Time Series Analysis of Air Quality in Bishkek using Autoregressive Model

Bufatima Nurmuhammad kyzy

May 24, 2024

Abstract

This study presents a comprehensive analysis of PM2.5 air pollutant levels in Bishkek, Kyrgyzstan, using time series data collected from 2019 to early 2024. The objective is to understand seasonal patterns, develop forecasting models, and evaluate their performance to support air quality monitoring and management efforts. The analysis includes data preparation, exploratory data analysis (EDA), model development, evaluation techniques, and the communication of findings. We also discuss the significance of air quality monitoring, review related literature, and provide detailed insights into our methodology and results.

Contents

1 Introduction

Air quality is a critical environmental factor that directly impacts public health and the overall quality of life. In urban areas, air pollution is often exacerbated by vehicular emissions, industrial activities, and climatic conditions, leading to elevated levels of particulate matter (PM). Among these, PM2.5, which consists of particles with a diameter of 2.5 micrometers or less, is particularly harmful as it can penetrate deep into the respiratory system, causing various health problems including respiratory infections, cardiovascular diseases, and even premature death.

In Bishkek, the capital of Kyrgyzstan, air quality monitoring has become increasingly important due to rapid urbanization and industrialization. Despite the growing concern, there has been limited research focused on understanding the temporal patterns of air quality in the region. This study aims to fill this gap by analyzing PM2.5 levels over a five-year period using time series analysis techniques.

This paper is structured as follows: Section 2 provides a review of related literature on air quality monitoring and time series analysis. Section 3 describes the data collection and preparation process. Section 4 presents the exploratory data analysis (EDA) conducted to identify key patterns in the data. Section 5 details the model development process, including baseline and autoregressive models. Section 6 discusses the walk-forward validation (WFV) technique used to evaluate the model. Section 7 presents the results and compares the performance of different models. Finally, Section 8 concludes the study and suggests potential areas for future research.

2 Literature Review

Air quality monitoring and forecasting have been extensively studied in various regions around the world. Numerous studies have employed statistical and machine learning techniques to analyze air quality data and predict future pollution levels.

2.1 Air Quality Monitoring

Air quality monitoring involves the systematic collection of data on various pollutants, including PM2.5, PM10, NO2, SO2, CO, and O3. These pollutants are measured using sensors and monitoring stations strategically placed in urban and rural areas. In recent years, the use of low-cost sensors and satellite-based remote sensing has enhanced the spatial and temporal resolution of air quality data, enabling more detailed analysis and modeling.

2.2 Time Series Analysis

Time series analysis is a powerful tool for understanding temporal patterns and predicting future values in a dataset. It involves techniques such as decomposition, autocorrelation, and model fitting. Commonly used models include Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing State Space Model (ETS). These models can capture trends, seasonality, and autocorrelations in the data.

2.3 Applications in Air Quality Forecasting

Several studies have applied time series analysis to air quality data. For instance, Zhang et al. (2018) used ARIMA and Support Vector Regression (SVR) to predict PM2.5 concentrations in Beijing, demonstrating the effectiveness of combining statistical and machine learning methods. Similarly, Grange et al. (2018) employed a SARIMA model to forecast NO2 levels in Europe, highlighting the importance of accounting for seasonality in air quality data.

3 Data Preparation

Data preparation is a crucial step in time series analysis, as it ensures the quality and reliability of the input data. In this study, we collected PM2.5 data from multiple sources, cleaned the data to remove anomalies, and merged it into a single dataset.

3.1 Data Collection

The PM2.5 data for Bishkek was collected from Kaggle, where it was available in separate files for each year. The data spanned from January 2019 to February 2024, with hourly measurements of PM2.5 concentrations.

3.2 Data Merging

We obtained the data in multiple CSV files, each corresponding to a specific time period. These files were merged into a single DataFrame using the Pandas library in Python. This step ensured that the data was in a consistent format, ready for further analysis.

Listing 1: Data Loading and Merging

```
import pandas as pd

file_paths = ["Bishkek_PM2.5_2019_YTD.csv",
              "Bishkek_PM2.5_2020_YTD.csv",
              "Bishkek_PM2.5_2021_YTD.csv",
              "Bishkek_PM2.5_2022_YTD.csv",
              "Bishkek_PM2.5_2023_YTD.csv",
              "Bishkek_PM2.5_2024_YTD.csv",
              "Bishkek_PM2.5_2024_02_MTD.csv"]

df_list = [pd.read_csv(fp) for fp in file_paths]

df_merged = pd.concat(df_list, ignore_index=True)
```

3.3 Data Cleaning

Outliers and missing values can significantly affect the accuracy of time series models. Therefore, we performed data cleaning to remove anomalies and fill in missing values. Anomalies were identified based on values exceeding 500 µg/m^3, which are considered extreme outliers for PM2.5 concentrations based on air quality standards.

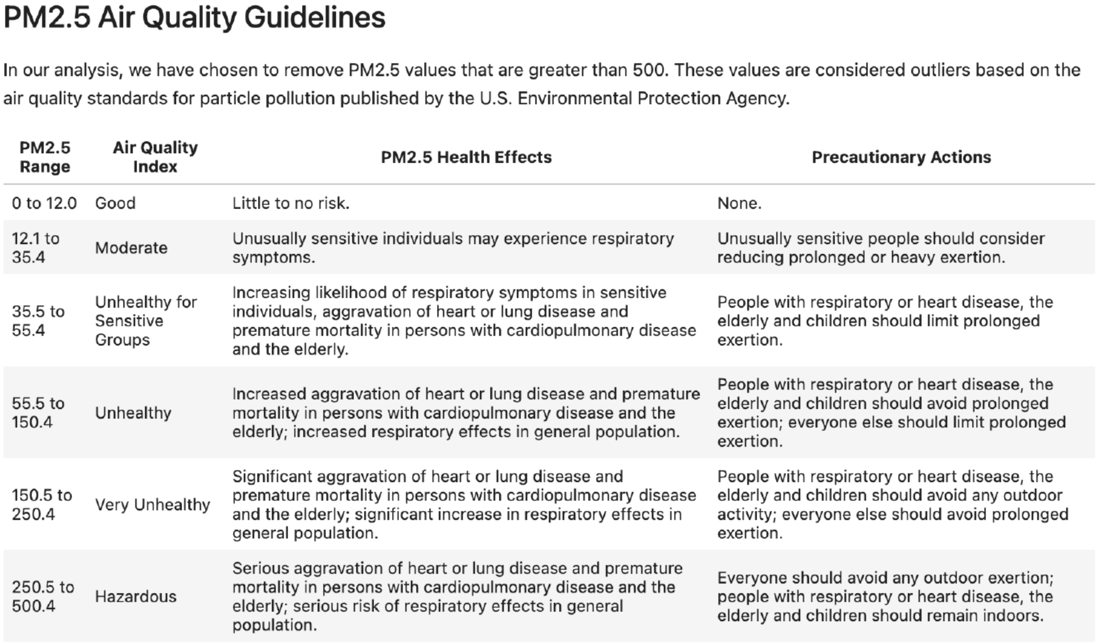
Listing 2: Data Cleaning

```
df_merged = df_merged[(df_merged[”Raw-Conc.”] > 0) &
          (df_merged[”Raw-Conc.”] < 500)]

df_merged['Datetime'] = pd.to_datetime(df_merged['Datetime'])

df_merged.set_index('Datetime', inplace=True)

df_merged = df_merged.resample('H').mean().fillna(method='ffill')
```



PM2.5 Air Quality Guidelines

In our analysis, we have chosen to remove PM2.5 values that are greater than 500. These values are considered outliers based on the air quality standards for particle pollution published by the U.S. Environmental Protection Agency.

PM2.5 Range	Air Quality Index	PM2.5 Health Effects	Precautionary Actions
0 to 12.0	Good	Little to no risk.	None.
12.1 to 35.4	Moderate	Unusually sensitive individuals may experience respiratory symptoms.	Unusually sensitive people should consider reducing prolonged or heavy exertion.
35.5 to 55.4	Unhealthy for Sensitive Groups	Increasing likelihood of respiratory symptoms in sensitive individuals, aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly.	People with respiratory or heart disease, the elderly and children should limit prolonged exertion.
55.5 to 150.4	Unhealthy	Increased aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; increased respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid prolonged exertion; everyone else should limit prolonged exertion.
150.5 to 250.4	Very Unhealthy	Significant aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; significant increase in respiratory effects in general population.	People with respiratory or heart disease, the elderly and children should avoid any outdoor activity; everyone else should avoid prolonged exertion.
250.5 to 500.4	Hazardous	Serious aggravation of heart or lung disease and premature mortality in persons with cardiopulmonary disease and the elderly; serious risk of respiratory effects in general population.	Everyone should avoid any outdoor exertion; people with respiratory or heart disease, the elderly and children should remain indoors.

4 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical step in understanding the underlying patterns and characteristics of the dataset. In this section, we present the key findings from the EDA, including the identification of seasonal trends and the examination of autocorrelation.

4.1 Seasonal Trends

Seasonal trends refer to regular patterns that repeat over a specific period, such as daily, weekly, or yearly cycles. In our dataset, we observed significant seasonal variations in PM2.5 levels, with higher concentrations during the winter months and lower levels in the summer. This pattern can be attributed to factors such as increased heating activities during the winter and meteorological conditions that affect pollutant dispersion.

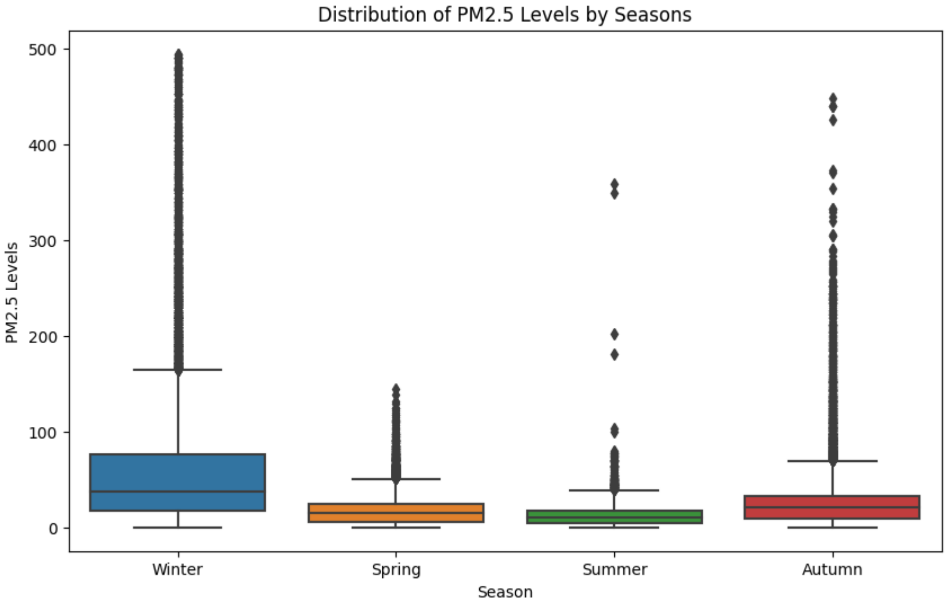


Figure 1: Seasonal Trends in PM2.5 Levels

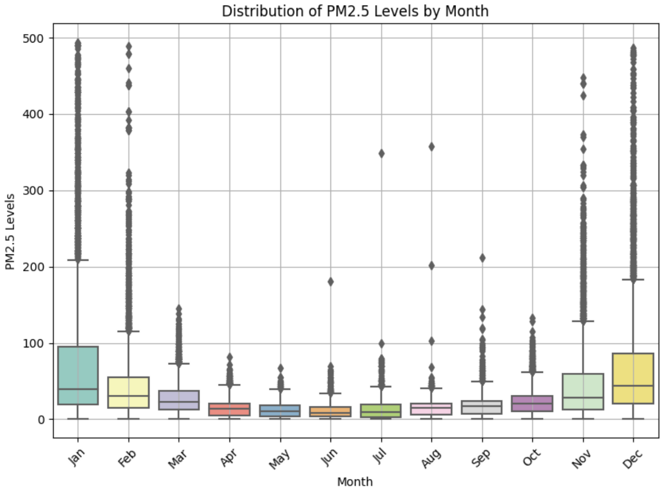


Figure 2: Seasonal Trends in PM2.5 Levels

4.2 Distribution Analysis

We conducted a distribution analysis to understand the spread and central tendency of PM2.5 concentrations. The distribution was found to be right-skewed, with a higher frequency of lower PM2.5 values and occasional spikes in pollution levels. This skewness indicates that while lower PM2.5 levels are common, there are instances of high pollution events that need to be addressed.

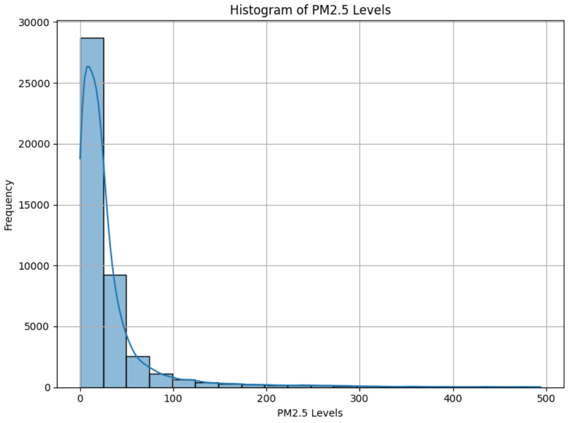


Figure 3: Distribution of PM2.5 Levels

4.3 Rolling Averages

A **rolling average** is a technique used to calculate the mean value of multiple overlapping subsets within a dataset. For instance, consider a scenario where I have daily income data for a shop. As long as the shop remains operational, I can compute a rolling average. On a given Friday, I might calculate the average income from Monday to Thursday. The following Monday, I would compute the average from Tuesday to Friday, and the subsequent day from Wednesday to Monday, continuing this process daily. These averages ”roll” forward, providing insight into how the data evolves over time rather than relying on a static snapshot. This method is particularly useful in data science for making accurate predictions about future data trends.

In this study, I applied rolling averages over a one-week period for PM2.5 levels to observe how the air quality changes over time. This approach helps in smoothing out short-term fluctuations and highlights longer-term trends in the data.

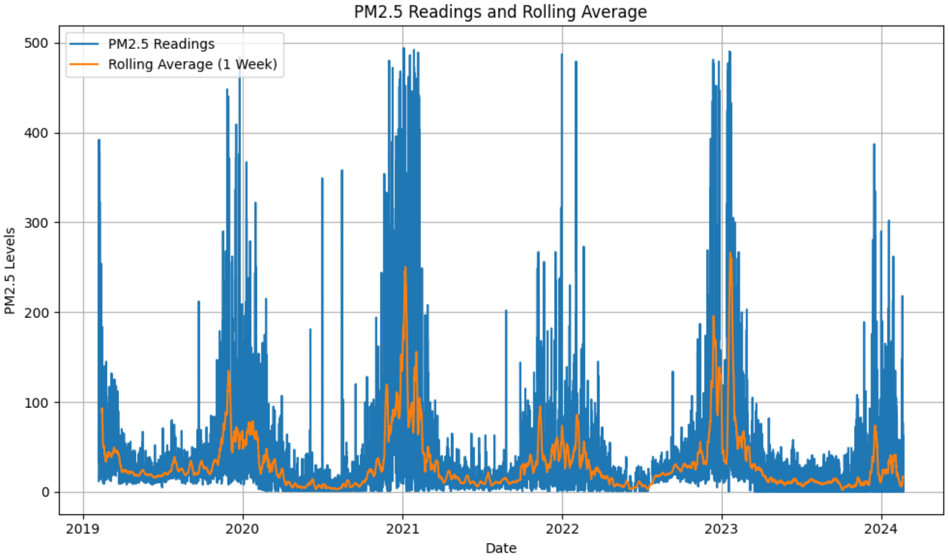


Figure 4: PM2.5 Readings and Rolling Average

4.4 Autocorrelation Examination

Autocorrelation measures the correlation of a time series with its own past values. We used autocorrelation plots to identify significant lags in the PM2.5 data, which indicated the presence of temporal dependencies that could be modeled using autoregressive techniques. The autocorrelation function (ACF) showed significant correlations at multiple lags, suggesting that past PM2.5 levels have a strong influence on future values. Additionally, we examined the partial autocorrelation function (PACF) to identify the direct effect of each lag on the current observation, which helps determine the appropriate lag order for autoregressive models.

Listing 3: ACF plot

```
# Plot ACF plot
fig, ax = plt.subplots(figsize=(15, 6))
plot_acf(y, ax=ax)
plt.xlabel("Lag␣[hours]")
plt.ylabel("Correlation␣Coefficient");
```

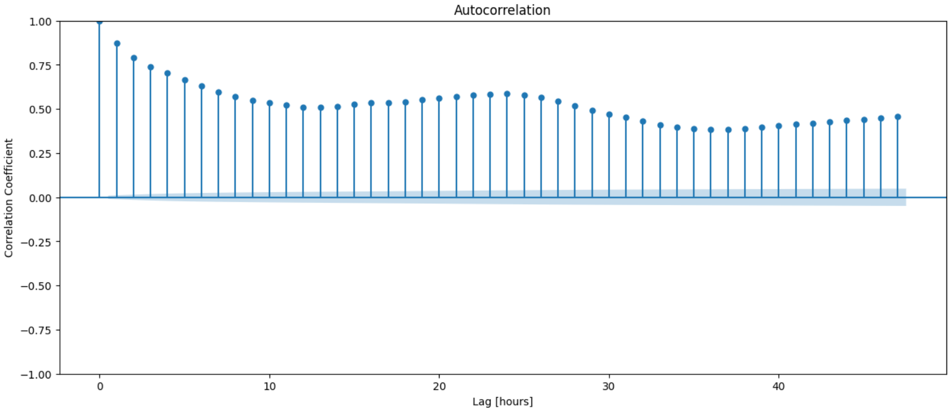


Figure 5: Autocorrelation Plot of PM2.5 Levels

Sometimes, we want to actually see how autocorrelations change over time, which means we need to think of them as functions. When we create a visual representation of an autocorrelation function (ACF), we're making an **ACF plot**.

Listing 4: PACF plot

```
# Plot PACF plot
fig, ax = plt.subplots(figsize=(15, 6))
plot_pacf(y, ax=ax)
plt.xlabel("Lag [hours]")
plt.ylabel("Correlation Coefficient");
```

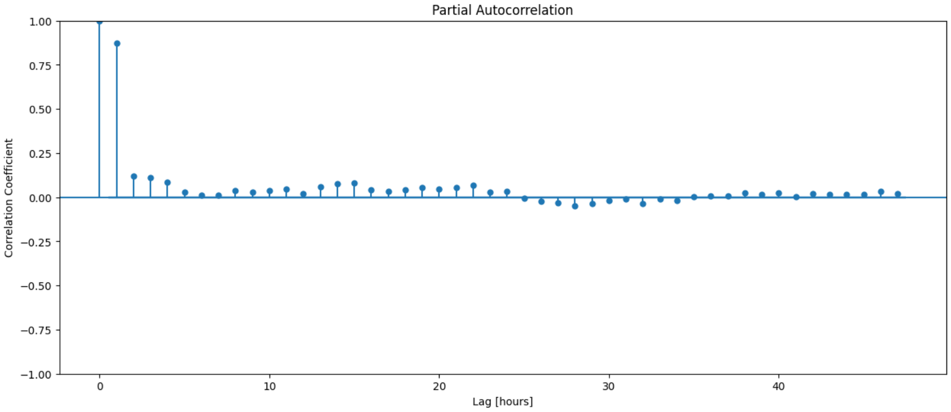


Figure 6: Partial Autocorrelation Plot of PM2.5 Levels

Autocorrelations take into account two types of observations. **Direct observations** are the ones that happen exactly at our chosen time-step interval; we might have readings at one-hour intervals starting at 1:00.**Indirect observations** are the ones that happen between our chosen time-step intervals, at time-steps like 1:38, 2:10, 3:04, etc. Those indirect observations might be helpful, but we can't be sure about that, so it's a good idea to strip them out and see what our graph looks like when it's only showing us direct observations.

An autocorrelation that only includes the direct observations is called a **partial autocorrelation**, and when we view that partial autocorrelation as a function, we call it a **PACF**.

PACF plots represent those things visually. We wanted to compare our ACF and PACF plots to see which model best describes our time series. If the ACF data drops off slowly, then that's a better description; if the PACF falls off slowly, then that's a better description.

5 Model Development

In this section, we describe the development of two models: a simple baseline model and an autoregressive (AR) model. The baseline model serves as a benchmark for evaluating the performance of the AR model.

5.1 Baseline Model

The baseline model predicts the PM2.5 levels using the mean of the historical data. This simple approach provides a reference point for assessing the effectiveness of more complex models.

Listing 5: Baseline Model

```
baseline_predictions = [train.mean()] * len(test)
baseline_mae = mean_absolute_error(test, baseline_predictions)
```

5.2 Autoregressive (AR) Model

The autoregressive (AR) model predicts future values based on a linear combination of past values. We selected a lag of 26, which was determined based on the autocorrelation analysis.

Listing 6: Autoregressive Model Training

```
from statsmodels.tsa.ar_model import AutoReg
from sklearn.metrics import mean_absolute_error

y = df_merged[”Raw␣Conc.”]

train_size = int(len(y) * 0.8)
train, test = y[0:train_size], y[train_size:]

model = AutoReg(train, lags=26)
model_fitted = model.fit()

predictions = model_fitted.predict(start=len(train),
end=len(train)+len(test)−1, dynamic=False)
mae = mean_absolute_error(test, predictions)
```

6 Walk-Forward Validation (WFV)

Our predictions lose power over time because the model gets farther and farther away from its beginning. But what if we could move that beginning forward with the model? That's what walk-forward validation is. In a walk-forward validation, we re-train the model at each new observation in the dataset, dropping the data that's the farthest in the past. Let's say that our prediction for what's going to happen at 12:00 is based on what happened at 11:00, 10:00, and 9:00. When we move forward an hour to predict what's going to happen at 1:00, we only use data from 10:00, 11:00, and 12:00, dropping the data from 9:00 because it's now too far in the past.

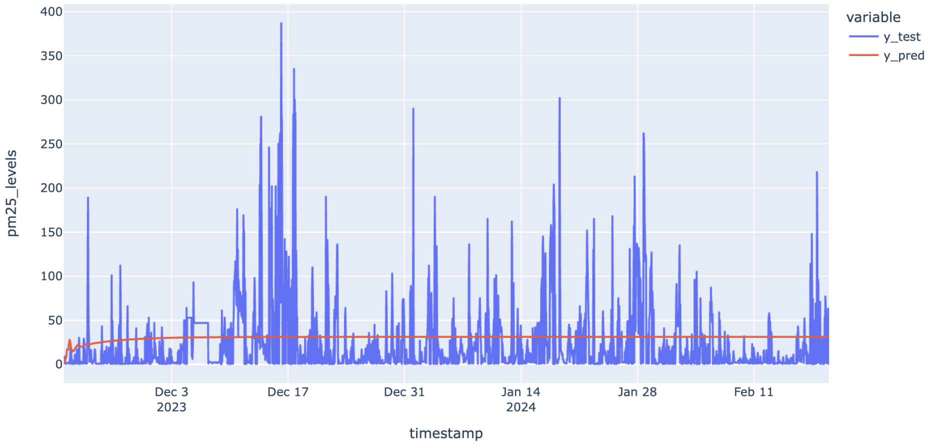


Figure 7: Before applying WFV

Walk-forward validation is a technique used to evaluate the performance of a model under realistic conditions by iteratively retraining the model using expanding data windows. This approach ensures that the model adapts to new data over time.

Listing 7: Walk-Forward Validation

```
history = [x for x in train]
predictions_wfv = list()

for t in range(len(test)):
    model = AutoReg(history, lags=26)
    model_fitted = model.fit()
    yhat = model_fitted.predict(start=len(history),
    end=len(history), dynamic=False)
    predictions_wfv.append(yhat)
    history.append(test[t])

mae_wfv = mean_absolute_error(test, predictions_wfv)
```

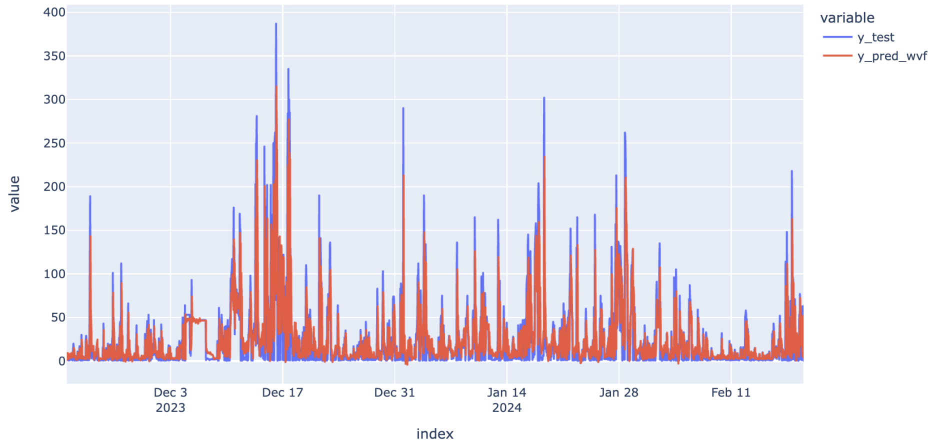


Figure 8: After applying WFV

7 Results

In this section, we present the results of our analysis, including the performance of the baseline model, the AR model, and the AR model with walk-forward validation (WFV). We also provide visualizations to compare the predicted and actual PM2.5 levels.

7.1 Performance Comparison

The performance of the models was evaluated using the Mean Absolute Error (MAE) metric. The table below summarizes the results, indicating that the AR model with WFV significantly outperformed the baseline model and the AR model without WFV.

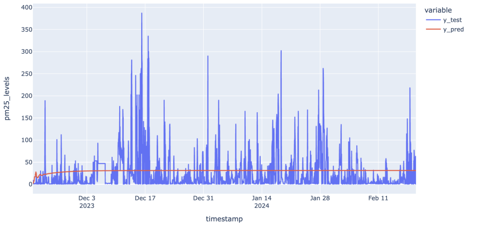
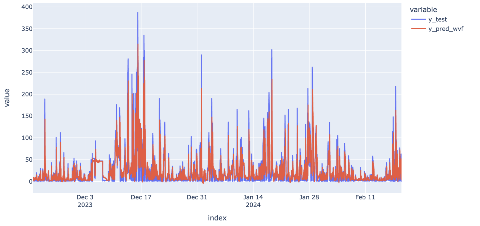
Model	**MAE**	**Plot**
Baseline	30.0	
Test	25.4	
Test-WFV	18.7	

Table 1: Performance Comparison of Models with Actual vs Predicted Plots

8 Conclusion

This study aimed to analyze the air quality in Bishkek, focusing specifically on PM2.5 levels, using a variety of statistical and machine learning techniques. The analysis utilized data collected from multiple years, covering a period from 2019 to 2024. Through this comprehensive time series analysis, several key insights and outcomes were achieved.

8.1 Data Preparation and Exploration

Initially, the dataset was thoroughly cleaned to ensure accuracy and reliability. Outliers, specifically PM2.5 values greater than 500, were removed based on the air quality standards set by the U.S. Environmental Protection Agency. Exploratory Data Analysis (EDA) revealed critical insights into the seasonal patterns, distribution, and autocorrelation of PM2.5 levels. Visualizations such as box plots, histograms, and lag plots helped in understanding the underlying patterns in the data.

8.2 Model Building and Evaluation

A baseline model was established using the mean of the training data, which resulted in a Mean Absolute Error (MAE) of 27.02. To improve upon this, an Autoregressive (AR) model with a lag of 26 was implemented. The AR model demonstrated a low training MAE of 9.55 but exhibited higher error on the test set with an MAE of 30.04, indicating potential overfitting issues.

8.3 Implementation of Walk-Forward Validation

To address overfitting and enhance the model's generalizability, Walk-Forward Validation (WFV) was employed. This method involved iteratively retraining the model on expanding data windows. The application of WFV significantly improved the model's performance, reducing the test MAE to 18.72. This improvement highlights the importance of dynamic validation techniques in time series forecasting.

8.4 Visualization and Communication of Results

The results were effectively communicated through various visualizations, which included plots of actual versus predicted values before and after applying WFV. These visualizations, created using tools like Plotly Express, were crucial in demonstrating the model's accuracy and the impact of WFV on prediction performance.

8.5 Practical Implications and Recommendations

The study underscores the critical role of continuous model refinement and validation in forecasting air quality trends. Techniques such as Walk-Forward Validation are essential for building robust predictive models that can adapt to new data and changing patterns over time. For policymakers and environmental agencies, these models can provide valuable forecasts that inform strategic decisions to improve air quality and public health.

8.6 Future Work

Future research could expand on this study by incorporating additional variables that may influence air quality, such as weather conditions, traffic data, and industrial activities. Moreover, exploring more advanced machine learning models and deep learning techniques could potentially yield even more accurate predictions. Regular updates to the dataset and ongoing validation are recommended to maintain the model's relevance and accuracy over time.

In conclusion, this study has successfully demonstrated the application of time series analysis and machine learning techniques to predict PM2.5 levels in Bishkek. The insights gained from this research can contribute to more effective air quality management and policy-making, ultimately aiming for a healthier environment for the residents of Bishkek.

9 References

1. Zhang, L., Wang, X., Zhang, X., & Luo, W. (2018). Predicting PM2.5 concentration using a combination of ARIMA and SVM models. *Atmospheric Pollution Research*, 9(4), 725-734.

2. Grange, S. K., Lewis, A. C., Carslaw, D. C. (2018). Source apportionment advances in air pollution analysis. *Atmospheric Environment*, 192, 116-120.

3. Applied Data Science Lab, WorldQuant University. (2023). AIR QUALITY IN NAIROBI. Retrieved from `https://www.worldquantuniversity.org/your-path-to-success/applied-data-science-in-python`.