Assessing Air Quality in Lombardia, Italy through Time Series Analysis: Implications for Public Health and Policy

Time Series Analysis project

Giovanni Costa - 880892

AY 2023/24

Contents

Introduction

This project aims to evaluate the air quality in Lombardia, Italy, using a comprehensive time series analysis. The data for this study will be derived from sensors located at various stations across the region, which can be accessed via the regional website.

The primary focus is to underscore the significance of air quality on public health. By analyzing the trends and patterns in air quality data over time, it's possible to identify periods of high pollution and correlate these with potential health risks. This analysis will provide valuable insights into how air quality fluctuations may impact the respiratory health of Lombardia's inhabitants.

Considering the data provided by the region, the following pollutants will be analyzed across three different stations placed in different areas:

- **Benzene** is a volatile organic compound (VOC) commonly used in the production of plastics, resins, and synthetic fibers, as well as in gasoline. It is released into the air through emissions from motor vehicles, industrial processes, and the evaporation of benzene-containing products. Exposure to benzene primarily occurs through inhalation and can lead to serious health effects, including bone marrow damage, which can cause blood disorders such as anemia and increase the risk of leukemia, a type of cancer.
- **Carbon Monoxide (CO)** is a colorless, odorless gas produced by incomplete combustion of carbon-containing fuels, such as gasoline, natural gas, oil, and wood. Major sources include motor vehicles, industrial processes, and residential heating systems. CO interferes with the body's ability to transport oxygen by binding to hemoglobin in the blood, forming carboxyhemoglobin. High levels of exposure can lead to symptoms such as headaches, dizziness, and even death due to oxygen deprivation.
- **Nitrogen Dioxide (NO_2)** is a reddish-brown gas with a sharp, biting odor. It is a significant air pollutant formed primarily from the combustion of fossil fuels in vehicles, power plants, and industrial processes. NO_2 can irritate the respiratory system, exacerbate asthma, and reduce lung function. It also contributes to the formation of ground-level ozone and fine particulate matter, which can have further adverse effects on human health and the environment.
- **Nitrogen Oxides (NO_2)** encompass a group of gases, including nitrogen dioxide (NO_2) and nitrogen monoxide (NO), produced during combustion processes, particularly at high temperatures. Major sources include motor vehicles, power plants, and industrial facilities. NO_x gases can cause respiratory problems, contribute to the formation of smog and acid rain, and lead to the secondary formation of fine particulate matter (PM2.5), all of which can have severe health and environmental impacts.
- **Particulate Matter 10 (PM10)** refers to airborne particles with a diameter of 10 micrometers or less. These particles can originate from a variety of sources, including construction sites, road dust, industrial emissions, and combustion processes. PM10 can be inhaled into the respiratory system, leading to health issues such as respiratory infections, lung inflammation, and aggravation of existing heart and lung diseases. Long-term exposure can decrease lung function and increase mortality from cardiovascular and respiratory diseases.

In particular, PM10 is selected for the main part of this analysis because it serves as a comprehensive indicator of particulate pollution from a wide range of sources, such as traffic emissions, industrial processes, and the secondary formation of particles from gaseous pollutants. Its direct link to adverse health effects, including respiratory and cardiovascular diseases, makes PM10 a crucial measure of air quality. Additionally, other pollutants like Benzene, CO, NO_2, and NO_x can contribute to PM10 levels through both primary emissions and secondary reactions, making them valuable predictor variables for analyzing fluctuations in PM10 concentrations.

For this study, daily data are considered, followed by aggregation into monthly averages to facilitate a long-term analysis and provide a clearer understanding of the underlying trends. Subsequently, forecasting models will be developed using pure daily data to predict air quality in the region.

Dataset description

The air quality dataset utilized is composed of daily observations of Benzene, CO, NO2, NOx, and PM10 coming from 3 different stations placed in different positions:

1. Station 548 - Milano v.Senato refers to the metropolitan area of Milan
2. Station 571 - Bormio v.Monte Braulio is located in the mountain zone
3. Station 703 - Schivenoglia v.Malpasso is placed in the rural plain of Lombardia

The analysis covers the period from January 1, 2014, to December 31, 2023. The original air quality data, which were recorded hourly, were aggregated to daily values by calculating the mean and excluding NA values. For additional details, refer to the ARPALData library manual [1].

A dataset with detailed information about the monitoring stations is also available. It includes the sensor ID, the pollutant measured by each sensor, as well as the location, altitude, and the start and stop dates of station operation. Since the stations measure different pollutants, columns in the dataset with all `NA` values indicate that the station does not have sensors for measuring that particular pollutant in the air.

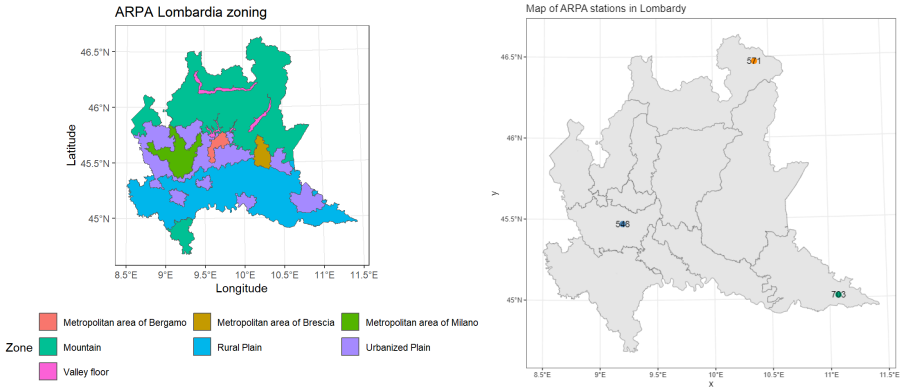


Figure 1: Stations zoning information

```
## [1] "The number of missing dates in the dataset from 2014-01-01 to 2023-12-31 is: 0"
```

```
## [1] "Common pollutants among the stations: Benzene, CO, NO2, NOx, PM10"
```

Data overview

Raw data

As previously mentioned, the air quality stations dataset provides details on the locations, sensors, and their operational status. Below is the table for the station located in Milan, while similar information is available for other stations, which may monitor different types of pollutants.

Table 1: Station 548 - Milano v.Senato information

IDSensor	Pollutant	Province	City	Latitude	Longitude	Altitude	DateStart	DateStop
12638	Arsenic	MI	Milano	45.47	9.197	118	2008-07-09	NA
6057	Benzene	MI	Milano	45.47	9.197	118	1999-01-01	NA
12641	Benzo_a_pyrene	MI	Milano	45.47	9.197	118	2008-07-10	NA
20005	BlackCarbon	MI	Milano	45.47	9.197	118	2013-09-30	2019-05-23
20465	BlackCarbon	MI	Milano	45.47	9.197	118	2020-03-12	NA
5834	CO	MI	Milano	45.47	9.197	118	1995-05-01	NA
12639	Cadmium	MI	Milano	45.47	9.197	118	2008-07-09	NA
12640	Lead	MI	Milano	45.47	9.197	118	2008-07-09	NA
5551	NO2	MI	Milano	45.47	9.197	118	1995-04-21	NA
6354	NOx	MI	Milano	45.47	9.197	118	1995-04-21	NA
12637	Nikel	MI	Milano	45.47	9.197	118	2008-07-09	NA
10320	PM10	MI	Milano	45.47	9.197	118	2007-08-10	NA
17122	PM2.5	MI	Milano	45.47	9.197	118	2012-09-04	NA

More than the stations' details, it is considered more insightful to focus on the air quality summary tables for the different stations. First, the number of missing values within the selected 10-year daily interval is significant, and these gaps must

[1] https://cran.r-project.org/web/packages/ARPALData/index.html

be addressed through imputation to correctly represent the time data. Additionally, while the selected pollutants generally show a small standard deviation (based on the range between the 1st and 3rd quartiles), some pollutants, such as PM10 and NOx, exhibit maximum values that are considerably higher than the 3rd quartile.

Table 2: Station 548 - Milano v.Senato daily data statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Ammonia	NA	NA	NA	NaN	NA	NA	3652
Arsenic	0.2000	1.0500	1.050	1.0632	1.050	6.305	3057
Benzene	0.0773	0.6708	1.254	1.5637	2.059	9.375	688
Benzo_a_pyrene	0.0070	0.0520	0.080	0.3476	0.429	4.143	2422
BlackCarbon	0.1083	0.9635	1.504	2.0178	2.582	13.800	586
Cadmium	0.0400	0.0850	0.168	0.2414	0.210	8.405	3057
CO	0.0958	0.6333	0.850	0.9206	1.137	2.792	174
Lead	0.8500	6.2670	10.498	15.5039	18.910	84.087	3057
Nikel	0.8500	2.1000	4.200	4.9678	6.300	84.002	3057
NO2	11.4667	32.9469	45.817	47.5396	59.665	125.229	132
NOx	12.0750	44.7490	68.501	90.1249	114.554	593.962	132
Ozone	NA	NA	NA	NaN	NA	NA	3652
PM10	2.0000	21.0000	30.000	36.2028	46.000	180.000	161
PM2.5	0.0000	12.0000	19.000	24.6229	32.000	159.000	194
Sulfur_dioxide	NA	NA	NA	NaN	NA	NA	3652

Table 3: Station 571 - Bormio v.Monte Braulio daily data statistics

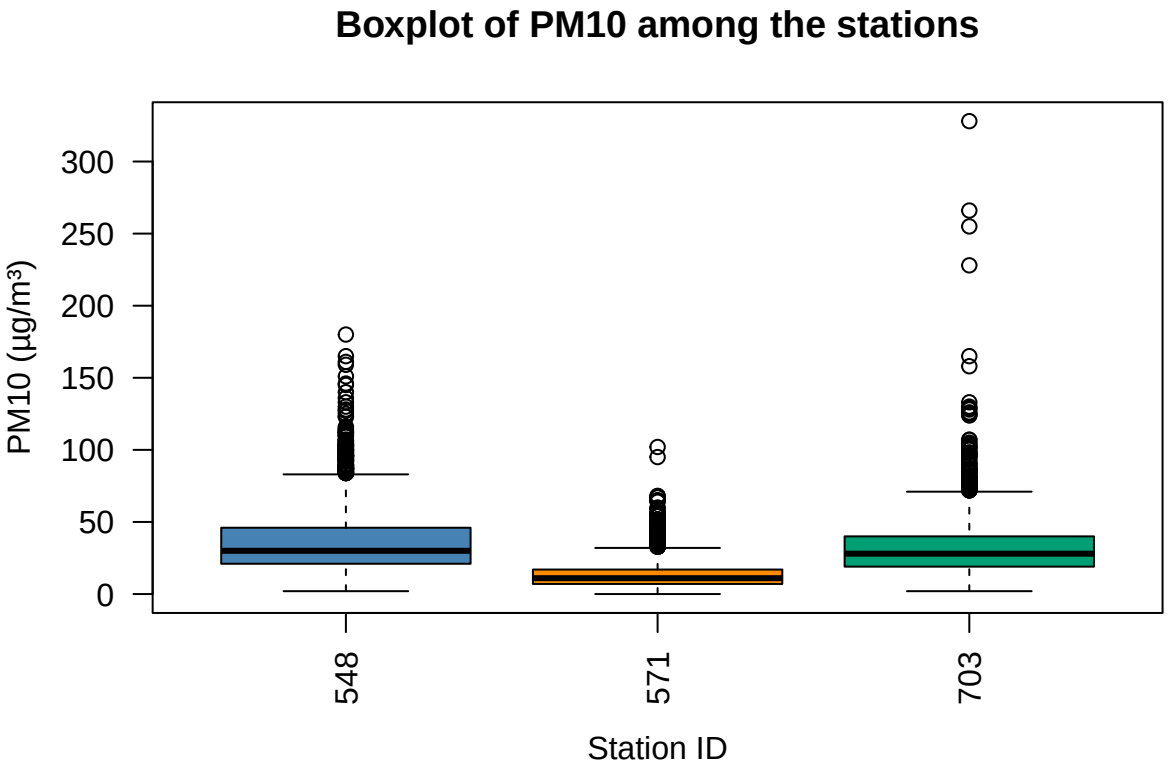
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Ammonia	NA	NA	NA	NaN	NA	NA	3652
Arsenic	NA	NA	NA	NaN	NA	NA	3652
Benzene	0.0000	0.2083	0.3417	0.6527	0.8542	7.358	1185
Benzo_a_pyrene	NA	NA	NA	NaN	NA	NA	3652
BlackCarbon	NA	NA	NA	NaN	NA	NA	3652
Cadmium	NA	NA	NA	NaN	NA	NA	3652
CO	0.0000	0.2667	0.3500	0.4152	0.4885	2.487	120
Lead	NA	NA	NA	NaN	NA	NA	3652
Nikel	NA	NA	NA	NaN	NA	NA	3652
NO2	0.1429	7.0917	10.3646	13.4804	17.3990	55.729	82
NOx	1.4000	9.9260	14.0604	20.8389	24.4490	190.375	82
Ozone	3.5000	45.4086	61.1021	60.4430	76.4333	143.192	108
PM10	0.0000	7.0000	11.0000	13.0044	17.0000	102.000	48
PM2.5	NA	NA	NA	NaN	NA	NA	3652
Sulfur_dioxide	0.0583	0.7583	1.1636	1.5001	1.8375	12.375	25

Table 4: Station 703 - Schivenoglia v. Malpasso daily data statistics

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
Ammonia	0.0000	9.9125	14.1500	16.3787	19.7292	448.041	415
Arsenic	0.1000	0.4500	1.0500	0.8343	1.0500	4.556	3057
Benzene	0.0000	0.1042	0.2417	0.4589	0.6833	2.946	55
Benzo_a_pyrene	0.0220	0.0500	0.0520	0.2818	0.3137	3.821	2430
BlackCarbon	NA	NA	NA	NaN	NA	NA	3652
Cadmium	0.0000	0.0400	0.0840	0.1480	0.1830	3.349	3057
CO	0.0000	0.3000	0.4957	0.5499	0.7208	1.883	91
Lead	0.4500	1.8260	4.1670	5.4383	7.1640	33.498	3057
Nikel	0.3500	1.1290	2.1000	3.0706	4.1850	33.491	3057
NO2	0.5833	9.7042	13.6208	16.3577	20.7439	55.583	84

NOx	2.8292	11.9233	16.6229	24.2250	27.6333	187.258	90
Ozone	1.2000	25.5927	54.8187	51.0362	73.4719	138.047	96
PM10	2.0000	19.0000	28.0000	31.9759	40.0000	328.000	246
PM2.5	0.0000	12.0000	18.0000	22.4005	29.0000	123.000	249
Sulfur_dioxide	0.0000	1.8167	2.7833	2.9466	3.8292	14.533	81

The data distribution of the PM10 can be better highlighted with a boxplot: the median values of the Milan station and Schivenoglia station are quite similar and all the stations present observations very distant from the others (in particular station 703). Indeed, PM10 concentrations can be higher in rural areas compared to urban metropolitan areas for several reasons. Agricultural activities, such as harvesting and livestock operations, contribute to elevated PM10 levels by generating bioaerosols rich in plant and animal matter. Additionally, rural dust often contains higher concentrations of crustal metals, exacerbated by drier climates and open spaces. Furthermore, rural areas typically experience more stable atmospheric conditions, leading to reduced dispersion of particulates compared to the more unstable, windier conditions often found in urban environments due to intense heat islands.



Standard regulatory levels for PM10 are as follows: the **Acceptable Level** is 50 µg/m³, which can be exceeded on up to 35 days per year without health concerns. The **Information Level** is set at 200 µg/m³, triggering public notifications about potential health risks. The **Alarm Level** is 300 µg/m³, prompting immediate public health actions, such as advising vulnerable groups to limit outdoor activities. Additionally, the **Annual Average** acceptable level is 40 µg/m³, ensuring that the yearly average concentration does not exceed this value to protect public health. Fortunately, as indicated in this table, the annual average concentration of Particulate Matter 10 remains only slightly above the limit across the years.

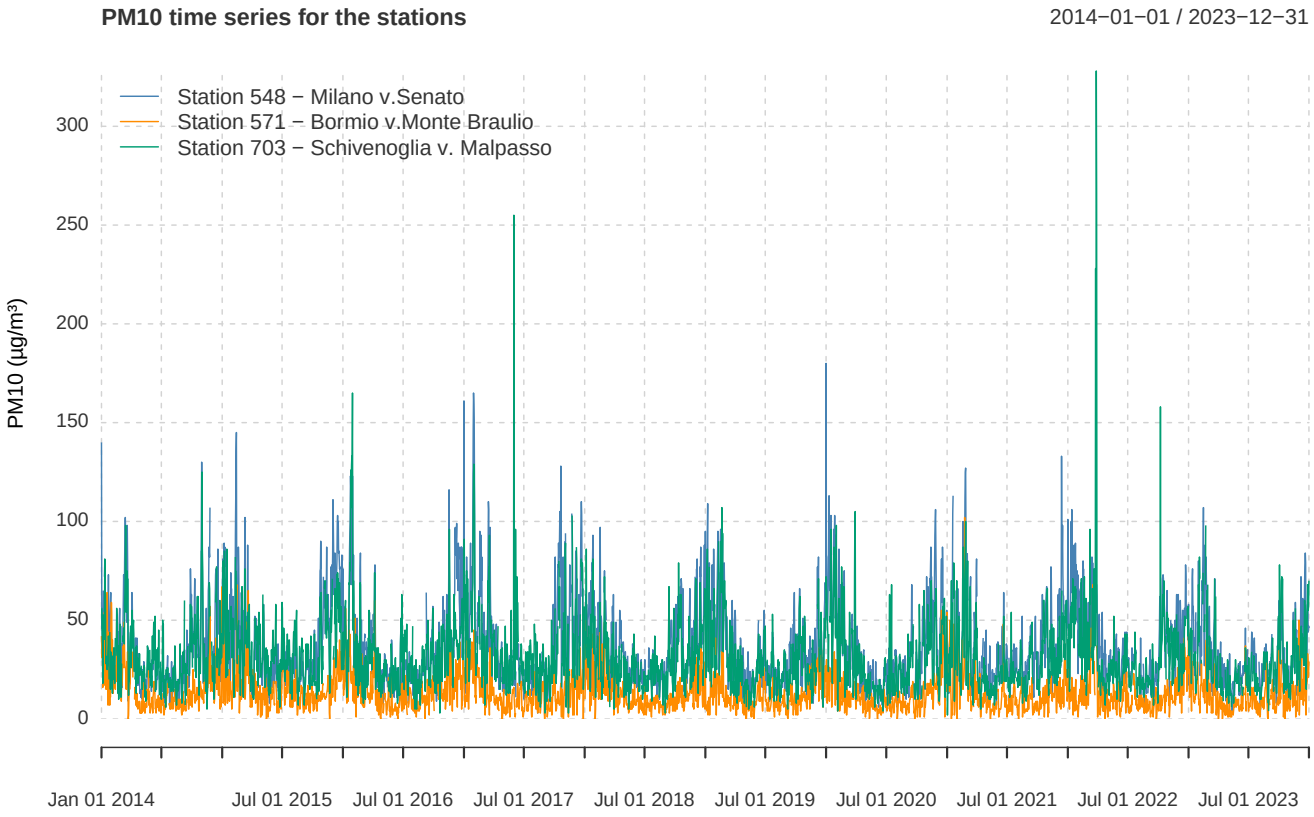
Table 5: Mean PM10 values by year

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

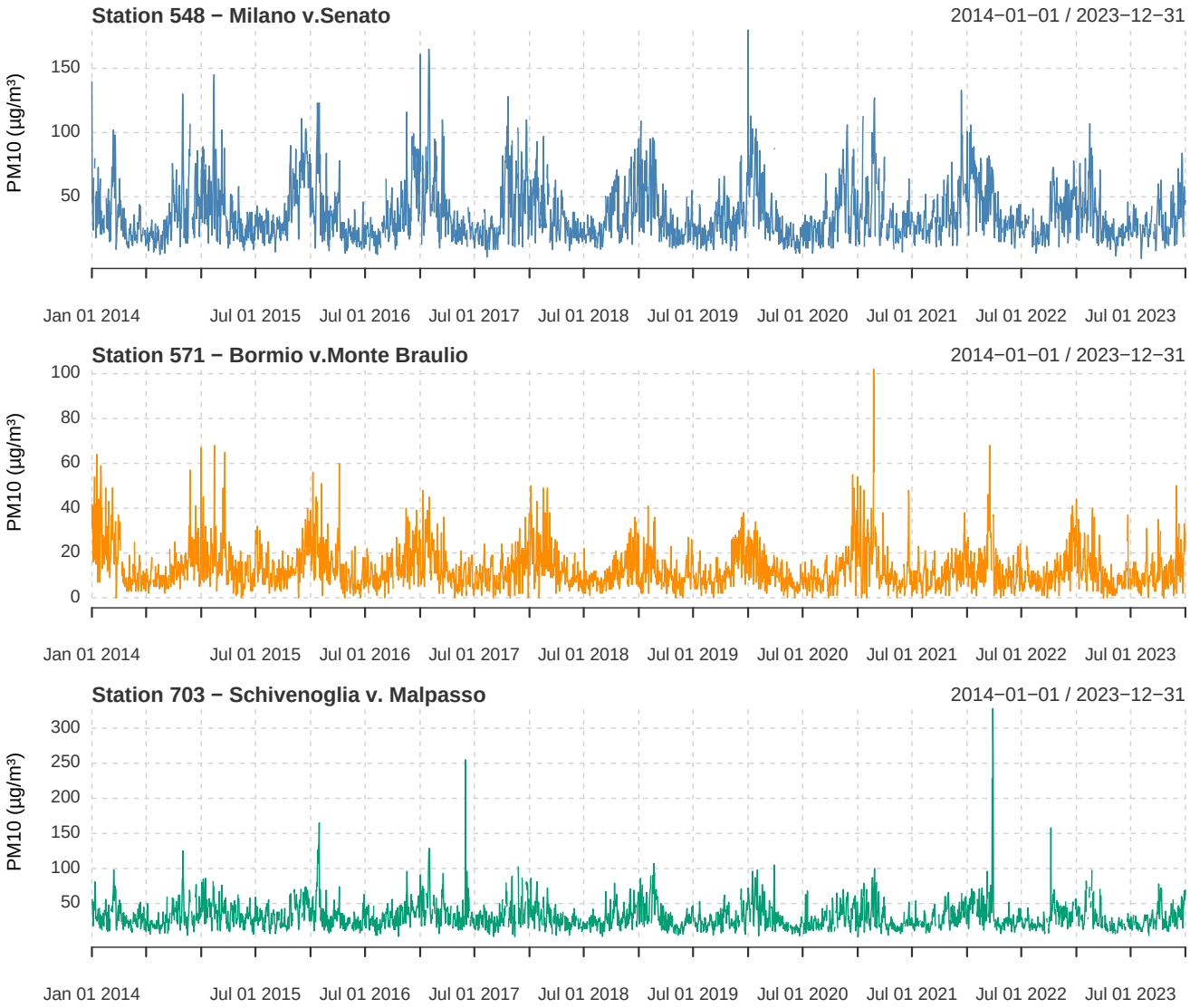
2020	36.49	11.65	31.21
2021	36.83	12.84	29.38
2022	38.96	12.30	33.57
2023	31.68	11.29	28.66

Time series data

Currently, the data are analyzed as time series rather than as raw values, and further considerations will be based on this interpretation. The plot displaying time series from all stations confirms previous observations about data distribution: the Milan station generally records higher PM10 values, whereas the Bormio station reports the lowest levels, and the Schivenoglia station shows numerous peaks.



When the time series are plotted individually, their patterns become clearer. Each series exhibits variability and is apparently non-stationary, with indications of some seasonal patterns across all stations.



Preprocessing

Missing values imputation

As previously noted, the data contain many missing values, and time series require consistent spacing without gaps. The following statistics provide insight into these missing values:

- "Number of Gaps" indicates the count of NA gaps, which are sequences of one or more consecutive missing values.
- "Average Gap Size" represents the average length of these consecutive NA gaps.
- "Longest NA Gap" shows the longest sequence of consecutive missing values in the time series.
- "Most Frequent Gap Size" identifies the most commonly occurring length of missing value sequences.

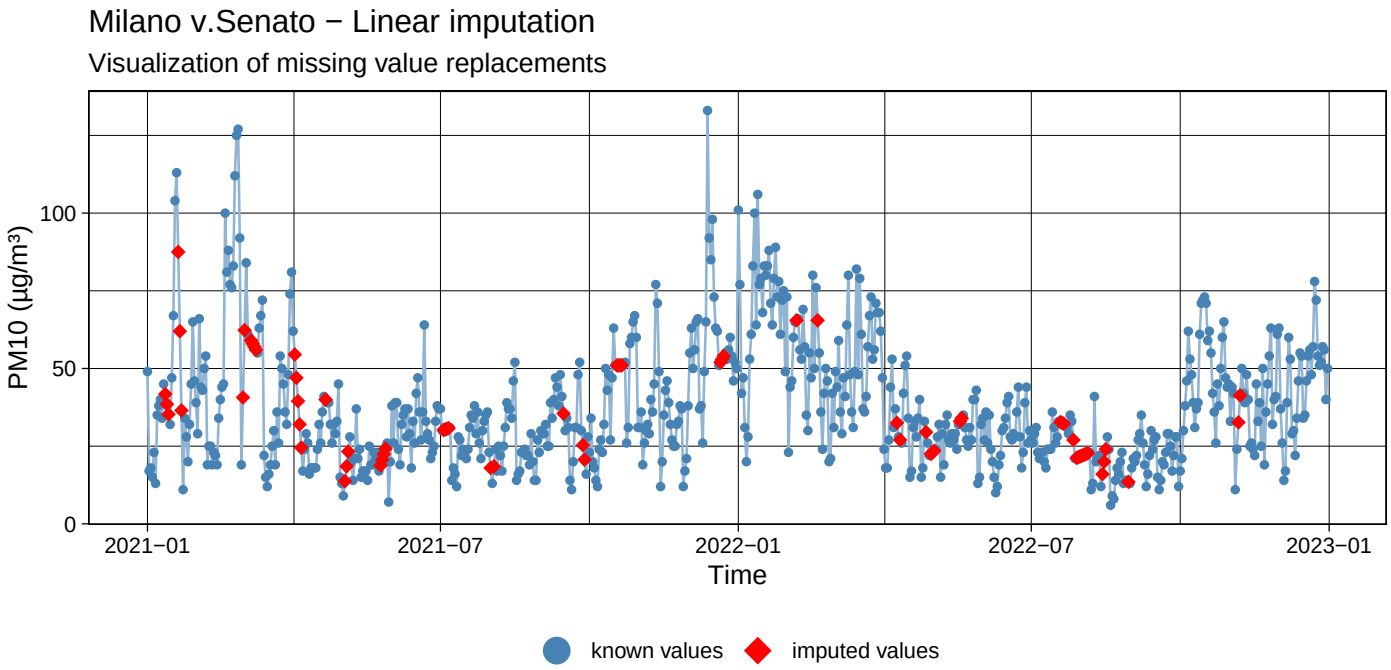
Table 6: Missing values statistics

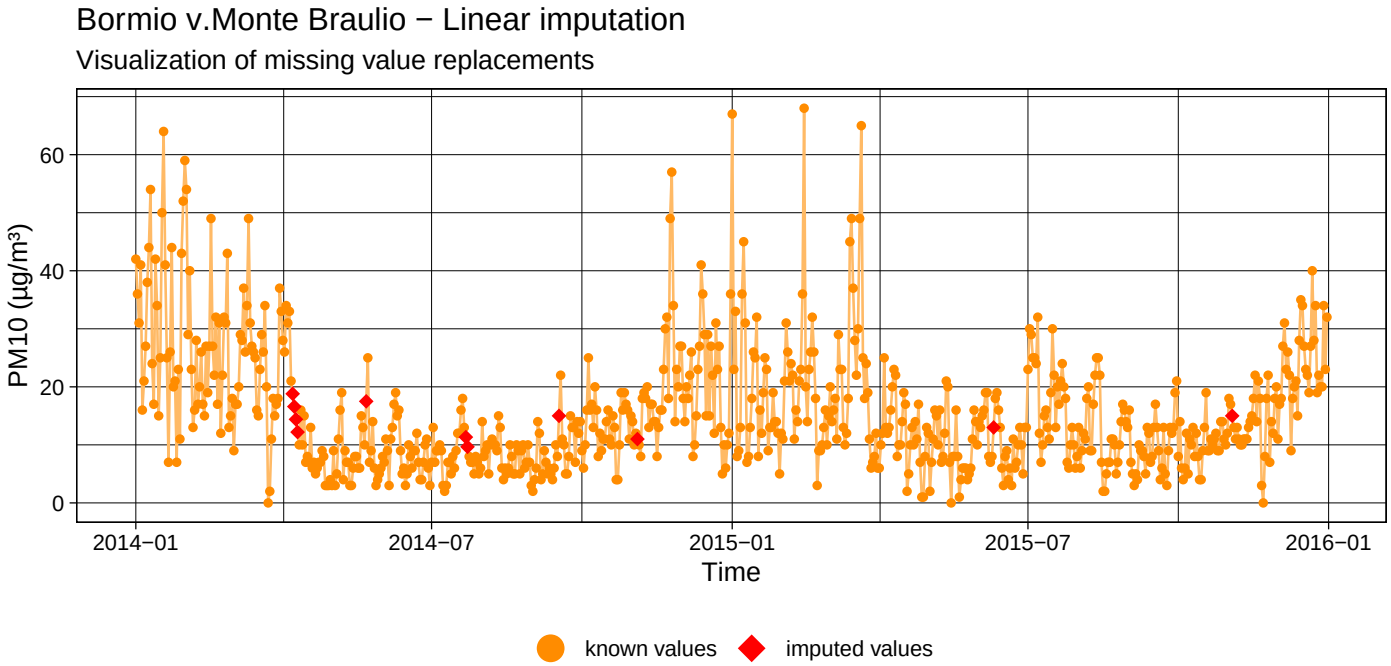
	Length TS	Number NAs	Number Gaps	Average Gap Size	Percentage NAs	Longest NA gap	Most frequent
Station 548	3652	161	79	2.038	4.41%	8	1
Station 571	3652	48	29	1.655	1.31%	9	1
Station 703	3652	246	135	1.822	6.74%	6	1

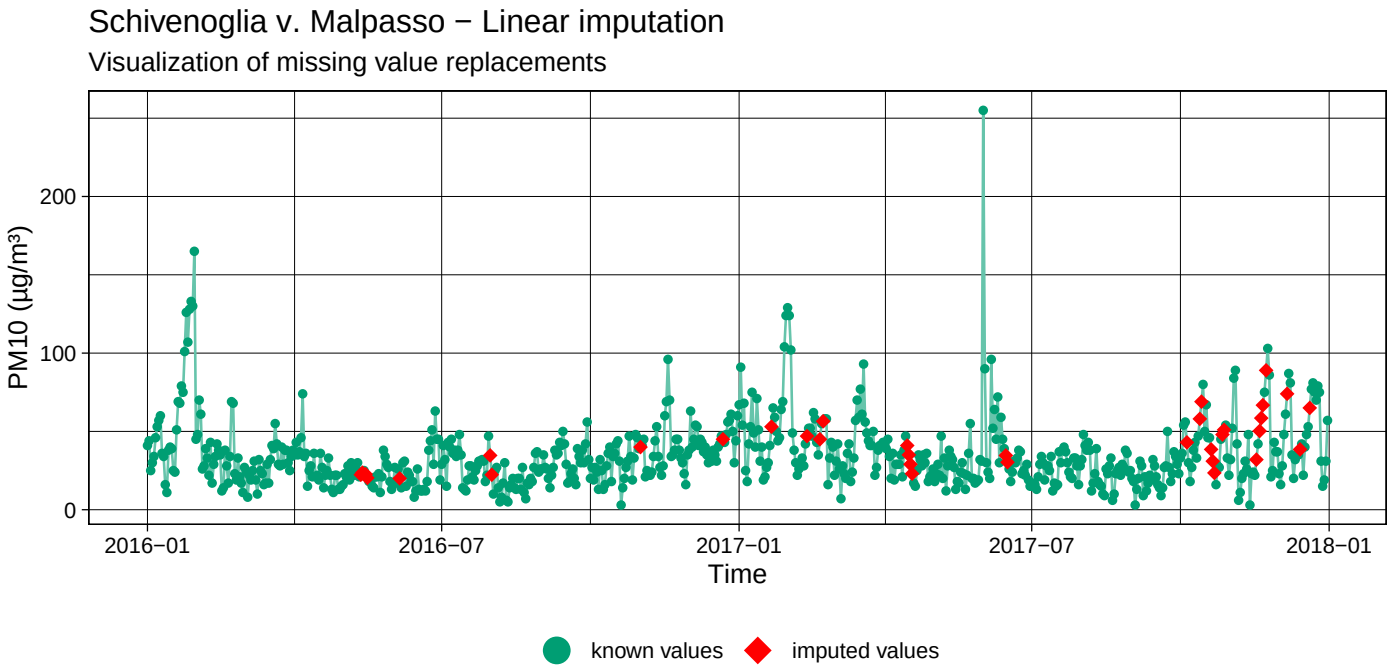
Fortunately, the most frequent gap length is one, and the average gap size is relatively small, likely resulting from occasional sensor malfunctions. This suggests that simple imputation techniques should be reasonably accurate and close to the actual values.

To address missing values in the time series, linear interpolation is used. This method assumes that missing values can be estimated by drawing a straight line between the known values on either side. For time series data, this means using the timestamps and values of the adjacent non-missing points to calculate the missing values.

As shown the the following plots, the values imputed among all the stations seem coherent.



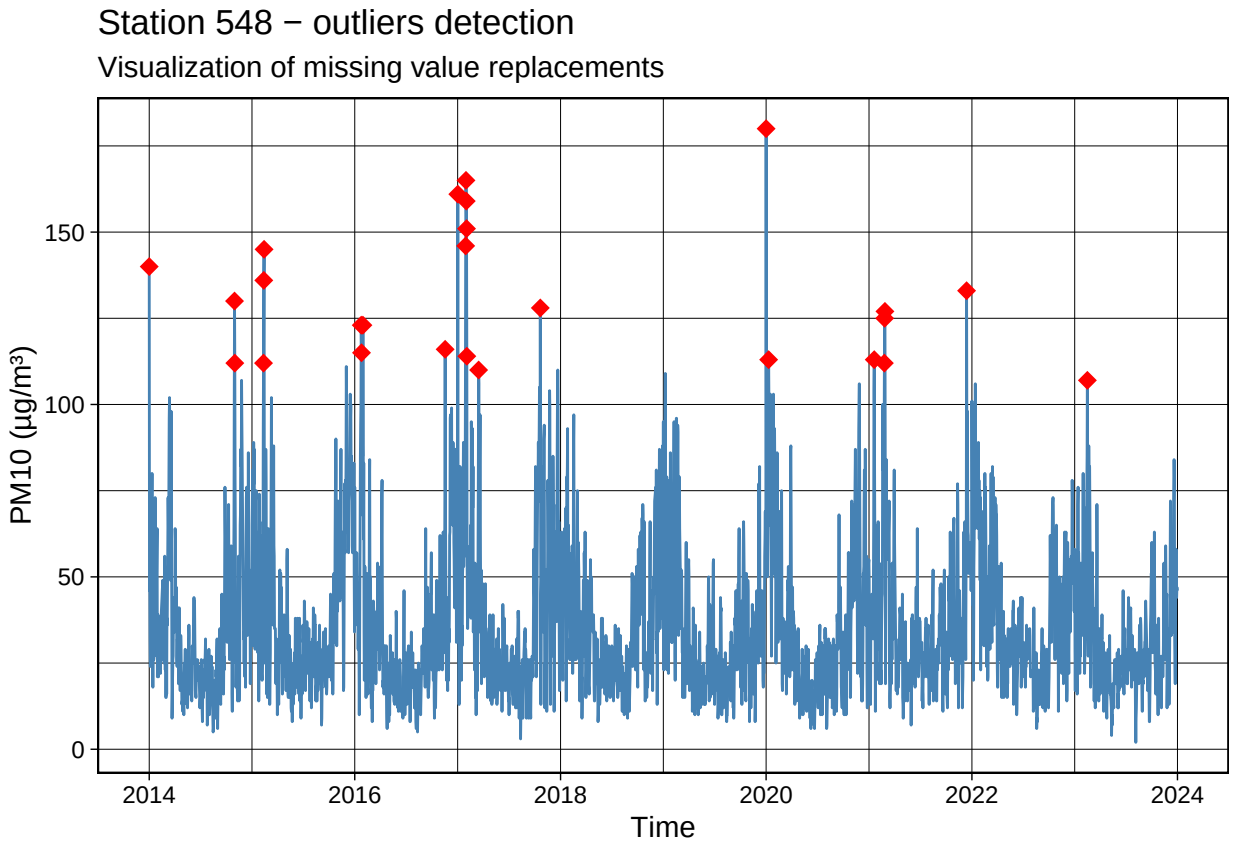


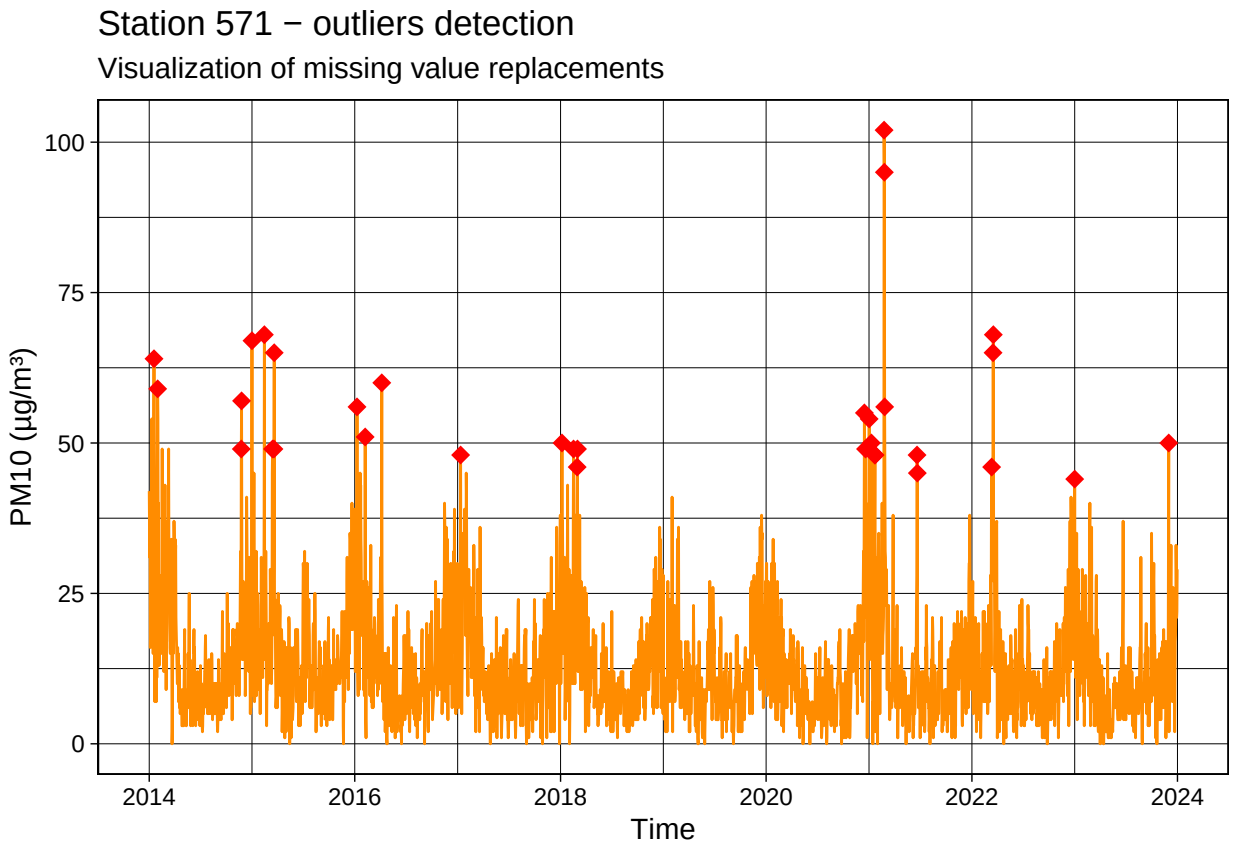


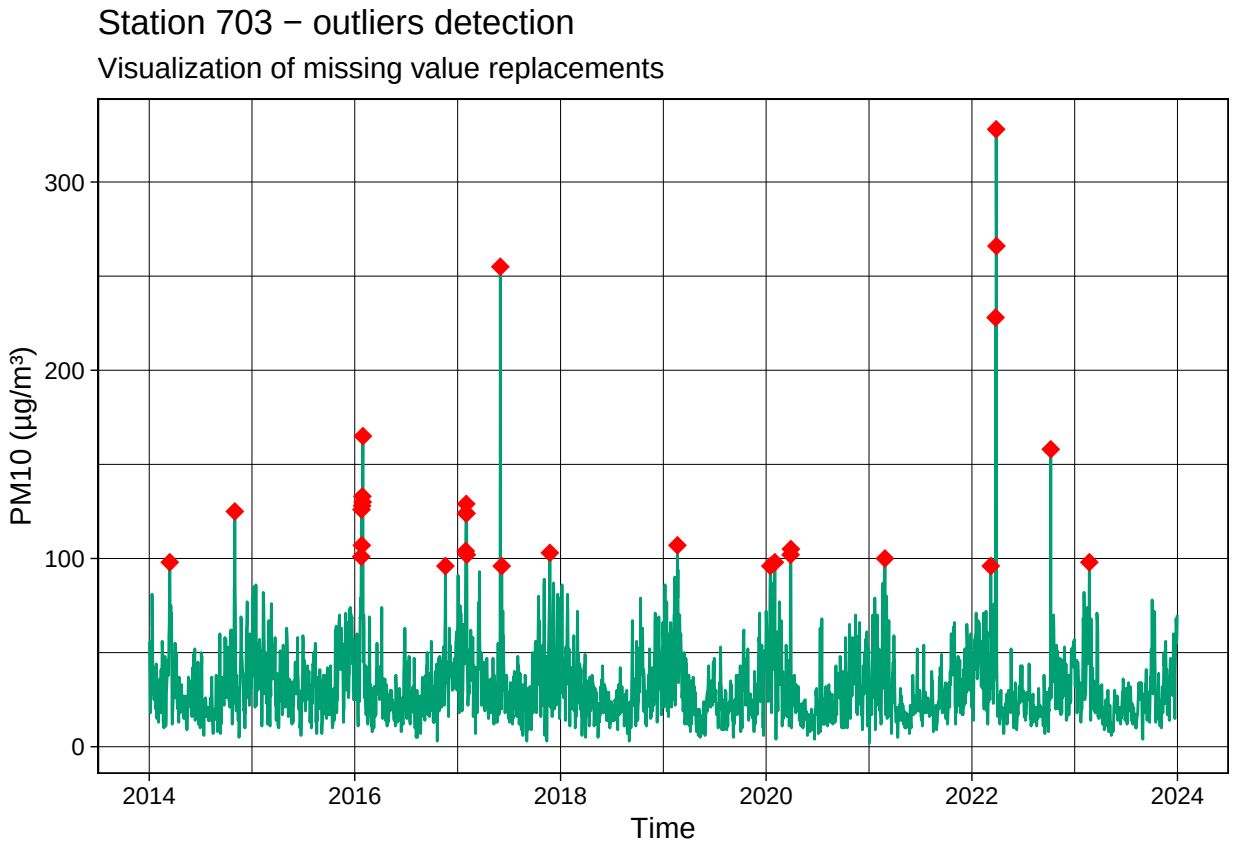

```
## [1] "Number of NA values in PM10 among the station are:  0"
```

Outliers detection

The previously identified outliers are also evident in the time series data, particularly as prominent peaks at the stations. These outliers have been left unaltered to preserve the integrity and semantics of the data.



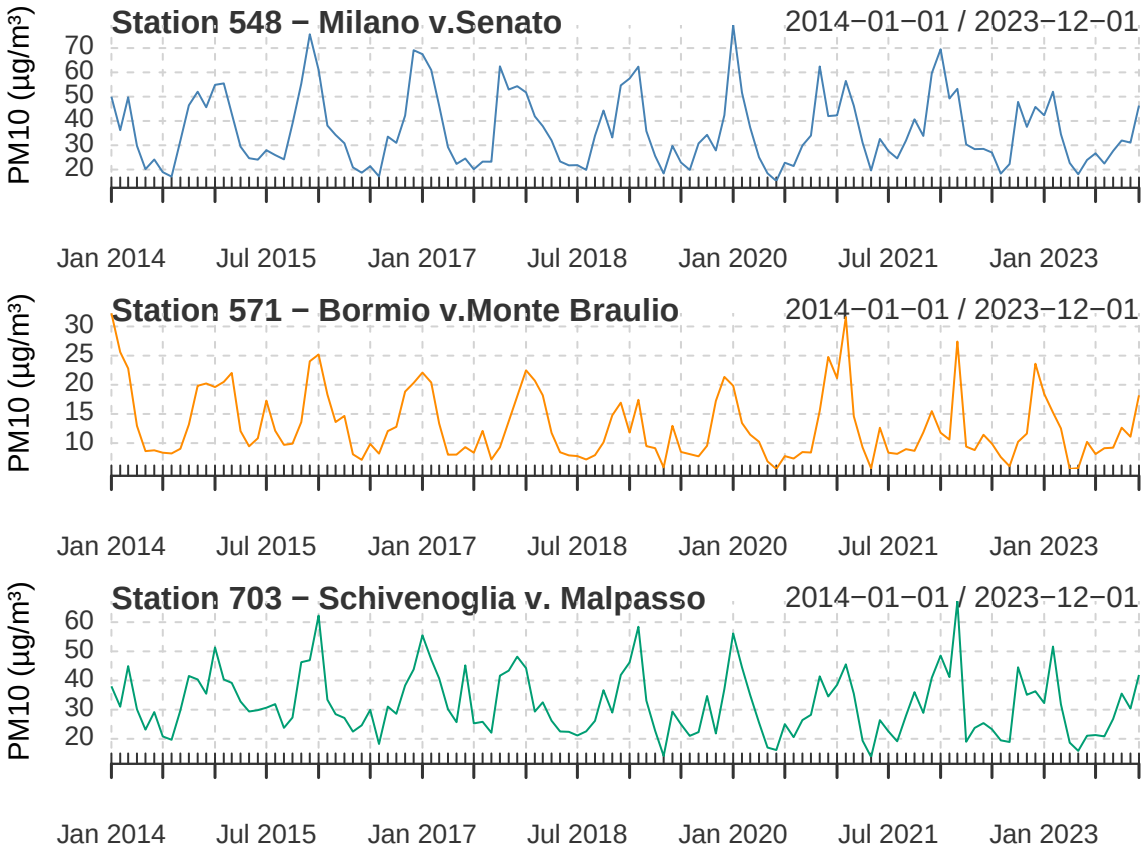




Time series data analysis

In this section data daily data are averaged across the months for performing a long-term analysis. Due to the lower number of observations, now the time series appear more clear. However, just by plotting the monthly values, the trend

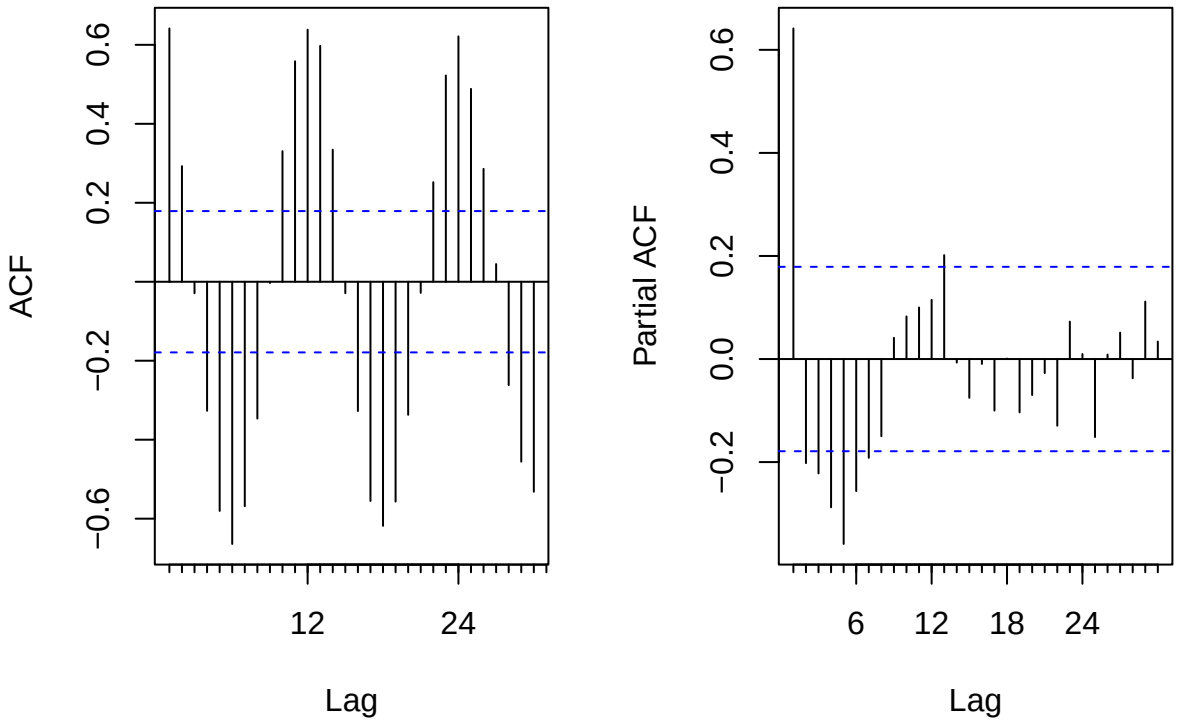
and the seasonal pattern don't emerge much.



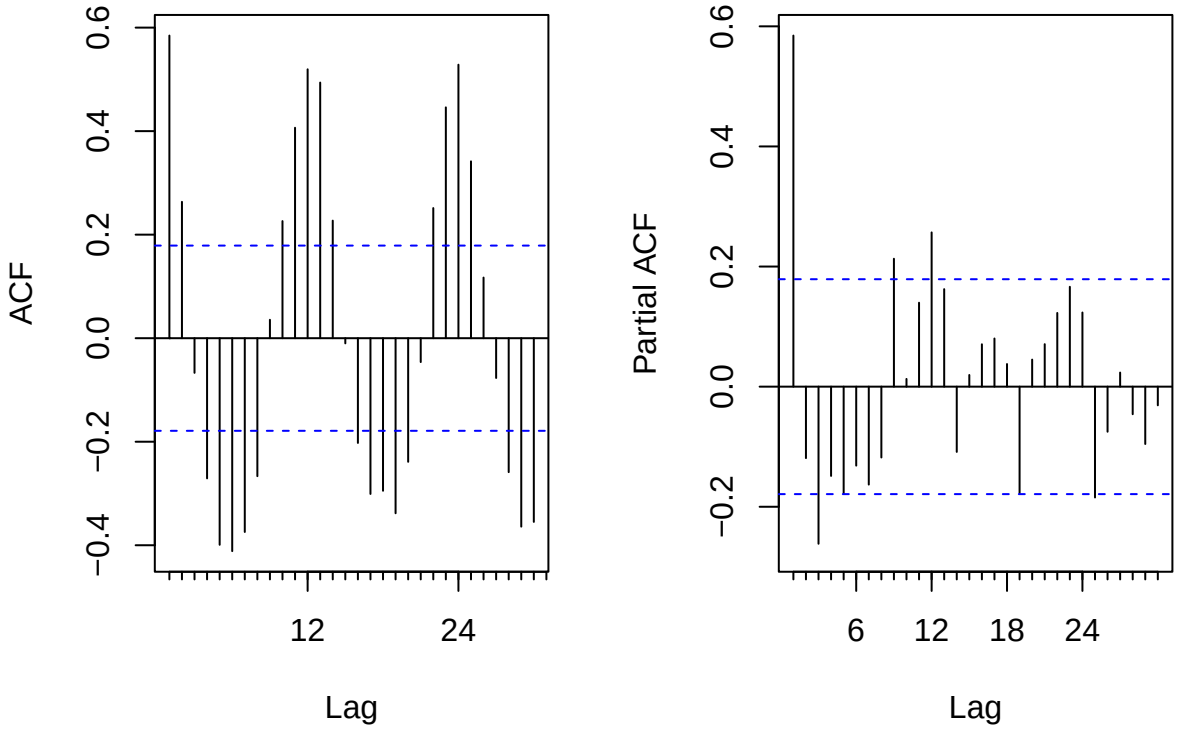
Autocorrelation and partial autocorrelation

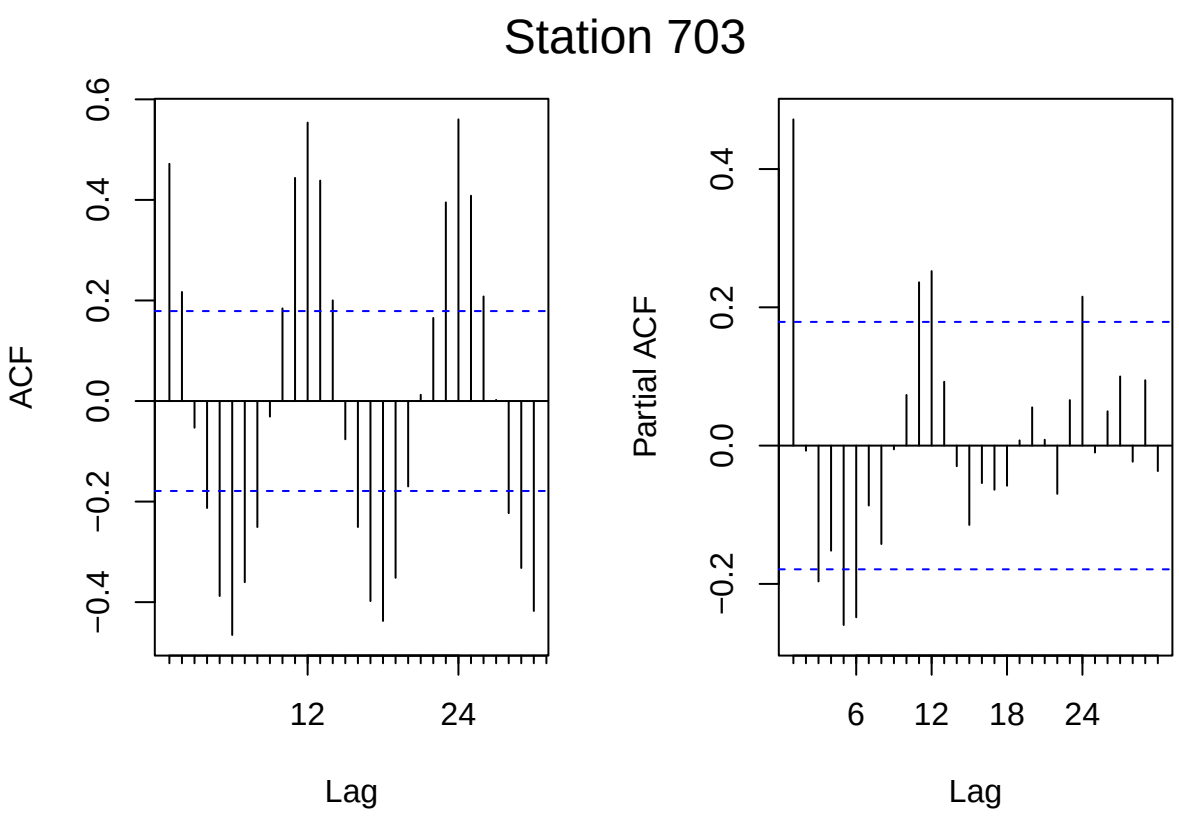
Autocorrelation (ACF) and partial autocorrelation (PACF) function plots provide additional insights into the data. The ACF plots reveal a sinusoidal pattern across all stations, with a pronounced peak every six lags, suggesting a recurring pattern approximately every six months. The PACF plots also show spikes around this period, indicating the presence of significant information during these intervals. Additionally, the analysis confirms with reasonable confidence that the time series is non-stationary.

Station 548



Station 571

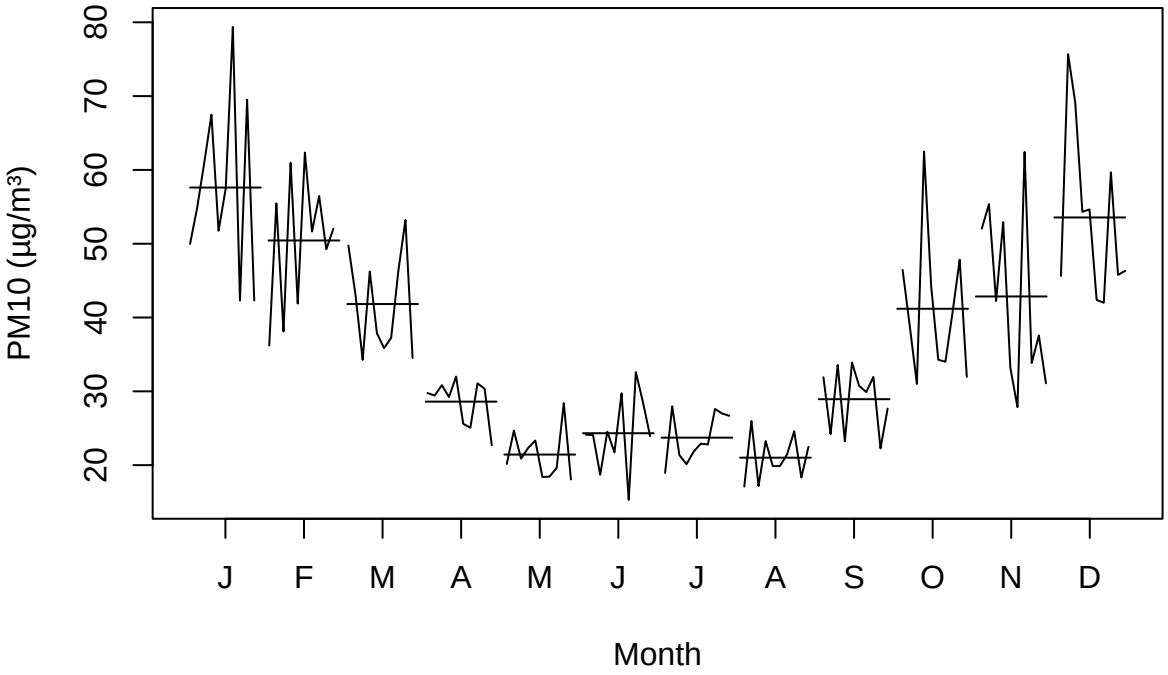




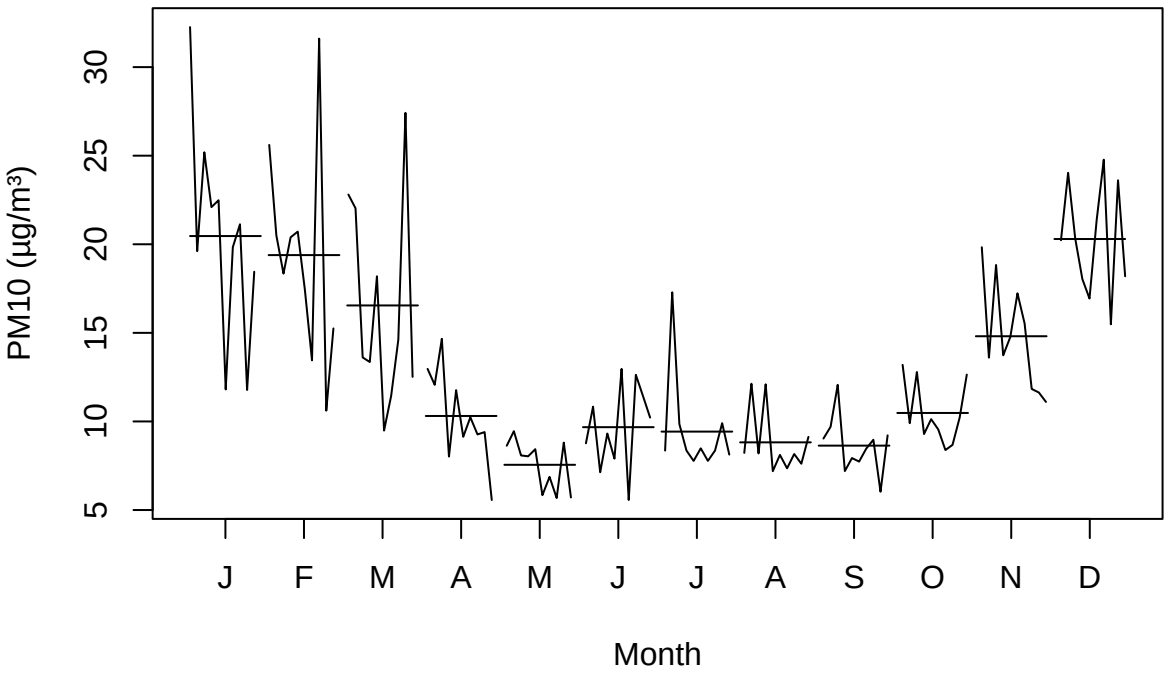
Monthplot

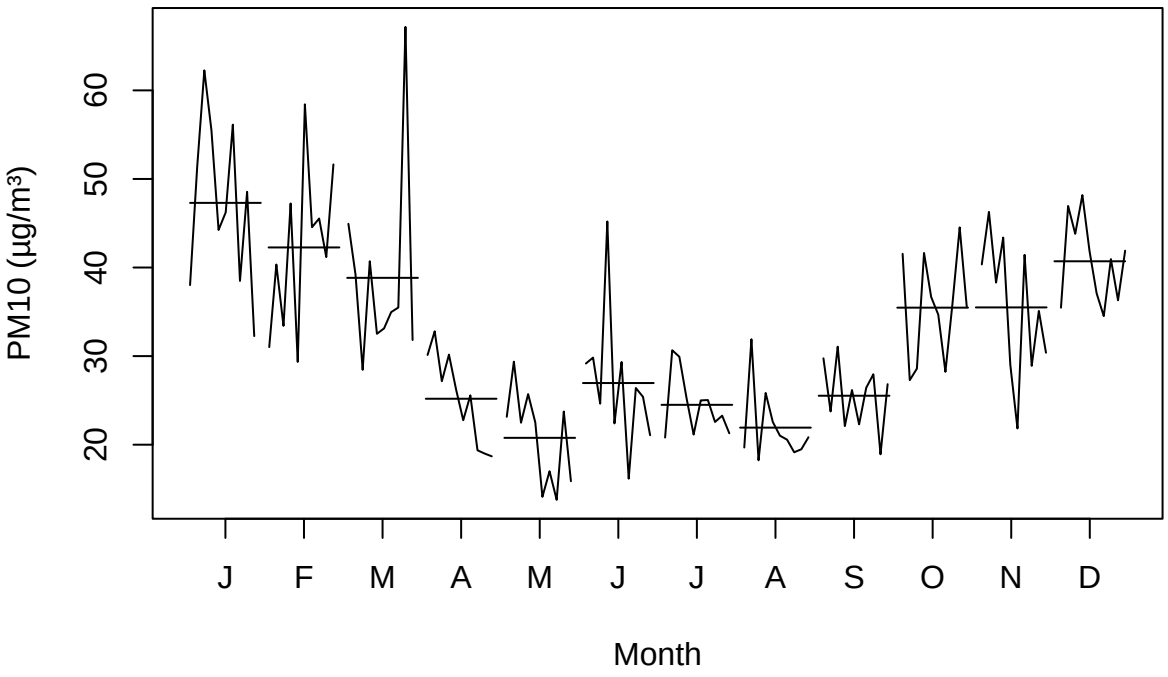
The monthplot function is a helpful tool for visualizing and analyzing the monthly patterns in a time series. It displays the average values for each month, making it easier to identify any seasonal trends. The resulting graph confirms the expected pattern: the observations show a pronounced monthly seasonality, with higher values typically occurring during the winter months and lower values during the summer. Additionally, the variability throughout the year is significant across all stations, indicating that the seasonal fluctuations are consistent yet varied in magnitude.

Monthly plot of PM10 for Milano v.Senato



Monthly plot of PM10 for Bormio v.Monte Braulio





Smoothing and decomposition

To better understand the trend component of the time series, this section employs smoothing techniques to estimate the underlying trend. Later, the time series from different stations will be decomposed using Seasonal and Trend decomposition using Loess (STL) to further isolate and highlight the trend, seasonal, and residual components.

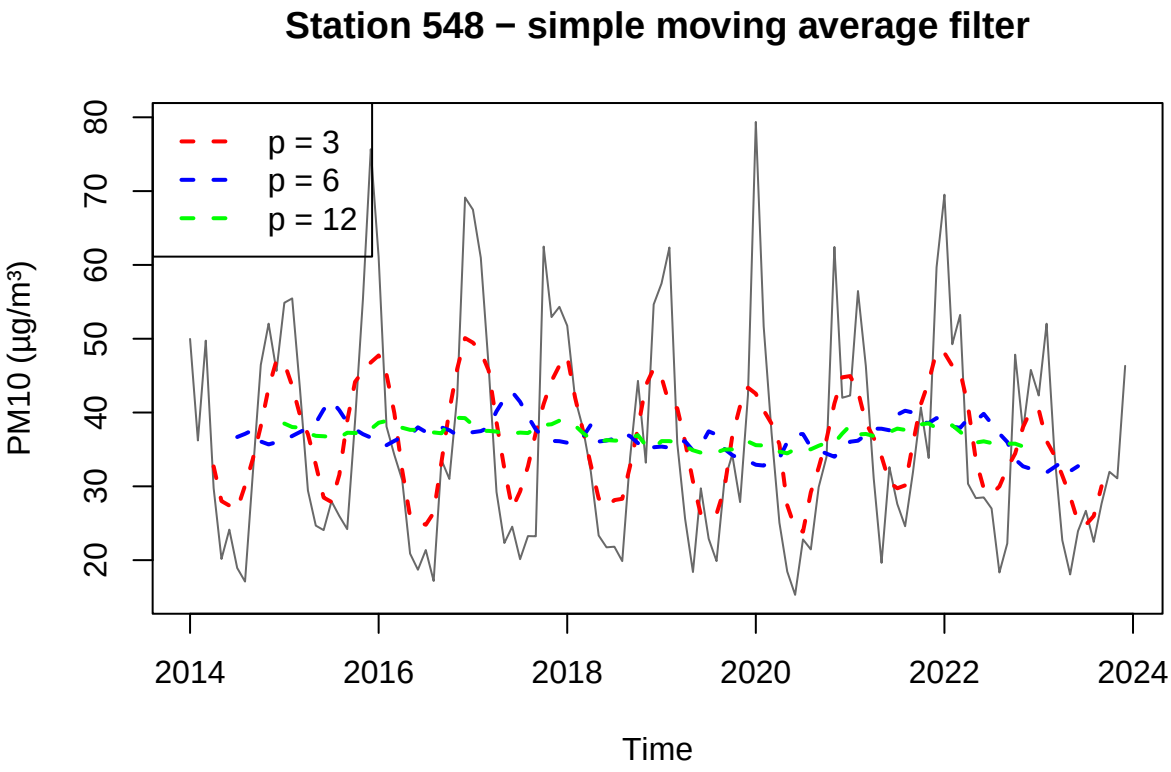
A widely used and straightforward method for smoothing is the **simple moving average** filter, which computes the arithmetic mean over a centered time window of size $2p + 1$. The filtered trend estimate at time t is given by:

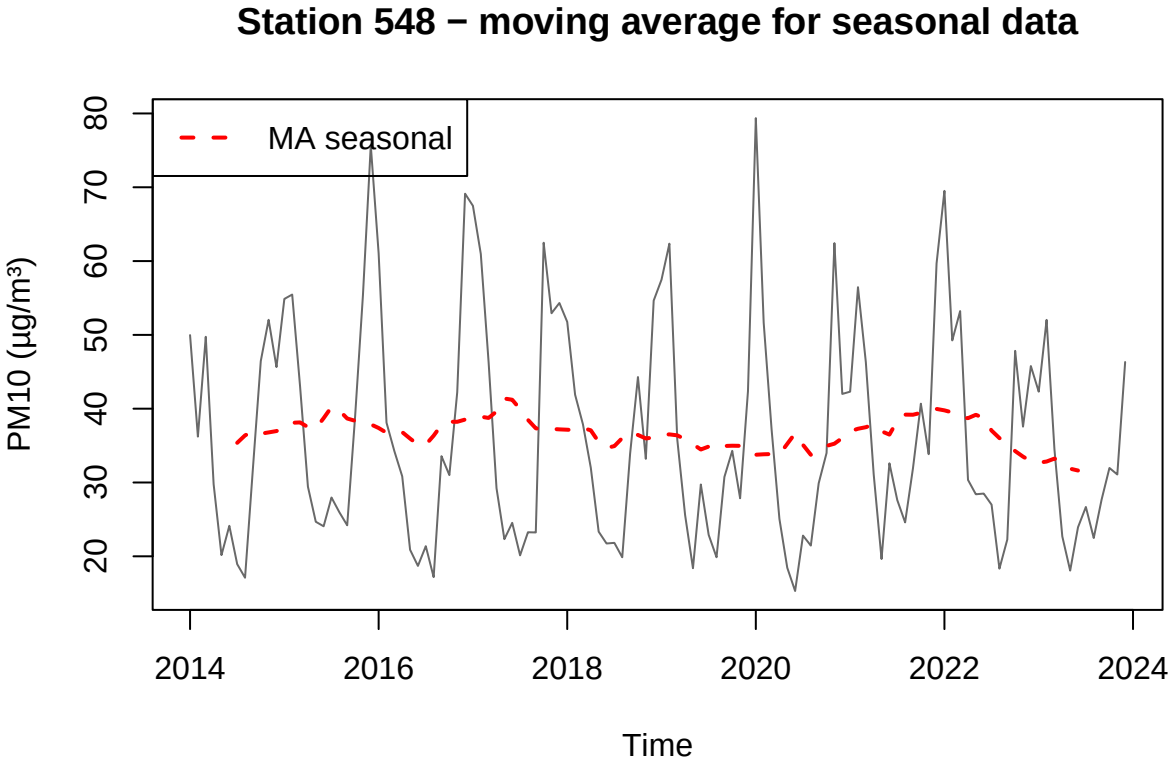
$$\hat{f}_t = \frac{1}{2p+1} \sum_{i=-p}^{p} y_{t+i}$$

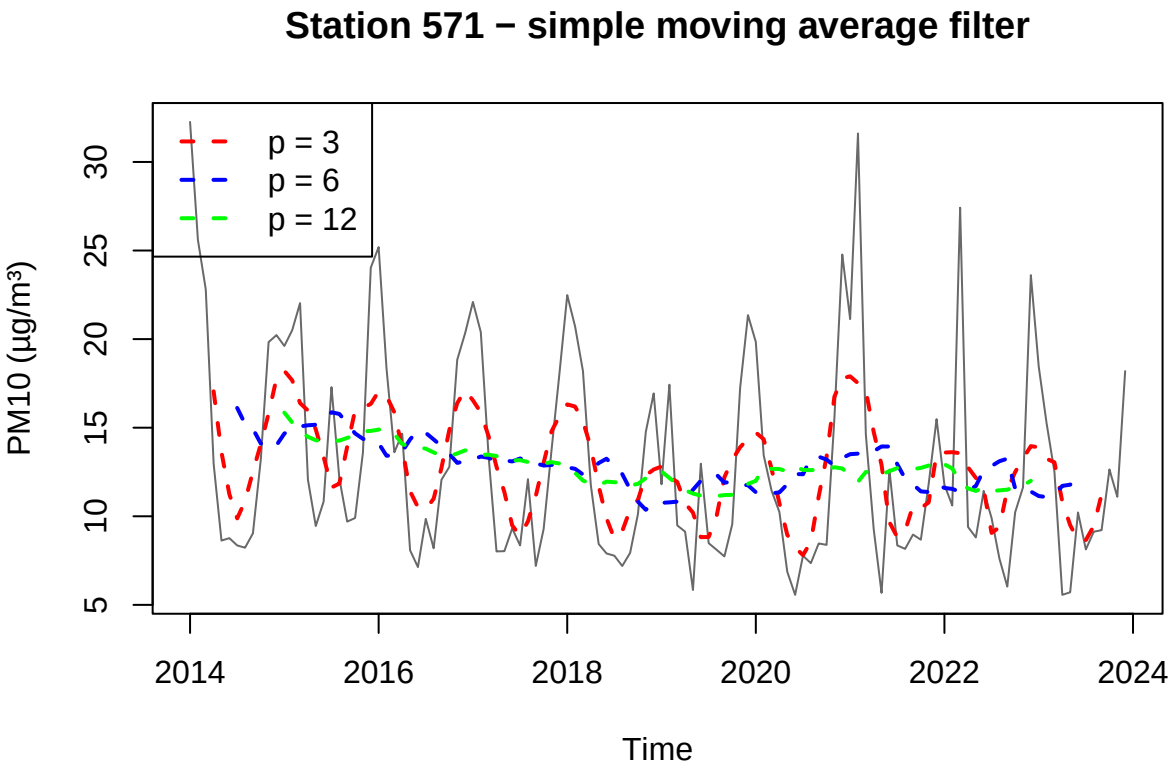
The choice of the window size p is crucial as it directly influences the degree of smoothing: larger values of p result in a smoother trend, while smaller values retain more of the original variability. Various values of p are tested to explore different levels of smoothing, each with a specific purpose:

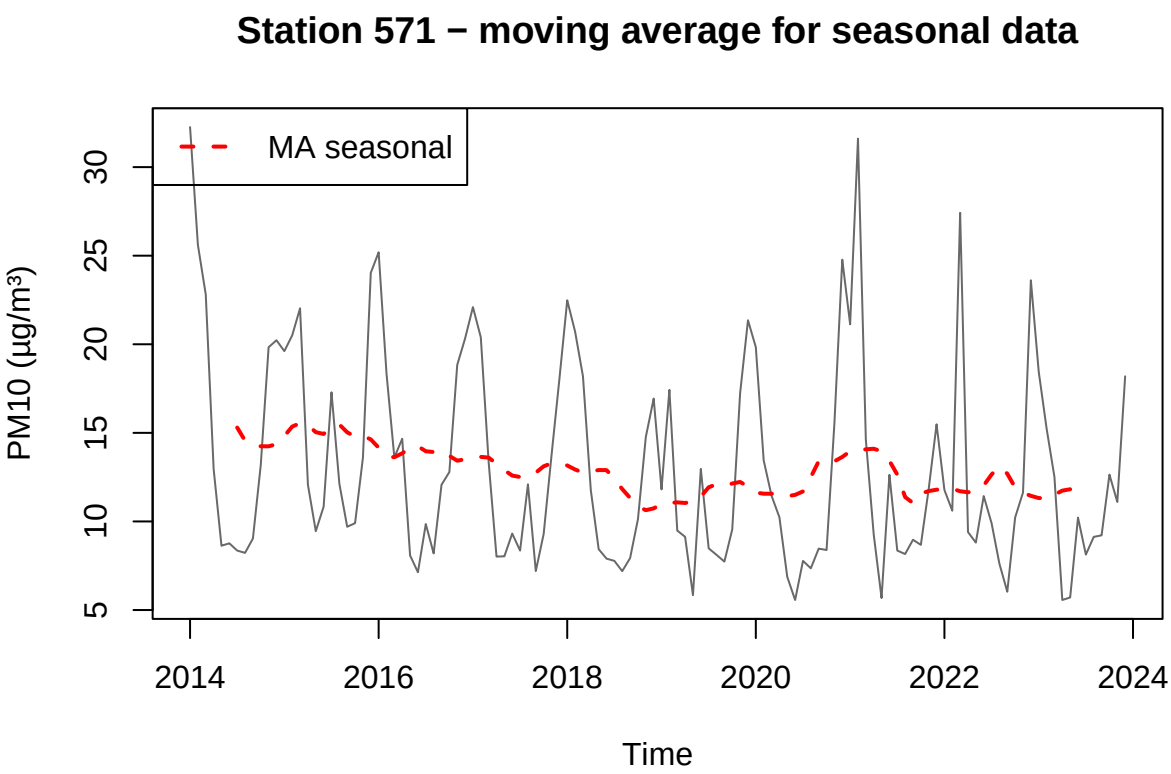
- $p = 3$**:** A smaller window that applies minimal smoothing, allowing short-term fluctuations to be visible while still reducing noise.
- $p = 6$**:** This value is chosen to remove the seasonal effect identified in the Auto-Correlation Function (ACF), particularly smoothing out variations that span over a half-year period.
- $p = 12$**:** A larger window size aimed at providing more significant smoothing, potentially eliminating yearly patterns and offering a clearer view of long-term trends.

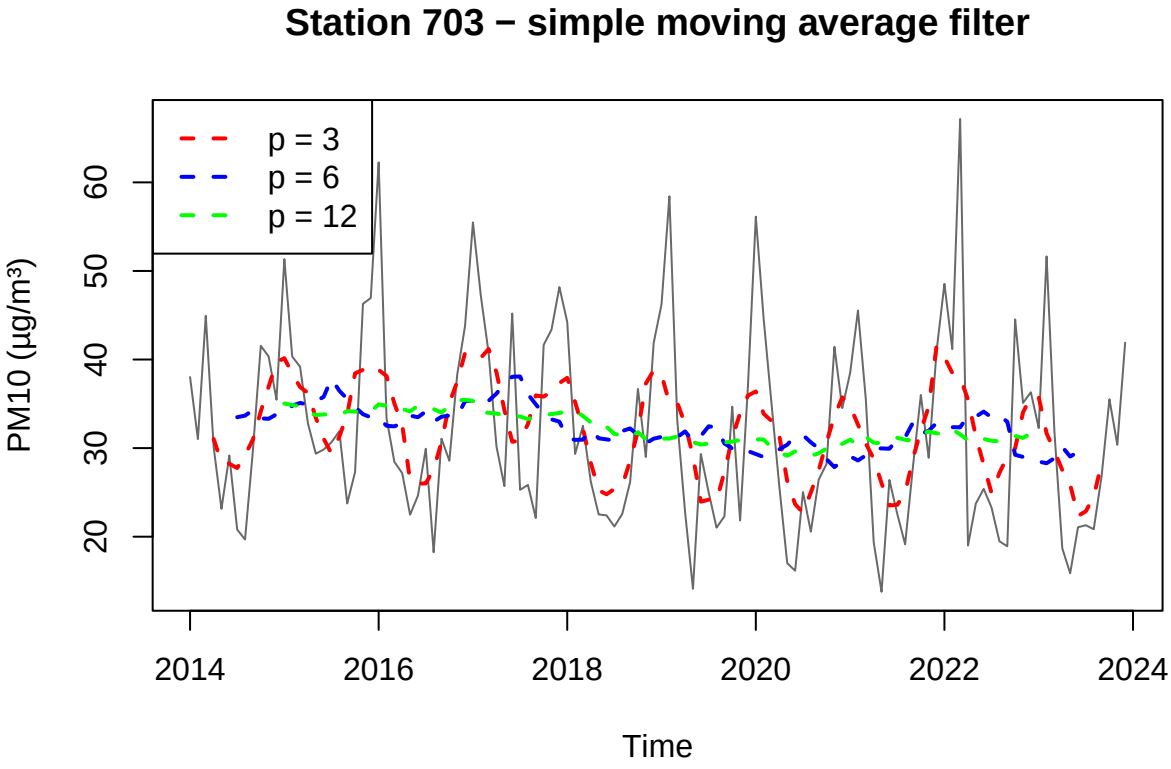
In addition, a moving average filter for seasonal data is tried to estimate the trend, given that the monthly time series exhibits a significant seasonal component.

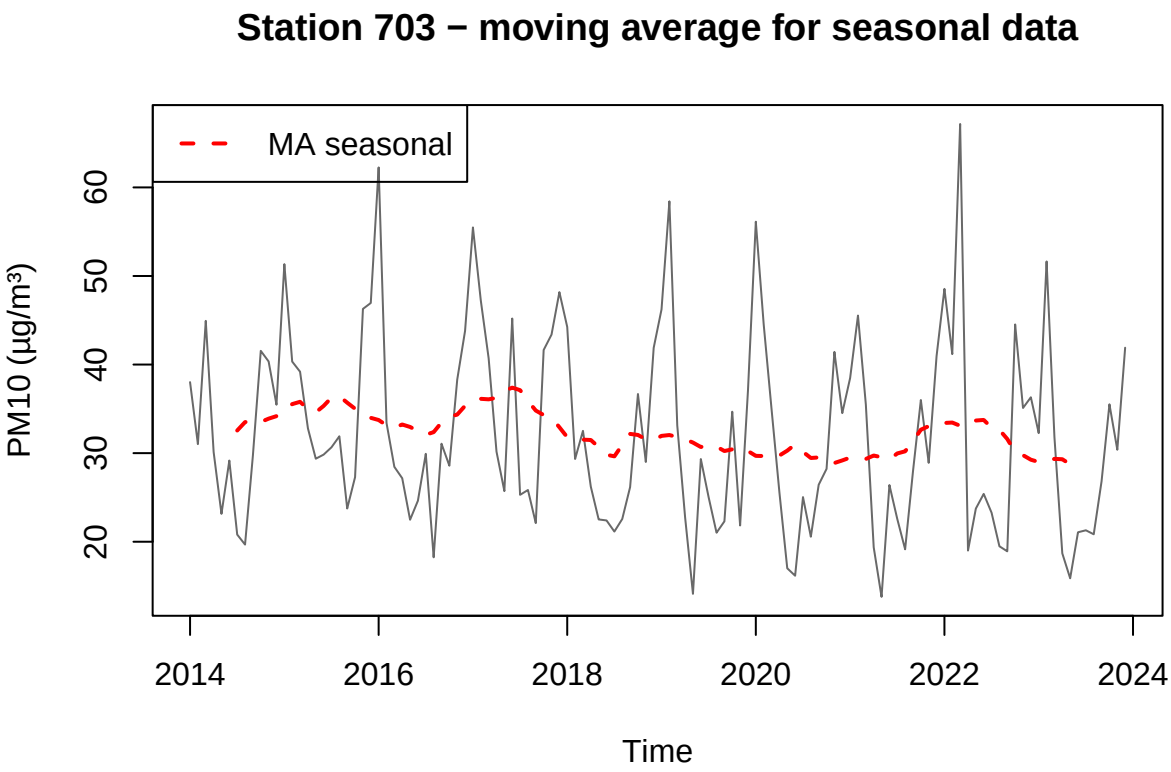

Station 548 − simple moving average filter
p = 3
p = 6
p = 12
PM10 (µg/m³)
20
30
40
50
60
70
80
2014
2016
2018
2020
2022
2024
Time


Station 548 − moving average for seasonal data
MA seasonal
PM10 (µg/m³)
20
30
40
50
60
70
80
2014
2016
2018
2020
2022
2024
Time


Station 571 − simple moving average filter
p = 3
p = 6
p = 12
PM10 (µg/m³)
5
10
15
20
25
30
2014
2016
2018
2020
2022
2024
Time

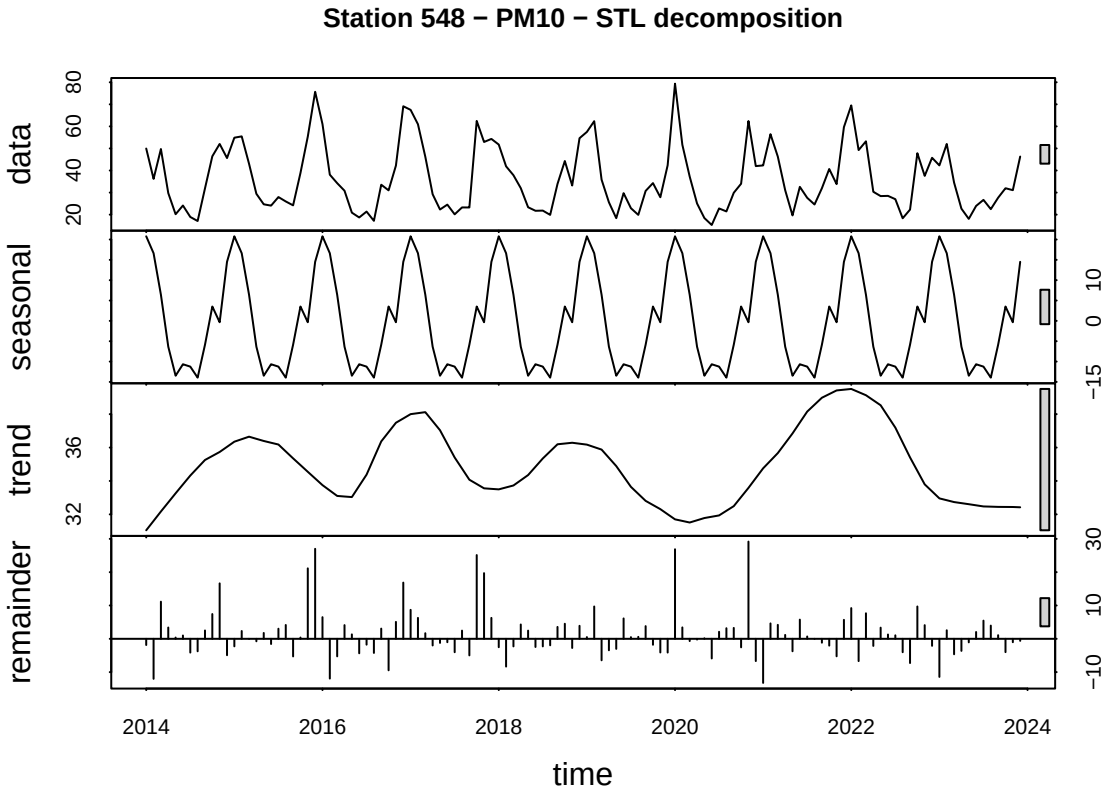

Station 571 − moving average for seasonal data
MA seasonal
PM10 (µg/m³)
5
10
15
20
25
30
2014
2016
2018
2020
2022
2024
Time

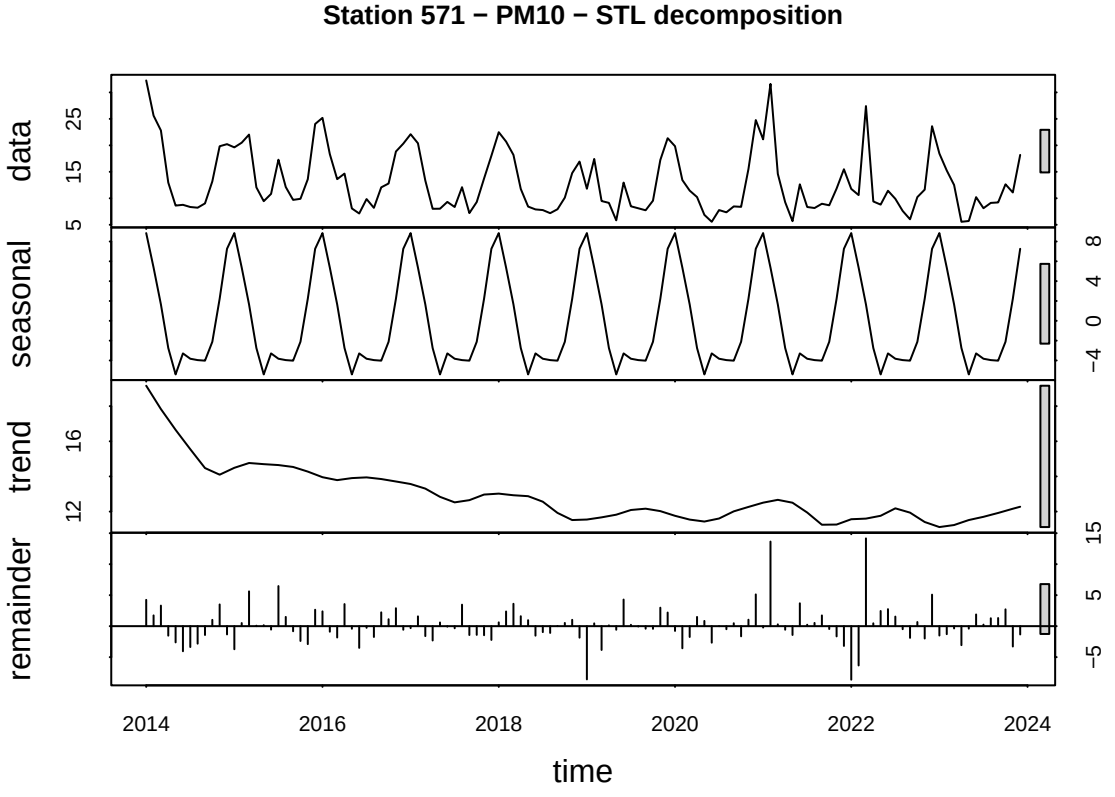


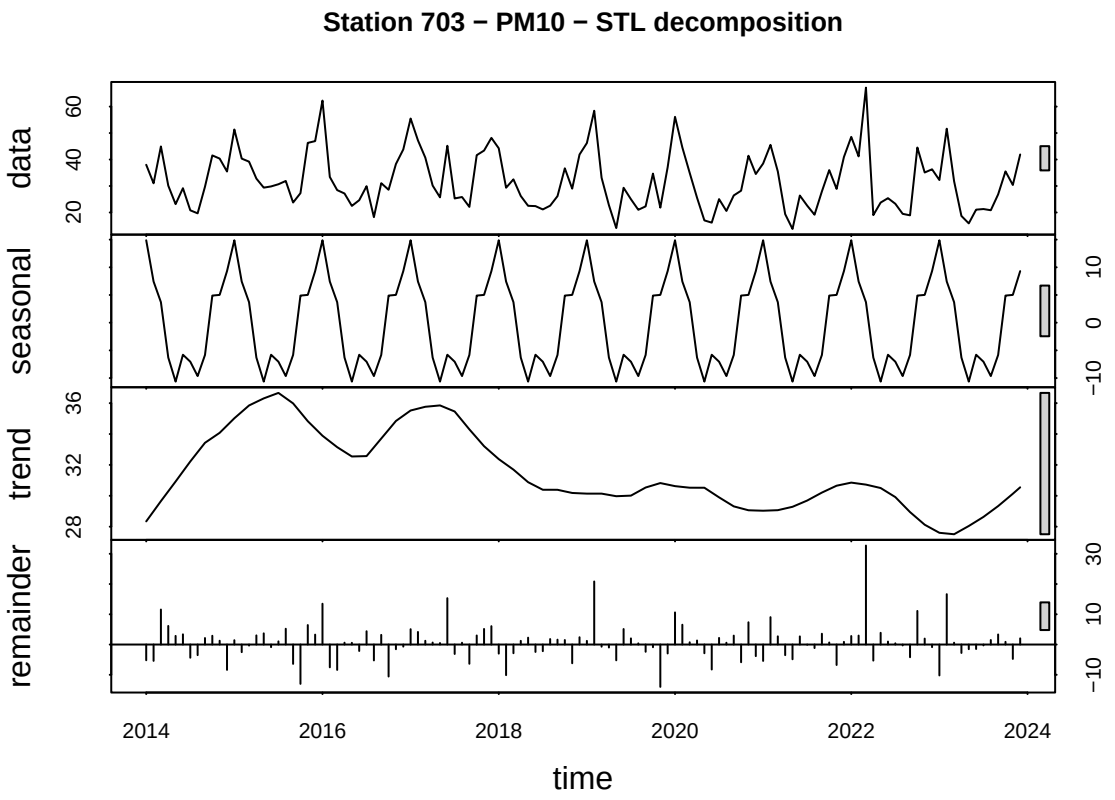


Overall, the estimated trend appears to be slowly decreasing. These data indicates also that during the COVID-19 restrictions in Italy from 2020 to 2022, the average PM10 levels did not significantly decrease. There were periods with increased PM10 values, such as at Station 571 in Bormio in 2021. This suggests that sources of particulate emissions unrelated to mobility, such as industrial activities or other local sources, played a substantial role in sustaining PM10 concentrations during this time.

To further inspect the behavior of the time series, STL (Seasonal and Trend decomposition using Loess) decomposition is applied. Given the previously observed anomalies, the robust version of the algorithm is used to mitigate their impact. STL also offers flexibility in defining the rate of change for the seasonal component. Since the seasonal pattern appears consistent over time, the seasonal window is set to "periodic" to ensure that the entire dataset is utilized for a comprehensive seasonal analysis.







All stations exhibit a strong seasonal component, with an overall decreasing trend. However, Station 548 shows a notable exception, as it experienced significant peaks in PM10 levels at the end of 2021 and throughout 2022.

The Ljung-Box test supports the validity of the decompositions, as it does not suggest rejecting the null hypothesis that the residuals are white noise for most stations. However, for station 571, the p-value is notably low, indicating that some adjustments to the parameters in the STL function might be necessary to improve the decomposition accuracy.

Table 7: Ljung-Box test for the noise component

	p-value
Station 548	0.1766
Station 571	0.0006
Station 703	0.2471

Models development

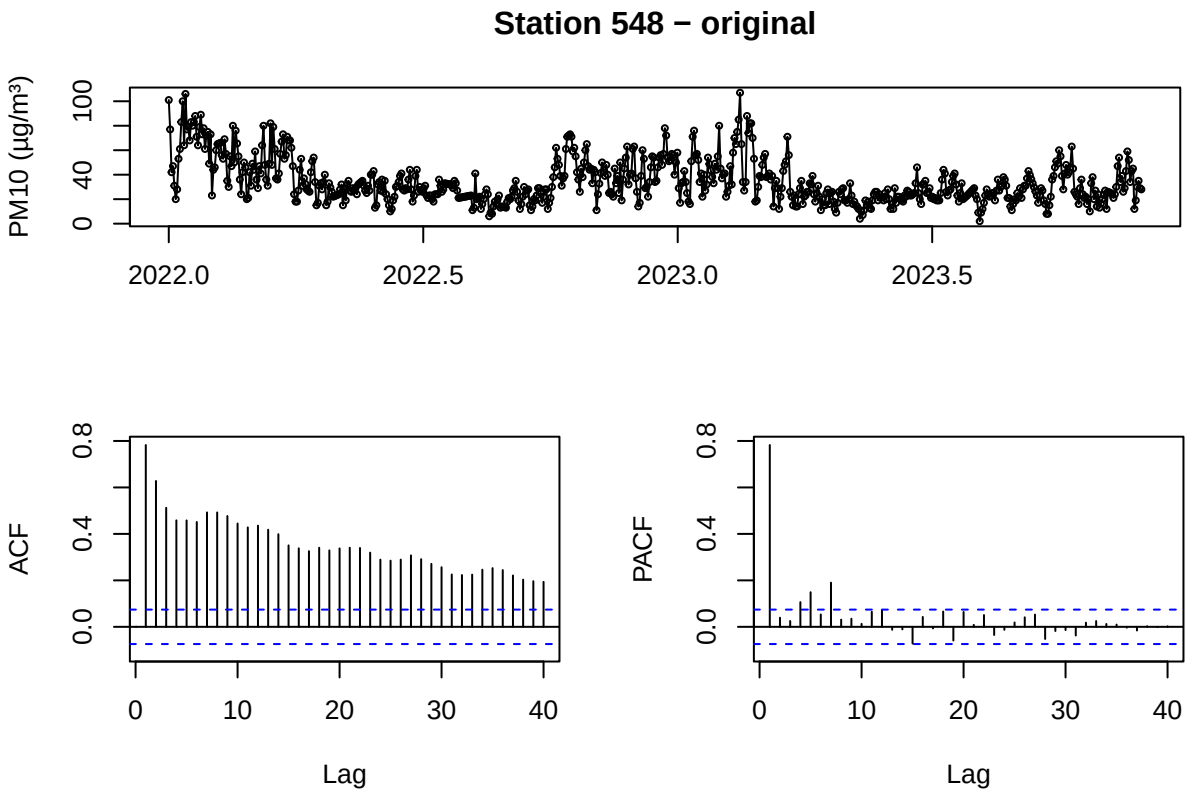
As previously mentioned, this section of the analysis focuses on daily data to develop forecasting models. Given the need to predict future PM10 values for implementing preventive health measures, a short-term analysis is essential.

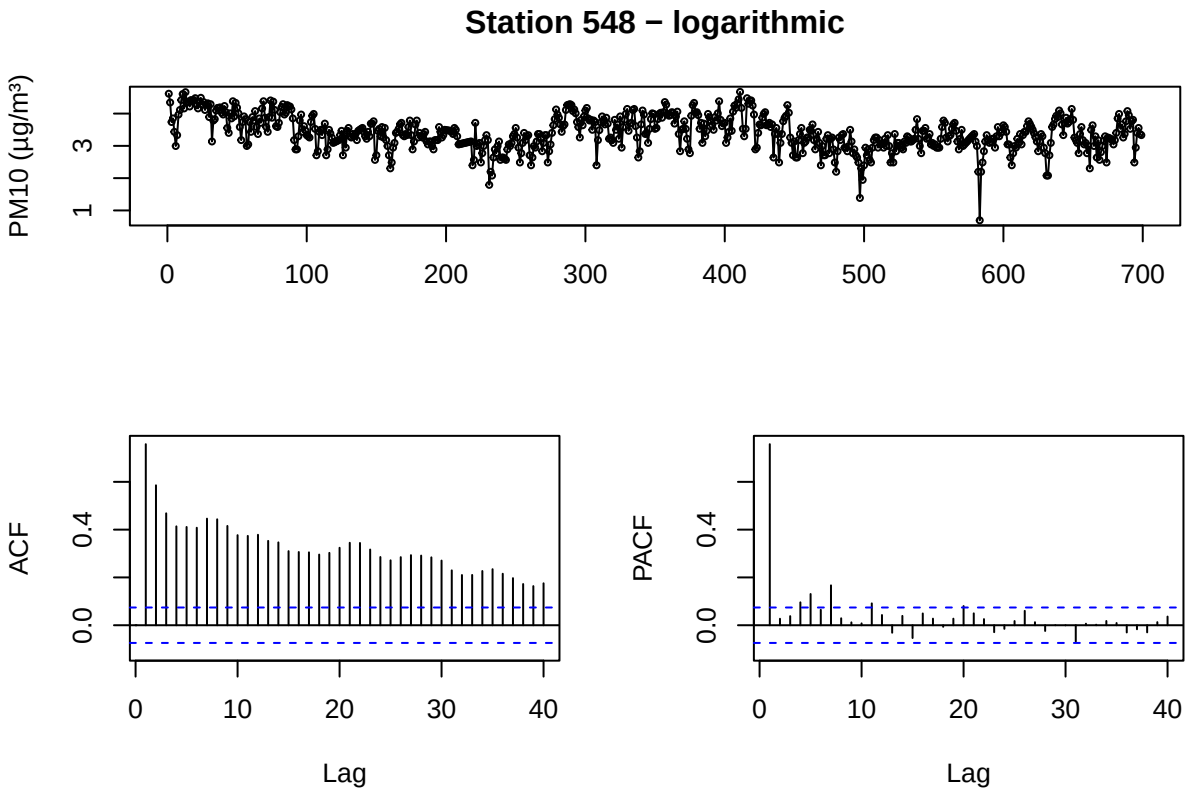
To simplify visualization and focus on recent data, the time series for this part of the study is limited to the period from January 1, 2022, to December 31, 2023. As shown previously, the choice of a time window at the end of the COVID-19 emergency period in Italy doesn't influence the PM10 values pattern

For an accurate evaluation, it is crucial to avoid using forecast data as training data. Therefore, the time series is divided into training and test sets, with the test period spanning from December 1, 2023, to the end of the period. A one-month test set is selected to assess how the model performs with a relatively long forecast horizon.

Stochastic models

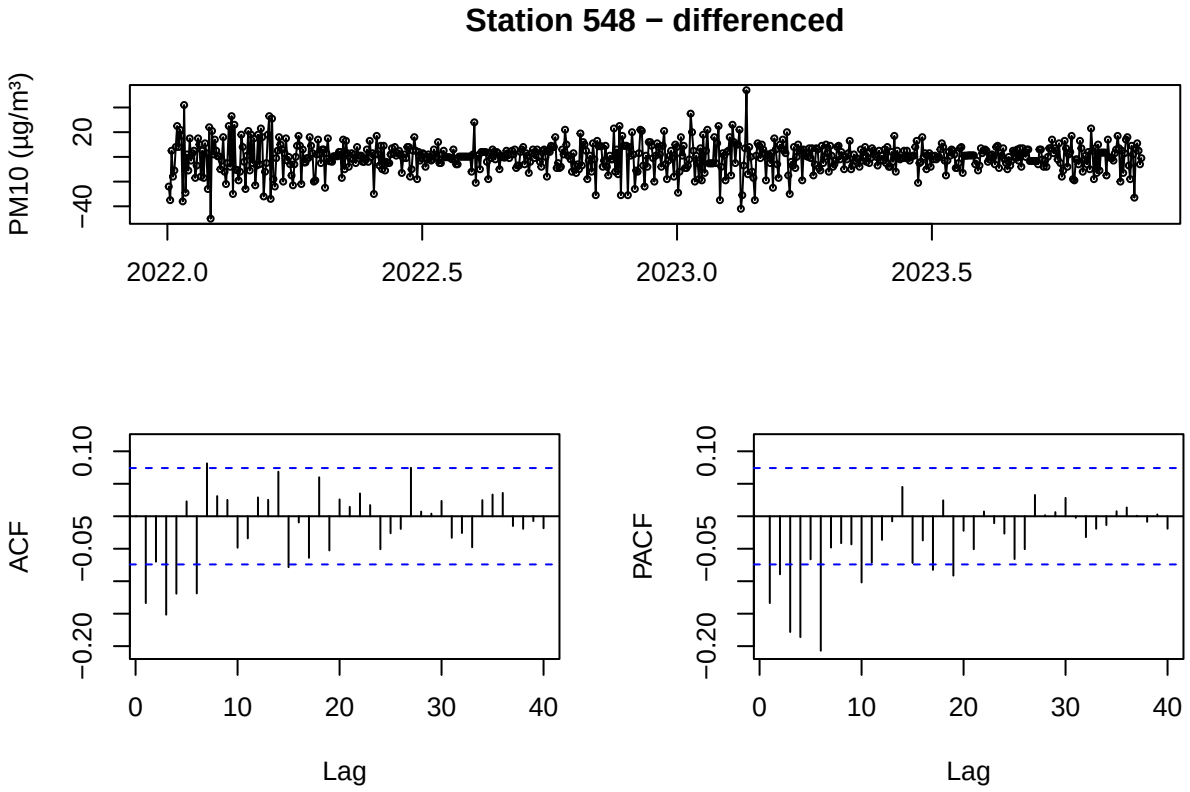
Station 548 - Milano v.Senato

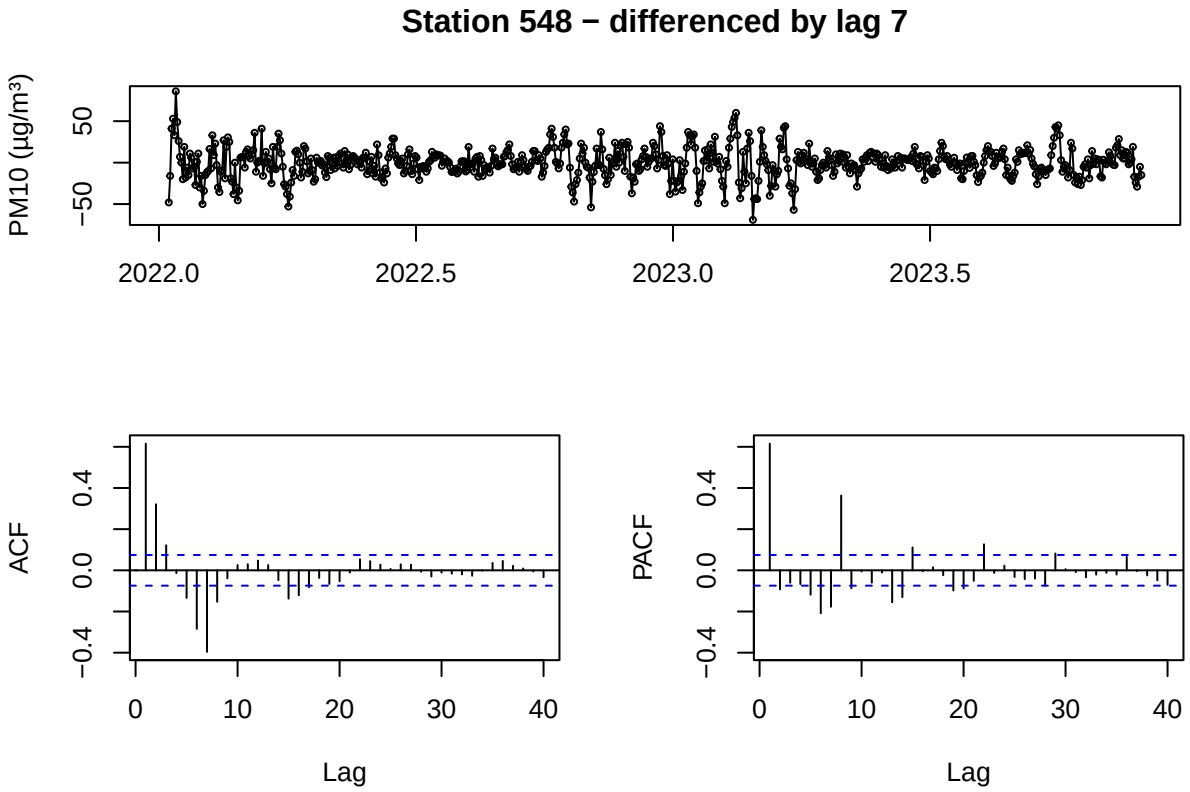


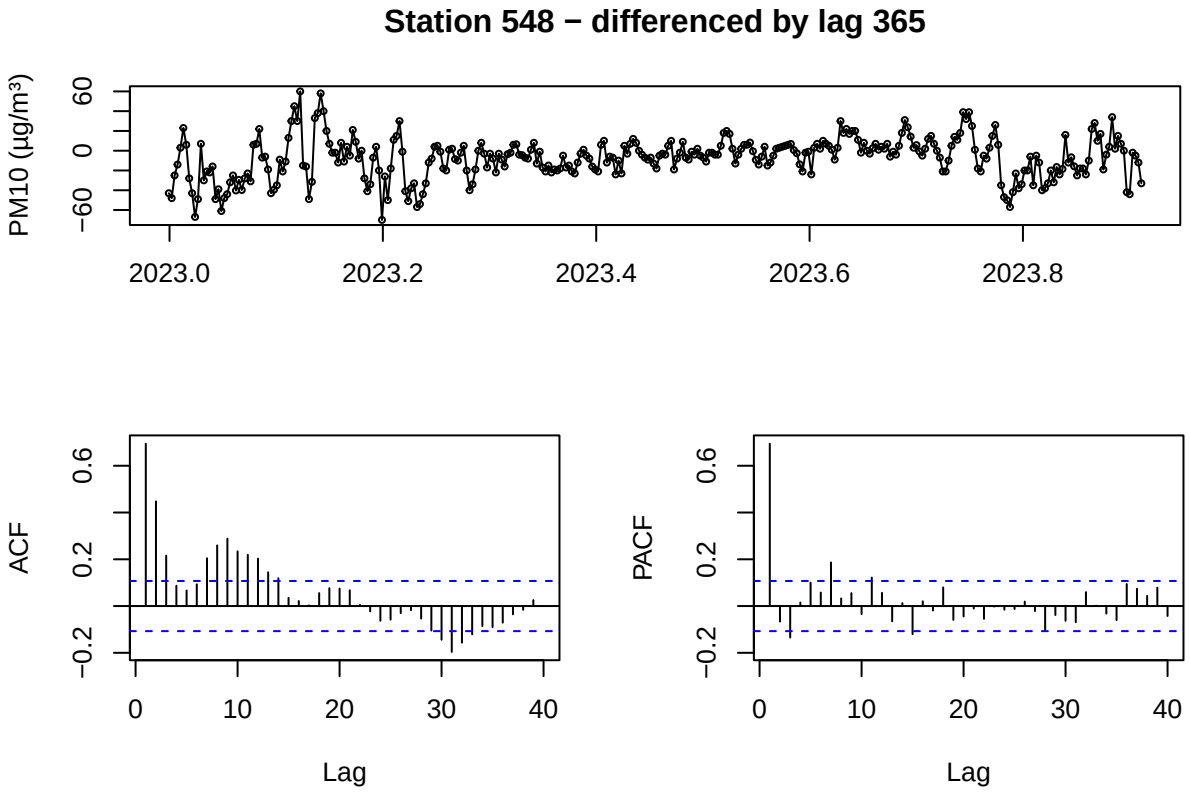


The plots indicate that the time series is not stationary, as evidenced by the slow decay of lags in the ACF plot. Even after

applying a logarithmic transformation to stabilize the variance, the overall situation remains largely unchanged. The ACF plot displays up to 40 lags since, beyond a month of daily correlations, the lags may become insignificant.

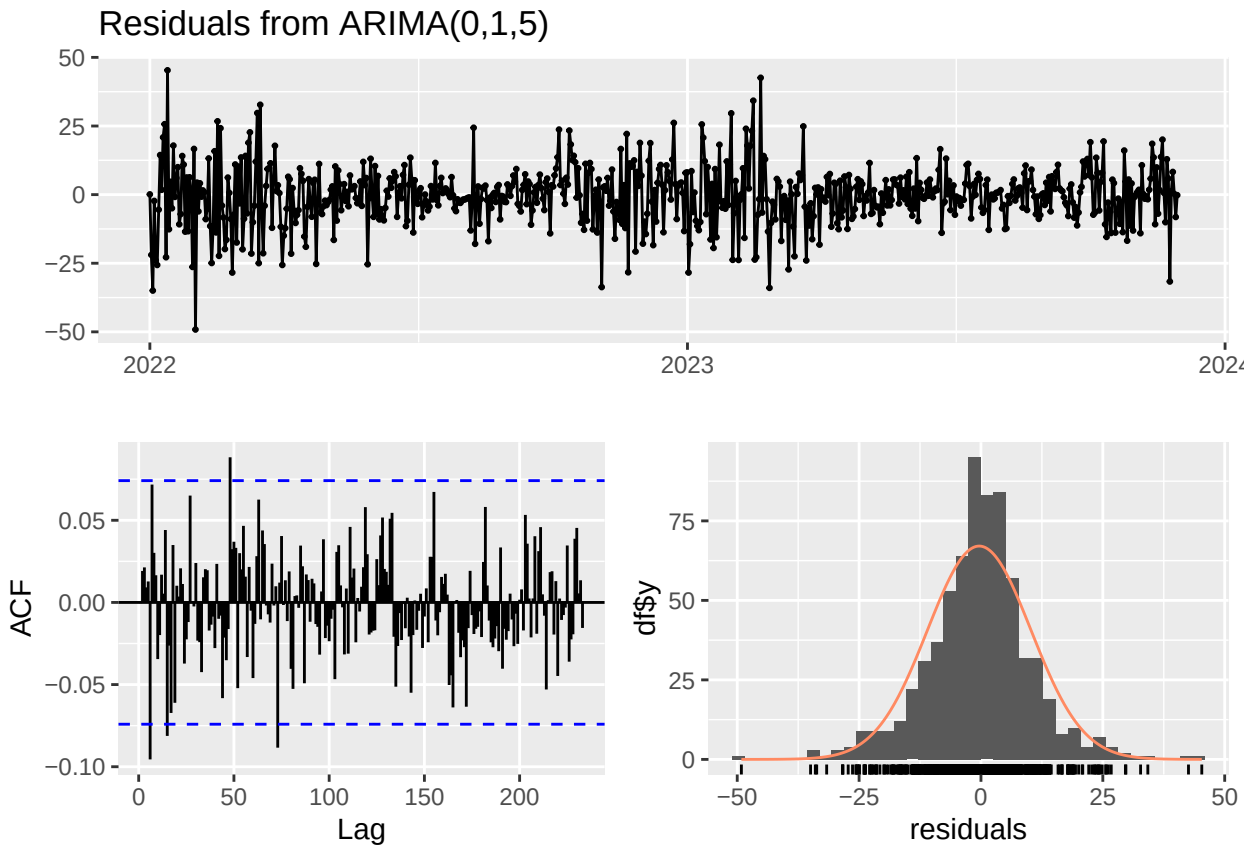




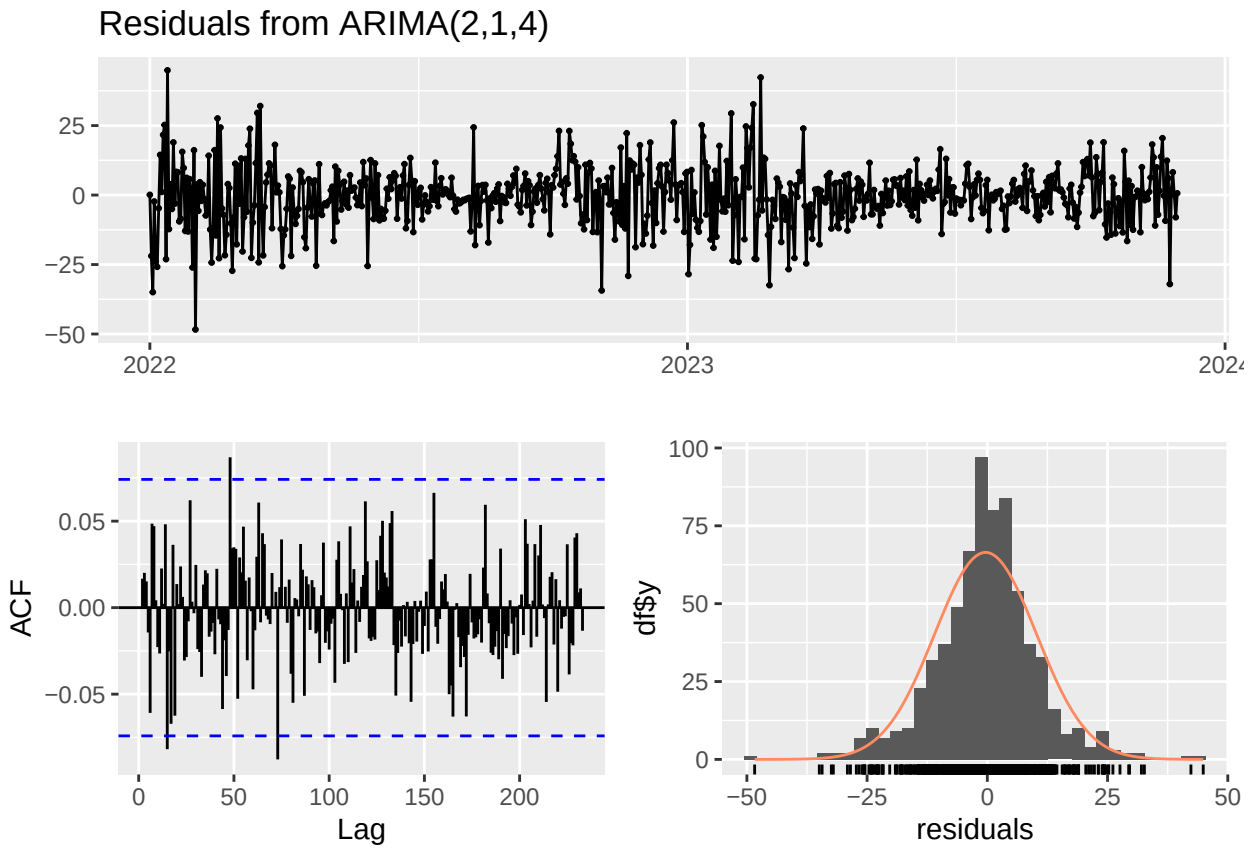


Differencing the time series appears to be highly effective, as the data now seem better aligned with the stationarity assumption. However, differencing by lag 7 proves unhelpful for improving stationarity and introduces an artificial seasonal pattern into the data. Similarly, differencing by the time series frequency results in the appearance of a seasonal pattern, with sinusoidal oscillations visible in the ACF plot. Despite this, applying differencing to remove potential periodic factors in daily observations can be impractical and risky. This approach overlooks time domains like months or weeks, where specific patterns are expected over longer periods. The variability between days across different years—affected by factors such as weather or the day of the week—renders such differencing uninformative and difficult to apply effectively.

Given that the ACF plot of the differenced time series shows more pronounced spikes up to lag 5 and the PACF exhibits a relatively fast decay, an ARIMA model with order (0, 1, 5) is fitted. The residuals are then examined to confirm they resemble white noise, ensuring the model's adequacy. Finally, this model is compared with an automatically selected ARIMA model to assess its performance.




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,5)
## Q* = 110, df = 135, p-value = 0.9
##
## Model df: 5.   Total lags used: 140
```

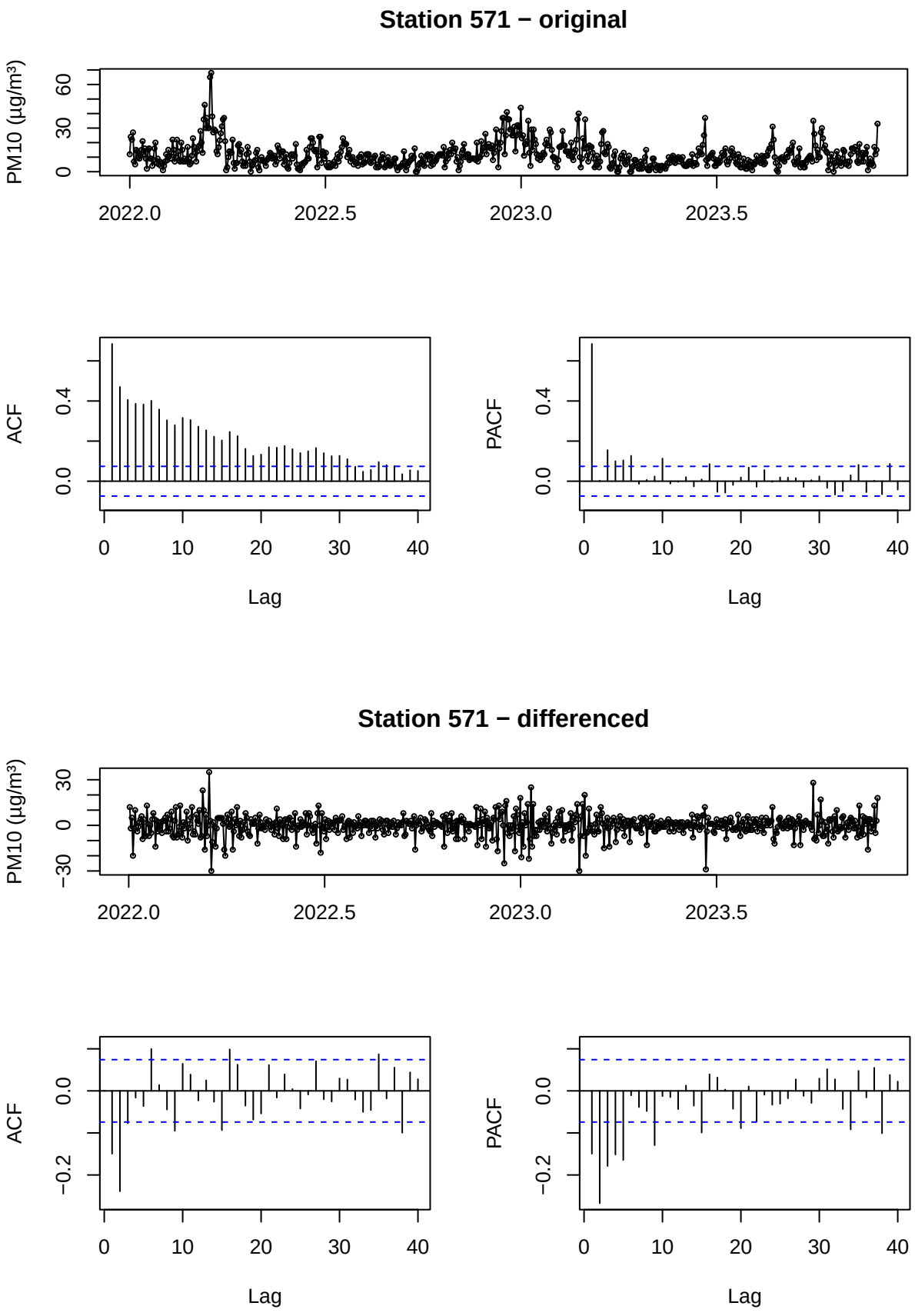



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(2,1,4)
## Q* = 105, df = 134, p-value = 1
##
## Model df: 6.   Total lags used: 140
```

The residuals of the automatically selected ARIMA model appear preferable. The Ljung-Box test yields a higher p-value, and both the residuals plot and ACF suggest that the residuals are closer to white noise.

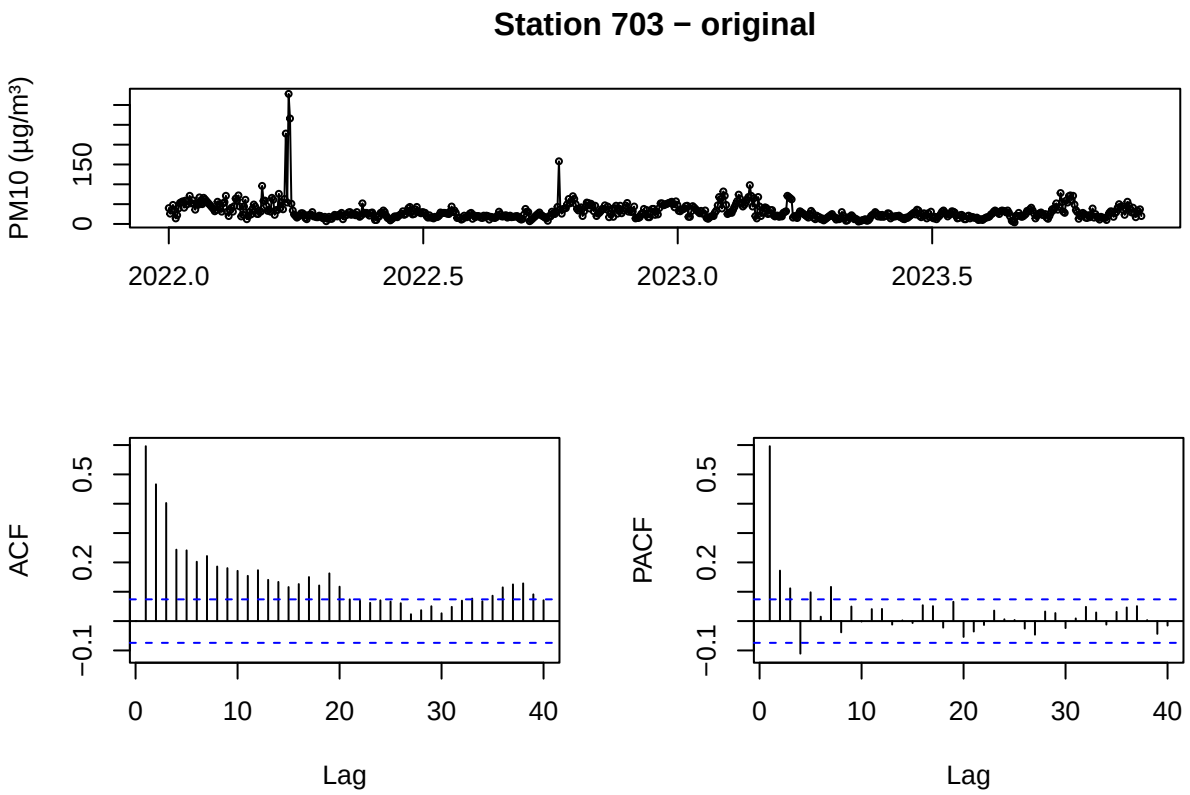
This procedure is also applied to the other stations. The time series are differenced once to improve stationarity, and the plots are inspected to determine the most suitable ARIMA model for each. In all cases, the models are compared with those selected automatically. While the residuals for both methods generally align with the assumptions, the automatically selected models consistently perform slightly better.

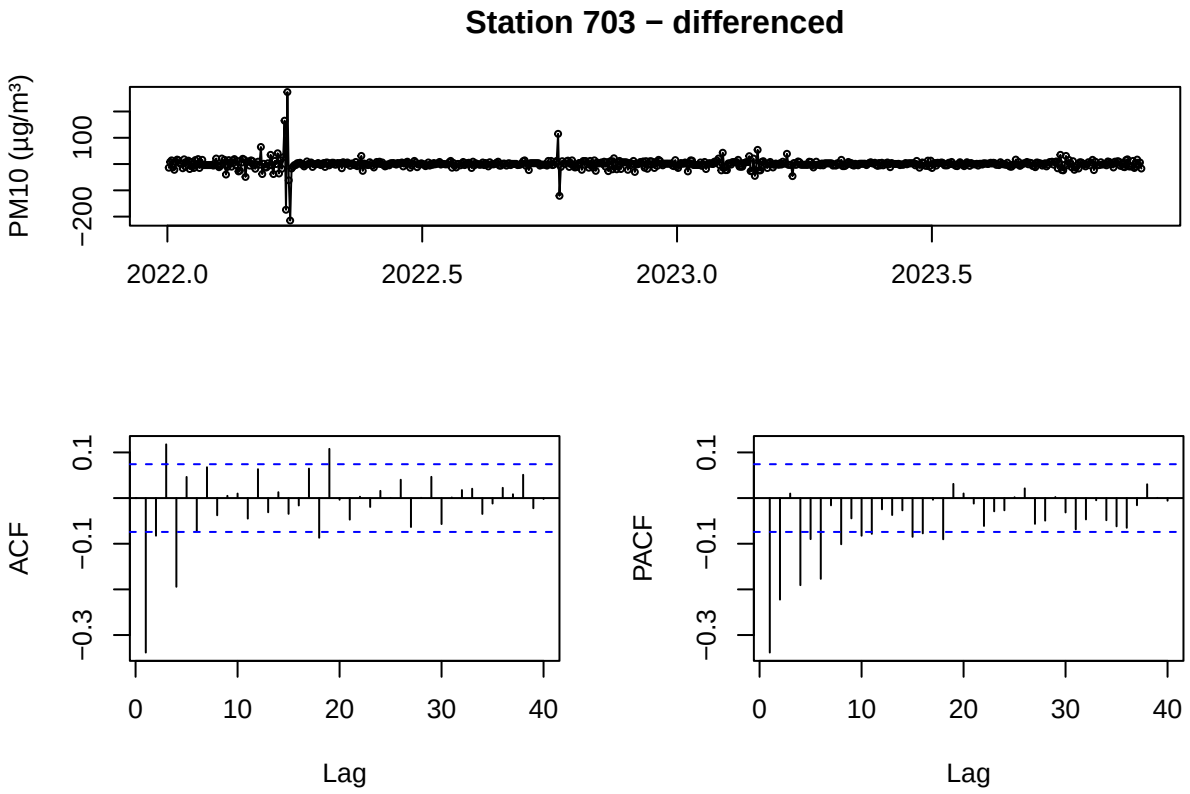
Station 571 - Bormio v.Monte Braulio



For the data from Station 571, more pronounced spikes are observed up to lag 2, along with indications of an autoregressive pattern in the additional spikes. As a result, an ARIMA(1,1,2) model is fitted to capture these characteristics.

Station 703 - Schivenoglia v. Malpasso



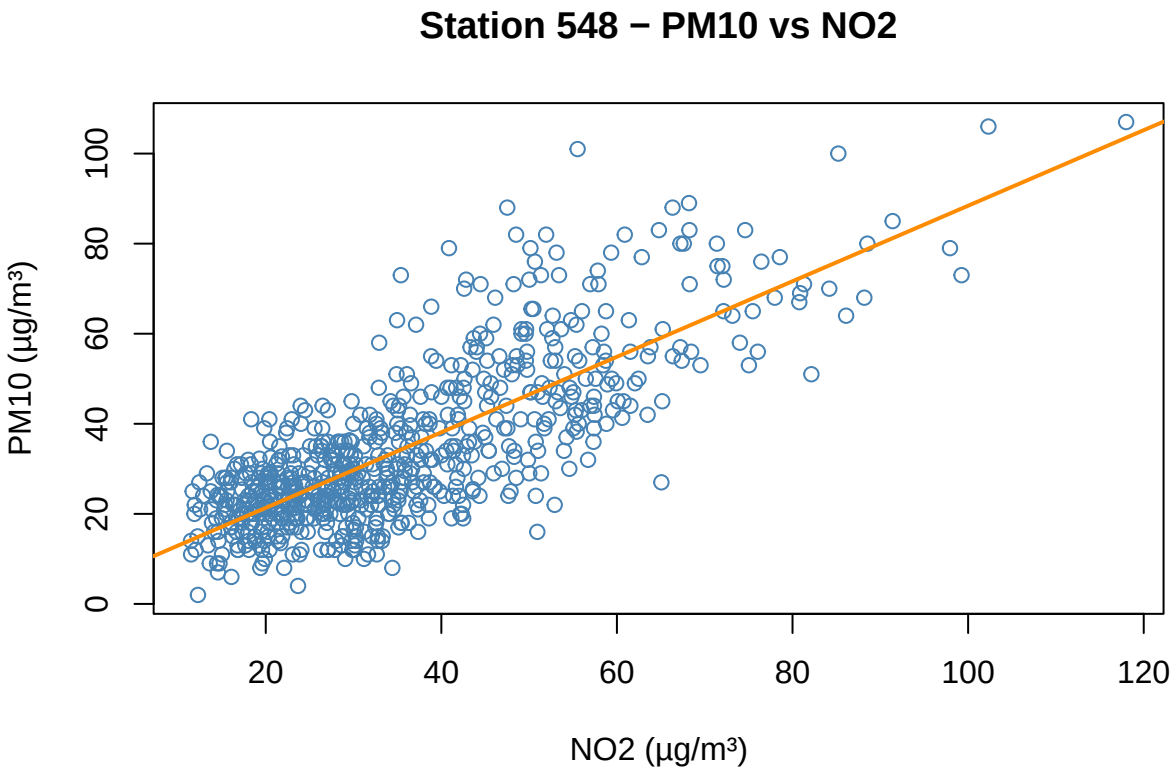


The same observation on spikes is applied also to the data of station 703.

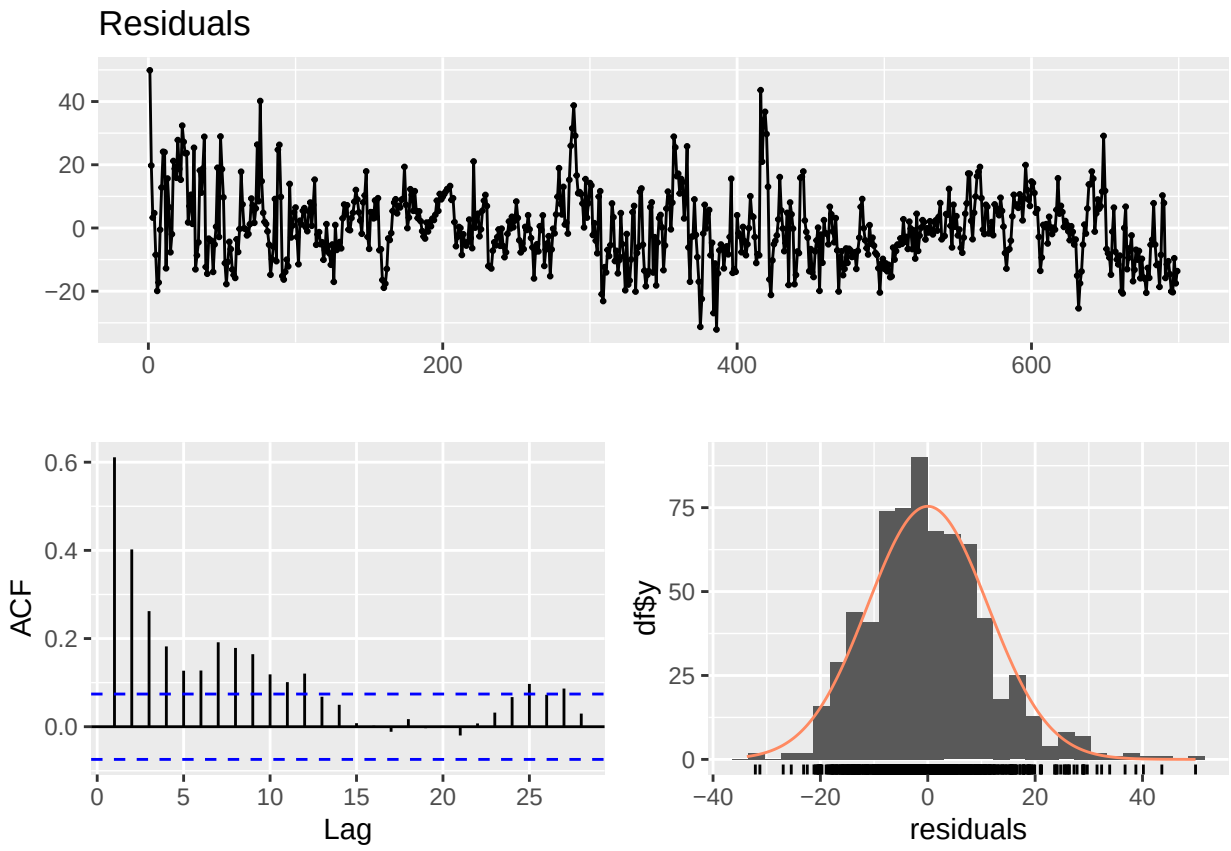
Dynamic regression

This section explores the potential benefits of including other pollutants in predicting PM10. NO_2 is a key precursor in particulate matter formation, as it reacts in the atmosphere to create secondary particles. NO_x (which includes both NO and NO_2) also significantly contributes to secondary particulate matter, originating mainly from combustion processes like vehicle and industrial emissions. While CO is primarily a gas, it can indirectly affect PM10 levels through atmospheric reactions that produce secondary pollutants, including particulate matter.

Inspecting the relation between NO2 and PM10 is decided, as NO2 appears to be the most relevant predictor. Missing values are imputed as before using linear interpolation.



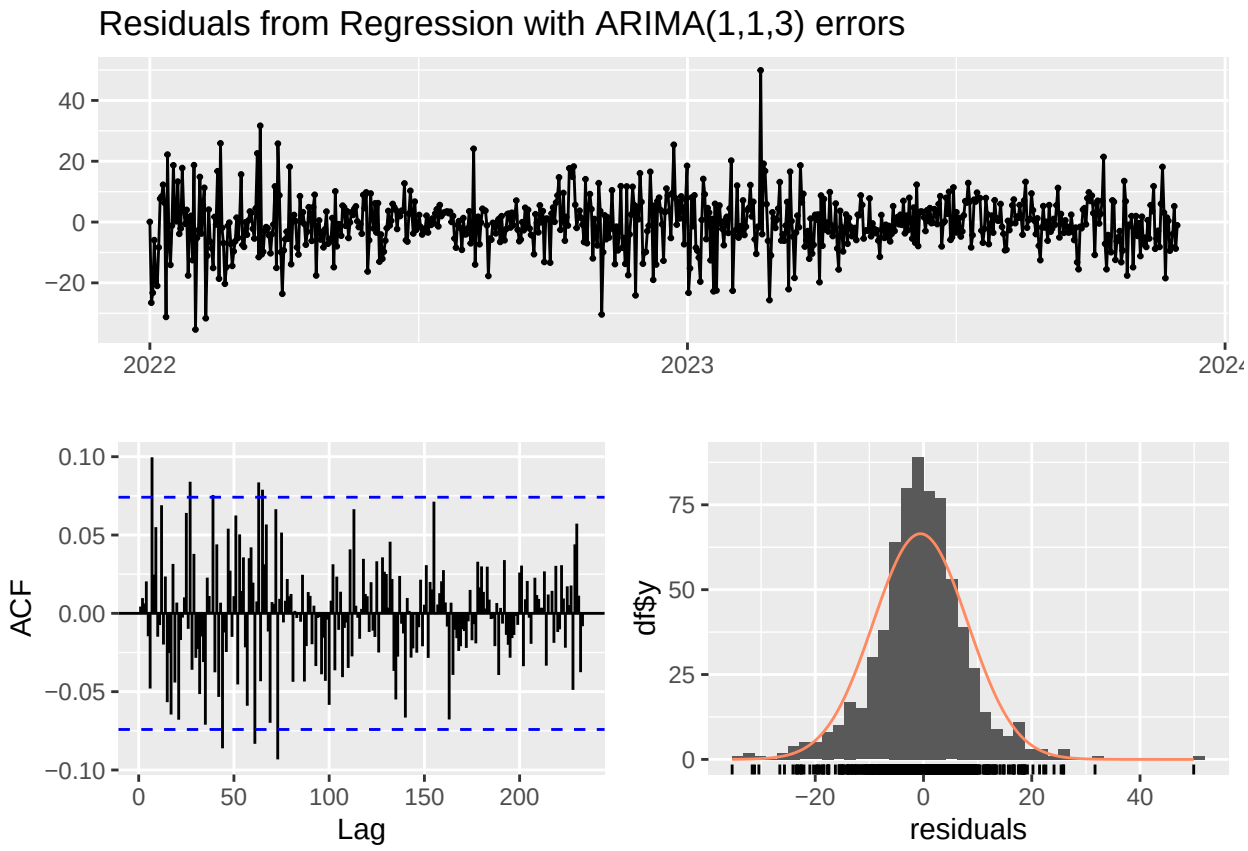

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32.12  -7.24  -0.71   6.71  49.88
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5635     1.0010    4.56  6.1e-06 ***
## x             0.8385     0.0255   32.86  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.2 on 697 degrees of freedom
## Multiple R-squared:  0.608,  Adjusted R-squared:  0.607
## F-statistic: 1.08e+03 on 1 and 697 DF,  p-value: <2e-16
```




```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 548, df = 10, p-value <2e-16
##
## Model df: 0.   Total lags used: 10
```

For Station 548, a linear relationship between NO_2 and PM10 is evident, supported by a linear regression model where the predictor is highly significant. However, the residuals violate the white noise assumption. Similar results are observed at other stations, though for brevity, these are not presented here.

A dynamic regression model is now fitted to the time series data using NO_2 as a predictor. The model is estimated using the `auto.arima` function, which selects the optimal ARIMA model based on the corrected Akaike Information Criterion (AICc). Residuals are then assessed for white noise using the `checkresiduals` function. Results for other stations are not shown for brevity but residuals appear satisfying.




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(1,1,3) errors
## Q* = 154, df = 136, p-value = 0.1
##
## Model df: 4.   Total lags used: 140
```

Residuals are in this case similar to white noise and the Ljung-Box test suggests that there's no evidence for rejecting this hypothesis.

ARIMA comparison

Table 8: Station 548 - ARIMA models IC values

	AIC	AICc	BIC
Regression with ARIMA(1,1,3) errors [auto]	5021	5021	5049
ARIMA(2,1,4) [auto]	5290	5290	5321
ARIMA(0,1,5)	5292	5292	5319

Table 9: Station 571 - ARIMA models IC values

	AIC	AICc	BIC
Regression with ARIMA(1,0,3) errors [auto]	4153	4153	4185
ARIMA(3,0,1) with non-zero mean [auto]	4486	4486	4513
ARIMA(1,1,2)	4489	4489	4507

Table 10: Station 703 - ARIMA models IC values

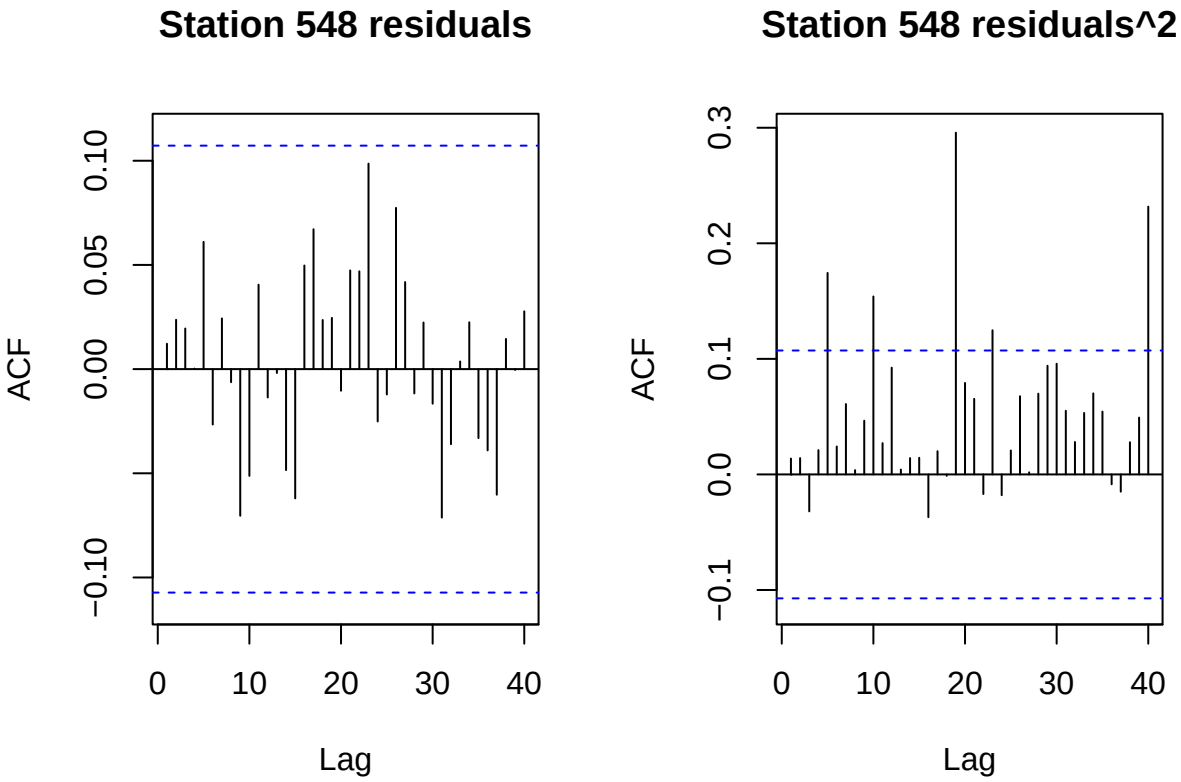
	AIC	AICc	BIC
Regression with ARIMA(3,0,1) errors [auto]	5952	5952	5983
ARIMA(3,1,2) [auto]	6002	6002	6029
ARIMA(1,1,3)	6023	6023	6046

In all the stations, models that use Nitrogen Dioxide (NO_2) as a predictor exhibit lower AICc values. Additionally, the BIC, which imposes a greater penalty for model complexity, is also lower for these models.

Non linear models

A nonlinear model is used to fit the data, specifically a neural network autoregressive model. This model is a feedforward neural network with a single hidden layer, estimated using the `nnetar()` function, which automatically selects the optimal neural network configuration. For non-seasonal data, the fitted model is represented as an $NNAR(p, k)$ model, where k denotes the number of hidden nodes. This model is analogous to an AR(p) model but incorporates nonlinear functions. For seasonal data, the model is denoted as an $NNAR(p, P, k)[m]$, analogous to an $ARIMA(p, 0, 0)(P, 0, 0)[m]$ model but with nonlinear components. According to the *Universal Approximation Theorem* [2], a neural network with a single hidden layer can approximate any continuous function on a compact subset, making it a powerful tool despite its reduced interpretability.

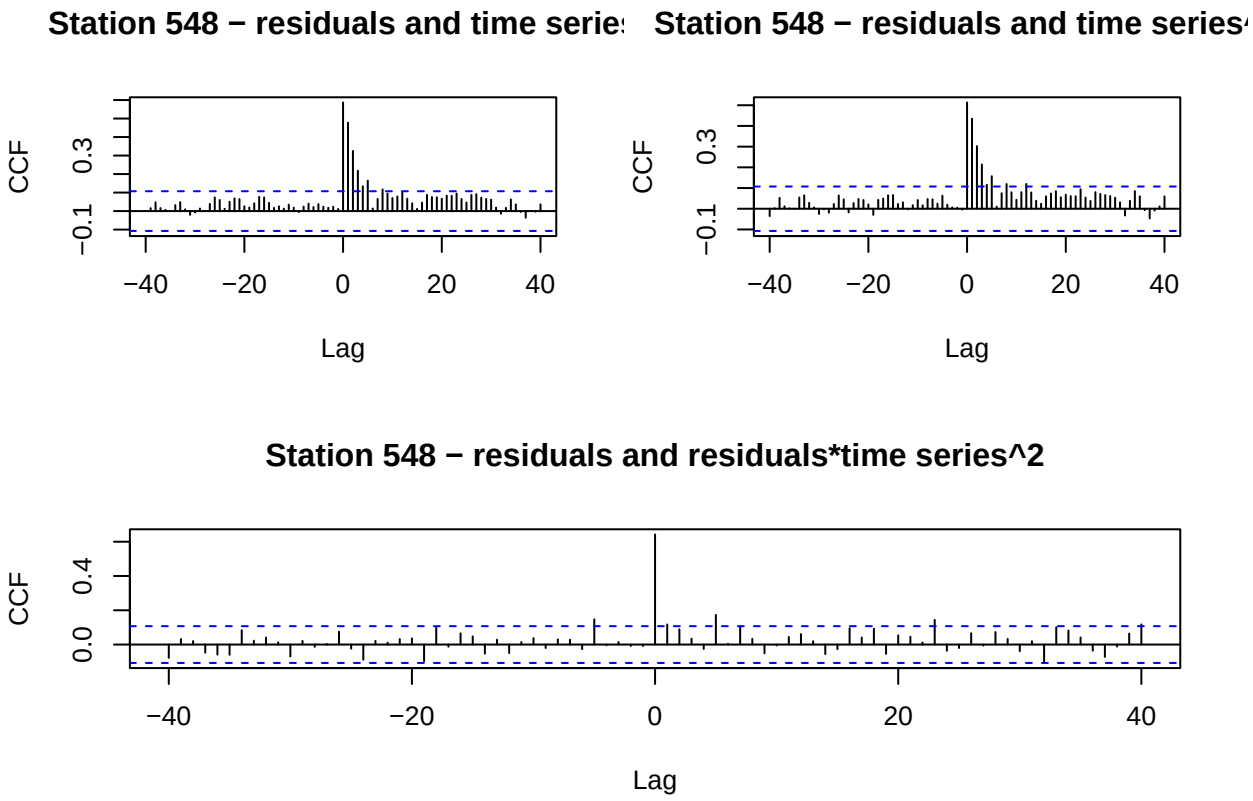
For nonlinear models, traditional residual diagnostics using autocorrelation functions may not be sufficient to assess model validity. Therefore, additional types of correlation are examined to ensure a comprehensive evaluation of the model's performance.




```
##
##  Box-Ljung test
##
## data:  r
## X-squared = 4.7, df = 10, p-value = 0.9
```

[2] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: Mathematics of Control, Signals and Systems 2 (1989), pp. 303-314

```
##
##
##  Box-Ljung test
##
## data:  r^2
## X-squared = 21, df = 10, p-value = 0.02
```

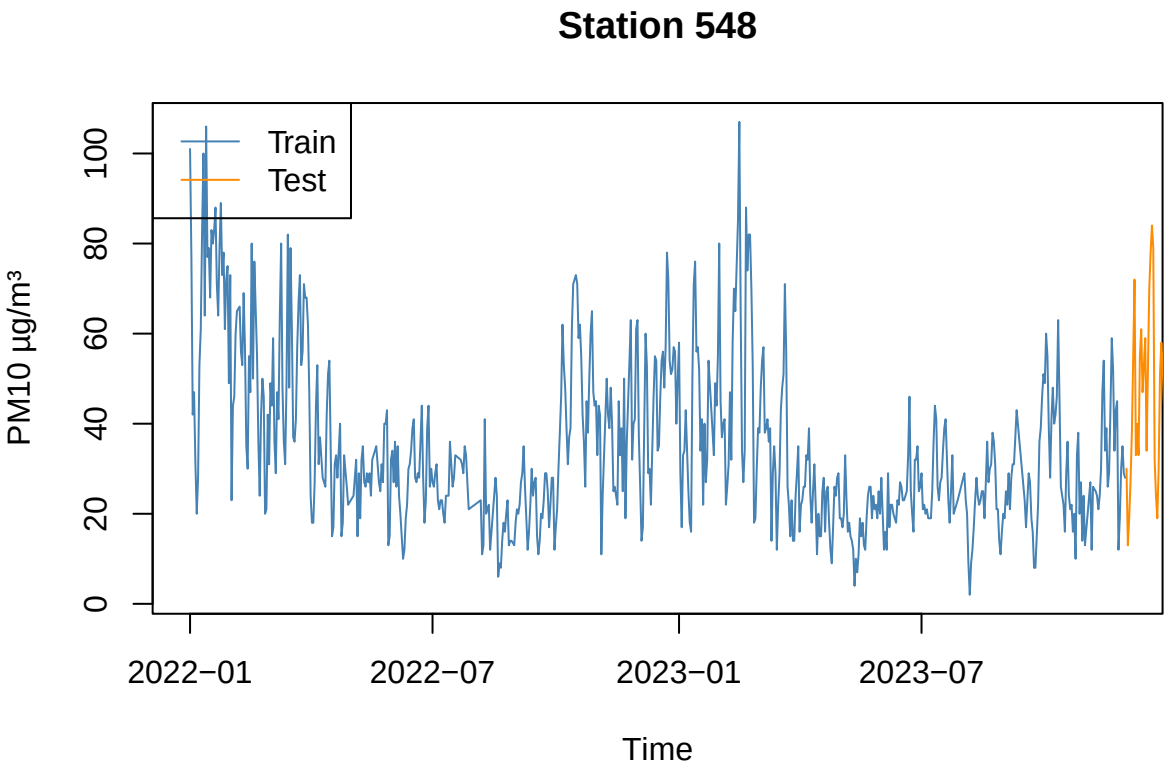


The ACF plots of the residuals suggest that they align with the white noise assumption. Additionally, the p-values from the Ljung-Box test indicate no significant evidence to reject the null hypothesis of white noise. However, the ACF of the residuals squared and cross-correlation function plots reveal potential remaining information, as some lags appear correlated. This suggests that a more advanced neural network model might be needed. The same procedure applied to other stations yields similar results.

Forecasting

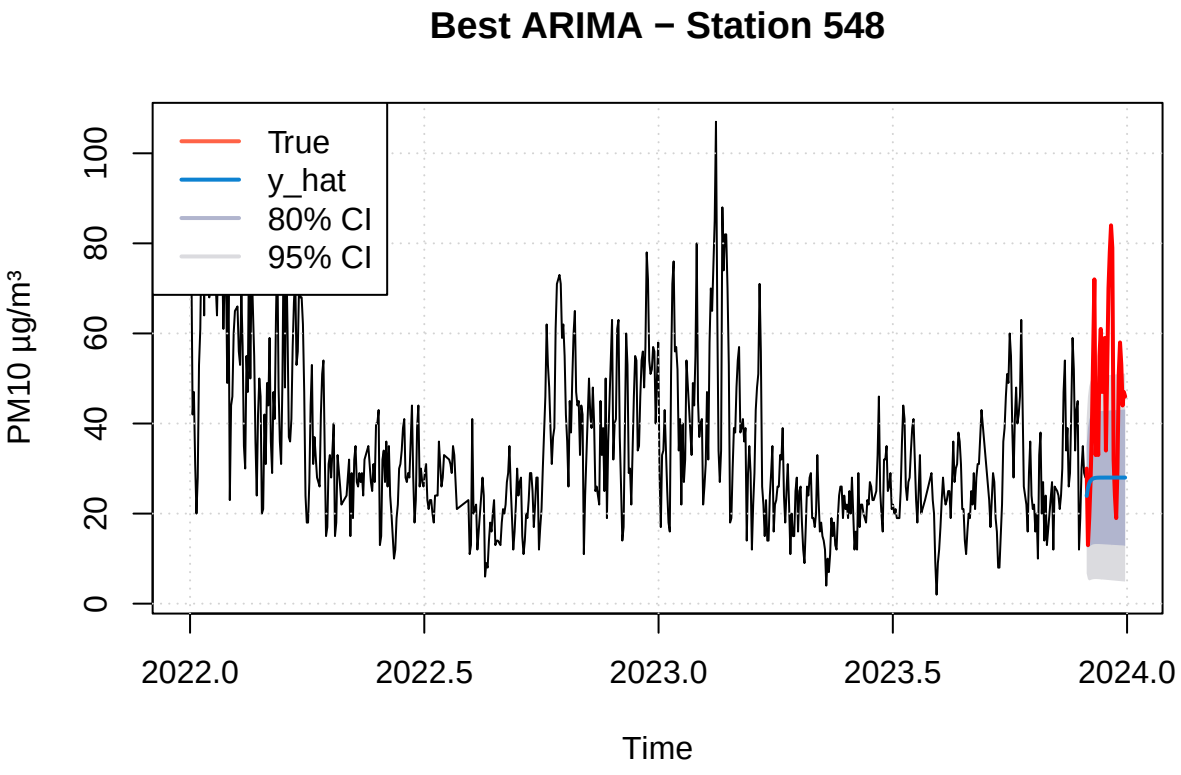
This section addresses the forecasting of PM10 levels using various models. An illustrative example of the train-test split for the Milan station data is presented in the plot below.

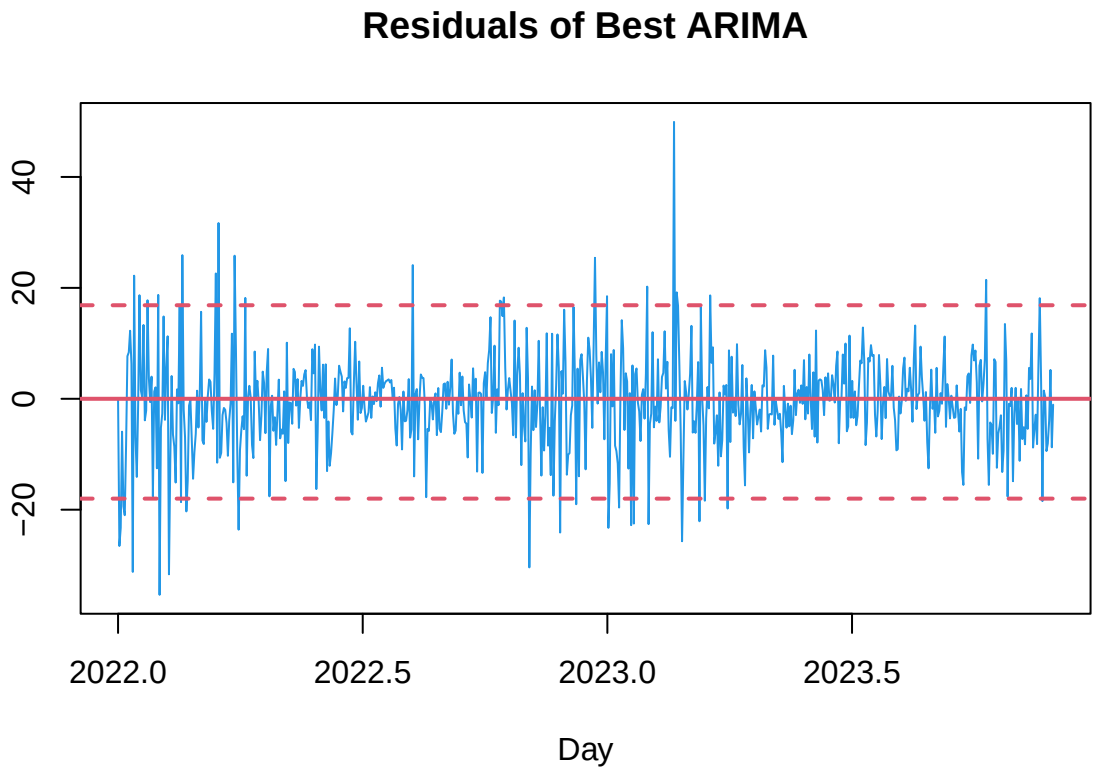
For convenience, the forecasting analysis is demonstrated using data from Station 548 in Milan. However, the same methodology can be extended to other stations.

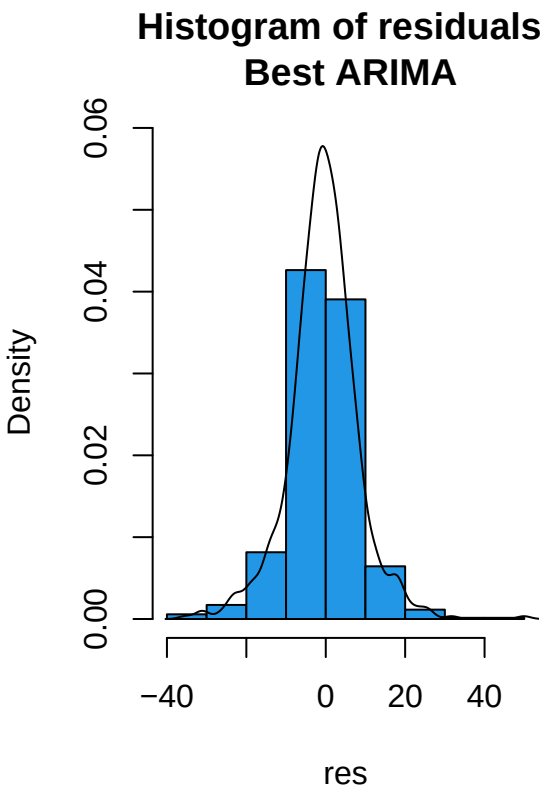


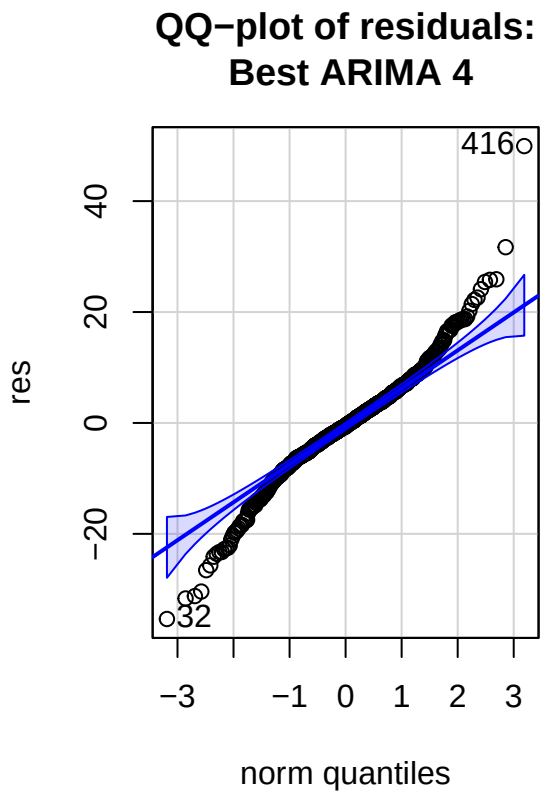
In the dynamic regression approach, the mean of the NO2 predictor observations in the training set is utilized for forecasting new values of PM10.

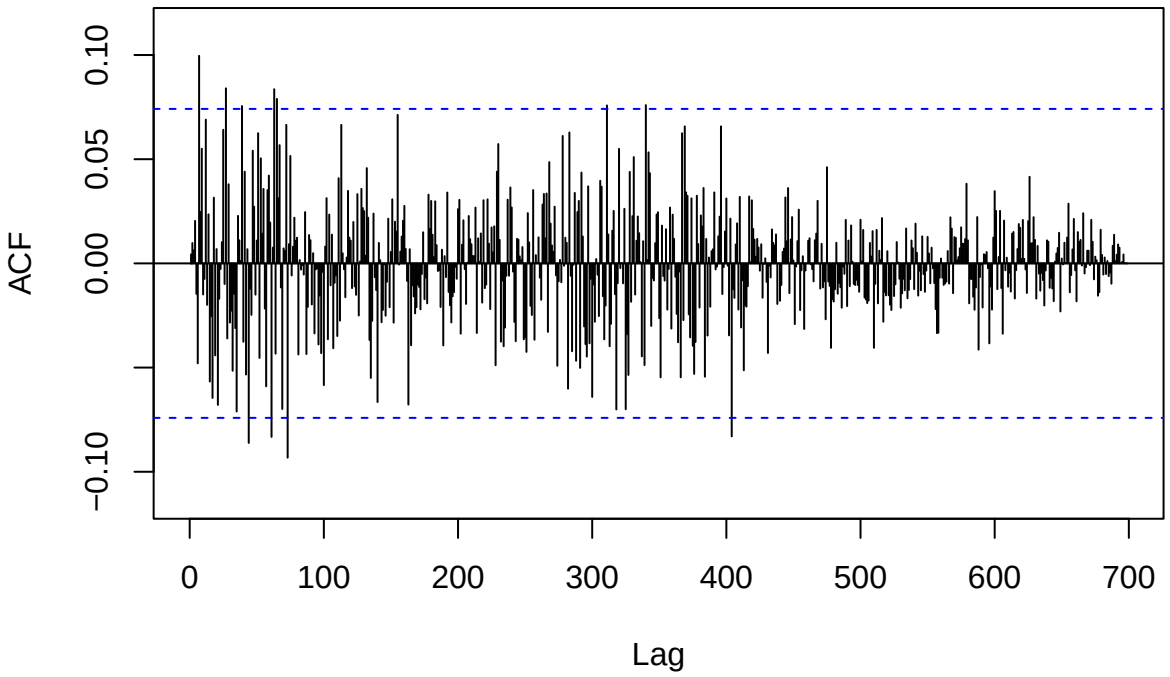
Prediction performance




Residuals of Best ARIMA
40
20
0
−20
2022.0
2022.5
2023.0
2023.5
Day

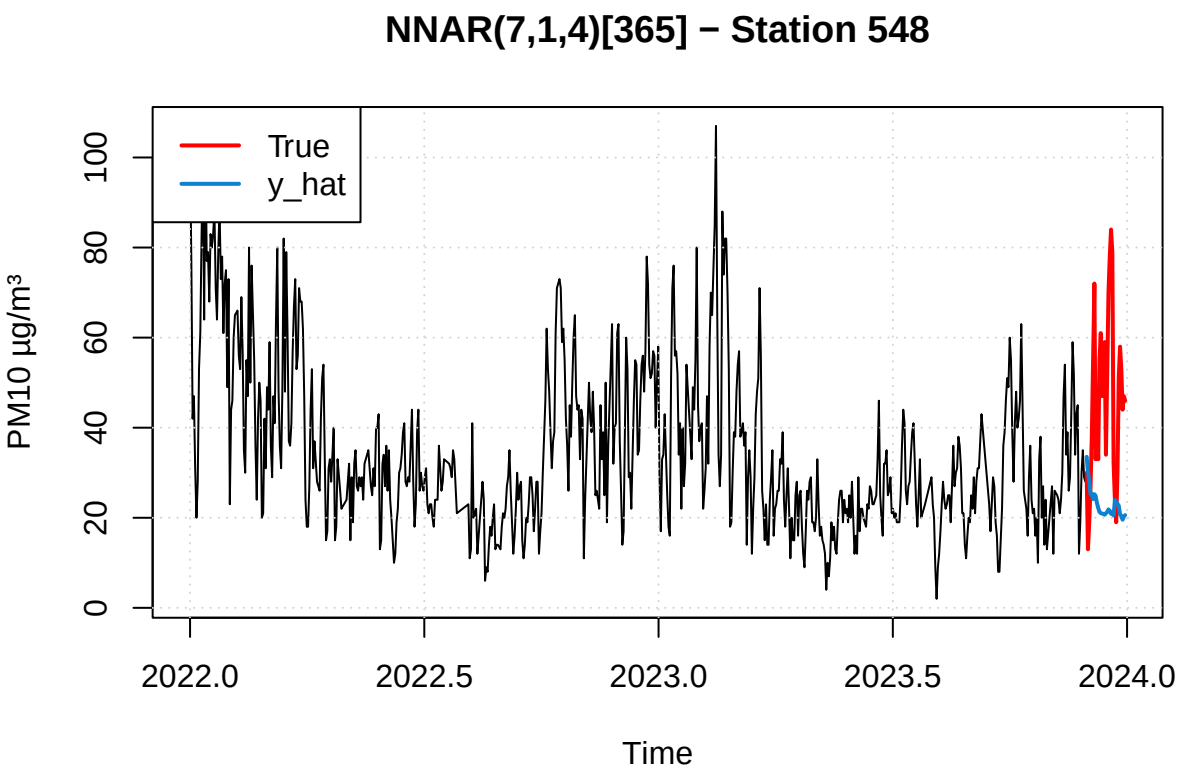

Histogram of residuals:
Best ARIMA
Density
0.06
0.04
0.02
0.00
−40
0
20
40
res

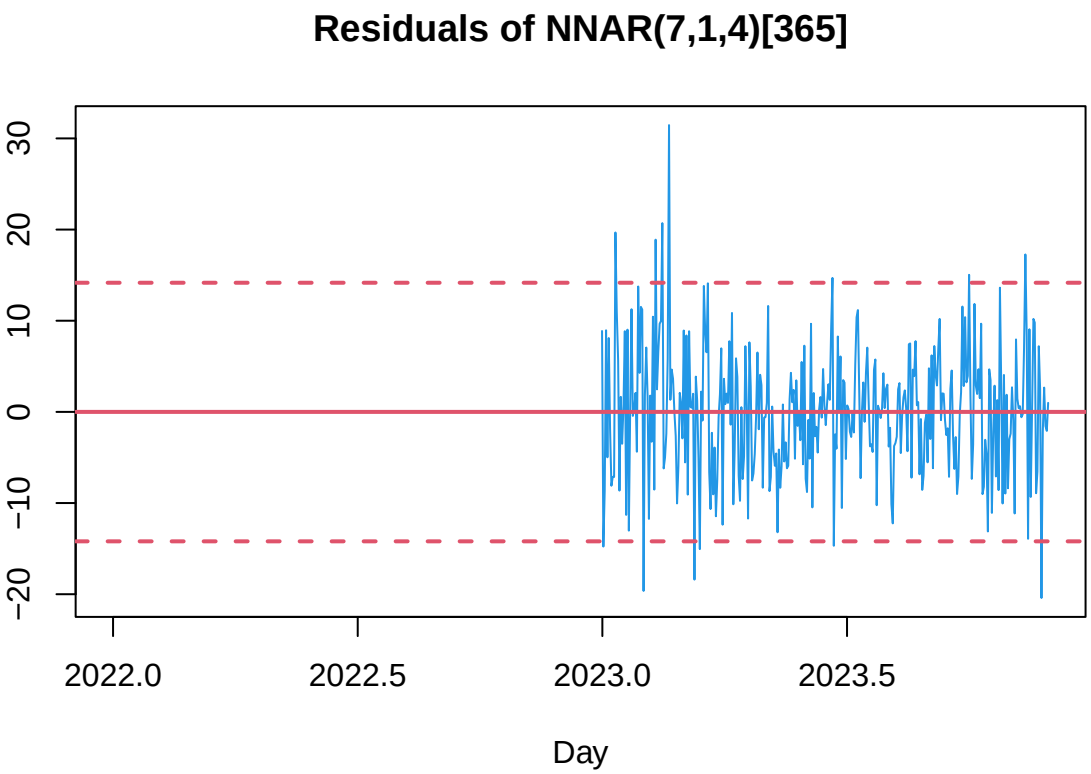

QQ−plot of residuals:
Best ARIMA 4
res
40
20
0
−20
416
32
−3
−1
0
1
2
3
norm quantiles

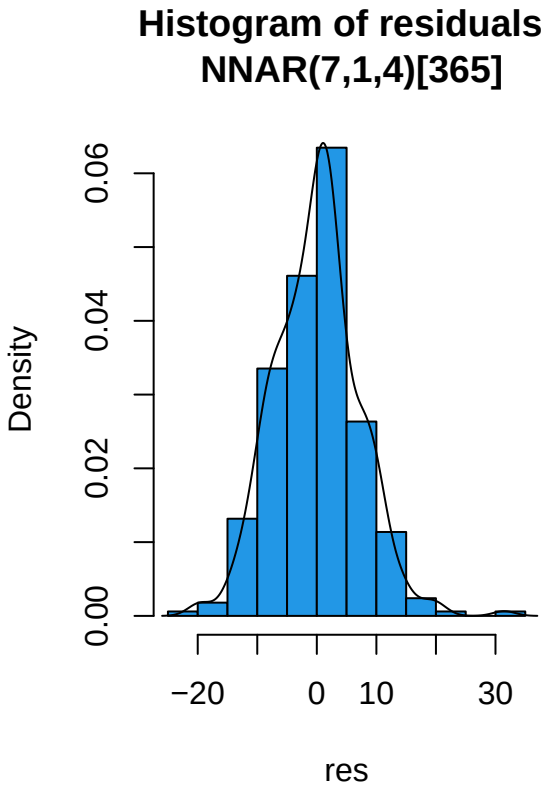


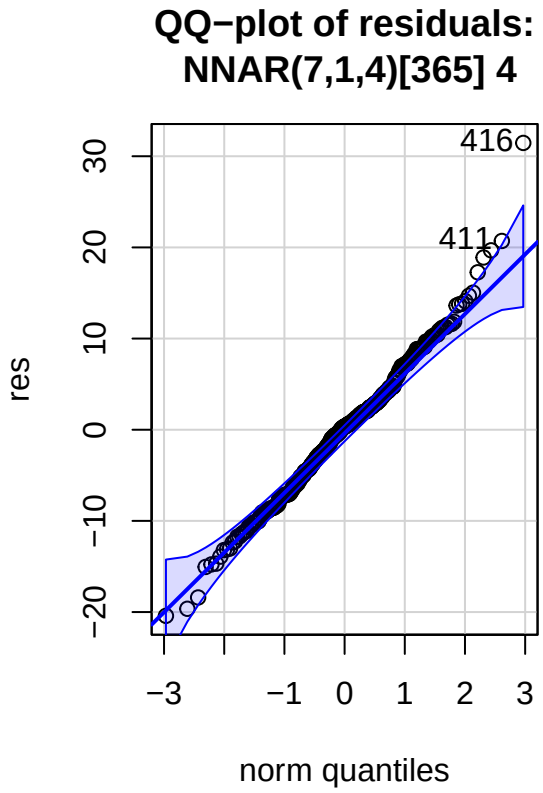

```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 154, df = 140, p-value = 0.2
##
## Model df: 0.   Total lags used: 140
```

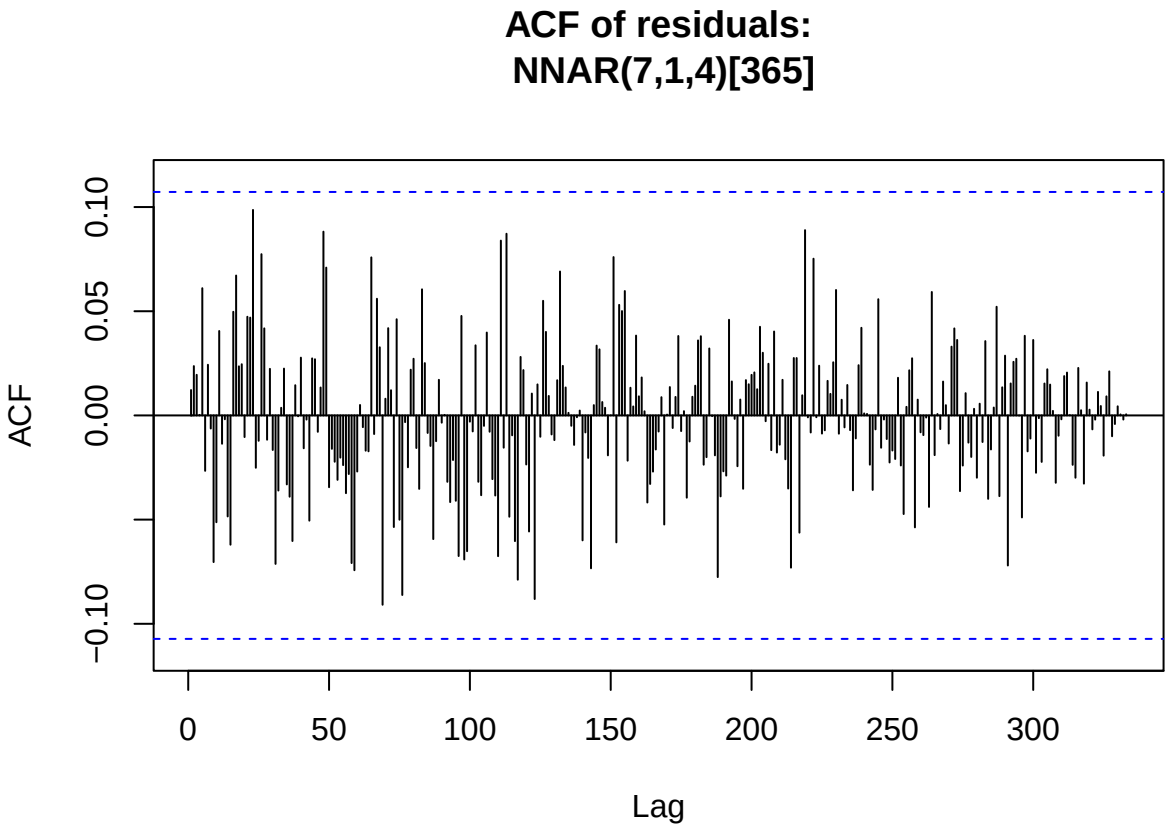
Attempting to forecast future values over a 31-day horizon using the best ARIMA model selected based on the AICc score and the dynamic regression model reveals suboptimal results for longer-term predictions. Although the residuals appear uncorrelated, as indicated by the ACF plots and the Ljung-Box test, which supports the hypothesis of white noise residuals, some issues persist. While the residuals have a mean close to zero, their variability is not constant, and the Q-Q plot shows unsatisfactory behavior in the tails.


NNAR(7,1,4)[365] − Station 548
True
y_hat
PM10 µg/m³
0
20
40
60
80
100
2022.0
2022.5
2023.0
2023.5
2024.0
Time


Residuals of NNAR(7,1,4)[365]
−20
−10
0
10
20
30
2022.0
2022.5
2023.0
2023.5
Day







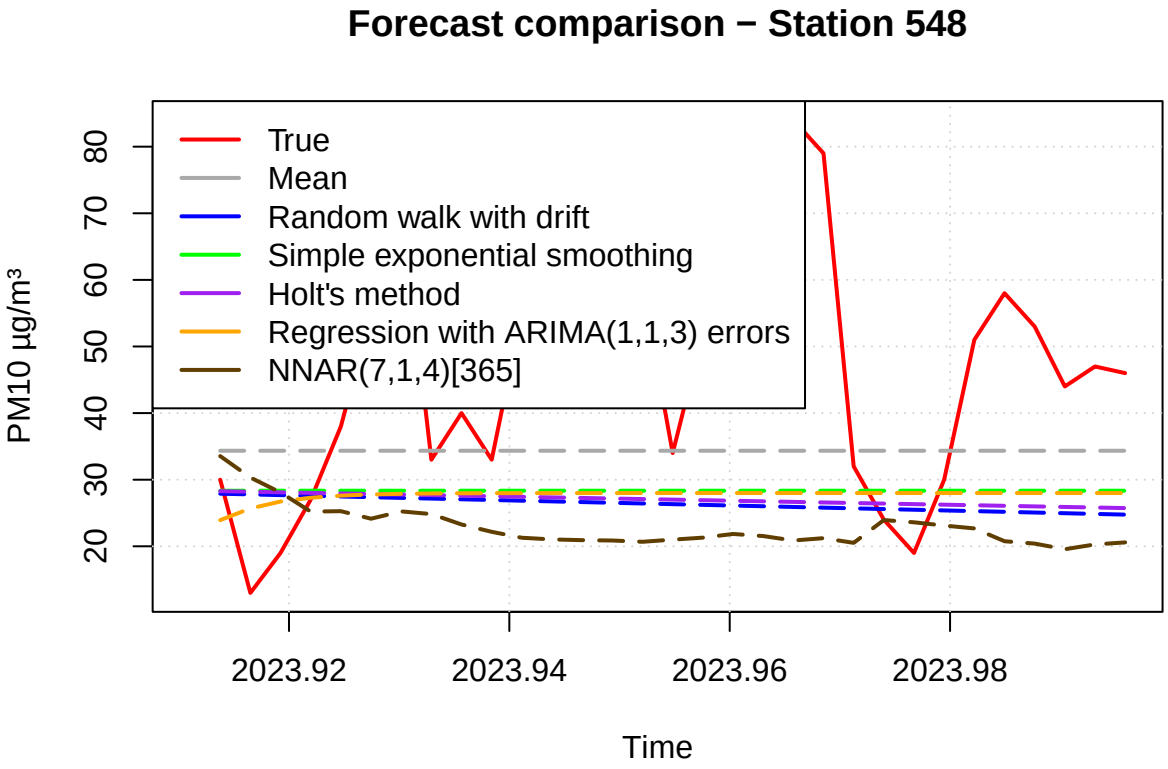

```
##
##  Ljung-Box test
##
## data:  Residuals
## Q* = 104, df = 140, p-value = 1
```

```
##
## Model df: 0.   Total lags used: 140
```

The forecasts generated by the neural network autoregressive model exhibit a distinctly different pattern compared to those produced by the previous model. The residual plot indicates that many values are NA. Despite this, the ACF is highly satisfactory, and the residuals appear to meet the assumptions of normality and a zero mean.

Forecasts comparison

A deeper evaluation in terms of forecasting ability is performed, assessing the two implemented methods with other simpler techniques.



For clearer visualization, the plot below shows only the test set, covering a 31-day forecast horizon. Initially, the drift method, simple exponential smoothing, and Holt's method perform similarly. However, over the long term, their predictions diverge.

The dynamic regression model with the NO_2 predictor initially underestimates values but then predicts accurately before eventually aligning more closely with the simpler methods. In contrast, the neural network-based model exhibits an oscillating behavior.

Tables with various metrics are provided to evaluate the models. Specifically, the models are tested across forecast horizons of 1, 3, 7, 14, and 31 days to assess their performance depending on the desired prediction length. As seen also above, the uncertainty in predictions increases with the length of the forecast horizon. Notably, for models based on daily data, very short-term predictions are particularly crucial.

Table 11: 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

Table 12: 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

Table 13: 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

Table 14: 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

Table 15: 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

For a 1-day ahead forecast, the simplest methods yield the lowest errors, with simple exponential smoothing achieving the best performance. The MASE score confirms that this method outperforms the naive approach. When the forecast window extends to 3 days, the dynamic regression model with ARIMA errors provides the best results, although simpler methods still perform reasonably well. As the forecast horizon increases beyond 7-14 days, prediction accuracy generally declines due to growing uncertainty and the challenges of capturing long-term trends. In these longer-term scenarios, the average method tends to perform better because it is less sensitive to short-term fluctuations.

Conclusions

Summarizing, this analysis highlights various aspects of the data related to air pollution in Lombardia, Italy, even if meteorological knowledge is not considered. Through exploratory data analysis on the monthly data, it was observed that rural areas may be affected by higher PM10 values more than other zones in some periods, and in general, Particulate Matter presents a strong seasonality. The global trend of this pollutant on the analyzed stations seems decreasing and during the emergency period in Italy due to COVID-19 PM10 values have not reduced much. Within the selected date range also the mean value of the pollutant among the zones didn't exceed the annual threshold so in general situation doesn't appear to be alarming.

In terms of forecasting daily data for station 548 in Milano v.Senato, the implemented models seem to not bring much improvement over the simplest methods, even if the white noise assumption for the residuals is satisfied. In particular the specified neural network autoregressive model seems to be not appropriate, considering the scores obtained on the test set. Using external predictors instead appears to be beneficial.

In conclusion, handling daily data likely requires more advanced models to better capture the underlying patterns. In addition, to effectively account for the seasonality which was not fully addressed in this study due to its complexity, incorporating Fourier terms could be a promising approach for improving the accuracy of forecasts[3].

[3] https://robjhyndman.com/hyndsight/longseasonality/