Multivariate Time Series Analysis of Air Quality Data in India

Aleksandr Jan Smoliakov [1*]

Abstract
This project blueprint outlines an investigation into the application of multivariate time series models, inspired by techniques from financial econometrics, to forecast the air quality in Indian cities. Using hourly measurements of seven pollutant concentrations across 20 cities collected between 2015 and 2020, the study will aim to assess the temporal dynamics between pollutants. The discussed methodologies will form the foundation for subsequent steps involving data transformation, model estimation, and interpretation of forecasting results.

Keywords
air pollution — multivariate time series

[1] *Department of Econometric Analysis, Faculty of Mathematics and Informatics, Vilnius University*
Corresponding author: aleksandr.smoliakov@mif.stud.vu.lt

Contents

Introduction

Urban air quality is a critical public health and environmental issue, especially in rapidly urbanizing regions. Accurate forecasting of air quality indicators, such as pollutant (e.g., PM2.5, PM10, NO2, CO) concentrations and their composite measures like the Air Quality Index (AQI), is essential for formulating timely policy interventions.

While univariate models (e.g., ARIMA) offer a baseline for prediction, complex datasets with interconnected variables require the advanced analytical capabilities of multivariate time series models. In financial econometrics, multivariate approaches – such as vector autoregression (VAR) and vector autoregressive moving average (VARMA) models – have been used to model the interdependencies among multiple time series, capturing interactions between variables and their lagged effects. These methodologies could be equally applicable to environmental datasets where pollutant concentrations interact over time.

In this project, I propose to analyze the *Air Quality Data in India (2015–2020)* dataset, which contains hourly observations of seven pollutant concentrations (as well as the Air Quality Index, AQI) measured in 20 Indian cities.

The objective of this project is to apply multivariate models to capture the dynamic interactions among pollutants (possibly incorporating spatial dependencies) and provide robust air quality forecasts for Indian cities.

1. Literature Review

Recent studies on air quality forecasting have adopted various time series modeling techniques, ranging from univariate to multivariate approaches. This section reviews relevant contributions in the field, focusing on the strengths and limitations of different methodologies.

Analysis of Air Quality using Univariate and Multivariate Time Series Models

Sethi and Mittal (2020) investigate the prediction of daily Air Quality Index (AQI) values in Gurugram, India by comparing univariate and multivariate time series methods [3]. Using ARIMA and VAR models, they demonstrate that, for their dataset, a simpler univariate ARIMA model gives more accurate predictions of AQI than the multivariate counterpart. The authors conclude that multivariate VAR models faced instability in handling the additional noise introduced by interdependent pollutant series. This paper provides valuable insights on the challenges of multivariate data modeling, especially the impact of differencing and parameter selection on model stability.

Multivariate Time Series Modelling for Urban Air Quality

Hajmohammadi and Heydecker (2021) focus on urban air quality forecasting in multiple meteorological stations in London. The authors compare seasonal ARMA (SARMA) models with vector ARMA (VARMA) models, which account for both spatial and temporal dependencies. The study demonstrates that while SARMA models capture individual station trends, VARMA outperforms it by accounting for cross-station interactions. The findings emphasize that in complex urban environments, accounting for spatial interactions can significantly enhance predictive performance. This concept could be applicable for modeling pollutant interdependencies in the

larger-scale Indian context.

Forecasting of Particulate Matter with a Hybrid ARIMA Model Based on Wavelet Transformation and Seasonal Adjustment

Aladağ (2021) introduces a hybrid modeling approach that integrates seasonal adjustment, wavelet transformation, and ARIMA [1]. Applied to forecasting PM10 in Erzurum, Turkey, this method combines a classical ARIMA model with wavelet decomposition, capturing both low-frequency trends and high-frequency variations present in the pollutant time series. This study provides a method for addressing seasonality and non-stationarity – features also present in air quality data – by decomposing the series before reassembling the forecasted components.

The insights from these studies suggest the use of multivariate frameworks in forecasting. Techniques from financial econometrics are shown to be effective when adapted to environmental datasets. Advanced data transformation methods, such as wavelet decomposition, can also enhance model performance. The project will thus build on these methodologies to analyze the air quality data in India.

References

[1] Aladağ, E. (2021). *Forecasting of Particulate Matter with a Hybrid ARIMA Model Based on Wavelet Transformation and Seasonal Adjustment*. Urban Climate. `https://doi.org/10.1016/j.uclim.2021.100930`.

[2] Hajmohammadi, H. and Heydecker, B. (2021). *Multivariate time series modelling for urban air quality*. Urban Climate. `https://doi.org/10.1016/j.uclim.2021.100834`.

[3] Sethi, J.K. and Mittal, M. (2020). *Analysis of Air Quality using Univariate and Multivariate Time Series Models*. `https://doi.org/10.1109/Confluence47617.2020.9058303`.