in the late 1970s and the GDP (gross domestic product) of China has touched almost ~ 10% through 2018 [3]. The $PM_{2.5}$ data over Beijing from 01 January 2010 to 31 December 2014 were archived from the Donald Bren School of Information and Computer Sciences (ICS), University of California, Irvine, USA (https://www.ics.uci.edu/). The header part of the downloaded file will look as shown in Table 1 and row 6 shows the hourly data of $PM_{2.5}$. The analysis and plotting part of this research is completed with the help of Python in Google Collaborator (Colab) and various libraries were imported including, 'numpy', 'pandas', 'matplotlib', 'sklearn.datasets', and 'ARIMA' respectively.

**Table1.** Header part of $PM_{2.5}$ data file

| | No | year | month | day | hour | pm2.5 | DEWP | TEMP | PRES | cbwd | Iws | Is | Ir |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 25 | 2010 | 1 | 2 | 0 | 129.0 | -16 | -4.0 | 1020.0 | SE | 1.79 | 0 | 0 |
| 25 | 26 | 2010 | 1 | 2 | 1 | 148.0 | -15 | -4.0 | 1020.0 | SE | 2.68 | 0 | 0 |
| 26 | 27 | 2010 | 1 | 2 | 2 | 159.0 | -11 | -5.0 | 1021.0 | SE | 3.57 | 0 | 0 |
| 27 | 28 | 2010 | 1 | 2 | 3 | 181.0 | -7 | -5.0 | 1022.0 | SE | 5.36 | 1 | 0 |
| 28 | 29 | 2010 | 1 | 2 | 4 | 138.0 | -7 | -5.0 | 1022.0 | SE | 6.25 | 2 | 0 |

## II. OBSERVATIONAL RESULTS

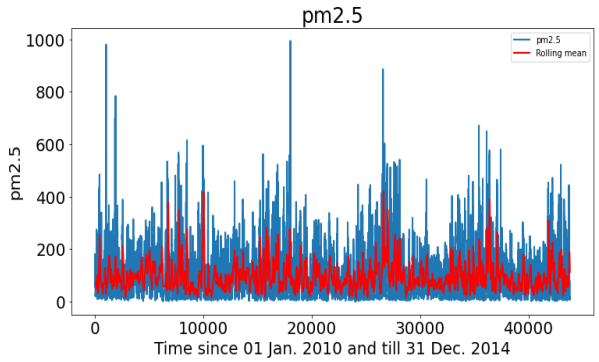


**Fig. 1** Temporal variations of PM 2.5 (blue) along with rolling mean trend (red)

The temporal variation of $PM_{2.5}$ (blue) along with 100 point running mean (red) is shown in figure 1 during 01 January 2010 and 31 December 2014. It is quite obvious that $PM_{2.5}$ shows dominant seasonal variations, particularly higher values during the winter seasons. The highest values during winter could be due to the presence of jet stream driving air masses from the west side to east, an important phenomenon that prevails during the winter season [4]. The important aspect regarding $PM_{2.5}$ is that when compared to larger particles, $PM_{2.5}$ can be breathed deep into the lungs and stay there for a longer duration [5]. Exposure to $PM_{2.5}$ and its sensitive health effects are linear if their concentrations are below 100 µg/m$^3$ [6]. A rough estimation from Figure 1 reveals that $PM_{2.5}$ concentration lies between ~ 30 and ~180 µg/m$^3$ most of the time, except during the winter seasons (large and sudden spikes in the Figure). On the other hand, few major cities of Japan (Yokohama) in 2007 and various Central and Eastern European countries too had shown relatively lower concentrations of $PM_{2.5}$ way back in the year 2001, which are only 20.6 and from 29 to 68 µg/m$^3$ [7, 8].

The entire dataset has been split into test and train data at 80:20 ratio to apply LSTM ML approach, which comes around 40000: 3870 hours of data since we have considered data between 01 January 2010 and 31 December 2014. The recurrent neural network (RNN) architecture named LSTM is applied, since this approach is capable of capturing the temporal dynamic behavior of $PM_{2.5}$ by feeding a time series of measurements. The following Figure 2 shows both training and test data sets. It can be seen that the test dataset shows distinct seasonal variation similar to train data set, which implies that the RNN/LSTM ML approach is an effective forecasting approach.

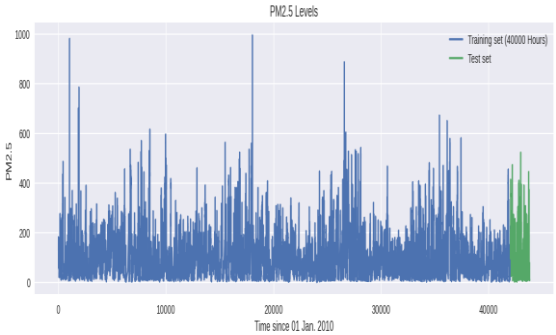


Fig. 2. Train and test PM2.5 datasets

Similarly, we have also implemented ARIMA ML approach on these data and the following figure 3 shows test and prediction data that shows almost a one-to-one correspondence. It may be worth mentioning here that the autoregressive order (p), the degree of differencing (d), and the moving average order (q) are found to be (2, 0, 1), which are essential and important parameters before implementing the ARIMA model on PM 2.5, of course, on any time series trend. It is also clear that both test and predictions show a great similitude.

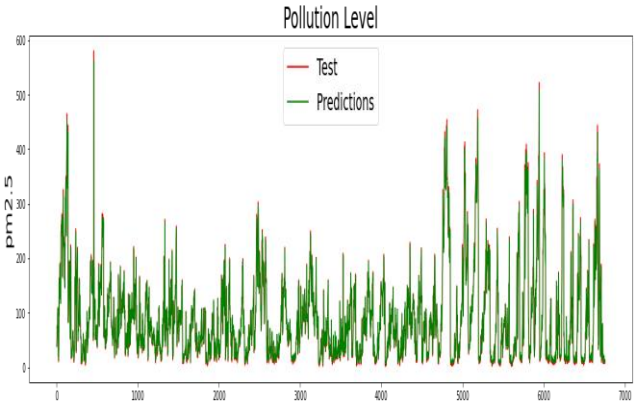


Figure 3. ARIMA predictions (green continuous line) and test data (red continuous line)

## III. CONCLUSIONS

The effective utilization of ML approaches including LSTM and ARIMA on PM2.5 time series has allowed us to verify their capability in the prediction of the dynamical trend of atmospheric pollutants. Fortunately, both approaches were able to predict pretty well, which implies that LSTM and ARIMA ML are effective, yet easy to implement, to predict temporal dynamical trends of pollutants that pose great ill-effects on humans, all living-beings and may have impacts on the Earth's atmosphere. The only way we have left with the immediate reduction of pollutants from the atmosphere, particularly PM $_{2.5}$-sized tiny particles. Otherwise, days are not far away, wherein the entire human fraternity has to wear masks permanently (24/7).

## REFERENCES

[1] Manisalidis I, Stavropoulou E, Stavropoulos A and Bezirtzoglou E, Environmental and Health Impacts of Air Pollution: A Review. Front. Public Health 8:14. doi: 10.3389/fpubh.2020.00014, 2020
[2] Tong, Y. Wan, B. Methods of forecasting air pollution and their development at home and abroad. In Proceedings of the sixth Natl Academic Conference on Envirnomental Monitoring B T, Chengdu, Sichuan, China, 10-12 October 2001
[3] CRS report, https://fas.org/sgp/crs/row/RL33534.pdf
[4] Luo, X. S., et al. "Spatial-temporal variations, sources, and transport of airborne inhalable metals (PM 10) in urban and rural areas of northern China." Atmospheric Chemistry and Physics Discussions, 14.9, 13133-13165, 2014
[5] P. Pandey, D. K. Patel, A. H. Khan, S. C. Barman, R. C. Murthy & G. C. Kisku, Temporal distribution of fine particulates ($PM_{2.5}$, $PM_{10}$), potentially toxic metals, PAHs and Metal-bound carcinogenic risk in the population of Lucknow City, India, Journal of Environmental Science and Health, Part A, 48:7, 730-745, 2013
[6] Schwela, D. Air pollution and health in urban areas. Rev. Environ. Health. 15, 13–42, 2000

[7] Khan, F.; Shirasuna, Y.; Hirano, K.; Masunaga, S. Characterization of PM2.5, PM2.5-10 and PM>10 in ambient air, Yokohama, Japan. Atmos. Res., 96 (1), 159–172, 2010
[8] Houthuijs, D.; Breugelmans, O.; Hoek, G.; Uzunova, E.;Marinescu, C.; Volf, J.; de Leeuw, F.; van de Wiel, H.; Fletcher, T.; Lebret, E.; Brunekreef, B. PM10 and PM2.5 concentrations in Central and Eastern Europe: results from the Cesar study. Atmos. Environ., 35, 2757–2771, 2001