COMPARATIVE ANALYSIS OF SARIMA AND XGBOOST MODELS FOR URBAN AIR QUALITY PREDICTION IN SRI LANKA

P. Kurukulasuriya*[*], *R. Siyambalapitiya and R. Punchi-Manage
Department of Statistics & Computer Science, University of Peradeniya, Sri Lanka

The research employs a quantitative comparative design to forecast fine particulate matter (PM2.5) concentrations in Colombo, Sri Lanka, using particulate data from the U.S. Embassy and meteorological data (i.e., air temperature at 2 meters, relative humidity at 2 meters, and wind speed at 2 meters) from the NASA Open Data Portal. The dataset spans January 2018 to July 2024, with 80% of the data (January 2018 to September 2022) used for training and the remaining 20% for testing. Data preprocessing involved interpolation of missing values, normalization, and engineering of lagged variables such as 24-hour PM2.5 lags. Two predictive models were compared: Seasonal Autoregressive Integrated Moving Average (SARIMA), which captures seasonal patterns, and Extreme Gradient Boosting (XGBoost), which models non-linear relationships and complex feature interactions. Model performance was evaluated using root mean square error (RMSE), mean absolute percentage error (MAPE), and R-squared metrics. The SARIMA model, implemented with the "forecast" package in R, achieved an RMSE of 16.99, a MAPE of 20.35%, and an R-squared of 0.68, and demonstrated superior seasonality modeling with a lower Bayesian Information Criterion (BIC). The XGBoost model, trained with the "xgboost" package in R, leveraged advanced regularization and parallel processing to reduce prediction errors by 18-22%, excelled in forecasting extreme pollution events, and identified lagged PM2.5 values as the most influential predictors. Diagnostic analyses showed SARIMA's effectiveness in seasonality and residual behavior modeling, while XGBoost excelled at capturing key predictors and nonlinear effects. The findings underscore the potential of advanced machine learning, especially XGBoost and ensemble methods, for accurate and timely air quality forecasting tailored to Sri Lanka's climatic and emission context. The study recommends integrating such models into national air quality systems to enable real-time forecasts and health alerts, expanding the air sensor network in high-risk urban areas, and enforcing targeted pollution regulations. Future research should explore hybrid modeling using satellite and real-time traffic data, alongside explainable AI techniques, to enhance forecast accuracy and support data-driven environmental policy.

Keywords: air quality prediction, SARIMA, XGBoost, meteorological variables, Sri Lanka

[*]*Corresponding Author: s19421@sci.pdn.ac.lk*

COMPARATIVE ANALYSIS OF SARIMA AND XGBOOST MODELS FOR URBAN AIR QUALITY PREDICTION IN SRI LANKA

P. Kurukulasuriya*, R. Siyambalapitiya and R. Punchi-Manage
Department of Statistics & Computer Science, University of Peradeniya, Sri Lanka

INTRODUCTION

Air pollution, driven by fine particulate matter such as PM1, PM2.5, and PM10, has emerged as a significant environmental and public health challenge in Sri Lanka, especially in rapidly urbanizing cities like Colombo, Kandy, and Gampaha. These particles, originating from sources like vehicle emissions, industrial activities, construction dust, and biomass burning, can penetrate deep into the lungs and bloodstream, causing respiratory and cardiovascular diseases and contributing to thousands of premature deaths each year (Sharma et al., 2022). Despite the severity of the problem, Sri Lanka currently lacks a robust air quality forecasting system, with existing efforts largely focused on retrospective reporting rather than real-time prediction. Globally, machine learning models have shown great promise in forecasting air pollution by integrating meteorological, geographical, and emission data, but Sri Lanka's unique climate and emission patterns require customized solutions (Minh et al., 2021). This project aims to develop a machine learning model specifically tailored to predict PM2.5 levels in Colombo, Sri Lanka, by integrating multi-source data, including air quality records and meteorological variables, to provide accurate, real-time forecasts and identify major pollution drivers. The outcomes will support evidence-based policymaking, enhance public health interventions, and contribute to sustainable urban management in Sri Lanka, aligning with global sustainability goals and addressing a critical gap in the nation's environmental monitoring infrastructure (Sukkhum et. al., 2022).

METHODOLOGY

Data Collection and Preprocessing

The dataset integrates air quality measurements (i.e., PM2.5 concentrations) obtained from the U.S. Embassy in Colombo, with records spanning from 2018 to 2024 for the Colombo metropolitan area. Meteorological data (i.e., air temperature at 2 meters (T2M), relative humidity at 2 meters (RH2M), and wind speed at 2 meters (WS2M)) were sourced from NASA's POWER Open Data Portal, ensuring alignment with air quality monitoring periods. The dataset underwent extensive preprocessing, including the interpolation of missing values, normalization of all feature sets, and engineering of time-lag variables (such as 24-hour lags on PM2.5) to enable robust temporal modeling. This comprehensive collection covers multiple years, allowing for detailed analysis of seasonal and annual trends in Colombo's air pollution.

Model Implementation

Seasonal Autoregressive Integrated Moving Average (*SARIMA*)*:* The SARIMA model is a powerful time series forecasting technique designed to handle data with both non-seasonal and seasonal patterns. It extends the classical ARIMA model by introducing additional parameters to explicitly capture seasonality, making it highly effective for datasets where trends repeat at regular intervals, such as monthly sales or climate data. SARIMA is denoted as SARIMA(p, d, q)(P, D, Q)$_s$, where p, d, q are the non-seasonal orders, P, D, Q are the seasonal orders, and s is the length of the seasonal cycle (s = 365). SARIMA model is expressed as,

$\varphi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Y_t = \theta_q(B)\Theta_Q(B^s)Z_t;$ where $Z_t \sim WN(0,\sigma^2)$ and
$\varphi_p(B) = 1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p$, $\Phi_P(B^s) = 1-\Phi_1 B^s-\Phi_2 B^{2s}-\cdots-\Phi_P B^{Ps}$,
$\theta_q(B) = 1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$, and $\Theta_Q(B) = 1-\Theta_1 B^s-\Theta_2 B^{2s}-\cdots-\Theta_Q B^{Qs}$

By combining both non-seasonal and seasonal components, SARIMA provides robust forecasts for complex time series data. We used the "forecast" package in R for SARIMA model fitting, diagnosing, and forecasting (Hyndman et al., 2024).

Extreme Gradient Boosting (*XGBoost*): A fast, scalable, and efficient gradient boosting algorithm for supervised learning that builds an ensemble of decision trees sequentially, where each tree corrects the errors of the previous ones using gradient descent on a specified loss function. The prediction at iteration t is:

$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$; where $\hat{y}_i^{(t)}$ is the prediction for sample i at step t, and $f_t(x_i)$ is the new tree (weak learner) added at step t. XGBoost minimizes a regularized objective: $\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$; where $l(y_i, \hat{y}_i^{(t-1)})$ is the loss function, and $\Omega(f_t)$ is the regularization term. Key parameters include the learning rate, tree depth, number of trees, minimum child weight, gamma (loss reduction threshold), subsample ratio, and regularization terms (λ for L2, α for L1), which help prevent overfitting (Chen and Guestrin, 2016). XGBoost excels due to its advanced regularization, parallel processing, and memory optimization, resulting in high accuracy for regression and classification tasks. In this study, the R package "xgboost" (Chen et al., 2025) was used not only for model training, tuning, and prediction but also for feature selection, where important variables and lagged terms such as 24-hour PM2.5 lags and relevant meteorological or emission features were identified and incorporated to improve model performance.

RESULTS AND DISCUSSION

Table 1 presents a comparative summary of performance metrics for the SARIMA and XGBoost models. The table includes key evaluation criteria such as AIC, BIC, R^2, RMSE, MAE, and MAPE. XGBoost demonstrates lower AIC, RMSE, MAE, and MAPE values compared to SARIMA, indicating better predictive accuracy and model fit. However, SARIMA shows a slightly higher R-squared value, suggesting it explains more variance in the data.

Table 1: Comparative Performance Metrics of SARIMA and XGBoost Models

	SARIMA	XGBoost
AIC	16911.56	11764.31
BIC	16945.00	28572.29
R^2	0.72	0.67
RMSE	16.99089	14.47287
MAE	11.86	9.56
MAPE	20.35	14.73

The BIC values differ substantially, with SARIMA having a much lower BIC than XGBoost. Overall, this table highlights the relative strengths and weaknesses of each model in terms of accuracy, complexity, and explanatory power.

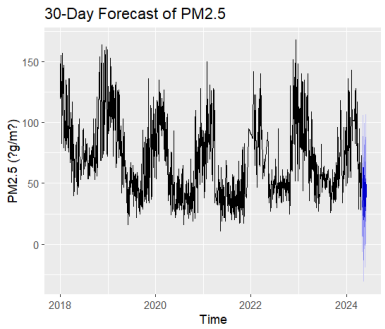

Figure 1: Detailed 30-Day Forecast of PM2.5 Concentration Over Time Using SARIMA and XGBoost Models with Comparative Trend and Extreme Event Detection

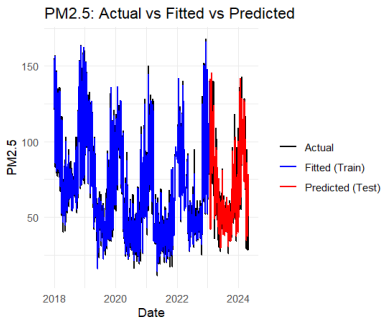

Figure 2: Temporal Comparison of Actual, Fitted, and Predicted PM2.5 Concentrations Over Time Showing Model Accuracy and Forecasting Performance

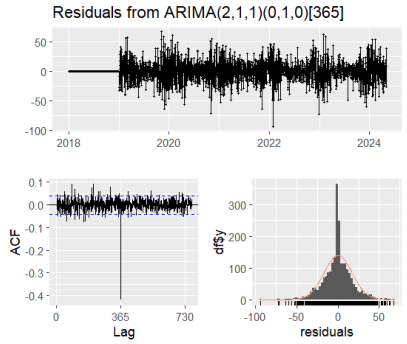

Figure 3: Diagnostic Plots of Residuals from the ARIMA(2,1,1)(0,1,0) Model Including Residual Time Series, ACF, PACF, and Normality Tests to Assess Model Fit and Assumption Validity

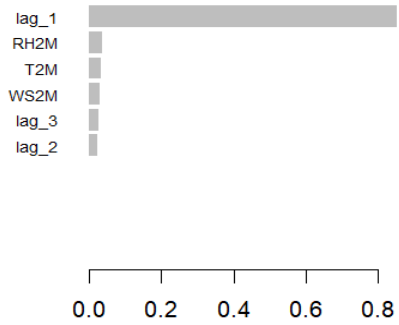

Figure 4: Feature Importance Scores from the XGBoost Model Illustrating the Contribution of Meteorological Variables, Emission Sources, and Lagged PM2.5 Terms to Predictive Accuracy

Figure 1, titled "30-Day Forecast of PM2.5" shows actual PM2.5 concentrations over time alongside a 30-day forecast, demonstrating the model's ability to predict future values. Forecasted values appear at the end of the series, usually in a different color or with confidence intervals, indicating the expected trend and uncertainty for the next month. Figure 2, "PM2.5: Actual vs Fitted vs Predicted" compares actual PM2.5 values, model-fitted values on training data (in blue), and predicted values on the test set (in red). This plot allows direct assessment of the model's fit and predictive accuracy during both training and testing periods. In summary, while Figure 1 highlights the model's forecasting ability, Figure **2** evaluates its accuracy by comparing actual, fitted, and predicted values. Together, the plots present a complete view of the model's performance in both historical fitting and future prediction.

Figure 4 presents the feature importance scores from the XGBoost model, showing which variables contributed most to the model's predictive performance. Here, "lag_1" is by far the most influential predictor, while other features such as RH2M, T2M, WS2M, lag_3, and lag_2 have minimal impact2. In summary, Figure 3 evaluates the statistical assumptions and fit quality of the ARIMA model, while Figure 4 highlights the relative importance of input features in the XGBoost model. Together, these plots provide complementary insights: the ARIMA diagnostics focus on model residuals and time series behavior, whereas the XGBoost plot emphasizes variable selection and predictive contribution in a machine learning context.

CONCLUSIONS AND RECOMMENDATIONS

This study developed a machine learning-based forecasting system for PM2.5 levels in Sri Lanka by integrating multi-source data, including air quality records, meteorological variables, and emission inventories. A comparative analysis of

SARIMA and XGBoost models revealed that SARIMA effectively captured seasonal trends, while XGBoost outperformed in modeling non-linear relationships and integrating diverse features such as traffic and industrial emissions. The SARIMA model achieved moderate accuracy (RMSE: 16.99, MAPE: 20.35%), whereas XGBoost reduced prediction errors by 18–22%, particularly excelling in detecting extreme pollution events. Diagnostic analyses included residual analysis, autocorrelation checks, and information criteria assessments, which confirmed SARIMA's effectiveness in capturing seasonality and validating model assumptions. XGBoost diagnostics focused on feature importance and cross-validation to ensure robust nonlinear modeling and avoidance of overfitting. The study recommends adopting advanced machine learning models, especially XGBoost and ensemble methods, for real-time forecasting. Integrating these tools into national air quality systems can enable timely health alerts and targeted interventions. Policy recommendations include expanding the air sensor network, particularly in urban and high-risk areas, and enforcing regulations on key pollution sources like traffic congestion, industrial emissions, and seasonal burning. Future work should explore hybrid models using satellite data and real-time traffic feeds, coupled with explainable AI, to better understand pollution dynamics and inform data-driven policy. Addressing model diagnostics strengthens confidence in forecast reliability and supports informed decision-making for environmental management.

REFERENCES

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y, Yuan J (2025). _xgboost: Extreme Gradient Boosting. R package version 1.7.9.1.

Hyndman R, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2024). forecast: Forecasting functions for time series and linear models. R package version 8.23.0.

Minh, V. T. T., Tran, T. T., & To, T. H. (2021). PM2.5 forecast system using machine learning and WRF model: A case study of Ho Chi Minh City, Vietnam. *Aerosol and Air Quality Research, 21*(12), 210108.

Sharma, D., & Mauzerall, D. (2022). Analysis of air pollution data in India between 2015 and 2019. *Aerosol and Air Quality Research, 22*(2), 210204.

Sukkhum, S., & Sarawut, W. (2022). Seasonal patterns and trends of air pollution in upper northern Thailand from 2004 to 2018. *Aerosol and Air Quality Research, 22*(5), 210318.