Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.777666

PREDICTION OF AIR QUALITY LEVEL USING SARIMA MODEL IN PUNE CITY OF INDIA

Sneha Khedekar#[1], Dr. Sunil Thakare*[2]
Research Scholar, Civil Engineering, DYPIT, Pimpri, Pune, Maharashtra, India
Assistant Professor, AISSMS College of Engineering, Pune, Maharashtra, India
* Research Guide, Civil Engineering, DYPIT, Pimpri, Pune, Maharashtra, India

ABSTRACT
One of the most important problems facing the entire planet is air pollution. There are various pollutants in the air that degrade the air and create a hazardous environment. This study analyses these contaminants and uses the Seasonal Auto Regressive Integrated Moving Average (SARIMA) model to predict them. One time series analysis model that predicts specific values based on past data is the SARIMA model. The data set utilised in this model includes various pollution values that were seen at a particular time and place. When the SARIMA model was used on the data set, the pollutants could be predicted. It is a reliable method for determining whether pollution values are higher than the World Health Organisation (WHO)-mandated limits. As a result, it raises public and governmental awareness so that appropriate steps can be made to reduce the levels of such dangerous pollutants. On the basis of the provided data set, this technique's efficacy is examined, and its performance is evaluated.
Keywords-air pollution, SARIMA, prediction, dataset, awareness.

I. INTRODUCTION
Both developed and developing nations, particularly those in Europe, America, and Asia, contribute significantly to the global status of environmental air quality. This air pollution can be found both indoors, where it occurs when pollutants are trapped in structures for an extended period of time, or outside, when it is caused by pollutants in the atmosphere. Because of this, contaminants can easily migrate from their point of origin to other locations. Air pollution is one of the greatest problems in the globe and has become the primary source of pollution in many regions.
There are two main causes of this air pollution: human-based activities like open burning, industrial processes, and fuel-burning vehicles, as well as possible natural disasters like volcanoes. Numerous air pollution incidents that have been reported around the world are the main factor in harmful effects on human health in the short term. The risk of the planet being affected by global warming and greenhouse gases tends to rise as a result of air pollution.
Particulate matter (PM10), ozone (O3), sulphur dioxides (SO2), nitrogen dioxides (NO2), and carbon monoxide (CO) are the five main pollutants that need our attention because they negatively impact the global ecosystem. Assuming that every item of data is suitably created as a time series, the method of choice in this study to analyse the air quality is time series modelling and forecasting. This strategy is chosen to be utilised in managing air quality in

order to aid in future planning and helps to design the better air quality since time series analysis is the key responsibility for researchers used in development. In India, there is currently a lack of time series modelling and forecasting research and development for the purpose of monitoring air pollution.

STUDIES AND METHODOLOGIES

Xiaojun Song, et.al. [37] suggest that by introducing multiple statistical techniques for assessing fictitious and indirect causal effects, it improves comprehension of the causal structure in a multivariate time series. Based on big data analysis, they suggested a number of statistical techniques to check for the existence of indirect or spurious causality. Claudio Guarnaccia, et.al. [5], analysed a set of data of carbon dioxide levels in San Nicolas de Garza, one of the twelve municipalities that make up the Metropolitan Area of Monterrey, using the ARIMA technique. The findings demonstrated that a model that provides a solid forecast on a short time horizon may be built using hourly data. The fundamental drawback of ARIMA models is that they use data from earlier eras to create future projections, making it unable to extrapolate the prediction to other future time periods. For policymakers to apply extraordinary countermeasures to keep pollutants below regulatory standards, this may be highly helpful. Snezhana Georgieva, et.al. [33] presented a statistical analysis of six contaminants affecting the overall air quality in the small Bulgarian town of Blagoevgrad. To help with daily air quality control and forecasts, the factor analysis and SARIMA technique are excellent tools for assessing the pollution levels in small towns and cities.. A. Jaiswal, et.al. [1] based on historical data from Varanasi, India's air quality index station, conducted a statistical analysis of trends of several air contaminants using the Mann-Kendall and Sen's slope estimator approach. Future air pollution levels are also forecast using the autoregressive integrated moving average model. K. Krishna Rani Samal, et.al. [16] employed a time series forecasting method with a wide confidence range to predict future levels of several contaminants. Both the prophet model and SARIMA offer a high level of accuracy. They indicated that a deep learning algorithm may be added to the proposed method to increase its degree of freedom, variety, adaptability, and accuracy. Uzair Aslam Bhatt, et.al. [34] focuses on Pakistan's Lahore city's ambient air quality. The detected pollutants ([CO], [NO], [SO2], and [O3]) were subjected to correlation and regression analyses, and the findings were compared to the sources of those pollutants' generation. By using trajectory approaches, this study provides a thorough understanding of the relationships between contaminants and the sources of their generations. They additionally predicted particulate matter concentrations using a time series model (SARIMA).

COLLECTED DATA

1. Sources of Data Collection

Data is gathered from the following sources, among others:

a) Website of MH Pollution Control Board.
b) Website of SAFAR-India.
c) Website of Indian Institute of Tropical Meteorology.

2. Data Collected

Following data is collected:

1. Meterological data and daily air pollution data for 2 years.
2. Pollutants - PM2.5, PM10, O3, NO2.
3. Meteorological parameters – temp., wind, rain, uv-index.

II. SEASONAL AUTOREGRESSIVE INTEGRATED MOVING AVERAGE

This is a simple but quite powerful model to use for analyzing time series with seasonality. The letter S stands for seasonality. Seasonal patterns are taken into account. This element must be taken into account in order to determine how pollutants are affected by the seasons.

By employing lag values for our target variable, or the AR, which stands for autoregressive, we can generate predictions. For instance, we could predict tomorrow's sales using the sales from today, yesterday, and the day before yesterday. As it makes its forecast using three lagged data, that would be an AR (3) model.

I refer to integration. It indicates that we are comparing the target values rather than taking the raw numbers. For instance, rather than merely predicting tomorrow's sales, our sales prediction model would also attempt to anticipate tomorrow's shift in sales (i.e. tomorrow's sales minus today's sales). Since many time series show a trend, which makes the raw numbers non-stationary, we require this. By subtracting, our Y variable becomes more stationary.

The term "moving average" refers to this (MA). The lag prediction errors are fed into a moving average model as inputs. Unlike the others, it cannot be observed directly, and it is not a stable parameter because it evolves along with the other parameters of the model. Overall, feeding the model's errors back to it causes it to be slightly pushed in the direction of the right value (the actual Y values).

III. PREDICTION BY SARIMA MODEL

a. SARIMA MODEL DATA CONSIDERATION

For SARIMA model, total duration of 2 years (730 days) were considered. The period considered is 1st January 2020 to 31st December, 2021. The training data considered for the model is of first 23½ months. The period for which the prediction is to be done is the last 15 days (15th December 2021 to 31st December, 2021).

b. SARIMA MODEL FOR PM2.5

Non-seasonal (trend) considered is: AR(3), I(2), MA(3). Seasonal (trend) considered is: AR (1) I(1) MA(4) Residual value is considered as 8.5. Model used for prediction is [SARIMAX (ts_train,order = (3, 2, 3) , seasonal_order = (1,1,4,275)]. The below fig 4.1 shows the comparison of actual and predicted values. We can see that the pattern of prediction is very much same as the actual values.

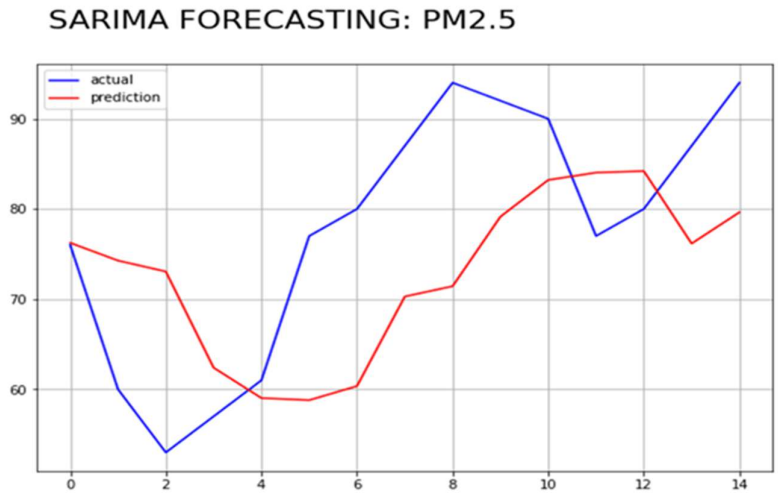

Fig. 4.1 SARIMA MODEL – PM2.5

c. SARIMA MODEL FOR PM10

Non-seasonal (trend) considered is: AR (1), I (2), MA (4). Seasonal (trend) considered is: AR (3) I (1) MA (4) Residual value is considered as 16. Model used for prediction is [SARIMAX (ts_train,order = (1, 2, 4) , seasonal_order = (3,1,4,275)]. The below fig 4.2 shows the comparison of actual and predicted values. We can see that the pattern of prediction is very much same as the actual values.

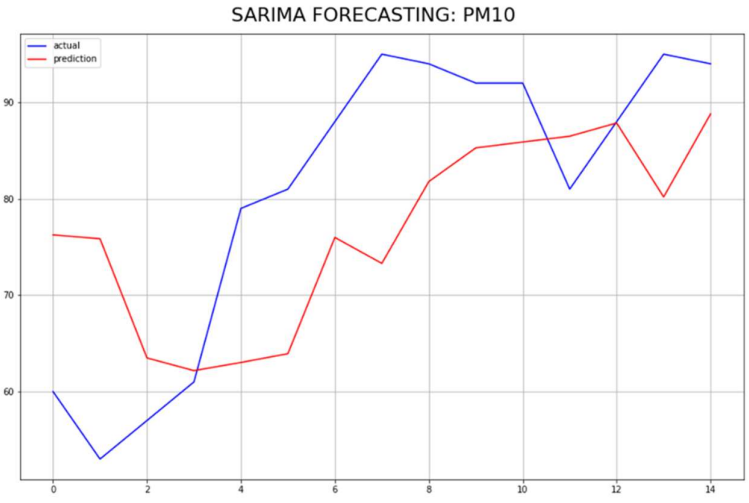

Fig. 4.2 SARIMA MODEL – PM10

d. SARIMA MODEL FOR O3

Non-seasonal (trend) considered is: AR (1), I (2), MA (3). Seasonal (trend) considered is: AR (1) I (1) MA (2) Residual value is considered as 8. Model used for prediction is [SARIMAX (ts_train,order = (1, 2, 3) , seasonal_order = (1,1,2,275)]. The below fig 4.2 shows the comparison of actual and predicted values. We can see that the pattern of prediction is very much same as the actual values.

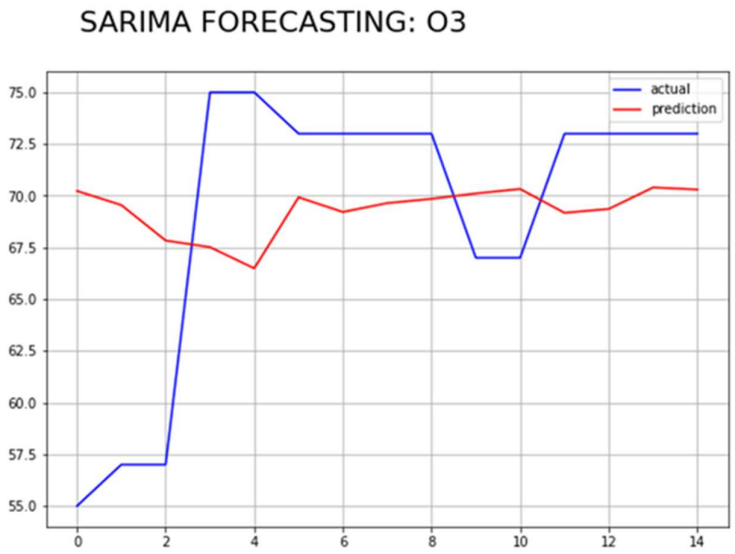

Fig. 4.3 SARIMA MODEL – O3

e. SARIMA MODEL FOR NO2

Non-seasonal (trend) considered is: AR (2), I (2), MA (2). Seasonal (trend) considered is: AR (2) I (1) MA (1) Residual value is considered as 7. Model used for prediction is [SARIMAX (ts_train,order = (2, 2, 2) , seasonal_order = (2,1,1,275)]. The below fig 4.2 shows the comparison of actual and predicted values. We can see that the pattern of prediction is very much same as the actual values.

VI.CONCLUSION

The SARIMA model is appropriate for making short-term forecasts since it allows for accurate forecasting using stationary data. A crucial instrument that aids in controlling, analysing, and monitoring the state of the air quality is the time series model utilised in forecasting. Prior to the eventual worsening of the issue, it is advantageous to act quickly.

For that reason, we need our model performance to be as accurate as possible so that good air quality forecasting can be achieved.

REFERENCES

[1] A. Jaiswal, C. Samuel, V.M. Kadabgaon, “Statistical trend analysis and forecast modeling of air pollutants”, Global J. Environ. Sci. Manage, 2018.

[2] B Ravi Kiran, Dilip Mathew Thomas, Ranjith Parakkal, “An overview of deep learning-based methods for unsupervised and semi-supervised anomaly detection in videos”, Journal of Imaging– volume 59 - Feb 2018, pp 204-207.

[3] Belagiannis, Vasileios et al. “Adversarial Network Compression.” ArXiv abs/1803.10750 (2018)

[4] Costa, Rômulo Fernandes & Yelisetty, Sarasuaty & Marques, Johnny & Tasinaffo, Paulo. (2019). A Brief Didactic Theoretical Review on Convolutional Neural Networks, Deep Belief Networks and Stacked Auto-Encoders. International Journal of Engineering and Technical Research. 9. 5-12. 10.31873/IJETR.9.12.35.

[5] Claudio Guarnaccia, Julia Griselda Ceron Breton, Rosa Maria, Carmine Tepedino, "ARIMA models application to air pollution data in Monterrey, Mexico", Mathematical Methods and Computational Techniques in Science and Engineering II- Feb 2018

[6] David Nu´ñez-Alonso, Luis Vicente Pe´rez-Arribas, Sadia Manzoor , Jorge O. Ca´ceres, "Statistical Tools for Air Pollution Assessment: Multivariate and Spatial Analysis Studies in the Madrid Region.", Journal of Analytical Methods in Chemistry, Feb 2019.

[7] Deng, Xueqing & Zhu, Yi & Newsam, Shawn. (2018). What Is It Like Down There? Generating Dense Ground-Level Views and Image Features from Overhead Imagery Using Conditional Generative Adversarial Networks.

[8] Devineau, Guillaume et al. "Convolutional Neural Networks for Multivariate Time Series Classification using both Inter- and Intra- Channel Parallel Convolutions." (2018)

[9] Elike Hodo, Xavier Bellekens, Andrew Hamilton, Christos Tachtatzis and Robert Atkinson, Shallow and Deep Networks Intrusion Detection System: A Taxonomy and Survey.

[10] Francis Chizoruo Ibea, Alexander Iheanyichukwu Opara, Chidi Edbert Durua ,Isiuku Beniah Obinnaa, Margaret Chinyelu Enedoh, "Statistical analysis of atmospheric pollutant concentrations in parts of Imo State, Southeastern Nigeria", Scientific African (Science Direct)- Nov 2019.

[11] Gurumurthy, Swaminathan et al. "DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 4941-4949.

[12] Huayi Zhan, A Survey on Deep Network.

[13] Ibrahim, Nehad. (2019). Text Mining using Deep Learning Article Review. International Journal of Scientific and Engineering Research. 9. 1916.

[14] Iqbal, Talha, and Hazrat Ali. "Generative Adversarial Network for Medical Images (MI-GAN)." Journal of medical systems vol. 42,11 231. 12 Oct. 2018, doi:10.1007/s10916-018-1072-9

[15] Janette Garcia, Akash Levy, Albert Tung, Ruomeng (Michelle) Yang, and Verena Kaynig-Fittkau, Applying Deep Learning to Petroleum Well Data, 2015.

[16] K. Krishna Rani Samal, Korra Sathya Babu, Santosh Kumar Das, Abhirup Acharaya, "Time Series based Air Pollution Forecasting using SARIMA and Prophet Model", ITCC 2019: Proceedings of the 2019 International Conference on Information Technology and Computer Communications.

[17] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng and F. -Y. Wang, "Generative adversarial networks: introduction and outlook," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588-598, 2017, doi: 10.1109/JAS.2017.7510583.

[18] Kowsari, Kamran & Heidarysafa, Mojtaba & Brown, Donald & Jafari Meimandi, Kiana & Barnes, Laura. (2018). RMDL: Random Multimodel Deep Learning for Classification. 10.13140/RG.2.2.22172.39046.

[19] Kuo Liao, Xiaohui Huang, Haofei Dang, Yin Ren, Shudi Zuo and Chensong Duan, "Statistical Approaches for Forecasting Primary Air Pollutants: A Review", Atmosphere, 2021.

[20] Lei Jiang & Ling Bai, "Spatio-temporal characteristics of urban air pollutions and their causal relationships: Evidence from Beijing and its neighboring cities", Scientific Reports, Jan 2018.

[21] Matiur Rahman Minar, Jibon Naher, Recent Advances in Deep Learning: An Overview, 2018.
[22] Mehrdad Yazdani, RemixNet: Generative Adversarial Networks for Mixing Multiple Inputs.
[23] Mohamed Amine Ferrag, Leandros Maglaras, Sotiris Moschoyiannis, Helge Janicke,Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study, Journal of Information Security and Applications,Volume 50, 2020, 102419, ISSN 2214-2126.
[24] Mojtaba Heidarysafa, Kamran Kowsari, Donald E. Brown, Kiana Jafari Meimandi, and Laura E. Barnes, An Improvement of Data Classification Using Random Multimodel Deep Learning (RMDL), International Journal of Machine Learning and Computing (IJMLC), Aug 2018.
[25] Nie, Dong et al. "Medical Image Synthesis with Context-Aware Generative Adversarial Networks." Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention vol. 10435 (2017): 417-425. doi:10.1007/978-3-319-66179-7_48
[26] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Data Synthesis based on Generative Adversarial Networks, Proceedings of the VLDB Endowment, Vol. 11, No. 10, Aug 2018.
[27] Palazzo, Simone & Spampinato, Concetto & Kavasidis, Isaak & Giordano, D. & Shah, M... (2017). Generative Adversarial Networks Conditioned by Brain Signals. 3430-3438. 10.1109/ICCV.2017.369.
[28] Pandya Mrudang Daxeshkumar, Dr. Jardosh Sunil, Applications & Challenges of Deep Learning in the field of bioinformatics, International Journal of Computer Science and Information Security (IJCSIS), Vol. 15, No. 7, July 2017.
[29] Patel, Jaynil & Pandya, Mr & Shah, Vatsal. (2018). Review on Generative Adversarial Networks. 4. 1230.
[30] Razzak M.I., Naz S., Zaib A. (2018) Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In: Dey N., Ashour A., Borra S. (eds) Classification in BioApps. Lecture Notes in Computational Vision and Biomechanics, vol 26. Springer, Cham.
[31] S. Shamshirband, T. Rabczuk and K. Chau, "A Survey of Deep Learning Techniques: Application in Wind and Solar Energy Resources," in IEEE Access, vol. 7, pp. 164650-164666, 2019, doi: 10.1109/ACCESS.2019.2951750.
[32] Schmidhuber, Jürgen. "Deep learning in neural networks: an overview." Neural networks: the official journal of the International Neural Network Society vol. 61 (2015): 85-117. doi: 10.1016/j.neunet.2014.09.003.
[33] Snezhana Georgieva Gocheva-Ilieva, Atanas Valev Ivanov, Desislava Stoyanova Voynikova , Doychin Todorov Boyadzhiev, "Time series analysis and forecasting for air pollution in small urban area: an SARIMA and factor analysis approach", Stochastic Environmental Research and Risk Assessment (Springer), 2016.
[34] Uzair Aslam Bhatt, Yuhuyan Yan, Mingquan Zhou, Sajid Ali, Aamir Hussain, Huo Qingsong, Zhaoyuan Yu, Linwang Yuan, "Time Series Analysis and Forecasting of Air

Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach", IEEE Access, Vol.9, Feb, 2021.

[35] Vlachostergiou A, Caridakis G, Mylonas P, Stafylopatis A. Learning Representations of Natural Language Texts with Generative Adversarial Networks at Document, Sentence, and Aspect Level. Algorithms. 2018; 11(10):164. https://doi.org/10.3390/a11100164.

[36] X. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," in IEEE Access, vol. 2, pp. 514-525, 2014, doi: 10.1109/ACCESS.2014.2325029.

[37] Xiaojun Song, Abderrahim Taamouti, "A better understanding of Granger causality analysis: A big data environment", Article in Oxford Bulletin of Economics & Statistics · August 2019.

[38] Zhang, Chaoyun ; Patras, Paul ; Haddadi, Hamed. / Deep Learning in Mobile and Wireless Networking: A Survey. In: IEEE Communications Surveys & Tutorials. 2019; Vol. 21, No. 3. pp. 2224-2287.

[39] Zhao, Zhong-Qiu et al. "Object Detection with Deep Learning: A Review." IEEE transactions on neural networks and learning systems vol. 30,11 (2019): 3212-3232. doi:10.1109/TNNLS.2018.2876865.