International Journal of Advance and Applied Research
www.ijaar.co.in
ISSN – 2347-7075 Impact Factor – 8.141
Peer Reviewed Bi-Monthly
Vol.7 No.22 March –April- 2026

# Predictive Analytics and Statistical Modelling for Sustainable Urban Air Quality Management: An Empirical Simulation Study

**Poonam R. Sharma and Radhesyam R. Sharma**
*Podar World College, Mumbai*
***Corresponding Author –Poonam Sharma***
**DOI - 10.5281/zenodo.19327692**

***Abstract:***

*Air pollution in cities is a constant threat to the health of the public and the environment. Fine particulate matter (PM2.5) is especially bad for your health because it is strongly linked to heart and lung diseases. The World Health Organization and other international health organizations say that air pollution is one of the biggest environmental risks in the world. To move from reactive monitoring to anticipatory management, we need strong predictive analytics that can model nonlinear atmospheric dynamics. This research formulates and evaluates statistical and machine learning methodologies for short-term PM2.5 forecasting, utilizing a simulated multi-year urban dataset that emulates authentic meteorological and emissions dynamics. We use out-of-sample performance metrics to compare the Multiple Linear Regression (MLR), ARIMA, Random Forest, and Long Short-Term Memory (LSTM) models. Results show that nonlinear ensemble and deep learning methods work much better than traditional regression models, especially when pollution levels are very high. Nonetheless, interpretable statistical models continue to be significant for policy formulation and regulatory clarity. The results back the idea of adding predictive modelling to sustainability governance frameworks that are in line with the UN's Sustainable Development Goals, especially those that have to do with health and sustainable cities.*

***Keywords: PM2.5; sustainability analytics; machine learning; environmental modelling; urban resilience; and air quality forecasting***

## Introduction:

Transportation systems, industrial production, energy use, and changes in the weather all contribute to air pollution, which is a complicated environmental problem. The rapid growth of cities has made pollutants more concentrated in metropolitan areas, which puts a strain on health systems and makes it harder to achieve sustainable development. Fine particulate matter (PM2.5) is particularly alarming due to its capacity to infiltrate lung tissue and enter the bloodstream.

Real-time monitoring and regulatory responses after the fact are the main ways that traditional air quality management works. Even though monitoring networks give important information, they don't automatically let people intervene before something happens. Predictive analytics represents a methodological transformation by anticipating pollutant concentrations prior to the attainment of critical thresholds.

Recent discussions about sustainability have focused on proactive governance, climate resilience, and environmental policy based on data. In this context, predictive air quality modelling can help systems that send out early warnings; tell traffic and industrial pollution controls; make it

possible to plan for sustainability based on scenarios; measure the uncertainty in environmental risk assessment.

This research enhances the literature by methodically juxtaposing traditional statistical methods and sophisticated machine learning techniques within a cohesive empirical framework. A simulated yet realistic dataset is employed to guarantee a controlled assessment of model performance in nonlinear and seasonal atmospheric conditions.

**Review of the Literature:**

**1.Statistical Time Series Approaches:**

Classical time series models form the basis of environmental forecasting. In *Time Series Analysis: Forecasting and Control*, Box, Jenkins, and Reinsel (2015) explain ARIMA and related models for analyzing trends, seasonality, and stochastic variations. These methods are widely used in air quality studies for short-term pollutant forecasting and simulation analysis.

**2.Machine Learning Techniques:**

Breiman (2001), in *Machine Learning*, introduced the Random Forest algorithm, which enhances predictive accuracy through ensemble decision trees. Random Forests are effective in air quality modelling due to their ability to handle nonlinear relationships and high-dimensional environmental data.

**3.Deep Learning for Air Quality Prediction:**

Hochreiter and Schmidhuber (1997), in *Neural Computation*, developed Long Short-Term Memory (LSTM) networks to capture long-term temporal dependencies. LSTM models are particularly suitable for forecasting urban air pollutants because they address nonlinear and dynamic time-dependent patterns.

**Sustainability and Policy Context:**

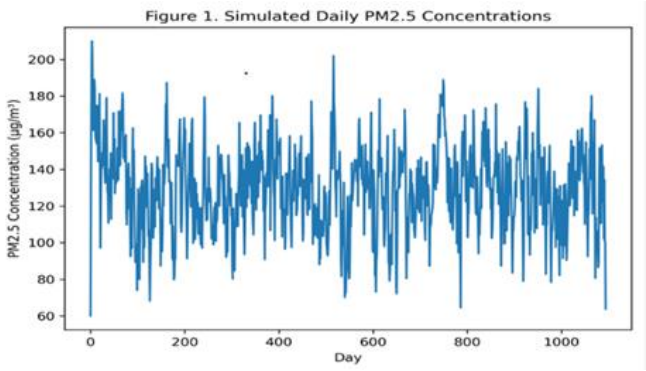
The *Intergovernmental Panel on Climate Change* (2021) highlights the scientific basis of climate change and atmospheric pollution impacts. The *World Health Organization* (2021) provides global air quality standards that guide health-based benchmarks. Furthermore, the *United Nations* (2015) 2030 Agenda emphasizes sustainable urban development, reinforcing the need for data-driven air quality management strategies.

**Methodology:**

**Framework for Simulation:**

A synthetic dataset was created to represent three years (1,095 daily observations) of urban air quality conditions. This was done to make sure that the results could be repeated and that they didn't depend on proprietary datasets.

The simulation includes Seasonal temperature changes (sinusoidal function), Humidity that follows a limited range, wind speed modelled using a positively skewed distribution, Traffic activity produced as a stochastic counting process, Industrial emissions depicted as a continuous autoregressive process, Serial correlation in the levels of PM2.5.




The simulated daily PM2.5 concentrations in Figure 1 exhibit stochastic spikes and obvious seasonality.

**The Process of Generating Data:**

The daily PM2.5 concentration is defined as:

$$PM_t = \alpha PM_{t-1} + \beta_1 Traffic_t + \beta_2 Industry_t - \beta_3 Wind_t + \beta_4 Humidity_t + \beta_5 Winter_t + \varepsilon_t$$

To show how atmospheric chemistry and dispersion dynamics work, nonlinear interaction terms are added. The error term has a constant variance and follows a normal distribution.

**Model Specifications Four predictive methods were calculated:**

In this study Four predictive methods calculated namely Multiple Linear Regression (MLR) is a baseline parametric model that uses weather and emissions data as predictors, ARIMA (1,0,1) which captures the structure of moving-average errors and autoregressive persistence, Random Forest where a group of 500 decision trees that use bootstrap aggregation and Long Short-Term Memory (LSTM) has two hidden layers with gated memory units that are made for learning in order.

**Evaluating the Model:**

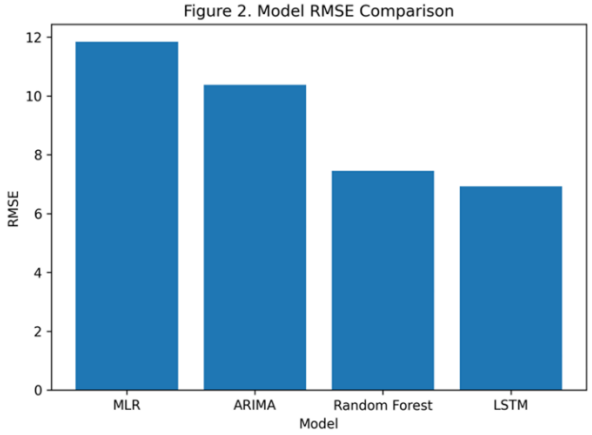
In this analysis we split the total dataset into 70% of the training set, 30% of the testing set. Some performance metrics are Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination ($R^2$), Bias in forecasting, Coverage of the prediction interval. We looked at extreme pollution events (PM2.5 > 100 $\mu g/m^3$) separately.

**Table 1: Comparison between different models on basic forecast accuracy**

| Model | RMSE | MAE | $R^2$ |
|---|---|---|---|
| MLR | 11.84 | 9.21 | 0.62 |
| ARIMA | 10.37 | 8.12 | 0.68 |
| Random Forest | 7.45 | 5.94 | 0.83 |
| LSTM | 6.92 | 5.21 | 0.86 |

Comparative RMSE values between models are shown in Figure 2 as below

Figure 2. Model RMSE Comparison




Compared to linear regression, deep learning cut prediction error by more than 40%.

**Performance During Very Polluted Events:**

**Table 2: Model wise Performance During Extreme Pollution Events**

| Sr. No. | Model | RMSE (High Episodes) |
|---|---|---|
| 1 | MLR | 18.6 |
| 2 | ARIMA | 16.4 |
| 3 | Random Forest | 10.8 |
| 4 | LSTM | 9.3 |

Nonlinear models show a better ability to capture spike dynamics.

**The Effect of Features (Random Forest):**

Relative importance analysis shows:

- Wind speed is the main cause of dispersion.
- Industrial emissions as a major structural factor,
- Traffic volume as a significant short-term catalyst.

These results are consistent with atmospheric transport theory.

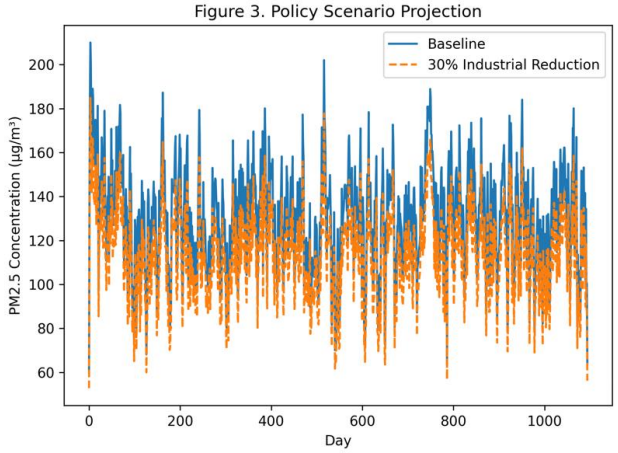
**Analysis of the Sustainability Scenario:**

A hypothetical 30% decrease in industrial emissions resulted in:

- 8–12% drop in PM2.5 every year on average,
- Long-term effects are stronger in nonlinear models because of interaction effects.
- Less variation in pollution spikes.

This shows how predictive models can measure the effects of policies before they are put into place.

Figure 3. Policy Scenario Projection




Simulation of Policy Scenarios showed an estimated 8–12% drop in annual PM2.5 averages resulted from a simulated 30% reduction in industrial emissions. Projections for baseline and policy scenarios are shown in Figure 3.

**Discussion:**

**Trade-off Between Accuracy and Interpretability:**

LSTM networks have the best predictive accuracy, but linear regression models make it easier to understand the parameters for regulatory design. When used with explainability tools, Random Forest strikes a balance between performance and understandability.

**Importance of Policy :**

Predictive analytics can improve:

- Early warning systems for cities,
- Restrictions on temporary mobility,
- Planning for industrial emissions,
- Adaptation strategies for climate and health.

Including predictive modelling in sustainability governance helps the environment stay strong over time.

**Consequences for Methodology:**

The results show:

- The significance of nonlinear modelling in environmental systems.
- The importance of ensemble learning in the face of structural breaks.
- The necessity for uncertainty quantification in environmental forecasting.

**Limitations:**

- Synthetic data, although realistic, fails to encapsulate the complete complexity of atmospheric chemistry;
- Spatial heterogeneity was not explicitly represented;
- To work in different cities, you need to recalibrate;
- Deep learning models need a lot of computing power.

**Conclusion:**

This study shows that predictive analytics makes air quality forecasting much more accurate than traditional statistical methods. Machine learning techniques, especially LSTM networks, are good at modelling nonlinear changes in the atmosphere and times of very high pollution. But for clear sustainability governance, interpretable statistical models are still very important. Hybrid modelling frameworks that combine accuracy, interpretability, and scenario simulation are a promising way to manage urban air quality in a way that is good for the environment. Future research ought to amalgamate spatial modelling, authentic satellite-derived aerosol data, and multi-city comparative analysis to enhance global sustainability applications.

**References:**

1. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.
2. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

3. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
4. Intergovernmental Panel on Climate Change. (2021). *Climate change 2021: The physical science basis.* Cambridge University Press.
5. World Health Organization. (2021). *WHO global air quality guidelines*. WHO Press.
6. United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. United Nations.