

Article

Predicting PM_{2.5}, PM₁₀, SO₂, NO₂, NO and CO Air Pollutant Values with Linear Regression in R Language

Zoltan Kazi , Snezana Filip  and Ljubica Kazi * 

Technical Faculty “Mihajlo Pupin”, University of Novi Sad, 23000 Zrenjanin, Serbia; zoltan.kazi@tfzr.rs (Z.K.); filipsnezana@gmail.com (S.F.)

* Correspondence: ljubica.kazi@gmail.com

Abstract: Air pollution is one of the most challenging and complex problems of our time. This research presents the prediction of air pollutant values based on using an R program with linear regression. The research sample consists of obtained values of air pollutants such as sulphur dioxide (SO₂), particulate matter (PM₁₀, PM_{2.5}), carbon monoxide (CO), nitrite oxides (NO, NO₂, and NO_x), atmospheric data pressure (p), temperature (T), and relative humidity (rh). The research data were collected from the city of Belgrade air quality monitoring reports, published by the Environmental Protection Agency of the Republic of Serbia. The report data were transformed into a form suitable for processing by the R program and used to derive prediction functions based on linear regression upon pairs of air pollutants. In this paper, we describe the R program that was created to enable the correlation of air pollutants with linear regression, which results in functions that are used for the prediction of pollutant values. The correlation of pollutants is presented graphically with diagrams created within the R GUI environment. The predicted data were categorized according to air pollution standard ranges. It has been shown that the derived functions from linear regression enable predictions that are well correlated with the data obtained by automatic acquisition from air quality monitoring stations. The R program was created by using R language statements without any additional packages, and, therefore, it is suitable for multiple uses in a diversity of application domains with minor adjustments to appropriate data sets.

Keywords: R language; programming; air pollution; prediction model; linear regression



Citation: Kazi, Z.; Filip, S.; Kazi, L. Predicting PM_{2.5}, PM₁₀, SO₂, NO₂, NO and CO Air Pollutant Values with Linear Regression in R Language. *Appl. Sci.* **2023**, *13*, 3617. <https://doi.org/10.3390/app13063617>

Academic Editors: Somandla Ncube and Precious Mahlambi

Received: 18 February 2023

Revised: 8 March 2023

Accepted: 10 March 2023

Published: 12 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Emissions of dangerous gases into the atmosphere due to accidents, human activities, natural disasters, or other reasons are great threats to the human population, nature, and infrastructure. Air pollution is one of the significant environmental problems that can cause adverse health effects, such as asthma, allergies, infections [1], cancer [2], and the risk of low birth weight [3]. These health-related issues are correlated with air pollution, particularly in traffic [4,5]. Therefore, air pollutants data acquisition, measurements, monitoring, evaluation model formulation [6], assessment [7], benchmarking [8], and forecasting [7,9] became increasingly important, particularly in the circumstances of pandemics [10]. Many efforts have been made in the development of appropriate air pollution-related methods and tools, as well as their integration [11].

Data science software tools and modern programming languages enable working with large amounts of data related to environmental problems by utilizing specific functions and commands for data analysis, advanced reasoning, and visualization [12], with languages such as Python [13] and R [14]. R was developed as a computational environment to enable statistical data analysis by Ross Ihaka and Robert Gentleman from the University of Auckland, New Zealand [15]. R is a free and open-source programming language and software environment. The development and maintenance are assigned to the R Core Team, i.e., the R Foundation for Statistical Computing [16]. One of the most important features of

R is its flexibility and scalability, i.e., the possibility to add a set of new functionalities to the base system. This is supported by tools for the development of additional packages [17], so, in this way, R could be used in a diverse set of applications, such as ecology [18].

The air pollution data analysis and the use of R in ecology were in focus in multiple previous research results [18–25]. However, the literature review of related work shows the research gap, which provides the basis for this research. Other related work has focused on particular air pollutants and their predictions [9,24,26–34], while this paper provides predictions for a variety of air pollutants. Other papers have used linear regression in air pollution prediction but not with R language [28,32–39]. In previously published papers, R was used for air pollution data analysis, but with other methods not with linear regression [23–25,40].

The aim of this research is related to using R for air pollution data correlation and prediction. R language statements are used for creating the program that performs data pre-processing with linear regression and establishes the mathematical model for the relationship between two variables. In this research, the dependent and independent variables are selected among air pollutant values for PM_{2.5}, PM₁₀, NO, NO₂, NO_x, CO, and SO₂, as well as meteorological parameters—pressure, temperature, and relative humidity. The obtained mathematical models are used for air pollution data prediction. Results of this study include: (1) a detailed presentation of the developed R program for linear regression and prediction of air pollution data; (2) results of correlation of particular air pollutants, particularly presented with linear regression diagrams; (3) results of linear fitting, i.e., statistical evaluation of linear functions preciseness; and (4) prediction results based on previously obtained and evaluated mathematical functions that correlate air pollutant values.

2. Related Work

The linear regression method has been applied to air pollution prediction within multiple research results. Syafei et al. present an application of a linear regression model regarding air pollutants (nitrogen dioxide, NO₂, particulate matter, PM₁₀, and ozone, O₃), as well as data related to meteorological and temporal factors [26]. These data were used to establish correlation between independent and dependent variables via the use of independent component analysis and principal component analysis. Input data to formulate the linear correlation function were obtained from monitoring stations in Indonesia for the period March–April 2002. It has been concluded that monitoring stations, used for data acquisition, provided data that resulted in different predictions due to meteorological factors (particularly wind) and pollutants interactions.

Before creating a prediction model, it is necessary to select predictors, i.e., air and other components and factors, to be used in the formation of linear regression functions, which, in turn, are used in prediction. In order to provide better air quality assessments, Olvera-García et al. propose a new air quality evaluation model where environmental parameters (PM_{2.5}, PM₁₀, O₃, CO, NO₂, SO₂) were evaluated with the application of fuzzy reasoning to compute and assign individual weights according to the pollutant importance on the air evaluation [6]. This model considers five air pollution score stages: excellent, good, regular, bad, and dangerous, based on data from the Mexico City Atmospheric Monitoring System. With the pollutant weights in place, a better evaluation model is proposed for the air quality assessments. Sethi and Mittal present the application of feature selection methods (based on Least Absolute Selection and Shrinkage Operator with the use of various machine learning techniques) in order to determine potentially significant predictors [27]. Conclusions were drawn based on exploratory data analysis in Delhi, India, and surrounding cities. It has been concluded that carbon monoxide, sulphur dioxide, nitrogen dioxide, and ozone are the most important factors affecting the air quality index.

Diverse solutions enable data acquisition and collection in research related to air pollution prediction via the Internet of Things (IoT), cloud computing, big data, and Geographic Information Systems (GIS). Iskandaryan et al. present smart city data collection through

the utilization of IoT system sensors, where air quality prediction is conducted with the use of machine learning technologies [41]. The Google Earth Engine (GEE) cloud computing platform was used to obtain CO, NO₂, SO₂, and aerosol optical depth (AOD) data in Shandong Province, China, during the 2018–2020 period [10]. Zheng et al. describe the air quality forecasting system based on monitoring stations located at large distances [28]. They collect a large quantity of meteorological and air quality data, and Microsoft's cloud platform Azure is used to integrate the obtained data. Geographic information systems were used as data sources in air quality prediction within the research of Briggs et al. [42] and Hochadel et al. [43].

The development of methods and tools for air pollution prediction has emerged as a significant domain in recent years, with research and professional results related to applied artificial intelligence, data mining, fuzzy systems, machine learning, and neural networks. Siwek and Osowski analyzed methods of data mining for the prediction of air pollution [29]. In their research, they selected data from atmospheric pollutants PM₁₀, SO₂, NO₂, and O₃ and performed prediction by integrating different methods, such as a genetic algorithm, the linear method of stepwise fit, decision trees, and neural networks. Rajat et al. propose a system for forecasting the air pollution index by utilizing a supervised machine learning approach [9]. This approach is applied to examine possibilities of having the best forecasting precision by contrasting four different supervised machine learning algorithms (decision tree, random forest tree, Naïve Bayes theorem, and K-nearest neighbor) for prediction calculations. Data sets from all Indian states were collected and used as the research sample. The trained extreme learning machine has been applied to eight air quality parameters, obtained from two Hong Kong monitoring stations over a six-year period [30]. This approach enabled the prediction of the concentration of air pollutants at an extremely fast learning speed and resulted in an appropriate level of prediction accuracy. The work of Ibarra-Berastegi et al. implements air pollution prediction based on the creation and application of neural networks [31]. The obtained data for model creation have been collected on an hourly basis for five air pollutants (SO₂, CO, NO₂, NO, and O₃) at six locations in the area of Bilbao, Spain, during the year 2000, and the prediction models were tested upon the data from 2001. The created prediction models show diversity in accuracy when comparing different pollutants and measuring sensors. Zhou et al. present a novel spatiotemporal interpolation model that combines data fusion techniques with a Long Short-Term Memory (LSTM) recurrent neural network (RNN) in order to achieve high estimation accuracy over a long time period [44]. Data fusion is performed upon the meteorological data, elevation data, land-use data, and daily PM_{2.5} data collected from China in 2016, to be used within four experiments to evaluate the efficiency and effectiveness of the proposed approach.

The linear regression method is implemented within the applied artificial intelligence methods and tools. Zheng et al. present the system that enables weather and air quality forecasts, based on the use of linear regression and neural networks [28]. Mani et al. present results in the forecasting of the Air Quality Index (AQI) by using machine learning techniques: linear regression and time series analysis with supervised machine learning [32]. Sensor output was related to NO₂, ozone (O₃), PM_{2.5}, and SO₂ data concentrations, and they were used to train a regression model. The obtained prediction model was validated with new sensor output data, and the performance has been analyzed. The Auto Regressive Integrated Moving Average (ARIMA) time series model is applied to forecast the AQI. Alsoltany and Alnaqash present the results of an application of the fuzzy linear regression method, with data collected daily from three air quality monitoring stations in Baghdad City [33]. Roy et al. propose a combination of linear regression and genetic algorithm methods for application, as well as multivariate polynomial regression [34]. Linear regression is used for the prediction of gas concentration, while a genetic algorithm approach is used to optimize results, i.e., to minimize errors in linear regression prediction. Predictive equations are formed for CO, O₃, and NO₂ based on data values for temperature, relative humidity, benzene, and NO_x.

Some research results are related to the application of the linear regression method to the prediction of particular air pollutant concentrations. A multivariate linear regression model was proposed to enable short-term period prediction of PM_{2.5} based on data on aerosol optical depth (AOD) obtained through remote sensing, meteorological factors from ground monitoring (wind velocity, temperature, and relative humidity), and other gaseous pollutants (SO₂, NO₂, CO, and O₃) [32]. The validity of the derived regression models has been measured, and it has been shown that annual data predictions have a lower fit compared to seasonal predictions (spring and winter data), which are more accurate. Choi and Choi presented results in creating multiple statistical models for prediction of PM₁₀, PM_{2.5}, and PM₁ based on local meteorological parameters (air temperature, wind speed, and relative humidity), PM₁₀ and PM_{2.5} concentrations, and dust periods [33]. This statistical modeling has been created based on the data from Beijing, China, and applied to Gangneung, Korea. The correlation among PM₁₀, PM_{2.5}, and PM₁ concentrations is represented as multiple correlation coefficients, and the prediction of PM concentration has shown a significant level of multi-regression significance when the observed and calculated PM concentrations were compared. The prediction of the next day's hourly ozone concentration was studied in [37]. In this research, feedforward artificial neural networks are proposed to use principal components as inputs. The developed model was compared with multiple linear regression, feedforward artificial neural networks based on the original data, and also with principal component regression. It has been shown that the proposed use of principal components reduced complexity in the application of compared methods and eliminated data collinearity. Land-use regression (LUR) models are used to estimate air pollution in epidemiologic research, and they use data obtained from a small set of locations, so geographical location could not impact the prediction results. Basagaña et al. presented one of such studies that use LUR, where the health effects of air pollution were examined [38]. Land-use regression has been used at the regional level (national level in the USA) regarding NO₂ measurements and predictions performed with satellite data [39].

The basic R system is constantly enhanced by several hundreds of contributed packages that cover a wide array of modern statistical methods and application areas [14]. With millions of lines of R code available in repositories stored on the Internet, researchers and programmers have an opportunity to use a combination of static and dynamic program features for various analyses [45]. In recent years, the R language has been frequently used in scientific research, especially in ecology and environmental protection. R was used for weather monitoring and rain gauge observation [19]. The result was a free and open-source R program with the purpose of computing merged data from radar and rain gauges. Another example of using R is described in the work of Kembel et al., who introduced Picante, a package created to extend R with tools for analyzing phylogenetic and trait diversity of ecological communities, calculating phylogenetic diversity metrics, performing trait comparative analyses, etc. [20]. Stanke et al. present an additional R package, rFIA, created for the purpose of forest data analysis with the use of the FIA database, which was created to support monitoring of changes in forests across the USA [21]. R has been used for creating an image analysis framework designed to detect land cover types and vegetation corresponding to the spectral reflectance of the objects represented on the Earth's surface [22]. The R model is used with images obtained from sensor scenes (Landsat-8 Operational Land Imager and Thermal Infrared Sensor). The image data were processed by using different auxiliary packages of R. Results of using the R-based image analysis framework were created based on time series of the images taken at various periods to monitor the landscape dynamics in the Congo River basin.

Seo et al. present results of using R as a statistical software to analyze a large amount of air pollution data in Korea [23]. Setiawan presents results that are related to utilizing R and R Studio for the prediction of nitrogen dioxide pollution, with particular emphasis on the comparison of the autoregressive integrated moving average and the exponential smoothing model [24]. Carslaw and Ropkins describe openair as an R package developed

for the analysis of air pollution measurement data, but it could also be used in broader atmospheric sciences [25]. The authors present the development of open-air additional features such as conditioning plots and inference possibilities, which enable better results in air pollution data analysis. The use of this package is illustrated with data obtained from UK air pollution monitoring networks. Selvi and Chandrasekaran present data mining with air pollution data by utilizing the open-air and ropenaq packages of R [40]. Derived patterns are used to enable the creation of predictive models for environmental issues.

3. Materials and Methods

Scientific data are often stored in formats not suitable for analysis and processing. Making applications that work with the diversity of data sources and growing databases has become an emerging topic since rapid data availability and processing could be a limiting factor for end-users [8]. The methodology and procedures used in this research include air pollution data collection, pre-processing with the use of the linear regression method, and processing in the prediction. Results of air pollutant values in this research are presented with the use of measurement units as mass units in the volume unit, taking them as related to time. These measurement units could be used over a long period of time for measuring concentrations of gases at shorter intervals [12]. Figure 1 presents a flowchart that visualizes the proposed method of using the R program within the air pollution prediction system. Data collection used in this research consists of air pollution data obtained from the official web site of the Environmental Protection Agency, affiliated with the Ministry of Environmental Protection, Belgrade, Republic of Serbia. Raw measurement data are presented at this website, since 2008, each hour of every day, obtained from automatic data acquisition stations that are located at different places, especially in cities (at crowded streets, industrial zones), but also at protected natural regions in the Republic of Serbia. Measurement data are presented comparatively for multiple measurement stations for each hour [46] or with a detailed data view for every measurement station, each day, and every hour [47].

The sample data for this research was downloaded from [47]. It consists of data related to the January-March period in 2021 and 2022 and particularly selected for Serbia's capital city, Belgrade. The observed period was from January to March in both years, since it was the winter period, when pollution is expected to be higher. Selecting a seasonal sample (winter) is aligned with the results of Zhao et al., where better accuracy of a mathematical model was created with linear regressions from seasonal data compared to annual data samples [32]. The City of Belgrade was selected as a very crowded urban area with large industrial and residential parts and heavy traffic. The sample consists of 378 measured air pollutant concentrations in 2021 and 567 measurements in 2022 for the following components: sulphur dioxide (SO_2), particulate matter (PM10, PM2.5), carbon monoxide (CO), and nitrite oxides (NO, NO_2 , and NO_x). All sample air pollutant values were obtained from the data source as raw data represented with the use of the measurement unit $\mu\text{g}/\text{m}^3$, while CO values were represented with the mg/m^3 measurement unit. In addition to these measurements, the sample data also contain the following meteorological conditions: air temperature (T, Celsius scale), relative humidity (rh, percentage), and atmospheric pressure (p, millibar measures).

The obtained sample data from the web site as a source was transformed into a tab-delimited text file, suitable for loading into the R GUI (Figure 2), with a data structure consisting of city/time/ SO_2 / NO_2 / NO_x /CO/NO/PM2.5/PM10/p/T/rh. The T symbol is regularly used for temperature, while in this research, the sample presents temperature with the symbol t. Data vectors, created from the loaded text file, are used to obtain a linear regression function that establishes the correlation of measured values of two air pollutants.

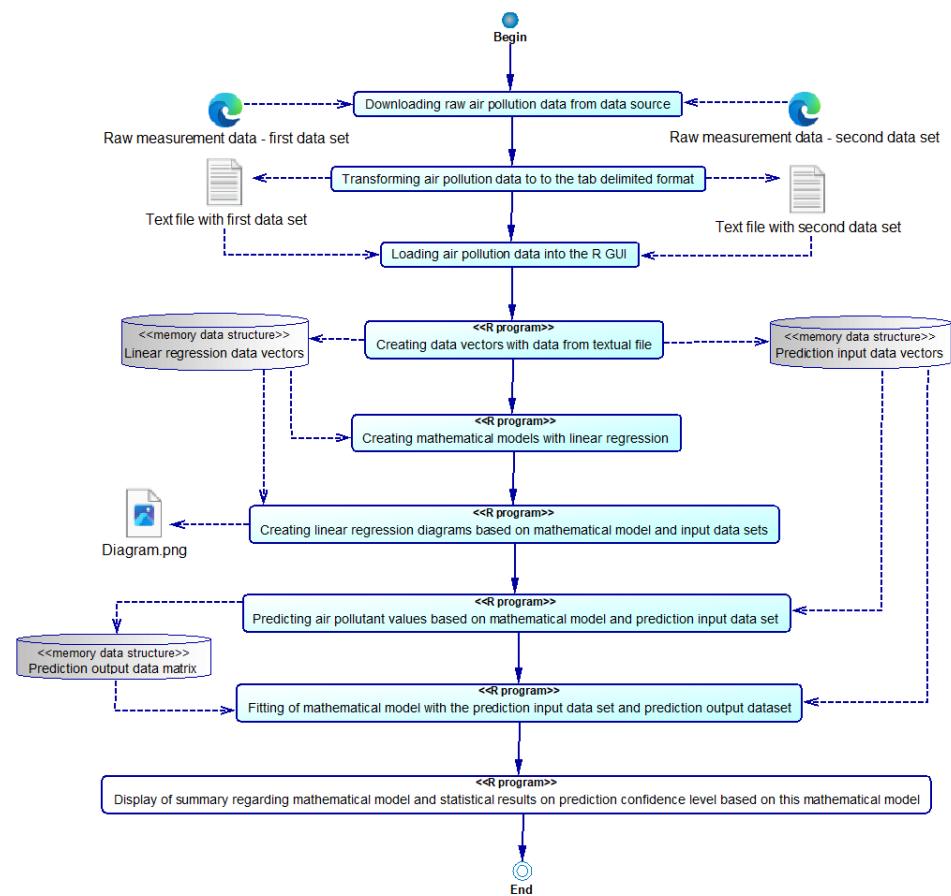


Figure 1. Flowchart presenting the proposed method in an air pollution prediction system, based on the developed R program.

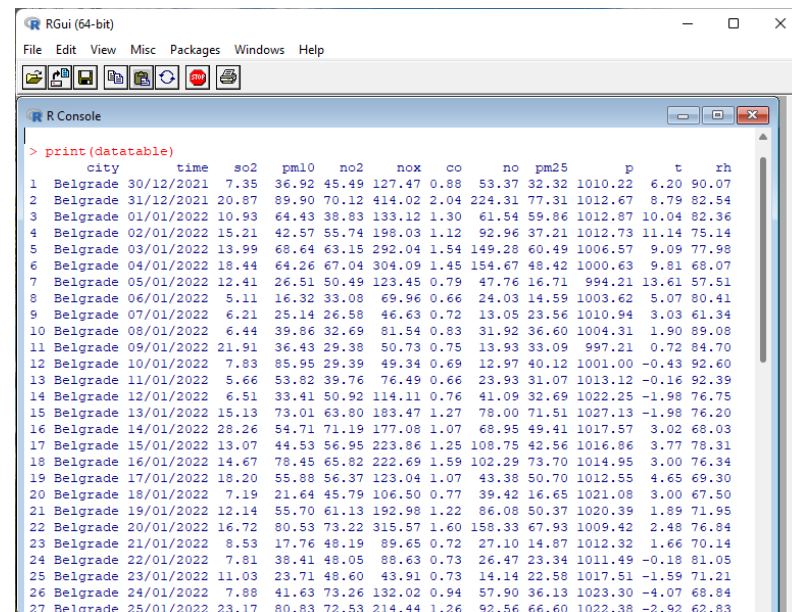
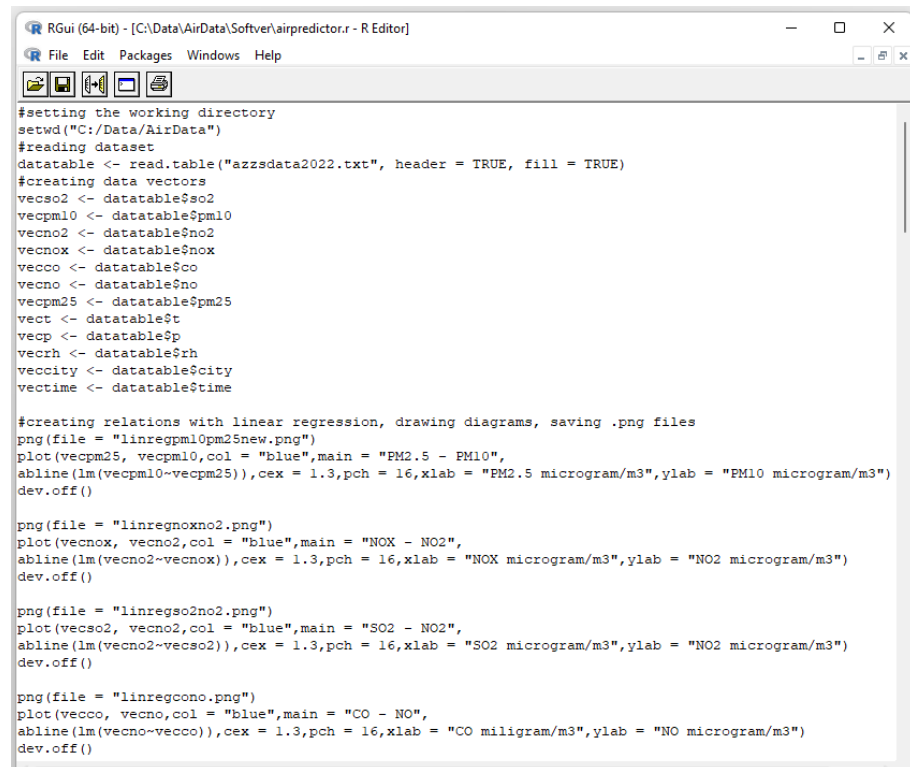


Figure 2. Air pollutants and meteorological measured values loaded into the R GUI.

The created R program enables the creation of linear regression functions (on all pairs of air pollutants and meteorological parameter values), graphical representation of data correlation in the form of colored correlation heat maps and linear regression diagrams, as well as the use of linear regression functions and R language functions for the prediction of

air pollution data. Finally, results of data processing (i.e., air pollutant value prediction) are organized according to categories of air pollutant concentrations and air quality: excellent, good, acceptable, polluted, very polluted, as defined by the Environmental Protection Agency [48].

The program used in this research was written using the R programming language within the R GUI (graphical user interface) editor, created by R Team (The R Foundation for Statistical Computing c/o Institute for Statistics and Mathematics, Vienna, Austria) [16]. Figures 3 and 4 present the first and second parts of the created R program within the R GUI development environment, respectively.



```
#setting the working directory
setwd("C:/Data/AirData")
#reading dataset
datatable <- read.table("azzsdata2022.txt", header = TRUE, fill = TRUE)
#creating data vectors
vecso2 <- datatable$so2
vecpm10 <- datatable$pm10
vecno2 <- datatable$no2
vecnox <- datatable$nox
vecco <- datatable$co
vecno <- datatable$no
vecpm25 <- datatable$pm25
vect <- datatable$t
vecp <- datatable$p
vecrh <- datatable$rh
veccity <- datatable$city
vectime <- datatable$time

#creating relations with linear regression, drawing diagrams, saving .png files
png(file = "linregpm10pm25new.png")
plot(vecpm25, vecpm10,col = "blue",main = "PM2.5 - PM10",
abline(lm(vecpm10~vecpm25)),cex = 1.3,pch = 16,xlab = "PM2.5 microgram/m3",ylab = "PM10 microgram/m3")
dev.off()

png(file = "linregnoxno2.png")
plot(vecnox, vecno2,col = "blue",main = "NOX - NO2",
abline(lm(vecno2~vecnox)),cex = 1.3,pch = 16,xlab = "NOX microgram/m3",ylab = "NO2 microgram/m3")
dev.off()

png(file = "linregso2no2.png")
plot(vecso2, vecno2,col = "blue",main = "SO2 - NO2",
abline(lm(vecno2~vecso2)),cex = 1.3,pch = 16,xlab = "SO2 microgram/m3",ylab = "NO2 microgram/m3")
dev.off()

png(file = "linregcono.png")
plot(vecco, vecno,col = "blue",main = "CO - NO",
abline(lm(vecno~vecco)),cex = 1.3,pch = 16,xlab = "CO miligram/m3",ylab = "NO microgram/m3")
dev.off()
```

Figure 3. Created the first part of the R program with creating data vectors, relations with linear regression, and drawing diagrams in the R GUI.

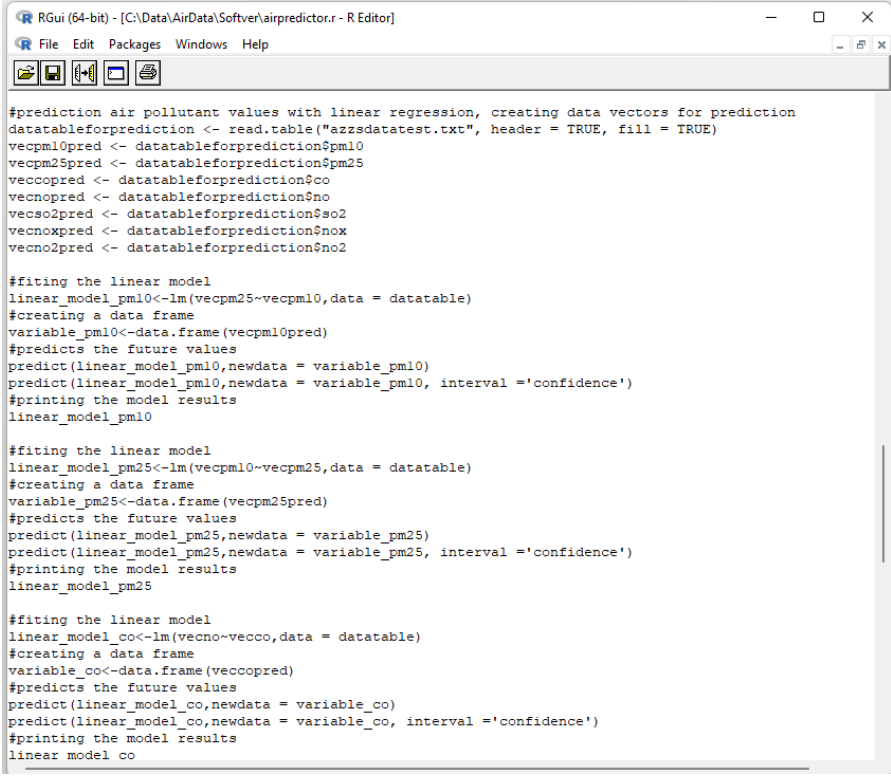
The first command in the newly created R program is used for setting the working directory in R with the `setwd(path)` procedure. The second command is reading data from an external tab-delimited .txt file that includes the stored data about air pollutants and meteorological conditions measured values with the `read.table()` function. It creates a table data structure with a title row that contains names for each column. The third step is creating vectors for measured pollutant values from the previously created table with the "<->" operator for assigning values to R structures:

vector <- datatable\$columnname

In this program, we created vectors for all pollutants and measured temperature, pressure, and relative humidity.

Establishing relations with linear regression between pollutants was conducted with the "`lm()`" function in R (Figure 3):

`lm (vectorpollutant2~vectorpollutant1)`, where `vectorpollutant1` is x and `vectorpollutant2` is y in the linear function $y = ax + b$.



```

RGui (64-bit) - [C:\Data\AirData\Software\airpredictor.r - R Editor]
File Edit Packages Windows Help

#prediction air pollutant values with linear regression, creating data vectors for prediction
datatableforprediction <- read.table("azzsdatatest.txt", header = TRUE, fill = TRUE)
vecpm10pred <- datatableforprediction$pm10
vecpm25pred <- datatableforprediction$pm25
veccopred <- datatableforprediction$co
vecnopred <- datatableforprediction$no
vecso2pred <- datatableforprediction$so2
vecnoxpred <- datatableforprediction$nox
vecno2pred <- datatableforprediction$no2

#fitting the linear model
linear_model_pm10<-lm(vecpm25~vecpm10,data = datatable)
#creating a data frame
variable_pm10<-data.frame(vecpm10pred)
#predicts the future values
predict(linear_model_pm10,newdata = variable_pm10)
predict(linear_model_pm10,newdata = variable_pm10, interval ='confidence')
#printing the model results
linear_model_pm10

#fitting the linear model
linear_model_pm25<-lm(vecpm10~vecpm25,data = datatable)
#creating a data frame
variable_pm25<-data.frame(vecpm25pred)
#predicts the future values
predict(linear_model_pm25,newdata = variable_pm25)
predict(linear_model_pm25,newdata = variable_pm25, interval ='confidence')
#printing the model results
linear_model_pm25

#fitting the linear model
linear_model_co<-lm(vecno~vecco,data = datatable)
#creating a data frame
variable_co<-data.frame(veccopred)
#predicts the future values
predict(linear_model_co,newdata = variable_co)
predict(linear_model_co,newdata = variable_co, interval ='confidence')
#printing the model results
linear_model_co

```

Figure 4. Created the second part of the R program, which deals with predicting air pollutants values, fitting the linear model, and printing the results.

Drawing diagrams for the obtained mathematical models was completed with three commands (Figure 3). First, we defined a .png image with a name:

```
png (file = "imagefilename.png")
```

After this command, the plot function creates the diagram based on the results of the lm function, while the dev.off function saves the diagram in a file named with the png command and stored at the location defined by the setwd command.

Predicting air pollutant values is possible with a predictor vector, response vector, and linear regression function (Figure 4). Commands for this purpose are listed below:

```

linear_model <- lm (vectorpollutantA~vectorpollutantB,data = datatable)
variable_vectorpollutantB<- data.frame(vectorpollutantBprediction)
predict (linear_model,newdata = variable_vectorpollutantB)predict (linear_model,
newdata = variable_vectorpollutantB, interval = 'confidence')

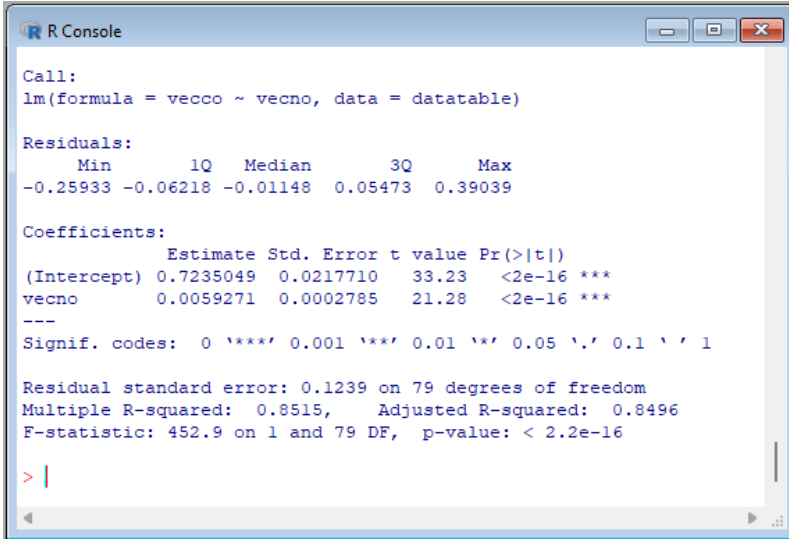
```

A linear model is created by applying the lm function to two datasets (data vectors), where the first parameter will have the y role and the second parameter data vector has the x role in the resulting linear function. The linear model is a mathematical function representing the data vectors correlation. The predict() function from the R language was used to predict the future values based on the previously created linear model and prediction input data vector. For the purpose of prediction, another data vector was prepared as an input data set of air pollutant values or meteorological data. Upon these data, as well as with the linear model, the predict() function of the R language was used to derive the predicted values. The next step was to execute the predict() function again, this time with the third parameter of the function call being used for checking the "confidence" level in predicted values, i.e., to enable the calculations of prediction accuracy, i.e., the preciseness of the mathematical model obtained with linear regression (Figure 4). Finally, the summary command in R provides statistics related to the obtained linear model as well as statistics related to prediction fitting (i.e., a statistical computation of the accuracy of the computed mathematical model that was used for prediction).

4. Results and Discussion

Comparing to previously published works, where R was used for other ecology-related research, air pollution data was processed with special R packages, and linear regression was used for other purposes, and the diversity of software tools and technologies, this work contributes with the detailed presentation of a specially created R program by using R Language statements, and this program enables data correlation with linear regression, fitting of the derived mathematical functions, and prediction of air pollution data.

Results of executing the R program in processing linear regression and prediction statistics are presented with one example for CO (dependent) and NO (independent) data vectors at Figures 5 and 6.



```

R Console

Call:
lm(formula = vecco ~ vecno, data = datatable)

Residuals:
    Min       1Q   Median       3Q      Max
-0.25933 -0.06218 -0.01148  0.05473  0.39039

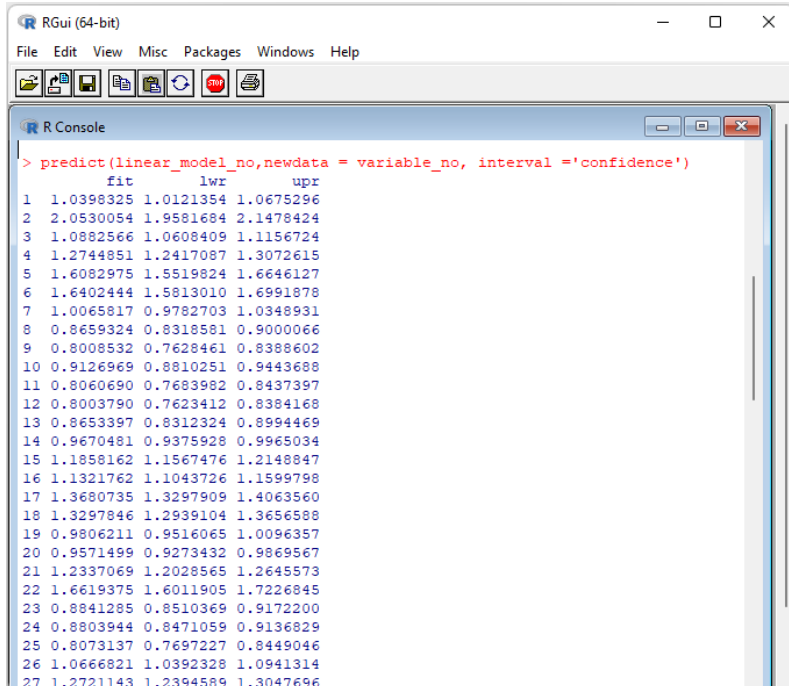
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.7235049   0.0217710   33.23  <2e-16 ***
vecno       0.0059271   0.0002785    21.28  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1239 on 79 degrees of freedom
Multiple R-squared:  0.8515,    Adjusted R-squared:  0.8496
F-statistic: 452.9 on 1 and 79 DF,  p-value: < 2.2e-16

> |

```

Figure 5. Example of R program execution for linear regression upon CO and NO pairs or vectors and the resulting statistics regarding mathematical model accuracy.



```

RGui (64-bit)
File Edit View Misc Packages Windows Help

> predict(linear_model_no, newdata = variable_no, interval = 'confidence')
      fit      lwr      upr
1 1.0398325 1.0121354 1.0675296
2 2.0530054 1.9581684 2.1478424
3 1.0882566 1.0608409 1.1156724
4 1.2744851 1.2417087 1.3072615
5 1.6082975 1.5519824 1.6646127
6 1.6402444 1.5813010 1.6991878
7 1.0065817 0.9782703 1.0348931
8 0.8659324 0.8318581 0.9000066
9 0.8008532 0.7628461 0.8388602
10 0.9126969 0.8810251 0.9443688
11 0.8060690 0.7683982 0.8437397
12 0.8003790 0.7623412 0.8384168
13 0.8653397 0.8312324 0.8994469
14 0.9670481 0.9375928 0.9965034
15 1.1858162 1.1567476 1.2148847
16 1.1321762 1.1043726 1.1599798
17 1.3680735 1.3297909 1.4063560
18 1.3297846 1.2939104 1.3656588
19 0.9806211 0.9516065 1.0096357
20 0.9571499 0.9273432 0.9869567
21 1.2337069 1.2028565 1.2645573
22 1.6619375 1.6011905 1.7226845
23 0.8841285 0.8510369 0.9172200
24 0.8803944 0.8471059 0.9136829
25 0.8073137 0.7697227 0.8449046
26 1.0666821 1.0392328 1.0941314
27 1.2721143 1.2394589 1.3047696

```

Figure 6. Example of results from executing the predict function for CO data vectors based on correlation with NO (within the execution of the created R program).

From Figure 5, it could be concluded that the r (correlation coefficient) for CO values being predicted from NO values has a value of 0.8515, which could be categorized as a high level of correlation.

Results of executing the created R program (that includes utilization of the predict() R function) within the R GUI are shown in Figure 6. This is an example of successfully predicting future values of a CO pollutant based on the previously generated linear model of correlated CO and NO. Figure 6 presents three columns of data generated by the predict function for each item of the input data vector; there are computed values of prediction fit, but also lwr (lower) and upr (upper) values for each fit value. This way, it is obvious that the prediction function does not provide only one value, but an interval of values that could be expected in the future, using the underlying linear function as a basis.

For this particular case of correlation between CO and NO values, Figure 7 presents a diagram that enables comparison between the computed prediction data and the real measurement data for CO as being dependent on the NO values data set in the previously presented prediction function.

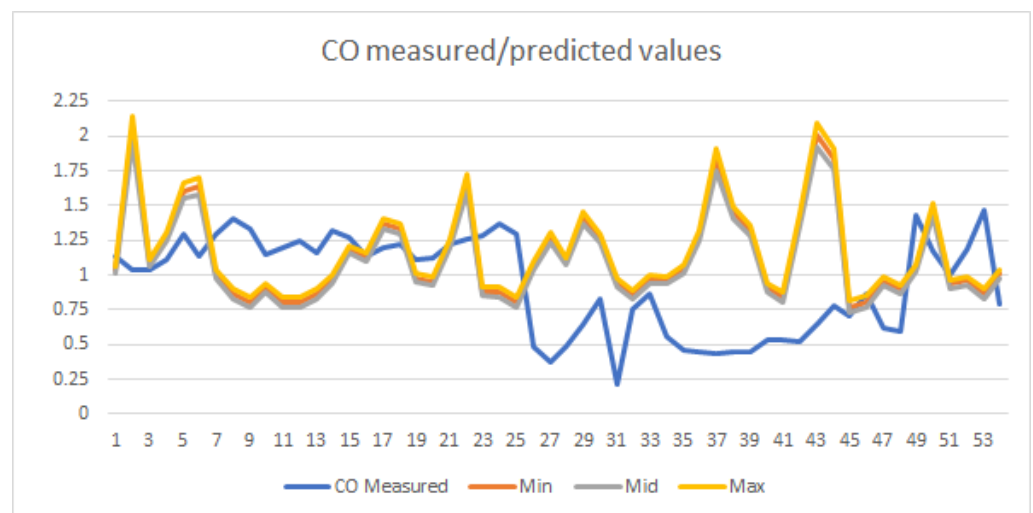


Figure 7. Diagram of comparatively presented real measurement data of CO and predicted values (range) of CO, based on linear correlation with NO values.

Figure 7 contains graphs representing predicted data for CO for each particular measurement of NO as a lower (min), fit (mid), and upper (max) predicted value, as well as a graph presenting CO measured values (at the same time and at the same monitoring station as the NO value that was used for prediction). Since the r correlation coefficient between NO and CO is 0.8515, it is obvious from Figure 7 that the measurements and prediction graphs for CO are not very closely aligned.

The correlation heat map, presented at Figure 8, shows values of r correlation coefficients computed upon pairs of data vectors for all obtained parameters, including air pollutants and meteorological parameters.

r	y	SO ₂	NO	NO ₂	NO _x	CO	PM2.5	PM10	p	T	rh	Colors:
x												
SO ₂	X		0.3159	0.4468	0.3654	0.3349	0.3626	0.3497	0.0017	0.0162	0.0140	0.0 – 0.1
NO	0.3159	X		0.6821	0.9809	0.8515	0.5053	0.4501	0.0004	0.0599	0.0001	0.1 – 0.2
NO ₂	0.4468	0.6821	X		0.7558	0.6305	0.4006	0.4114	0.0049	0.0284	0.0426	0.2 – 0.3
NO _x	0.3654	0.9809	0.7558	X		0.8667	0.5204	0.4725	0.0004	0.0601	0.0006	0.3 – 0.4
CO	0.3349	0.8515	0.6305	0.8667	X		0.6269	0.5373	0.0140	0.0651	0.0197	0.4 – 0.5
PM2.5	0.3626	0.5053	0.4006	0.5204	0.6269	X		0.7895	0.0506	0.0236	0.0196	0.5 – 0.6
PM10	0.3497	0.4501	0.4114	0.4725	0.5373	0.7895	X		0.0226	0.0001	0.0001	0.6 – 0.7
p	0.0017	0.0004	0.0049	0.0004	0.0140	0.0506	0.0226	X		0.1217	0.0645	0.7 – 0.8
T	0.0162	0.0599	0.0284	0.0601	0.0651	0.0236	0.0001	0.1217	X		0.1218	0.8 – 0.9
rh	0.0140	0.0001	0.0426	0.0006	0.0197	0.0196	0.0001	0.0645	0.1218	X		0.9 – 1.0

Figure 8. Correlation heat map with all obtained air pollution parameters from the sample.

According to the correlation heat map, the high correlation (r values from 0.9 to 0.7) was computed with pairs NO-NO_x ($r = 0.9809$), CO-NO_x ($r = 0.8667$), NO-CO ($r = 0.8515$), PM2.5-PM10 ($r = 0.7895$), NO₂-NO_x ($r = 0.7558$), while for moderate correlation (r values from 0.7 to 0.5), there are also six air pollutant pairs. It could be concluded that the computed correlation between meteorological parameters and air pollutants is very low, with an r value less than 0.2. Therefore, in the rest of this paper, these pairs of parameters will not be presented with linear regression diagrams, predictions, and detailed statistics. The linear equations (as results of the linear regression method) are presented for selected (from the high, moderate, and low correlation categories) pairs of air pollutants as bivariate graphical plots (Figure 9a–f) that describe their mutual dependence, i.e., inter-variable correlations. Dots represent data from vectors, while lines, plotted between dots are the graphical representation of a derived linear equation from the data in vectors. A detailed statistical analysis of the correlation between the model and the data, apart from the graphical representation (quantity of dots placed near the line), can also be conducted on the basis of the statistical data presented in Tables 1–6.

Table 1. Summary statistics from linear regression model for PM2.5 and PM10.

$Y = 3.413 + 0.698x$					$r = 0.7895$
PM2.5-PM10 (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	0.0405	17.214	$<2 \times 10^{-16} ***$	7.945	296.3
Signif. codes: 0 ***.					

Table 2. Summary statistics from linear regression model for CO and NO.

$Y = 143.66x - 94.94$					$r = 0.8515$
CO-NO (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	6.751	21.28	$<2 \times 10^{-16} ***$	19.3	452.9
Signif. codes: 0 ***.					

Table 3. Summary statistics from linear regression model for SO₂ and NO₂.

$Y = 0.229x - 0.328$					$r = 0.4468$
SO ₂ -NO ₂ (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	0.02873	7.988	$9.18 \times 10^{-12} ***$	4.084	63.82
Signif. codes: 0 ***.					

Table 4. Summary statistics from linear regression model for CO and NO₂.

$Y = 8.428 + x39.489$					$r = 0.6305$
CO-NO ₂ (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	3.401	11.609	$<2 \times 10^{-16} ***$	9.722	134.8
Signif. codes: 0 ***.					

Table 5. Summary statistics from linear regression model for NO and NO_x.

$Y = 36.108 + 1.755x$					$r = 0.9809$
NO-NO _x (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	0.02757	63.64	$<2 \times 10^{-16} ***$	12.27	4050
Signif. codes: 0 ***.					

Table 6. Summary statistics from linear regression model for NO₂ and NO_x.

$Y = 4.822x - 104.375$					$r = 0.7558$
NO _x -NO ₂ (x-Y)	Std. error	t-value	p-value	Residual Std. error	F-statistic
	0.3084	15.637	$<2 \times 10^{-16} ***$	43.83	244.5
Signif. codes: 0 ***.					

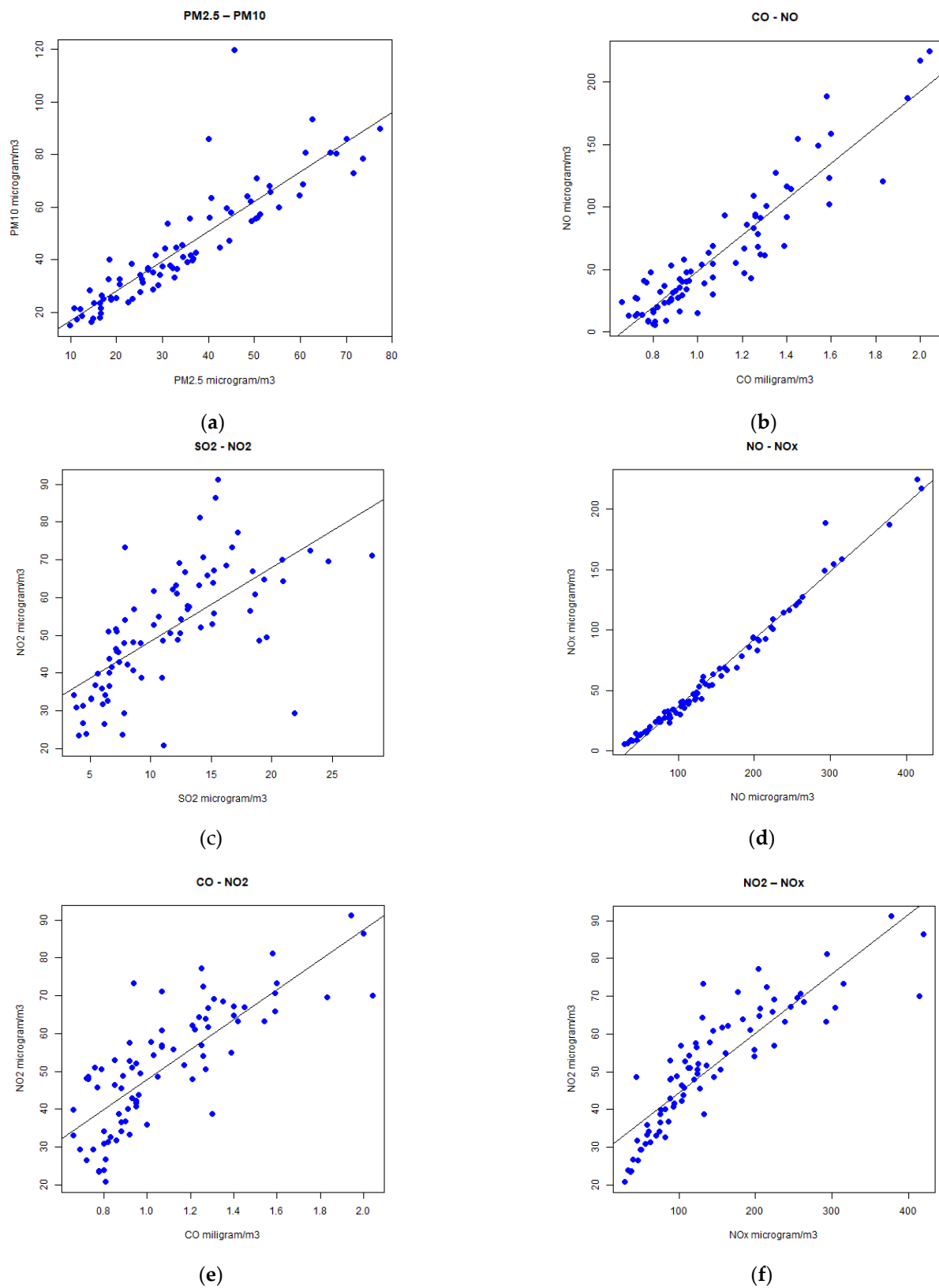


Figure 9. Linear regression diagrams created within R GUI with R program (a) PM10-PM2.5; (b) CO-NO; (c) SO₂-NO₂; (d) NO-NO_x; (e) CO-NO₂; (f) NO₂-NO_x.

Standard statistical parameters such as Fisher's criterion (F), correlation coefficient (r), estimate and residual standard error, *p*-value, and *t*-value were used for model validation. According to statistical parameters, it can be seen from Tables 1–6 that we have a moderate, high, and very high positive correlation with the model. The best correlation and fit to the data is with NO-NO_x pollutants (*r* = 0.9809), which had a high F-value and low *p*-values and standard errors. The correlation between CO and NO pollutants was strong (*r* = 0.8515), which can be seen by statistical parameters and a graph. The very high correlation was in the case of PM2.5-PM10 (*r* = 0.7895) and NO₂-NO_x (*r* = 0.7558), as were the high F-value and low *p*-value and standard errors. CO-NO₂ air pollution showed a moderate correlation with the model (*r* = 0.6305). The weakest interdependence was shown by SO₂-NO₂ pollutants (*r* = 0.4468), who had the lowest F-value and the greatest dispersion of data.

A good correlation between NO₂, NO, and NO_x (*r*~0.68 to 0.98) can be explained by the chemical similarity of the compounds and the chemical pathway of their formation [49].



According to the correlation heat map (Figure 8), there is a weak correlation (*r* < 0.5) between chemically different oxides of carbon, sulphur, and nitrogen. The specific case is CO, which establishes a high and medium correlation with all air pollutants, except with SO₂ (*r* = 0.3349).

Table 7 shows the summarized and categorized results of predicting air pollutant values for SO₂, PM10, PM2.5, CO, and NO₂. Part of the data that is the source for Table 7 was previously presented in Figure 8. The categorization of program execution results (prediction data), which is presented in Table 7, has been performed according to air pollutant concentration indexes as excellent, good, acceptable, polluted, much polluted, and colored according to criteria defined by the Environmental Protection Agency of the Ministry of Environmental Protection, Republic of Serbia [48], which are aligned with the normative defined in the European Union. These values are, of course, different for each pollutant.

Table 7. Summarized and categorized results of predicting air pollutant values with R.

Air Quality Index	Excellent	Good	Acceptable	Polluted	Very Polluted
PM2.5 concentration intervals number of predicted values (relation with PM10) (µg/m ³)	0–15 3	15.01–30 36	30.01–55 33	55.01–110 9	>110 0
NO ₂ concentration intervals number of predicted values (relation with SO ₂) (µg/m ³)	0–50 46	50.01–100 35	100.01–150 0	150.01–400 0	>400 0
CO concentration intervals number of predicted values (relation with NO) (mg/m ³)	0–5 81	5.01–10 0	10.01–25 0	25.01–5 0	>50 0
SO ₂ concentration intervals number of predicted values (relation with NO ₂) (µg/m ³)	0–50 81	50.01–90 0	90.01–180 0	350.01–500 0	>500 0
PM10 concentration intervals Number of predicted values (relation with PM2.5) (µg/m ³)	0–25 16	25.01–50 37	50.01–90 27	90.01–180 1	>180 0

Note: Categories colors according to [48].

According to the results presented in Table 7, it could be concluded that the great majority of predicted values could be categorized as excellent or acceptable.

5. Conclusions

This paper contributes to the R program that enables data transformation, linear regression, mathematical model fitting, and the processing of prediction functions. This program also enables the graphical presentation of results in the form of diagrams that illustrate the correlation between pollutants. This research also contributes to the evaluation of linear regression functions for their accuracy. It has been shown that the resulting functions enable predictions with high precision; the predicted values correlate very well with the obtained data. The developed R program was created with R language statements without using specific R packages. Therefore, it could be used or easily adapted to diverse application domains. The second contribution is related to the results of data correlation with all types of air pollutants, which are included in most frequent air quality monitoring systems. It has been shown that certain air pollutant pairs of data have a high level of correlation since their linear regression function has a high level of fitting. The limitations of this work are related to several aspects. The R program was not presented entirely in this paper (because of the program's length), but it is available at the public repository GitHub, and only the most important lines of R code are presented and explained. The second limitation is related to the sample characteristics: one city, two years, and only a winter period of three months in both years; not all atmospheric parameters were included (e.g., data from monitoring stations in this sample do not provide wind-related data). Future work could be related to the utilization of the created R program to make predictions based on a wider set of parameters, larger data sets taken over longer periods of time, a diversity of monitoring locations, adaptation to other application domains, and improving the program to use other statistical methods supported by the R language, with a special emphasis on supporting further chemical analysis of complex interactions and processes with gasses in the atmosphere.

Author Contributions: Conceptualization, Z.K. and L.K.; methodology, Z.K. and L.K.; software, Z.K.; validation, Z.K., S.F. and L.K.; formal analysis, Z.K. and S.F.; investigation, Z.K. and L.K.; resources, Z.K. and L.K.; data curation, Z.K. and S.F.; writing—original draft preparation, Z.K., S.F. and L.K.; writing—review and editing, Z.K., L.K. and S.F.; visualization, Z.K.; supervision, Z.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The R GUI tool is open source software that has been used in this research. It is available at <https://www.r-project.org>. The created R program is open-sourced and the code is available at <https://github.com/AirPolWRL/APPWRL> (accessed on 1 September 2022). Data for this research are obtained from <http://www.amskv.sepa.gov.rs/pregledpodatakazbirni.php?lng=en> (accessed on 1 January 2021). The datasets analyzed in this study are available in the repository: <https://github.com/AirPolWRL/APPWRL> (accessed on 1 September 2022).

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

References

1. Brauer, M.; Hoek, G.; Smit, H.A.; De Jongste, J.C.; Gerritsen, J.; Postma, D.S.; Kerkhof, M.; Brunekreef, B. Air pollution and development of asthma, allergy and infections in a birth cohort. *Eur. Respir. J.* **2007**, *5*, 879–888. [CrossRef]
2. Tusnio, N.; Fichna, J.; Nowakowski, P.; Tofilo, P. Air Pollution Associates with Cancer Incidences in Poland. *Appl. Sci.* **2020**, *10*, 7489. [CrossRef]
3. Balogun, H.A.; Rantala, A.K.; Antikainen, H.; Siddika, N.; Amegah, A.K.; Ryti, N.R.I.; Kukkonen, J.; Sofiev, M.; Jaakkola, M.S.; Jaakkola, J.J.K. Effects of Air Pollution on the Risk of Low Birth Weight in a Cold Climate. *Appl. Sci.* **2020**, *10*, 6399. [CrossRef]
4. McConnell, R.; Berhane, K.; Yao, L.; Jerrett, M.; Lurmann, F.; Gilliland, F.; Kunzli, N.; Gauderman, J.; Avol, E.; Thomas, D.; et al. Traffic, susceptibility, and childhood asthma. *Environ. Health Persp.* **2006**, *114*, 766–772. [CrossRef] [PubMed]
5. Morgenstern, V.; Zutaver, A.; Cyrys, J.; Brockow, I.; Koletzko, S.; Kramer, U.; Behrendt, H.; Herbarth, O.; von Berg, A.; Bauer, P.C.; et al. Atopic diseases, allergic sensitization, and exposure to traffic-related air pollution in children. *Am. J. Respir. Crit. Care Med.* **2008**, *177*, 1331–1337. [CrossRef] [PubMed]

6. Olvera-García, M.A.; Carbajal-Hernández, J.J.; Sánchez-Fernández, L.P.; Hernández-Bautista, I. Air quality assessment using a weighted Fuzzy Inference System. *Ecol. Inform.* **2016**, *33*, 57–74. [\[CrossRef\]](#)
7. Morley, D.W.; Gulliver, J. A land use regression variable generation, modelling and prediction tool for air pollution exposure assessment. *Environ. Modell. Softw.* **2018**, *105*, 17–23. [\[CrossRef\]](#)
8. Betancourt, C.; Hagemeyer, B.; Schroder, S.; Schultz, M.G. Context aware benchmarking and tuning of a TByte-scale air quality database and web service. *Earth Sci. Inform.* **2021**, *14*, 1597–1607. [\[CrossRef\]](#)
9. Rajat, R.R.; Vaibhav, D.; Ridam, G.; Rahul, P.; Pratik, G.; Mukul, S.; Ritik, J.; Preetee, K. Prediction of Air Quality Index Using Supervised Machine Learning. *Int. J. Res. Appl. Sci. Eng. Tech.* **2022**, *10*, 1371–1382.
10. Xing, H.; Zhu, L.; Chen, B.; Niu, J.; Li, X.; Feng, Y.; Fang, W. Spatial and temporal changes analysis of air quality before and after the COVID-19 in Shandong Province, China. *Earth Sci. Inform.* **2022**, *15*, 863–876. [\[CrossRef\]](#)
11. Carmichael, G.R.; Sandu, A.; Chai, T.; Daescu, D.N.; Constantinescu, E.M.; Tang, Y. Predicting air quality: Improvements through advanced methods to integrate models and measurements. *J. Comput. Phys.* **2008**, *227*, 3540–3571. [\[CrossRef\]](#)
12. Ilijazi, V.; Jacimovski, S.; Milic, N.; Popovic, B. Software-Supported Visualization of Mathematical Spatial-Time Distribution Models of Air-Pollutant Emissions. *J. Sci. Ind. Res.* **2021**, *80*, 915–923. Available online: <http://op.niscair.res.in/index.php/JSIR/article/view/46963/465479886> (accessed on 30 August 2022).
13. Kadivala, A.; Kumar, A. Applications of Python to evaluate environmental data science problems. *Environ. Prog. Sustain.* **2017**, *16*, 1580–1586. [\[CrossRef\]](#)
14. Dutang, C.; Goulet, V.; Pigeon, M. Actuar: An R package for actuarial science. *J. Stat. Softw.* **2008**, *25*, 1–37.
15. Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **2012**, *5*, 299–314.
16. R Foundation for Statistical Computing. R Core Team. R: A Language and Environment for Statistical Computing. Available online: <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf> (accessed on 7 September 2022).
17. Csárdi, G.; Salmon, M. rhub: Connect to ‘R-hub’. Available online: <https://r-hub.github.io/rhub/authors.html> (accessed on 7 September 2022).
18. Frichot, E.; Francois, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **2015**, *6*, 925–929. [\[CrossRef\]](#)
19. Guenzi, D.; Fratianni, S.; Boraso, R.; Cremonini, R. CondMerg: An open source implementation in R language of conditional merging for weather radars and rain gauges observations. *Earth Sci. Inform.* **2017**, *10*, 127–135. [\[CrossRef\]](#)
20. Kembel, S.W.; Cowan, P.D.; Helmus, M.R.; Cornwell, W.K.; Morlon, H.; Ackerly, D.D. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **2010**, *26*, 1463–1464. [\[CrossRef\]](#)
21. Stanke, H.; Finley, A.O.; Weed, A.S.; Walters, B.F.; Domke, G.M. rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. *Environ. Modell. Softw.* **2020**, *127*, 104664. [\[CrossRef\]](#)
22. Lemenkova, P.; Debeir, O. R Libraries for Remote Sensing Data Classification by K-Means Clustering and NDVI Computation in Congo River Basin, DRC. *Appl. Sci.* **2022**, *12*, 12554. [\[CrossRef\]](#)
23. Seo, J.Y.; Lee, H.M. A study on statistical map of air pollution in Korea using R. In Proceedings of the 4th International Conference on Computer Applications and Information Processing Technology CAIPT2017, Kuta Bali, Indonesia, 8–10 August 2017.
24. Setiawan, I. Time series air quality forecasting with R Language and R Studio. *J. Phys. Conf. Ser.* **2020**, *1450*, 012064. [\[CrossRef\]](#)
25. Carslaw, D.C.; Ropkins, K. openair—An R package for air quality data analysis. *Environ. Modell. Softw.* **2012**, *27–28*, 52–61. [\[CrossRef\]](#)
26. Syaefi, A.D.; Fujiwara, A.; Zhang, J. Prediction model of Air Pollutant Levels Using Linear Model with Component Analysis. *Int. J. Environ. Sci. Dev.* **2015**, *6*, 519–525. [\[CrossRef\]](#)
27. Sethi, J.K.; Mittal, M. An efficient correlation based adaptive LASSO regression method for air quality index prediction. *Earth Sci. Inform.* **2021**, *14*, 1777–1786. [\[CrossRef\]](#)
28. Zheng, Y.; Xiuwen, Y.; Ming, L.; Ruiyan, L.; Zhangping, S.; Eric, C.; Tiannui, L. Forecasting Fine-Grained Air Quality Based on Big Data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 2267–2276.
29. Siwek, K.; Osowski, S. Data Mining Methods for Prediction of Air Pollution. *Int. J. Appl. Math. Comput. Sci.* **2016**, *26*, 467–478. [\[CrossRef\]](#)
30. Zhang, J.; Ding, W. Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. *Int. J. Environ. Res. Pub. He.* **2017**, *14*, 114. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Ibarra-Berastegi, G.; Elias, A.; Barona, A.; Saenz, J.; Ezcurra, A.; Diaz de Argandona, J. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. *Environ. Modell. Softw.* **2008**, *23*, 622–637. [\[CrossRef\]](#)
32. Zhao, R.; Gu, X.; Xue, B.; Zhang, J.; Ren, W. Short period PM_{2.5} prediction based on multivariate linear regression model. *PLoS ONE* **2018**, *13*, e0201011. [\[CrossRef\]](#) [\[PubMed\]](#)
33. Choi, S.-M.; Choi, H. Statistical Modeling for PM₁₀, PM_{2.5} and PM₁ at Gangneung Affected by Local Meteorological Variables and PM₁₀ and PM_{2.5} at Beijing for Non- and Dust Periods. *Appl. Sci.* **2021**, *11*, 11958. [\[CrossRef\]](#)
34. Young, M.T.; Bechle, M.J.; Sampson, P.D.; Szpiro, A.A.; Marshall, J.D.; Sheppard, L.; Kaufman, J.D. Satellite-Based NO₂ and Model Validation in a National Prediction Model Based on Universal Kriging and Land-Use Regression. *Environ. Sci. Technol.* **2016**, *50*, 3686–3694. [\[CrossRef\]](#)

35. Mani, G.; Viswanadhapalli, J.K.; Stonier, A.A. Prediction and forecasting of air quality index in Chennai using regression and ARIMA time series models. *J. Eng. Res.* **2022**, *10*, 179–194. [\[CrossRef\]](#)
36. Alsoltany, S.N.; Alnaqash, I.A. Estimating Fuzzy Linear Regression Model for Air Pollution Predictions in Baghdad City. *J. Al-Nahrain Univ.* **2015**, *18*, 157–166. [\[CrossRef\]](#)
37. Roy, S.S.; Paraschiv, N.; Popa, M.; Lile, R.; Naktode, I. Prediction of air-pollutant concentrations using hybrid model of regression and genetic algorithm. *J. Intell. Fuzzy Syst.* **2020**, *38*, 5909–5919. [\[CrossRef\]](#)
38. Sousa, S.I.V.; Martins, F.G.; Alvim-Ferraz, M.C.M.; Pereira, M.C. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environ. Modell. Softw.* **2007**, *22*, 97–103. [\[CrossRef\]](#)
39. Basagaña, X.; Aguilera, I.; Rivera, M.; Agis, D.; Foraster, M.; Marrugat, J.; Elosua, R.; Künzli, N. Measurement Error in Epidemiologic Studies of Air Pollution Based on Land-Use Regression Models. *Am. J. Epidemiol.* **2013**, *178*, 1342–1346. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Selvi, S.; Chandrasekaran, M. Performance evaluation of mathematical predictive modeling for air quality forecasting. *Cluster. Comput.* **2019**, *22*, 12481–12493. [\[CrossRef\]](#)
41. Iskandaryan, D.; Ramos, F.; Trilles, S. Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. *Appl. Sci.* **2020**, *10*, 2401. [\[CrossRef\]](#)
42. Briggs, D.J.; Collins, S.; Elliot, P.; Fischer, P.; Kingham, S.; Lebre, E.; Pryl, K.; Van Reeuwijk, H.; Smallbone, K.; Van der Veen, A. Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 699–718. [\[CrossRef\]](#)
43. Hochadel, M.; Heinrich, J.; Gehring, U.; Morgenstern, V.; Wichmann, H.E.; Kuhlbusch, T.; Link, E.; Kramer, U. Predicting long-term average concentrations of traffic-related air pollutants using GIS-based information. *Atmos. Environ.* **2006**, *40*, 542–553. [\[CrossRef\]](#)
44. Zhou, X.; Tong, W.; Li, L. Deep learning spatiotemporal air pollution data in China using data fusion. *Earth Sci. Inform.* **2020**, *13*, 859–868. [\[CrossRef\]](#)
45. Morandat, F.; Hill, B.; Osvald, L.; Vitek, J. Evaluating the Design of the R language. In *ECOOP 2012—Object-Oriented Programming; Lecture Notes in Computer Science*; Noble, J., Ed.; Springer: Berlin/Heidelberg, Germany, 2012; Volume 7313, pp. 104–131.
46. Environmental Protection Agency, Ministry of Environmental Protection, Republic of Serbia. National Network of Automatic Stations for Air Quality Monitoring—Raw Data Obtained from Measuring Stations. Available online: <http://www.amskv.sepa.gov.rs/stanicepodaci.php> (accessed on 1 January 2021).
47. Environmental Protection Agency, Ministry of Environmental Protection, Republic of Serbia. National Network of Automatic Stations for Air Quality Monitoring—Data View. Available online: <http://www.amskv.sepa.gov.rs/pregledpodatakazbirmi.php?lng=en> (accessed on 1 January 2021).
48. Environmental Protection Agency, Ministry of Environmental Protection, Republic of Serbia. National Network of Automatic Stations for Air Quality Monitoring—Criteria for Pollution Classification. Available online: <http://www.amskv.sepa.gov.rs/kriterijumi.php?lng=en> (accessed on 31 August 2022).
49. Jacob-Lopes, E.; Queiroz Zepka, L.; Costa Deprá, M. Methods of evaluation of the environmental impact on the life cycle. In *Sustainability Metrics and Indicators of Environmental Impact, Industrial and Agricultural Life Cycle Assessment*; Elsevier: Amsterdam, The Netherlands, 2021; pp. 29–70.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.