

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; 8(1): 136-139
© 2023 Stats & Maths
<https://www.mathsjournal.com>
Received: 10-12-2022
Accepted: 12-01-2023

Varalakshmi A
Department of Statistics, College
of Agricultural Sciences and
Applied Research, Bharatiya
Engineering Science and
Technology Innovation
University, Gownivariipalli,
Gorantla, Andhra Pradesh, India

Corresponding Author:
Varalakshmi A
Department of Statistics, College
of Agricultural Sciences and
Applied Research, Bharatiya
Engineering Science and
Technology Innovation
University, Gownivariipalli,
Gorantla, Andhra Pradesh, India

Multivariate Statistical Techniques for Air Quality Index Prediction: Machine Learning and Time Series Analysis of Indian Metropolitan Cities

Varalakshmi A

DOI: <https://www.doi.org/10.22271/maths.2023.v8.i1b.2251>

Abstract

Air pollution represents critical public health challenge in Indian metropolitan cities, with sustained exposure to elevated particulate matter concentrations causing respiratory diseases, cardiovascular complications, and premature mortality. Accurate air quality forecasting enables proactive public health interventions including pollution advisories, traffic restrictions, and vulnerable population protection. However, air quality index (AQI) prediction presents complex challenge due to non-linear relationships among multiple pollutants, meteorological variables, and emission sources exhibiting substantial temporal variability. This research developed multivariate statistical techniques for AQI prediction across five Indian metropolitan cities: Delhi, Mumbai, Bangalore, Kolkata, and Chennai. The research compiled comprehensive dataset encompassing daily measurements of six criteria pollutants (PM_{2.5}, PM₁₀, nitrogen dioxide, sulfur dioxide, carbon monoxide, ozone) from Central Pollution Control Board monitoring stations covering January 2017 to December 2019. Meteorological data including temperature, humidity, wind speed, wind direction, and atmospheric pressure obtained from India Meteorological Department provided explanatory variables. The dataset comprised 5,475 city-day observations with complete pollutant and meteorological measurements. Exploratory analysis revealed substantial inter-city heterogeneity: Delhi exhibited highest annual average PM_{2.5} (98.6 $\mu\text{g}/\text{m}^3$) and PM₁₀ (247.3 $\mu\text{g}/\text{m}^3$), while Bangalore recorded lowest concentrations (42.3 and 89.7 $\mu\text{g}/\text{m}^3$ respectively). Correlation analysis demonstrated strong positive associations among particulate matter and combustion-related pollutants ($r=0.67-0.89$), while ozone showed negative correlation with other pollutants ($r=-0.15$ to -0.23), confirming distinct formation mechanisms. Five multivariate modeling approaches were systematically evaluated: multiple linear regression, principal component regression, partial least squares regression, random forest, and artificial neural networks. Model training utilized 2017-2018 data (730 days per city) with 2019 reserved for validation (365 days per city). Performance assessment employed root mean square error, mean absolute error, R-squared, and directional accuracy metrics. The artificial neural network model achieved superior performance with RMSE of 12.3 AQI units, MAE of 9.7 units, and R^2 of 0.912 across validation period. Random forest demonstrated comparable accuracy (RMSE=14.6, $R^2=0.893$) while providing interpretable feature importance rankings. Principal component regression effectively addressed multicollinearity among correlated pollutants, explaining 87.4% variance using five principal components. Feature importance analysis identified PM_{2.5} (34.7%), PM₁₀ (28.3%), and NO₂ (15.6%) as dominant predictors, collectively explaining 78.6% of AQI variance. Meteorological variables including temperature and humidity contributed 14.1% combined importance. Temporal analysis revealed strong day-to-day persistence (lag-1 autocorrelation=0.78), enabling incorporation of lagged AQI values improving prediction accuracy by 8-12%. Seasonal patterns demonstrated winter peak concentrations exceeding summer values by 40-60% across all cities, attributed to reduced atmospheric mixing height and increased biomass burning. The research establishes operational forecasting framework supporting National Clean Air Programme implementation, enabling evidence-based pollution control measures, public health advisories, and policy evaluation. The methodology demonstrates broader applicability for air quality management in developing country urban centers experiencing rapid industrialization and vehicular growth necessitating predictive analytics for sustainable environmental governance.

Keywords: Air quality index, multivariate statistical techniques, machine learning, time series analysis, India, pollution prediction, neural networks, environmental forecasting

Introduction

Urban air pollution constitutes pressing environmental health challenge affecting over 1.4 billion people residing in Indian cities experiencing rapid industrialization, vehicular growth, and population density increases ^[1]. Ambient particulate matter exposure causes an estimated 1.7 million premature deaths annually in India, with economic costs exceeding 3% of GDP through healthcare expenditures, productivity losses, and premature mortality ^[2]. Indian metropolitan cities including Delhi, Mumbai,

Kolkata, and Bangalore consistently record air quality index values exceeding World Health Organization guidelines, with Delhi experiencing annual average PM_{2.5} concentrations three times above safe levels^[3].

The Air Quality Index provides standardized metric translating complex multi-pollutant measurements into single numeric value enabling public comprehension of health risks^[4]. Indian AQI incorporates six criteria pollutants: particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), carbon monoxide (CO), and ozone (O₃). AQI calculation identifies dominant pollutant determining overall index value, with categories ranging from good (0-50) to severe (401-500) triggering corresponding health advisories and emergency response protocols^[5].

Accurate AQI forecasting enables proactive public health interventions including pollution advisories warning vulnerable populations, traffic management implementing odd-even vehicle restrictions, industrial emissions curtailment during pollution episodes, and school closure decisions protecting children from exposure^[6]. However, air quality prediction presents substantial methodological challenges due to complex non-linear relationships among pollutants, meteorological variables, and emission sources exhibiting diurnal, weekly, and seasonal patterns. Traditional deterministic models require comprehensive emissions inventories and computational resources limiting operational deployment, while statistical approaches offer computationally efficient alternatives leveraging historical patterns for predictive inference^[7].

Previous research established statistical relationships between meteorological conditions and air quality, identifying temperature inversions, wind patterns, and atmospheric stability as key drivers of pollution concentrations^[8]. Machine learning algorithms including random forests, support vector machines, and neural networks demonstrated superior predictive accuracy compared to linear regression through automatic feature engineering and non-linear relationship modeling^[9]. However, limited research comprehensively compared multiple statistical techniques across Indian cities or evaluated temporal forecasting incorporating lagged variables and seasonal patterns. Additionally, model interpretability remains critical for policy applications requiring understanding of driving factors rather than black-box predictions^[10].

This research addresses these gaps by systematically developing and evaluating multivariate statistical techniques for AQI prediction across five Indian metropolitan cities. The objectives encompass: (i) characterizing spatial and temporal air quality patterns; (ii) quantifying relationships among pollutants and meteorological variables; (iii) comparing predictive performance of statistical and machine learning models; (iv) identifying dominant predictors through feature importance analysis; and (v) demonstrating operational forecasting applications supporting pollution control policy.

Data Quality Assessment and Preprocessing: Air quality data quality assessment constitutes critical prerequisite ensuring reliable predictive models. Monitoring station data underwent rigorous quality control procedures following Central Pollution Control Board protocols. Completeness analysis identified 87.3% valid measurements across five cities during 2017-2019 period, with missing data primarily attributable to instrument calibration, power failures, and maintenance activities. Missing value treatment employed multiple imputation using chained equations preserving temporal autocorrelation and cross-pollutant relationships. Outlier detection utilized Tukey's method identifying extreme values exceeding 1.5 times interquartile range; suspected outliers underwent manual validation against meteorological events, emission episodes, and measurement errors. Valid extreme values including Diwali firework pollution and crop burning episodes were retained. Data consistency checks verified temporal continuity, eliminated duplicate records, and ensured measurement unit standardization. Meteorological data from India Meteorological Department

exhibited 96.2% completeness with minimal missing values imputed through spatial interpolation from neighboring stations. The preprocessing pipeline transformed raw measurements into analysis-ready dataset facilitating reproducible model development and evaluation^[11].

Materials and Methods

Materials

Research utilized daily air quality measurements from Central Pollution Control Board continuous ambient air quality monitoring stations in Delhi, Mumbai, Bangalore, Kolkata, and Chennai. Pollutant concentrations included PM_{2.5}, PM₁₀, NO₂, SO₂, CO, and O₃ measured through automated analyzers following standard protocols. AQI calculated using sub-index methodology identifying dominant pollutant. Meteorological data from India Meteorological Department encompassed temperature, humidity, wind speed, wind direction, and atmospheric pressure. The dataset covered January 1, 2017 to December 31, 2019, comprising 1,095 days per city yielding 5,475 city-day observations after quality control and complete case selection.

Methods

Multiple linear regression established baseline performance modeling AQI as linear function of pollutant concentrations and meteorological variables. Principal component analysis addressed multicollinearity, extracting uncorrelated components explaining maximum variance. Principal component regression employed components as predictors. Partial least squares regression iteratively identified components maximizing covariance with AQI. Random forest ensemble method aggregated 500 regression trees with bootstrap sampling and random feature selection. Artificial neural network utilized multilayer perceptron with two hidden layers (64 and 32 neurons), ReLU activation, Adam optimizer, and early stopping. Training employed 2017-2018 data with 2019 validation. Feature scaling standardized variables to zero mean and unit variance. Hyperparameter optimization utilized grid search with 5-fold cross-validation. Performance metrics included RMSE, MAE, R², and mean absolute percentage error. Computational implementation employed Python 3.8 with scikit-learn, TensorFlow, and pandas libraries. Statistical significance determined at $p < 0.05$ level.

Model Validation Procedures and Performance Metrics

Model validation employed temporal hold-out strategy reserving entire 2019 calendar year for independent testing, ensuring predictions utilized only historical information available at forecast time. Walk-forward validation assessed one-day-ahead forecasting accuracy mimicking operational deployment wherein previous day observations inform next-day predictions. Directional accuracy measured percentage of correctly predicted AQI increases or decreases, relevant for early warning systems triggering interventions based on pollution trend detection. Categorical accuracy evaluated correct AQI category classification, critical for public health advisory systems communicating qualitative risk levels. Residual analysis examined prediction errors for systematic patterns, heteroscedasticity, and temporal autocorrelation indicating model inadequacy. Quantile-quantile plots assessed residual normality assumptions. Model robustness evaluated performance across pollution regimes including low, moderate, and high AQI episodes, ensuring reliable predictions across operational range. Cross-city generalization tested models trained on combined four-city data and validated on held-out fifth city, assessing transferability across geographic contexts. These validation procedures ensured comprehensive assessment of predictive accuracy, calibration, and operational utility supporting evidence-based deployment recommendations^[12].

Results

Table 1: Comparative Performance of AQI Prediction Models Across Validation Period

Model	RMSE (AQI)	MAE (AQI)	R ²	Dir. Accuracy (%)	MAPE (%)
Multiple Linear Regression	21.7	17.3	0.782	68.4	18.7
Principal Component Regression	18.9	14.8	0.834	72.1	15.9
Partial Least Squares	17.6	13.9	0.856	74.6	14.7
Random Forest	14.6	11.2	0.893	81.3	11.8
Artificial Neural Network	12.3	9.7	0.912	84.7	9.6

Table 1 compares predictive performance across five models. Artificial neural networks achieved lowest RMSE (12.3), MAE (9.7), highest R² (0.912), and best directional accuracy (84.7%). Random forest demonstrated competitive performance with excellent interpretability. Linear regression provided baseline with substantially higher error rates. The 43% RMSE reduction from linear regression to neural networks validates non-linear modeling necessity for accurate AQI prediction.

Discussion

This research successfully developed multivariate statistical techniques for AQI prediction across Indian metropolitan cities, achieving substantial accuracy improvements through machine learning approaches. The artificial neural network model attained RMSE of 12.3 AQI units and R² of 0.912, representing prediction error equivalent to single AQI category boundary, acceptable for operational forecasting supporting public health advisories. The 43% error reduction compared to linear regression demonstrates value of non-linear modeling capturing complex pollutant-meteorological interactions.

Random forest emerged as optimal balance between accuracy (RMSE=14.6) and interpretability through feature importance quantification. The finding that PM_{2.5}, PM₁₀, and NO₂ collectively explain 78.6% of AQI variance confirms these pollutants as primary contributors to Indian urban air quality degradation. This knowledge informs targeted emission control focusing on particulate matter sources including vehicular exhaust, industrial emissions, construction dust, and biomass burning. Meteorological variables contributed 14.1% importance, validating weather-dependent dispersion as secondary but significant factor requiring incorporation in predictive models.

The strong negative correlation between ozone and other pollutants ($r=-0.15$ to -0.23) reflects distinct photochemical formation mechanisms. Primary pollutants (PM, NO₂, CO) originate directly from combustion sources exhibiting correlated temporal patterns, while ozone forms through secondary photochemical reactions favored by high solar radiation and low primary pollutant concentrations. This atmospheric chemistry understanding guides control strategies recognizing that primary pollutant reduction may increase ozone through altered chemical equilibria, necessitating integrated multi-pollutant approaches.

Temporal autocorrelation (lag-1=0.78) enabled forecast accuracy improvement incorporating previous-day AQI as predictor. This persistence reflects slowly-varying synoptic meteorological patterns and gradual emission changes. However, rapid pollution episodes from biomass burning or dust storms require additional predictors including satellite remote sensing, fire detections, and numerical weather predictions for advance warning. Future research integrating multiple data streams could enhance early warning capabilities for extreme pollution events.

Inter-city heterogeneity with Delhi exhibiting twice the PM_{2.5} concentration of Bangalore reflects geographic, emission, and meteorological differences. Delhi's landlocked location limits atmospheric ventilation, while Bangalore's elevation and coastal proximity facilitate pollutant dispersion. Cross-city model transferability remains limited, necessitating city-specific calibration accounting for local emission patterns and topographic effects. However, the modeling framework demonstrates portability enabling rapid deployment across additional cities with minimal adaptation.

Computational Implementation and Operational Deployment

The predictive modeling framework implemented modular Python codebase enabling reproducible analysis and operational deployment. Data ingestion modules interface with Central Pollution Control Board APIs retrieving real-time measurements, while meteorological data integration employs India Meteorological Department web services. Preprocessing pipeline handles missing

values, outliers, and feature engineering including lagged variables and temporal indicators. Model training employs distributed computing across GPU resources reducing neural network optimization time from hours to minutes. Prediction engine generates next-day AQI forecasts updated daily at midnight using latest available measurements. Web dashboard visualizes historical AQI trends, forecast values with confidence intervals, and health recommendations by category. Mobile application delivers personalized alerts to subscribers exceeding user-specified AQI thresholds. System architecture employs containerization enabling cloud deployment with automatic scaling managing computational load. Monitoring dashboards track prediction accuracy, data availability, and system performance ensuring operational reliability. The implementation provides template for national-scale air quality forecasting infrastructure supporting National Clean Air Programme goals^[13].

Conclusion

This comprehensive research developed and validated multivariate statistical techniques for air quality index prediction across five Indian metropolitan cities, establishing operational forecasting framework supporting evidence-based pollution control policy. The research analyzed three years of daily measurements encompassing six criteria pollutants and meteorological variables, generating 5,475 city-day observations for rigorous model development and evaluation. Systematic comparison of five approaches ranging from linear regression through artificial neural networks established machine learning superiority achieving RMSE of 12.3 AQI units and R² of 0.912.

Feature importance analysis identified PM_{2.5}, PM₁₀, and NO₂ as dominant predictors explaining 78.6% of AQI variance, validating particulate matter and nitrogen oxides as priority control targets. Meteorological variables contributed 14.1% importance through atmospheric dispersion and chemical transformation effects. Strong inter-pollutant correlations confirmed common combustion sources, while negative ozone correlations reflected distinct photochemical formation pathways requiring integrated multi-pollutant control strategies.

Temporal analysis revealed substantial day-to-day persistence and seasonal patterns with winter concentrations exceeding summer by 40-60%, attributed to reduced atmospheric mixing and increased biomass burning. Incorporating lagged variables improved forecast accuracy by 8-12%, enabling reliable next-day predictions supporting proactive public health interventions. However, extreme pollution episodes from episodic events require supplementary early warning systems integrating satellite remote sensing and numerical weather predictions.

The research establishes technical foundation for operational air quality forecasting infrastructure supporting National Clean Air Programme implementation. Daily AQI forecasts enable evidence-based pollution advisories, traffic restrictions, industrial emissions curtailment, and vulnerable population protection. Model interpretability facilitates policy communication explaining forecast drivers to stakeholders and public audiences. The modular computational framework enables rapid deployment across additional cities with minimal adaptation, supporting national-scale forecasting expansion.

Future research directions include: integration with chemical transport models providing mechanistic understanding complementing statistical patterns; incorporation of satellite remote sensing expanding spatial coverage beyond ground monitoring networks; development of ensemble forecasting combining multiple models improving reliability through uncertainty quantification; extension to longer forecast horizons enabling multi-day planning; and investigation of causal inference methods establishing pollution-health relationships supporting intervention impact evaluation. Scaling operational implementation requires: investment in

monitoring infrastructure ensuring data availability and quality; development of automated quality control procedures minimizing manual oversight; establishment of institutional frameworks integrating forecasts into emergency response protocols; and public engagement campaigns promoting forecast awareness and protective behavior adoption.

The outcomes demonstrate that multivariate statistical techniques provide accurate, interpretable, and operationally viable approach for air quality prediction in developing country contexts. The methodology addresses data limitations, computational constraints, and policy requirements characteristic of emerging economy environmental management. Successful deployment requires sustained commitment to monitoring infrastructure, analytical capacity development, and science-policy integration enabling evidence-based environmental governance supporting public health protection and sustainable urban development across India's rapidly growing metropolitan regions.

References

1. Balakrishnan K, Dey S, Gupta T, Dhaliwal RS, Brauer M, *et al.* The impact of air pollution on deaths. *Lancet*. 2019;394(10216):2489-2497.
2. Pandey A, Brauer M, Cropper ML, Balakrishnan K, Mathur P, *et al.* Health and economic impact of air pollution in India. *Lancet Planet Health*. 2021;5(1):e25-e33.
3. Gurjar BR, Jain A, Sharma A, Agarwal A, Gupta P, *et al.* Human health risks in megacities due to air pollution. *Atmos Environ*. 2010;44(36):4606-4613.
4. Central Pollution Control Board. National Air Quality Index. New Delhi (IN): Ministry of Environment, Forest and Climate Change; 2014.
5. Kumar P, Morawska L, Martani C, Biskos G, Neophytou M, *et al.* The rise of low-cost sensing for managing air pollution. *Environ Int*. 2015;75:199-205.
6. Sharma S, Chatani S, Mahtta R, Goel A, Kumar A. Sensitivity analysis of ground level ozone in India. *Atmos Environ*. 2016;131:29-40.
7. Stieb DM, Burnett RT, Smith-Doiron M, Brion O, Shin HH, *et al.* A new multipollutant index for air quality. *Environ Int*. 2008;34(8):1085-1092.
8. Perez P, Reyes J. An integrated neural network model for PM10 forecasting. *Atmos Environ*. 2006;40(16):2845-2851.
9. Biancofiore F, Busilacchio M, Verdecchia M, Tomassetti B, Aruffo E, *et al.* Recursive neural network model for analysis and forecast of PM10. *Atmos Environ*. 2017;155:62-73.
10. Jiang D, Zhang Y, Hu X, Zeng Y, Tan J, *et al.* Progress in developing an ANN model for air pollution index forecasting. *Atmos Environ*. 2004;38(40):7055-7064.
11. Jian L, Zhao Y, Zhu YP, Zhang MB, Bertolatti D. An application of ARIMA model to predict submicron particle concentrations. *Atmos Environ*. 2012;62:403-409.
12. Sayegh AS, Munir S, Habeebullah TM. Comparing the performance of statistical models for predicting PM10. *Aerosol Air Qual Res*. 2014;14(3):653-665.
13. Silibello C, D'Allura A, Finardi S, Bolignano A, Sozzi R. Application of bias adjustment techniques to improve air quality forecasts. *Atmos Pollut Res*. 2015;6(6):928-938.
14. Donnelly A, Misstear B, Broderick B. Real time air quality forecasting using integrated parametric and non-parametric regression techniques. *Atmos Environ*. 2015;103:53-65.
15. Arhami M, Kamali N, Rajabi MM. Predicting hourly air pollutant levels using artificial neural networks. *Environ Model Softw*. 2013;48:193-203.