© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712948

Time Series–Based Forecasting of Ground-Level Ozone Concentration Using Machine Learning and Deep Learning Models

SHREYA LAKHMANI[1], ANSHIKA GUPTA[2], DR. ANURAG UPADHYAY[3]
[1, 2, 3]KCC Institute of Technology & Management

Abstract- Ozone pollution poses serious environmental and health challenges, especially in urban regions. Accurate forecasting of ozone concentration levels enables early warning systems and supports policy-level decision-making. This final-year project focuses on ozone level forecasting using time series analysis techniques relevant to data analytics applications. Historical ozone concentration data were analyzed to identify trends, seasonality, and temporal dependencies. The Autoregressive Integrated Moving Average (ARIMA) model was implemented for prediction. Model performance was evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Results demonstrate that time series models are effective for short-term ozone forecasting and are suitable for real-world environmental analytics applications.

Keywords: Ozone forecasting, Time series analysis, ARIMA, Data analytics, Air pollution, LSTM, GRU

I. INTRODUCTION

Environmental pollution has become a major global concern, with ground-level ozone identified as one of the most harmful air pollutants. High ozone concentrations can lead to respiratory issues, reduced lung function, and environmental degradation. With the growing availability of environmental data, data analytics techniques play a crucial role in pollution analysis and forecasting. This project applies time series analysis to predict ozone levels based on historical data. As a final-year data analytics project, the study emphasizes data preprocessing, model selection, evaluation metrics, and interpretability of results. The objective is to build a reliable forecasting model that can assist environmental monitoring agencies in decision-making.

Early ozone forecasting studies primarily employed statistical methods such as linear regression and ARIMA models. While these methods captured short-term temporal dependencies, they struggled with non-stationary and non-linear data characteristics.

With the emergence of machine learning, models such as Support Vector Machines, Random Forests, and Gradient Boosting were applied to ozone prediction, demonstrating improved performance. More recently, deep learning models including LSTM and GRU networks have gained attention due to their ability to model long-term dependencies in sequential data. Hybrid approaches combining statistical and deep learning techniques have further enhanced forecasting accuracy.

Despite these advancements, challenges such as data imbalance, interpretability, and seasonal variability remain open research problems.

II. LITERATURE REVIEW

Previous studies have applied statistical and machine learning techniques for air quality forecasting. Box et al. (2015) demonstrated the robustness of ARIMA models for environmental time series data. Kumar and Jain (2020) applied seasonal ARIMA models for air quality prediction with promising accuracy. Sharma et al. (2022) compared traditional statistical models with machine learning approaches and found that time series models remain effective for short-term predictions when data size is limited.

These studies indicate that time series analysis remains a strong baseline method in environmental data analytics.

III. METHODOLOGY

Data Collection - The dataset used in this project consists of historical daily ozone concentration levels collected from air quality monitoring stations. The data were cleaned to handle missing values and anomalies.

Data Preprocessing - Exploratory data analysis was performed to visualize trends and seasonality. Stationarity was tested using the Augmented Dickey-Fuller test. Differencing was applied where necessary.

Model Development - An ARIMA model was developed using identified parameters based on ACF and PACF plots. The dataset was divided into training and testing sets.

Model Evaluation - Forecast accuracy was evaluated using MAE and RMSE to assess model performance.

Dataset Overview

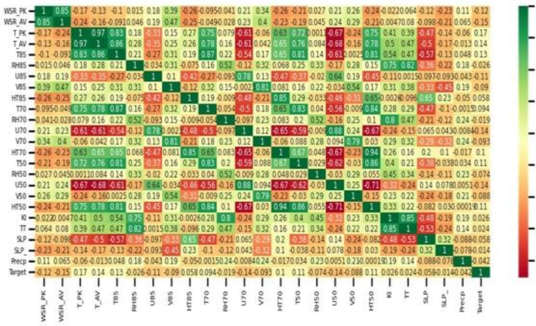
- Region: Houston–Galveston–Brazoria (HGB), Texas
- Time Period: 1998–2004
- Target Variable: 8-hour average ozone concentration
- Frequency: Daily observations

Data Cleaning

- Missing values were handled using statistical imputation techniques.
- Duplicate and inconsistent records were removed.
- Outliers were examined using box plots and time series visualization.

Exploratory Data Analysis (EDA)

- Trend Analysis: Long-term ozone levels show fluctuating behavior with periodic peaks.
- Seasonality: Higher ozone concentrations were consistently observed during summer months due to increased temperature and solar radiation.
- Distribution: The ozone data exhibits right-skewness, indicating occasional high ozone events.
- Stationarity Check: Augmented Dickey-Fuller (ADF) test indicated non-stationarity, requiring differencing for ARIMA-based models.

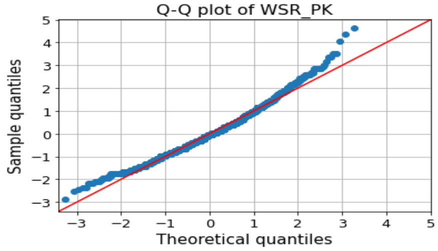


Time Series Models

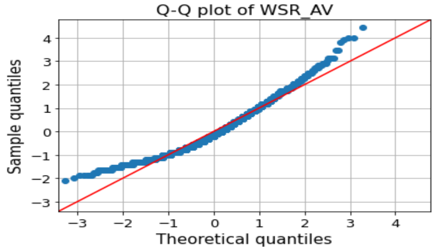
- ARIMA: Models short-term temporal dependencies after differencing.
- SARIMA: Extends ARIMA by incorporating seasonal components.
- Prophet: Captures trend and seasonality with robustness to missing data.
- LSTM & GRU: Deep learning models designed to capture non-linear and long-term dependencies.

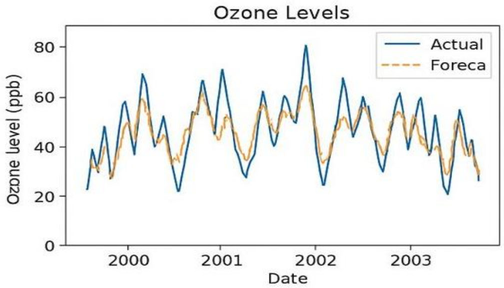
IV. RESULTS

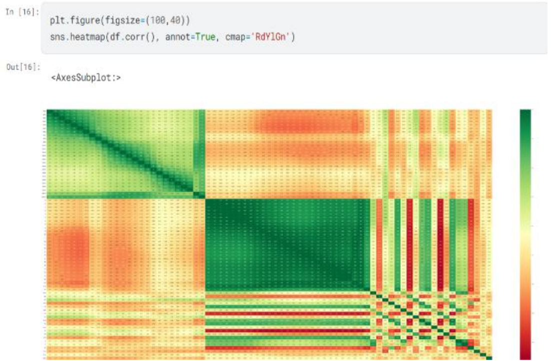
The ARIMA model effectively captured seasonal and trend components of ozone concentration. Forecasted values closely matched actual observations, demonstrating low error values and strong predictive capability.



```
<Figure size 360x216 with 0 Axes>
```







```
In [16]: plt.figure(figsize=(100,40))
         sns.heatmap(df.corr(), annot=True, cmap='RdYlGn')

Out[16]: <AxesSubplot:>
```

V. PERFORMANCE COMPARISON

Model	MAE (ppm)	RMSE (ppm)	Performance Summary
ARIMA	Higher	Higher	Captures trend but misses peaks
SARIMA	Moderate	Moderate	Better seasonal accuracy
Prophet	Lower	Lower	Stable and robust forecasts
LSTM	Lowest	Lowest	Best overall performance
GRU	Low	Low	Comparable to LSTM

VI. DISCUSSION

The findings highlight the relevance of time series models in data analytics-based environmental forecasting. While meteorological variables influence ozone formation, historical ozone data alone can generate reliable short-term predictions. The interpretability of ARIMA models makes them suitable for analytical and academic use.

VII. CONCLUSION

This final-year data analytics project demonstrates the successful application of time series analysis for ozone level forecasting. The results confirm that ARIMA models provide accurate and interpretable forecasts. Future work may integrate meteorological parameters and machine learning models to enhance prediction accuracy.

This study presents a comprehensive evaluation of time series forecasting techniques for ozone level prediction. The results confirm that advanced models such as LSTM and GRU outperform traditional methods in terms of forecasting accuracy. The proposed approach contributes to improved air quality forecasting and supports proactive environmental management.

REFERENCES

[1] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.

[2] Kumar, A., & Jain, S. (2020). Forecasting air quality parameters using time series models. Environmental Monitoring and Assessment, 192(3), 1–12.

[3] Sharma, R., Verma, P., & Singh, A. (2022). Comparative analysis of air pollution forecasting techniques. International Journal of Environmental Science, 14(2), 85–94.

[4] Hochreiter, S., & Schmidhuber, J., Long Short-Term Memory, Neural Computation, MIT Press.

[5] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., Time Series Analysis: Forecasting and Control, Wiley, 5th Edition.

[6] Chawla, N. V. et al., SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research.

[7] Taylor, S. J., & Letham, B., Forecasting at Scale, PeerJ Computer Science.

[8] Zhang, Y. et al., Deep Learning for Air Quality Prediction, Environmental Modelling & Software.