



Multivariate Time Series Analysis of Air Quality Data in Delhi

Aleksandr Jan Smoliakov¹

¹Vilnius University, Faculty of Mathematics and Informatics

2025-05-27

Table of Contents

Introduction: The Air Quality Challenge

- Urban air quality is a critical public health and environmental issue, especially in rapidly urbanizing regions like Delhi.
- Accurate forecasting of pollutants (e.g., $PM_{2.5}$, PM_{10} , NO_2 , CO) is essential for timely policy interventions.
- Univariate models (e.g., ARIMA) offer a baseline but may not capture complex interdependencies.
- Multivariate time series models (e.g., VAR, VARMA), prominent in financial econometrics, can model interactions between multiple pollutant series.

Focus of this Project

Analyze air quality in Delhi (2018-2019) using daily data for five key pollutants: $PM_{2.5}$, PM_{10} , NO_2 , CO, and NH_3 .

Project Objectives

- To apply multivariate time series models (VAR and VARMA) to understand the dynamic interactions among key air pollutants in Delhi.
- To generate forecasts for these pollutant concentrations.
- Key steps involved:
 - Data preprocessing and Exploratory Data Analysis (EDA).
 - Stationarity testing.
 - VAR and VARMA model estimation and diagnostics.
 - Granger causality analysis.
 - Impulse Response Function (IRF) analysis.
 - Forecast Error Variance Decomposition (FEVD).
 - Forecast evaluation.

Brief Literature Review

- **Sethi and Mittal (2020):** Compared ARIMA and VAR for AQI in Gurugram. Found ARIMA more accurate, highlighting challenges with VAR stability and noise in interdependent series.
- **Hajmohammadi and Heydecker (2021):** Compared SARMA and VARMA for London air quality. VARMA outperformed by capturing cross-station interactions, emphasizing spatial dependency benefits.
- **Aladağ (2021):** Proposed a hybrid ARIMA model with wavelet transformation and seasonal adjustment for PM10 forecasting, addressing nonstationarity and seasonality.

Key Takeaway

Multivariate frameworks are promising. Adapting techniques from fields like finance and using advanced data transformations can enhance model performance for environmental data.

Data Source and Preparation

- **Dataset:** *Air Quality Data in India (2015–2020)*.
- **Focus:** Delhi, Jan 1, 2018 – Jan 1, 2020 (732 daily observations).
- **Pollutants:** $PM_{2.5}$, PM_{10} , NO_2 , CO , NH_3 .
- **Reasons for Delhi focus:**
 - One of the world's most polluted cities.
 - Relatively complete data ($< 1\%$ missing values for the selected period).
- **Missing Value Imputation:** Linear interpolation (`na.interp`).
- **Data Transformation:** $\log(x + 1)$ (`log1p`) to stabilize variance and normalize distributions.
- **Data Aggregation:** Hourly data aggregated to daily means (of `log1p`-transformed values) to reduce noise.

Exploratory Data Analysis (EDA)

- Distributions after transformation:

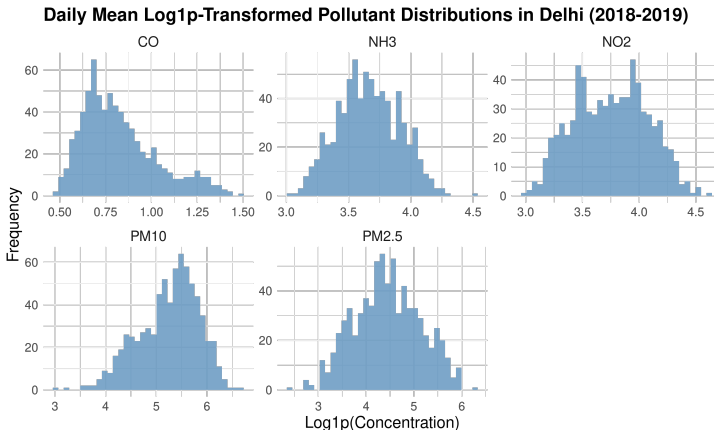
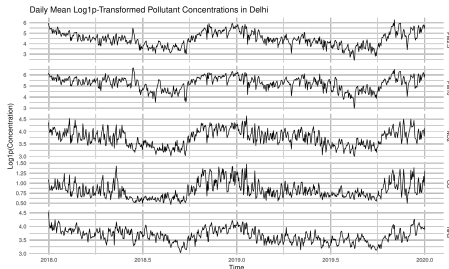


Figure: Histograms of Daily Mean log1p-Transformed Pollutants (Delhi, 2018-2019).

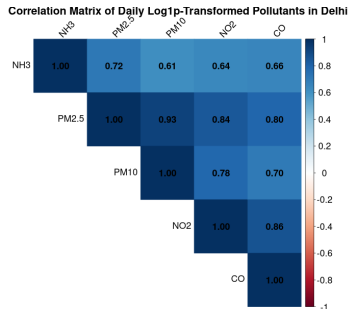
EDA: Time Series and Correlations

Time Series Behavior:



Daily Log1p-Transformed Pollutants.

Correlation Matrix:



Correlations (Log1p-Transformed).

Note: Strong interdependencies are evident, motivating multivariate analysis.

Stationarity Testing Model Choices

- **Stationarity Testing:** Augmented Dickey-Fuller (ADF) test on log1p-transformed daily series.

Pollutant	Test Statistic	Stationary ($p < 0.01$)
$PM_{2.5}$	-5.874	TRUE
PM_{10}	-7.268	TRUE
NO_2	-7.986	TRUE
CO	-8.681	TRUE
NH_3	-7.559	TRUE

→ *All series are I(0), allowing direct VAR/VARMA application.*

- **Vector Autoregression (VAR) Model:**

$$Y_t = c + A_1 Y_{t-1} + \dots + A_p Y_{t-p} + \epsilon_t$$

Optimal lag p via AIC (VARselect). Estimated via vars package.

- **Vector Autoregressive Moving Average (VARMA) Model:**

$$Y_t = A_1 Y_{t-1} + \dots + A_p Y_{t-p} + B_1 \epsilon_{t-1} + \dots + B_q \epsilon_{t-q} + \epsilon_t$$

Estimated via MTS package. VARMA(1,1) explored.

VAR(4) Model Analysis: Lag Selection Causality

- **Lag Order Selection for VAR:**

- AIC and FPE suggested $p = 4$. HQ suggested $p = 2$, SC suggested $p = 1$.
- **VAR(4) model selected** based on AIC.
- Model stable (all roots of characteristic polynomial < 1). High R^2 values (0.70-0.84).

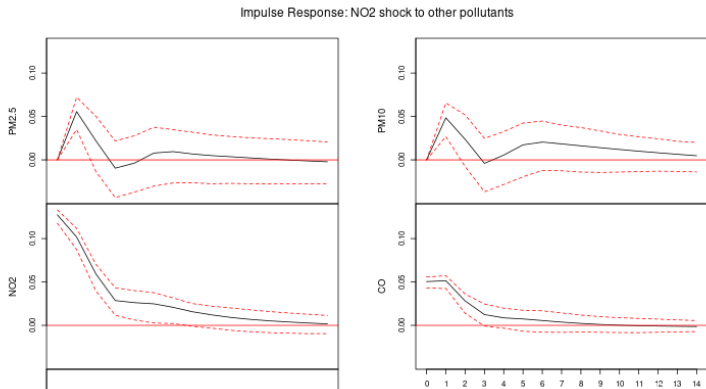
- **Granger Causality (from VAR(4) model):**

Causality Direction	p -value
$PM_{2.5} \rightarrow \text{Others}$	4.80×10^{-5} ***
$PM_{10} \rightarrow \text{Others}$	9.81×10^{-4} ***
$NO_2 \rightarrow \text{Others}$	0.0327 *
$CO \rightarrow \text{Others}$	0.174
$NH_3 \rightarrow \text{Others}$	8.56×10^{-7} ***

→ *Significant predictive relationships, especially from $PM_{2.5}$ and NH_3 .*

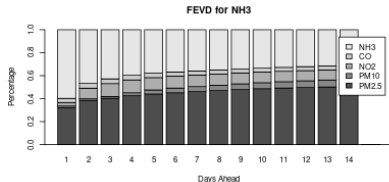
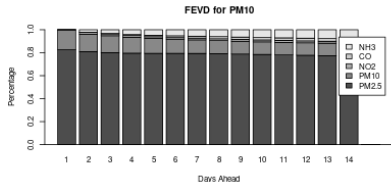
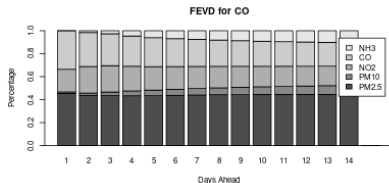
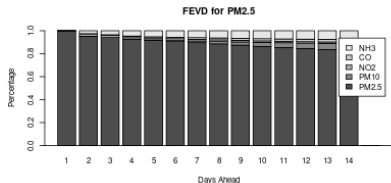
VAR(4) Analysis: Impulse Response Functions (IRFs)

- IRFs trace the effect of a one-standard-deviation shock in one variable on others.
- Example: Response of other pollutants to a shock in NO_2 .



VAR(4) Analysis: Forecast Error Variance Decomposition (FEVD)

- FEVD shows the proportion of forecast error variance of each variable attributable to its own shocks versus shocks from other variables.



Forecasting Evaluation: VAR(4) vs. VARMA(1,1)

- **Setup:**

- Data split: Training (first 718 days), Test (last 14 days).
- Forecast horizon: 14 days ahead.
- Metric: Root Mean Squared Error (RMSE) on log1p-transformed scale.

- **RMSE Comparison on Test Set:**

Model	$PM_{2.5}$	PM_{10}	NO_2	CO	NH_3
VAR(4)	0.721	0.543	0.179	0.202	0.150
VARMA(1,1)	0.605	0.446	0.169	0.189	0.142

→

VARMA(1,1) showed lower RMSE for all five pollutants, suggesting better performance.

Example: VAR(4) Forecasts for Delhi Pollutants

14 -Day Ahead Forecasts for Delhi Pollutants (from VAR on full data)

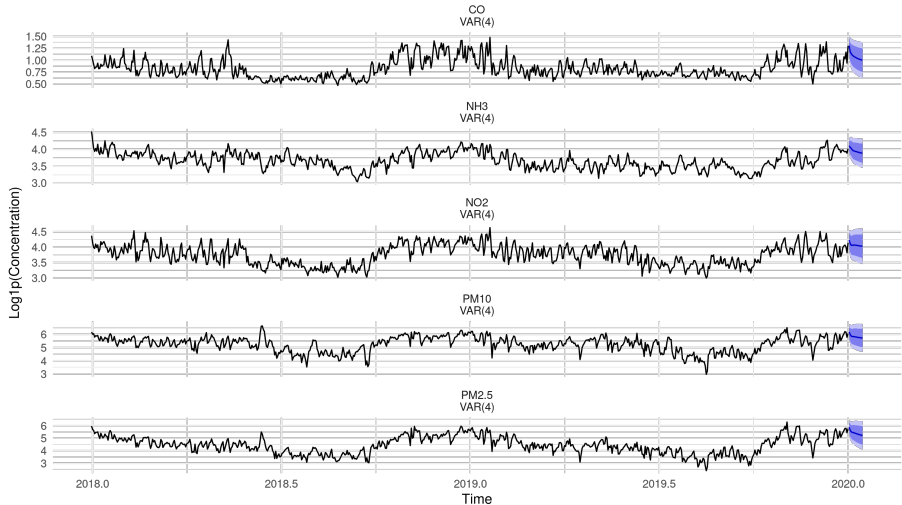


Figure: 14-Day Ahead Forecasts from VAR(4) Model (Trained on Full

Conclusion

- Successfully applied VAR and VARMA models to analyze multivariate dynamics of 5 key air pollutants in Delhi (2018-2019).
- Daily log1p-transformed pollutant series were found to be stationary $I(0)$.
- VAR(4) model revealed:
 - Significant Granger causalities (e.g., $PM_{2.5}$, NH_3 influencing others).
 - Dynamic interactions via IRFs (e.g., NO_2 shocks affect other pollutants).
 - FEVD showed importance of own shocks and $PM_{2.5}$ in forecast error variance.
- For 14-day ahead forecasting, VARMA(1,1) outperformed VAR(4) in terms of RMSE.
- The study highlights the interconnected nature of air pollution and the utility of multivariate models.

Limitations and Future Work







Limitations

- Focus on a single city (Delhi).
- Limited set of pollutants.
- Use of linear models (VAR/VARMA).
- Daily aggregation might mask hourly dynamics.
- VARMA order selection was illustrative (1,1), not exhaustive.






Future Work

- Extend to multiple cities (spatial dependencies).
- Incorporate exogenous variables (e.g., meteorological data).
- Explore non-linear multivariate models.
- Investigate seasonal decomposition methods prior to modeling.

References I

-  Aladağ, E. (2021). *Forecasting of Particulate Matter with a Hybrid ARIMA Model Based on Wavelet Transformation and Seasonal Adjustment*. Urban Climate.
-  Hajmohammadi, H. and Heydecker, B. (2021). *Multivariate time series modelling for urban air quality*. Urban Climate.
-  Sethi, J.K. and Mittal, M. (2020). *Analysis of Air Quality using Univariate and Multivariate Time Series Models*. IEEE Confluence.
-  R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
-  Wickham, H., et al. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.
-  Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

References II

-  Wickham, H., et al. (2024). *tidyr: Tidy Messy Data*. R package version 1.3.1.
-  Hyndman, R., et al. (2025). *forecast: Forecasting functions for time series and linear models*. R package version 8.24.0.
-  Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*. Second Edition. Springer, New York.
-  Pfaff, B. (2008). *VAR, SVAR and SVEC Models: Implementation Within R Package vars*. Journal of Statistical Software.
-  Tsay, R. S., et al. (2022). *MTS: All-Purpose Toolkit for Analyzing Multivariate Time Series (MTS) and Estimating Multivariate Volatility Models*. R package version 1.2.1.

Thank You!

Thank you for your attention!