From ARIMA to Transformers: Navigating the Landscape for AQI Prediction

Pratik Dwivedi
dept. name of organization (of Aff.)
Bennett University
Greater Noida, India
prateekdwivedi30@gmail.com

Manishit Singh
dept. name of organization (of Aff.)
Bennett University
Greater Noida, India
singhmanishit95@gmail.com

Anant Teotia
dept. name of organization (of Aff.)
Bennett University
Greater Noida, India
anant.teotia@outlook.com

***Abstract*—The air quality index (AQI) is a metric used to assess the impact of air pollution on health within a short duration of timeframe. Its aim is to inform the public about the detrimental health consequences of local air pollution. The levels of air pollution in Indian urban areas (especially Delhi) have markedly risen, prompting various approaches to devising a mathematical formula for calculating the AQI. However, with the advancements in AI, researchers have used AI models for establishing a correlation between exposure to air pollution and negative health outcomes among the populace. This study offers an in-depth investigation of time series data analysis and forecasting methods, covering deep learning models, traditional machine learning (ML) methods, classical mathematical models, and cutting-edge methods. Starting with a discussion of traditional mathematical models, with a focus on their fundamental significance in capturing temporal patterns. Examples of these models are ARIMA and exponential smoothing. It does, however, recognize their shortcomings when it comes to managing intricate, non-linear relationships found in time series data. Entering the domain of traditional machine learning, it then explores deep learning models and looks into how well RNNs and LSTMs can capture complex temporal dependencies, while also exploring capabilities of Liquid Neural Networks (LNN) for the same . The research examines current developments in time series forecasting techniques, going beyond accepted methods. Comparative analysis, which assesses each technique's performance under various circumstances, is at the core of this study. Considerations include robustness to various data properties, interpret ability, and computational efficiency. The results help practitioners choose the best method for their particular set of time series data requirements.**

***Index Terms*—Time Series Forecasting, Data Analysis, Forecasting, Liquid Neural Networks, LNN**

I. Introduction

As metropolitan cities continue to grow and develop, a multitude of environmental issues, including deforestation, toxic material release, solid waste disposal, and air pollution, have garnered unprecedented attention. Among these concerns, the issue of air pollution in cities has reached alarming levels, necessitating timely monitoring of pollution levels. The World Health Organization (WHO) cautions that air pollution, particularly particulate matter (PM), poses a threat that is severe enough to increase mortality rates [Fan+19], [Gle+20]. Furthermore, pollution from automobiles is contributing to the rise in NO2, CO, NH3, PM2.5, and PM10, while industrial sources are the source of pollutants including SO2, CO, O3, B (benzene), T (toluene), and X (xylene). Based on increased Particulate Matter (PM) concentrations, India ranks second after Kuwait in terms of pollution severity [CMW18], [GAF18], [MZH20].

Predicting air quality has now become a focal point in air pollution research and the creation of accurate forecasting models for Air Quality Index (AQI) regarding major air pollutants in urban settings is imperative. Analyzing and forecasting time series data has special opportunities and problems because of its temporal interdependence and sequential structure. In this research work, we take a deep dive into time series data processing and forecasting strategies for AQI prediction, covering everything from traditional procedures to state-of-the-art methodologies. This paper examines both theoretical underpinnings and practical applications, examining them via the prism of an actual real-world data set—the Air Quality Index (AQI) of Delhi, which government agencies registered from 2011 to 2014.

The paper is divided into 6 sections. .....

II. Literature Review

Time series forecasting of air quality index (AQI) in urban environments, particularly in cities like Delhi, is essential for understanding and mitigating the impact of air pollution on public health and the environment. Various studies have explored different modeling approaches and techniques to predict AQI levels, ranging from linear and nonlinear methods to deep learning architectures.

Linear and nonlinear modeling approaches have been extensively studied in the context of urban air quality prediction. For instance, [**4**] compared the performance of linear modeling techniques such as Partial Least Squares Regression (PLSR) with nonlinear methods including Multivariate Polynomial Regression (MPR) and Artificial Neural Networks (ANNs). They found that ANNs outperformed linear models, with Generalized Regression Neural Network (GRNN) yielding high correlations for pollutants like Respirable Suspended Particulate Matter (RSPM), Nitrogen Dioxide (NO2), and Sulfur Dioxide (SO2). Additionally, [**5**] focused on forecasting AQI using regression models, with Support Vector Regression (SVR) demonstrating high performance compared to linear models. They evaluated the models using statistical criteria

such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

Deep learning approaches have also gained traction in AQI forecasting due to their ability to capture complex temporal patterns. [**6**] trained a deep learning model using a Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) to predict Tropospheric Ozone (O3) concentrations, considering its impact on human health. They highlighted the complexity of O3 formation processes and sources, emphasizing the suitability of RNNs for such predictions. Furthermore, [**10**] proposed a spatio-temporal deep learning model for air quality prediction, achieving superior performance compared to traditional time series models like Autoregressive Integrated Moving Average (ARIMA) and Support Vector Regression (SVR). They emphasized the importance of considering spatial and temporal correlations in air quality prediction, which the proposed model effectively captured.

Incorporating time-series forecasting techniques for AQI prediction has been a subject of interest in recent research. [**7**] evaluated four time-series analysis methods, including SARIMAX, ARIMA, AR, and LSTM, for predicting logistics companies' staffing needs and order volume. They found that SARIMAX outperformed other methods in predicting order volumes, highlighting its applicability in resource planning and management for logistics companies. Additionally, [**8**] provided a comprehensive literature review on time series forecasting methods, discussing various techniques such as Autoregression (AR), Moving Average (MA), ARIMA, and LSTM. They emphasized the importance of stationarity in time series modeling and discussed the performance of different methods in forecasting AQI.

In conclusion, time series forecasting of AQI in cities requires a holistic approach that combines linear and non-linear modeling techniques, deep learning architectures, and advanced time series analysis methods. These studies provide valuable insights into the modeling approaches and techniques for accurately predicting AQI levels, contributing to better air quality management and public health protection.

III. Methodology Used in This Paper

Understanding how time series data behaves over time is crucial. One thing we check is whether the data is "stationary" or not. There are two tests we use - Augmented Dickey-Fuller (ADF) and is Kwiatkowski-Phillips-Schmidt-Shin (KPSS). ADF is to checking if our data has some patterns that mess with its consistency over time, while KPSS is looking at the overall trend and if maintains consistency. Their results aren't the same because they focus on different things meaning we cannot just swap one for another.

A. Traditional Methods

We're starting with the classics—ARIMA and exponential smoothing—known to be the go-to tools in time series analysis.

¡——— add ARIMA explanation and Exponential Smoothing explantion ———-¿

B. Basic Machine Learning Models

Diving deeper into our exploration, we now delve into the world of conventional machine learning (ML) techniques. Specifically, we're shining a spotlight on the versatility of linear regression and decision trees. These techniques are kind of like the all-rounders of the ML game – they can adapt to all sorts of situations. We're not just looking at how good they are at predicting stuff; we're also checking out how fast they can do it and whether we can make sense of their predictions in real-world situations.

But, here's the catch: like all superheroes, they have their limitations. These methods might struggle when our data gets too messy or has a lot of twists and turns. Also, they might not be the quickest when we're dealing with a ton of data. So, while we're putting them to the test, we're also keeping an eye on how they handle the tricky parts and whether they're practical for real-life situations. Our goal here is to really understand how well these familiar ML techniques can handle the tricky world of time series forecasting.

C. Deep Learning Techniques

Embarking on a more advanced leg of our journey, we delve into the world of deep learning (DL) models. In this exciting phase, we unleash the power of simple deep neural networks (DNN), alongside the heavyweights - Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs). Our mission: to grapple with the intricate patterns in the AQI dataset, taking advantage of these cutting-edge technologies.

As we step into this high-tech territory, our main goal is to uncover what these DL techniques can really do. We're curious about whether these more sophisticated models can do a better job than the classic ones, especially when things get complicated and non-linear in real-world time series data. It's like we're peeling back the layers of complexity to understand how these deep learning models can supercharge our analysis.

These DL models, while powerful, aren't without their limitations. They can get computationally hungry and might need a lot of data to flex their muscles properly. So, as we explore their potential, we're also keeping an eye on the challenges they might throw our way. This part of our journey adds a whole new level of sophistication to our analytical toolbox, and we're excited to see what these high-tech models can bring to the table.

D. Recent Advancements

Embarking beyond the usual trails of classical and regular machine learning models, our study takes a dive into the cutting-edge world of time series forecasting. We're not just sticking to the classics; we're turning our gaze to the cool, new developments in the field. Think of it as peeking into the future of forecasting! In this exciting phase, we're checking out innovative techniques like Time GPT, Patch TST, N-hits, N-beats, TimesNet, and TSmixer and Liquid Neural Networks. Our goal? To figure out how these new methods might shake

things up and push the boundaries of what we can do with time series data.

Liquid Neural Networks are new additions to the stale but growing list of intelligent systems which are designed to be adaptable to future data that they might see during inference or deployment using an ODE solver as its weight adjusting module that changes the network's learned set of weights by whatever factor it calculates during inference

While we're excited about these techniques, we're also looking into their limitations. We want to know where they might stumble or struggle, and if they can really outshine the tried-and-true methods we've been using. It's like having a critical eye on the latest tech gadgets – they're cool, but are they really better than what we already have? Our analysis here is not just about the shiny and new; it's about figuring out if they can truly outperform the trusty tools we've been using for time series forecasting.

IV. Exploring the Data set

A. Description and Raw Features - AQI Dataset

This sub-section describes the dataset Air Quality Index(AQI). This data set consists of AQI data collected over days spanning 1 to 5 years from 2015-2020 for various metropolitan cities, namely Amaravati, Chandigarh, Hyderabad, Kolkata, Patna, Delhi, Amritsar, Gurugram, Vishakhapatnam, allowing a complete representation of the entire 5 years AQI data.This data set comprises of the attributes such as: City name, Date of measurement, Respirable Suspended Particulate Matter (RSPM) or Particulate Matter (PM10), Particulate matter(PM2.5), Nitric Oxide(NO),Nitrous oxide(NO2), nitrogen oxides(NOx), Ammonia(NH3), Carbon Oxide(CO), Sulphur Dioxide(SO2),Ozone(O3), Benzene, Toluene, Xylene, AQI, AQI Bucket. In handling missing data, we have chosen to discard records where significant attributes, such as AQI Bucket and AQI, are empty for the specified time periods, and consequently, those columns have been dropped. However, certain periods and cities have empty data records due to various reasons. For instance, in Ahmedabad, data from November 2015 to May 2016 and November 2016 to October 2017 is missing. In Amravati, there are gaps in the data for August 2019 to October 2019. Patna's data records were empty from April 2017 to September 2017, while Vishakhapatnam's data was unavailable from April 2017 to October 2017. Additionally, for some cities like Chandigarh, the data is available from 2019 onwards, not from 2015. Similarly, data for the city of Kolkata is available from 2018.

B. Observations

As observed in the graph visualizations of the RSPM/PM10 data points for each year (from *Fig. 1, Fig. 2, Fig. 3, Fig. 4, Fig. 5*), we observe a much denser visualization for *Fig.3* and *Fig. 4* since the data for 2014 and 2015 is in considerably greater quantity than 2012 and 2013 in comparison.

Within a year no visual patterns can be discerned, interestingly enough there is no conclusive pattern in the entire 2012 to 2015 data visualization as well, which maybe a cause of the imbalance in the data readings, which calls for a mathematical test for this data to check if there indeed a pattern in these RSPM/PM10 readings.

Observing the complete plot of 2012 to 2015 data for SO2 *Fig. 6* and NO2 *Fig. 7* we can come to the same conclusion that there are no evident patterns in the raw data, calling for appropriate pre-processing techniques to mold the data into a time series.

C. Data Preparation

First we split the data into two sets one comprising of the 2012 and 2013 readings and the second comprising of the remaining 2014 and 2015 to create a balance in both sets according to the number of observations.

One thing we should always check if the data is "stationary", in broad terms it means the data is independent of time and there is a clear pattern that it follows, in more detail what we are looking to observe through the stationary characteristic is that if the data is dependent on seasons, which is referred to as "seasonality" in the data, or follows a trend over time which is good and bad depending how you look at it.

Being stationary means independence from the bound of time, the easiest set to predict would be a stationary data, which is good but usually not the case in environmental data just like our case of the AQI dataset. So, for our case we need to employ the seasoned techniques of checking for stationary readings - the Augmented Dickey-Fuller Test (ADF), and the Kwiatkowski-Phillips-Schmidt-Shin test (KPSS).

1) ADF test for stationarity: The *Augmented Dickey-Fuller* test is a type of statistical test called a *unit root test*. In probability theory and statistics, a unit root is a feature of some stochastic processes (such as random walks) that can cause problems in statistical inference involving time series models.*In simple terms, the unit root is non-stationary* but does not always have a trend component.

ADF test is conducted with the following assumptions:

- ***Null Hypothesis (HO)***: Series is non-stationary, or series has a unit root.
- ***Alternate Hypothesis(HA)***: Series is stationary, or series has no unit root.

If the null hypothesis is failed to be rejected, this test may provide evidence that the series is non-stationary.

Conditions to Reject Null Hypothesis(HO):

If Test statistic is **less than** Critical Value and p-value **less than** 0.05 – *Reject Null Hypothesis(HO) i.e.* time series does not have a unit root, **meaning it is stationary**. It does not have a time-dependent structure.

So, from this test we need each reading in our data to have a P-value of less than 0.05, here is a table of results for each unique column reading for each set of the bifurcated data.

2) KPSS test for stationarity: A key difference from the ADF test is the null hypothesis of the KPSS test is that the series is stationary. So practically, the interpretation of p-value is just the opposite of each other. That is, if the p-value is ¡ significance level (say 0.05), then the series is non-stationary.

ADF Statistics	2012 RSPM/PM10	2012 SO2	2012 NO2
Test Statistic	-3.828587	-1.445353	-2.306111
p-value	0.002628	0.560351	0.170005
Lags Used	3.000000	16.000000	5.000000
Observations Used	284.000000	271.000000	282.000000
Critical Value (1%)	-3.453587	-3.454713	-3.453754

TABLE I
ADF TEST RESULTS FOR EACH FEATURE FOR 2012

ADF Statistics	2013 RSPM/PM10	2013 SO2	2013 NO2
Test Statistic	-2.018454	-7.625891	-2.198024
p-value	0.278541	2.069277e-11	0.206963
Lags Used	7.000000	1.000000	5.000000
Observations Used	263.000000	269.000000	265.000000
Critical Value (1%)	-3.455461	-3.454896e	-3.455270

TABLE II
ADF TEST RESULTS FOR EACH FEATURE FOR 2013

ADF Statistics	2014 RSPM/PM10	2014 SO2	2014 NO2
Test Statistic	-2.018454	-7.625891	-2.198024
p-value	0.278541	2.069277e-11	0.206963
Lags Used	7.000000	1.000000	5.000000
Observations Used	263.000000	2.690000e+02	265.000000
Critical Value (1%)	-3.455461	-3.454896	-3.455270

TABLE III
ADF TEST RESULTS FOR EACH FEATURE FOR 2014

KPSS Statistics	2012 RSPM/PM10	2012 SO2	2012 NO2
Test Statistic	-3.828587	-1.445353	-2.306111
p-value	0.002628	0.560351	0.170005
Lags Used	3.000000	16.000000	5.000000
Observations Used	284.000000	271.000000	282.000000
Critical Value (1%)	-3.453587	-3.454713	-3.453754

TABLE IV
KPSS TEST RESULTS FOR EACH FEATURE FOR 2012

Whereas in the ADF test, it would mean the tested series is stationary.

The KPSS test is conducted with the following assumptions.

- ***Null Hypothesis (HO)***: Series is trend stationary or series has no unit root.
- ***Alternate Hypothesis(HA)***: Series is non-stationary, or series has a unit root.

Note: The hypothesis is reversed in the KPSS test compared to ADF Test.

If the null hypothesis is failed to be rejected, this test may provide evidence that the series is trend stationary.

Conditions to Fail to Reject Null Hypothesis(HO)

If the Test Statistic is **less than** Critical Value and p-value is less than 0.05 – Fail to *Reject Null Hypothesis(HO),* i.e., time series does not have a unit root, meaning it is trend stationary.

Conditions to Reject Null Hypothesis(HO):

If Test statistic is **less than** Critical Value and p-value **less than** 0.05 – *Reject Null Hypothesis(HO) i.e.* time series does not have a unit root, **meaning it is stationary**. It does not have a time-dependent structure.

V. IMPLEMENTATION AND RESULTS

A. SARIMAX

Here are the results of applying SARIMAX on the AQI dataset of 2012

KPSS Statistics	2013 RSPM/PM10	2013 SO2	2013 NO2
Test Statistic	-2.018454	-7.625891e+00	-2.198024
p-value	0.278541	2.069277e-11	0.206963
Lags Used	7.000000	1.000000e+00	5.000000
Observations Used	263.000000	2.690000e+02	265.000000
Critical Value (1%)	-3.455461	-3.454896e+00	-3.455270

TABLE V
KPSS TEST RESULTS FOR EACH FEATURE FOR 2013

KPSS Statistics	2014 RSPM/PM10	2014 SO2	2014 NO2
Test Statistic	-2.607770	-6.478201e+00	-2.588092
p-value	0.091402	1.314939e-08	0.095483
Lags Used	16.000000	6.000000e+00	16.000000
Observations Used	712.000000	7.220000e+02	712.000000
Critical Value (1%)	-3.439568	-3.439440e+00	-3.439568

TABLE VI
KPSS TEST RESULTS FOR EACH FEATURE FOR 2014

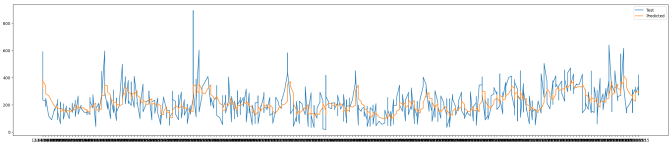

Fig. 1. SARIMAX results on 2012-2015 RSPM/PM10

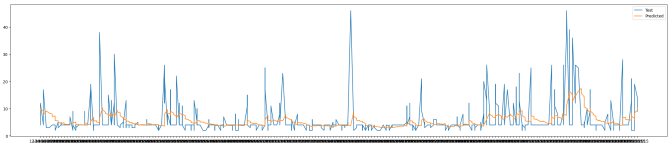

Fig. 2. SARIMAX results on 2012-2015 SO2

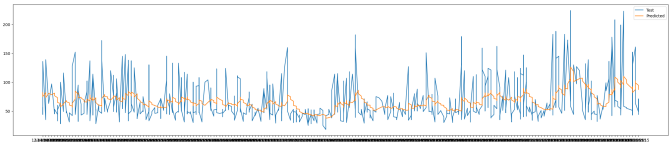

Fig. 3. SARIMAX results on 2012-2015 NO2

B. Linear Regression

The results are underwhelming, on the lower count dataset of 2012 which had only about.....

VI. Conclusion

References

[CMW18] Shisheng Chen, Kuniaki Mihara, and Jianxiu Wen. "Time series prediction of CO2, TVOC and HCHO based on machine learning at different sampling points". In: *Building and Environment* 146 (2018), pp. 238–246.

[GAF18] Zeinab Ghaemi, Abbas Alimohammadi, and Mahdi Farnaghi. "LaSVM-based big data learning system for dynamic prediction of air pollution in Tehran". In: *Environmental monitoring and assessment* 190 (2018), pp. 1–17.

[Fan+19] Delin Fang et al. "Clean air for some: Unintended spillover effects of regional air pollution policies". In: *Science advances* 5.4 (2019), eaav4707.

[Gle+20] Drew A Glencross et al. "Air pollution and its effects on the immune system". In: *Free Radical Biology and Medicine* 151 (2020), pp. 56–68.

[MZH20] Jiaqing Miao, Xiaobing Zhou, and Ting-Zhu Huang. "Local segmentation of images using an improved fuzzy C-means clustering algorithm based on self-adaptive dictionary learning". In: *Applied Soft Computing* 91 (2020), p. 106200.