JOURNAL OF
EMERGING INVESTIGATORS

Article

Forecasting air quality index: A statistical machine learning and deep learning approach

Manvik Pasula[1], Atul Dubey[2]
[1]Prospect High School, Saratoga, California
[2]AIClub, Santa Clara, California

SUMMARY

Air pollution is a serious issue that affects many people around the globe. It has many negative effects, especially on health, and is measured by the air quality index (AQI). India is one of the most polluted countries in the world, with over 660 million people living in areas with air pollution above the standard AQI. We hypothesize that traditional time series processing algorithms for forecasting, such as the Seasonal Autoregressive Integrated Moving Average (SARIMA), capture seasonal variations and can forecast future AQI levels better than the more common complex deep-learning models, like long short-term memory (LSTM) models. We used a dataset from the Central Pollution Control Board, the official portal of the Government of India that contains time series data for different cities. We created a forecasting model using the SARIMA and LSTM models. Our findings reveal that the SARIMA model effectively captures seasonal patterns in the data for all cities, except Chennai, predicting values with a minimal error margin. In contrast, the LSTM model, while comparable in some cases, generally exhibits poorer performance across more cities and underperforms compared to SARIMA even in its better scenarios. This trend is further evidenced by the root-mean-squared error (RMSE) results, where SARIMA consistently outperforms LSTM in all cities. Overall, our methodology demonstrates high accuracy, holding significant potential to positively impact numerous lives, and supports our hypothesis.

INTRODUCTION

Air pollution is a serious issue, with almost 99.82% of land area having PM2.5 (particulate matter with a diameter <2.5 µm) levels higher than the World Health Organization (WHO) recommends, and almost 99.999% of people breathing in levels higher than what the WHO recommends (1). In addition to being very harmful to health. Air pollution causes almost 9 million deaths worldwide and over 100 million disability-adjusted life years (3,4). These numbers are only increasing as time goes on: for example, from 2008–2012 the amount of deaths attributable to air pollution almost tripled in size, making it the fourth-highest risk factor for human health and is associated with three of the leading causes of death (4,5). Air pollution also has a link to many diseases across different body systems, including cardiovascular diseases (e.g., stroke), respiratory diseases (e.g., asthma), and diseases of the nervous, digestive, and urinary systems (5,6).

Air pollution is also closely related to negative environmental effects, such as climate change, acid rain, and the greenhouse effect, alongside the contamination of both indoor and outdoor environments with particles that change the natural format of the air (7,8). This includes different gases (ozone, carbon monoxide, sulfur dioxide, etc.) and particulate matter (PM) (7) which influence environmental effects. There are two main types of PM considered, which are defined by the size of the particles in micrometers (µm): PM2.5 and PM10 (1). Air pollution is generally caused by processes such as the combustion of gas and oil, fires, and diesel/gasoline emissions (9).

India is one of the worst-polluted countries in the world in terms of air pollution. In 2010, 45% of Indian cities exceeded standard Air Quality Index (AQI) values, and Indian cities took 9 out of the 10 spots from the WHO list of cities with the worst air pollution (10). In 2015, 660 million Indians lived in areas that exceeded National Ambient Air Quality Standard levels for air pollution, around half the population at the time (11). If India were to achieve the air quality standards that the WHO sets, then almost five years of life would be saved per person (11). India also has a significant economic impact from air pollution, where it loses almost half a trillion dollars a year from the issue (11). To combat the severity of this issue, there are several different methods, including analyzing geographic trends to create policies (11). Our focus in this paper is making sure that the public is aware of air pollution levels both in the present and in the future (via AQI forecasting) so they can better plan for increased pollution, and reduce the impact it has on their lives (6). Such information can also help governments create policies to reduce the negative impact of air pollution (6).

We hypothesize that traditional time series processing algorithms for forecasting, such as Seasonal Autoregressive Integrated Moving Average (SARIMA), capture seasonal variations and can forecast future AQI levels using current and past AQI levels better than complex deep-learning models, like Long Short-Term Memory (LSTM) models, which are increasingly common in forecasting (12,13). One of the most common methods that others have tried for AQI forecasting is the usage of autoregressive models like Auto-Regressive Integrated Moving Average (ARIMA) or Auto Regressive Moving Average (ARMA), which are models that apply statistical analysis for forecasting (14–16). However, this method fails to consider seasonality, which is a key factor of the Indian AQI time series. A simple solution is to use similar models that do account for seasonality, like SARIMA. In addition, previous work only accounts for very few cities. Most only account for a single city, which is limited in its practicality, and others only account for three at a maximum

https://doi.org/10.59720/24-079

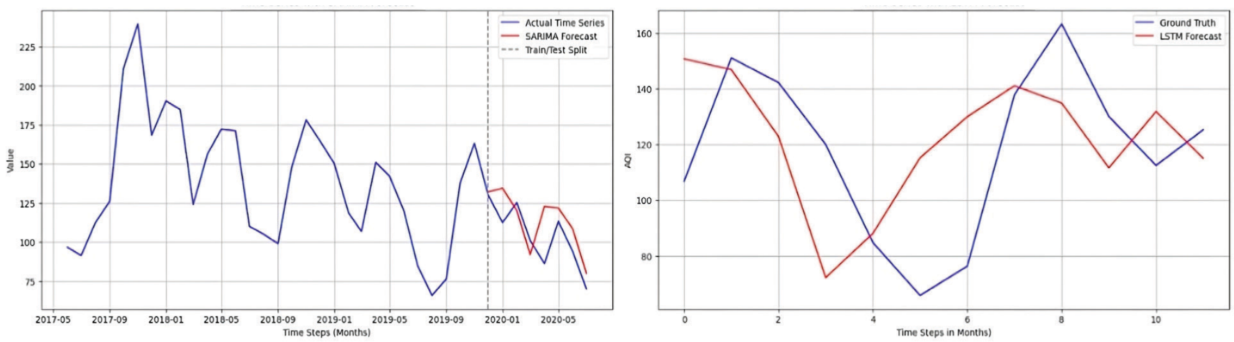

Figure 1: SARIMA and LSTM AQI Forecasts for the city of Jaipur. Line graphs showing the actual AQI value and the forecast from the SARIMA (left) or LSTM (right) model for the city of Jaipur (n=1). We use “matplotlib” to graph the values. We use “statsmodels” to make the SARIMA graph and “keras” to make the LSTM graph.

(14–16). Also, the cities that these studies focus on are generally ones larger in population. This makes sense, as the data availability for such cities is greater. However, though air pollution is a serious issue in those cities, a large majority of the Indian population does not reside in these cities yet still suffer from air pollution, and thus this work does not have much of an impact on them. Another method for AQI forecasting is using deep learning models such as the LSTM model (17,18), which passes data through itself multiple times to predict future values. Though this approach does account for seasonality, these works also have the same issue as previously mentioned of only accounting for one or two regions, thus reducing the impact they can make. Neither approach tests their statistical or deep learning counterpart, and thus it is difficult to determine which approach is the best way to tackle the problem.

In this paper, we present a method that uses ARIMA, SARIMA, and the LSTM models to tackle the task of AQI forecasting. We focus on 5 major cities in India (Jaipur, Bangalore, Delhi, Hyderabad, and Mumbai) that have a combined population of around 100 million people (19). We can forecast the AQI values with small amounts of error, thus providing accurate predictions as to the air pollution conditions of the future. Our research tests multiple approaches that also account for seasonality and is thus the best at finding the most optimal approach when compared to other methods. In addition, the scope of our impact is much greater, as we consider 5 major cities across India, and our research has the potential to make a significant impact on around 100 million people (19). Through evaluation with mean absolute error (MAE) and mean squared error (MSE) metrics, we find that the SARIMA methods performed better than the LSTM models and performed well in showing seasonal patterns in the data, supporting our hypothesis.

RESULTS

Experiments for five major cities in India were performed and the results for each experiment are presented below. To forecast the AQI values, statistical methods such as ARIMA and SARIMA, along with deep learning techniques such as LSTM, were used. We present the results from using these algorithms individually for each city, as well as its overall performance.

Jaipur

The SARIMA forecast captures the seasonal patterns and trends for the data from Jaipur. However, it follows a shorter seasonal window and has a higher trend. The LSTM forecast can be seen to follow the seasonal pattern; however, the model follows a shorter seasonal window than the actual time series (**Figure 1**).

Bangalore

The SARIMA forecast matches seasonality very well, but it does have a slightly larger seasonal window than the actual time series. It also predicts a higher trend than the actual time series. The LSTM forecast doesn’t appear to predict the data's seasonal patterns and does not perform well in general (**Figure 2**).

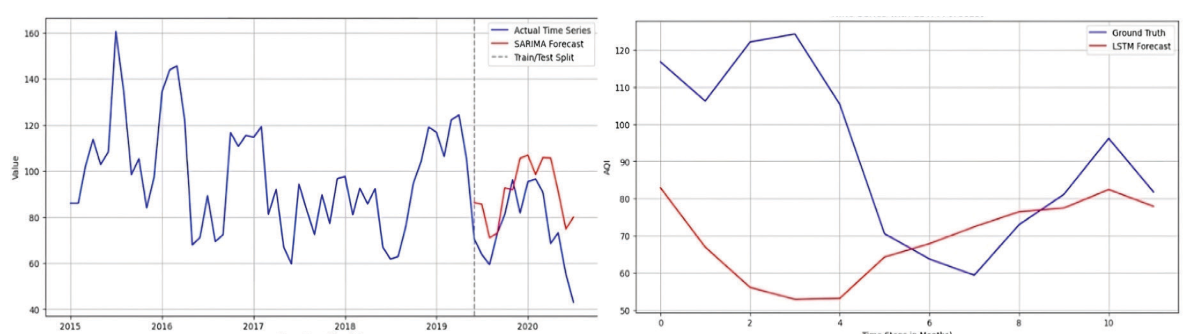

Figure 2: SARIMA and LSTM AQI Forecasts for the city of Bangalore. Two line graphs showing the actual AQI value and the forecast from the SARIMA (left) or LSTM (right) model for the city of Bangalore (n=1). We use “matplotlib” to graph the values. We use “statsmodels” to make the SARIMA graph and “keras” to make the LSTM graph.

https://doi.org/10.59720/24-079

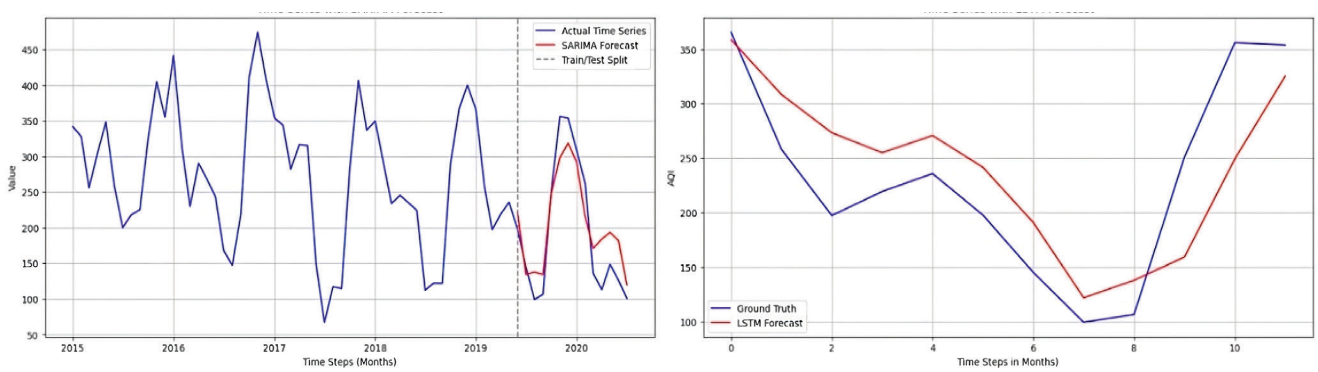

Figure 3: SARIMA and LSTM AQI Forecasts for the city of Delhi. Two line graphs showing the actual AQI value and the forecast from the SARIMA (left) or LSTM (right) model for the city of Delhi (n=1). We use "matplotlib" to graph the values. We use "statsmodels" to make the SARIMA graph and "keras" to make the LSTM graph.

Delhi

The SARIMA forecast follows seasonal patterns and trends very well, and its predictions match with the actual time series. However, it doesn't have the same time series amplitude as the actual time series, with its amplitude being slightly smaller. The LSTM forecast seems to follow seasonal patterns of the data. However, it predicts with a larger seasonal window than the actual time series and the values follow a higher trend than the actual time series (**Figure 3**).

Hyderabad

The SARIMA forecast follows seasonal patterns and trends well. However, the trend is slightly higher than the actual time series. The LSTM Forecast appears to follow the seasonal patterns well, but its performance decreases the more it predicts. There is also more fluctuation in the actual time series data than in the forecast (**Figure 4**).

Mumbai

The SARIMA forecast seems to predict seasonal patterns well, though a full determination cannot be made due to a lack of data. In addition, it predicts a higher trend than the actual time series. The LSTM forecast follows seasonal patterns but does not have the same amplitude as the actual time series. There is also a seasonal window that is much smaller than the actual time series (**Figure 5**).

Across All Cities

Most of the models have similar values for the (p, d, q)(P, D, Q) parameters. The value p represents the autoregressive order, the "q" represents the degree of first differencing, and the q value represents the order of the moving average. The capital counterparts in SARIMA represent their seasonal counterparts. Most models have a p value of 1, a d value of 0, a q value of 1, a P value of 1, a D value of 0, and a Q value of 0. For each of the values stated, there are always 1–2 exceptions. Regarding the MAE and RMSE values, one can see that both values are fairly good. For all of the MAE and RMSE values, the predictions are less than the standard deviation of the dataset. The MAE is less than the RMSE for all values, which is to be expected. However, the difference between the two is small. The RMSE value is affected more by larger errors, so a smaller difference between the two means there are few large errors, which can be seen in our results (**Table 1**).

For all LSTM models, using an LSTM shape of 50 and a dropout value of 0 works best, as determined by RMSE. All models are trained with early stopping, and the only model that needed additional training was the one for Jaipur, thus the increase in epochs trained. All models used the same learning rate. It can be seen from the MAE and RMSE categories that the model performances are decent, as they are not far from the standard deviation for the city but do surpass it. In addition, there is a larger gap between the MAE

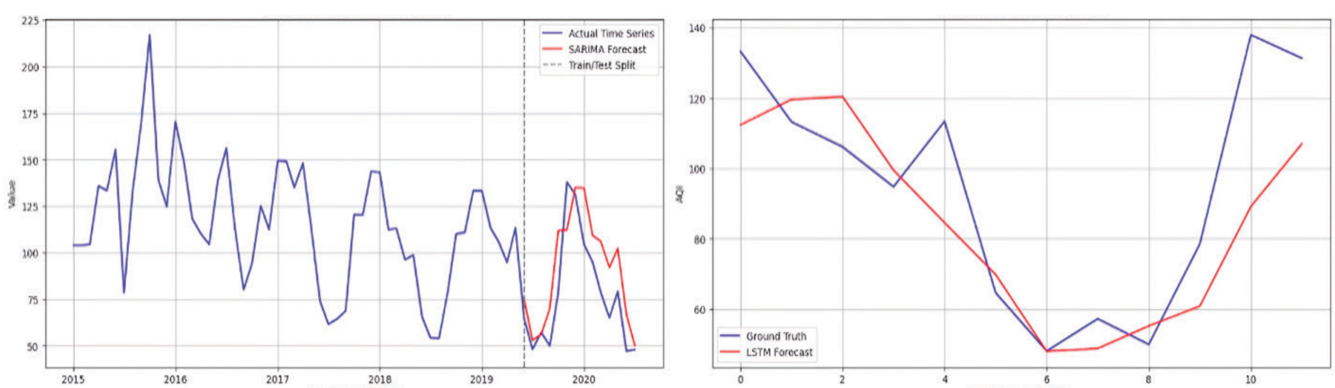

Figure 4: SARIMA and LSTM AQI Forecasts for the city of Hyderabad. Two line graphs showing the actual AQI value and the forecast from the SARIMA (left) or LSTM (right) model for the city of Hyderabad (n=1). We use "matplotlib" to graph the values. We use "statsmodels" to make the SARIMA graph and "keras" to make the LSTM graph.

https://doi.org/10.59720/24-079

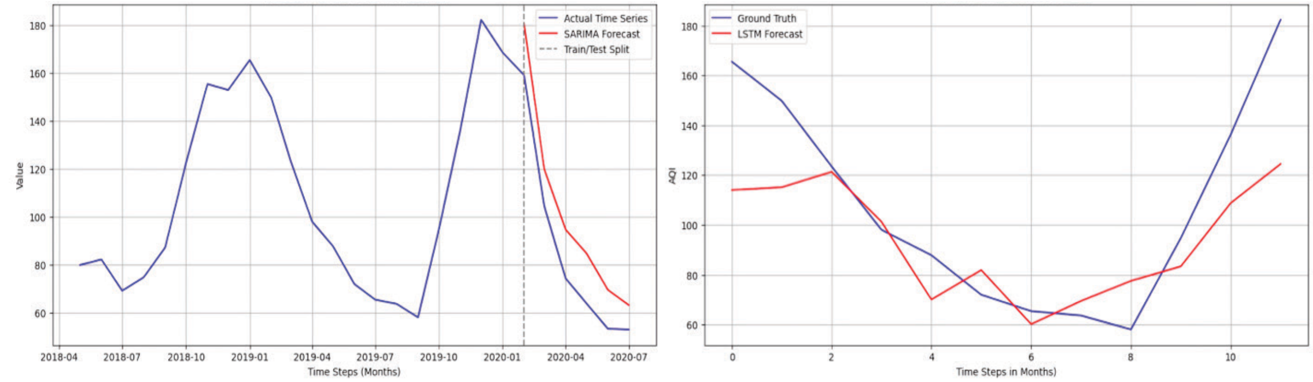

Figure 5: SARIMA and LSTM AQI Forecasts for the city of Mumbai. Two line graphs showing the actual AQI value and the forecast from the SARIMA (left) or LSTM (right) model for the city of Mumbai (n=1). We use "matplotlib" to graph the values. We use "statsmodels" to make the SARIMA graph and "keras" to make the LSTM graph.

and RMSE values than what is seen in the SARIMA results. This means that there are more values of larger errors for the LSTM predictions (**Table 2**).

It can also be seen that both SARIMA and LSTM models perform well, with all SARIMA MAE and RMSE values less than the respective standard deviation for that category, and all but one of the LSTM MAE and RMSE values being less than the respective standard deviation for that category (**Tables 1-2**). It can also be seen that for both MAE and RMSE and all cities, the SARIMA model performs better than the LSTM as there is a statistical improvement in both values for each city. This shows that for our case, the SARIMA model is better suited for the task (**Tables 1-2**). Since the MAE and MSE values for SARIMA are always less than those for LSTMs, the results support our hypothesis.

DISCUSSION

The overall results show that our approach performs well based on its MAE and RMSE values and could be applicable in helping people predict AQI values for the future. Since SARIMA's MAE and MSE values are always below those of LSTM's, our results support our hypothesis. Combined with the graphs of the results, we can see that for all cities our method of using SARIMA is an effective way of forecasting AQI. Some interesting results were that the best "d" and "D" values almost always worked at a value of 0, except for Hyderabad. This means the differencing for the time series would be best at 0. Another interesting observation is that most models found it optimal to have a "q" value of 1, except for Mumbai. This exception could have been due to a lack of data for Mumbai. A similar pattern was observed for the other SARIMA parameter values. These similarities show that the overall task of AQI forecasting in these major Indian cities has similar, core patterns with slight differences depending on intricacies found within each city (Results and parameters for SARIMA across major cities, **Table 1**).

When models follow seasonal patterns but are off in the height of peaks or time of values, this could be due to noise in the dataset or because of previous patterns that were not present during the time period they forecasted for. The latter of which could be due to changes in policy or events like the COVID-19 pandemic. One can also see that the problem is not overly complex, as a simple LSTM trained for a relatively small number of epochs can perform well, and that SARIMA, a statistical model that is not as complex as its deep learning counterparts, can perform very well on the dataset. Such improved performance with statistical models generally means a less complex problem, where the deep learning model begins to overfit the information it is trained on (**Figures 1–5**).

This research can be improved by including more data to train the models (meaning more information they can learn) or by including more cities, thus increasing the scope of the study. Further, including additional information as input, such as wind patterns or the AQI values of nearby cities, may improve model performance.

We hypothesize that traditional time series processing algorithms for forecasting, such as SARIMA, capture seasonal variations and can forecast future AQI levels using current and past AQI levels better than complex deep-learning models, like LSTMs, which are increasingly common in forecasting. The results we attain support our hypothesis, as the SARIMA

City Name	p	d	q	P	D	Q	s	MAE	RMSE	STD
Jaipur	1	0	1	1	0	1	12	12.44	16.02	40.21
Bangalore	0	0	1	1	0	1	12	16.38	19.62	23.84
Delhi	1	0	1	1	0	0	12	34.64	39.40	99.74
Hyderabad	1	1	1	1	1	0	12	17.38	20.44	35.42
Mumbai	0	0	0	1	0	0	12	17.43	17.87	26.1

Table 1: Parameters for the SARIMA model and its forecasting results across the major cities we consider. The values p, d, q, P, D, Q, and S refer to their respective SARIMA parameters. MAE stands for mean absolute error. RMSE stands for root mean squared error. STD stands for standard deviation.

City Name	LSTM Shape	Dropout	Epochs Trained	Learning Rate	MAE	RMSE	STD
Jaipur	50	0	750	0.001	21.54	26.05	40.21
Bangalore	50	0	300	0.001	20.12	25.43	23.84
Delhi	50	0	300	0.001	56.60	73.48	99.74
Hyderabad	50	0	300	0.001	21.57	28.07	35.42
Mumbai	50	0	300	0.001	15.44	22.73	26.1

Table 2: Parameters for the LSTM model and its forecasting results across the major cities we consider. MAE stands for mean absolute error and RMSE stands for root mean squared error. STD stands for standard deviation.

https://doi.org/10.59720/24-079

model is seen to forecast better than the LSTM model for both MAE and RMSE.

Our approach allows for relatively accurate prediction of AQI, which if implemented would others to prepare better and take care of their health as . The approaches we test use the statistical models SARIMA and ARIMA, as well as the deep learning model LSTM. Our results show that SARIMA performs the best forecasting and is thus most effective for the problem presented. SARIMA's higher performance also shows that the task of forecasting AQI in these Indian cities may not be so complex that it requires models such as LSTM. Our results also show that forecasting of AQI across India has many similarities regardless of location. Further technology can be developed from our research which can then make future AQI knowledge accessible to the general public so they can make informed decisions and stay safe from air pollution.

MATERIALS AND METHODS

Datasets

The dataset used in this study was made publicly available by the Central Pollution Control Board, which is the official portal of the Government of India (20). It is a dataset that contains AQI values over time for a multitude of cities in India. It is organized into multiple files based on the information they convey. The file "city_day.csv" is used from this dataset to create a time series. The "city_day.csv" file contains a table with the name of the city, the date for each row, the particle concentration in air recorded by an air quality sensor, and the AQI value. This part of the dataset has daily AQI values from around 220 sensors from their respective cities. This means that there are almost 190,000 samples in total. We used data for the 5 major cities we focused on: Jaipur, Bangalore, Delhi, Hyderabad, and Mumbai. These cities have 3,342; 20,090; 76,342; 12,036; and 20,090 samples, respectively. We chose to use Bangalore, Delhi, Hyderabad, and Mumbai because they are some of the largest cities in India and they have a large data quantity (Dataset distribution across cities, **Figure 6**). Though not being one of the top cities in the dataset, we chose to include Jaipur as well because it is a large Indian city, and to show how our algorithms are not limited to cities with significant amounts of data. For the time series approach, the dataset is filtered based on each city and only the date and AQI values were used. The dataset is then pre-processed by calculating the monthly mean. The models are trained on the first 80% of the time series data, and are tested on the last 20% of the time series data. The dataset distribution shows that the cities we consider have some of the most samples out of all the cities in the dataset (Dataset distribution across cities, **Figure 6**).

Training and Implementation Details

After preprocessing our data for each major city, the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) were performed and graphed to find if the city's data has a trend or seasonality. All major cities were found to have seasonality, so the SARIMA model was used rather than ARIMA. We chose to use SARIMA because it is a simpler statistical forecasting technique that uses autoregressive, moving average, and differencing elements to generate a forecast. We use the "statsmodels" Python package to implement SARIMA. To find optimal parameters for the model, we looped over all combinations of SARIMA

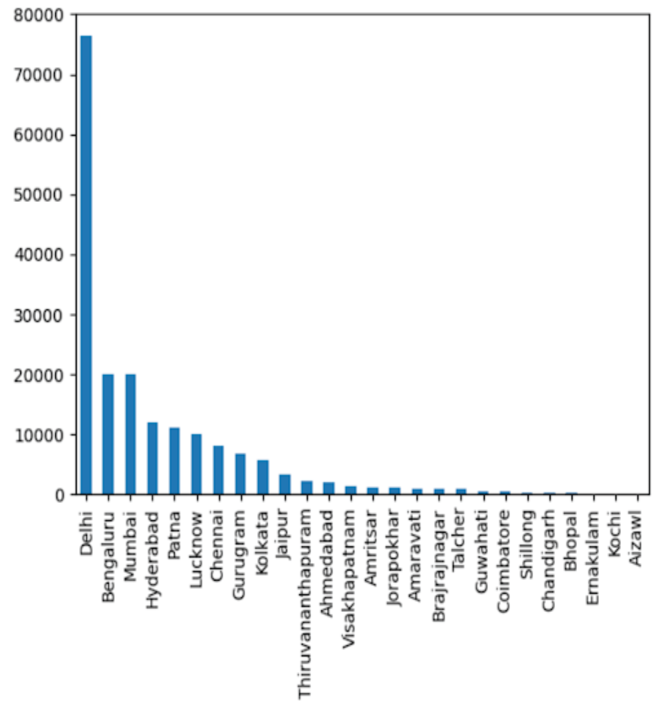

Figure 6: Distribution of AQI dataset across the different cities it contains. Bar graph showing number of values for each city present in the dataset (n=1). We use "pandas" and "matplotlib" to graph the number of samples per city.

parameters that are at most 1 value away from the values gained from the analysis of ACF and PACF. The seasonality value was kept constant at 12, as there are 12 months in a year. The model is trained on 80% of its time series data, and the rest is used for testing. This method was used to find the model with the least mean squared error (MSE) for its prediction, and the model was then saved to a "pkl" file.

Also, the performance of an LSTM is observed for these cities. We chose to use the LSTM model because it is a more complex deep learning technique built upon the Recurrent Neural Network framework, which could allow it to understand more complex patterns in the data. We use the "keras" Python package to implement LSTMs. For all cities, the LSTM must forecast for 12 months, but the input size varies depending on the availability of data. The model is trained with a learning rate of 0.001. It is trained for 300 epochs with the early stopping callback enabled. Then, the MSE of this model's predictions is found for a 12-month prediction. This MSE is then compared with the MSE calculated for the predictions of the best model from SARIMA, and the best is used in subsequent steps.

Forecasting Techniques

For each time series, we generate the PACF and ACF graphs which help us better understand the patterns of the data. We use the model ARIMA, which is a common statistical model for time series forecasting. It uses three parameters: (p, d, q). This model works well for trend-based time series data, however, is at a disadvantage in accommodating seasonality. When seasonality is present in the time series, the SARIMA model is used. The SARIMA model includes 4 additional parameters from the ARIMA model, which are (P, D, Q, s)

https://doi.org/10.59720/24-079

(21). Information from the PACF and ACF graphs helps us find parameters for SARIMA and ARIMA. The Long Short-Term Memory (LSTM) model is also used. Its structure is an advancement on the structure for recurrent neural networks (RNNs) that considers both long-term aspects of the data as well as short-term information (22).

Received: February 17, 2024
Accepted: September 6, 2024
Published: February 17, 2025

REFERENCES

1. Yu, Wenhua, *et al.* "Global Estimates of Daily Ambient Fine Particulate Matter Concentrations and Unequal Spatiotemporal Distribution of Population Exposure: a Machine Learning Modelling Study." *The Lancet Planetary Health*, vol. 7, no. 3, 2023, pp. 209-218, https://doi.org/10.1016/S2542-5196(23)00008-6.
2. Fang, Zhong, *et al.* "Air pollution's impact on the economic, social, medical, and industrial injury environments in China." *Healthcare*. vol. 9. no. 3, 2021, https://doi.org/10.3390/healthcare9030261.
3. Fuller, Richard, *et al.* "Pollution and health: a progress update." *The Lancet Planetary Health*, vol. 6, No. 6, 2022, pp. 535-547, https://doi.org/10.1016/s2542-519622)00090-0.
4. Juginović, Alen, *et al.* "Health impacts of air pollution exposure from 1990 to 2019 in 43 European countries." *Scientific Reports*, vol. 11, no. 1, 2021, pp. 22516, https://doi.org/10.1038/s41598-021-01802-5.
5. Jiang, Xu-Qin, Xiao-Dong Mei, and Di Feng. "Air pollution and chronic airway diseases: what should people know and do?." *Journal of thoracic disease*, vol. 8, no. 1, 2016, pp. 31, https://doi.org/10.3978%2Fj.issn.2072-1439.2015.11.50.
6. Babatola, Samuel Soledayo. "Global burden of diseases attributable to air pollution." *Journal of public health in Africa*, vol. 9, no. 3, 2018, https://doi.org/10.4081/jphia.2018.813.
7. "Air Pollution." World Health Organization, www.who.int/health-topics/air-pollution. Accessed 18 January 2024.
8. Manisalidis, Ioannis, *et al.* "Environmental and health impacts of air pollution: a review." *Frontiers in public health*, vol. 8, 2020, pp. 1-13, https://doi.org/10.3389%2Ffpubh.2020.00014.
9. Calderón-Garcidueñas, Lilian, *et al.* "Air pollution and your brain: what do you need to know right now." *Primary health care research & development*, vol. 16, no. 4, 2015, pp. 329-345, https://doi.org/10.1017/s146342361400036x.
10. Sharma, Rohit, *et al.* "Inferring air pollution from air quality index by different geographical areas: case study in India." *Air Quality, Atmosphere & Health*, vol. 12, 2019, pp. 1347-1357, https://doi.org/10.1007/s11869-019-00749-x.
11. Greenstone, Michael, *et al.* "The solvable challenge of air pollution in India." *India Policy Forum*. vol. 12, 2017.
12. Chang, Xinghua, *et al.* "Seasonal autoregressive integrated moving average model for precipitation time series." *Journal of Mathematics & Statistics*, vol. 8, no. 4, 2012, http://doi.org/10.3844/jmssp.2012.500.505.
13. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation*, vol. 9, no. 8, 1997, pp. 1735-1780, https://doi.org/10.1162/neco.1997.9.8.1735.
14. Kaur, Jatinder, Sarbjit Singh, and Kulwinder Singh Parmar. "Forecasting of AQI (PM2.5) for the three most polluted cities in India during COVID-19 by hybrid Daubechies discrete wavelet decomposition and autoregressive (Db-DWD-ARIMA) model." *Environmental Science and Pollution Research*, vol. 30, no. 45, 2023, pp. 101035-101052, https://doi.org/10.1007/s11356-023-29501-w.
15. Kumar, Anikender, and P. Goyal. "Forecasting of daily air quality index in Delhi." *Science of the Total Environment*, vol. 409, no. 24, 2011, pp. 5517-5523, https://doi.org/10.1016/j.scitotenv.2011.08.069.
16. Pant, Alka, *et al.* "Predictive Modeling for Forecasting Air Quality Index (AQI) Using Time Series Analysis." *Avicenna Journal of Environmental Health Engineering*, vol. 10, no. 1, 2023, pp. 38-43, http://doi.org/10.13052/jrss0974-8024.1621.
17. Belavadi, Sagar V., *et al.* "Air quality forecasting using LSTM RNN and wireless sensor networks." *Procedia Computer Science*, vol. 170, 2020, pp. 241-248, https://doi.org/10.1016/j.procs.2020.03.036.
18. Krishan, Mrigank, *et al.* "Air quality modelling using long short-term memory (LSTM) over NCT-Delhi, India." *Air Quality, Atmosphere & Health*, vol. 12, 2019, pp. 899-908, https://doi.org/10.1007/s11869-019-00696-7.
19. "Population of Cities in India 2024." Population of Cities in India 2024, *World Population Review*, worldpopulationreview.com/countries/cities/india. Accessed 18 Jan. 2024.
20. Sarkar, Mani. *Air Quality Data in India Extended*, 8 Nov. 2020, Accessed 26 June 2023.
21. Permanasari, Adhistya Erna, Indriana Hidayah, and Isna Alfi Bustoni. "SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence." *2013 international conference on information technology and electrical engineering*, 2013, pp. 203-207, http://doi.org/10.1109/ICITEED.2013.6676239.
22. Absar, Nurul, *et al.* "The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases." *Infectious Disease Modelling*, vol. 7, no. 1, 2022, pp. 170-183, https://doi.org/10.1016/j.idm.2021.12.005.