DS 299 Capstone in Data Science
Project Report

AIR POLLUTION DYNAMICS IN YEREVAN

May 9, 2025

Ina Karapetyan and Elen Galoyan
BS in Data Science, AUA
Supervisor: Houry Keoshkerian

Contents

1 Abstract

Air pollution is one of the most pressing environmental threats in Armenia, particularly in urban areas such as Yerevan. The rising levels of harmful pollutants, such as PM2.5, pose significant risks to public health, contributing to increased mortality rates and long-term health conditions. This study applied time series analysis using SARIMAX and AI models, including XGBoost and LSTM, to predict PM2.5 levels in Yerevan. The performances of multiple models were compared to choose the most effective method for trend estimation and forecasting. Although much research exists on pollution causes, limitations such as missing time-series pollutant data and inaccessible health records constrained our analysis of health outcomes. Using real-time sensor data to detect pollution trends and predict possible effects, this project aims to inform the public and policymakers about the critical need for immediate and effective actions, along with attempting to increase awareness of the increasing dangers caused by air pollution.

Keywords: air quality; time series analysis; forecasting models; prediction of pollutants' levels

2 Introduction

2.1 Background and Significance

Air pollution, particularly the presence of fine particulate matter (PM2.5), is a growing environmental and public health concern in Armenia. Urban areas like Yerevan are especially vulnerable due to dense traffic, rapid urban development, and limited regulatory enforcement. [1] PM2.5 particles, which are less than 2.5 micrometers in diameter, can penetrate deep into the lungs and bloodstream, contributing to the increased mortality rates among exposed populations. Ambient air pollution also poses a significant threat [2] The ever-growing population of cities, rapid urbanization, and increased levels of motorization contribute to increasing emissions. The estimated annual number of deaths of all causes in the world as a result of air pollution from fine particles and ozone is **8.34 million** (95% confidence interval: **5.63 to 11.19 million**). Cardiometabolic conditions contribute to the majority (**52%**) of the mortality burden, with ischaemic heart disease accounting for *30%*. **16%** of the mortality load is attributed to stroke and chronic obstructive pulmonary disease. Approximately **20%** of all-cause mortality is unknown, and *neurodegenerative diseases* and *arterial hypertension* may be involved [3]. These alarming statistics demand urgent action not only from policymakers but also from the scientific community, whose tools and insights can support effective decision-making.

2.2 Research Motivation

While European Union programs and strategic plans offer a path toward long-term air quality improvement, the results of these actions often take years to materialize. This lag highlights the need for adaptive tools, such as air quality forecasting systems, that can help bridge the gap between policy implementation and real-world outcomes.

The deterioration of air quality in Yerevan in recent years underscores the urgent need for data-driven research to understand and mitigate the impact of $PM_{2.5}$ pollution. Yet, despite widespread public concern and discussion—as seen in recent studies such as *Air Quality in Yerevan in the Context of Climate Change* [4]—there remains a lack of concrete, time series-based analysis specifically focused on $PM_{2.5}$. Our review found no recent forecasting studies addressing this pollutant in Yerevan.

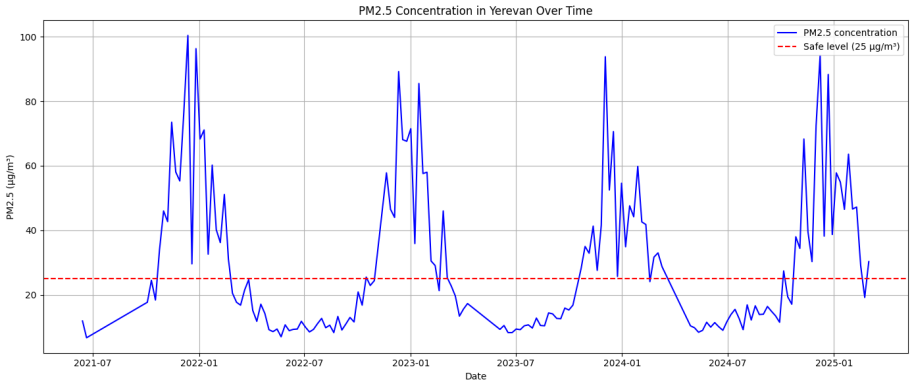


Figure 1: $PM_{2.5}$ levels in Yerevan exceeding the safe threshold.

The time series plot indicates that $PM_{2.5}$ concentrations in Yerevan consistently exceed the WHO's recommended threshold of 25 $\mu g/m^3$, indicating an ongoing issue with air pollution in multiple regions. [5] This persistent exceedance poses a health risk, particularly for vulnerable groups such as children, the elderly, and individuals with respiratory or cardiovascular conditions.

To address this gap, we collected weekly $PM_{2.5}$ concentration data spanning 2019 to 2025 from publicly available sources, specifically the PurpleAir API and AQI.org. This dataset enables us to construct a robust time series model for predicting air quality trends in 2026. The period and granularity offer a valuable opportunity to capture both seasonal patterns and long-term shifts.

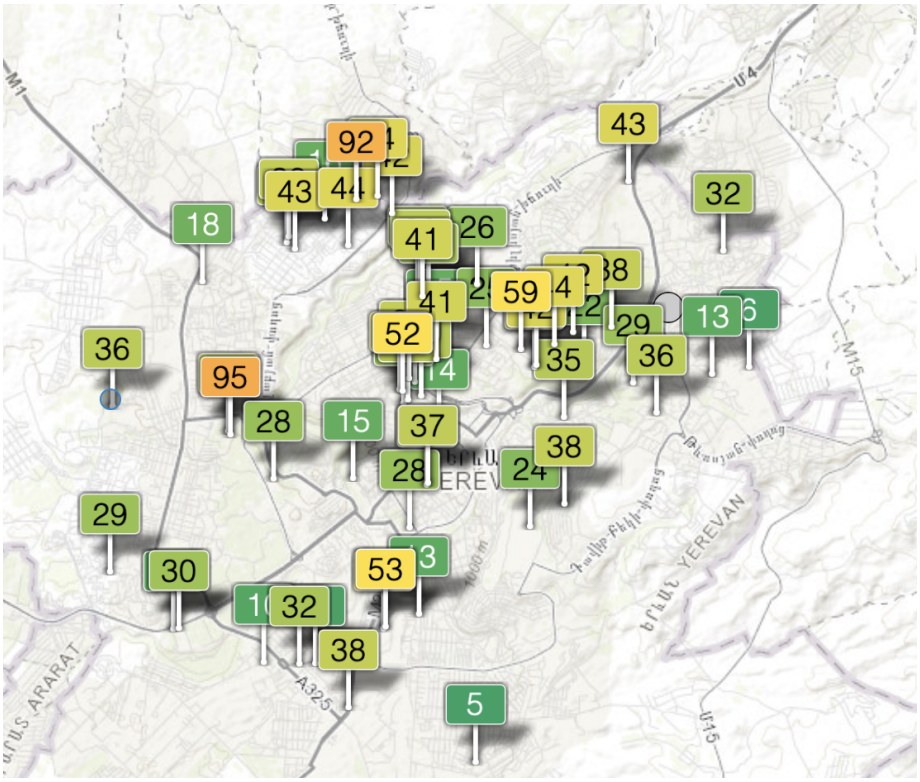


Figure 2: Sensors for $PM_{2.5}$ in Yerevan

By combining advanced forecasting methods with sensor data from various Yerevan districts, our project seeks to close this gap. We concentrate on using artificial intelligence algorithms to forecast $PM_{2.5}$ concentrations and a Monte Carlo simulation to estimate the associated mortality risks. Additionally, we were motivated by a time series study carried out in Sofia [**?**], where similar pollution problems are being addressed using techniques that support our approach. This cross-city comparability supports the findings' regional significance and shows how adaptable our approach is.

Although our data constraints prevented us from modeling additional pollutants, the same analytical framework can be extended to study other harmful substances in the future. Ultimately, our goal is to provide timely, evidence-based insights that both inform policy development and raise public awareness around air quality and health in Armenia.

2.3 Objectives and Contributions

This study aims to:

- Forecast $PM_{2.5}$ levels in Yerevan using real-time sensor data and AI-based predictive models;

- Estimate the potential impact of air pollution on mortality rates using Monte Carlo simulations;
- Offer data-driven insights to support environmental health interventions and policy recommendations.

3 Literature Review

Air pollution forecasting is a vital research area due to its impact on public health and urban sustainability. A variety of methods have been employed to model pollutant behavior, ranging from traditional time series models like *ARIMA* and *SARIMA* to more advanced machine learning techniques. While classical models are effective at capturing linear patterns, they often fall short in modeling nonlinear relationships and accounting for external influences. To overcome these limitations, machine learning algorithms such as *Random Forest*, *Support Vector Regression (SVR)*, and deep learning architectures like *Recurrent Neural Networks (RNNs)* and *Long Short-Term Memory (LSTM)* networks have gained traction. These approaches offer improved accuracy by incorporating complex temporal dynamics and exogenous variables such as meteorological and traffic data.

Marinov et al. (2023) [6] conducted a time series forecasting study on air pollution in Sofia, Bulgaria, focusing on four key pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and fine particulate matter ($PM_{2.5}$). The study employed the *AutoRegressive Integrated Moving Average (ARIMA)* model, utilizing five years of data (2015–2019) from multiple monitoring stations and evaluating forecasts across various temporal granularities (3h, 6h, 12h, 24h). Although ARIMA effectively identified temporal trends and facilitated short-term forecasts, its reliance on historical pollutant values was noted as a limitation. The authors recommended the incorporation of *exogenous variables*, such as temperature, humidity, and wind speed, as well as the use of *machine learning models* to enhance predictive performance.

Among the various machine learning models applied to air quality prediction, *Long Short-Term Memory (LSTM)* networks have demonstrated outstanding performance in modeling complex time series data such as $PM_{2.5}$ concentrations [7]. As an advanced variant of *Recurrent Neural Networks (RNNs)*, LSTMs are specifically designed to mitigate the vanishing gradient problem through a gated architecture—comprising forget, input, and output gates—that enables selective retention of relevant information across long temporal sequences. This architecture makes LSTMs particularly effective at capturing long-range dependencies and nonlinear patterns that are characteristic of environmental datasets [8].

Recent advancements have further improved the predictive capabilities of LSTMs. Notably, the integration of *attention mechanisms* allows the model to dynamically focus on the most relevant features at each time step, while the inclusion of *meteorological variables* (e.g., temperature, humidity, wind speed) enhances the model's contextual awareness and accuracy [9]. Given the increasing complexity, volume, and multidimensional nature of

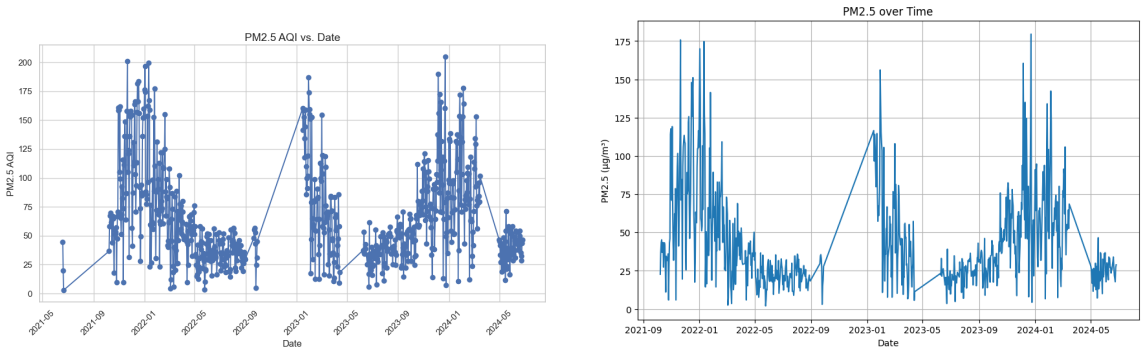
air quality data, LSTM networks remain among the most powerful and adaptable tools for forecasting fine particulate matter concentrations in urban environments. **Liu et al. (2018)** [10] extended the capabilities of time series modeling through regional numerical approaches in Hong Kong. Using stochastic time series methods, including ARIMA, the study successfully forecasted air pollutant concentrations and highlighted the model's strength in capturing short-term temporal dynamics. Similarly, in Beijing, ARIMA was applied to predict $PM_{2.5}$ concentrations across 35 monitoring stations within a 24-hour window. A *sliding window technique* was employed to address the constraints of data continuity and volume, expanding the training set to millions of observations and enabling more effective feature extraction and generalization.

In addition to these projects, other research carried out a correlation analysis between other environmental factors and pollutants. Through the clarification of the relationships between pollution indicators and outside variables, these studies attempted to improve the interpretability of prediction models. A version of ARIMA that takes seasonality into account, the *Seasonal ARIMA (SARIMA)* model was used to estimate yearly $PM_{2.5}$ levels in order to further improve predictions [11]. SARIMA gave predicted values for both the minimum and maximum, giving an in-depth understanding of long-term changes in air quality. Particularly in urban areas that are quickly industrializing, these models are crucial for environmental policy-making and public health planning.

4 Data Collection and Preprocessing

4.1 Air Quality Data Sources

The primary air quality dataset was retrieved using API keys from the **PurpleAir** platform [12]. Specifically, we accessed real-time pollution data from the **TUMO Center for Creative Technologies** sensor in Yerevan, collecting one data point per day from 2021 to March 2025. To enhance spatial and temporal coverage, data from an additional sensor in the Kentron district of Yerevan was obtained via the **AQICN.org** platform [13], spanning daily data from 2019 to 2025.



(a) AQI data distribution

(b) PurpleAir $PM_{2.5}$ distribution

Figure 3: Comparison of AQI and PurpleAir $PM_{2.5}$ distribution

4.2 Raw Data Structure and Merging

The Kentron dataset included daily metrics such as *minimum*, *maximum*, *median*, first and third quartiles (Q1, Q3), *standard deviation*, and *count of observations*. In contrast, the PurpleAir dataset—collected from the TUMO sensor—contained additional meteorological variables such as `humidity`, `temperature`, and `pressure`, along with $PM_{2.5}$ and PM_{10}. For consistency and analytical focus, we primarily used the $PM_{2.5}$ values from both datasets.

To create a more robust time series, we merged the Kentron and PurpleAir datasets by matching timestamps. After alignment, we computed the average $PM_{2.5}$ value across the two sources every day. All original columns, including the meteorological features from PurpleAir, were kept in the merged dataset for subsequent analysis and modeling. Additionally, AQI did not have exact $PM_{2.5}$ values but provided the median. Given the right-skewed distribution of the data, we decided to use the median for consistency in our analysis.

Table 1: Quality of the raw dataset for PurpleAir

Variable	Min	Mean	Max	Missing (%)	Negatives (%)
Humidity (%)	7.13	40.20	70.40	0.0	0.0
Temperature	18.50	61.86	108.33	0.0	0.0
Pressure	858.83	897.08	913.50	0.0	0.0
$PM_{2.5}$	0.20	29.20	230.80	0.0	0.0
PM_{10}	0.00	53.13	423.98	0.0	0.0

4.3 Meteorological Variables and Correlation

As mentioned, the PurpleAir dataset provided meteorological variables such as `humidity`, `temperature`, and `pressure`, which were preserved during the merging process. Correlation analysis on the merged dataset revealed statistically significant associations between $PM_{2.5}$ and humidity and temperature. Given their influence on air quality dynamics, we incorporated these two variables as exogenous regressors in a **Seasonal ARIMA (SARIMA)** model using the SARIMAX framework. This allowed for improved forecasting accuracy by accounting for environmental drivers of pollution variability.

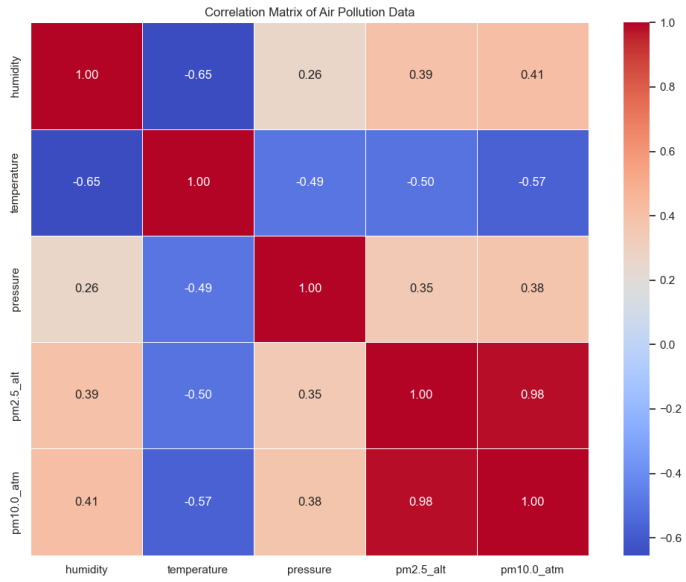


Figure 4: Correlation Matrix

4.4 Data Granularity

We have two use cases for the data granularity: one where the data points are averaged per week, and another where the data is averaged daily. For the LSTM model, the higher granularity (daily averaged data) worked better, leading to smaller error results. In contrast, for the time series analysis, we used both the weekly and daily datasets and performed comparisons based on these different granularities.

5 Data Simulation for mortality rates

Since there is no direct data on the medical conditions associated with air pollution in Yerevan, simulation-based methods provide a suitable alternative for calculating possible

health effects. This section presents a probabilistic technique for modelling mortality rates related to exposure to $PM_{2.5}$. We want to quantify the uncertainty and unpredictability inherent in pollution-related health consequences by applying Monte Carlo techniques and using acknowledged risk coefficients from global studies on health. Also, through modifying the risk factors, we consider several hypothetical scenarios, which enables us to assess how changes in population vulnerability or pollutant sensitivity can affect anticipated death rates.

To estimate potential health outcomes without direct medical data, we conducted both *deterministic* and *Monte Carlo simulations*. Baseline mortality rates and risk coefficients were sourced from the **World Health Organization (WHO)**. The $PM_{2.5}$ probability distribution was visualized and found to resemble a positively skewed distribution.

5.1 Distribution Testing and Noise Injection

5.1.1 Simulation

To estimate the potential health impact of PM2.5 exposure, we adopted two approaches to model daily mortality in Yerevan: a deterministic baseline estimation and a stochastic Monte Carlo simulation based on a lognormal relative risk distribution.

Deterministic Mortality Estimation. We began with a WHO-aligned baseline mortality rate of 26 deaths per day in Yerevan. The health risk from PM2.5 was assumed to increase linearly by 4% per 10 micrograms per cubic meter above a safe threshold of 25 micrograms per cubic meter, as suggested by epidemiological studies. For each observed daily PM2.5 concentration, the expected excess mortality was computed by scaling the baseline using this risk increase.

To reflect statistical uncertainty, we included a confidence interval around each mortality estimate. We assumed a fixed 5% margin of variability to represent possible differences in measurement accuracy and how populations might respond to pollution exposure. This allowed us to construct a normal-based 95% confidence interval using a standard error calculated as $\sigma = 0.05 \times$ expected mortality, and applying the 1.96 multiplier to define the bounds. This deterministic approach assumes a fixed linear risk and does not reflect variability in the relative risk estimates or underlying environmental randomness.

Stochastic Monte Carlo Simulation. To introduce a more realistic degree of uncertainty, we performed a Monte Carlo simulation in which the relative risk (RR) of mortality per 10 micrograms per cubic meter increase in $PM_{2.5}$ was treated as a lognormally distributed random variable. This was based on an average relative risk of $RR_{mean} = 1.04$, with the standard deviation defined as 5% of the log-transformed mean, i.e., $\sigma = 0.05 \times \log(RR_{mean})$, reflecting uncertainty in the epidemiological risk estimates.

For each observed $PM_{2.5}$ concentration, we generated $n = 1000$ samples from the lognormal distribution defined by $\mu = \log(RR_{mean})$ and $\sigma = 0.05 \cdot \mu$. The excess risk associated with a given day's pollution level was computed using the following transformation:

$$\text{Excess Risk} = (RR_{sample} - 1) \cdot \frac{PM_{2.5}^{obs} - PM_{2.5}^{safe}}{10}$$

Subsequently, simulated mortality was computed for each RR sample as:

$$\text{Mortality}_{sim} = \text{Baseline Mortality} \cdot (1 + \text{Excess Risk})$$

We constrained the mortality estimates to non-negative values using $\max(0, \text{Mortality}_{sim})$. From these simulations, we calculated the mean and the 2.5^{th} and 97.5^{th} percentiles to form a non-parametric 95% confidence interval for daily mortality.

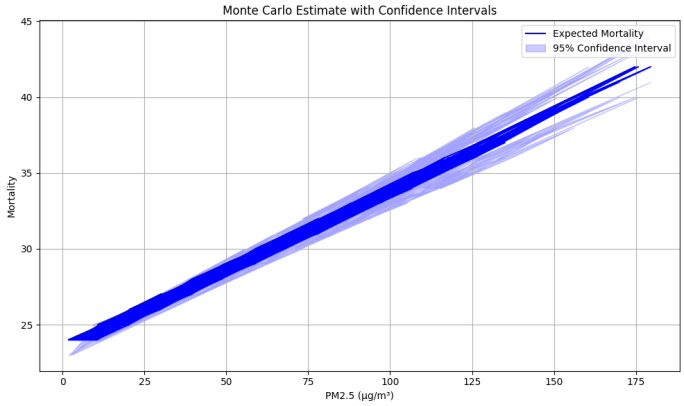


Figure 5: Monte Carlo simulation of mortality rate with 95% confidence interval shaded in gray.

This simulation method provides a more nuanced and probabilistic estimate of mortality by integrating both observed pollution levels and the uncertainty in the exposure-response relationship. Compared to the deterministic method, it yields wider confidence intervals, better capturing real-world uncertainty in both air quality and health effect estimation.

5.1.2 Hypothetical scenarios

To evaluate the health impacts of variable levels of air pollution, we simulated changes in daily mortality under hypothetical scenarios by modifying the risk coefficient associated with $PM_{2.5}$ exposure. We used a Monte Carlo simulation approach to estimate daily mortality across a range of adjusted risk increases, including $\pm 10\%$ and $\pm 20\%$ changes relative to a baseline risk increase of 0.04 (i.e., a 0.04% increase in mortality per 10 $\mu g/m^3$ of $PM_{2.5}$). We kept constant $PM_{2.5}$ concentrations constant and varied only the `risk_increase_per_10ug`

parameter to assess the sensitivity of mortality outcomes to uncertainties in the exposure-response relationship.

The relative risk of mortality per 10 μg/m^3 increase in PM2.5, denoted as `risk_increase_per_10ug`, was adjusted for each scenario using the formula:

$$\texttt{modified\_risk} = \texttt{risk\_increase\_per\_10ug} \times (1 + \texttt{change}) \tag{1}$$

where `change` represents the relative percentage change in PM2.5 (e.g., -0.1 for a 10% reduction).

For each value of PM2.5 observed in the dataset, we calculated the expected mortality using:

$$\texttt{expected\_mortality} = \texttt{baseline\_mortality} \times \left(1 + \texttt{modified\_risk} \times \frac{(\text{PM}_{2.5} - \texttt{safe\_PM}_{2.5})}{10}\right) \tag{2}$$

This expected mortality was used as the mean in a lognormal distribution to simulate daily mortality over `n_simulations` iterations. Negative values were excluded by clipping at zero.

The plot presents the resulting mean daily mortality under each hypothetical scenario, with confidence intervals 95% based on the simulated distributions. As expected, mortality rates decreased under PM2.5 reduction scenarios and increased under elevated exposure conditions. This approach provides a useful tool for estimating public health impacts of air quality interventions.

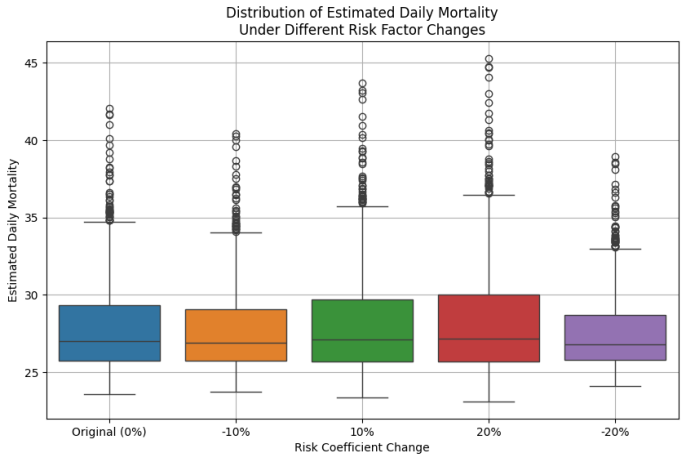


Figure 6: Mortality numbers using hypothetical scenarios

The plot titled *Mortality numbers using hypothetical scenarios* shows how changes in $\text{PM}_{2.5}$ concentrations affect estimated daily mortality. The x-axis represents percentage

changes in $PM_{2.5}$, ranging from -20% to $+20\%$, while the y-axis shows the estimated number of deaths per day. Each vertical line in the plot corresponds to a scenario and represents the 95% confidence interval of estimated mortality from Monte Carlo simulations.

Reductions in $PM_{2.5}$ levels (e.g., -10% and -20%) are associated with fewer deaths per day, whereas increases in $PM_{2.5}$ levels (e.g., $+10\%$ and $+20\%$) lead to higher estimated mortality. The baseline scenario (0% change) is the reference point. The results show a clear relationship: as $PM_{2.5}$ increases, so does daily mortality. Confidence intervals for the most extreme scenarios are largely nonoverlapping, indicating statistically significant differences. This plot highlights the potential public health benefits of reducing $PM_{2.5}$ pollution and emphasizes the importance of air quality management.

5.1.3 Noise Injection

To address the limitation of low variability and better reflect real-world fluctuations, we introduced synthetic stochastic variation into the dataset. Specifically, we added 5% Gaussian noise to the PM2.5 values, simulating sensor error and natural environmental variability. This step helped broaden the distribution tails, allowing the Monte Carlo simulation to more effectively explore uncertainty in pollution exposure and its potential health consequences.

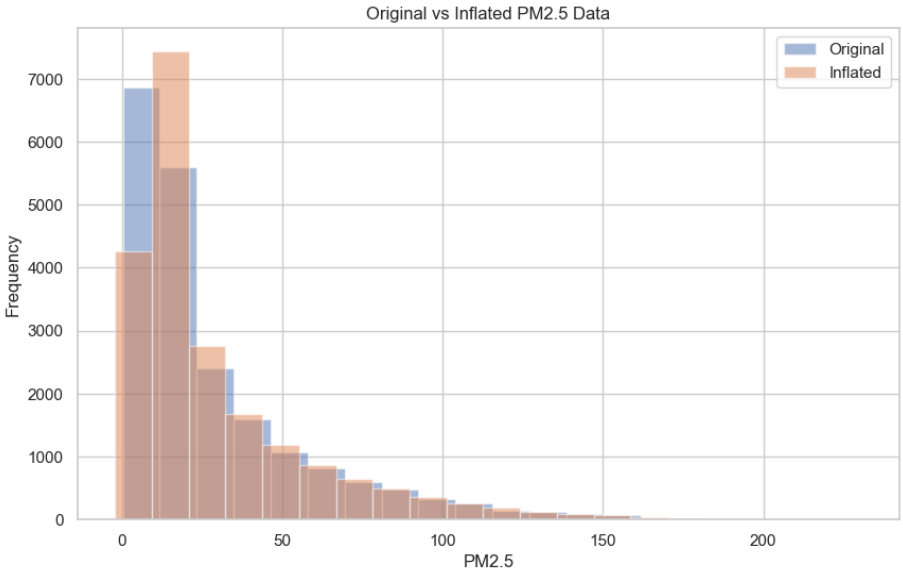


Figure 7: Imputation results

6 Theoretical Background

This section provides an extensive theoretical foundation for air pollution forecasting, with a focus on machine learning and time series analytical methods. Time series models are often used to better understand the temporal dynamics of air quality indicators since they are especially made to analyze data points gathered over time. Reliable forecasting depends on the ability to identify underlying patterns, including trends, seasonality, and cyclical variations, which these models help with finding.

The AutoRegressive Moving Average (ARMA) model is a popular classical method for modelling and forecasting stationary time series data. It captures the connection between an observation and its historical values and errors. Air quality data that shows seasonal fluctuations, such as greater pollution levels during specific months, can be effectively modelled using the SARIMA (Seasonal ARIMA) model, which extends ARMA by adding seasonal components. Recently, machine learning models that can handle complicated, non-linear interactions and enhance predicting performance, such as XGBoost and LSTM (Long Short-Term Memory networks), have become more and more popular. In high-dimensional datasets where traditional statistical models might not be able to accurately represent intricate patterns and relationships, these models are very effective.

In the subsequent sections, we will focus on the theoretical foundations of these models, beginning with the ARMA and SARIMA models and progressing to machine learning approaches, including LSTM and XGBoost.

6.1 Traditional time series methods

6.1.1 ARIMA Model

The ARIMA model, denoted as ARIMA(p, d, q)(P, D, Q), is a powerful tool for analyzing and forecasting time series data that show both trend and seasonal patterns. In this notation, the parameters p, d, and q correspond to the non-seasonal components: p represents the order of the autoregressive (AR) term, d is the degree of differencing required to achieve stationarity, and q indicates the order of the moving average (MA) term. The seasonal components P, D, and Q define similar orders but over seasonal lags (e.g., monthly or yearly cycles), enabling the model to capture seasonal patterns.

The ARIMA model integrates three key components: the Autoregressive (AR) process, which models the relationship between an observation and a number of lagged observations; the Integrated (I) component, which applies differencing to remove trends and ensure stationarity; and the Moving Average (MA) process, which models the relationship between an observation and past forecast errors.

6.1.2 Autocorrelation and Partial Autocorrelation

Autocorrelation and partial autocorrelation are important tools in time series analysis used to identify the internal structure and dependencies within a dataset. These measures help determine the order of the autoregressive (AR) and moving average (MA) components of ARIMA and SARIMA models.

Autocorrelation Function (ACF): The autocorrelation function measures the correlation between observations of a time series separated by a lag of k time units. For a weakly stationary time series $\{X_t\}$ with mean μ and autocovariance function $\gamma(k)$, the autocorrelation at lag k is defined as:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\mathbb{E}[(X_t - \mu)(X_{t-k} - \mu)]}{\mathbb{E}[(X_t - \mu)^2]}$$

The ACF provides an understanding into how current values of the series are related to past values. In practice, the sample autocorrelation is computed and plotted to visualize these dependencies. [11].

Partial Autocorrelation Function (PACF): The partial autocorrelation function (PACF) measures the direct relationship between a value and its lagged value at lag k, after accounting for the effects of all shorter lags (lags 1 to $k-1$). [14] In other words, it shows how much of the current value can be explained by the value k steps before, without interference from the values between. The PACF at lag k, denoted ϕ_{kk}, is the coefficient of X_{t-k} in the regression of X_t on $X_{t-1}, X_{t-2}, ..., X_{t-k}$:

$$X_t = \phi_{k1}X_{t-1} + \phi_{k2}X_{t-2} + \cdots + \phi_{kk}X_{t-k} + \varepsilon_t$$

Interpretation and Use: Plotting the ACF and PACF helps to understand whether the time series follows an AR, MA, or ARMA process:

- **AR**(p): PACF cuts off after lag p, while the ACF decays gradually.
- **MA**(q): ACF cuts off after lag q, while the PACF decays gradually.
- **ARMA**(p, q): Both ACF and PACF decay gradually.

To include statistical significance, confidence intervals are generally set at:

$$\pm\frac{2}{\sqrt{n}}$$

where n is the sample size. Any autocorrelation coefficient that drops outside these bounds indicates a statistically significant correlation at the corresponding lag.

6.1.3 General Formula of ARIMA Using the Backshift Operator

The ARIMA model, short for AutoRegressive Integrated Moving Average, is a powerful tool for modeling and forecasting time series data.[15] It combines three components:

- **AR (AutoRegressive)**: The current value depends linearly on its previous values.
- **I (Integrated)**: Differencing is used to make the series stationary.
- **MA (Moving Average)**: The current value depends linearly on past error terms.

To express these components concisely, we use the **backshift operator** B, defined as:

$$B^k X_t = X_{t-k}$$

This allows us to write lagged terms efficiently.

The general ARIMA(p, d, q) model using the backshift operator is:

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\varepsilon_t$$

where:

- $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the autoregressive (AR) polynomial,
- $(1 - B)^d$ represents differencing d times to achieve stationarity,
- $\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ is the moving average (MA) polynomial,
- ε_t is white noise.

For seasonal time series, the model can be extended to ARIMA$(p, d, q)(P, D, Q)_s$ by incorporating seasonal AR, MA, and differencing components with seasonal period s.

6.2 Evaluation Metrics: AIC and BIC

In time series forecasting, selecting the best model involves evaluating its performance based on statistical criteria. Two commonly used evaluation metrics for model comparison are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). [16] These criteria penalize models for complexity (number of parameters) and help to avoid overfitting.

6.2.1 Akaike Information Criterion (AIC)

The AIC is a widely used metric for model selection, defined as:

$$\text{AIC} = -2\log(\hat{L}) + 2k$$

where:

- $\hat{L}$ is the likelihood of the model, representing how well the model fits the data,
- k is the number of estimated parameters in the model.

A lower AIC value indicates a better model, balancing the trade-off between goodness of fit and model complexity.

6.2.2 Bayesian Information Criterion (BIC)

The BIC, also known as the Schwarz Information Criterion (SIC), is similar to AIC but applies a stronger penalty for the number of parameters. It is defined as:

$$\text{BIC} = -2\log(\hat{L}) + k\log(n)$$

where:

- n is the number of observations,
- k is the number of estimated parameters.

Like AIC, a lower BIC value indicates a more favorable model, but since BIC penalizes complexity more heavily than AIC, it tends to prefer simpler models.

In our research, the `auto_arima` function from the `pmdarima` library was employed to automatically select the best-fitting ARIMA model. The function conducts a search across various combinations of model parameters and identifies the model that minimizes the AIC criterion. This approach guarantees that the selected model strikes a balance between goodness of fit and model complexity, thereby reducing the risk of overfitting while enhancing predictive accuracy.

7 Results and Discussion

In this section, we present and discuss the results of our air pollution prediction models. We explore various forecasting methods, including SARIMA, ETS, SARIMAX with exogenous variables, XGBoost, Random Forest, SVR, comparing their performance in predicting PM2.5 levels for the future. We also evaluate how accurately each model performs by looking at their strengths and weaknesses in the context of forecasting air quality in Yerevan.

Also we explain the methodology used to select and evaluate the models, providing insights into which approaches were the most effective and why.

The graph below displays the results of a naÃ¯ve prediction model and highlights the limitations of simple forecasting approaches. The model fails to capture key patterns such as seasonality, trends, and variability in $PM_{2.5}$ concentrations, resulting in poor predictive accuracy. This underperformance underscores the need for more advanced models that can account for the complex temporal dynamics of air pollution data.

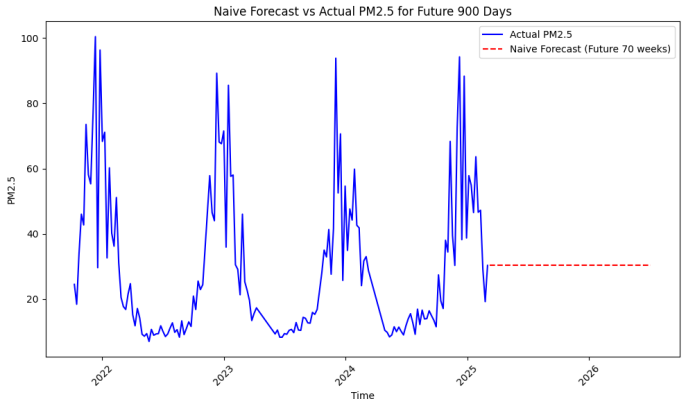


Figure 8: Naive prediction model

For some methods, we utilized different data granularities. In one case, we calculated weekly averages of the PM2.5 levels, while in the other, daily data was used. To evaluate the effect of time scale on model accuracy and performance, this variance in data granularity was taken into consideration.

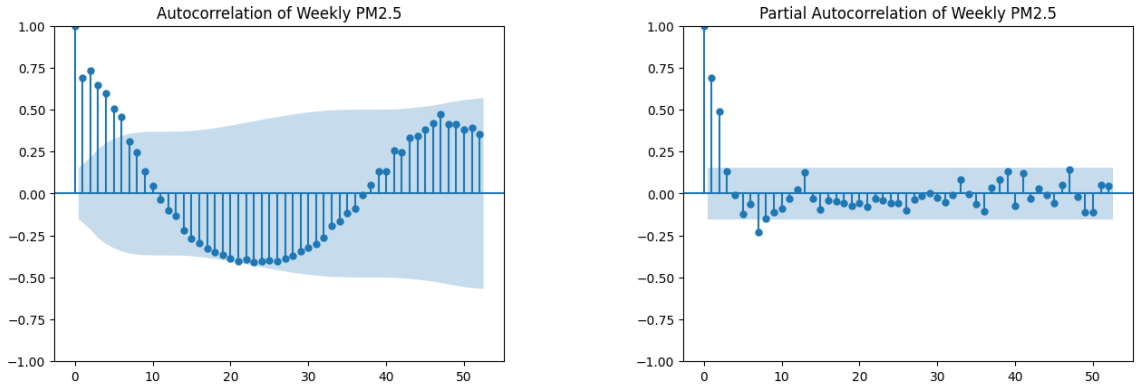
7.1 Traditional time series analysis

We began our analysis using traditional time series methods, starting with the examination of the autocorrelation function (ACF) and partial autocorrelation function (PACF). Since the dataset contains one observation per week, all lag-based conclusions are interpreted on a weekly basis.

Observations from the ACF and PACF plots: The lag-1 autocorrelation is close to 0.9, indicating a very strong dependence on the previous week's PM2.5 value. The ACF indicates a slow, sinusoidal decay, crossing the zero line around lags 13–14 and rising again near lags 26 and 52. This pattern reflects a seasonal component, and the prominent peak at lag 52 suggests the presence of an annual seasonal cycle.

The PACF shows significant spikes at lags 1, 2, and 3, which indicates that the current PM2.5 level is directly influenced by the values from the previous one to three weeks. This

helps inform the potential autoregressive structure of models such as ARIMA.



(a) PACF (b) ACF

Figure 9: Autocorrelation and Partial Autocorrelation plots of PM2.5 weekly values

Observations from the Time Series Plot:

- **Seasonality:** There is a clear seasonal pattern, with peaks in PM2.5 levels typically occurring in late fall to winter (November–January). These elevated levels may be due to increased heating, more stagnant air during cold weather, and higher emissions. In contrast, PM2.5 levels drop significantly during spring and summer, likely due to improved atmospheric dispersion, reduced heating demand, and increased rainfall.

- **Trend:** No clear long-term increasing or decreasing trend is visually apparent. The PM2.5 levels exhibit a cyclical pattern rather than a consistent upward or downward trajectory.

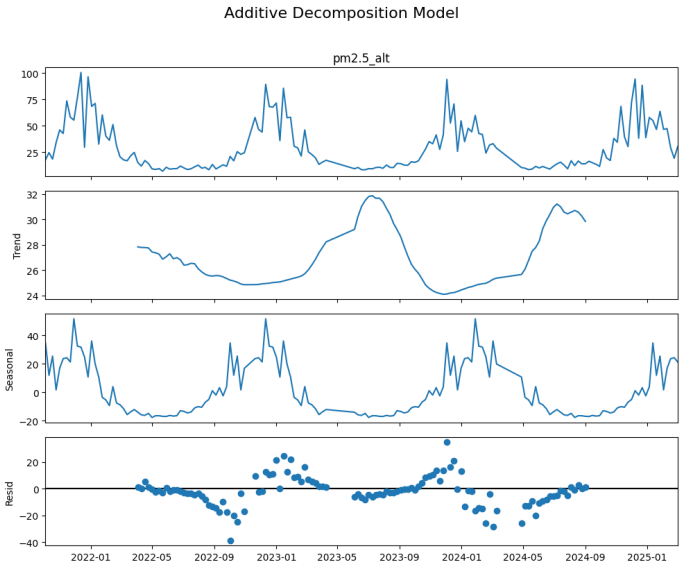


Figure 10: Correlation Matrix

Residual Analysis and Additive Decomposition: To assess the variance stability of the PM2.5 time series, we applied the Levene test by splitting the dataset into two equal parts (early and late periods) and comparing their variances. The resulting p-value of 0.496 shows that there is no important difference in variance between the two parts of the data ($p \geq 0.05$). This means the assumption of constant variance is reasonable, so using an **additive decomposition model** is suitable for further analysis.

The additive decomposition reveals several key insights. The **original PM2.5 series** exhibits strong seasonal spikes that occur annually, peaking during late fall to winter (November–January) with values near 100 μg/m^3, and dropping below 20 μg/m^3 in warmer months. The **trend component** indicates smooth fluctuations, with a gradual decline from late 2021 to mid-2022, followed by an increasing pattern from late 2022 into early 2024.

Residual Component: The residual component captures short-term variations not explained by the trend or seasonality. While many residuals are minor, some spikes reach up to 25 μg/m^3, indicating significant deviations. The residuals are not completely randomâcertain periods, such as early 2023 and mid-2024, show systematic over- or under-estimations, suggesting the presence of external influences or limitations in the model.

7.1.1 SARIMA Model

To see if there is a necessity of differentiation, we employed the `auto_arima` function from the `pmdarima` library. This function evaluates various combinations of ARIMA parameters

and selects the best-fitting model based on the Akaike Information Criterion (AIC). In our case, it identified the following model:

```
Total fit time: 121.114 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  166
Model:             SARIMAX(3, 0, 2)x(0, 0, [1], 52)   Log Likelihood               -657.055
Date:                              Tue, 06 May 2025   AIC                           1330.111
Time:                                      00:09:58   BIC                           1355.007
Sample:                                           0   HQIC                          1340.216
                                                - 166
Covariance Type:                                opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.7483      0.127      5.905      0.000       0.500       0.997
ar.L1          1.6182      0.087     18.503      0.000       1.447       1.790
ar.L2         -0.3023      0.169     -1.786      0.074      -0.634       0.029
ar.L3         -0.3438      0.083     -4.143      0.000      -0.506      -0.181
ma.L1         -1.5640      0.096    -16.246      0.000      -1.753      -1.375
ma.L2          0.5718      0.102      5.599      0.000       0.372       0.772
ma.S.L52       0.1362      0.111      1.227      0.220      -0.081       0.354
sigma2       143.8459     12.447     11.557      0.000     119.450     168.241
==========================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):               126.08
Prob(Q):                              0.83   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                             0.54
Prob(H) (two-sided):                  0.91   Kurtosis:                         7.13
==========================================================================================
```

Figure 11: SARIMA model

ARIMA(3,0,2)(0,0,1)[52] with intercept

The interpretation of this model is:

- **Non-seasonal part:** AR(3), I(0), MA(2) — three autoregressive terms and two moving average terms with no differencing required.
- **Seasonal part:** MA(1) — one seasonal moving average term with a periodicity of 52 (reflecting weekly data with yearly seasonality).
- **Stationarity:** No differencing needed ($d = 0$, $D = 0$), confirming that the series is stationary

The model diagnostics support the adequacy of this SARIMA specification:

- **Ljung-Box test:** $Q = 0.05$, $p = 0.83$ — residuals show no significant autocorrelation and resemble white noise.
- **Heteroskedasticity test:** $H = 1.03$, $p = 0.91$ — residuals show no significant change in variance, indicating constant variance and eliminating the need for GARCH modeling.

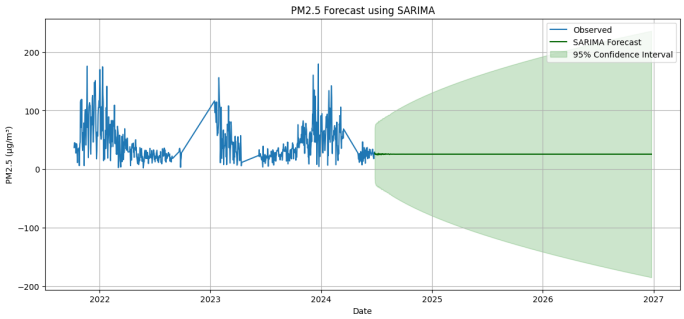


Figure 12: SARIMA model using daily average

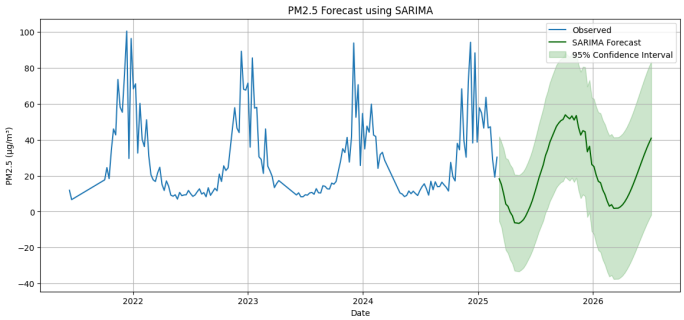


Figure 13: SARIMA model using weekly average

In the context of SARIMA models, higher data granularity, such as daily frequency, can introduce excessive noise and obscure the underlying seasonal and trend components. This can make model identification and parameter estimation more challenging. Aggregating the data to a weekly frequency reduces short-term variability, enhances the stability of seasonal patterns (e.g., annual cycles in weekly data), and allows the SARIMA model to better capture the pattern of the time series, leading to more accurate forecasts.

7.1.2 Holt-Winters Exponential Smoothing

Exponential smoothing is a classical time series forecasting technique that assigns exponentially decreasing weights to past observations. By giving more importance to recent data points, this method allows for faster adaptation to changes in the time series. Depending on the characteristics of the data, different forms of exponential smoothing can be applied. Simple Exponential Smoothing (SES) is used when the data exhibits no clear trend or seasonality. Holt's method extends SES to capture linear trends, while the Holt-Winters method further incorporates seasonal patterns, either additively or multiplicatively.

In our analysis, exponential smoothing demonstrated clear advantages in terms of simplicity and computational speed. It effectively captured short-term fluctuations and seasonal cycles, particularly when the series displayed stable patterns. However, one potential limitation of this method is its tendency to overfit, especially when the smoothing parameters are not regularized adequately.

A key insight from our research was that exponential smoothing performed significantly better on weekly data compared to daily data. Daily observations introduced excessive noise and irregular fluctuations, which often overwhelmed the seasonal signal. In contrast, aggregating the data to a weekly level reduced variance and highlighted the underlying structure of the series, resulting in more stable and accurate forecasts.

Overall, exponential smoothing proved to be a valuable baseline model for short-term forecasting, especially when data were aggregated to a coarser time resolution such as weekly intervals.

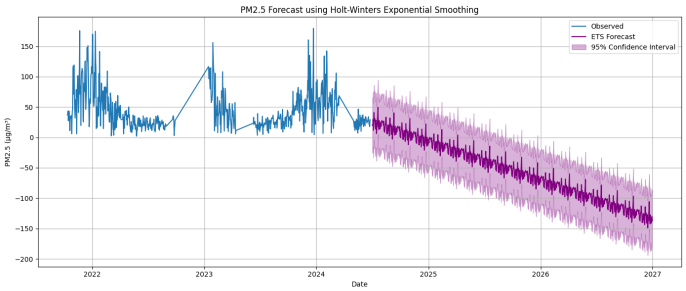


Figure 14: ETS model using daily average

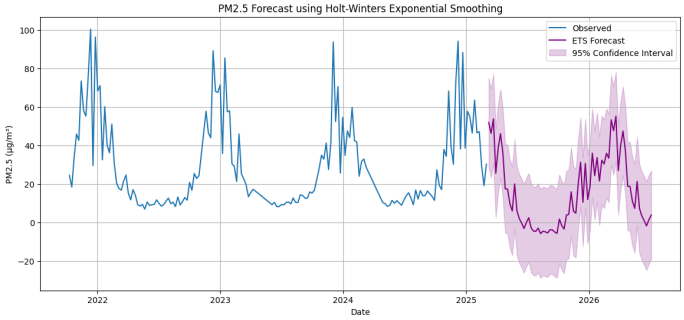


Figure 15: ETS model using weekly average

Higher granularity in time series data, such as daily observations, can lead to increased noise, greater variability, and less reliable model performance. In contrast, aggregating the data to a weekly level smooths out short-term fluctuations, captures broader patterns more effectively, and often results in improved forecasting accuracy.

7.1.3 XGBoost Model

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on gradient boosting decision trees. Unlike traditional time series models like ARIMA or exponential smoothing, XGBoost is non-parametric and capable of modeling complex nonlinear relationships, as well as interactions between features.

In our study, XGBoost was applied to forecast PM2.5 levels using both past pollution levels. Compared to classical time series models, XGBoost demonstrated superior predictive performance, generating the lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) among all models tested.

One key advantage of XGBoost is its flexibility—it does not require the data to be stationary, and it can easily incorporate multiple lagged variables and moving averages. Likewise, its built-in regularization helps prevent overfitting, especially in high-dimensional settings.

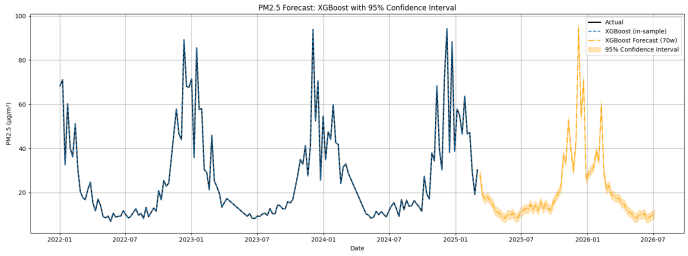


Figure 16: XGboost model using weekly average

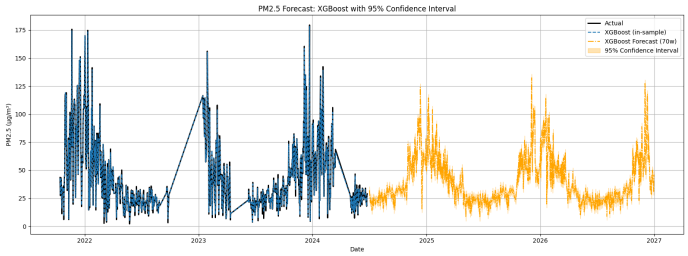


Figure 17: XGboost model using daily average

For XGBoost, the predicted values for weekly and daily averages do not look much different from each other visually. This means that XGBoost is better at reducing prediction errors by capturing complex patterns and using regularization, rather than making the forecasts look very different from simpler methods.

7.1.4 SARIMAX using exogenous variables

Despite adding temperature and humidity as extra variables, the improvement in forecasting results was small. The error values, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), showed that the SARIMAX model did not perform much better than the simpler models. This means that while temperature and humidity might affect PM2.5 levels, their effects may be too small or too complex to be fully captured by the model we used.

Weather conditions like temperature and humidity can influence air pollution, but their impact may change depending on other factors, such as wind or traffic. To improve the model, we may need to include more types of outside information or use a more advanced method that can better handle these complicated relationships.

It is important to note that the SARIMAX model has certain limitations:

- **Assumes linearity:** SARIMAX assumes a linear relationship between the dependent and independent variables, which may not capture more complex, non-linear patterns.
- **Struggles with sudden spikes or irregular patterns:** The model may not perform well in situations with sharp changes or highly irregular patterns in the data, as it is designed to model smoother trends and seasonality.

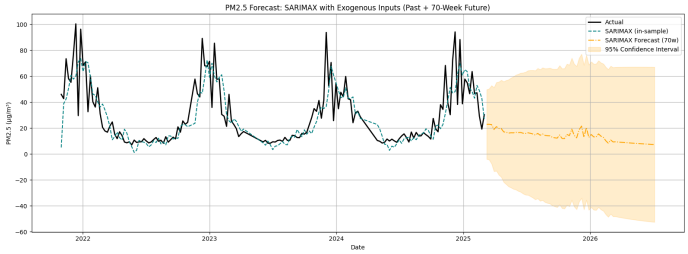


Figure 18: SARIMAX using exogenous variables using weekly average

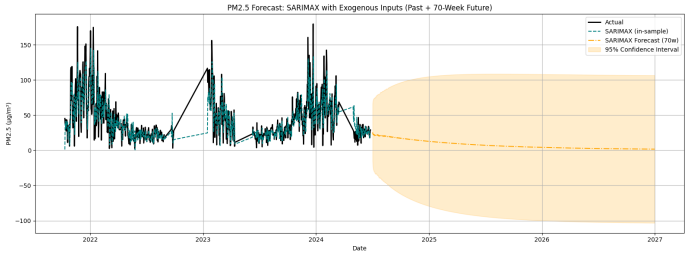


Figure 19: SARIMAX using exogenous variables using daily average

7.1.5 Random forest

The Random Forest model is the second best model indicated by the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. It does not outperform XGBoost, but provides a strong predictive capability for forecasting PM2.5 levels.

One of the key reasons Random Forest performs well is its ability to capture complex relationships between the features without assuming any particular form of the relationship (e.g., linearity). This flexibility allows it to model non-linear patterns in the data effectively. Additionally, by using multiple decision trees, Random Forest can reduce the risk of overfitting, especially when dealing with a large number of features or noisy data.

When comparing the results across different granularities, we observed that the model performs better on weekly data than on daily data. The RMSE and MAE metrics were both lower when using weekly averages, which suggests that the model benefits from the smoother, more aggregated data. Weekly data helps reduce noise and random fluctuations that might be present in daily measurements, allowing the Random Forest model to make more accurate predictions.

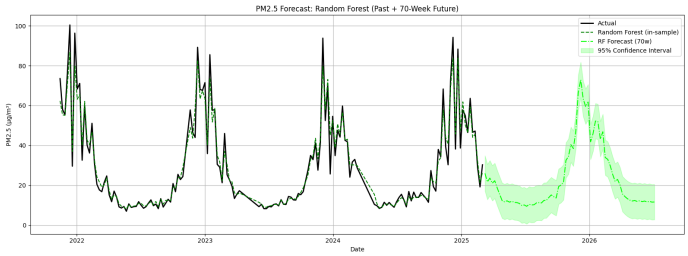


Figure 20: Random forest using weekly average

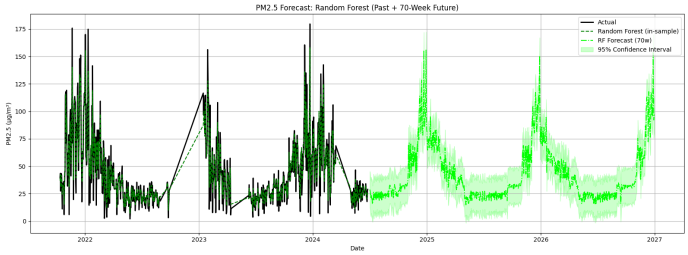


Figure 21: Random forest using daily average

7.1.6 Support Vector Regression (SVR)

Support Vector Regression (SVR) is a powerful machine learning technique that uses support vector machines for regression tasks. It is particularly practical in capturing

complex non-linear relationships by mapping input data into a higher-dimensional space using a kernel function. Despite its potential, the performance of the SVR model in this analysis was not as satisfactory compared to other models, particularly in terms of the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

One challenge with Support Vector Regression (SVR) is that it depends heavily on its settings, such as the kernel, regularization parameter (C), and epsilon. If these settings are not chosen carefully, the model might overfit or underfit the data, which means it may not work well on new data. Additionally, SVR can be slow, especially with large datasets, making it more time-consuming to train and predict results.

In terms of results, the SVR model performed better on daily data rather than weekly data, which is unusual since many models benefit from the smoother, aggregated nature of weekly data. The daily data provided more granular information, allowing the SVR model to capture short-term variations and better align with the actual values. Both the MAE and RMSE for the daily data were lower compared to the weekly data.

However, the model still faced limitations in handling sudden spikes or irregular patterns, as SVR tends to struggle with deviations from expected behavior. Also, SVR assumes a smooth relationship between the predictors and the target variable, which may not always hold in real-world data.

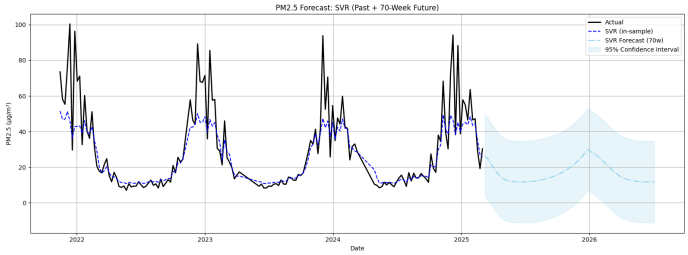


Figure 22: SVR using weekly average

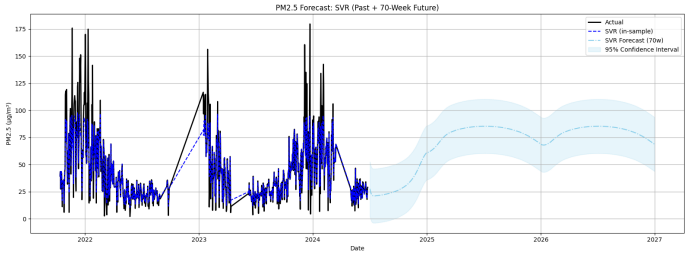


Figure 23: SVR using daily average

7.2 Long Short-Term Memory (LSTM)-based neural networks

Air pollution, particularly fine particulate matter, poses serious health risks in urban environments. Accurate forecasting of $PM_{2.5}$ levels is essential for early warnings and effective environmental policy planning. In this research, we developed several methods for predicting future $PM_{2.5}$ values, including a deep learning forecasting framework based on *Long Short-Term Memory* (LSTM) networks to model and predict $PM_{2.5}$ concentration levels in Yerevan, Armenia.

The goal was to use the temporal dynamics of both pollution and meteorological data to forecast $PM_{2.5}$ trends up to 2026–12–25. We designed two primary LSTM architectures: a standard Sequence-to-Sequence (Seq2Seq) LSTM model and an improved Bidirectional Seq2Seq model.

7.2.1 LSTM Architecture and Relevance

LSTM is a form of Recurrent Neural Network (RNN) that captures long-term dependencies in sequential data. Unlike traditional RNNs, LSTMs address the problem of vanishing gradients through regulated mechanisms such as the forget gate, the input gate, and the output gate. These components allow the network to retain or discard past information based on its importance, which is particularly valuable in predicting air quality, where historical data can have lasting effects.

The unique temporal characteristics of pollution data make LSTMs well suited to forecast $PM_{2.5}$ levels. $PM_{2.5}$ concentrations are shaped by cyclical (seasonal), meteorological, and anthropogenic factors, all of which interact with time in complex, non-linear ways. Such patterns are difficult to capture using shallow regressors or conventional linear time-series models (e.g., ARIMA). In contrast, LSTMs are ideal for this task because they can handle multivariate inputs and model complex dependencies over long time sequences.

7.2.2 Challenges with Temporal Aggregation

Our forecasting methodology started by utilizing the $PM_{2.5}$ weekly averages data. However, this tactic soon turned out to be inefficient. Even after parameter adjustments, our first LSTM models trained on weekly-aggregated data generated much higher error metrics, with RMSE values regularly above 40 $\mu g/m^3$. The main cause of the high Mean Squared Error (MSE) was the Temporal Granularity Loss. Significant sharp increases in $PM_{2.5}$ after temperature shifts or wind decline were removed when aggregating to weekly resolution. Consistent and fine-grained patterns are essential for learning meaningful representations in LSTM models, which are intended to capture complex sequential relationships. These constraints made the forecasting performance on weekly data unstable, inaccurate, and unable to capture daily dynamics that are essential for public health actions and air quality warnings. Following much testing, it was determined that daily resolution not only provided

superior learning signals but also more closely matched the practical use of pollution predictions, where daily variations are significant.

As a result, daily-level data was used for all ensuing analysis, feature engineering, and modeling. As noted in the final models, this change significantly enhanced performance, resulting in RMSE values between 16 and 24 $\mu g/m^3$.

7.2.3 Dataset and Feature Engineering

The dataset used in this study included key meteorological variables such as humidity, temperature, and air pressure. To enhance model performance, we applied targeted feature engineering techniques as follows:

- **Lag Features:** To capture short-term autocorrelations, we added 1-day (`pm2.5_lag1`) and 7-day (`pm2.5_lag7`) lagged $PM_{2.5}$ values.
- **Rolling Statistics:** A 7-day rolling mean (`pm2.5_roll7`) and standard deviation (`pm2.5_std7`) were computed to reflect smoothed trends and volatility in pollution levels.
- **Seasonality Encoding:** The `month` feature was transformed into its cyclical representations using sine and cosine functions (`month_sin`, `month_cos`) to effectively model annual seasonality.
- **Weather Covariates:** Meteorological factors such as humidity, temperature, and pressure were included as direct predictors due to their known causal influence on pollutant concentrations.

To ensure consistency and comparability among features, all variables were normalized using a MinMaxScaler prior to model training.

7.2.4 Model Architectures

Vanilla Seq2Seq LSTM Our baseline model employed a Sequence-to-Sequence (Seq2Seq) architecture, which is commonly used for multistep forecasting tasks. The model consists of two encoder LSTM layers, with 128 and 64 units, respectively, followed by a `RepeatVector` layer and two decoder LSTM layers. The output sequence was transformed into one-dimensional $PM_{2.5}$ predictions using a `TimeDistributed` dense layer. To prevent overfitting, the model incorporated a dropout rate of 30% and L2 regularization.

Bidirectional Seq2Seq LSTM To enhance performance and stability, we implemented a bidirectional version of the Seq2Seq model. Bidirectional LSTM layers process the input sequence in both forward and backward directions, capturing temporal dependencies that may be missed when considering only one direction. This feature is particularly

advantageous for $PM_{2.5}$ forecasting, as the influence of meteorological events on pollution levels may exhibit symmetric temporal patterns.

Each encoder and decoder LSTM layer in the bidirectional model was wrapped with a `Bidirectional()` layer. Despite the added complexity, the training time remained reasonable due to early stopping and the use of small batch sizes.

7.2.5 Training Procedure and Settings

We trained both models using the Adam optimizer and employed mean squared error (MSE) as the loss function. The dataset was split such that 80% of the time-indexed data was used for training, while the remaining 20% was reserved for evaluation.

Each model was trained with a specific lookback window and an x-day prediction horizon. This design was based on the assumption that annual cycles and long-term dependencies significantly influence pollution trends in Yerevan.

Training was conducted for up to 100 epochs, with early stopping triggered if the validation loss did not improve for 10 consecutive epochs. A batch size of 16 was chosen to balance the speed of training and the stability of the model.

7.2.6 Forecasting Strategy and Evaluation

To simulate real-world forecasting scenarios, we adopted a forward iterative prediction strategy. At each step, the model predicted a future span of x days (adjusted based on validation accuracy) for $PM_{2.5}$ levels. These predictions were then recursively fed back into the model as inputs for the next forecast window. This iterative process continued until a total of 913 future days were forecasted, covering approximately 2.5 years.

Model evaluation was performed using the predefined test set. The predicted values were inverse-transformed from their scaled form, and performance metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were computed.

The best performance metrics for each model were as follows:

- **Vanilla Seq2Seq:** MSE = 563.07, RMSE = 23.73
- **Bidirectional Seq2Seq:** MSE = 276.50, RMSE = 16.63

The Bidirectional Seq2Seq model showed a nearly 30% reduction in RMSE compared to the vanilla model, confirming the hypothesis that bidirectional temporal modeling improves the networkâs ability to learn symmetric dependencies in pollution patterns.

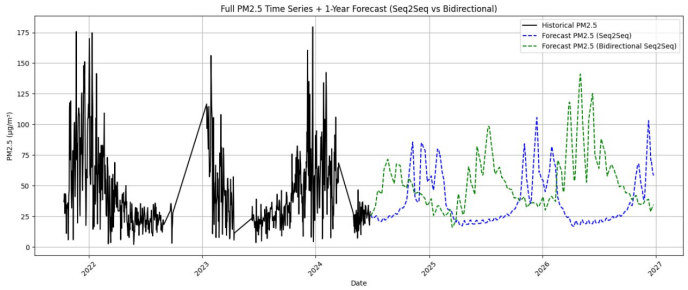


Figure 24: LSTM predictions

7.2.7 Seasonal and Trending Behaviour

During the course of our project, we observed that incorporating sine and cosine transformations of the `month` variable greatly enhanced the model's ability to understand seasonal variations. These cyclic features allowed the models to capture yearly patterns more effectively.

Additionally, the inclusion of a large lookback window (365 days) enabled the models to detect long-term trends, such as gradual yearly increases or declines in $PM_{2.5}$ levels. These trends were likely driven by climatological factors or changes in environmental policies. The ability to accurately model these long-term patterns was crucial in generating reliable future projections, significantly improving the temporal precision of the forecasts.

7.2.8 Model Reliability and Justification

The reliability of the LSTM-based models was assessed through both statistical and visual analyses. The models demonstrated strong performance, with RMSE values of less than 24 $\mu g/m^3$ for the Seq2Seq model and 17 $\mu g/m^3$ for the BiLSTM model. These results indicated that the models were able to accurately distinguish between high-risk pollution periods and normal conditions.

We justified the selection of these models based on the following factors:

- **Temporal structure of the $PM_{2.5}$ data:** The sequential nature of $PM_{2.5}$ time series data made it particularly suitable for LSTM-based modeling.
- **Feature richness:** The inclusion of carefully selected variables enhanced the model's ability to capture important context and underlying patterns in the data.
- **Recursive forecasting capability:** LSTMs are inherently capable of handling extended sequential input/output formats, making them ideal for recursive forecasting tasks.

- **Peak-aware tuning potential (future expansion):** These models can be fine-tuned to place more emphasis on critical pollution peaks, allowing for future enhancements focused on public health and pollution control.

7.2.9 LSTM improvements

The use of LSTM architectures, particularly the Bidirectional Seq2Seq version, proved to be highly effective for long-term $PM_{2.5}$ forecasting in a complex urban environment. By incorporating meteorological data, engineered lag features, and seasonal signals, the models were able to generate reliable and realistic forecasts over multiple years.

Future research could focus on the integration of advanced techniques such as:

- **Attention mechanisms:** Incorporating attention processes could allow the model to focus on more relevant time periods, improving forecasting accuracy.
- **Transformer-based temporal models:** These models have demonstrated strong performance in capturing long-range dependencies and could enhance spatiotemporal resolution.
- **Multivariate convolutional layers:** These could be integrated to model more complex interactions among multiple environmental variables.
- **Quantile-based loss functions:** Implementing such loss functions might provide better performance under heavy-tailed distributions, commonly seen in pollution data.
- **Uncertainty modeling (e.g., Monte Carlo Dropout):** This approach could offer probabilistic forecasts, which are critical for issuing reliable public health alerts.

8 Conclusion

Using real-time sensor data and time-series modelling approaches, this study compares several models for predicting $PM_{2.5}$ concentrations in Yerevan, Armenia. Using MSE and RMSE criteria, nine modelsâincluding statistical, machine learning, and deep learning techniquesâwere assessed on daily and weekly prediction tasks. XGBoost was the most accurate and reliable model out of all of them, with a remarkably low daily RMSE of 1.49 and weekly RMSE of 0.014. These outcomes show how well XGBoost uses time-based features and can handle nonlinear, high-variance pollution data.

With a daily RMSE of 2.27, Support Vector Regression (SVR) is the second, but its performance deteriorated on weekly averages. With a daily RMSE of 14.51 and a weekly RMSE of 12.45, Random Forest provided a balance between generalisability and accuracy, making it a reliable model at both resolutions. Traditional time series models like SARIMA and SARIMAX showed a limited ability to predict complex and dynamic pollution patterns

in real-world situations, which is reflected in their higher RMSE values of 21.38 and 28.51 (daily) and 21.47 and 32.00 (weekly). The performance difference between traditional methods and contemporary AI models was demonstrated by the Naive model, which were used as baselines.

Deep learning models provided insightful information. With an MSE of 563.07 and an RMSE of 23.73 for daily predictions, the Sequence-to-Sequence (Seq2Seq) model was unable to understand complex long-term connections in this application. With an MSE of 276.50 and an RMSE of 16.63, the Bidirectional LSTM (BiLSTM) outperformed the majority of machine learning and statistical models on the daily dataset. This result shows the advantage of using bidirectional temporal processing for sequence modeling. Overall, the results show that the three best-performing models are Random Forest, BiLSTM, and XGBoost. While BiLSTM provides a deep learning option for capturing sequence dynamics, Random Forest guarantees stable predictions, and XGBoost is the most accurate and adaptable across data granularities.

These findings support the use of advanced deep learning and ensemble methods for monitoring air quality in cities. They show how predictive models can help guide long-term environmental planning, inform policy decisions, and trigger timely public health alerts. To make predictions even more accurate, future research could focus on combining different models, using transfer learning, and including additional factors like traffic and industrial activities.

Model	**MSE**	**RMSE**
Naive	600.17	24.50
SARIMA	13.58	21.38
ETS	18.63	25.83
SARIMAX	18.84	28.51
XGBoost	1.08	1.49
Random Forest	10.77	14.51
SVR	1.48	2.27
Seq2Seq	563.07	23.73
BiLSTM	276.50	16.63

Table 2: Model Performance: MSE and RMSE Values using daily averages

Model	MSE	RMSE
Naive	319.5708	17.8765
SARIMA	15.4172	21.4771
ETS	31.7786	38.5492
SARIMAX	25.6747	32.0003
XGBoost	0.0104	0.0140
Random Forest	8.1587	12.4542
SVR	7.9351	14.0034

Table 3: Model Performance: MSE and RMSE Values using weekly averages

9 Future Work

Future studies could expand the analysis of air pollution by including other pollutants that also affect air quality and human health, such as PM10, NO_2, CO, and O_3 [6]. These pollutants, along with their intercorrelations, could be studied using similar time series models to predict their levels over time and understand how they change throughout the year.

A key improvement for future studies would be the use of longer past data for training the models. Increasing the range of historical data would make it easier to identify cyclical patterns, seasonal fluctuations, and long-term trends that are essential for accurate forecasting. To manage the larger volume of data and the complexity of more advanced models, more computational resources would be required, especially powerful GPUs.

Further research on the relationship between air pollution and other urban factors, such as traffic jams and construction projects, could also provide insights into the root causes of pollution. Access to detailed health data would enable a deeper investigation into how air pollution directly impacts public health. For example, linking pollution levels with medical records on cardiovascular diseases, asthma, chronic obstructive pulmonary disease (COPD), and other pollution-related conditions could help identify patterns and potential causal relationships.

References

[1] JJP News, "What is the Relationship Between Pollution and Human Health?" 2023.

[2] G. Tylor, "Wildfire Smoke: The Silent Killer and AI Battling Its Threat," 2024.

[3] The BMJ, "Global mortality from outdoor fine particle and ozone air pollution," 2023.

[4] A. Mkrtchyan, M. Avagyan, and A. Harutyunyan, "AIR QUALITY IN YEREVAN IN THE CONTEXT OF CLIMATE CHANGE," 2024.

[5] World Health Organization, "WHO Air Quality Guidelines," 2021.

[6] E. Marinov, D. Petrova-Antonova, and S. Malinov, "Time series forecasting of air quality: A case study of sofia city," *atmosphere*, 2022.

[7] J. Brownlee, "Long Short-Term Memory Networks with Python: Develop Sequence Prediction Models with Deep Learning," 2017, p. 229.

[8] Y. Zhukovskiy, N. Korolev, I. Babanova, and . Boikov, "The prediction of the residual life of electromechanical equipment based on the artificial neural network," 2017, iP Conference Series: Earth and Environmental Science 87, 032056.

[9] G. Kumar, "Machine Learning with Applications," 2021, international Journal for Research in Applied Science and Engineering Technology 9(4), 369–373.

[10] T. Liu, A. K. H. Lau, K. Sandbrink, and J. C. H. Fung, "Time series forecasting of air quality based on regional numerical modeling in hong kong," *Journal of Geophysical Research: Atmospheres*, 2018. [Online]. Available: https://doi.org/10.1002/2017JD028052

[11] J. Fattah, L. Ezzine, Z. Aman, H. E. Moussami, and A. Lachhab, "Forecasting of Demand Using ARIMA Model," pp. 1–9, 2018.

[12] PurpleAir, "PurpleAir Real-Time Air Quality Monitoring Map," 2025.

[13] AQICN, "Air Quality Index (AQI) - Armenia, Yerevan Kentron Station," 2025.

[14] U. A. Yakubu and M. P. A. Saputra, "Time Series Model Analysis Using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for E-wallet Transactions during a Pandemic," pp. 80–85, 2022.

[15] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting*. Springer-Verlag, 1996.

[16] M. Z. Hossain, "AIC and BIC â The two competitive information criteria for model selection in economics and statistics," 1998.

[17] V. G. Margaryan and G. A. Khachatryan, "About the Assessment of Atmospheric Air Pollution of the City of Yerevan," 2022.

[18] K. C. N. Dozie and M. Ijomah, "A Comparative Study on Additive and Mixed Models in Descriptive Time Series," 2020, american Journal of Mathematical and Computer Modelling, DOI: 10.11648/j.ajmcm.20200501.12.

[19] M. Morf, A. Vieira, and T. Kailath, "Covariance Characterization by Partial Autocorrelation Matrices," 1978.

[20] U.S. Environmental Protection Agency, "Quality assurance guidance document 2.12: Monitoring pm2.5 in ambient air using designated reference or class i equivalent methods," 2016, office of Air Quality Planning and Standards, Air Quality Assessment Division, RTP, NC 27711.

[21] B. Kim, E. Kim, S. Jung, M. Kim, J. Kim, and S. Kim, "Enhanced sequence-to-sequence attention-based pm2.5 concentration forecasting using spatiotemporal data," *Atmosphere*, 2024.

[22] O. E. Taylor and P. S. Ezekiel, "A model for forecasting air quality index in port harcourt nigeria using bi-lstm algorithm," *arXiv preprint*, 2023.