Assessing Air Quality in Lombardia, Italy through Time Series Analysis

a.y. 2023/2024

Giovanni Costa
880892@stud.unive.it

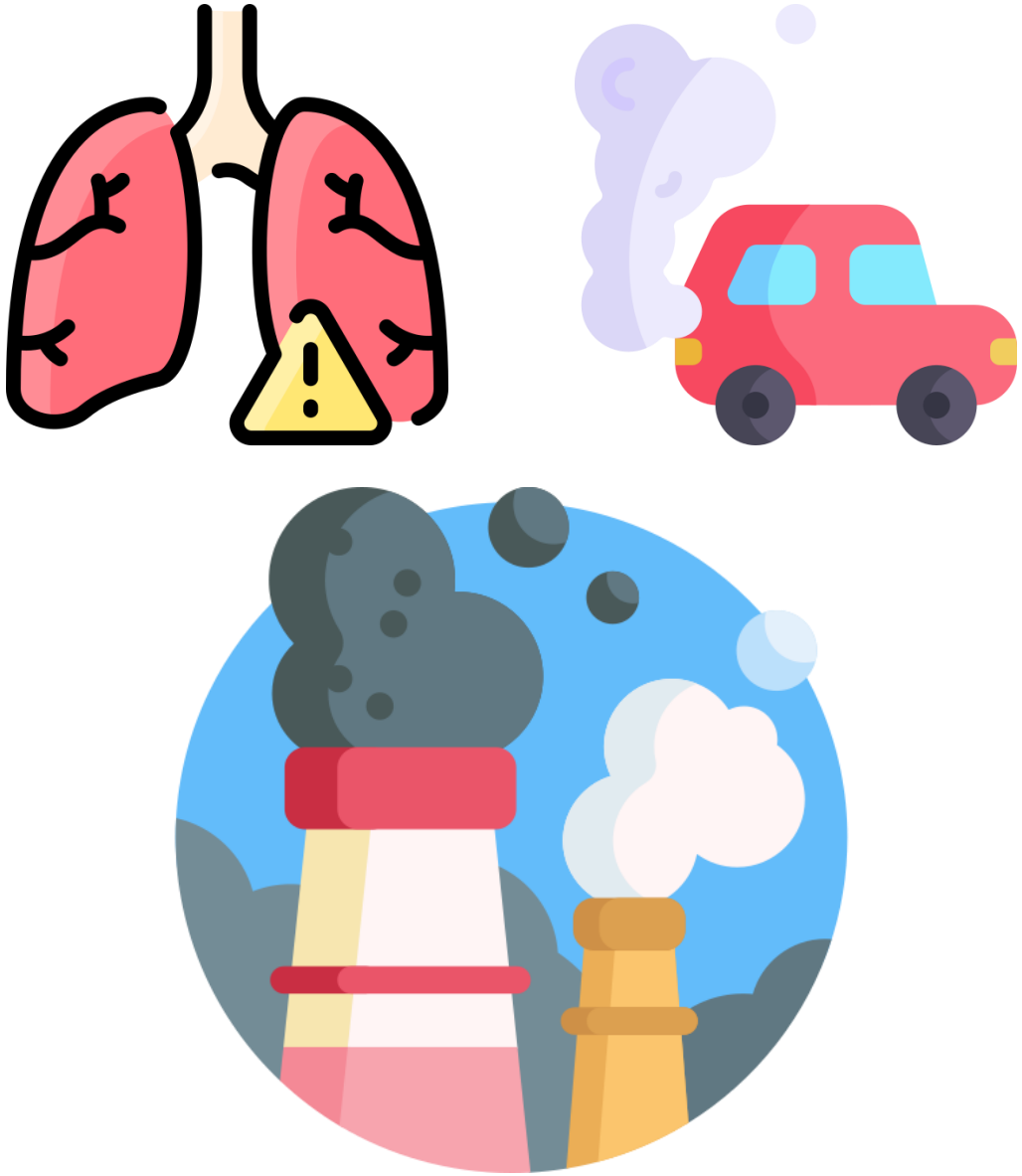
Presentation outline

- Introduction
- Dataset description
- Long-term data analysis
- Models implementation
- Forecasting assesment
- Conclusions

Introduction

Motivations

- Evaluate air quality in Lombardia using open data derived from various stations across the region
- Underscore the significance of air quality on public health
- Identify possible correlations between pollutants
- Analyze trends and patterns in air quality data over time to enable prevention measures



Pollutants

- **PM10 (Particulate Matter 10)**: Airborne particles ≤10 µm in diameter from sources like construction, road dust, and industrial emissions. Health effects include respiratory infections, lung inflammation, and cardiovascular diseases. Long-term exposure reduces lung function and increases mortality
- **NO_2 (Nitrogen Dioxide)**: A reddish-brown gas from fossil fuel combustion. It irritates the respiratory system, worsens asthma, and lowers lung function. NO_2 also leads to ozone and fine particulate matter formation, further harming health

PM10 as a Key Indicator: Selected for analysis due to strong link to adverse health effects. Other pollutants like Benzene, CO, and NO_2 influence PM10 levels, making them critical predictors in the study

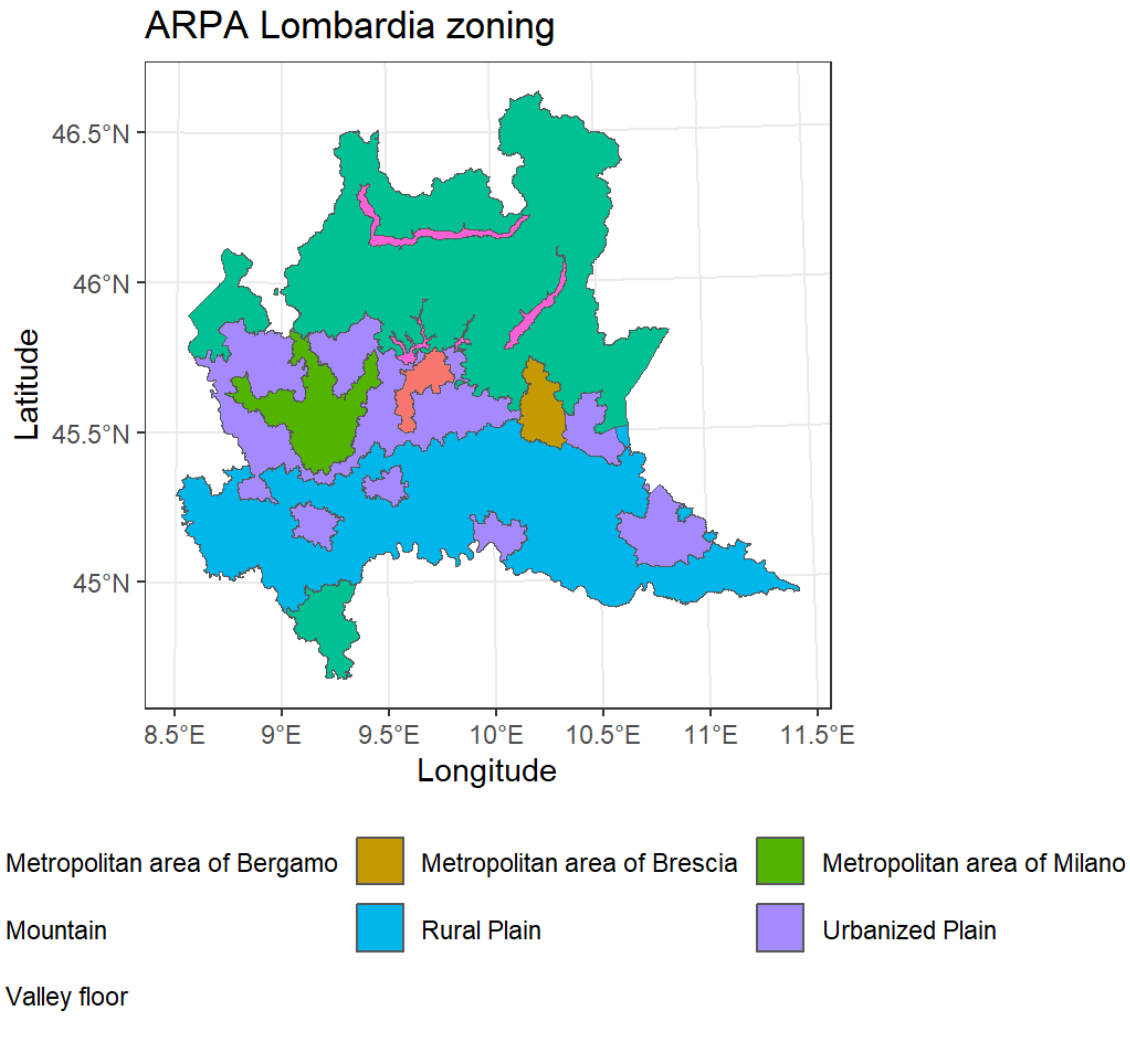
Dataset description

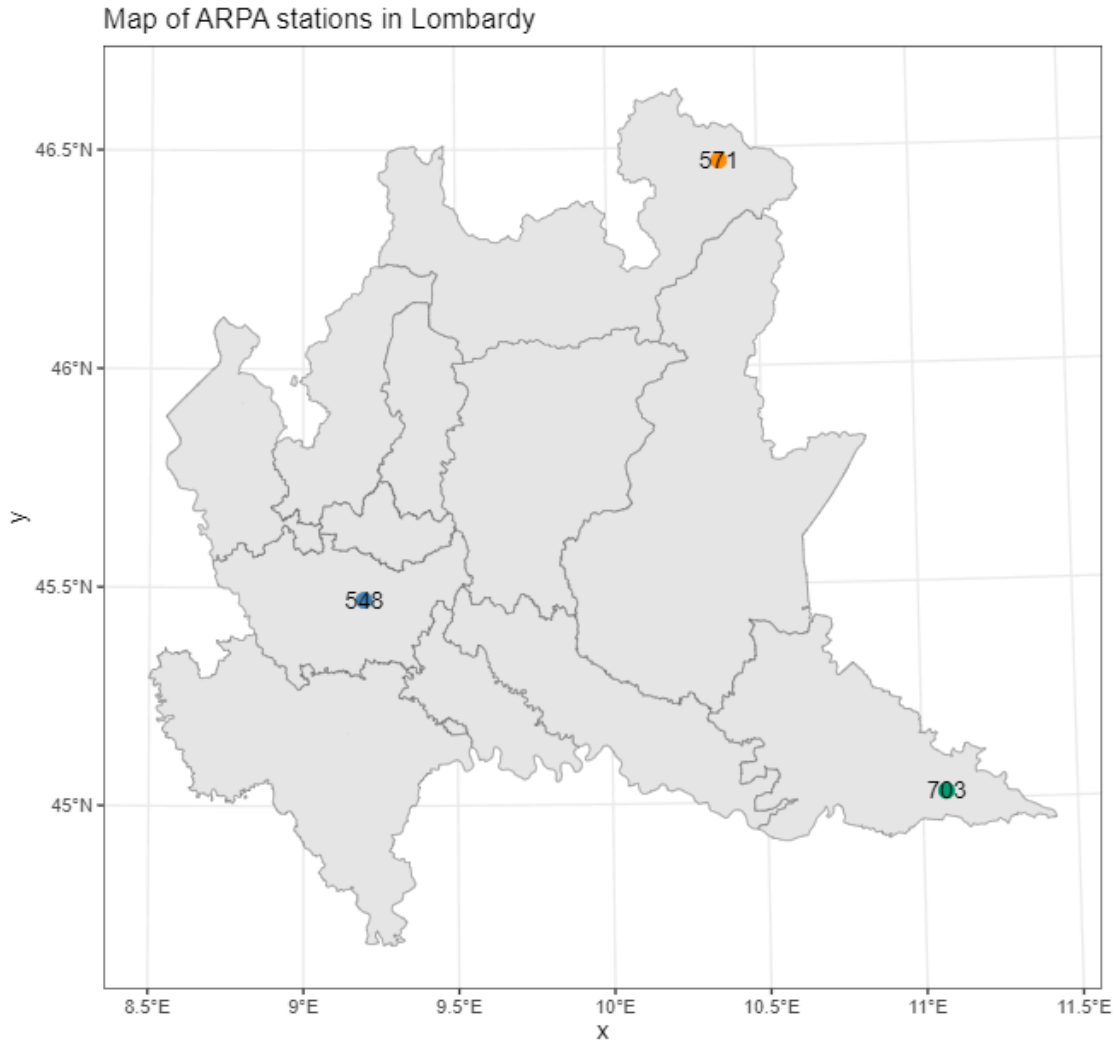
Air quality data

- **Daily observations** of pollutants (e.g. Benzene, CO, NO2, NOx, and PM10) coming from 3 different stations placed in different positions
- Data interval are **from January 1, 2014, to December 31, 2023**
- For **long-term** analysis and clearer understanding of the underlying trends data are monthly averaged
- For model implementation and **short-term** prediction original daily data are used

- Stations
  - Milano v.Senato (Station 548): **Metropolitan area**
  - Bormio v.Monte Braulio (Station 571): **Mountain zone**
  - Schivenoglia v.Malpasso (Station 703): **Rural plain**

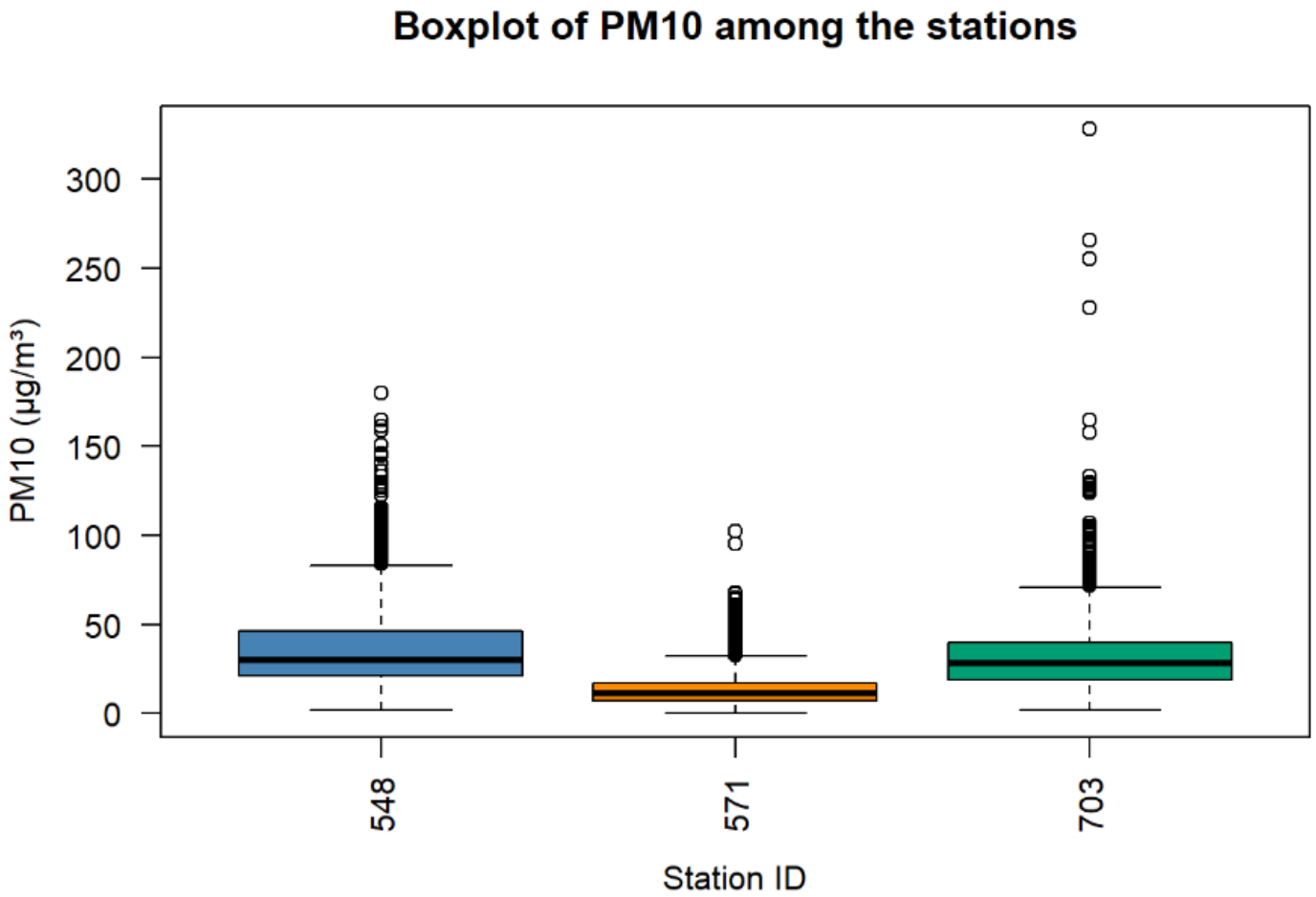
- Dataset includes station details like sensor ID, pollutant types, and location information


ARPA Lombardia zoning
46.5°N
46°N
45.5°N
45°N
Latitude
8.5°E
9°E
9.5°E
10°E
10.5°E
11°E
11.5°E
Longitude
Metropolitan area of Bergamo
Metropolitan area of Brescia
Metropolitan area of Milano
Mountain
Rural Plain
Urbanized Plain
Valley floor


Map of ARPA stations in Lombardy
46.5°N
46°N
45.5°N
45°N
y
571
548
703
8.5°E
9°E
9.5°E
10°E
10.5°E
11°E
11.5°E
x

PM10 statistics

- Median values for Milan station 548 and Schivenoglia station 703 are quite similar
- All the stations present observations very distant from the others (in particular station 703)
- PM10 concentrations can be higher in rural areas compared to urban metropolitan areas for several reasons e.g. agricultural activities and rural dust



Standard regulatory

- Annual Average acceptable level is 40 μg/m^3, to protect public health
- PM10 **only slightly above the limit** across the years

NA gaps

- Sequences of one or more consecutive missing values.
- Most frequent gap length is one
- Average gap size is relatively small
- **Simple imputation** techniques should be **reasonably accurate**
- Linear interpolation is used

Mean PM10 values by year

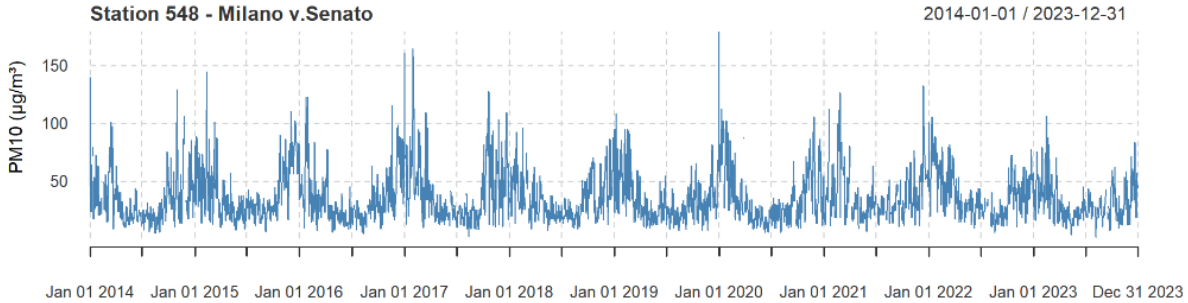
Year	Milano v.Senato	Bormio v.Monte Braulio	Schivenoglia v. Malpasso
2014	34.28	15.82	31.93
2015	39.84	15.10	35.47
2016	35.28	14.25	32.49
2017	40.43	12.50	36.77
2018	34.31	12.80	29.61
2019	33.95	11.56	30.43
2020	36.49	11.65	31.21
2021	36.83	12.84	29.38
2022	38.96	12.30	33.57
2023	31.68	11.29	28.66

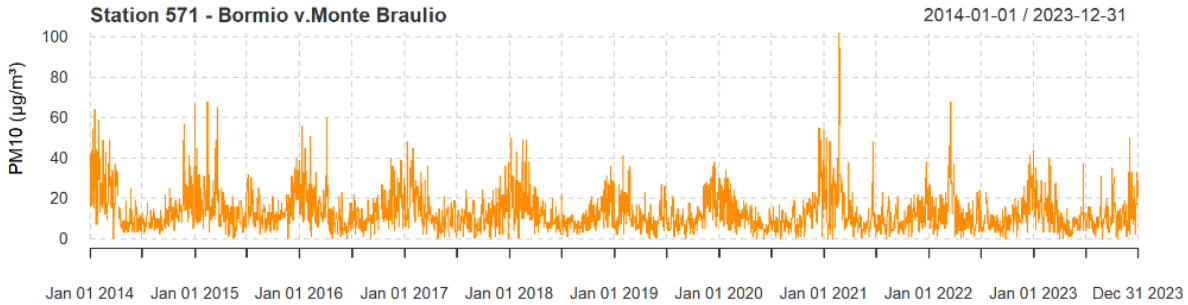
Missing values statistics

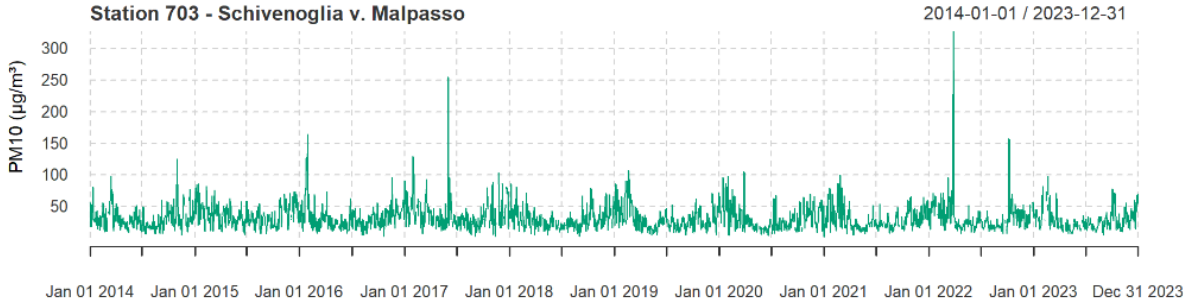
	Length TS	Number NAs	Number Gaps	Average Gap Size	Percentage NAs	Longest NA gap	Most frequent gap size
Station 548	3652	161	79	2.038	4.41%	8	1
Station 571	3652	48	29	1.655	1.31%	9	1
Station 703	3652	246	135	1.822	6.74%	6	1

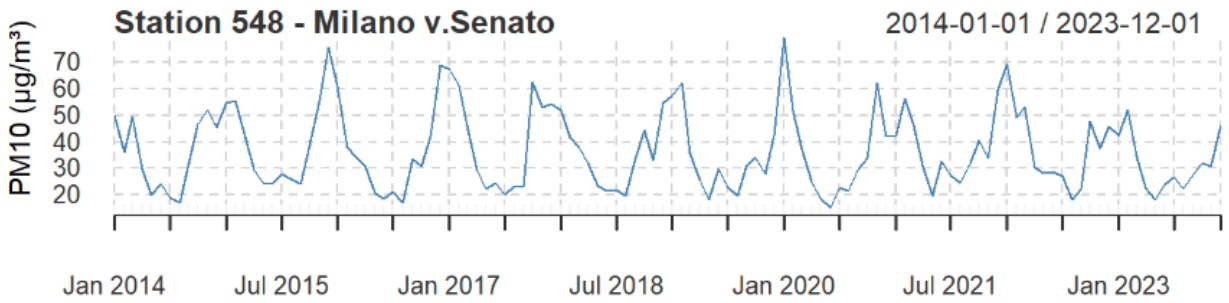
Long-term data analysis

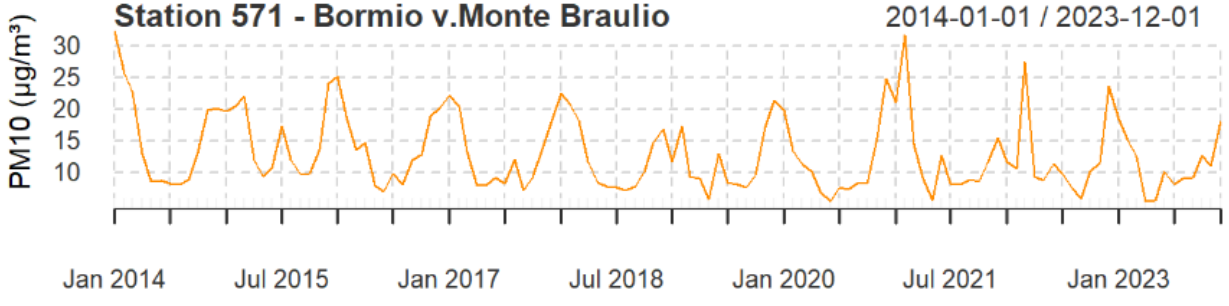
Time series data

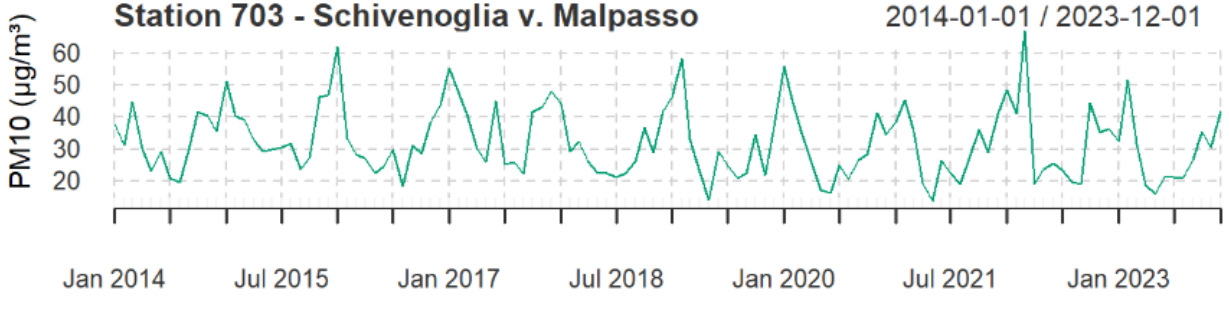






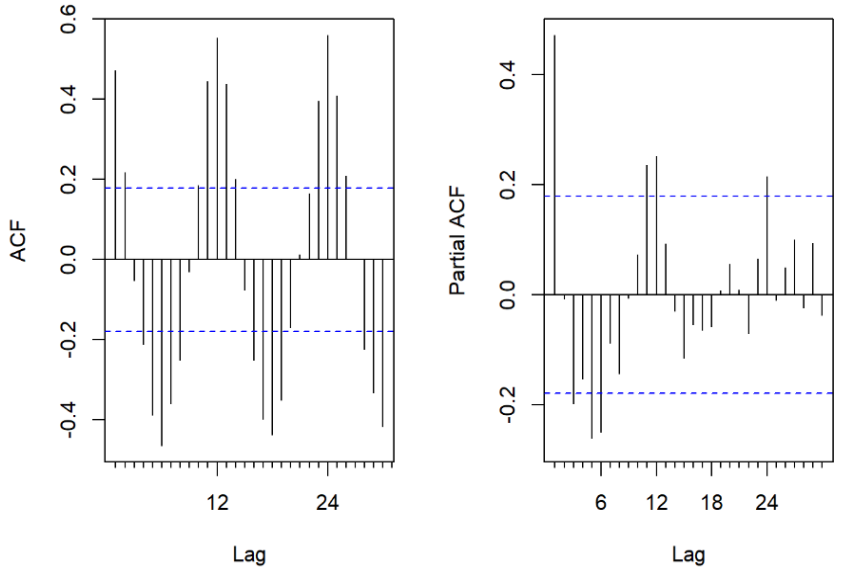




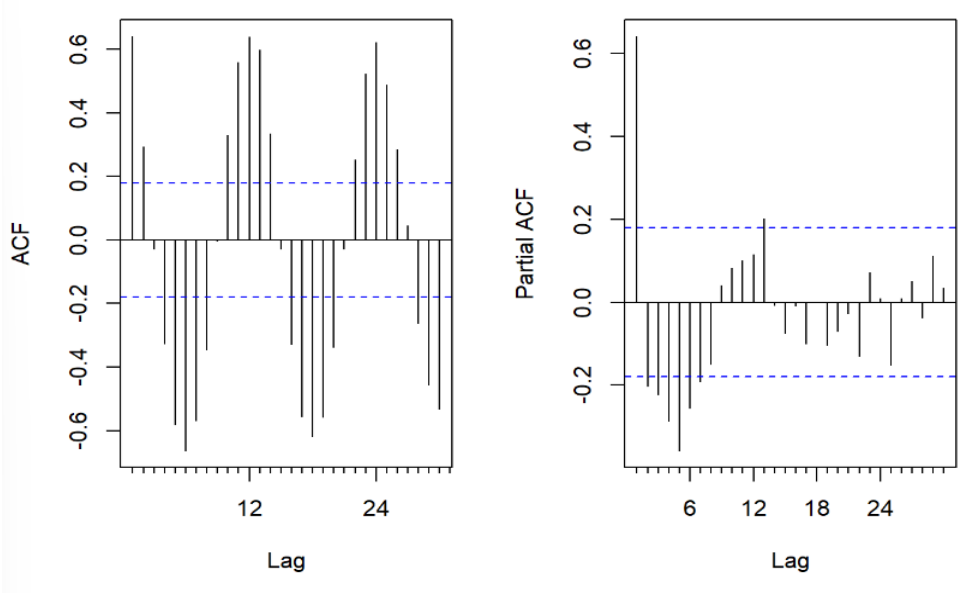


- ACF plots reveal a sinusoidal pattern across all stations. Peaks every six lags, suggesting a **recurring pattern** every six months
- PACF plots show spikes around the same period, indicating significant information during these intervals
- Time series is non-stationary with reasonable confidence

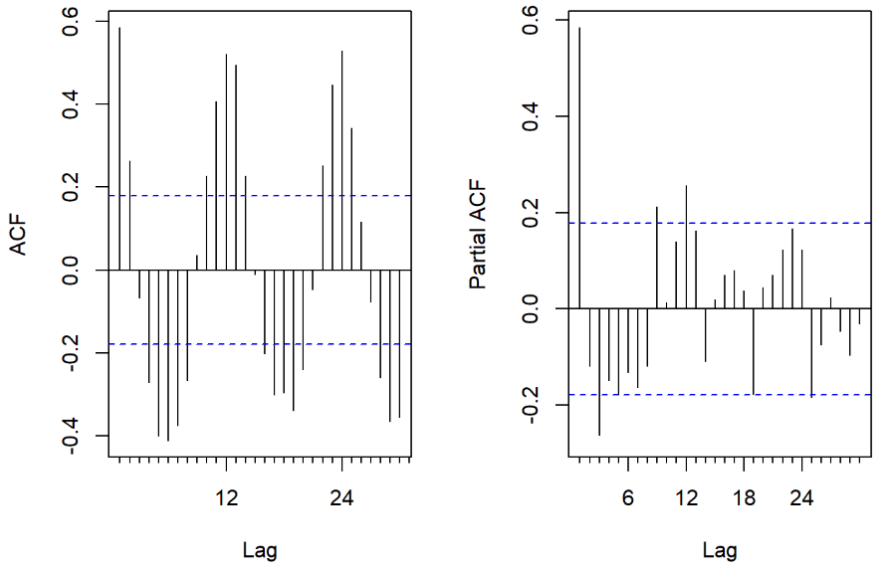
Station 703



Station 548



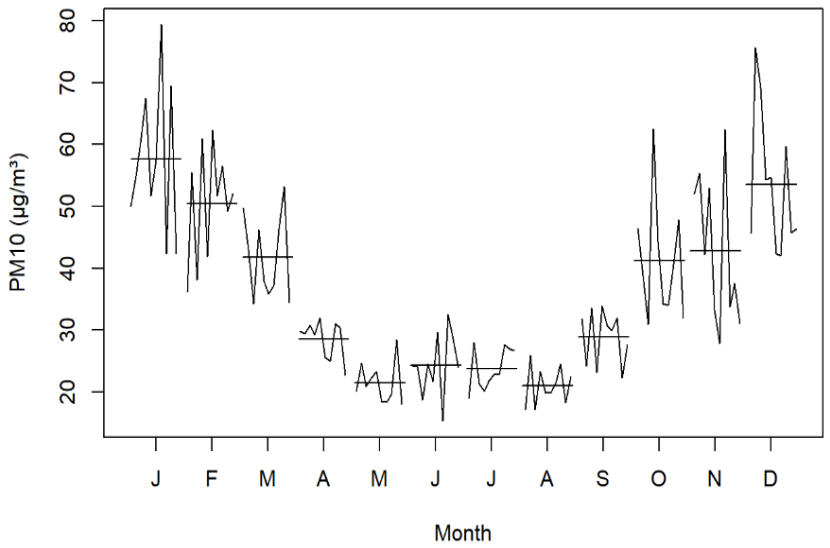
Station 571



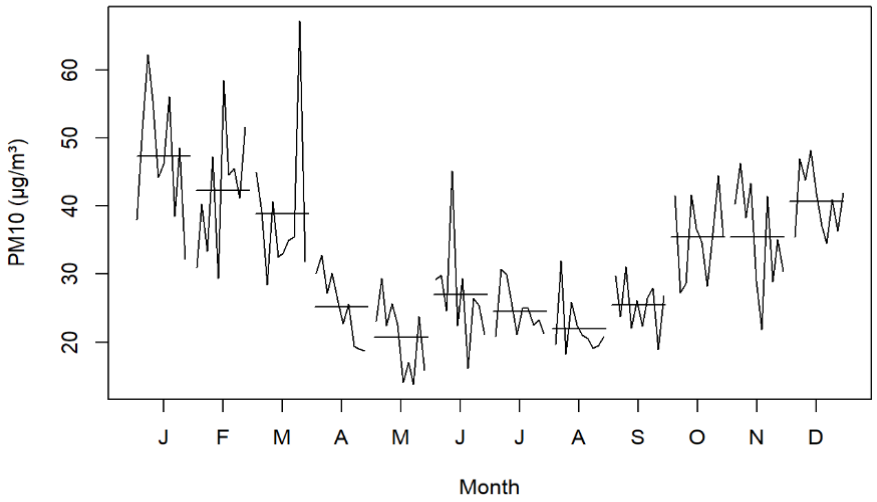
Monthplot visualization

- The monthplot visualizes monthly patterns by displaying average values for each month
- It helps identify seasonal trends, with **higher values in winter and lower in summer**
- **Significant variability** is observed throughout the years across all stations, indicating consistent yet varied seasonal fluctuations

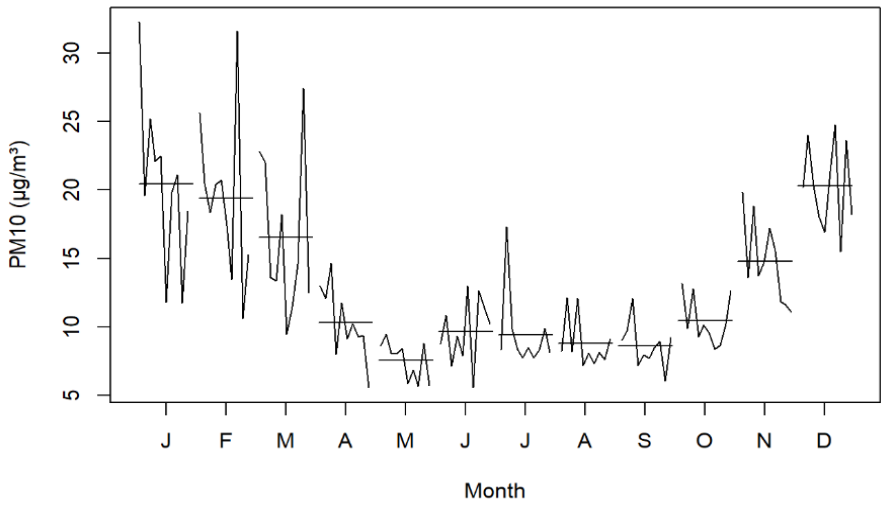
Station 548



Station 703



Station 571



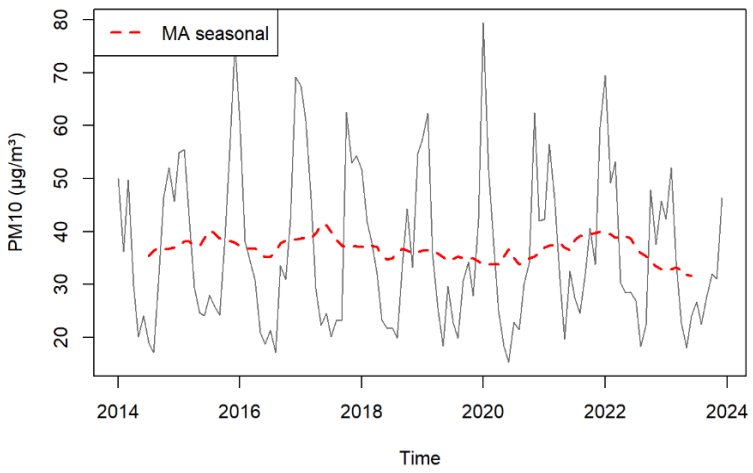
Trend estimation

- Moving average filter for seasonal data is used to estimate the trend-cycle

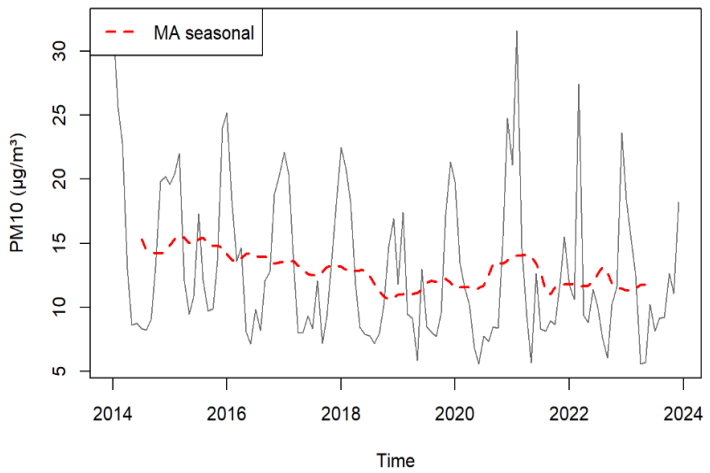
$$\widehat{f_t} = \frac{0.5 \cdot y_{t-p/2} + y_{t-p/2+1} + \cdots + y_0 + \cdots + y_{t+p/2-1} + 0.5 \cdot y_{t+p/2}}{p}$$

- The overall trend shows a slow decrease in PM10 levels

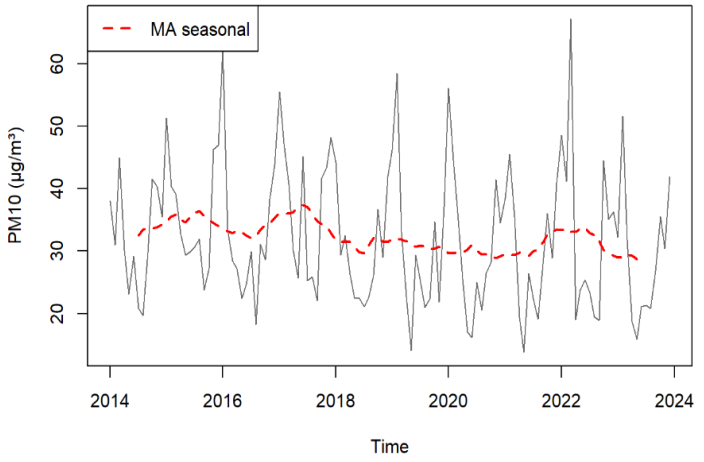
Station 548



Station 571



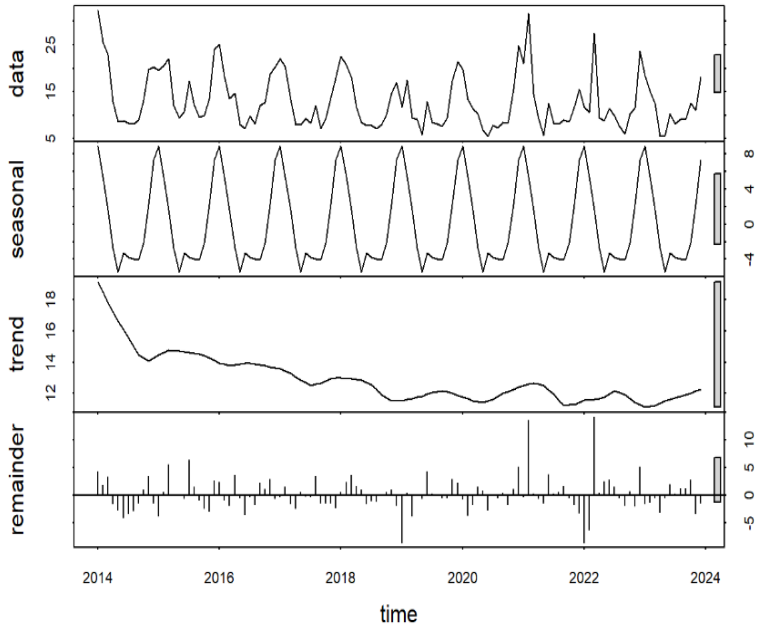
Station 703



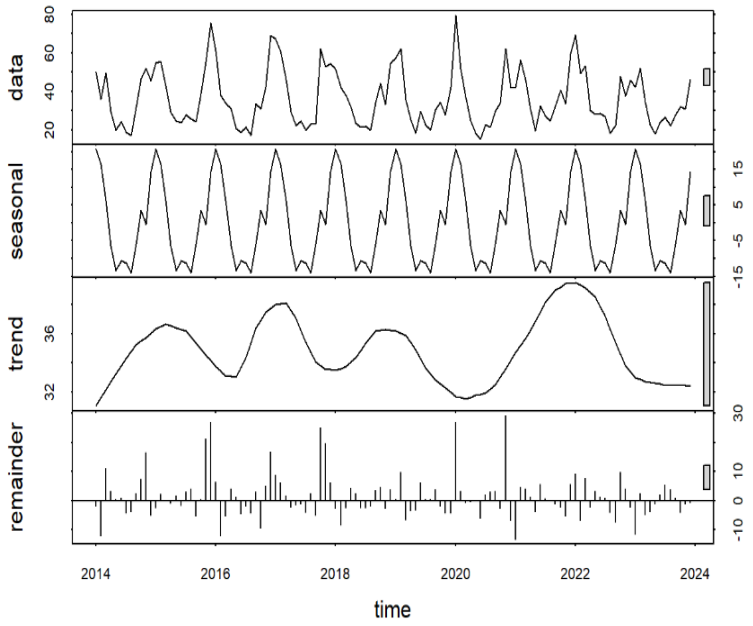
STL decomposition

- **Strong seasonal component** and an overall **decreasing trend**
- Station 548 shows significant PM10 peaks at the end of 2021 and throughout 2022
- The Ljung-Box test confirms the validity of the decompositions for most stations
- During **COVID-19 restrictions** (2020-2022), average PM10 levels did **not significantly decrease**
- Some stations, like Station 571 in Bormio, saw increased PM10 levels in 2021. This suggests that non-mobility-related sources (e.g. industrial activities, local emissions) contributed to sustained PM10 concentrations during the pandemic

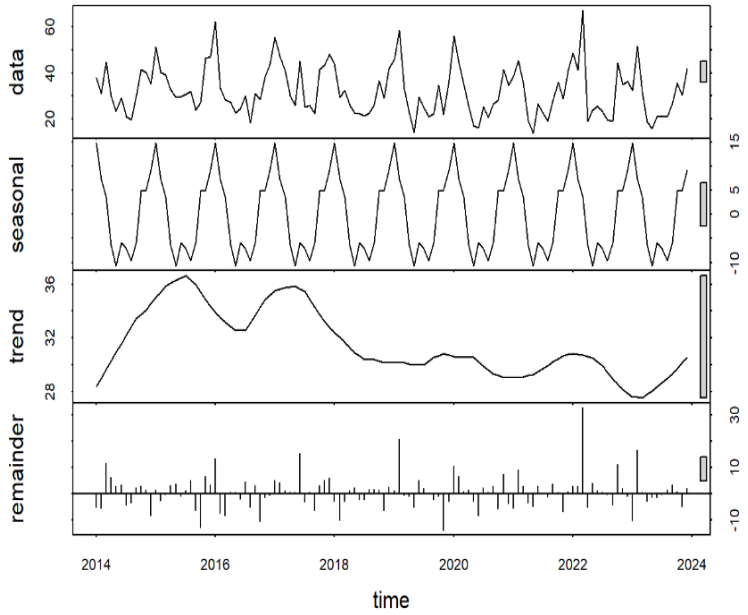
Station 571



Station 548



Station 703

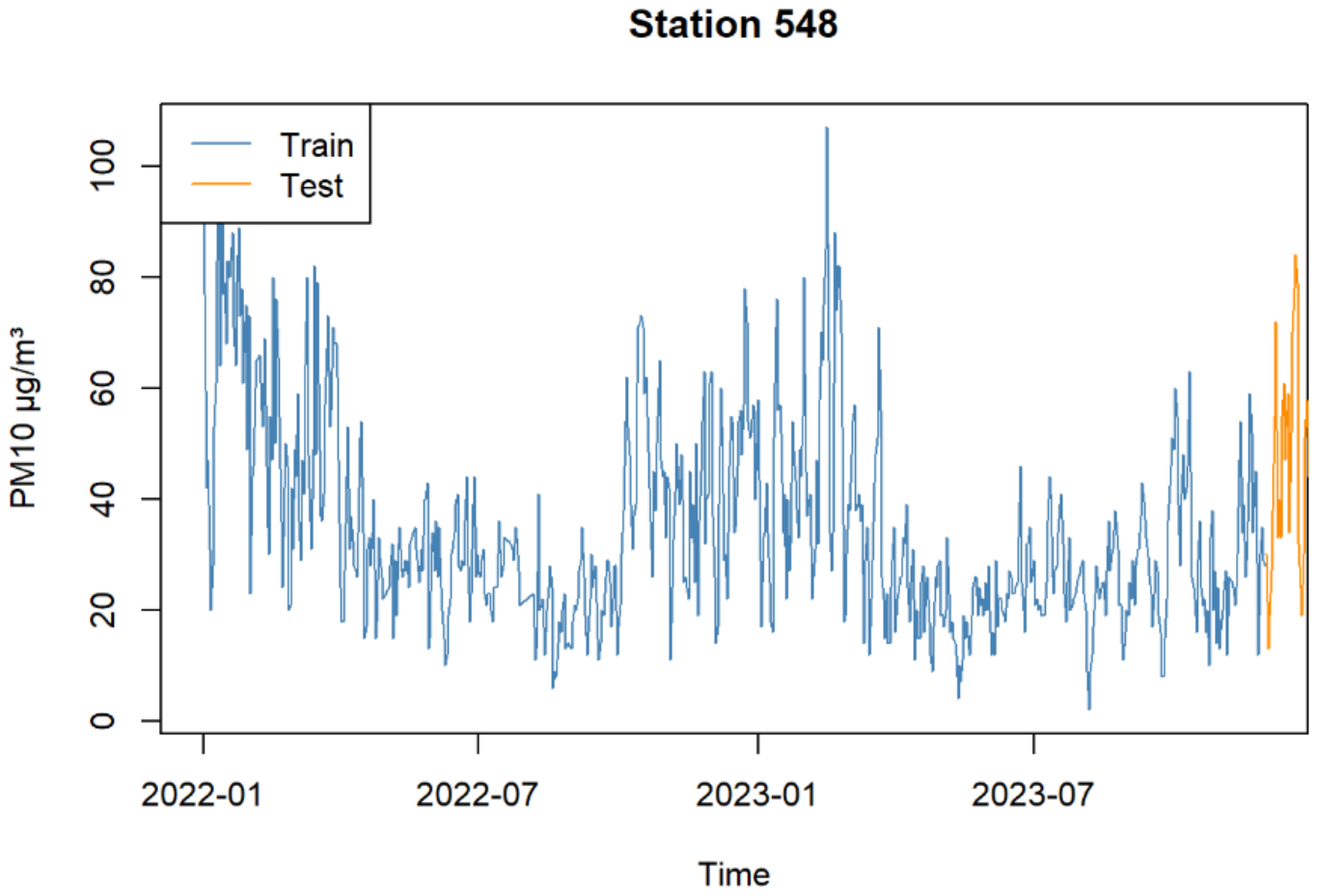


Models implementation

Data partitioning

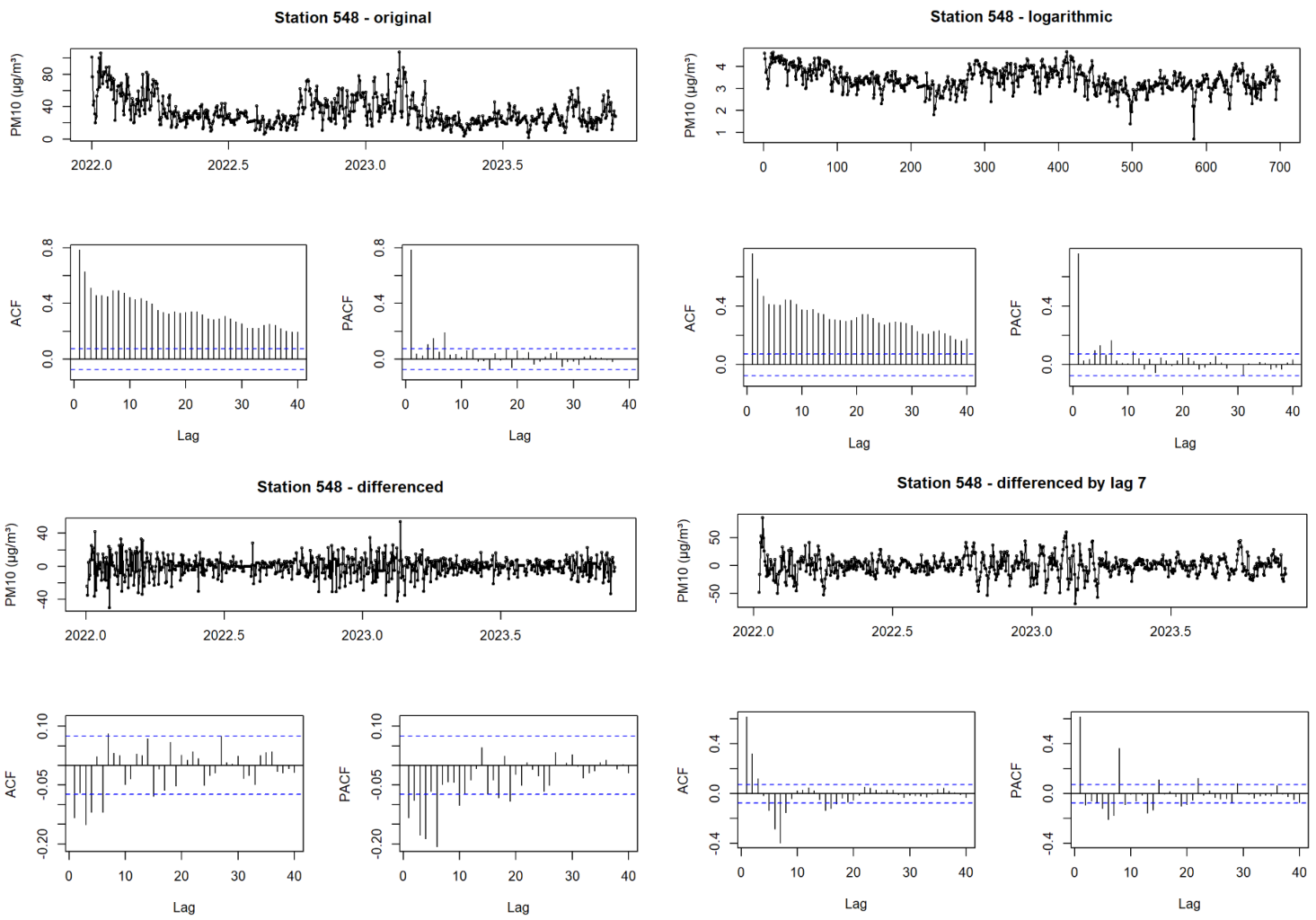
- The analysis covers the period **from January 1, 2022, to December 31, 2023**, to simplify visualization and focus on recent daily data
- The chosen time window at the end of the COVID-19 period does not affect the PM10 pattern
- The time series is divided into training and test sets, with the test period from December 1, 2023, to the end of the study
- A one-month test set is used to evaluate the model over a longer forecast horizon

Examples are shown only for station 548 for convenience



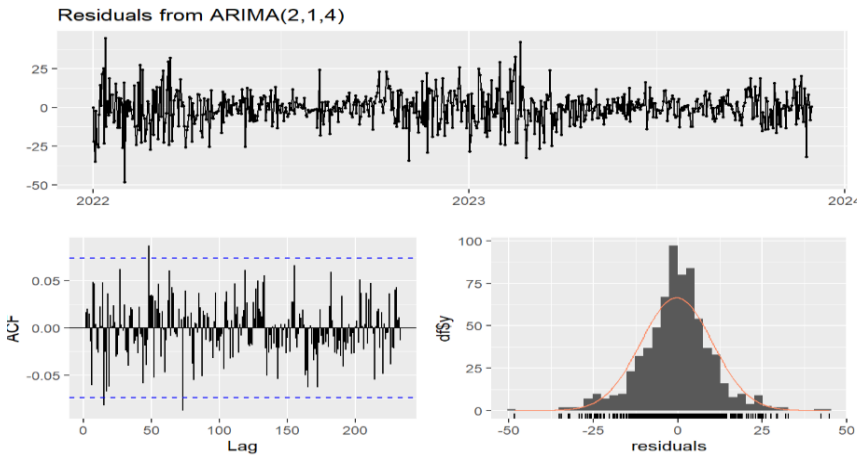
Data transformations

- The time series is not stationary, as shown by the slow decay of lags in the ACF plot
- Logarithmic transformation does not significantly improve stationarity
- Differencing improves stationarity but differencing by lag 7 or by time series frequency introduces artificial seasonal patterns
- Differencing may be impractical and risky, as it ignores time domains like months or weeks and does not account for variability across different years

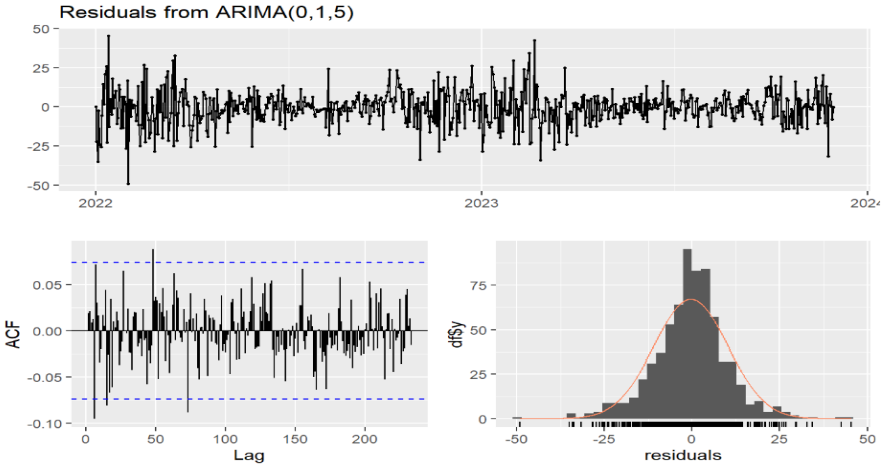


ARIMA models

- An ARIMA(0,1,5) is fitted based on the ACF and PACF plots of the differenced time series
- The ARIMA model is compared with an automatically selected model
- The automatically selected model shows **preferable residuals**
- Automatically selected models consistently perform slightly better than manually selected one



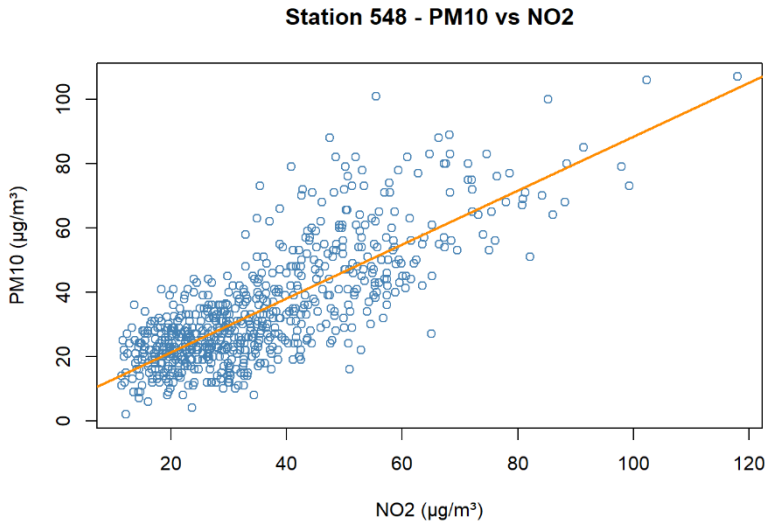
Ljung-Box test p-value: 1

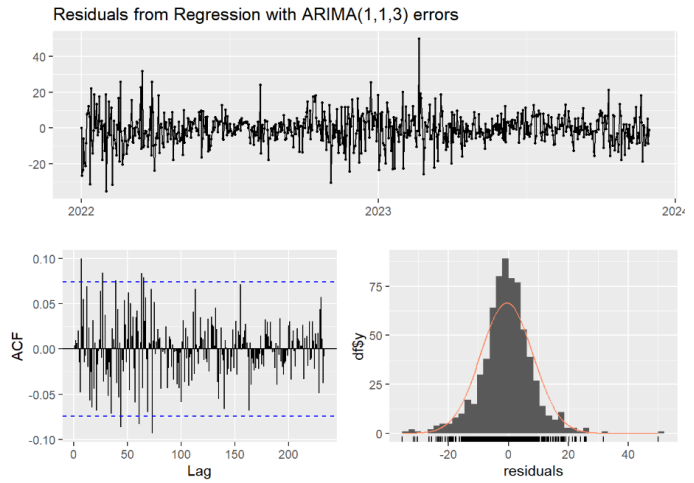


Ljung-Box test p-value: 0.9

Dynamic regression model

- Including other pollutants in PM10 prediction is explored, focusing on NO2
- NO_2 is a key precursor for particulate matter formation
- Missing values are imputed using linear interpolation
- A linear relationship is evident with NO_2 being a significant predictor, though residuals in linear regression do not meet white noise assumptions
- Implemented regression with ARIMA errors

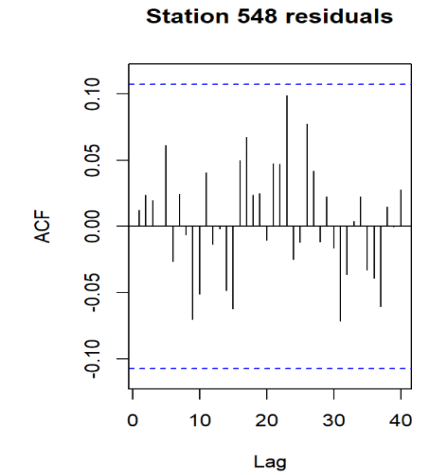




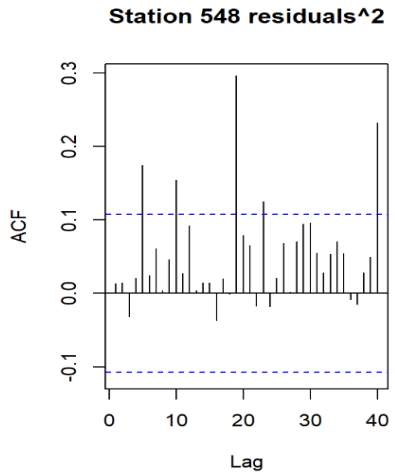
Ljung-Box test p-value: 0.1

Neural Network Autoregressive model

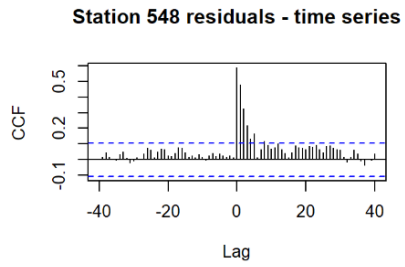
- The model used for seasonal data is an NNAR(p,P,k)[m], similar to an ARIMA(p,0,0)(P,0,0)[m] but with nonlinear components
- In this case NNAR(7,1,4)[365]
- ACF plots of residuals and Ljung-Box test p-values suggest residuals align with white noise
- Residuals squared and cross-correlation function plots indicate some correlated lags, suggesting the need for a more advanced neural network model

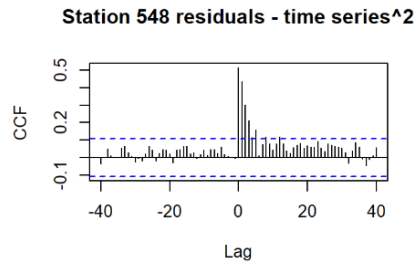


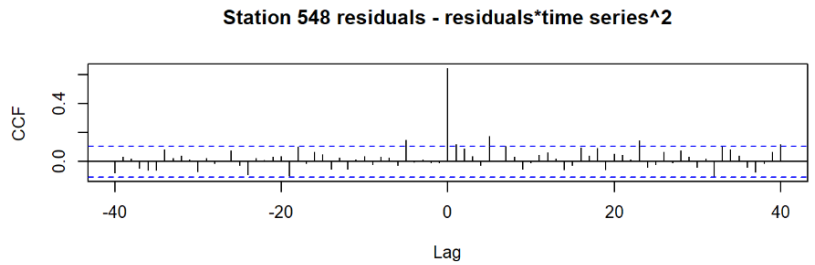
Ljung-Box test p-value: 0.9



Ljung-Box test p-value: 0.02



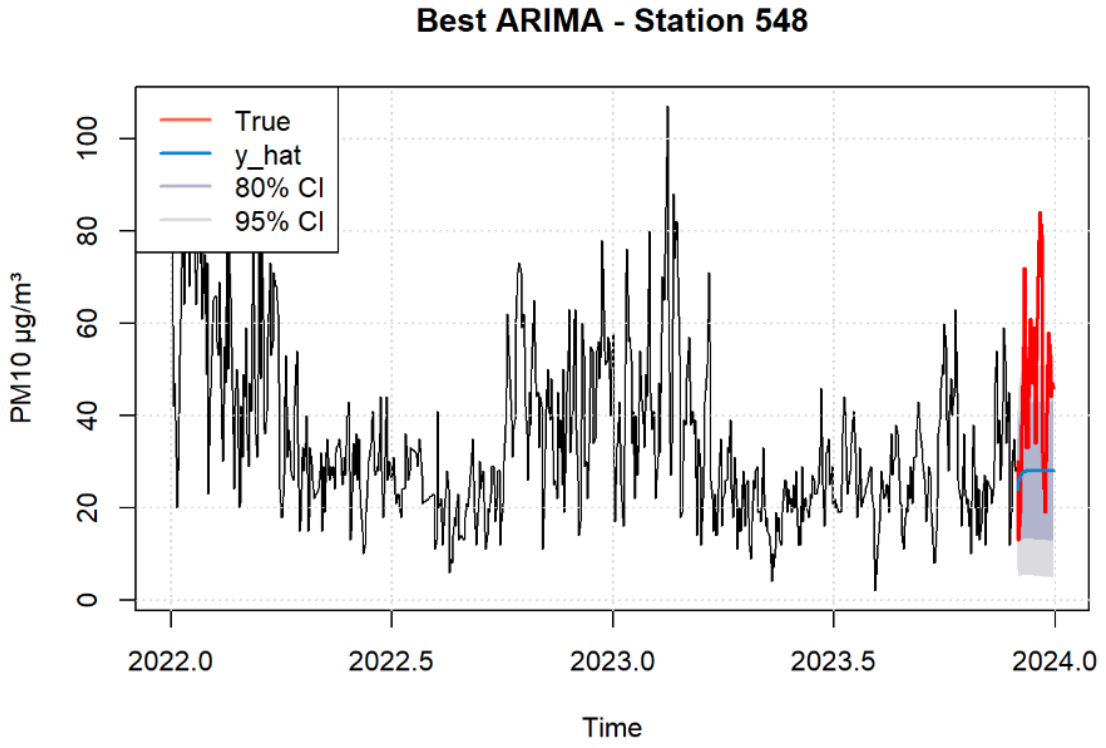




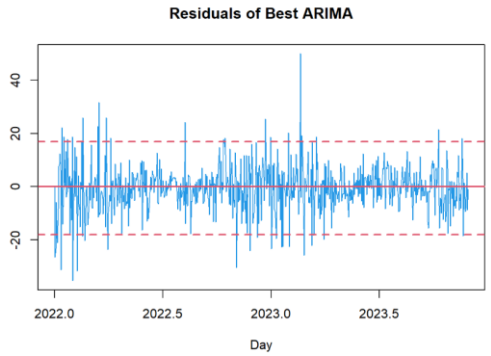
Forecasting assesment

Prediction performance best ARIMA

- Forecast future PM10 values over a 31-day horizon
- The best ARIMA model selected based on AICc score
- Model also have lower BIC values, indicating better performance despite model complexity
- In dynamic regression models, the mean of NO_2 predictor observations from the training set is used to forecast PM10 values



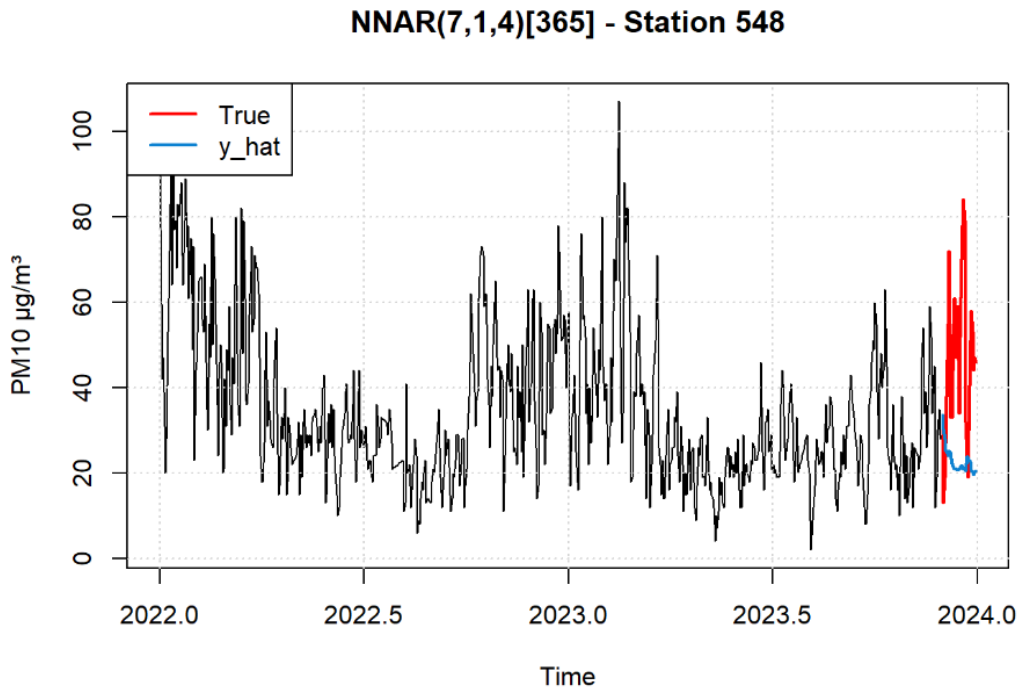
- Suboptimal results for longer-term predictions
- Residuals appear uncorrelated by ACF plots
- Residuals have a mean close to zero
- Residuals variability is not constant
- Q-Q plot shows unsatisfactory behavior in the tails

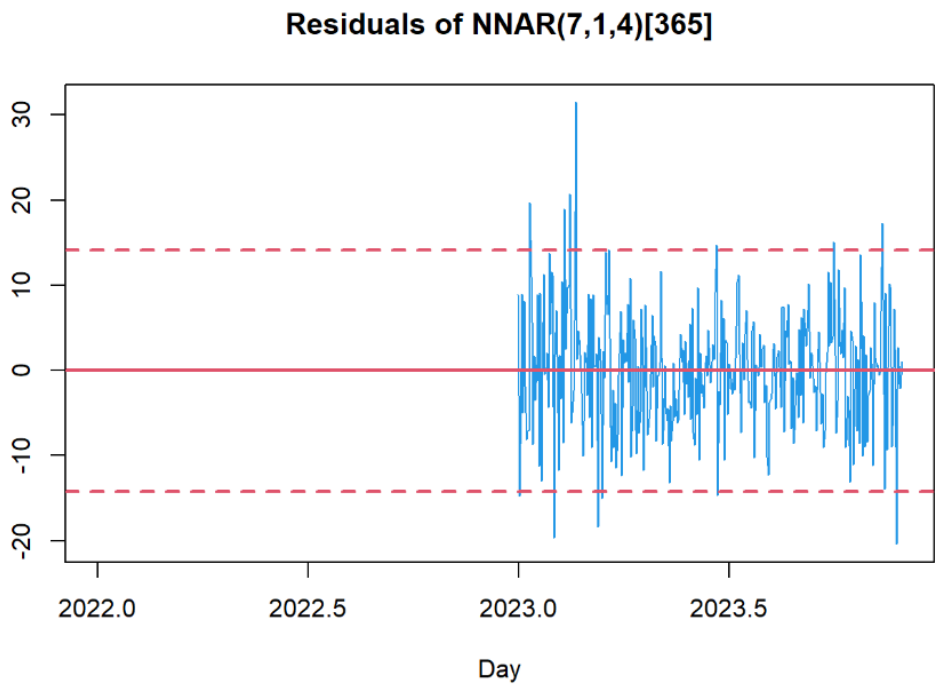


Ljung-Box test p-value: 0.2

Prediction performance NNAR

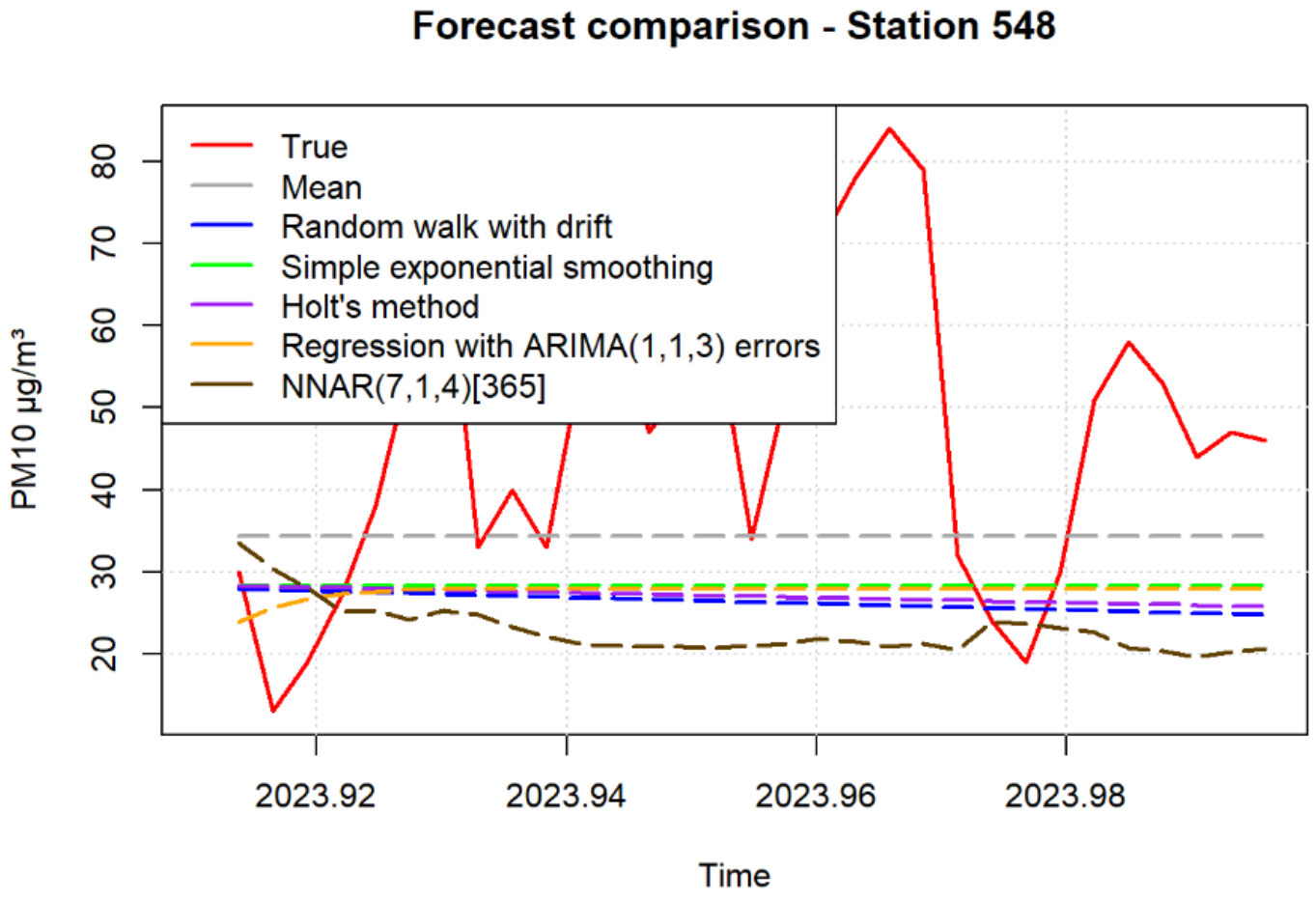
- Neural network autoregressive model forecasts show a distinctly different pattern compared to previous models
- The residual plot contains many NA values
- ACF is satisfactory and Ljung Box test p value is 1
- Residuals generally meet the assumptions of normality and a zero mean





Forecasts comparison

- The plot shows only the test set for a 31-day forecast horizon
- Drift method, simple exponential smoothing, and Holt's method perform similarly, but diverge over the long term
- The dynamic regression model with NO_2 initially underestimates values but aligns more closely with simpler methods over time
- The neural network-based model exhibits oscillating behavior



- Models are evaluated across forecast horizons of 1, 3, 7, 14, and 31 days
- Prediction uncertainty increases with the forecast horizon length. Short-term predictions are especially important for daily data
- For a 1-day forecast, simple methods like exponential smoothing achieve the lowest errors, outperforming the naive approach (MASE score)
- For a 3-day forecast, the dynamic regression model with ARIMA errors performs best, though simpler methods are still effective
- Beyond 7-14 days, prediction accuracy generally declines due to increased uncertainty. The average method often performs better in these longer-term scenarios

Forecast metrics - Station 548 - 1 day ahead

	RMSE	MAE	MAPE	MASE
Mean	4.345	4.345	14.483	0.2649
Random walk with drift	2.105	2.105	7.015	0.1283
Simple exponential smoothing	1.656	1.656	5.521	0.1010
Holt's method	1.724	1.724	5.748	0.1051
Regression with ARIMA(1,1,3) errors	6.069	6.069	20.232	0.3700
NNAR(7,1,4)[365]	3.541	3.541	11.804	0.2159

Forecast metrics - Station 548 - 3 days ahead

	RMSE	MAE	MAPE	MASE
Mean	15.383	13.678	86.48	0.8339
Random walk with drift	9.977	8.527	55.50	0.5199
Simple exponential smoothing	10.416	8.781	57.58	0.5354
Holt's method	10.274	8.674	56.84	0.5288
Regression with ARIMA(1,1,3) errors	9.264	8.827	52.83	0.5381
NNAR(7,1,4)[365]	11.477	9.982	64.29	0.6086

Forecast metrics - Station 548 - 7 days ahead

	RMSE	MAE	MAPE	MASE
Mean	19.35	15.76	55.16	0.9610
Random walk with drift	21.29	15.58	44.10	0.9497
Simple exponential smoothing	20.83	15.38	44.60	0.9376
Holt's method	21.09	15.50	44.51	0.9448
Regression with ARIMA(1,1,3) errors	20.88	15.50	42.53	0.9450
NNAR(7,1,4)[365]	22.99	17.44	50.58	1.0632

Forecast metrics - Station 548 - 14 days ahead

	RMSE	MAE	MAPE	MASE
Mean	17.53	14.09	39.41	0.8592
Random walk with drift	21.54	17.36	41.26	1.0586
Simple exponential smoothing	20.61	16.52	39.84	1.0070
Holt's method	21.17	17.03	40.81	1.0385
Regression with ARIMA(1,1,3) errors	20.80	16.76	39.23	1.0219
NNAR(7,1,4)[365]	24.80	20.69	49.74	1.2615

Forecast metrics - Station 548 - 31 days ahead

	RMSE	MAE	MAPE	MASE
Mean	21.85	17.38	38.05	1.060
Random walk with drift	27.26	22.07	43.32	1.345
Simple exponential smoothing	25.64	20.54	40.82	1.252
Holt's method	26.73	21.58	42.56	1.315
Regression with ARIMA(1,1,3) errors	25.87	20.80	40.78	1.268
NNAR(7,1,4)[365]	30.81	25.64	50.97	1.563

Conclusions

Summary

- The analysis **provides insights** into air pollution in Lombardia, Italy, **without considering meteorological factors**
- Exploratory data shows **rural areas may experience higher PM10** values at times, with PM10 exhibiting strong seasonality
- The **overall trend of PM10 is decreasing**, and values did not significantly drop during the COVID-19 emergency
- Mean annual values did not exceed thresholds, suggesting no alarming situation
- For station 548 in Milano v.Senato, the models do **not significantly improve upon simpler methods**, and the neural network autoregressive model performed poorly
- External predictors show potential benefits
- Advanced models are needed to capture daily data patterns better, and incorporating Fourier terms could improve forecast accuracy by addressing seasonality

Credits

- https://cran.r-project.org/web/packages/ARPALData/index.html
- https://www.flaticon.com/free-icons/