TIME SERIES FORECASTING

ILSIN SU SAHIN

Introduction

Air pollution poses a significant threat to public health and the environment, prompting governments worldwide to focus on monitoring and mitigating its anthropogenic sources. In this report, the concentration of PM10, a well-known airborne pollutant, in an Italian region is projected. PM10 is connected to mechanical operations, atmospheric chemical reactions, and combustion processes. Accurate forecasting of PM10 concentrations is crucial for effective pollution control strategies.

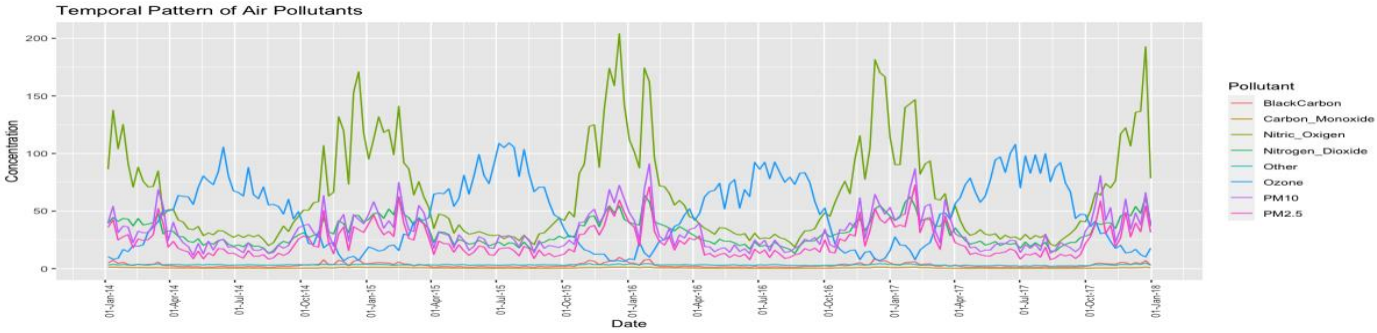
209 weekly observations from 2014 to 2017 were included in the dataset for this analysis, which was obtained from the Italian Environmental Agency ARPA. The analysis begins with exploring the dataset and examining the temporal patterns of PM10 concentration.

An ARIMA model will be created to predict PM10 concentrations. The temporal patterns and autocorrelation in time series data are captured by this model. Recent observations will be preserved for evaluation. Evaluation metrics like AIC will be used to choose the "best-fitting" ARIMA model. Residual analysis, autocorrelation diagnostics, and statistical tests will be used to evaluate the goodness-of-fit.

Alternative models, such as exponential smoothing, will be taken into consideration in addition to the ARIMA model. Based on the accuracy of their forecasts and the dependability of their confidence intervals, the models will be contrasted. The chosen model will be used to project PM10 concentrations over the ensuing 36 weeks and 120 weeks with the predictions being contrasted with provided real-world data. RMSE, MAE and MAPE will be used to assess forecast accuracy, and coverage probability and confidence interval width will be used to assess reliability.

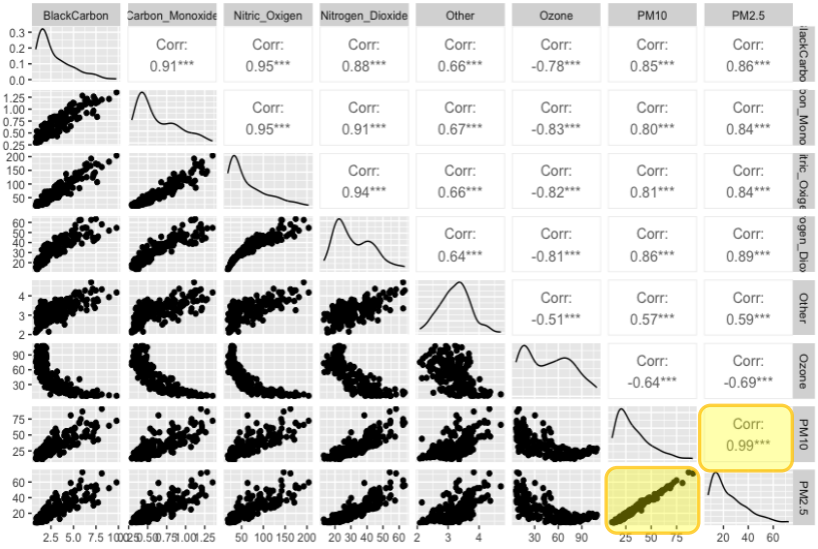
Exploratory Data Analysis

We first take a view of the case volumes on all air pollutants in Italy. Apart from Ozone, all pollutants exhibit peaks during the same time of the year. Among them, Nitric Oxygen demonstrates the highest concentration levels, while Black Carbon, Carbon Monoxide, and other pollutants exhibit comparatively lower levels.

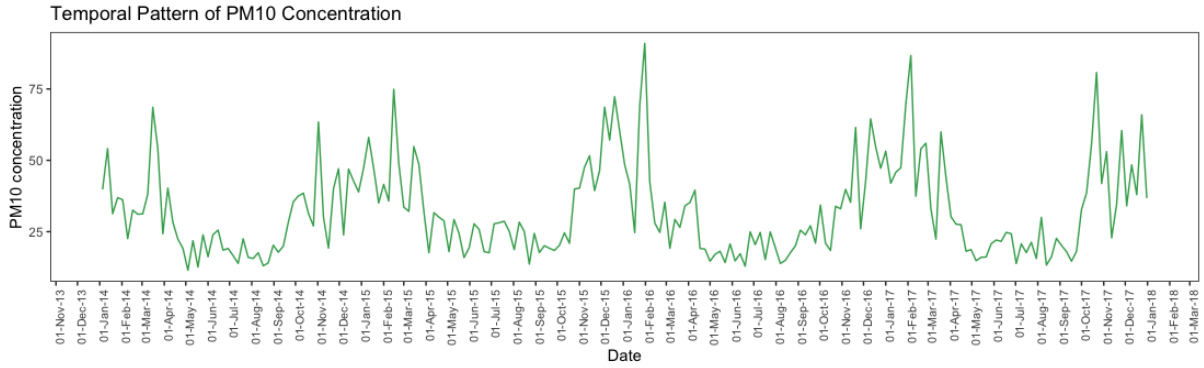


Similar patterns and volumes are displayed by PM10, Nitrogen Dioxide, and PM2.5, pointing to a possible relationship between these pollutants. These results offer insightful information about the temporal trends and relative sizes of air pollutant concentrations in Italy.

Based on the pairs plot analysis, it is evident that air pollutants exhibit strong correlations with each other. Specifically, PM10 demonstrates a robust linear association with Black Carbon, Carbon Monoxide, Nitric Origenian Nitrogen Dioxide, with correlation coefficients ranging from 0.80 to 0.86. Furthermore, PM2.5 exhibits and almost perfect positive linear association with a correlation coefficient of 0.99. In contrast, Ozone displays a moderate negative correlation of -0.67 with PM10. These findings suggest that there are significant relationships and dependencies between different air pollutants in the dataset.



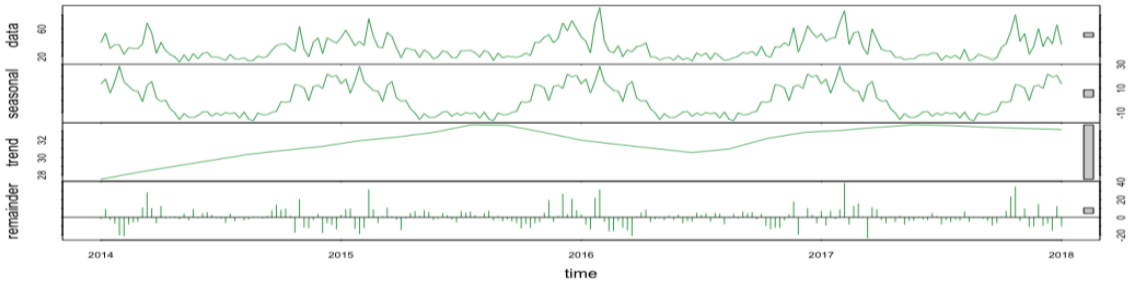
The PM10 concentration in Italy is the sole subject of this analysis. A visual examination of the plot reveals that the PM10 data clearly displays a seasonal pattern that occurs every year around February or March. Consistent peaks are visible on the plot during these times, pointing to a stationary and constant seasonal component. In addition, regardless of the trend, the size of the seasonal fluctuations stays fairly constant. This finding lends more support to the idea that the PM10 data have an additive structure.



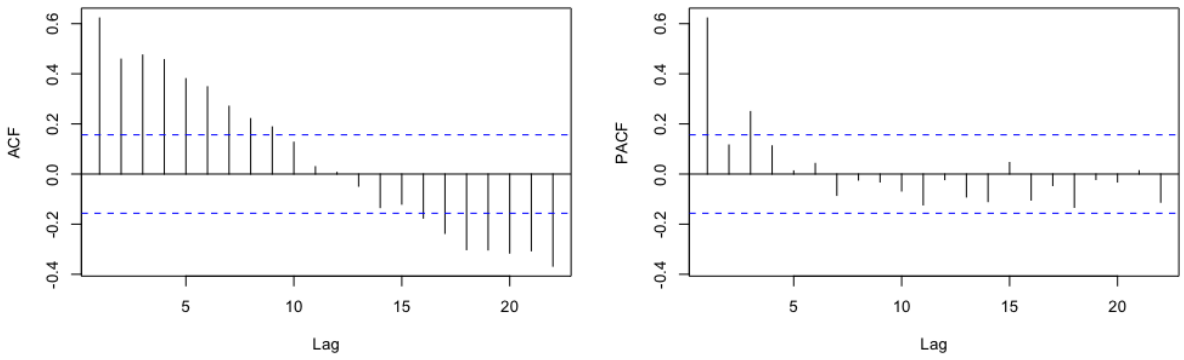
The identified seasonal pattern suggests that the seasonal component, regardless of trend level, has a stable impact on the PM10 data, with consistent peaks occurring around the same time each year. The case for an additive structure is strengthened by the consistency of the seasonal behaviour; since the seasonal fluctuations add to the overall variation in the data, Box-Cox transformation technique is not required.

Testing for Stationarity

We can see from the plots below that it does appear there is some trend and seasonality in the data.

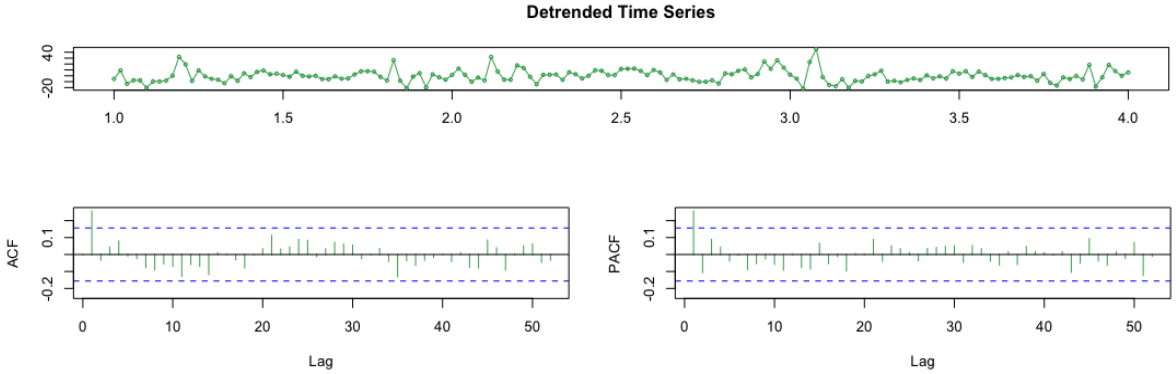


A time series is considered stationary if the mean and variance are constant and if there is no seasonality. We can confirm that there is a trend. With regards to seasonality, we can see a repeating cycle of higher number of concentrations of PM10 in certain moths and lower in others. *Additionally, the autocorrelation plot (ACF) and the partial autocorrelation plot (PACF) demonstrates that the data is non-stationary. In the ACF lines are above the blue lines. This line indicates the 95% confidence level that the data is auto-correlated and do not need adjustments. From the PACF we can see that there are auto-correlated lags present.*



In order to make our data stationary before considering applying ARIMA models to it, we will need to try to eliminate trends and seasonality from it. We will use a linear trend with a harmonic component to remove the trend and seasonality from the data and then use the Augmented Dickey-Fuller (ADF) test to determine whether the detrended data is stationary.

The application of a linear trend with a harmonic component to the time series data results in a noticeable improvement in stationarity and stability.



The ADF test is a unit root test for stationarity, we want to determine the roots as their presence can cause unpredictable results for our time series analysis and make our predictions potentially not dependable. The null and alternative hypothesis of the ADF test are as follows:

$$\text{Null hypothesis}(H_0) = \text{There is a unit root (data is not stationary)}$$

$$\text{Alternative Hypothesis } (H_1) = \text{Time series is stationary}$$

	Dicky-Fuller Test Stat	P-Value
Original Data	-2.9483	0.1785
Detrended Data	-4.8235	0.01

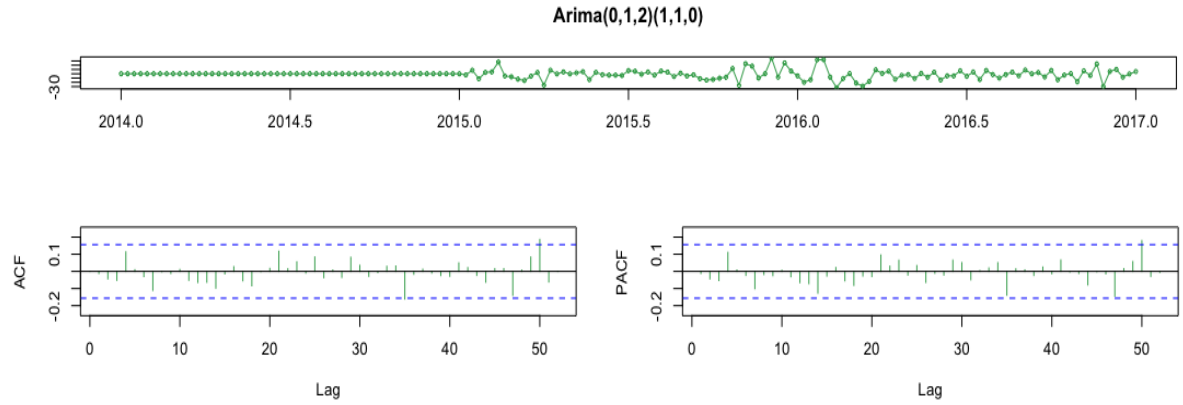
We can confirm that the data are not stationary by looking at the test's results, which show that the original data's p-value, 0.18, is greater than 0.05. As a result, we cannot reject the null hypothesis. The p-value, however, is less than 0.05 when we proceed to analyse the detrended data. Consequently, the null hypothesis can be rejected, and it is safe to assume that the detrended data is stationary. According to the Dicky-Fuller score, the more stable the data is, the higher the negative value must be. Therefore, it further confirms that the detrended is stable.

Modelling of time series of PM10

For model training and evaluation, we will divide the dataset into a training set and a testing set as follows: The first three years (2014-2017) of data will be used as the training set, and the last 52 weeks (1 year) will be used as the testing set. This partitioning results in a 75%-25% split of the data.

Auto-Arima

We start with the ARIMA auto.arima() function, which suggests that the best fitting model for the training data is ARIMA(0,1,2)(1,1,0)[52]. We can observe the residual plots for this process to see if it fists the model well.



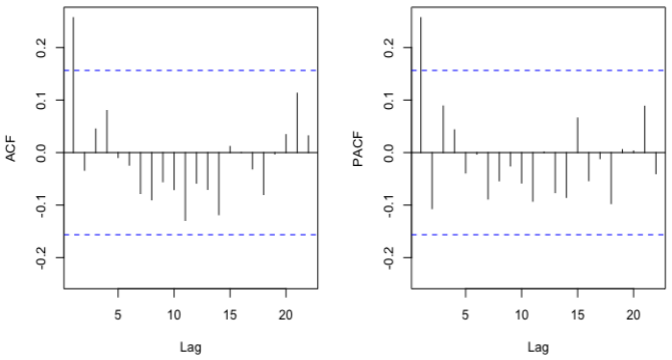
All of the autocorrelation values are within the 95% confidence intervals, according to the autocorrelation function (ACF) plot. This finding suggests that the data do not contain any statistically significant autocorrelation. Because there is little to no significant autocorrelation, it is likely that the time series' observations are independent and do not exhibit a consistent linear relationship with their prior values.

Arima

To model the time series data using an ARIMA approach, we begin by detrending the time series. In this case, a regression approach is chosen to capture both the seasonal and trend patterns, where these patterns are modelled using two cosine functions. The linear additive model can be formally represented as follows:

$$X_t = \beta_0 + \beta_1 t + \beta_2 \cos\left(\frac{2\pi t}{52}\right) + \beta_3 \cos\left(\frac{4\pi t}{52}\right) + e_t$$

Where e_t will be modelled by an Arima model. Based on the analysis of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the detrended data, it appears appropriate to consider either an ARIMA (0,0,1) or ARIMA (1,0,0) model. After thorough analysis and consideration, we have selected the ARIMA (0, 0, 1) model for time series forecasting. This choice is based on the recommendations of the auto.arima function, which evaluated detrended data and identified this model as the most suitable option. It achieved lower RMSE and AIC values compared to alternative models, ensuring accurate and robust forecasting for our data.



The coefficients of the model can be seen below:

Table 1: Linear Regression coefficients

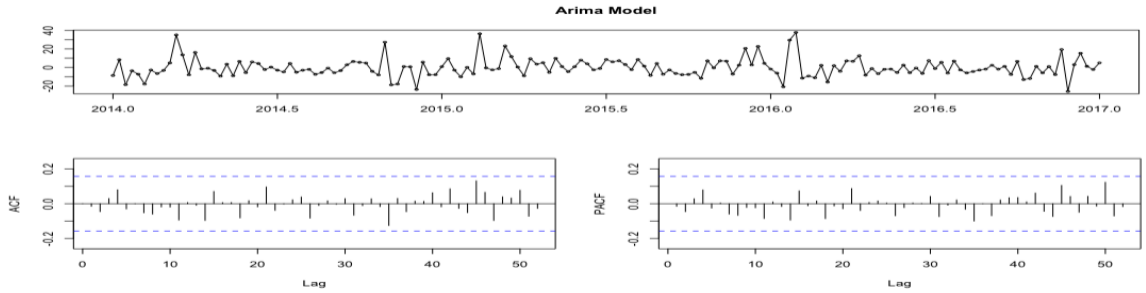
Variable	*Coef.*	*Lower CI (95%)*	*Upper CI (95%)*
(Intercept)	30.992	27.615	34.370
t	0.003	-0.034	0.040
Cos1	14.819	12.449	17.189
Cos2	3.063	0.693	5.434

Table 2: Arima coefficients

Variable	*Coef.*	*Lower CI (95%)*	*Upper CI (95%)*
(Intercept)	30.992	29.140	33.246
ma1	0.003	0.143	0.456
Cos1	14.819	11.910	17.685
Cos2	3.063	0.163	5.918

The two cosine functions, "Cos1" and "Cos2," are both statistically significant and are crucial in explaining the seasonal patterns in the pollutant concentration, according to both the linear regression and ARIMA models. The 't' coefficient's lack of significance indicates that there is no obvious linear trend in the pollutant concentration over time. The importance of the cosine functions in capturing seasonality and the lack of a significant trend in the pollutant concentration are supported by the coefficients' significance. These results must be taken into account when interpreting the models and applying them to forecasting and additional research.

After fitting the regression model with ARIMA, it is essential to examine the residuals of the model to assess whether short-term correlations have been effectively removed. The residual plots are displayed below:



It is clear from looking at the residual plots that there is no longer any apparent correlation in the data, and the residuals display characteristics resembling white noise. This finding suggests that the trend, seasonality, and short-term correlation from the initial time series data have been successfully eliminated by the model. We used the Box-Ljung test on the residuals, though, to formally ratify this conclusion.

A statistical test called the Box-Ljung test is used to determine whether autocorrelation is present in the residuals of a time series model. The test statistic in this situation assesses how much the residuals has autocorrelation.

Box-Ljung test	
X-squared	P-Value
30.742	0.9917

We can confidently claim that the residuals are uncorrelated based on the high p-value, indicating that the model has successfully captured the underlying patterns in the data. The result is residuals that resemble white noise after successfully removing the trend, seasonality, and short-term correlation.

It is crucial to evaluate our ARIMA model's performance against that of the suggested auto.arima model now that we have created and validated it using the two cosine functions as exogenous variables. Two important metrics—the Root Mean Squared Error (RMSE) and the Akaike Information Criterion (AIC)—will be used to evaluate the model comparison.

Model	**AIC**	**RMSE**
Auto-Arima	847.81	10.34044
Model 1	1182.35	10.1183

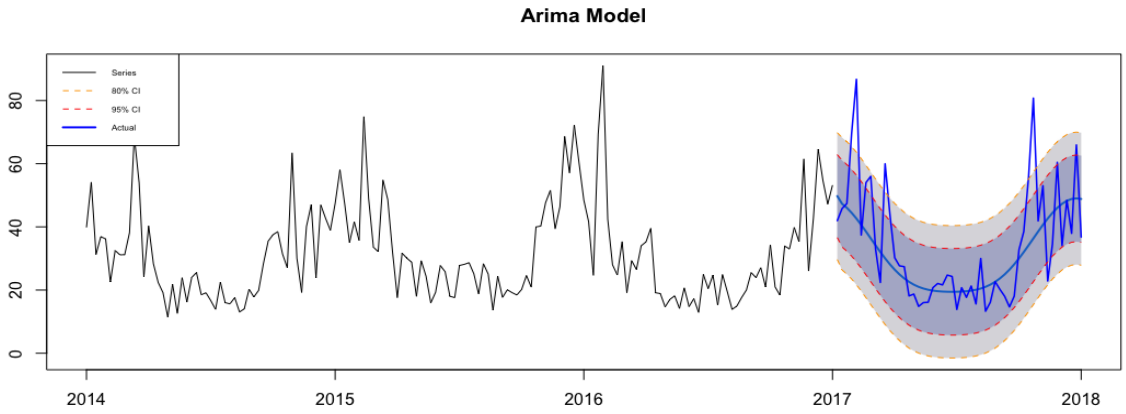
The auto-Arima model outperforms "Model 1" in terms of goodness-of-fit, with a significantly lower AIC, suggesting better overall model fit and complexity. However, our model shows slightly better predictive accuracy, possibly due to the inclusion of two cosine functions capturing additional seasonality and short-term dynamics.

Model Testing, Prediction & Comparison

We will now compare the actual values to the predicted values using the Root Mean Square Error (RMSE), to evaluate the performance of our chosen ARIMA model. To compare with our ARIMA model, we will also create a second model using exponential smoothing techniques. We will compare their respective RMSE results to decide which model performs better.

ARIMA Prediction

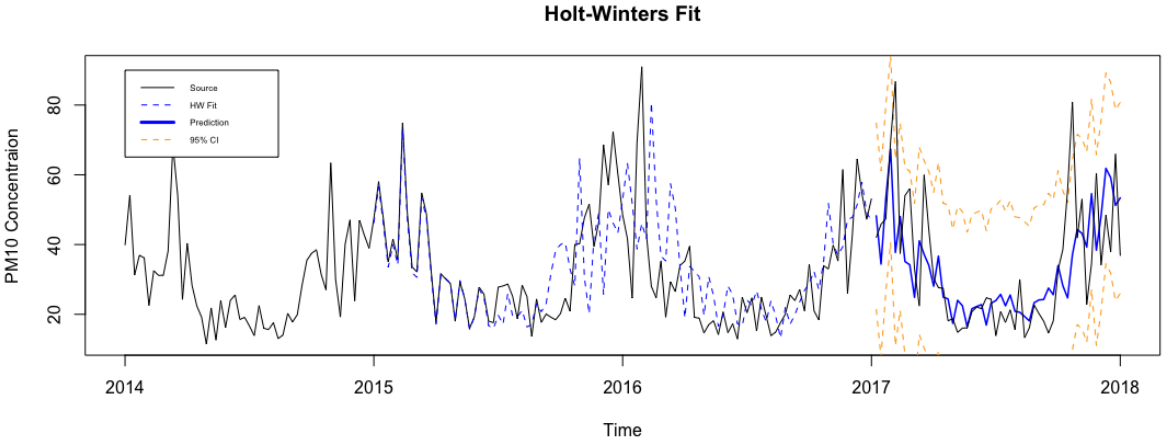
We will now incorporate the 52-week test set into our evaluation of the model's performance. The predicted PM10 concentration for the following 52 weeks is shown on the forecast plot along with the 80% and 95% confidence intervals. The blue line represents the model's predictions, and the black line represents the actual observations. We can evaluate the model's precision in predicting PM10 concentration for the test period by contrasting the actual values with the predictions and taking the confidence intervals into account.



The forecasted plot reveals that the model's coverage is not at 100%. It fails to capture two peaks in the data that fall outside the 95% confidence interval. However, for most of the data, the model provides accurate predictions within the 80% confidence interval. While the model performs well for most observations, it may need improvement to handle extreme variations beyond the 95% confidence interval. Further analysis and potential model refinement can address these limitations and enhance forecasting accuracy.

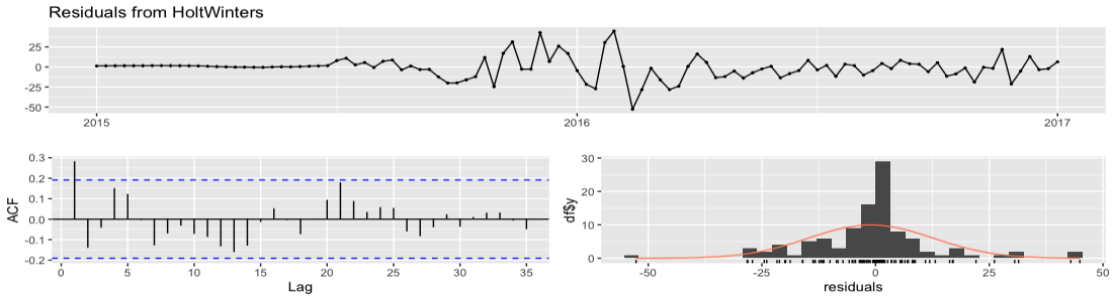
Exponential Smoothing Model

To compare the performance of our ARIMA model, we will now fit an exponential smoothing model to the data. As the data presents non-stationarity, as previously discussed, we will fit a Holt-Winters exponential smoothing model.



The forecast plot makes it clear that, when compared to our ARIMA model, the Holt-Winters model performs better at predicting the data peaks. It is crucial to remember that not all forecasts fall within the 95% confidence interval. However, the Holt-Winters model performs better in identifying the peak patterns in the data. However, just like the ARIMA model, it might have trouble handling extreme variations outside of the 95% confidence interval.

Upon examining the residuals plot, we observe that the ACF plot displays significant correlation at lag 1, indicating that the model has not entirely removed the autocorrelation from the data.



It is clear from the residuals plot that the autocorrelation in the data has not entirely been eliminated by the model because the ACF plot shows a significant correlation at lag 1. The ACF plot's presence of a significant correlation at lag 1 suggests that the residuals may still contain some short-term dependencies or patterns. These correlations suggest that some underlying temporal dynamics of the time series data have not been fully accounted for by the model, resulting in residual patterns that are not entirely uncorrelated.

Model	RMSE	MAE	MAPE
ARIMA Model	13.40686	9.255690	26.54516
Exponential Smoothing	14.59868	10.564972	30.80392

We can compare the effectiveness of the ARIMA model and the Exponential Smoothing model for forecasting PM10 concentration based on the accuracy statistics that have been provided. In comparison to the Exponential Smoothing model, the ARIMA model exhibits higher predictive accuracy across all three metrics (RMSE, MAE, and MAPE). It achieves a lower RMSE of 13.41, demonstrating that its forecasts are more accurate than the actual values. Additionally, the ARIMA model displays a lower MAE of 9.26, which suggests improved accuracy in predicting the precise magnitude of the forecast errors. In addition, the ARIMA model achieves a more precise percentage accuracy in predicting the forecast errors relative to the actual values with a lower MAPE of 26.55. Based on these results, the ARIMA model outperforms the Exponential Smoothing model in forecasting the PM10 concentration. Its ability to provide more accurate and reliable forecasts makes it the preferred choice for this specific forecasting task to move forward as our final model.

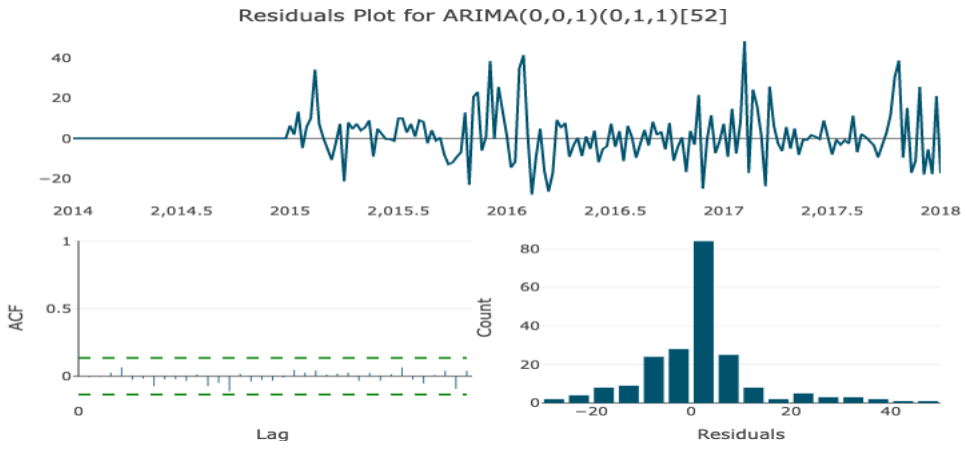
Out-of-Sample Forecasting and Performance Evaluation

Out-of-Sample Forecasting: 36 weeks ahead

We will conduct an out-of-sample forecast in this section for the pollutant concentration for the 36 weeks prior to the last date available (December 31, 2017). Additionally, an auto-Arima model will be developed for comparison. We will compare the predicted values with the actual data after obtaining the forecasted values. By comparing the two forecasts' accuracy using RMSE, MAE, and MAPE, we can determine whether the selected model is still appropriate for this 36-week forecasting horizon.

Auto – Arima Model

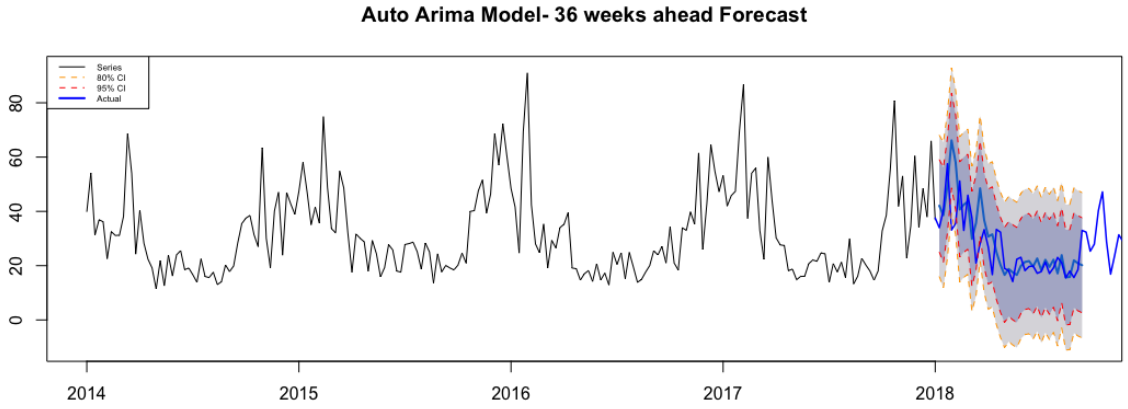
The automatic model selection process, specifically the Auto-ARIMA model, suggests that the best fitting model for the entire dataset is ARIMA(0,0,1)(0,1,1)[52]. To validate the goodness-of-fit of this model, we examine the residual plots, which offer insights into the model's performance.



We can see from the residual plots of the first auto-generated model that it adequately fits the data. The model has successfully eliminated the temporal dependencies from the original data, leaving only white noise in the residuals, which are devoid of any discernible correlation. The residual distribution also seems to be close to normal, supporting the model's suitability for identifying the underlying trends in the data. The model coefficients are shown as follows:

Variable	*Coef.*	*Lower CI (95%)*	*Upper CI (95%)*
ma1	0.237	0.830	0.390
Sma1	-0.577	-0.879	-0.275

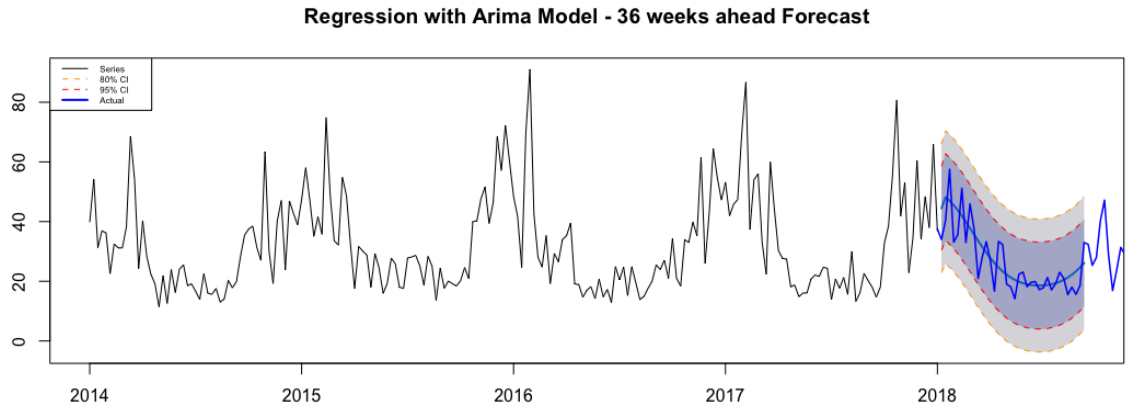
The confidence intervals for the coefficients provide a range within which the true values of the parameters are likely to fall with a certain level of confidence. The intervals suggest that the estimates are statistically significant, not containing zero and provide additional insights into the precision of the model parameters.



As all actual observations fall within the corresponding confidence intervals, the 36-week forecast plot demonstrates that the Auto-ARIMA model exhibits a coverage probability of 100% for the forecasted values. The plot clearly shows that the actual values closely match the 80% confidence interval, demonstrating a fitting within this interval that is reasonably accurate. The model is confirmed to have the capacity to accurately capture the expected range of outcomes for hospital admissions in the upcoming fortnight as the actual values comfortably fall within the wider 95% confidence interval.

Regression with Arima Model

We will now make a 36-week forecast using our final model that was chosen. The final model, which incorporates the ARIMA model and the regression model with two cosine functions, successfully captures the patterns and trends in the data for the 36-week forecast. The forecasted values fall comfortably within the corresponding confidence intervals, indicating that the model has a high coverage probability. It is noteworthy that the actual observations closely match the 80% confidence interval, demonstrating a good fit to this range.



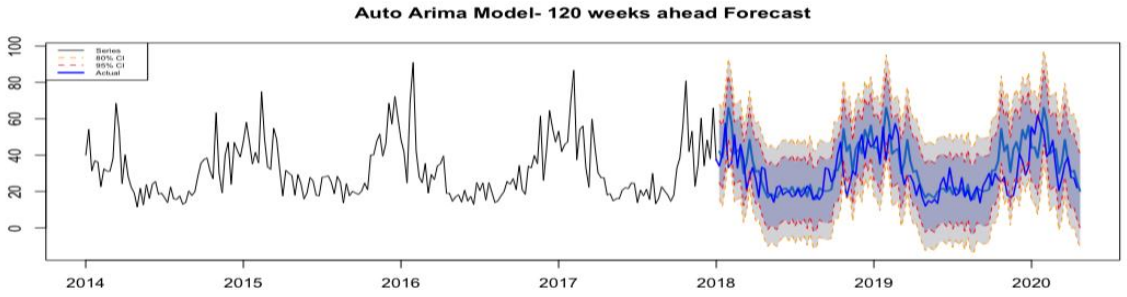
We will now compare the accuracy statistics of the two models. The statistics are displayed as below:

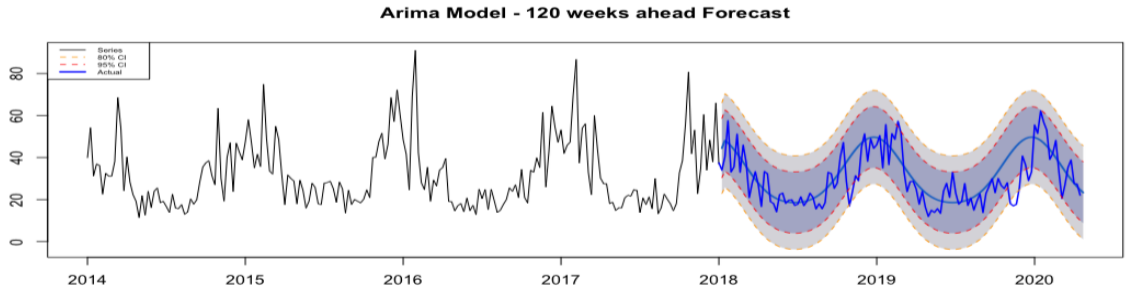
Model	RMSE	MAE	MAPE
Auto-Arima	9.814445	6.854907	25.57864
Final Model	6.318434	5.17602	20.15632

The Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) values for the Final Model are significantly lower than those for the other two metrics, demonstrating the Final Model's superior accuracy. The RMSE for the Final Model was 6.32, which was significantly less than the Auto-ARIMA's RMSE of 9.81. The MAE for the Final Model is 5.18, which is superior to the Auto-ARIMA's MAE of 6.854907 in both instances. Additionally, the Final Model's MAPE is 20.16, which indicates more precise predictions when compared to the Auto-ARIMA's MAPE of 25.58.
These results demonstrate that the Final Model provides more precise and reliable forecasts for the pollutant concentration. The incorporation of the two cosine functions in the regression component has proven to be advantageous, capturing seasonality and enhancing the model's predictive capabilities. Additionally, the ARIMA component effectively models temporal dependencies, further enhancing the forecast accuracy. Based on the accuracy statistics, the Final Model emerges as the preferred choice for forecasting pollutant concentrations over the Auto-ARIMA model.

Out-of-Sample Forecasting: 120 weeks ahead

In this section, we will conduct another out-of-sample forecast for the pollutant concentration, this time for the 120 weeks ahead of the last available observation. Once again, we will use the chosen best model and compare it with an auto-Arima mode. We will assess if the chosen model is still suitable for this extended forecasting horizon.





The majority of the actual observations fell within the 80% confidence intervals, which showed that both models did a fairly good job of capturing the underlying patterns in the pollutant concentration data when viewed as forecast plots. Most of the data points in the Final Model were captured within the 80% confidence interval, with only a few points outside.

Considering the accuracy statistics shown below, the Final Model consistently outperformed the Auto-ARIMA model in terms of RMSE, MAE, and MAPE for this extended forecasting horizon. The Final Model's lower RMSE and MAE indicate better predictive accuracy and reduced forecast errors.

Model	RMSE	MAE	MAPE
Auto-Arima	11.37556	8.477413	30.91901
Final Model	9.153928	7.049821	27.44973

In conclusion, the Final Model continues to be a good option for forecasting pollutant concentration even with the expanded forecasting horizon. Its accuracy is better compared to the Auto-ARIMA model, and it can capture the majority of data points within the 80% confidence interval. However, it is essential to consider that unforeseen events might affect air quality, potentially causing deviations in the forecast.

Limitations to Approach

- **Seasonal Component:** The current approach relies on two cosine functions to capture the seasonal component. However, during the analysis, it was observed that the Fourier terms (sine and cosine functions) did not yield statistically significant coefficients for sine, indicating that this method might not be the most suitable for capturing the seasonal variations in the PM10 concentration data. *(See appendix table).*

- **Linear Relationship with Other Variables**: Black carbon, PM2.5, carbon monoxide, and nitrogen dioxide all exhibit a strong linear association with PM10, according to research. The forecasting precision of the model might be improved by including these correlated variables as extra exogenous variables in the regression component. The model will perform better overall and be able to capture underlying patterns if more trend and seasonality are removed from the data.

- **External Factors:** The current model primarily focuses on temporal patterns and autocorrelations in the data. However, air pollutant concentrations can be influenced by external factors such as weather conditions, geographical features, and industrial activities. Incorporating these external factors as additional predictors in the forecasting model could enhance its predictive capabilities and provide a more comprehensive understanding of pollutant concentration fluctuations.

- **Data Granularity**: Weekly observations make up the dataset that was used for the analysis. Data with a higher level of detail, like daily or hourly measurements, might reveal more information about short-term fluctuations. With the help of this data, the forecasting model could be improved to produce more precise predictions, particularly for capturing smaller-scale variations.

Conclusion

In conclusion, this report focused on forecasting PM10 concentrations in an Italian region using an ARIMA model with two cosine functions as exogenous variables. The chosen model demonstrated good predictive accuracy and outperformed alternative models, such as exponential smoothing. To improve the forecasting approach, exploring more flexible methods for seasonality modelling and considering additional correlated variables could enhance accuracy. Furthermore, integrating external factors like weather conditions and industrial activities as predictors might provide more comprehensive forecasts. Despite these considerations, the developed ARIMA model showed promising results both for 36 weeks ahead and 120 weeks ahead forecasting and can serve as a valuable tool for air pollution forecasting.

Appendix

Table 1: Fourier Regression with Arima coefficients

	Lower CI (95%)	*Upper CI (95%)*
ma1	0.125	0.445
(intercept)	29.178	33.203
S1-52	-0.375	5.335
C1-52	11.9628	17.624
S2-52	-2.637	3.050
C2-52	0.215	5.858