atmosphere

MDPI

Article

Time Series Forecasting of Air Quality: A Case Study of Sofia City

Evgeniy Marinov [1,*], Dessislava Petrova-Antonova [1,2,*] and Simeon Malinov [1]

[1] GATE Institute, Sofia University "St. Kliment Ohridski", 1504 Sofia, Bulgaria; simeon.malinov@gate-ai.eu
[2] Faculty of Mathematics and Informatics, Sofia University "St. Kliment Ohridski", 1164 Sofia, Bulgaria
* Correspondence: evgeniy.marinov@gate-ai.eu (E.M.); dessislava.petrova@gate-ai.eu (D.P.-A.)

Abstract: Air pollution has a significant impact on human health and the environment, causing cardiovascular disease, respiratory infections, lung cancer and other diseases. Understanding the behavior of air pollutants is essential for adequate decisions that can lead to a better quality of life for citizens. Air quality forecasting is a reliable method for taking preventive and regulatory actions. Time series analysis produces forecasting models, which study the characteristics of the data points over time to extrapolate them in the future. This study explores the trends of air pollution at five air quality stations in Sofia, Bulgaria. The data collected between 2015 and 2019 is analyzed applying time series forecasting. Since the time series analysis works on complete data, imputation techniques are used to deal with missing values of pollutants. The data is aggregated by granularity periods of 3 h, 6 h, 12 h, 24 h (1 day). The AutoRegressive Integrated Moving Average (ARIMA) method is employed to create statistical analysis models for the prediction of pollutants' levels at each air quality station and for each granularity, including carbon oxide (CO), nitrogen dioxide (NO_2), ozone (O_3) and fine particles (PM2.5). In addition, the method allows us to find out whether the pollutants' levels exceed the limits prescribed by the World Health Organization (WHO), as well as to investigate the correlation between levels of a given pollutant measured in different air quality stations.

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

Citation: Marinov, E.; Petrova-Antonova, D.; Malinov, S. Time Series Forecasting of Air Quality: A Case Study of Sofia City. *Atmosphere* **2022**, *13*, 788. https://doi.org/10.3390/atmos13050788

Academic Editors: Kostadin Ganev and Georgi Gadzhev

Received: 24 April 2022
Accepted: 10 May 2022
Published: 12 May 2022

1. Introduction

Ambient air pollution poses a significant threat to public health. The ever-growing population of cities, rapid urbanization and the increasing level of motorization contribute to the constantly growing gas emissions. The latest reports by the European Environmental Agency show that in 2019 more than 350,000 premature deaths were attributed to chronic exposure to fine particulate matter, nitrogen dioxide and ozone [1]. As Figure 1 shows, the current state of ambient air pollution in some eastern European cities is becoming increasingly problematic where the statistics show more than 200 deaths per 100,000 inhabitants placing cities such as Sofia and Plovdiv at the highest mortality rates related to poor air quality [2].

These highly worrying statistics call for immediate action both from a political and regulatory standpoint. However, the role of applied research solutions and comprehensive scientific analyses cannot be overlooked as they can empower decision makers and give the necessary tools to act quickly and effectively.

Both the European Union and the World Health Organization have developed air quality standards and goals. However, despite all global efforts, a survey showed that no country met the World Health Organization (WHO) air quality standards in 2021 [3]. In order to address this highly complex problem, the European Union has created a substantial system of programs, funds and strategic plans comprising specific measures and actions in order to help countries and cities meet those standards. The implementation of those measures and actions would inevitably lead to better air quality in the long run. However, the results of these actions can take a relatively long time to take effect which leaves a gap between actions and the significant real-world mitigation of air pollution. This creates

the need for the creation of tools to fill this gap and create a system for adaptation to the current levels of air pollution. An important piece of this system would be an advanced air quality forecasting tool.

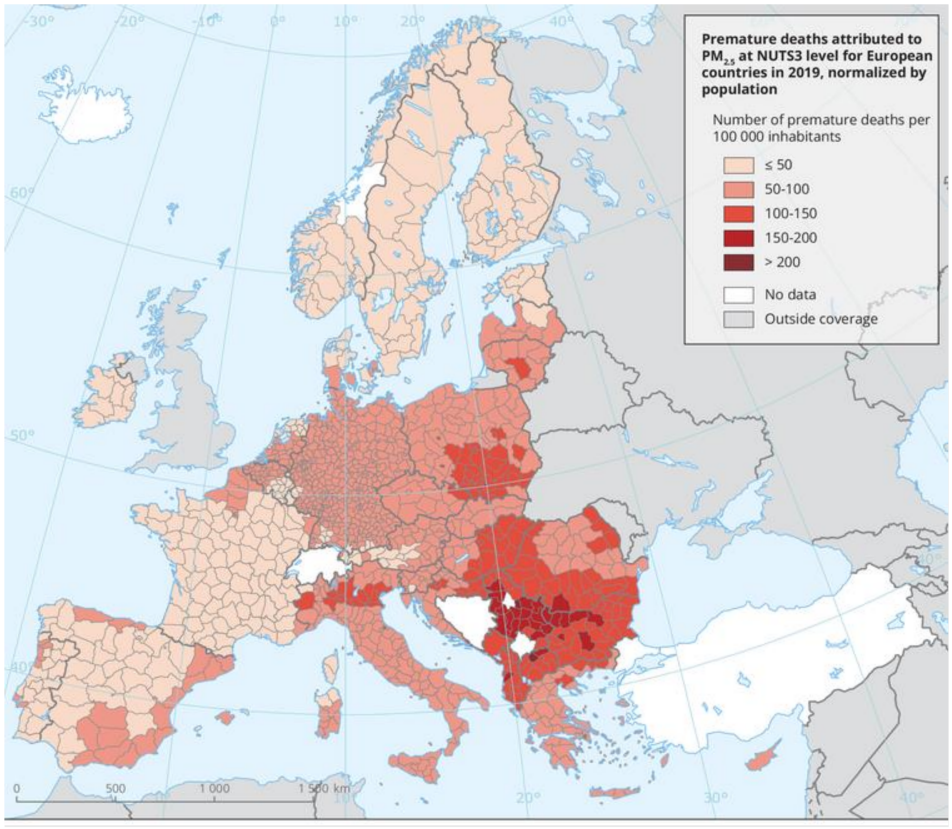


Figure 1. Premature deaths attributed to PM 2.5 for European countries in 2019, normalized by population (source: European Environmental Agency).

To keep its inhabitants safe from harmful air, many cities have adopted air quality forecasting programs to estimate and predict the concentrations of a variety of air pollutants. Those forecasting systems give decision-makers and citizens the opportunity to take preventive measures in order to protect the most vulnerable and sensitive groups and immediately and adaptively implement measures such as temporarily stopping primary emission sources to reduce air pollution, stimulating the use of public transportation, enforcing "No-burn days", and so on.

Accurate air quality forecasting can result in great social and economic benefits by providing the capabilities for advanced planning for citizens, families, companies and governing entities. Various analytical models are proposed for air quality prediction, which can be categorized as traditional, based on statistical models and non-traditional using Artificial Intelligence (AI) approaches. Time-series methods and regression analysis are examples of traditional techniques, while k-nearest-neighbors and artificial neural networks (ANNs) are representatives of the AI approaches. ANNs have been used in several research works for air quality forecasting [4–8]. A hybrid wavelet model was developed to investigate hourly data of Taiyuan, China between 2016 and 2017 [9]. The results show that ANNs and support vector machines (SVMs) have better forecasting accuracy on the decomposed data compared to raw data. The higher accuracy of the models on the transformed data is proved also by Cheng et al. [10], who employed the AutoRegressive Integrated Moving Average (ARIMA), SVN and ANN methods. The requirement of less computational power, satisfactory prediction ability, relatively easy implementation and increased availability of air quality data make the regression analysis a preferable technique among the statistical

methods. The time series regression is a statistical method that predicts the future based on the past, known also as autoregressive dynamics. It evaluates historical data to establish a valid forecasting model. The AARIMA method has been extensively studied and implemented in recent years, proving to be an effective approach in providing accurate results and it has been expounded in many previous publications.

In the air quality forecasting field, the time-series approach is generally used to understand the cause-and-effect relationships. The ARIMA method has been widely applied in the last few decades to analyze and predict air pollutant concentrations [11–15]. One study accounted seasonal non-stationarity in time series models for short-term ozone level forecasts [16]; and simulation of the daily average PM10 concentrations were done at Ta-Liao [17], while another used ARIMA to forecast a full range of pollutants-ozone (O_3), nitrogen oxide (NO), nitrogen dioxide (NO_2) and carbon dioxide (CO) [18]. Hidden periodicities of the fine particle (PM10) time series were identified and used to increase the performance of the time series models [19]. The duration of cycle is calculated for observed PM10 levels in London as 365 days corresponding to a year, 7 days corresponding to a weak, 456 days corresponding to 15 months, and 183 days corresponding to 6 months and 25 days. In East Central Florida, nonlinear regression and ARIMA models were used for precipitation chemistry from 1978 to 1997 [20]. The trends in PM2.5 concentrations of Fuzhou, China between August 2014 and July 2016 were studied using the ARIMA [21]. Seasonal fluctuations of two years were identified as a result, where lower concentrations appear in warm days, while higher concentrations appear in cold days. Periodogram-based time series methodology was utilized to find the hidden periodic structure of monthly PM2.5 data, available for Paris between January 2000 and December 2019 [22].

Different studies focus on a wide variety of applications from predicting Air Quality Index (AQI) comprising every major pollutant to individual pollutant concentrations predictions. A study in Taiwan [23] mainly predicted multi-level AQI classifications, which helped forecast qualitative AQI at time t based on information up to t−1 when considering seasonal variables and three weather covariates: daily wind direction, daily average temperature and daily accumulated precipitation. Henry et al. [24] located nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction.

Numerical air quality forecasts in Hong Kong were improved using stochastic time series approach, namely ARIMA models [25]. The ARIMA model was used to predict fine particle (PM2.5) concentrations at the 35 air quality monitoring stations in Beijing in a span of 24 h resolving the issue of processing high-dimensional large-scale data and taking the forecasting data as one of the data sources for predicting the air quality [26]. Acknowledging the lack of data, the study employed the sliding window mechanism to deeply mine the high-dimensional temporal features for increasing the training dimensions to millions. Another study provided an in-depth analysis of all factors of air pollutants by correlation between those factors. Using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a prediction of future concentration of PM2.5 was made which gives the increasing value and provides the lowest and highest predictions of PM2.5 concentrations in the next year [27].

The main goal of this study is to find trends and forecast the air pollution in Sofia City, Bulgaria thorough development of time series predictive models for different air pollutants such as CO, NO_2, O_3 and fine particles (PM2.5). The ARIMA method is employed for predictive analytics at five local air quality monitoring stations with the aim of improving forecast accuracy at roadside, rural and urban places.

Sofia City was chosen for the case study since it is vulnerable to air pollution due to three main factors as follows:

- Geographic—The city is located in a valley surrounded by mountains, which keep the pollutants over the city for prolonged periods of time, especially during the winter, when fog and thermal inversions appear.

- Urban traffic—The city has over 1 million inhabitants, and there are almost 0.8 cars per adult. The cars in Sofia are more than the average in big cities in Europe (0.5 cars per adult), and the trend for car traveling continues to grow [28,29].
- Domestic heating—Part of the city's population is still using solid fuel for heating, which is mostly low-quality coal producing a lot of smoke and ash [30].

The rest of the paper is organized as follows. Section 2 describes the methods used for data preparation and analysis. Section 3 describes the obtained results of the air quality forecasting models. Section 4 discusses the main findings and forecasts, while Section 5 concludes the paper and gives directions for future work.

2. Materials and Methods

The data is collected by five air quality stations located within and near Sofia City. The air quality stations, shown in Figure 2, are managed by the Bulgarian Executive Environmental Agency (ExEA), which is a national reference within the European Environment Agency (EEA). The data ranges from 1 January 2015 to 31 December 2019, providing values for the levels of the polluters CO, NO_2, ozone O_3, PM2.5.

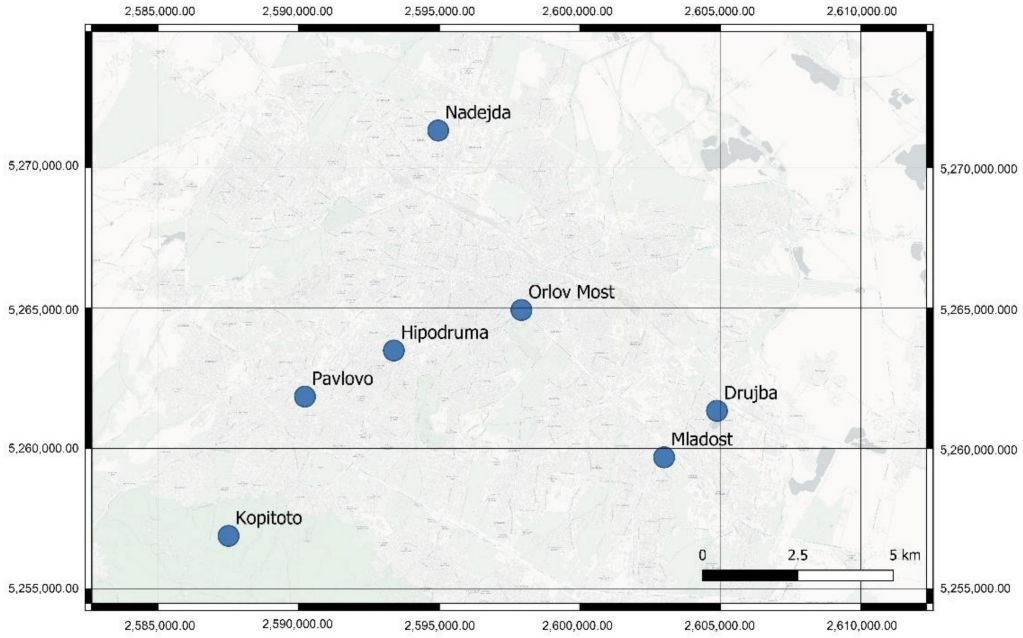


Figure 2. Air quality stations in Sofia City.

Since the measurements are not available for every pollutant in all air quality stations (see Table 1), the study is focused on the data obtained by the air quality stations at the neighborhoods of Pavlovo and Hipodruma and the region of Kopitoto, which is located close to Vitosha mountain. The measurements of $PM_{2.5}$ are available only for the neighborhood of Hipodruma.

2.1. Data Preparation

The initial raw data contains hourly measurements. Due to missing values and presence of noise in the data, both imputation and aggregation of data is performed. Granularity periods of 3 h, 6 h, 12 h and 24 h (1 day) are used.

2.1.1. Quality of Raw Data

The raw datasets are explored to check their quality. The results are summarized in Tables 2–5. The exploratory analysis performed looks to identify the percentage of the missing and the negative values. Such missing and negative values may be due to

temporary failure or bias of the sensors measuring the level of the polluter for the current time frame (1 h in the case of the current study). A threshold of 15% for missing or negative values is applied to discard the specific air quality station for the corresponding polluter. If the sum of the percentages of the missing and negative values for each line in Tables 2–5 is equal to or above 15%, the data is considered as incomplete or of low quality and not taken for further exploratory and time series analysis.

Table 1. Availability of polluters regarding districts of and near Sofia City. The available measurements are marked with black circles.

Neighborhood/Polluter	NO_2	CO	O_3	$PM_{2.5}$
Hipodruma	●	●	●	●
Pavlovo	●	●	●	○
Kopitoto	●	●	●	○
Mladost	●	●	○	○
Drujba	●	○	●	○
Nadejda	●	○	●	○
Orlov most	●	●	○	○

Table 2. Quality of the raw dataset for NO_2.

NO_2	Min	Mean	Max	Missing	Negatives
Hipodruma	−0.59	32.99	212.45	0.7%	0.1%
Pavlovo	−2.93	32.01	275.2	6.0%	0.4%
Kopitoto	−1.51	4.97	72.3	9.8%	0.4%
Mladost	0.0	30.31	229.13	24.1%	0.0%
Drujba	0.45	25.10	185.91	0.6%	0.0%
Nadejda	0.0	24.56	226.55	5.0%	0.0%
Orlov most	1.88	43.33	240.04	85.3%	0.0%

Table 3. Quality of the raw dataset for CO.

CO	Min	Mean	Max	Missing	Negatives
Hipodruma	0.0	0.63	7.86	1.9%	0.0%
Pavlovo	−0.14	0.66	7.1	3.1%	0.2%
Kopitoto	−0.24	0.32	2.46	8.9%	3.4%
Mladost	0.0	0.60	6.28	18.0%	0.0%
Drujba	-	-	-	-	-
Nadejda	-	-	-	-	-
Orlov most	0.14	0.82	7.59	85.0%	0.0%

Table 4. Quality of the raw dataset for O_3.

O_3	Min	Mean	Max	Missing	Negatives
Hipodruma	0.01	34.84	152.23	0.5%	0.0%
Pavlovo	−4.43	45.30	199.3	2.8%	0.5%
Kopitoto	−7.32	79.24	195.22	10.8%	0.2%
Mladost	-	-	-	-	-
Drujba	0.0	42.07	254.0	2.5%	0.0%
Nadejda	0.0	43.55	184.16	1.4%	0.0%
Orlov most	-	-	-	-	-

Table 5. Quality of the raw dataset for PM2.5.

PM2.5	Min	Mean	Max	Missing	Negatives
Hipodruma	0.0	23.95	580.27	12.0%	0.0%

When the threshold of 15% is applied, the data related to NO_2 at "Orlov most" (85.3% of missing values) and "Mladost" (24.1% of missing values) air quality stations and the data related to CO at "Orlov most" (85.0% of missing values) and "Mladost" (18.0% of missing values) air quality stations are discarded. For the rest of the data, the negative values are fixed to 0.0, supposing that there were some bias or failure in the sensors measuring the level of the polluter in a small portion of the time frame.

2.1.2. Imputation of Missing Values

Since for the chosen method in time series analysis, applied in the current research paper, it is essential to avoid missing values, they are handled in the raw dataset after the corrections described in previous subsection. It is worth mentioning that there are methods for time series analysis that do not require imputation, if certain conditions are met [31] and corresponding software implementations are available [32]. However, different methods for time series imputation have been tried and some of them are included in the interpolate functionality for data frames manipulation in the programming language Python [33] and in the more sophisticated library dedicated to missing values imputation [34]. Besides the simpler forward and backward imputation method, linear, quadratic and spline interpolation have been applied but the results were unsatisfactory. The issue is solved by applying a method of imputation by averaging the present values by year, month, day and hour. For example, in the raw dataset for NO_2 at "Kopitoto" air quality station there are missing values for the period from 4 p.m. 20 August 2015 until 6 p.m. 14 September 2015. Thus, all the values from the same time frame for the years 2016, 2017, 2018 and 2019 are taken to substitute the missing values. Since the data for these years in this time frame may again have missing values, only the present ones are used, and the missing values are filled by averaging of the values for the corresponding slot in the selected time frames.

2.1.3. Data Granularity

For the purpose of the study, data is aggregated using granularity of 3 h, 6 h, 12 h, 24 h (1 day). The time stamp for each aggregated value is taken from the end of the granularity period. For every present polluter and air quality station the data is aggregated, and separate forecasting time series models are created for each combination of polluter, district and granularity of data.

2.2. Analytical Methods

ARIMA (p, d, q) (P, D, Q) model is applied to perform data analysis, where p, d, q and P, D, Q represent continuity and seasonal auto-regression differences, respectively. It consists of three components Autoregressive (AR) model, Integrated (I) and Moving Average (MA) model. AR and MA are combined on data differenced for stationarity. Autocorrelation function (ACF) and Partial Autocorrelation (PACF) are used to select the best values for the models' parameters.

2.2.1. Autocorrelation and Partial Autocorrelation

In order to identify how the observations are correlated in the time series, the autocorrelation is used. The correlation coefficient is plotted by the ACF against the lag measured in terms of periods. The lag is related to a certain observation in time after which the first value is observed in the time series. Thus, ACF is calculated by the autocovariance of x_t and x_{t-n} as follows [35]:

$$ACF(n) = \frac{Cov(x_t,\ x_{t-n})}{Var(x_t)} \quad (1)$$

where $Cov(x_t,\ x_{t-n})$ is a covariance of variables x_t and x_{t-n}, and $Var(x_t)$ is a variance of the variable x_t. The covariance measures the relationship between two random variables, while the variance measures the variability The correlation coefficient ranges from -1 (negative relationship) to 1 (positive relationship). If there is no relationship between the variables, the correlation coefficient is 0.

Partial autocorrelation summarizes the relationship between an observation and observations at prior time period without taking into account the correlations between the intervening observations. The PAC is a simple correlation of x_t and x_{t-n}, which can be calculated as follows [36]:

$$PACF(n) = Corr[x_t - E^*(x_t|x_{t-1}, \ldots, x_{t-n+1}), x_{t-n}] \quad (2)$$

where $E^*(x_t|x_{t-1}, \ldots, x_{t-n+1})$ is the expected value of x_t. The optimal solution model for the time series is obtained according to the results from ACF and PACF.

2.2.2. Stationarity

The stationarity of a time series is an important condition that is required by the majority of time series forecasting algorithms. The stationarity means that the statistical characteristics of the time series such as mean, variance, autocorrelation, etc. do not change over time. In case of missing stationarity, the time series need to be stabilized before further analysis. A Stable History Period (SHP), namely a time period when the time series remain stable, can be used to resolve such issue [31]. The stability of the time series in this study is determined by ACF and PACF. Time series is stable, if ACF fluctuates around a fixed horizontal line with a gradual decay trend.

A differential transformation can be applied to the time series T_t to produce stationary time series S_t [26] as follows:

$$T_t = S_t - S_{t-1} \quad (3)$$

An alternative approach is to calculate the percent of change, as follows:

$$T_t = \frac{S_t - S_{t-1}}{S_{t-1}} \quad (4)$$

For example, the time series of CO can be non-stationary, and a differential transformation can be applied to achieve stationarity. The first-order differencing determines the difference in observations between two subsequent periods. If stationarity is not achieved, a second-order differencing is applied on the first-order differencing. The process can be repeated until the time series become stationary.

2.2.3. ARIMA Method

The ARIMA employs an AR model in combination with a MA model to perform time series forecasting. The main parameters considered are as follows:

- Number of previous observations (p);
- Degree of differencing (d);
- Size of the moving average (q).

The AR model shows the dependence of one observation on an earlier time period. The AR model obtains p previous observations as follows [36]:

$$y_t = \alpha + \sum_{i-1}^{p} \varphi_i y_{t-i} + e_t \quad (5)$$

where y_t is predictand for time t from a normal distribution and y_{t-i} defines p previous observations of the same time series. φ_i denotes the regression coefficients, α is a constant and e_t is the random error term. The order p for the AR(p) model is selected based on the significant spikes of PACF. An additional indicator is the slow decay of ACF.

The MA performs forecasting relying on moving averages of the past random error terms as follows:

$$y_t = \mu + \sum_{i-1}^{p} \theta_i e_{t-i} \quad (6)$$

where θ_t represents the regression coefficients, q is the order of moving average, and μ is a constant. The order q for the MA(q) model is obtained from the ACF, if it has a sharp cut-off after lag q. The PACF decays slowly in this case.

The ARIMA model can be defined as follows:

$$\varphi_p(B)(1-B)^d y_t = \theta_q(B)e_t \tag{7}$$

where B is a backshift operator, p is the order of autoregression, d is the order of differencing and q is the order of moving average.

2.2.4. Evaluation Metrics

The best combination of p and q can be found using an objective function, which can measure the performance of the model on a validation set. Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are mathematical methods for scoring statistical models and selection of one that best one in a candidate model space. AIC evaluates the model's goodness-of-fit to the data, as follows [37]:

$$\text{AIC} = -2l + 2k \tag{8}$$

where l is a log-likelihood, and k is a number of parameters in the candidate model. Additional indicator n is added to the BIC, which defines the number of samples used for fitting as follows [27]:

$$\text{BIC} = -2l + k \log n \tag{9}$$

The BIC is derived under Bayesian network, while the AIC is elaborated by adjusting biased empirical information. It tends to select a less complex model in comparison to AIC. The coefficient of determination, called R-squared, measures the proportion of the variance of the dependent variable, predictable form the dependent variables in the model. It can also be used for evaluation of the model's goodness-of-fit to the data. R-squared varied between 0% and 100%, where 0% means that the model does not explain the variability of the response data around its mean and 100% means that the model explains all the variability.

2.2.5. Software Libraries

The air quality models are implemented using the Python programming language and libraries matplotlib and seaborn for data visualization and pandas, numpy and statsmodel for the data analysis.

3. Experimental Results

The results from the exploratory analysis of the data are shown in Figures 3 and 4.

3.1. Analytical Methods

Figures 5–7 show that there are some seasonalities for the NO_2 data aggregated by 1 h, 3 h and 12 h. The plot of Figure 5 shows a seasonality with lag 24, corresponding to the pick in the autocorrelation lower part, and that is expected because of the repeated daily activity by hour within the city. The second significant pick is observed 24 h after the first one, and namely, near the 48th tick. A similar trend is observed in Figures 6 and 7 and the most significant picks in the autocorrelation part of the plots are at 24/3 = 8 and 24/12 = 2, respectively. As expected, the second significant picks shown in for Figures 6 and 7 are observed at the 16th and 4th ticks, which corresponds to the 2-day lags after the first tick.

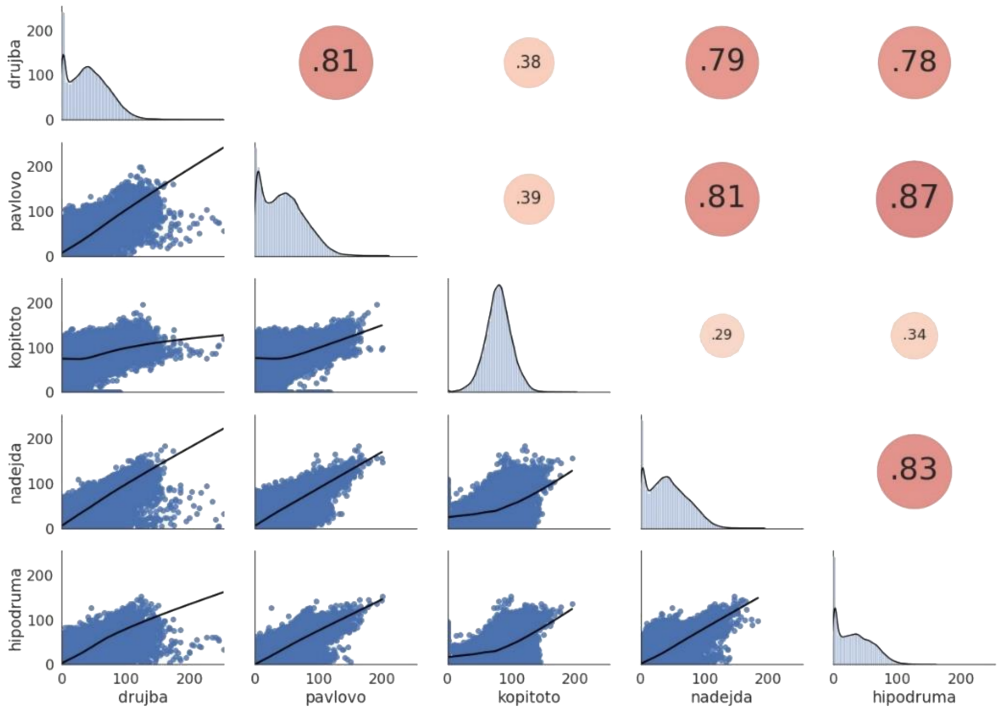


Figure 3. Interaction relationships—scatterplot (lower left triangular area) and Pearson correlation values (upper right triangular area) and histogram (on the diagonal) of the O_3 at each air quality station measured per hour.

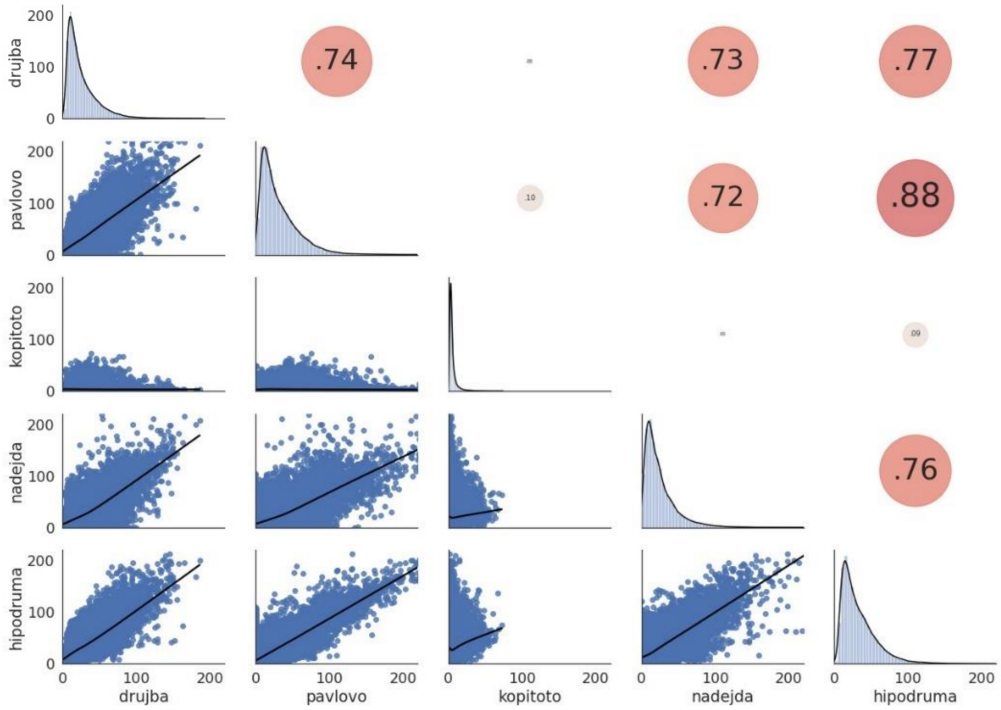


Figure 4. Interaction relationships—scatterplot (lower left triangular area) and Pearson correlation values (upper right triangular area) and histogram (on the diagonal) of the NO_2 at each air quality station measured per hour.

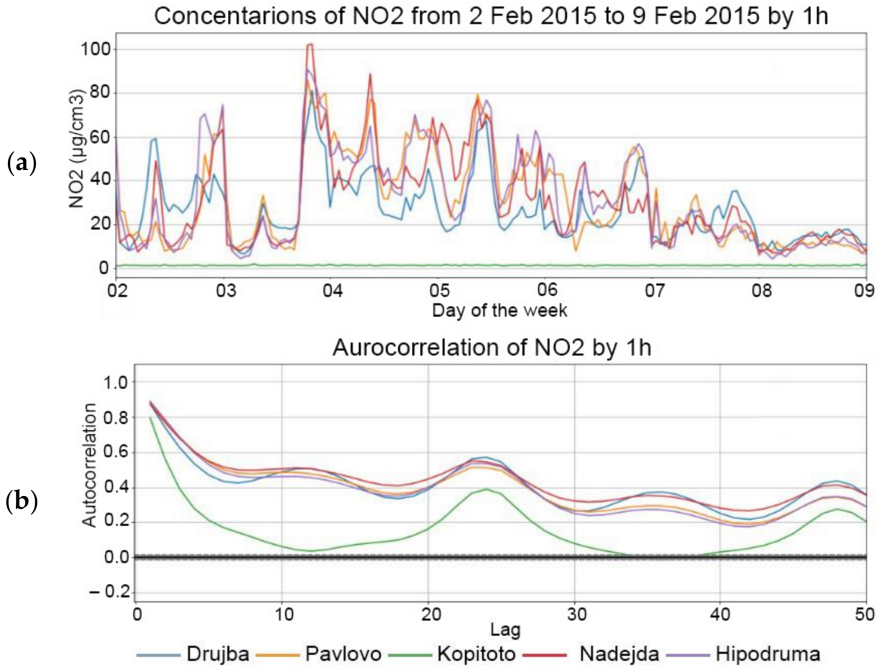


Figure 5. The signal for the different stations corresponding to NO_2 measured by 1 h (**a**) and the corresponding autocorrelation (**b**).

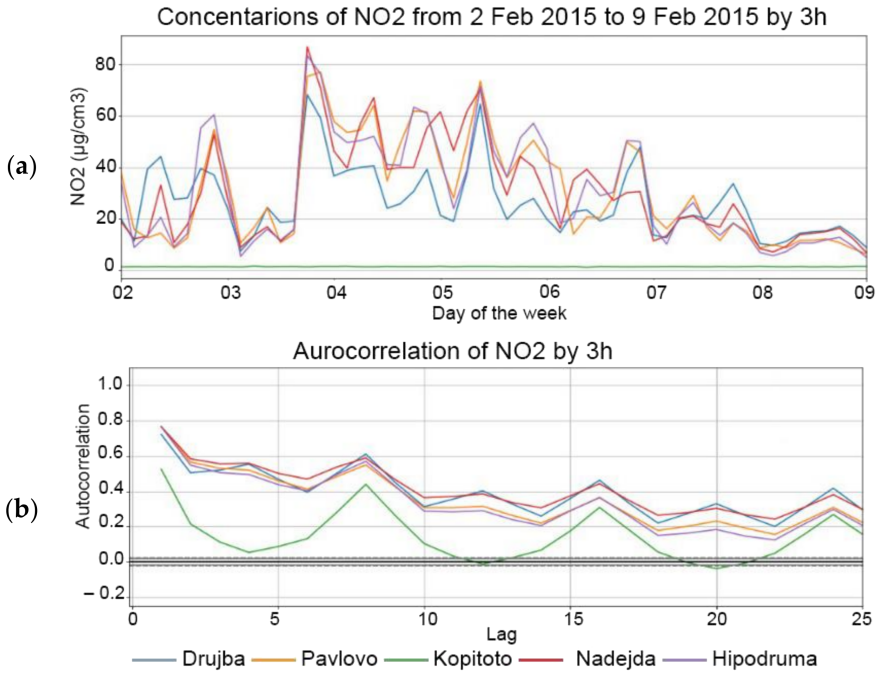


Figure 6. The signal for the different stations corresponding to NO_2 measured by 3 h (**a**) and the corresponding autocorrelation (**b**).

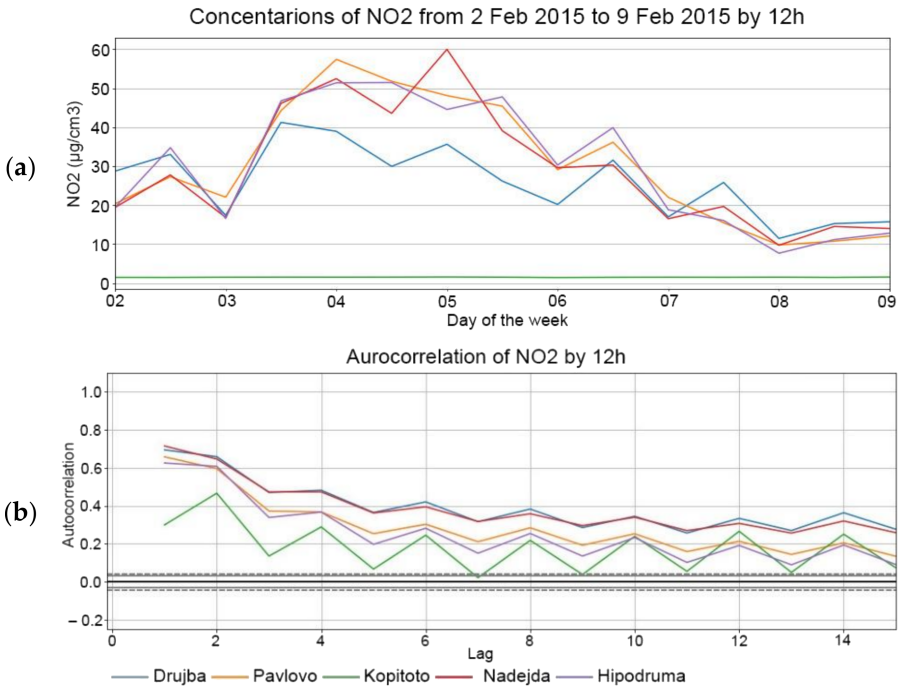


Figure 7. The signal for the different stations corresponding to NO_2 measured by 12 h (**a**) and the corresponding autocorrelation (**b**).

For the sake of brevity, plots for NO_2 are shown only but one observes similar behavior for the rest of the pollutants.

Partial autocorrelation, as described in Section 2.2.2, was also applied to check what is the direct relation of lagged members to the first member of the series (without taking into account the ones in between). Figure 8 presents an example for NO_2 measurements in Drujba air quality station. It can be seen that for one 1 h aggregated data there is a direct relationship of the members of the series at and immediately before the 24th tick, which is expected since the granularity of the data by hour and the daily activities withing the city. Similar behavior can be observed in the plots corresponding to the 3 h and 12 h granularity of the data. There are direct relationships of the first lag and the lags on and immediately before the 24/3 = 8th lag for 3 h granularity and the first lag and the lags on and immediately before the 24/12 = 2nd lag for the 12 h granularity. Similarly, as in the autocorrelation part, further picks are observed in the partial autocorrelation plots, which are located immediately before the 48th, 6th and 4th lags.

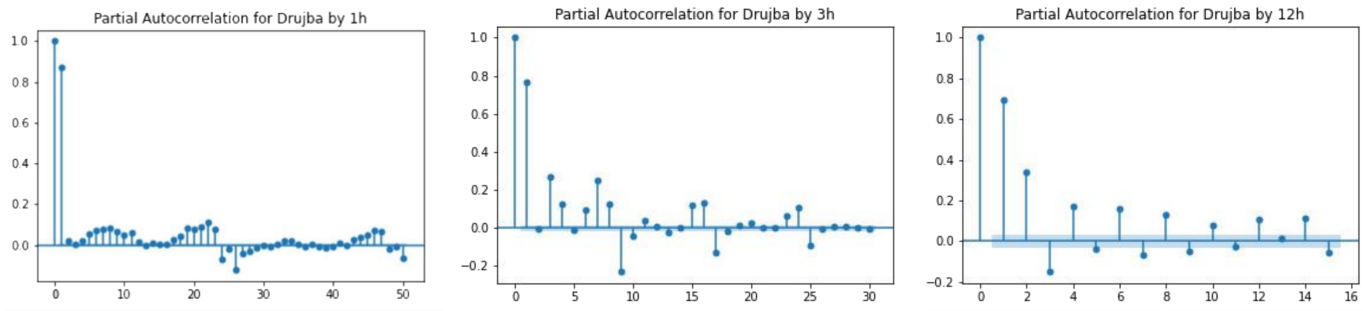


Figure 8. Partial autocorrelation of NO_2 in "Drujba" air quality station measured at 1 h granularity (**a**), 3 h granularity (**b**) and 12 h granularity (**c**).

As described in Section 2.2.2, two statistical tests are used to check for (non) stationarity of the signals, namely the augmented Dickey–Fuller (ADF) and the Kwiatkowski–Philips–Schmidt–Shin (KPSS) tests. A common misconception is that both tests can be used interchangeably, which can lead to contradictory conclusions about the stationarity of the tested signal. A key difference between ADF and KPSS is how one states the null hypothesis and consequently the interpretation of the *p*-value, which is the opposite in each other.

The ADF test should be interpreted as follows:

- If *p*-value > 0.05 (test value > 5% test critical value), one fails to reject the null hypothesis (H0) and the data is considered as non-stationary;
- If *p*-value $\leq$ 0.05 (test value $\leq$ 5% test critical value), one rejects the null hypothesis (H0) and the data is considered as stationary.

On the other hand, the *p*-value of the KPSS test is interpreted in the opposite side.

The critical values for ADF test are 1% (−3.430), 5% (−2.862), 10% (−2.567) and for the KPSS test: 1% (0.739), 5% (0.463), 10% (0.347).

The results of applying both tests on the data into consideration are stated in Tables 6–8.

Table 6. Values of the ADF and KPSS tests applied on the NO_2 dataset for granularities: 1 h.

NO_2/1 h	ADF	*p*-Value	Lags	KPSS	*p*-Value	Lags
Hipodruma	−15.455	0.0	52	0.165	0.1	109
Pavlovo	−17.265	0.0	55	0.174	0.1	110
Kopitoto	−19.228	0.0	50	5.618	0.01	82
Drujba	−10.814	0.0	40	0.30	0.1	109
Nadejda	−15.580	0.0	52	0.337	0.1	111

Table 7. Values of the ADF and KPSS tests applied on the NO_2 dataset for granularities: 3 h.

NO_2/3 h	ADF	*p*-Value	Lags	KPSS	*p*-Value	Lags
Hipodruma	−12.864	0.0	40	0.119	0.1	68
Pavlovo	−12.160	0.0	40	0.123	0.1	67
Kopitoto	−12.449	0.0	41	3.586	0.01	64
Drujba	−10.814	0.0	40	0.191	0.1	69
Nadejda	−10.479	0.0	40	0.220	0.1	69

Table 8. Values of the ADF and KPSS tests applied on the NO_2 dataset for granularities: 12 h.

NO_2/12 h	ADF	*p*-Value	Lags	KPSS	*p*-Value	Lags
Hipodruma	−7.306	0.0	30	0.087	0.1	31
Pavlovo	−6.745	0.0	27	0.084	0.1	32
Kopitoto	−6.511	0.0	25	2.361	0.01	30
Drujba	−5.763	0.0001	28	0.116	0.1	34
Nadejda	−6.334	0.0	28	0.141	0.1	33

The Python's library statsmodels is used, where the output for ADF and KPSS tests is:

- Value of the test statistic
- The *p*-value
- Number of lags considered for the test
- The critical value cut-offs

For the purpose of the current paper, we consider the typical critical test value 5%, that is, *p*-value of 0.05. Based on the assumptions made for the ADF and KPSS tests, the main observation is that even without significant analytical transformation on the signal for NO_2, for all levels of granularity except Kopitoto for KPSS, all other air quality stations show significant level of signal stationarity. Additionally, that significant stationarity is confirmed by both tests. That is, for all considered air quality stations, the ADF *p*-value < 0.05, and

again for all air quality stations except Kopitoto, the KPSS p-value > 0.05. Furthermore, the last consideration is valid not only for the 5% critical values of the corresponding tests, but also for the further 10% critical values. However, for the station obtained from Kopitoto and the KPSS test, we observe that for all levels of granularity, the test statistic > critical% value and p-value < 0.05, which rejects the null hypothesis, and we conclude that the signal is rather non-stationary.

As a conclusion, both tests show significant levels of stationarity for all air quality stations except Kopitoto. For Kopitoto, ADF shows stationarity and KPSS shows non-stationarity. Therefore, the series for Kopitoto are non-stationarity to some extent and lag difference is to be used to make the series more stationary.

Lag difference for the 1 h, 3 h and 12 h granularity of the initially preprocessed data is applied. The number of lag differences for the corresponding granularity levels are based on the observations regarding the autocorrelation and partial autocorrelation plots. The results after applying the lag difference procedure on the corresponding signals and their autocorrelation plots are shown in Figures 9–11. The plots clearly show the improvement of stationarity compared to the plots in Figures 5–7.

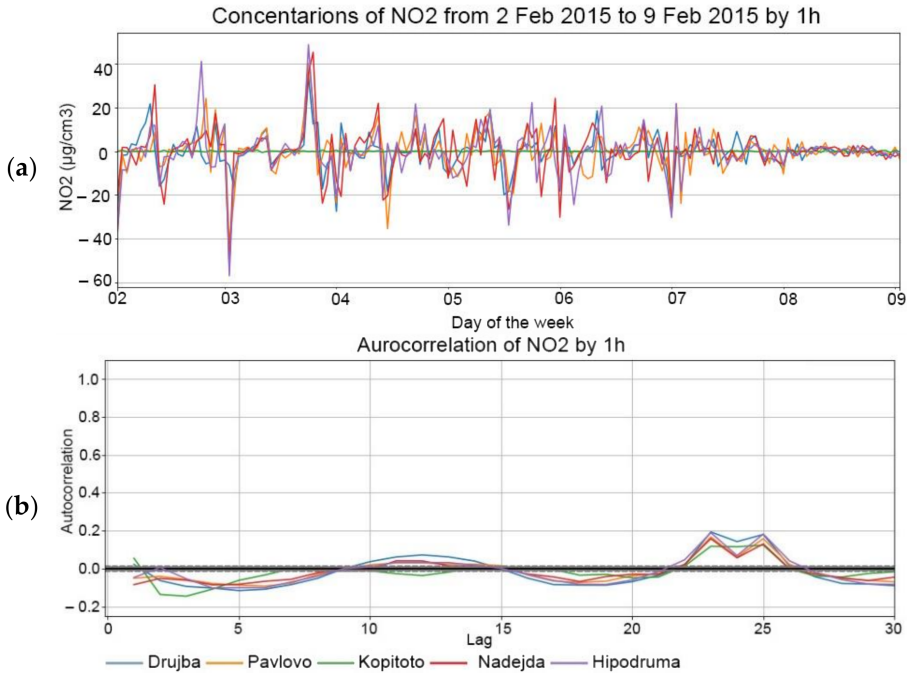


Figure 9. The signal for the different stations corresponding to NO_2 for granularity 1 h with lag difference 1 (**a**) and its autocorrelation (**b**).

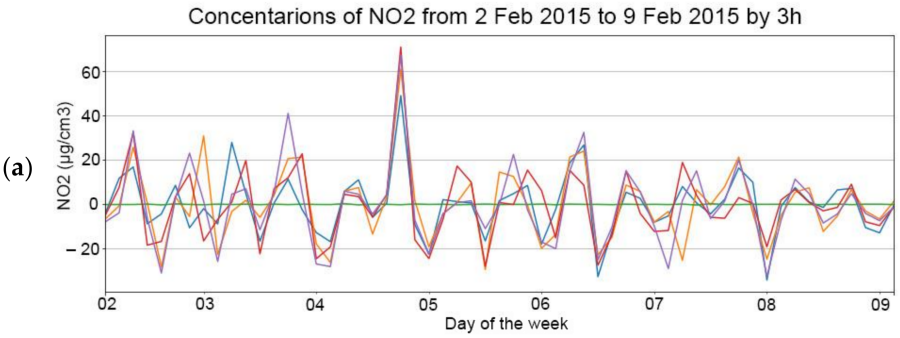


Figure 10. *Cont.*

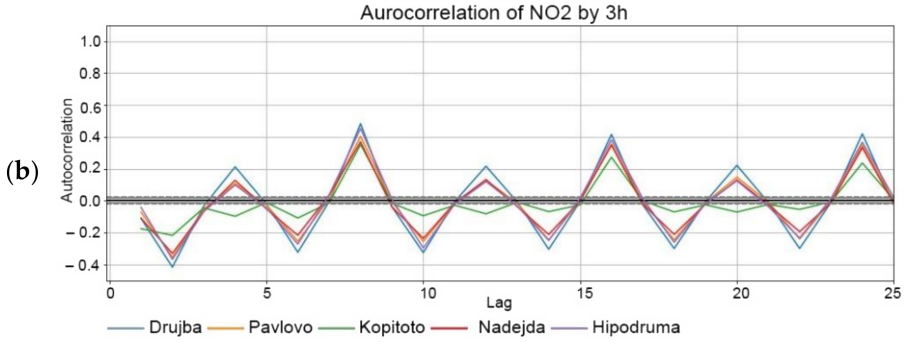


Figure 10. The signal for the different stations corresponding to NO_2 for granularity 3 h with lag difference 1 (**a**) and its autocorrelation (**b**).

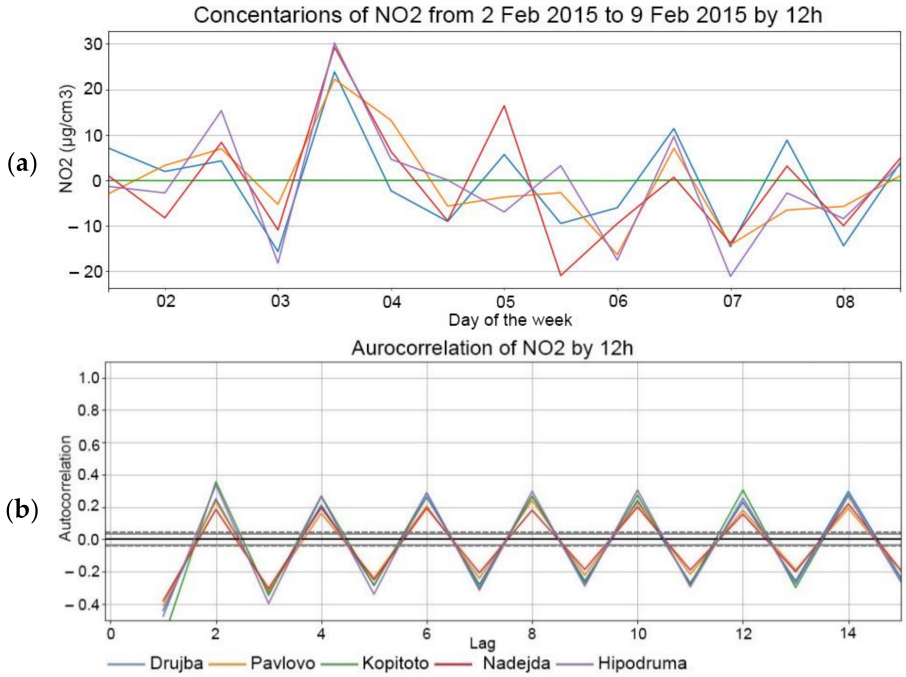
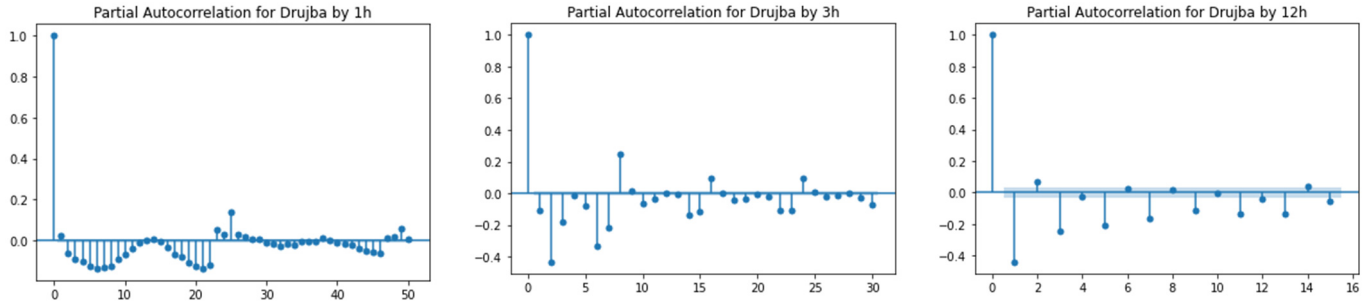


Figure 11. The signal for the different stations corresponding to NO_2 for granularity 12 h with lag difference 1 (**a**) and its autocorrelation (**b**).

Figure 12 shows the partial autocorrelation of NO2 in "Drujba" air quality station measured at 1 h, 3 h and 12 h granularity with lag difference 1. Numerically, the improvement of stationarity levels after applying the lag differences are shown in Table 9, where one can check the ADF and KPSS tests applied on the NO_2 dataset for granularities: 1 h with lag difference 24, 3 h with lag difference 8 and 12 h with lag difference 2. In Table 6, only the corresponding test values are shown, where the improvement consists of the reduction of the values of ADF and KPSS tests for the corresponding data granularities and lag differences. For Kopitoto air quality station, the KPSS test, having shown non-stationarity with critical *p*-value 0.05 before, after the lag difference, has a test value significantly lower than the critical test value. Therefore, the lag difference turned the signals corresponding to Kopitoto from non-stationary to stationary for all granularity levels.



(**a**) (**b**) (**c**)

Figure 12. Partial autocorrelation of NO_2 in "Drujba" air quality station measured at 1 h granularity (**a**), 3 h granularity (**b**) and 12 h granularity (**c**) with lag difference 1.

Table 9. Values of the ADF and KPSS tests applied on the NO_2 dataset for granularities: 1 h with lag difference 1, 3 h with lag difference 1 and 12 h with lag difference 1.

NO_2	ADF 1 h/1	ADF 3 h/1	ADF 12 h/1	KPSS 1 h/1	KPSS 3 h/1	KPSS 12 h/1
Hipodruma	−33.392	−25.186	−17.470	0.002	0.002	0.031
Pavlovo	−33.471	−24.916	−16.753	0.001	0.001	0.025
Kopitoto	−38.502	−27.869	−18.243	0.003	0.008	0.055
Drujba	−32.502	−25.718	−16.040	0.002	0.001	0.021
Nadejda	−32.888	−25.421	−16.512	0.003	0.002	0.026

3.2. ARIMA Models

There is a rule of the thumb on how to choose the parameters for AR (p), lag difference (d) and MA (q). From the autocorrelation graphs one may conclude that there is a significant number of consequent positive lags; therefore, one needs to apply more differencing and, on the other hand, if more lags are negative, the series may be over-differenced already. Higher degree lag differences have been tested but they produced too many negative autocorrelation lags. Therefore, for that data more than 1 degree lag difference may not be needed. The parameter p can be chosen to correspond to the most significant lag in the partial autocorrelation plot. From Figure 11, one concludes that $p = 7$ may be a good choice for 1 h granularity, and 2 and 1 for 3 h and 12 h granularity, respectively. The parameter q, corresponding to the MA degree can be chosen roughly by counting the number of lags crossing the threshold of $-0.2/+0.2$ threshold in the autocorrelation plots. Therefore, for 12 h and 3 h granularity q can be chosen larger compared to 1 h granularity.

Because the above rule is not mathematically grounded, a grid search is applied based on the Python's library pmdarima. For the grid search of the parameters the AIC and BIC metrics were used in combination with the ADF test. As a final evaluation metric mean absolute error (MAE) is applied.

Figures 13 and 14 present plots with predictions from ARIMA models of NO_2 for optimally chosen parameters p, d and q for granularity 1 h and 3 h, respectively. Table 10 shows the corresponding numerical values.

The results for CO in the stations located in the neighborhoods Hipodruma, Pavlovo and Kopitoto passed through the same grid search and ARIMA based time series analysis pipeline are presented in Table 11.

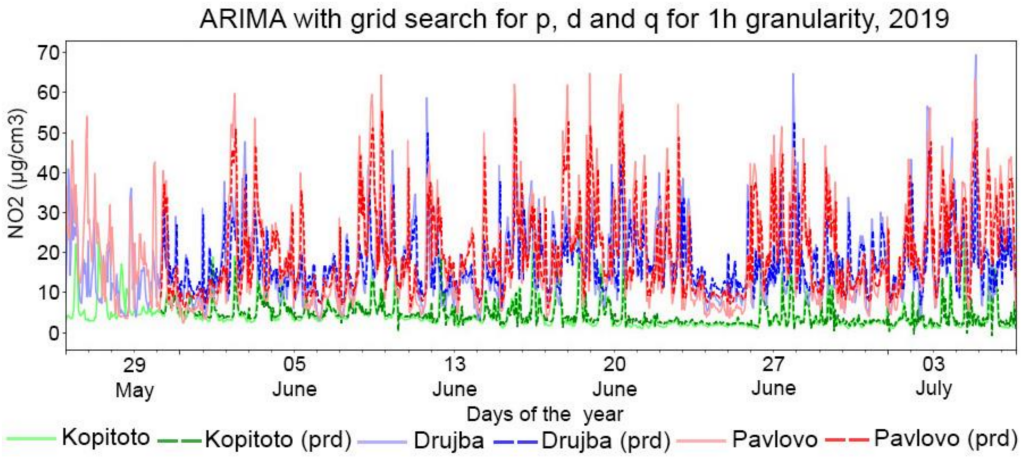


Figure 13. ARIMA model with predictions for NO_2, 1 h granularity, based on grid search and ACF, PACF lags consideration.

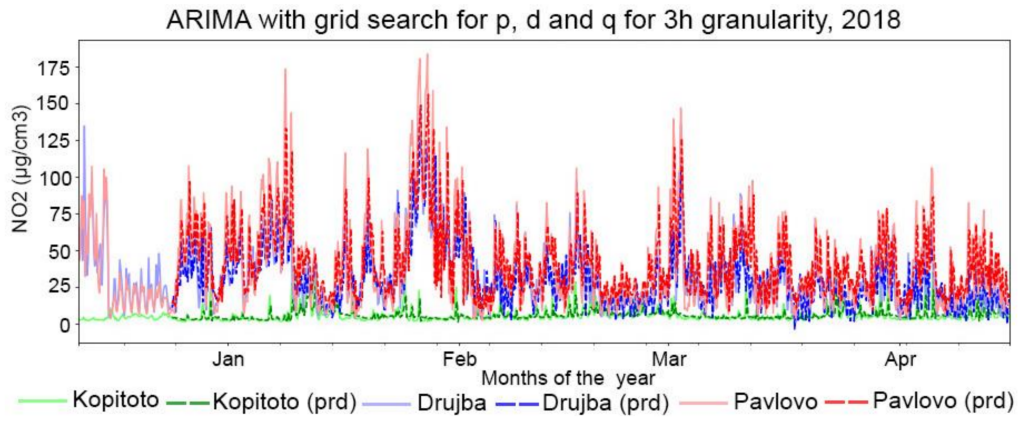


Figure 14. ARIMA model with predictions for NO_2, 3 h granularity, based on grid search and ACF, PACF lags consideration.

Table 10. Mean absolute error of NO_2 for the chosen parameters p, d, q, for 1 h and 3 h granularity.

NO_2	MAE 1 h	RMSE 1h	*p*	*d*	*q*	MAE 3 h	RMSE 1h	*p*	*d*	*q*
Hipodruma	5.12	8.17	2	0	2	12.2	16.23	5	0	1
Pavlovo	5.77	8.39	3	1	1	14.12	19.18	3	1	2
Kopitoto	1.52	2.57	5	1	3	2.53	4.29	2	1	1
Drujba	4.69	6.84	5	1	1	9.80	7.56	5	0	5
Nadejda	4.42	7.11	2	1	1	9.45	7.11	4	1	3

Table 11. Mean absolute error of CO for the chosen parameters p, d, q, for 1 h and 3 h granularity.

NO_2	MAE 1 h	RMSE 1h	*p*	*d*	*q*	MAE 3 h	RMSE 1h	*p*	*d*	*q*
Hipodruma	0.046	0.074	3	0	2	0.23	0.37	3	0	1
Pavlovo	0.044	0.064	3	0	1	0.25	0.40	4	0	1
Kopitoto	0.018	0.029	4	1	3	0.029	0.045	3	1	1

4. Discussion

The ARIMA method is one of the most commonly used ones to forecast the future values of time series. The predictions obtained from the application of this method would converge the mean of the time series for a further period if a stationarity is ensured [17].

The ARIMA method gives consistent results for short-term predictions. It is not preferable for long-term predictions, especially in the case of periodicity in time series. Long-term predictions, e.g., annual, might be more useful for city authorities to define long-term policies and measures. Thus, the periodicity of the time series has to be considered in order to provide reasoning based on the hidden data cycles. The knowledge about the underlying structure can be used as the measurement of effectiveness of the measures taken by policy makers [21].

ADF and KPSS statistical tests are used to check stationarity of the data for granularities of 1 h with lag difference from 1 until 24, 3 h with lag difference from 1 until 8 and 12 h with lag difference 1 and 2. Both tests show significant level of stationarity in NO_2 measurements for all air quality stations except "Kopitoto", where ADF shows stationarity and KPSS shows non-stationarity. The stationarity levels are improved after applying the corresponding optimal lag differences for all granularity levels, obtained from the Python's library pmdarima and optimizing the ADF statistics. The inter- and intra-annual fluctuations in data could be further explored by the least-squares wavelet analysis (LSWA), which avoids interpolation and/or gap filling and decomposes the time series into the time-frequency domain [32]. The SHP representing both inter- and intra-annual fluctuations can be successfully determined in the time series [31]

The ARIMA method is successfully applied to perform forecasting of the air pollutants considering different granularities. The developed ARIMA models show the powerfulness of the time series to predict pollutant concentrations. They work quite well, as it can be seen in Figures 13 and 14, especially for 1 h granularity. The verification of the results show that the predicted values are close to the actual ones. The mean absolute error varies between 1.52 and 5.12 for 1 h granularity and 2.53 and 14.12 for 3 h granularity. The ARIMA method often outperforms sophisticated structural methods with short-term predictions for one-step forecasting on univariate datasets. Therefore, the developed ARIMA models within this study can be used as a benchmark for evaluation of different forecasting methods. They also can be combined with other techniques fully mine time series characteristics [38].

The ARIMA method requires data on time series in question only. This is an advantage, especially in case of large volume of time series as it is in the current case. Thus, it is simple and efficient. Furthermore, additional predictive methods will be explored, such as Long Short-Term Memory [LSTM], which are proven to provide predictions closer to ARIMA [4].

5. Conclusions

The paper presents the results from the application of time series analysis for the prediction of air quality in Sofia, Bulgaria. Specifically, the ARIMA method is employed to predict the levels of CO, NO_2, O_3 and PM2.5 at five local air quality monitoring stations. The proposed approach aims to improve forecast accuracy at roadside, rural and urban places. ADF and KPSS statistical tests are successfully applied to improve the stationarity levels of the data.

The proposed autoregressive forecasting models predict the future values based on the past values measured for the air pollutants. Thus, the implicit assumption is made that the future will resemble the past, without considering the influence of weather conditions. Future work should include the development of predictive models using machine learning methods. Meteorological data such as temperature, humidity and atmosphere pressure will be used to enrich the air quality datasets to obtain deeper insights about the causes and trends of air pollution. Further applications of ARIMA models will be done on respiratory diseases to find trends in seasonal variation, which can support the effective management of disease burden.

Author Contributions: Conceptualization, E.M. and D.P.-A.; methodology, E.M.; software, E.M.; formal analysis, E.M.; investigation, E.M. and S.M.; resources, E.M.; data curation, E.M.; writing—original draft preparation, E.M., D.P.-A. and S.M.; writing—review and editing, D.P.-A.; visualization, E.M.; supervision, D.P.-A.; project administration, D.P.-A.; funding acquisition, D.P.-A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Horizon 2020 WIDESPREAD-2018-2020 TEAMING Phase 2 Programme under grant agreement no. 857155, by Operational Programme Science and Education for Smart Growth under Grant Agreement No. BG05M2OP001-1.003-0002-C01 and by the Bulgarian National Science fund under agreement no. KP-06-N 32/5.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for the analysis is available from the Executive Environment Agency in Bulgaria upon a request (http://eea.government.bg/en) (accessed on 18 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. EEA. Air pollution: How It Affects Our Health. Available online: https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution (accessed on 21 April 2022).
2. EEA. Premature Deaths Attributed to PM2.5 at NUTS3 Level for European Countries in 2019, Normalized by Population. Available online: https://www.eea.europa.eu/data-and-maps/figures/premature-deaths-attributed-to-pm2 (accessed on 21 April 2022).
3. Reuters. No Country Met WHO Air Quality Standards in 2021, Survey Shows. Available online: https://www.reuters.com/business/environment/no-country-met-who-air-quality-standards-2021-data-2022-03-22 (accessed on 21 April 2022).
4. Spyrou, E.D.; Tsoulos, I.; Stylios, C. Applying and Comparing LSTM and ARIMA to Predict CO Levels for a Time-Series Measurements in a Port Area. *Signals* **2022**, *3*, 235–248. [CrossRef]
5. Agarwal, S.; Sharma, S.; Suresh, R.; Rahman, M.H.; Vranckx, S.; Maiheu, B.; Blyth, L.; Janssen, S.; Gargava, P.; Shukla, V.; et al. Air quality forecasting using artificial neural networks with real time dynamic error correction in highly polluted regions. *Sci. Total Environ.* **2020**, *735*, 139454. [CrossRef] [PubMed]
6. Zhang, K.; Thé, J.; Xie, G.; Yu, H. Multi-step ahead forecasting of regional air quality using spatial-temporal deep neural networks: A case study of Huaihai Economic Zone. *J. Clean. Prod.* **2020**, *277*, 123231. [CrossRef]
7. Eslami, E.; Salman, A.K.; Choi, Y.; Sayeed, A.; Lops, Y. A data ensemble approach for real-time air quality forecasting using extremely randomized trees and deep neural networks. *Neural Comput. Appl.* **2020**, *32*, 7563–7579.
8. Palvanov, A.; Cho, Y.I. Visnet: Deep convolutional neural networks for forecasting atmospheric visibility. *Sensors* **2019**, *19*, 1343. [CrossRef] [PubMed]
9. Wang, P.; Zhang, G.; Chen, F.; He, Y. A hybrid wavelet model applied for forecasting PM2. 5 concentrations in Taiyuan city, China. *Atmos. Pollut. Res.* **2019**, *10*, 1884–1894. [CrossRef]
10. Cheng, Y.; Zhang, H.; Liu, Z.; Chen, L.; Wang, P. Hybrid algorithm for short-term forecasting of PM2.5 in China. *Atmos. Environ.* **2019**, *200*, 264–279. [CrossRef]
11. Shi, J.P.; Harrison, R.M. Regression modeling of hourly NOx and NO2 concentrations in urban air in London. *Atmos. Environ.* **1997**, *31*, 4081–4094. [CrossRef]
12. Milionis, A.E.; Davies, T.D. Regression and stochastic models for air pollution—I, review, comments and suggestions. *Atmos. Environ.* **1994**, *28*, 2801–2810. [CrossRef]
13. Zennetti, P. *Air Pollution Modeling: Theories, Computational Methods and Available Software*, 1st ed.; Springer Science + Business Media: New York, NY, USA, 1990.
14. Goyal, P.; Chan, A.T.; Jaiswal, N. Statistical models for the prediction of respirable suspended particulate matter in urban cities. *Atmos. Environ.* **2006**, *40*, 2068–2077. [CrossRef]
15. Tripathi, O.P.; Jennings, S.G.; O'Dowd, C.D.; Coleman, L.; Leinert, S.; O'Leary, B.; Moran, E.; O'Doherty, S.J.; Spain, T.G. Statistical analysis of eight surface ozone measurement series for various sites in Ireland. *J. Geophys. Res.* **2010**, *115*, 1–20.
16. Kim, S.E.; Kumar, A. Accounting seasonal nonstationarity in time series models for short-term ozone level forecast. *Stoch. Environ. Res. Risk Assess.* **2005**, *19*, 241–248.
17. Liu, P.W.G. Simulation of the daily average PM10 concentrations at Ta-Liao with Box–Jenkins time series models and multivariate analysis. *Atmos. Environ.* **2009**, *43*, 2104–2113. [CrossRef]
18. Kumar, U.; Jain, V.K. ARIMA forecasting of ambient air pollutants (O_3, NO, NO_2 and CO). *Stoch. Environ. Res. Risk Assess.* **2010**, *24*, 751–760. [CrossRef]
19. Okkaoğlu, Y.; Akdi, Y.; Ünlü, K.D. Daily PM10, periodicity and harmonic regression model: The case of London. *Atmos. Environ.* **2020**, *238*, 117755. [CrossRef]
20. Nickerson, D.M.; Madsen, B.C. Nonlinear regression and ARIMA models for precipitation chemistry in East Central Florida from 1978 to 1997. *Environ. Pollut.* **2005**, *135*, 371–379. [CrossRef]
21. Zhang, L.; Lin, J.; Qiu, R.; Hu, X.; Zhang, H.; Chen, Q.; Wang, J. Trend analysis and forecast of PM2. 5 in Fuzhou, China using the ARIMA model. *Ecol. Indic.* **2018**, *95*, 702–710. [CrossRef]
22. Akdi, Y.; Gölveren, E.; Ünlü, K.D.; Yücel, M.E. Modeling and forecasting of monthly PM2.5 emission of Paris by periodogram-based time series methodology. *Environ. Monit. Assess.* **2021**, *193*, 622. [CrossRef]
23. Chen, C.W.S.; Chiu, L.M. Ordinal Time Series Forecasting of the Air Quality Index. *Entropy* **2021**, *23*, 1167. [CrossRef]

24. Henry, R.C.; Chang, Y.; Spiegelman, C.H. Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction. *Atmos. Environ.* **2002**, *36*, 2237–2244. [CrossRef]
25. Liu, T.; Lau, A.K.; Sandbrink, K.; Fung, J.C. Time Series Forecasting of Air Quality Based on Regional Numerical Modeling in Hong Kong. *J. Geophys. Res. Atmos.* **2018**, *123*, 4175–4196. [CrossRef]
26. Avinash, S.B.; Chaluvaraj, M.; Devanand, N.V.; Raju, H.; Kousar, M.G. Review on Air Quality Prediction Using ARIMA and Neural Network. *Int. Res. J. Eng. Technol.* **2021**, *8*, 473–476.
27. Bhatti, U.A.; Yan, Y.; Zhou, M.; Ali, S.; Hussain, A.; Qingsong, H.; Yu, Z.; Yuan, L. Time Series Analysis and Forecasting of Air Pollution Particulate Matter (PM2.5): An SARIMA and Factor Analysis Approach. *IEEE Access* **2021**, *9*, 41019–41031. [CrossRef]
28. Hurst, L. Bulgarian Citizens Try to Challenge Sofia in Court over Air Pollution Levels. Available online: https://www.euronews.com/green/2021/06/02/bulgarian-citizens-try-to-challenge-sofia-in-court-over-air-pollution-levels (accessed on 21 April 2022).
29. Lee, K.; Bernard, Y.; Dallmann, T.; Braun, C.; Miller, J. Impacts of a Low-Emission Zone in Sofia. The Real Urban Emissions Initiative. Available online: https://www.trueinitiative.org/media/792101/impacts-of-lez-in-sofia-true-report-en.pdf (accessed on 21 April 2022).
30. EEA. Eco-Innovation for Air Quality. 21st European Forum on Eco-Innovation. 2018. Available online: http://eea.government.bg/en/news/EcoAP-report.pdf (accessed on 21 April 2022).
31. Ghaderpour, E.; Vujadinovic, T. The Potential of the Least-Squares Spectral and Cross-Wavelet Analyses for Near-Real-Time Disturbance Detection within Unequally Spaced Satellite Image Time Series. *Remote Sens.* **2020**, *12*, 2446. [CrossRef]
32. Ghaderpour, E. JUST: MATLAB and python software for change detection and time series analysis. *GPS Solut.* **2021**, *25*, 85. [CrossRef]
33. Pandas. Available online: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html (accessed on 21 April 2022).
34. Impyute. Available online: https://impyute.readthedocs.io/en/master/api/time_series_imputation.html (accessed on 21 April 2022).
35. Morf, M.; Vieira, A.; Kailath, T. Covariance characterization by partial autocorrelation matrices. *Ann. Statist.* **1978**, *6*, 643–648. [CrossRef]
36. Akaike, H. Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer Series in Statistics; Springer: New York, NY, USA, 1998; pp. 199–213.
37. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
38. Chen, S.; Nong, Y.; Chen, Z.; Liang, D.; Lu, Y.; Qin, Y. The CEEMD-LSTM-ARIMA Model and Its Application in Time Series Prediction. *J. Phys. Conf. Ser.* **2022**, *2179*, 012012.