Multivariate Time Series Analysis of Air Quality Data in Delhi

Aleksandr Jan Smoliakov[1]

[1]Vilnius University, Faculty of Mathematics and Informatics

2025–05–27

Table of Contents

1. Introduction
2. Methodology
3. Results and Discussion
4. Conclusion & Limitations

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) may not capture complex interdependencies.
- Multivariate time series models (e.g. VAR, VARMA) 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

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

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.

- **Distributions after transformation:**

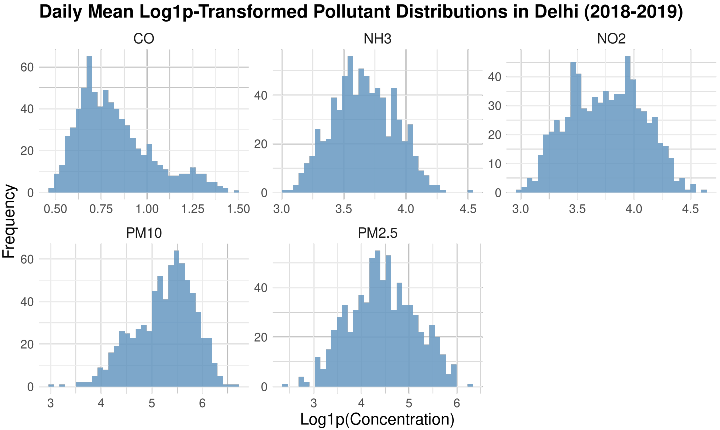
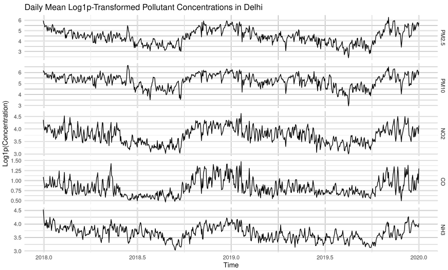


Figure: Histograms of Daily Mean log1p-Transformed Pollutants (Delhi, 2018–2019).

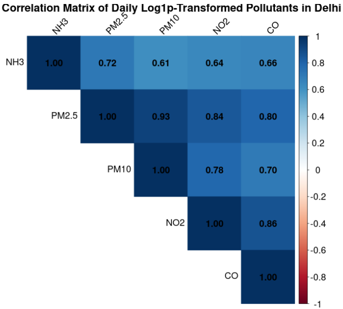
EDA: Time Series and Correlations

Time Series Behavior:



Daily Log1p-Transformed Pollutants.

Correlation Matrix:



Correlations (Log1p-Transformed).

Note: Strong correlations are evident, there may be multivariate dependencies.

Stationarity Testing & Model Choices

- **Stationarity Testing:** Augmented Dickey-Fuller (ADF) test on log1p-transformed daily series.
  Conclusion: all log1p-transformed series are stationary (I(0)) with $p < 0.01$, allowing for VAR/VARMA modeling.
- **Vector Autoregression (VAR) Model:**

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

  Optimal lag p via AIC (`vars::VARselect`).
- **Vector Autoregressive Moving Average (VARMA) Model:**

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

  Only VARMA(1,1) explored, larger orders did not converge.

VAR(4) Model Analysis: Lag Selection & Causality

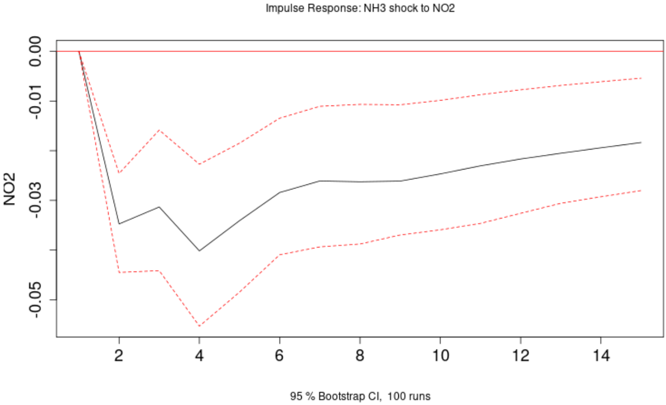
- **Lag Order Selection for VAR:**
  - VAR(4) model selected as suggested by AIC.
  - Model stable (all roots of characteristic polynomial < 1).
- **Granger Causality (from VAR(4) model):**

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

Observation: Significant predictive relationships, especially from $PM_{2.5}$ and NH_3.

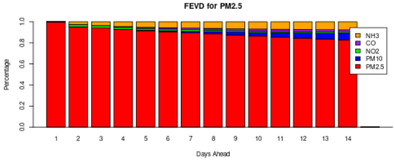
VAR(4) Analysis: Impulse Response Functions

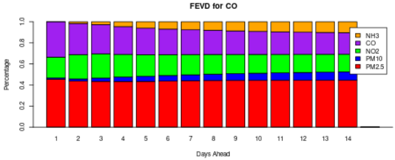
- IRFs trace the effect of a one-standard-deviation shock in one variable on others.
- Shown below: Response of NO_2 to a shock in NH_3.

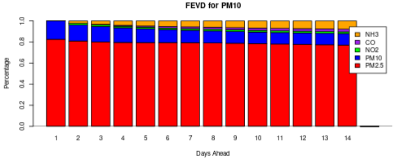


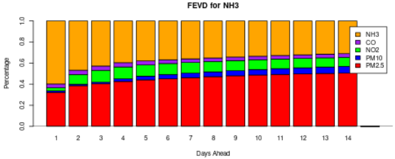
VAR(4): Forecast Error Variance Decomposition

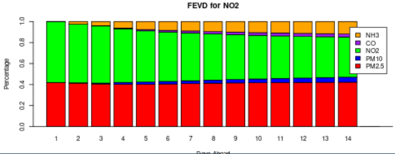
- FEVD shows the proportion of forecast error variance of each variable attributable to shocks to itself versus other variables.











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

- **Setup:**
  - Data split: Training (first 718 days), Test (last 14 days).
  - Forecast horizon: 14 days ahead.
- **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**

Observation: VARMA(1,1) showed lower RMSE for all five pollutants, suggesting better forecast accuracy for this dataset and horizon.

Example: VAR(4) Forecasts for Delhi Pollutants

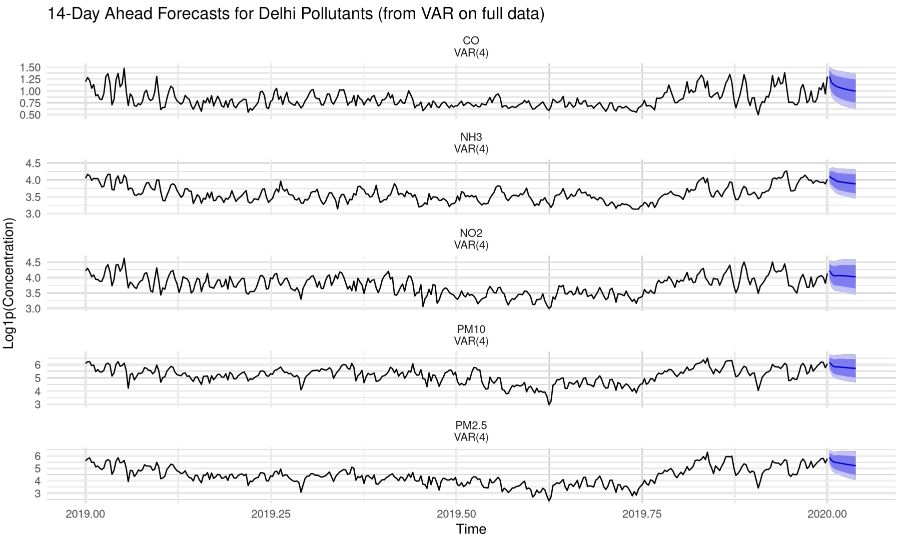


Figure: 14-Day Ahead Forecasts from VAR(4) Model.

Conclusion

- Successfully applied VAR and VARMA models to analyze multivariate dynamics of 5 key air pollutants in Delhi.
- 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 potential of multivariate time series models for air quality forecasting and understanding pollutant interactions.

Limitations

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

Thank You!

Thank you for your attention!