



INNOAIR

INDEX: UIA05-202

**Innovative demand responsive green
public transportation for cleaner air in
urban environment**

**Data analytics modelling of traffic
conjunction and air pollution**

11.2021



Abbreviations

Abbreviation	Explanation
AR	AutoRegression
ARMAX	AutoRegressive-Moving Average with exogenous (external) predictors
BIC	Bayesian information criterion
comps	components
ExEA	Executive Environment Agency, Ministry of Environment and Waters, Bulgaria
fcst	forecast
GAMs	Generalized Additive Models
MAE	Mean absolute error
MBE	Mean bias error
MLR	Multiple linear regression
MoE	Mixtures-of-Experts
NAs	Not Available values (missing values)
RMSE	Root mean square error
sm	smoothed

Authors

Name
Nadya N. Neykova
Prof. Neyko M. Neykov
Preslava Dimitrova
Stanislava Tsvetkova

Key data

Keywords	traffic air pollution, nitrogen dioxide, ozone, particulate matter, log-normal regression, mixtures-of-experts, multinomial logistic regression, neural networks, lasso penalization, time series forecasting, numerical weather prediction, WRF model
Lead Editor	Dimiter Atanassov
Internal Reviewer(s)	Dimiter Syrakov, Tatiana Spassova

Statement of originality

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Abstract

In this project study we developed several data analytics models to forecast 48 hours ahead average hourly concentrations of nitrogen dioxide (NO₂), tropospheric ozone (O₃) and particulate matter (PM₁₀) using data collected by the ExEA monitoring urban stations located at Hipodruma and Pavlovo in Sofia for the period 01.01.2019-07.05.2021. The models are hybrid between numerical weather prediction (WRF) model and stochastic models. The stochastic models are based on the methodology of the so called mixtures-of-experts time series regression models. The WRF output delivers 72 hour forecast which characterize the atmospheric conditions in Hipodruma and Pavlovo from ground surface up to 5500 m height. The WRF output derivatives, the lags of these derivatives, lags of NO₂, O₃ and PM₁₀ serve as input predictors into the stochastic models. The models provide reasonably good fits in terms of the standard forecast accuracy metrics such as MAE, RMSE and correlation coefficients between observed and forecast values. In addition, models including simulated traffic density as a relevant additional predictor are developed to forecast 48 hours ahead NO₂ and PM₁₀ average hourly concentrations using data for the period 01.01.2021-30.09.2021 and this inclusion shows an improvement of the results.



Table of Contents

ABSTRACT	3
TABLE OF CONTENTS	4
LIST OF TABLES.....	5
LIST OF FIGURES.....	6
1 INTRODUCTION TO DATA ANALYTICS AIR POLLUTION MODELS.....	7
2 DATA.....	10
3 EXPLORATORY DATA ANALYSIS.....	10
4 MISSING DATA IMPUTATION	14
5 STOCHASTIC AIR POLLUTION MODEL.....	17
6 MODEL PREDICTORS.....	18
6.1 THE WEATHER RESEARCH AND FORECAST (WRF) PREDICTORS	18
6.2 FOURIER SERIES PREDICTORS	19
6.3 TRAFFIC VEHICLE PREDICTORS AT STATION PAVLOVO	19
6.4 AUTOREGRESSIVE PREDICTORS.....	20
6.5 MODEL SPECIFICATION	20
7 DATA SPLITTING FOR TIME SERIES VALIDATION.....	21
8 ACCURACY MEASURES.....	21
9 MAXIMUM LIKELIHOOD ESTIMATES AND MODEL SELECTION.....	22
10 GRAPHICAL MODELS ASSESSMENT	22
11 CONCLUSIONS.....	27
12 APPENDIX	29
13 REFERENCES.....	33

List of tables

Table 1: Pattern of missing NO ₂ , O ₃ and PM ₁₀ hourly data values at Pavlovo	15
Table 2: Distribution of missing NO ₂ hourly data values at station Pavlovo.....	15
Table 3: Correlation coefficients between NO ₂ , PM ₁₀ and O ₃ at Hipodruma and Pavlovo.....	16
Table 4: Weather Research and Forecasting predictors.....	19
Table 5: Finite Fourier terms.....	19
Table 6: Autoregressive pollutant predictors	20
Table 7: Data splitting for time series	21
Table 8: Forecast accuracy measures	21
Table 9: Model quality forecasts based on test data set	22
Table 10: Pavlovo NO ₂ MoE log-normal time series regression model coefficients estimates.....	29



List of figures

Figure 1: Distributions of PM10 values at Pavlovo station	11
Figure 2: Distributions of NO2 values at Pavlovo station	12
Figure 3 Distributions of O3 values at Pavlovo station	13
Figure 4: Histogram of PM10 at Pavlovo and Hipodruma.....	14
Figure 5: Histogram of NO2 at Pavlovo and Hipodruma.....	14
Figure 6: Observed vs forecast NO2 at Hipodruma and Pavlovo.....	24
Figure 7: Observed vs forecast O3 at Hipodruma and Pavlovo	25
Figure 8: Observed vs forecast NO2 at Pavlovo with simulated traffic vehicle as predictor.....	26
Figure 9: Observed vs forecast PM10 at Pavlovo with simulated traffic vehicle as predictor.....	27

1 Introduction to data analytics air pollution models

Vehicular traffic plays an important role in atmospheric pollution and can be used as one of the key predictors in air-quality forecasting models. The models that can account the role of traffic are especially valuable in urban areas where high level of pollution concentrations are often observed during certain hours of the day (peak time) and especially during the cold half of the year. In this project, we develop a data analytics models based on mixtures-of-experts (MoE) of time series regression models to analyze the behavior of concentrations of nitrogen dioxide (NO₂), particulate matter (PM₁₀) and ozone (O₃) collected by the Executive Environment Agency (ExEA) stations located at Pavlovo and Hipodruma in the city of Sofia. The main reason to employ this class of models is their flexibility and ability to account for nonlinearities and other complexities such as heterogeneity in the data due to the weather conditions in Sofia area.

Thousands of papers about air pollution modeling have been published during the last 40 years because of its strong impact to the environment and human health. It is impossible an exhaustive review to be made. Thus only few papers related to this project will be considered because the forecasting pollution technology is well known - it is a synthesis between weather prediction models and stochastic models. Numerical prediction models provide the weather forecast a few days ahead whereas the pollution forecast is made by means of the stochastic models trained on the past pollution and weather forecast data. The more accurate weather forecast and air pollution data quality the more reliable air pollution forecast could be expected. Standard stochastic model techniques that have been used are based on the ordinary methodologies of: 1) multiple linear regression; 2) principal component regression; 3) nonparametric regression models using splines, kernels or wavelets to account nonlinearities of the predictors; 4) time series autoregression models of types ARX, ARMAX; 5) generalized linear models and generalized additive models in order handle non Gaussian distributed data; 6) various neural network types. Many of these models are well-accepted and proved in practice in the case of almost stationary and homogeneous air pollution data. However, in case of highly heterogeneous data these models are not adequate. In order to overcome this deficiency, the so called stochastic ensemble models techniques have been developed, details can be found in McLachlan and Peel (2000), Seni and Elder (2010), and Zhou (2012), just to name a few. Mixture-of-experts is an ensemble learning methodology developed in the field of statistics and neural networks. A short introduction can be found in Brownlee (2021). This methodology seeks to explicitly address predictive modeling problem in terms of several models such that any one of them accommodate locally to subsample of the data. In consequence, nonlinearities and other complexities of the data can be accounted for.

Ordieres et al. (2005) compared three different topologies of neural networks (multilayer perceptron, radial basis function and square multilayer perceptron) to the prediction of daily averages PM_{2.5} concentrations so as to identify their potential strengths and weaknesses within the prediction. The classical linear autoregression time series and multiple linear regression were included in the study as well. It is shown that the neural approach outperforms the classical models.

A comparison of 3 predictive models used to forecast 24 hours in advance the hourly ozone concentration is considered by Ballester et al. (2002). The following models were included in their study: (i) autoregressive-moving average with exogenous inputs (ARMAX); (ii) multilayer perceptron

and (iii) finite impulse response neural network. They focused on the O₃ summer peaks recorded by three Spanish monitoring stations situated in urban or rural environment, from 1996 to 1999. The five performance criteria calculated yield reasonably good results and indicated that the multilayer perceptron neural networks were more effective than the linear ARMAX models, which performed better than the dynamic finite impulse response neural networks.

The objective of the work of Paschalidou et al. (2009) is to compare various techniques for modeling the dependence of the tropospheric ozone concentrations on several meteorological and pollutant parameters. Their study focuses on two different sites in the metropolitan area of Athens, Greece, one in the city center and another one in the suburbs. These authors reduce the multicollinearity using the principal component analysis and then derive a prediction model via the stepwise regression analysis. They reported that the derived model is validated over test data and the multiple R-square coefficient is approximately 0.8.

Vlachogianni et al. (2011) developed forecast models based on stepwise multiple linear regression (MLR) for Athens and Helsinki to forecast the maximum hourly concentrations of PM₁₀ and NO_x, and the daily average PM₁₀ concentration of the next day.

Russo et al. (2015) present a simple neural network and data preselection framework, discriminating the most essential input data for accurately forecasting the concentrations of PM₁₀, based on observations for the years between 2002 and 2006 in the metropolitan region of Lisbon, Portugal. The daily values of PM₁₀ concentrations measured by 12 monitoring stations in the agglomeration of Lisbon between 2002 and 2006, which record the atmospheric concentrations of major pollutants, such as gases (e.g. NO₂, NO and CO) and PM₁₀ were considered.

Multiple linear regression is widely used method to find statistical relationships among PM_{2.5} concentrations and other pollutants, including CO, NO₂, SO₂, and O₃. As the relationship between PM_{2.5} and these predictors is usually complicated, Song et al. (2015) recommend the use of Generalized Additive Models (GAMs, Wood, 2017) which are capable to estimate and non-linear statistical relationships between the potential predictors and PM_{2.5} concentrations. Their model includes linear functions of SO₂ and CO, univariate smoothing non-linear functions of NO₂, O₃, aerosol optical thickness and temperature, and bivariate smoothing non-linear functions of location and winds. It is reported that the model can explain 69.50% of PM_{2.5} concentrations variability, the correlation coefficient between observed and predicted values equals to 0.83, which is an improvement against the stepwise linear regression as the corresponding correlation coefficient equals to 0.76.

Bertaccini et al. (2012) consider vehicular traffic as important predictors in air-quality forecasting models. They develop a GAMs approach (Wood, 2017) to analyze the behavior of concentrations NO₂ and PM₁₀, collected at the environmental monitoring stations distributed throughout the city of Turin, Italy, from 01.12.2003-30.04.2005. Although GAMs are highly flexible the authors reported that they get higher correlations between forecasted and observed data when developed their models separately within each seasons.

Perez et al. (2000) compare the multilayer neural network, linear regression and persistence methods. The methods were applied to hourly averaged PM_{2.5} data for the years 1994-1995 measured at a location in the downtown area of Santiago, Chile. It is shown that the neural network gave overall the best results in the prediction of the PM_{2.5} hourly concentrations, although the prediction errors for the hourly PM_{2.5} data were found to range from 30% to 60% for the neural network, from 30% to 70% for the persistence approach, and from 30% to 60% for the linear regression.

Abdul-Wahab et al. (2002) and Antanasijevic et al. (2013) also use artificial neural network in order to predict tropospheric O₃ concentration levels and PM₁₀, respectively. In fact, these authors employ linear logistic regression to predict occurrence of high pollution events such as Neykova and Neytchev (2019).

Chen et al. (2015) discussed the usage of the ordinal logistic regression in order to predict the risk of occurrence of high PM₁₀ values in Beijing, China. For this reason, the original PM₁₀ values are transformed into three ordered categories using two standard levels of pollution. The WRF model forecast of reasonable meteorological derivatives based on principal component analysis serve as input to the logistic regression model. To assess the forecast accuracy standard metrics were used such as the mean bias error, normalized mean bias error, mean absolute values and root mean squares error. The authors reported that using the hybrid forecast model they are able to predict high PM₁₀ pollution events more accurately than standard statistical methods such as multiple linear regression.

The influence of synoptic-scale circulations on air quality with interest to air quality management in regards to future climate change is studied by Pearce et al. (2011). The dominant synoptic types over the region of Melbourne, Australia are determined and linked to regional air quality. For this reason, the self-organizing map as an unsupervised tool is used to generate an automatic time series of synoptic charts classification the annual daily circulation affecting Melbourne into 20 different synoptic types. This classification is then employed within the framework of GAMs (Wood, 2017) to identify links between synoptic-scale circulations and observed changes air pollutant concentrations. A study similarly to Pearce et. al. (2011), however, instead of self-organizing map ensemble models based on mixture is discussed by Misiti et al. (2015). These authors consider mixtures of linear regression models for the statistical forecasting of the daily mean PM₁₀ concentration and discuss also the usage of GAMs (Wood, 2017) for the same purpose. Several meteorological variables as the daily mean temperature, atmospheric pressure and wind speed, maximum gradient of temperature and the average concentration measured on the previous day are included as predictors. The evaluation of these models is based on the forecast accuracy for several horizons starting from some hours to one day ahead.

Synoptic classification is a standard approach used in synoptic climatology where the atmospheric state is partitioned into several types usually in terms of the spatial patterns associated with the 500 hPa height. The synoptic classification can be performed manually by experts or automatically using the recent powerful unsupervised learning techniques for classification such as mixture of multivariate Gaussian distributions (McLachlan and Peel, 2001). These synoptic types are then used to provide insight into the influence of large scale processes on local environmental conditions. For instance, the time series vector of synoptic classification can be used as predictor within the time series regression type models like surface temperature, daily precipitation totals, daily air pollution concentrations, just to name a few.

Jacobs et al. (1991) discuss the usage of the so called mixtures-of-experts (MoE) within the framework of modelling nonlinear time series data. This model type is more flexible than the standard mixtures model with constant probabilities. In this way, the mixture probabilities are allowed to depend on the history of the process and external covariates (predictors, explanatory variables), a consideration that seems extremely natural in a time series setting. On the other hand, the move from model to model is far smoother. Furthermore, the MoE model despite its highly non-linear nature, is able to exploit the simplicity of the individual auto-regression components with external predictors in order to keep the model computationally tractable. More details can be found in McLachlan and Peel (2001).

2 Data

The 1st data set consists of both air pollutants and meteorological variables for the period 01.01.2019-07.05.2021. Air pollutants include hourly average concentrations of NO₂, PM₁₀ and O₃ measured at Pavlovo and Hipodruma stations of the Executive Environment Agency (ExEA), Sofia. The meteorological data are based on the numerical weather prediction model WRF (Weather Research and Forecasting), more details are given in section 6.1. The 2nd data set consists of simulated hourly vehicle traffic for the period 01.01-30.09.2021 for Pavlovo station only.

3 Exploratory data analysis

In order to get an impression of the NO₂, PM₁₀ and O₃ data distributions at Pavlovo stations, boxplots of the hourly concentrations at different time scales - months, week days, day hours, and weekend are presented on the plots of Figures 1 and 2, respectively. It is seen that PM₁₀ and NO₂ values are higher during the winter season and exhibit strong seasonal and diurnal variation, the extreme values during the week days are larger than those measured during the weekends. The extreme PM₁₀ and NO₂ values are observed during the dark and bright part of the day, respectively. Similar results are obtained for PM₁₀ and NO₂ data series at station Hipodruma.

The distributions of O₃ at station Pavlovo at different time scales are presented on the plots in Figure 3. It is seen that O₃ values are higher during the warm months of the year and exhibit strong seasonal and diurnal variation, the extreme values during the weekends are larger than those measured during the week days which is well known fact because of the lower values of NO₂ during the weekends. The extreme O₃ values are observed at noon and in the afternoon.

The histograms of PM₁₀ and NO₂ at Pavlovo and Hipodruma stations are presented in Figures 4 and 5. It is seen that these data series are right skewed whereas their logarithmic transformations are more symmetric but far from the histogram of the normally distributed data.

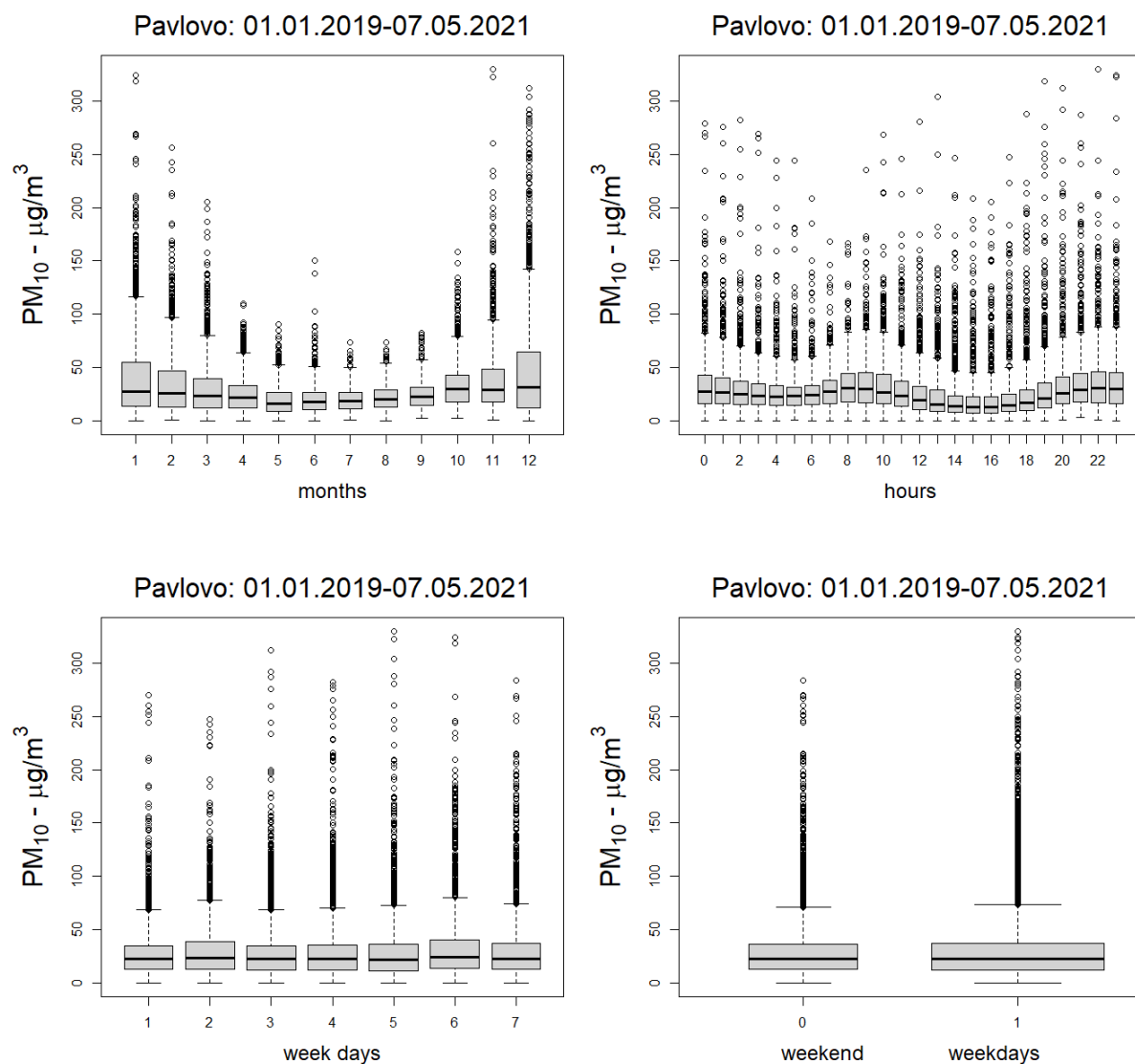


Figure 1: Distributions of PM₁₀ values at Pavlovo station: monthly, diurnal, week days and weekend data for the period 01.01.2019-07.05.2021.

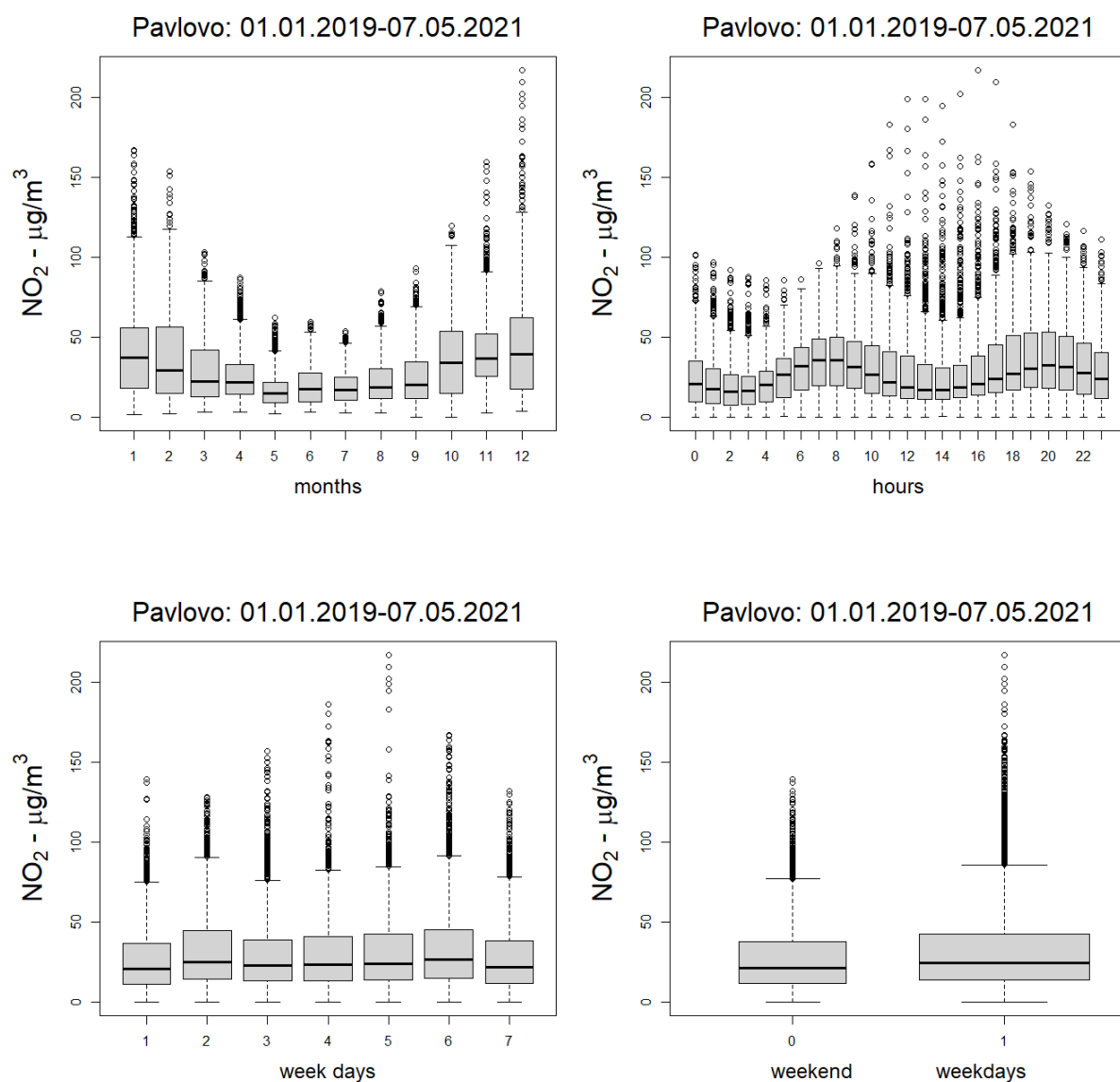


Figure 2: Distributions of NO₂ values at Pavlovo station - monthly, diurnal, week days and weekend data for the period 01.01.2019-07.05.2021.

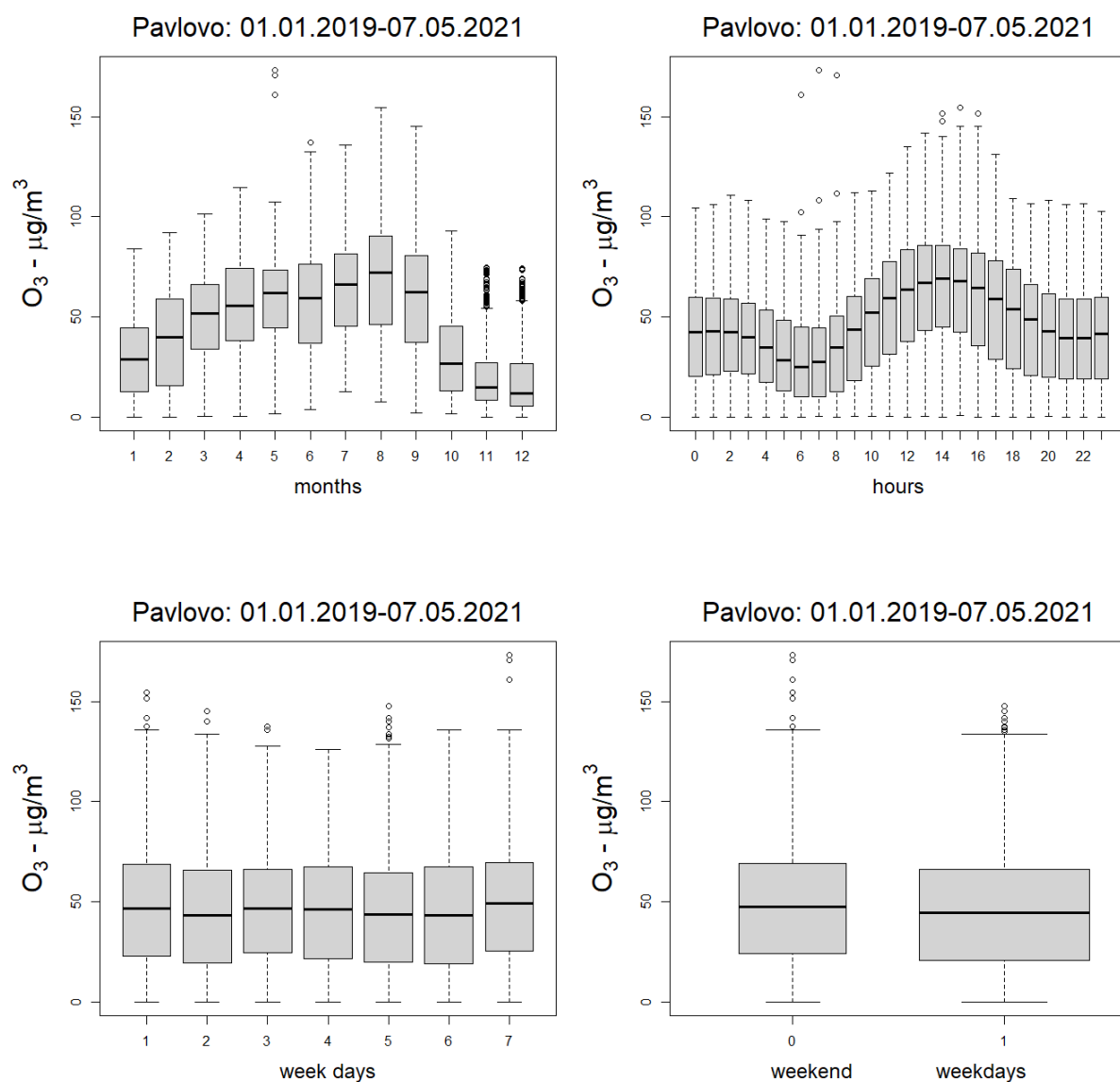


Figure 3: Distributions of O₃ values at Pavlovo station - monthly, diurnal, week days and weekend data for the period 01.01.2019-07.05.2021.

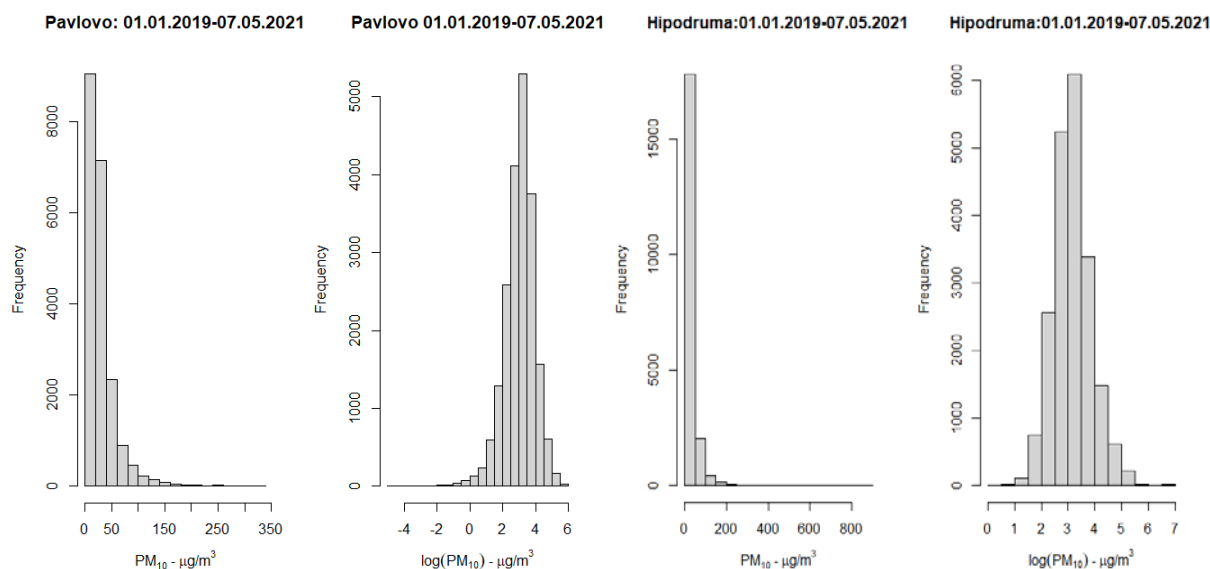


Figure 4: Histogram of PM₁₀ and logarithmic transformation of PM₁₀ data series at Pavlovo (left panel) and Hipodruma (right panel) for the period 01.01.2019-07.05.2021.

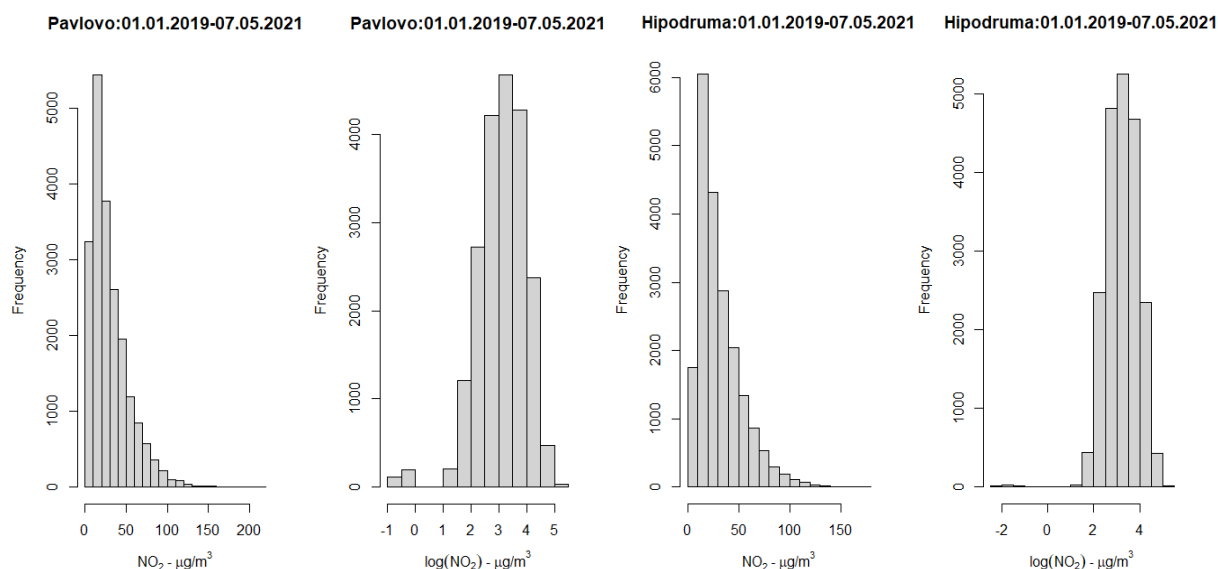


Figure 5: Histogram of NO₂ and logarithmic transformation of NO₂ data series at Pavlovo (left panel) and Hipodruma (right panel) for the period 01.01.2019-07.05.2021.

4 Missing data imputation

Missing values will be denoted by NA (Not Available). Missing data is a very common problem for all kinds of data. The presence of missing values in air quality data is mainly due to machine failures and human error. Therefore, it is extremely important to fill the missing data values otherwise the corresponding samples (data rows, cases) will be dropped out from the statistical and analytics data

processing software and this can reduce the sample size drastically. Incomplete data set usually causes bias due to differences between observed and unobserved data and inefficiency in data analytics modelling.

In order to get a better understanding of the pattern of missing data the *md.pattern* function from *mice* package in R is used on the input data. The output of this function is presented in Table 1. It tells us that 10637 samples are complete, 8304 samples miss only PM₁₀, 125 samples miss only the O₃ measurement, 50 samples miss simultaneously O₃ and PM₁₀, 2 samples miss only NO₂, 23 samples miss simultaneously NO₂ and PM₁₀, 17 samples miss simultaneously NO₂ and O₃, and 1434 samples miss simultaneously NO₂, O₃ and PM₁₀. This function is useful for investigating any structure of missing observations in the data. Each cell of Table 1 corresponds to a missing data pattern – 1 means observed and 0 means missing. The rows and columns are sorted in increasing amounts of missing information. The last column and row contain row and column counts of zeros, respectively. The *md.pattern* function do not tell us anything about the number and/or percentage of missing values for the split into bins, the longest NA gap, i.e., the longest series of consecutive missing values (NAs in a row) and most frequent occurring series of missing values, the series of consecutive missing values that account for most missing values overall in the time series, overview about how often each series of consecutive missing values occurs etc.

Table 1. Pattern of missing NO₂, O₃ and PM₁₀ hourly data values at station Pavlovo

	NO ₂	O ₃	PM ₁₀	
10637	1	1	1	0
8304	1	1	0	1
125	1	0	1	1
50	1	0	0	2
2	0	1	1	1
23	0	1	0	2
17	0	0	1	2
1434	0	0	0	3
	1476	1626	9811	12913

This is the reason to employ the *imputeTS* package in R. The *statsNA* function of this package calculates several missing data statistics of the input data. This includes overall percentage of missing values, absolute amount of missing values, amount of missing value in different sections of the data, longest series of consecutive NAs and occurrence of consecutive NAs. The summary statistics concerning the missing values in the input NO₂ hourly time series at station Pavlovo are shown on Table 2. From the rows of Table 2 it is seen that 1 (one) NA in a row occurs 341 times, two consecutive NA series in a row occur 17 times, etc. Finally, 297 consecutive NA in a row is the longest NA gap, i.e., series of 297 consecutive hours without NO₂ measurements which is due to a breakdown of the Pavlovo station.

Table 2. Distribution of missing NO₂ hourly data values at station Pavlovo

statsNA(NO ₂) from library(imputeTS)	
Length (sample number) of time series: 20592	
Number of Missing Values: 1476 (see 1 st and 2 nd columns of Table 1: 2+23+17+1434=1476)	
Percentage of Missing Values: 7.17%	
Number of Gaps: 381	
Average Gap Size: 3.874016Stats for Bins	
Bin 1 (5148 values from 1 to 5148) :	12 NAs (0.233%)

Bin 2 (5148 values from 5149 to 10296) :	312 NAs (6.06%)
Bin 3 (5148 values from 10297 to 15444) :	768 NAs (14.9%)
Bin 4 (5148 values from 15445 to 20592) :	384 NAs (7.46%)
Longest NA gap (series of consecutive NAs): 297 in a row	
Most frequent gap size (series of consecutive NA series)	
1 NA in a row (occurring 341 times)	
Gap size accounting for most NAs	
1 NA in a row (occurring 341 times, making up for overall 341 NAs)	
Overview NA series	
1 NA in a row: 341 times	26 NA in a row: 1 times
2 NA in a row: 17 times	27 NA in a row: 1 times
3 NA in a row: 5 times	30 NA in a row: 1 times
5 NA in a row: 1 times	36 NA in a row: 1 times
6 NA in a row: 2 times	71 NA in a row: 1 times
12 NA in a row: 1 times	129 NA in a row: 1 times
14 NA in a row: 2 times	156 NA in a row: 1 times
20 NA in a row: 1 times	187 NA in a row: 1 times
25 NA in a row: 2 times	297 NA in a row: 1 times

Lots of packages have been developed for data imputation. For instance, the *imputeTS* package offers the following function for univariate time series imputation techniques based on linear interpolation, imputation by spline interpolation, imputation by last observation carried forward, imputation by next observation carried backward, missing value imputation by simple moving average, missing value imputation by linear weighted moving average, missing value imputation by exponential weighted moving average, missing value imputation by mean value, missing value imputation by median value, just to name a few. However, all these imputation techniques are appropriate for stationary time series, i.e., for short NA gaps, e.g., 1, 2 or 3 series of consecutive NAs. Therefore, few of these packages are appropriate in general for air pollution time series data. This is because NO₂, O₃ and PM₁₀ time series are seasonal and non-stationary and require highly nonlinear models to be employed for data imputation. Thus we employ three single imputation techniques: 1) linear interpolation for data imputation with short gaps up to 2 consecutive NAs; 2) imputation by last observation carried forward and imputation by next observation carried backward with short gaps between 3 to 5 consecutive NAs; and 3) imputation by means of log-normal time series regression models between Pavlovo and Hipodruma stations data series in case of larger than 5 consecutive NAs gaps.

Table 3. Correlation coefficients between NO₂, PM₁₀ and O₃ data series at Hipodruma and Pavlovo.

	NO ₂ .H	PM ₁₀ .H	O ₃ .H	NO ₂ .P	PM ₁₀ .P	O ₃ .P
NO ₂ .H	1.00	-0.64	0.65	0.91	-0.59	0.64
PM ₁₀ .H	-0.64	1.00	-0.44	-0.60	0.93	-0.44
O ₃ .H	0.65	-0.44	1.00	0.63	-0.40	0.83
NO ₂ .P	0.91	-0.60	0.63	1.00	-0.63	0.69
PM ₁₀ .P	-0.59	0.93	-0.40	-0.63	1.00	-0.45
O ₃ .P	0.64	-0.44	0.83	0.69	-0.45	1.00

To perform the imputation concerning the 1st and 2nd approaches the corresponding *imputeTS* package functions were used. The 3rd approach is based on the high correlation coefficient values, 0.91, 0.93 and 0.83, respectively between NO₂, PM₁₀ and O₃ data series at these two stations according to Table

3. There is also a good reason to use this approach especially for O₃ as the majority of tropospheric ozone is generated through anthropogenic sources attributed to the photo-disassociation of NO₂. More details about can be found in Neykova and Neytchev (2020).

5 Stochastic air pollution model

A great deal of research on nonlinear time series regression models has been done during the last 30 years, as described by De Gooijer (2017). Among the several models proposed in the literature, one can find an important class of mixture of distributions denoted as mixtures-of-experts (MoE), see Jacobs et al. (1991), McLachlan and Peel (2001) and Carvalho and Tanner (2005). This class of models has been used in many different areas in order to account for nonlinearities and other complexities such as heterogeneity in the data. Mixture of experts is an ensemble learning technique developed in the field of statistics and neural networks. In contrast to ordinary statistical and neural network approaches which try to develop one model from the training data, ensemble methods try to develop a set of models and combine them. A short introduction to MoE can be found in Brownlee (2021) whereas McLachlan and Peel (2000), Seni and Elder (2010) and Zhou (2012) offer deep understanding of the theory and efficient computational algorithms.

Let $\mathbf{y}_t = \log(\text{PM}_{10})$ for $t = 1, \dots, T$ and $\mathbf{x}_t = (\mathbf{y}_{t-r}, \dots, \mathbf{y}_{t-d}, \mathbf{x}_{t1}, \dots, \mathbf{x}_{tm})$ is a vector of predictors including lags of transformations of the observed response $\mathbf{y}_t$ and relevant external meteorological (atmospheric) predictors, their lags and/or derivatives. The distribution of $\mathbf{y}_t$ is highly complex due to the weather conditions (states) $\mathbf{x}_t$. It can be approximated by the following class of mixtures-of-experts (Markov switching models within the econometrics) defined by

$$\psi(\mathbf{y}_t, \mathbf{x}_t; \Psi) = \sum_{j=1}^J \pi_j(\mathbf{z}_t, \beta_j) \varphi_j(\mathbf{y}_t, \mu_j(t), \sigma_j),$$

where $\varphi_j(\mathbf{y}_t, \mu_j(t), \sigma_j)$ is the density of the normal distribution with mean (expectation) $\mu_j(t) = \mathbf{x}_t^T \theta_j + \theta_{j0}$ and dispersion σ_j^2 , $\pi_j(\mathbf{z}_t, \beta_j) \in (0, 1)$ is the j th probability (weight) the so-called mixing proportions which satisfy the condition $\sum_{j=1}^J \pi_j(\mathbf{z}_t, \beta_j) = 1$,

$\Psi = (\beta_1^T, \dots, \beta_J^T, \theta_1^T, \dots, \theta_J^T, \theta_{1,0}, \dots, \theta_{J,0}, \sigma_1, \dots, \sigma_J)^T$ is the vector of unknown coefficients such that $\beta_j^T = (\beta_{j0}, \beta_{j1}, \dots, \beta_{jq})$ and $\theta_j = (\theta_{j0}, \theta_{j1}, \dots, \theta_{jp})$ and the coordinates of vector $\mathbf{z}_t$ are part of the $\mathbf{x}_t$ coordinates.

In other words, the distribution of the pollution potential $\mathbf{y}_t$ at time t which depends on the weather conditions $\mathbf{x}_t$ and additional predictors is characterized by the probability distribution $\varphi_j(\mathbf{y}_t, \mu_j(t), \sigma_j)$ with probability $\pi_j(\mathbf{z}_t, \beta_j)$

$$\psi(\mathbf{y}_t, \mathbf{x}_t; \Psi) = \begin{cases} \varphi_1(\mathbf{y}_t, \mu_1(t), \sigma_1) & \text{with probability } \pi_1(\mathbf{z}_t, \beta_1) \\ \varphi_2(\mathbf{y}_t, \mu_2(t), \sigma_2) & \text{with probability } \pi_2(\mathbf{z}_t, \beta_2) \\ \dots & \dots \\ \varphi_J(\mathbf{y}_t, \mu_J(t), \sigma_J) & \text{with probability } \pi_J(\mathbf{z}_t, \beta_J) \end{cases}.$$

Estimation of the unknown vector of coefficient Ψ of the mixture of distributions of time series models can be performed by maximizing the likelihood function

$$\max_{\Psi} \left(L(\Psi) = \prod_{t=d}^T \psi(\mathbf{y}_t, \mathbf{x}_t; \Psi) = \prod_{t=d}^T \sum_{j=1}^J \pi_j(\mathbf{z}_t, \beta_j) \varphi_j(\mathbf{y}_t, \mu_j(t), \sigma_j) \right).$$

From numerical point of view instead of maximization of the likelihood function the negative logarithm of the likelihood function is minimized. The optimization is performed with the so called *lasso* type penalization on the coordinates of the vector of coefficients Ψ in order to select the most significant predictors and overcome multicollinearity problem. Details can be found in McLachlan and Peel (2001) and Städler et al. (2010). Standard software based on the expectation-maximization (EM, divide and conquer) algorithm can be used to handle the computation. The R (R Core Team, 2018) packages MoEClust (Murphy and Murphy, 2020), MEclustnet (Gormley and Murphy, 2018), mixreg (Turner, 2014), mixtools (Benaglia et al., 2009), flexCWM (Mazza et al., 2017) can fit MoE whereas Brownlee (2021a) discuss Python libraries for fitting MoEs to models.

Note 1: In general, the probabilities $\pi_j(\mathbf{z}_t, \beta_j)$ are assumed to have a multinomial logistic (classification neural network) regression form:

$$\pi_j(\mathbf{z}_t, \beta_j) = \frac{\exp(\mathbf{z}_t^T \beta_j)}{\sum_{l=1}^J \exp(\mathbf{z}_t^T \beta_l)} \quad \text{for } j=1, \dots, J.$$

Note 2: In real applications the number of components J is unknown and has to be estimated. A classical approach is to then fit models with an increasing number of components and to compare them. The optimal number of mixture components is determined as minimum of the Bayesian Information Criteria (BIC) of Schwartz which is defined by

$$BIC = -2 \log(L(\hat{\Psi})) + ncoef * \log(T - d),$$

where $\hat{\Psi}$ is the maximum likelihood estimate of the vector of unknown model coefficients Ψ , $ncoef$ is the number of model coefficients (the length of Ψ) and $T-d$ is the data sample size, d is the maximal lag of $\mathbf{x}_t$ and $\mathbf{y}_t$.

Selecting a model that minimizes the BIC provides a relatively parsimonious model that fits the data well. However, BIC should be used as a guide only as the final decision must be evaluated in terms of physical realism.

6 Model predictors

6.1 The Weather Research and Forecast (WRF) Predictors

Numerical weather predictions come from the Weather Research and Forecasting (WRF) system. WRF is a mesoscale numerical weather prediction system designed to serve both atmospheric research and operational forecasting needs. The WRF output delivers 72 hour forecast which characterizes the atmospheric conditions in Sofia area for the next 3 days from ground surface up to 5500 m height. Some additional fields are calculated such as Monin-Obukhov scale, Pasquill stability parameter, temperature gradients in 7 layers, amount of low cloudiness, average vertical speed in the 50-meter ground layer, etc. The WRF predictors included within the stochastic pollutant forecast models are shown on Table 4. In order to account more strongly the meteorological impact on the pollution process dynamics lags of transformations of the WRF predictors such as the lags of surface temperature t_{2-t-6} , t_{2-t-12} , t_{2-t-18} , t_{2-t-24} , ... are also included in the stochastic models.

Table 4. Weather Research and Forecasting predictors.

t2	[C]	surface temperature at 2m
td2	[C]	dewpoint temperature
ts	[C]	soil temperature
ts/t2	[C]	surface temperature ratio
wd10	[deg]	wind direction at 10m
ws10	[m/s]	wind speed at 10m
Q2	[g/kg]	specific humidity
U*	[m/s]	roughness velocity
Moi	[1/m]	inverse Monin-Obukhov length
PBLH	[m]	Planetary Boundary Layer (PBL) height
PBLT	[m ² /s ²]	PBL kinetic turbulent energy
PBLL	[m]	PBL horizontal scale
CL		low cloudiness level (ceiling level)
Pasq		Pasquil stability parameter
W	[m/s]	mean vertical velocity in the lowest 50m layer
G20	[C/100m]	temperature gradient of layer 0-20m
G50	[C/100m]	temperature gradient of layer 20-50m
G120	[C/100m]	temperature gradient of layer 50-120m
G200	[C/100m]	temperature gradient of layer 120-200m
G300	[C/100m]	temperature gradient of layer 200-300m
G500	[C/100m]	temperature gradient of layer 300-500m
G800	[C/100m]	temperature gradient of layer 500-800m
RAINC	[mm]	convective precipitation
RAINNC	[mm]	nonconvective precipitation

6.2 Fourier series predictors

Finite Fourier terms are included as deterministic predictors in the model in order to account various hidden periodicity and trends of the pollution process such as seasonality, diurnality, the week days and weekends. For this purpose, the following couples of Fourier harmonics presented on Table 5 are used as model predictors:

Table 5. Finite Fourier terms.

Fourier sine and cosine harmonics		harmonics arguments	
m11=sin(arg11)	w11=sin(arg31)	arg11=(2 · π · month)/12	month=1,...,12
m12=cos(arg11)	w12=cos(arg31)	arg12=(4 · π · month)/12	month=1,...,12
m21=sin(arg12)	w21=sin(arg32)	arg21=(2 · π · hour)/24	hour=1,...,24
m22=cos(arg12)	w22=cos(arg32)	arg22=(4 · π · hour)/24	hour=1,...,24
h11=sin(arg21)	y11=sin(arg41)	arg31=(2 · π · wkday)/7	wkday=1,...,7
h12=cos(arg21)	y12=cos(arg41)	arg32=(4 · π · wkday)/7	wkday=1,...,7
h21=sin(arg22)	y21=cos(arg42)	arg41=(2 · π · yrday)/365.25	yrday=1,...,365.25
h22=cos(arg22)	y22=sin(arg42)	arg42=(4 · π · yrday)/365.25	yrday=1,...,365.25

6.3 Traffic vehicle predictors at station Pavlovo

Hourly vehicle traffic was simulated for the period 01.01.-30.09.2021 for Pavlovo station. Several NO₂ and PM₁₀ mixture models with different components number were developed including predictors

such as: 1) lags of simulated vehicle traffic; 2) WRF data and their lags described in section 6.1; 3) Finite Fourier series in section 6.2; and 4) NO₂ and PM₁₀ and their lags described in section 6.3.

6.4 Autoregressive predictors

In order to account for the evolution of the corresponding pollution process lags of transformations of the observed response NO₂, O₃ and PM₁₀ measured at Pavlovo and Hipodruma stations are included in the corresponding model as predictors at the following discrete times in hours 48, 52, 56, 60, 64, 68 and 72. This means that past values of any of the above pollutants could be included in the model as predictors with a delay at least 48 hours. Lags transformation was applied to simulated vehicle traffic (Veh.lag) at Pavlovo at well at the same discrete times. The potential predictors are given on Table 6.

Table 6. Autoregressive pollutant predictors

Autoregressive pollutant predictors					
NO ₂ .lag.48	NO ₂ .lag.52	NO ₂ .lag.56	NO ₂ .lag.60	NO ₂ .lag.64	NO ₂ .lag.72
O ₃ .lag.48	O ₃ .lag.52	O ₃ .lag.56	O ₃ .lag.60	O ₃ .lag.64	O ₃ .lag.72
PM ₁₀ .lag.48	PM ₁₀ .lag.52	PM ₁₀ .lag.56	PM ₁₀ .lag.60	PM ₁₀ .lag.64	PM ₁₀ .lag.72
Veh.lag.48	Veh.lag.52	Veh.lag.56	Veh.lag.60	Veh.lag.64	Veh.lag.72

6.5 Model specification

A great variety of models can be defined through a combination of the mixtures-of-experts components number and the potential predictors such as those discussed in sections 6.1, 6.2 and 6.3.

1. Denote by the acronyms m1.Hipodruma and m2.Pavlovo the following two O₃ models defined as MoE log-normal time series regression models including all the relevant predictors in sections 6.1 and 6.2. Let's add to them the predictors from the 1st and 2nd rows of Table 6 measured at Hipodruma and Pavlovo stations.
2. Similarly, denote by the acronyms m3.Hipodruma and m4.Pavlovo the following two NO₂ models defined as MoE log-normal time series regression models including all the relevant predictors in sections 6.1 and 6.2. Let's add to them the predictors from the 1st row of Table 6 measured at Hipodruma, respectively Pavlovo, stations to m3.Hipodruma, respectively to m4.Pavlovo, models.
3. Denote by the acronyms m5.Pavlovo.NO₂.Veh and m6.Pavlovo.PM₁₀.Veh the following two NO₂ and PM₁₀ models defined as MoE log-normal time series regression models including all the relevant predictors in sections 6.1 and 6.2. Let's add to them the predictors from the 1st, 3rd and 4th rows of Table 6 measured at Pavlovo stations only.

Models m1.Hipodruma, m2.Pavlovo, m3.Hipodruma and m4.Pavlovo are based on training data for the period 01.01.2019-31.12.2020 whereas the period 01.01.2021-07.05.2021 is reserved for the test data. Models m5.Pavlovo.Veh and m6.Pavlovo.Veh are based on training data for the period 01.01.2021-31.08.2021 whereas the period 01.09.2021-30.09.2021 is reserved for the test data. The "rolling forecasting origin" cross-validation procedure is used to assess the forecast accuracy on the test data. More details are given in section 7.

Various combinations of MoE log-normal time series regression models were fitted to the training data. Due to the modern regularization (penalization) techniques one can afford to handle models including many predictors. Table 9 contains only those models selected by optimal BIC. The observed and forecasts values of any of these models are presented on the plots in Figures 6 – 9 in section 10.

7 Data splitting for time series validation

The standard random sampling of data is not suitable to resample times series data as pointed out by Hyndman and Athanasopoulos (2013) because the temporary dependence between observations would be destroyed. Thus, we follow the so called “rolling forecasting origin” cross-validation procedure offered by these authors that move the training and test sets in time. Because we are interested in models that produce 48-hours-ahead forecasts then these techniques can be adapted to allow multi-step errors to be used. The corresponding data splitting of training and test sets is presented schematically on Table 7. The cells in blue color indicate samples that are included in the training set, the cells in red indicate samples in the test set, and the green ones indicate samples that are not used yet in the cross-validation analysis. In this way 63 folds are formed as the period 01.01.2021- 07.05.2021 is reserved for testing the quality of the forecasts.

Table 7. Data splitting for time series

		01.01.2019-31.12.2020 yrs (17544 hours)										01.01.2021-07.05.2021 yrs (3048 hours)									
		48 hours		...		...		...		...		48 hours		...		...		...		...	
s	1	●	●	●	●	...	●	●	...	●	●	●	●	●	...	●	●	●	●	●	●
a	2	●	●	●	●	...	●	●	...	●	●	●	●	●	...	●	●	●	●	●	●
m	3	●	●	●	●	...	●	●	...	●	●	●	●	●	...	●	●	●	●	●	●
p	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
l	62	●	●	●	●	...	●	●	...	●	●	●	●	●	...	●	●	●	●	●	●
e	63	●	●	●	●	...	●	●	...	●	●	●	●	●	...	●	●	●	●	●	●

8 Accuracy Measures

Common metrics to evaluate the forecast accuracy include: (1) mean absolute error (MAE); (2) root mean square error (RMSE); (3) mean bias error (MBE); and (4) correlation coefficient (R). These measures are evaluated on the test data, the predicted values are based on the test data set, which is independent from the training data set. Definitions of these measures are given on Table 8. Mean absolute error is the average of the difference between predicted and observed values of the data. MAE ranges from 0 to infinity and a perfect fit is obtained when MAE equals to 0. The RMSE gives the error value of the same dimensionality as the actual and predicted values. The smaller value of RMSE indicates the better performance of the model. The bias or MBE is the average forecast error representing the systematic error of a forecast model to under or overforecast and it takes values on the whole real line, value closer to zero implying a perfect fit. The coefficient of correlation (R) takes on values between -1 and 1, with values closer to 1 implying a better fit.

Table 8. Forecast accuracy measures

$MAE = \frac{1}{n} \sum_{i=1}^n y_i^{test} - \hat{y}_i^{test} $	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{test} - \hat{y}_i^{test})^2}$
$MBE = \frac{1}{n} \sum_{i=1}^n (y_i^{test} - \hat{y}_i^{test})$	$cor(y^{test}, \hat{y}^{test}) = \frac{\sum_{i=1}^n (y_i^{test} - \bar{y}^{test})(\hat{y}_i^{test} - \bar{\hat{y}}^{test})}{\sqrt{\sum_{i=1}^n (y_i^{test} - \bar{y}^{test})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i^{test} - \bar{\hat{y}}^{test})^2}}$

9 Maximum likelihood estimates and model selection

For completeness, the coefficient estimates of the model with acronym m2.Pavlovo based on NO₂ mixtures-of-experts log-normal time series regression model with 5 components are presented on Table 10. One can see that due to lasso penalization many of the coefficient estimates are zeros. This means that the impact of the corresponding predictor is negligible within the corresponding component. As NO₂ mixture model is comprised by 5 log-normal time series regression models one could associated these components by 5 typical weather types over Sofia which are latent (unobserved, hidden) in the data matrix of predictors. The evaluation of the models is based on the accuracy forecast for several horizons starting from some hours to two days ahead. Similar results were found about O₃ and PM₁₀ for these two stations. Summary about the optimal models selected by BIC among several competing models comprised by different components and predictors as discussed in section 6.5 are given on Table 9.

Table 9. Model quality forecasts based on test data set

Model.Station	pollutant	MoE comps	BIC	R	MAE	RMSE	MBE
m1.Hipodruma	O ₃	5	30920.71	0.91	6.24	6.42	2.99
m2.Pavlovo	O ₃	5	17794.83	0.86	8.82	12.08	-2.59
m3.Hipodruma	NO ₂	5	23953.43	0.74	10.26	14.45	1.68
m4.Pavlovo	NO ₂	5	23625.37	0.76	9.82	13.78	1.59
m5.Pavlovo.NO ₂ .Veh	NO ₂	5	10402.73	0.82	7.12	10.27	3.56
m6.Pavlovo.PM ₁₀ .Veh	PM ₁₀	4	8182.54	0.77	3.69	6.35	2.81

10 Graphical models assessment

Figures 6-9 show a comparison of measured and forecast values of NO₂, and O₃ pollutants on the test data sets. The observed (obs) and forecast (fcst) values of these pollutants on any of the plots of these Figures are given by red and blue continuous lines whereas their corresponding smoothed (sm) values are given with dotted lines in red and light green, respectively.

Three test data subsamples of length 144 consecutive hours are given on the plots in Figures 6 and 7 in order to get an impression about the forecast accuracy of the pollutants NO₂ and O₃ measured at the monitoring stations Hipodruma and Pavlovo. The starting times of any of these samples begin at 00:00h on 01.01.2021, 20.02.2021 and 11.04.2021 which correspond to (a) and (b), (c) and (d), and (e) and (f) plots of Figures 6 and 7, respectively. The NO₂ and O₃ forecasts are due to the models with acronyms m1.Hipodruma and m2.Pavlovo, and m3.Hipodruma and m4.Pavlovo, respectively defined in section 6.5. The metrics characteristics of these forecasts are given on Table 9. The high values of the correlation coefficients in both monitoring stations between observed and forecast values of O₃ can be explained by the usage of NO₂ lags of transformation included as predictors of the developed models.

The plots of Figure 8 show four different forecasts due to four competitive NO₂ MoE log-normal time series regression models developed for Pavlovo station in order to study the traffic effect on the air pollution. These models were fitted to the training time series data set for the period 01.01-31.07.2021 whereas the test sample was reserved for the period 01.08-30.09.2021. The BIC criterion is minimal for model m5.Pavlovo.NO₂.Veh defined in section 6.5. The coefficient of correlations between observed and forecast values of this model equals to 0.82 whereas the other accuracy metrics are given on Table 9. The alternative models of m5.Pavlovo.NO₂.Veh were based on MoE of log-normal time series regression with 5 components including the following predictors: finite Fourier trends and WRF predictors and their lags (m5.Pavlovo.NO₂.-FF.WRF.lags), lags of logarithm of NO₂ and simulated vehicle number (m5.Pavlovo.NO₂.laglogNO₂Veh), lags of logarithm of simulated vehicle number (m5.Pavlovo.NO₂.laglogVeh). The BIC values of these models are 11765.91, 11595.46 and 12198.96, respectively. The forecasts presented on the plots (a), (b), (c) and (d) of Figures 8 correspond to these models, respectively. The starting time of any of these forecasts begin at 00:00 h on 06.08.2021 and end at 23:00 h on 11.08.2021. Observed and forecast values of NO₂ are given in red and blue continuous lines whereas their respective smoothed values are given with dotted lines, respectively in red and light green. The comparison between these models shows that additional traffic vehicle information improves NO₂ forecast at Pavlovo station although lags transformation of NO₂ are included in the models, which implicitly could be related with the traffic vehicles.

The PM₁₀ model m6.Pavlovo.PM₁₀.Veh defined in section 6.5 is optimal according BIC criterion among several competitive models. The lags transformations of the simulated hourly traffic number are included as predictors in order to account their effects on PM₁₀ pollution at station Pavlovo. The three PM₁₀ forecasts and their corresponding test data subsamples of length 144 consecutive hours due to the m6.Pavlovo.Veh model are presented on the plots in Figures 9. The starting times of any of these samples begin at 00:00 on: (a) 01.09.2021 and end at 23:00 h 06.09.2021; (b) 00:00 h on 13.09.2021 and end on 23:00 h on 18.09.2021; (c) 00:00 h on 21.09.2021 and end on 23:00 h 26.09.2021. The forecast values presented on the plot (d) of this figure are due to a model with 3 components only, but with the same predictors as model m6.Pavlovo.PM₁₀.Veh, however with larger BIC value.

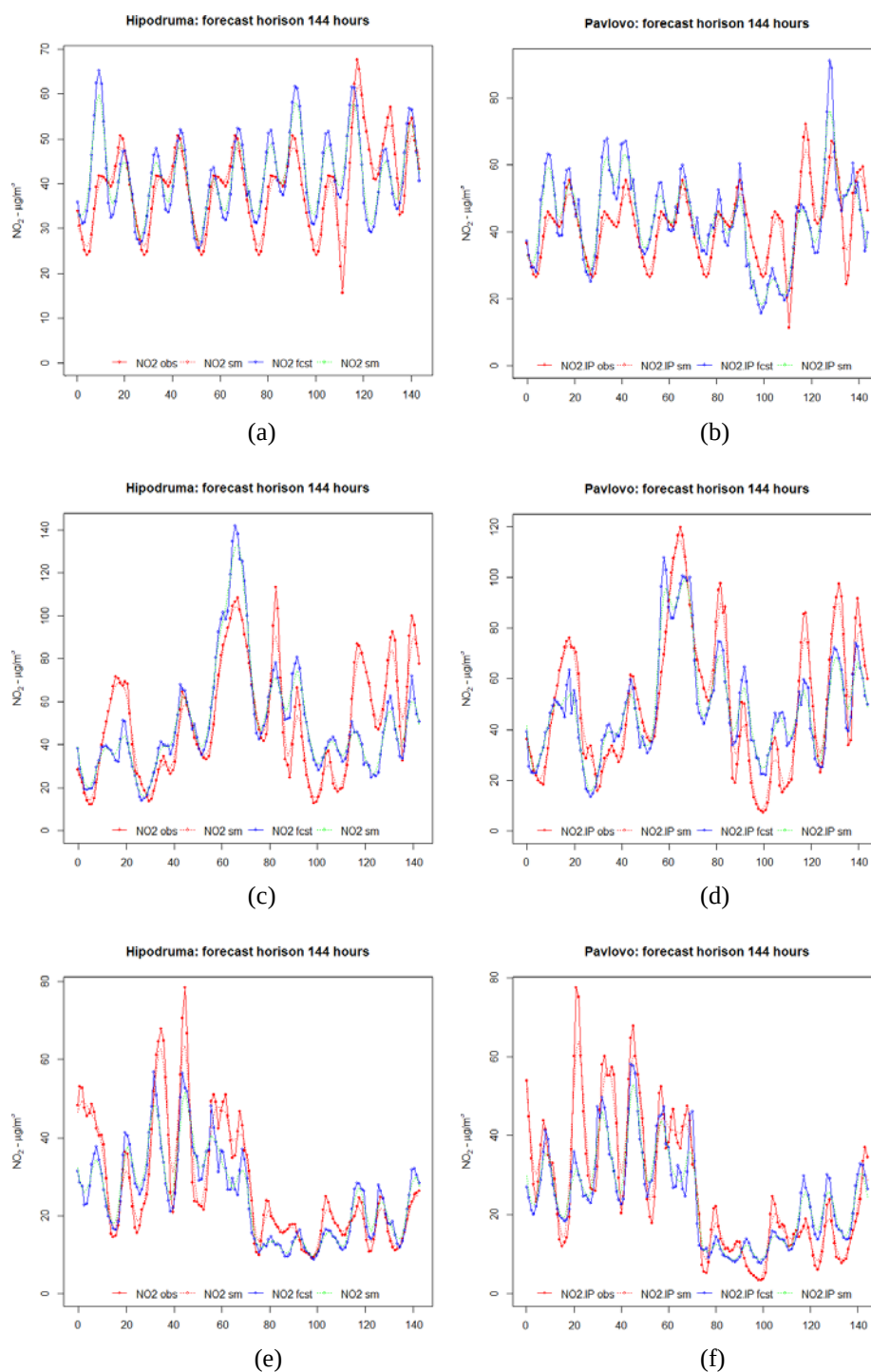


Figure 6. NO₂ test data samples measured at Hipodruma (left) and Pavlovo (right). Observed (red lines) and forecast (blue lines) values due to m1.Hipodruma and m2.Pavlovo models whereas their respective smoothed values are given with dotted lines, respectively in red and light green.

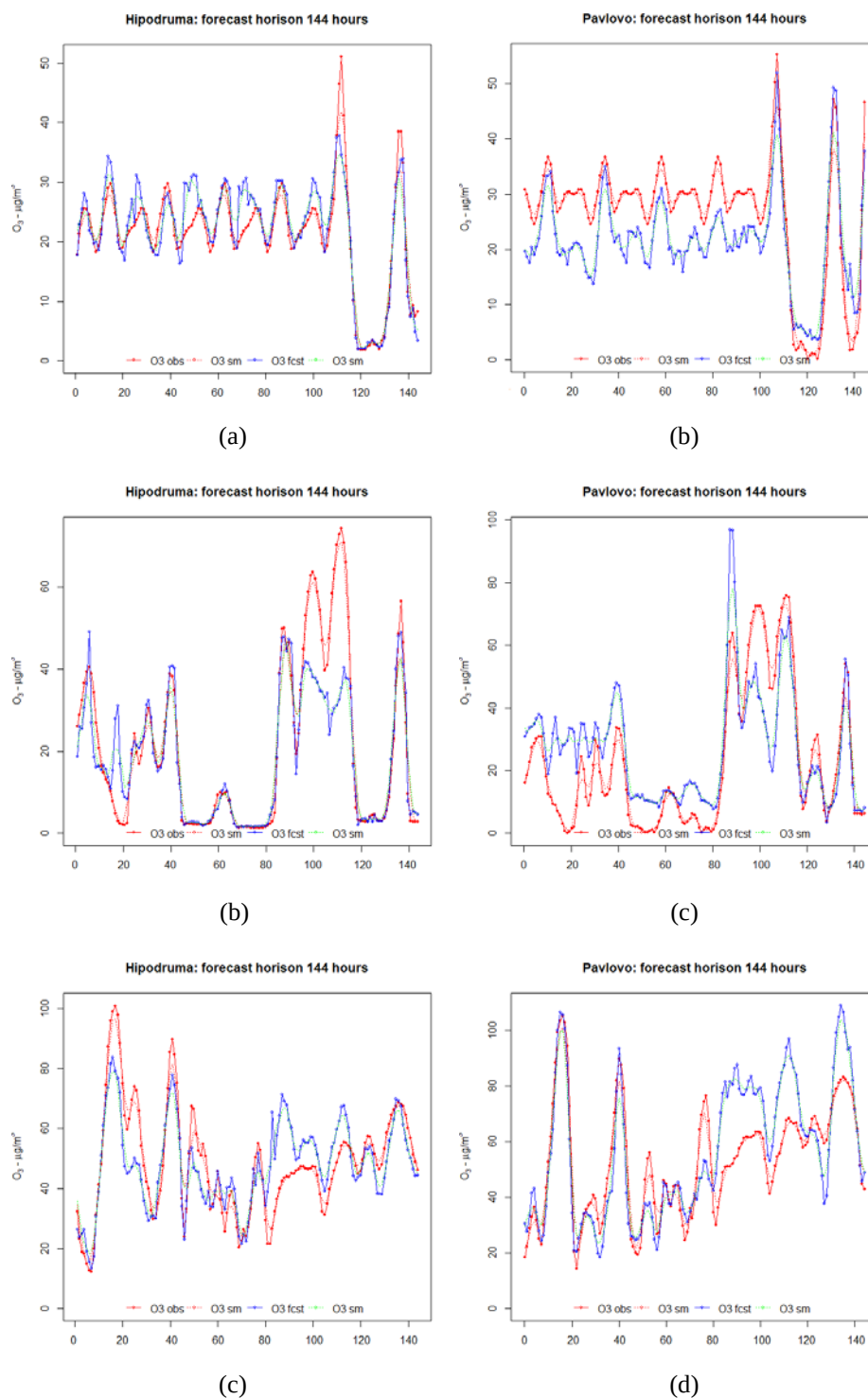


Figure 7. O₃ test data samples measured at Hipodruma (left) and Pavlovo (right). Observed (red lines) and forecast (blue lines) values due to m1.Hipodruma and m2.Pavlovo models whereas their respective smoothed values are given with dotted lines, respectively in red and light green.

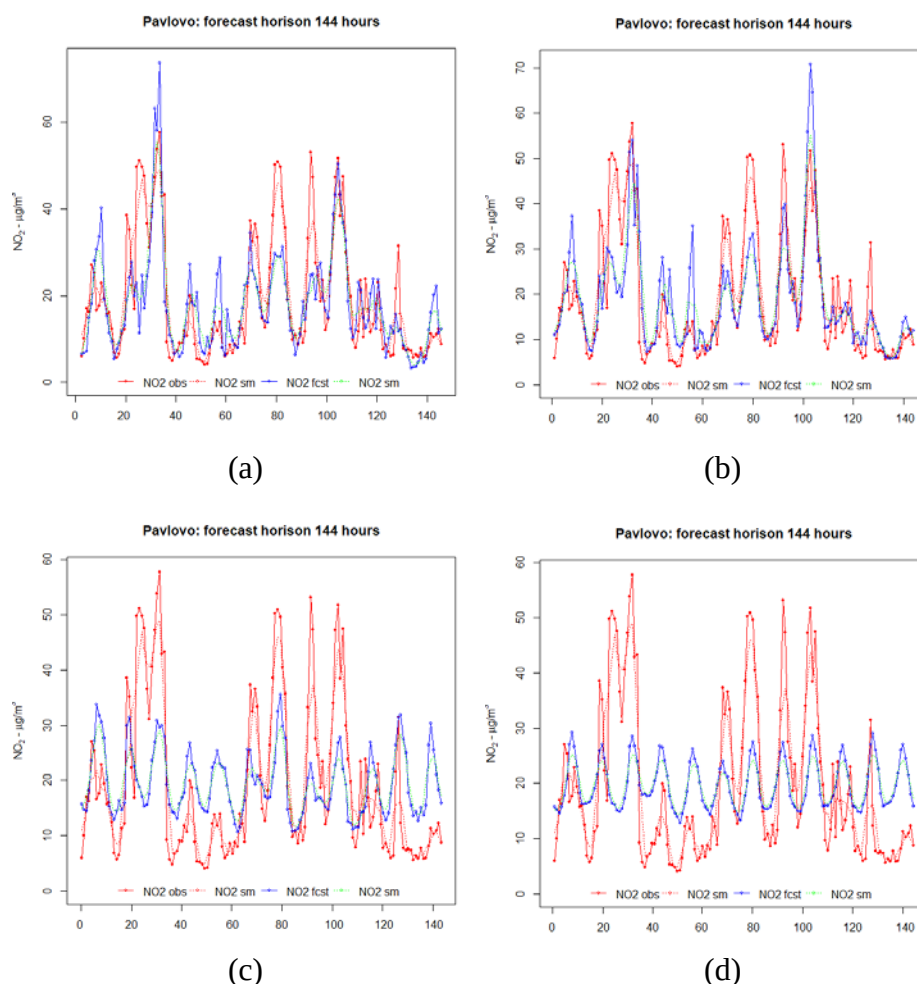


Figure 8. Traffic vehicle impact: NO₂ test data samples measured at Pavlovo station. Observed (red lines) and forecast (blue lines) values are due to models *m5.Pavlovo.NO2.Veh* (a), *m5.Pavlovo.NO2.FF.WRF.lags* (b), *m.Pavlovo.NO2.laglogNO2Veh* (c), and *m5.Pavlovo.NO2.laglogVeh* (d), whereas their respective smoothed values are given with dotted lines, respectively in red and light green. The starting time of the forecast begin at 00:00h on 06.08.2021 and end at 23:00h on 11.08.2021.

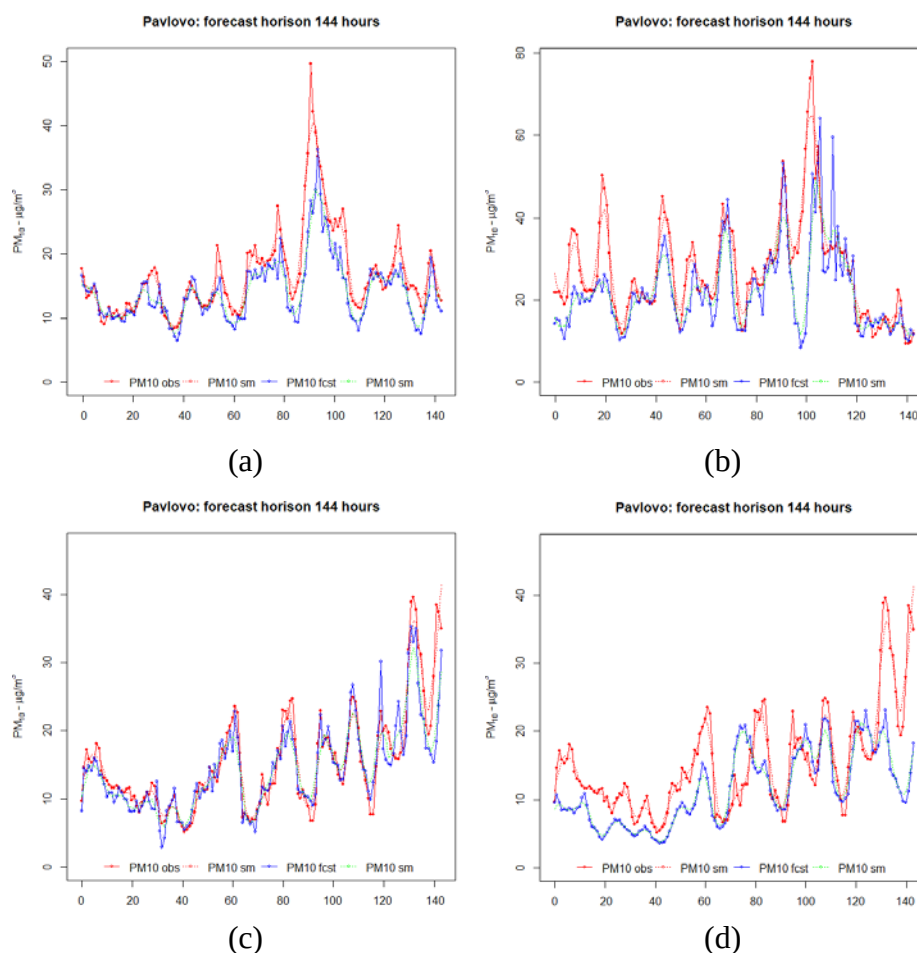


Figure 9. Traffic vehicle impact: PM10 test sample at Pavlovo station. Observed (red lines) and forecast (blue lines) values are due to m6.Pavlovo.Veh model whereas their corresponding smoothed values are given with dotted lines, respectively in red and light green. The starting times of any of these samples begin at: (a) 00:00h on 01.09.2021 and end at 23:00h on 06.09.2021; (b) 00:00h on 13.09.2021 and end on 23:00h on 18.09.2021; (c) and (d) 00:00h on 21.09.2021 and end on 23:00h 26.09.2021.

11 Conclusions

In this project we developed data analytics models to forecast 48 hours ahead hourly NO₂, O₃ and PM₁₀ concentrations using data collected by the ExEA monitoring stations located at Hipodruma and Pavlovo for the periods 01.01.2019-07.05.2021 and 01.01.2021-30.09.2021. These models are hybrid between numerical weather prediction (WRF) model and stochastic models. The stochastic model is based on the methodology of the so called mixtures-of-experts time series regression models. The WRF output delivers 72 hour forecast which characterize the atmospheric conditions in Hipodruma and Pavlovo from ground surface up to 5500 m height. The WRF output derivatives, the lags of these derivatives, lags of NO₂, O₃ and PM₁₀ serve as input predictors into the stochastic models. The models provide reasonably good fits in terms of the standard forecast accuracy metrics such as MAE, RMSE and correlation coefficients between observed and forecast values. The project report presents a new



methodology for estimation of the relationship between each air pollution ingredient NO₂, O₃ and PM₁₀ and the numerical weather prediction output meteorological parameters for two locations of the urban area of Sofia, Bulgaria. The estimated models for these two locations share a common trait and exhibit a general trend. Hence, similar models could be developed for other locations. The traffic density effect on the pollutants NO₂ and PM₁₀ as a relevant additional variable is based on simulated data and this inclusion shows an improvement of the results.

12 Appendix

Table 10. Pavlovo NO₂ mixtures-of-experts log-normal time series regression model with 5 components and minimal BIC.

Mixtures-of-experts log-normal NO ₂ time series regression model with 5 components and BIC=17794.83						
predictors	coefficients	Component.1	Component.2	Component.3	Component.4	Component.5
intercept	coef1	3.732242678	2.396554e+00	2.578184602	3.0553419160	3.3734748087
m11	coef2	-0.175600391	2.663747e-01	0.837280730	-0.2814112626	0.0871046602
m12	coef3	0.104484252	0.000000e+00	-0.004076924	0.0000000000	0.0305555266
m21	coef4	-0.055305466	0.000000e+00	0.0000000000	-0.1129847535	0.0000000000
m22	coef5	-0.042374060	-2.200649e-01	0.158003487	-0.1830393849	-0.0482604587
h11	coef6	-0.056492858	-3.162206e-01	0.0000000000	0.4431115253	0.1515254514
h12	coef7	0.039455054	-3.735947e-01	-0.229618560	0.3290658646	0.0871797273
h21	coef8	0.061427356	1.087193e-01	0.137576467	0.0000000000	0.0687299413
h22	coef9	0.059675848	0.000000e+00	0.262659478	0.0000000000	0.0423021638
w11	coef10	0.022242283	1.181852e-01	0.048605319	0.0965298271	0.0379067527
w12	coef11	0.033432234	5.829326e-02	0.062157577	0.0295718593	0.0388039393
w21	coef12	0.0000000000	1.431096e-01	0.0000000000	0.0099473237	0.0167656728
w22	coef13	0.017180379	3.884106e-02	0.042270213	0.0300344281	0.0470523543
y11	coef14	0.274376210	-6.264495e-01	-0.481836527	0.4308153798	0.0000000000
y12	coef15	-0.231499035	-1.330905e+00	-0.864542807	0.0802866217	-0.1168319115
y21	coef16	-0.039752164	0.000000e+00	0.0000000000	0.0000000000	-0.0815836986
y22	coef17	0.135608332	-2.078957e-01	0.339143383	0.2503252335	0.0536828875
CL	coef18	-0.149641710	2.511716e-01	0.373719685	0.0000000000	-0.0492731889
t2	coef19	0.079879580	-1.146738e+00	0.0000000000	0.2357395654	0.1384327162
td2	coef20	-0.158321351	-1.286656e+00	-1.011856782	0.0000000000	-0.2896432953
ts	coef21	0.0000000000	8.362046e-01	0.074955456	0.0000000000	0.0000000000
tst2	coef22	0.0000000000	0.000000e+00	0.0000000000	0.0000000000	0.0000000000
u10	coef23	0.0000000000	7.791314e-02	0.102133845	0.0265376577	0.0007396989
v10	coef24	0.017963964	1.225613e-01	0.0000000000	0.0531853949	0.0162121308
Q2	coef25	0.075634035	9.215274e-01	0.493751167	0.0030163420	0.1997475485
Us	coef26	0.046176851	0.000000e+00	0.088171588	0.0000000000	0.0010529649
MOi	coef27	0.0000000000	0.000000e+00	0.0000000000	0.0000000000	0.0033612717
PBLH	coef28	0.0000000000	0.000000e+00	0.0000000000	0.0000000000	-0.0116749601
PBLT	coef29	0.0000000000	2.671525e-01	-0.254138449	0.0000000000	0.0000000000
PBLL	coef30	0.013133121	1.169480e-01	0.150222298	0.1563536461	0.1316957181
W	coef31	0.0000000000	5.427385e-02	-0.135099328	0.0285747382	0.0000000000
Pasq	coef32	0.017117789	-7.588284e-02	0.0000000000	-0.0281110283	0.0367735905
G20	coef33	0.028812893	2.639247e-01	0.124298272	0.2004417639	0.1541395487
G50	coef34	0.0000000000	0.000000e+00	-0.008951876	-0.1160405005	-0.0315808789
G120	coef35	0.0000000000	3.915013e-02	0.080716057	0.1789708216	0.1097545581
G200	coef36	0.049735827	2.186988e-01	0.074497083	-0.0287185561	0.2122899067
G300	coef37	0.0000000000	0.000000e+00	0.084761895	0.0258868018	0.0644497446
G500	coef38	0.014968724	-1.309746e-01	0.112458929	0.1349207792	0.0000000000
G800	coef39	0.016515998	1.685337e-01	0.028621578	0.0656996334	0.0000000000
RAIN	coef40	-0.002103759	0.000000e+00	0.0000000000	0.0000000000	0.0000000000
RAINNC	coef41	0.0000000000	0.000000e+00	0.0000000000	0.0000000000	-0.0042210554
O3.H.48	coef42	0.057995497	1.243531e-01	0.568989413	0.2849957224	0.0176754277
O3.H.52	coef43	0.0000000000	0.000000e+00	0.0000000000	0.0000000000	0.0000000000



O3.H.56	coef44	0.008653029	2.999228e-01	0.074957513	0.0000000000	0.0010955193
O3.H.60	coef45	0.000000000	-8.244861e-02	0.000000000	0.0000000000	0.0000000000
O3.H.64	coef46	0.000000000	2.098649e-01	0.069674013	0.1184873441	0.0424621235
O3.H.68	coef47	0.000000000	-1.217786e-04	0.205719215	0.0000000000	0.0781020026
O3.H.72	coef48	0.034565593	1.999277e-01	0.000000000	0.0000000000	0.0579690483
t2.6	coef49	0.000000000	-2.658531e-01	-0.290285944	0.0000000000	0.0000000000
t2.12	coef50	0.043681434	0.000000e+00	0.273382047	0.4210161285	0.1227975498
t2.18	coef51	0.000000000	0.000000e+00	0.364999629	0.5932172900	0.0000000000
t2.24	coef52	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
td2.6	coef53	0.057642391	0.000000e+00	0.000000000	0.0000000000	0.0000000000
td2.12	coef54	0.000000000	3.390074e-05	0.000000000	-0.3140317821	0.0279960801
td2.18	coef55	0.000000000	-2.891896e-01	0.000000000	-0.1733258625	-0.0086888080
td2.24	coef56	-0.052728029	4.643951e-01	0.004498699	0.0000000000	0.0159602622
ts.6	coef57	0.000000000	0.000000e+00	0.239209847	0.0023515075	0.1062210518
ts.12	coef58	0.000000000	-1.898587e-01	0.000000000	-0.3137936921	0.0000000000
ts.18	coef59	-0.030253673	0.000000e+00	0.000000000	-0.4776006731	-0.1314286143
ts.24	coef60	-0.006200648	5.538524e-03	0.000000000	0.0000000000	0.0000000000
tst2.6	coef61	0.000000000	0.000000e+00	0.000000000	0.0004713838	-0.0015810817
tst2.12	coef62	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
tst2.18	coef63	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
tst2.24	coef64	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
u10.6	coef65	0.019998049	1.168859e-01	0.000000000	0.0000000000	-0.0463784377
u10.12	coef66	0.000000000	-3.657322e-02	0.000000000	0.0000000000	0.0471091487
u10.18	coef67	0.018694885	0.000000e+00	0.000000000	0.0529655775	0.0000000000
u10.24	coef68	0.000000000	6.920541e-02	-0.111740114	0.0000000000	0.0005562355
v10.6	coef69	0.014934374	0.000000e+00	-0.037289411	0.0000000000	-0.0224367883
v10.12	coef70	0.000000000	-1.430183e-01	-0.114420372	0.0000000000	0.0000000000
v10.18	coef71	0.000000000	0.000000e+00	-0.052336118	0.0134666389	0.0000000000
v10.24	coef72	0.000000000	0.000000e+00	-0.010086938	0.0000000000	-0.0292203898
Q2.6	coef73	0.000000000	1.327614e-01	0.000000000	0.0000000000	-0.0527531071
Q2.12	coef74	0.000000000	2.889095e-01	-0.248191291	0.0000000000	0.0000000000
Q2.18	coef75	0.000000000	0.000000e+00	0.000000000	0.1216926962	0.0006204782
Q2.24	coef76	0.043134109	-7.731316e-04	0.000000000	0.0438615157	0.0000000000
Us.6	coef77	0.000000000	0.000000e+00	-0.084986934	0.0000000000	0.0145758520
Us.12	coef78	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.1060469039
Us.18	coef79	0.065644210	0.000000e+00	0.000000000	0.1054852938	0.0954889002
Us.24	coef80	0.000000000	-6.436051e-01	0.000000000	0.0630479060	0.0000000000
MOi.6	coef81	0.000000000	0.000000e+00	0.000000000	0.0088742381	0.0000000000
MOi.12	coef82	0.000000000	0.000000e+00	0.005287410	0.0000000000	0.0020925693
MOi.18	coef83	0.000000000	2.867343e-02	0.005159202	0.0000000000	0.0000000000
MOi.24	coef84	-0.004523267	0.000000e+00	0.000000000	-0.0043587712	0.0000000000
PBLH.6	coef85	0.000000000	1.031341e-01	0.000000000	0.0000000000	-0.0114544635
PBLH.12	coef86	-0.022017504	0.000000e+00	0.000000000	0.0000000000	-0.0616826563
PBLH.18	coef87	-0.013316571	0.000000e+00	0.000000000	0.0000000000	0.0000000000
PBLH.24	coef88	0.000000000	0.000000e+00	0.000000000	-0.0429015582	-0.0028068053
PBLT.6	coef89	-0.016466373	1.419827e-01	0.302446873	0.1156645396	0.0250836608
PBLT.12	coef90	0.000000000	2.050778e-01	0.078515748	0.0517637027	-0.0573481044
PBLT.18	coef91	-0.031417496	6.541722e-03	0.000000000	-0.0739804106	-0.0498469742
PBLT.24	coef92	0.037698455	5.072072e-01	0.000000000	0.0000000000	0.0049551995
PBLT.6	coef93	0.027186226	0.000000e+00	0.000000000	0.0880845374	0.0000000000



PBLL.12	coef94	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
PBLL.18	coef95	0.000000000	0.000000e+00	0.000000000	0.0246522507	0.0000000000
PBLL.24	coef96	0.000000000	0.000000e+00	0.084805578	0.0535587853	0.0201642778
W.6	coef97	0.000000000	4.861848e-02	0.000000000	0.0000000000	-0.0623935421
W.12	coef98	0.000000000	0.000000e+00	-0.092742007	0.0200102310	0.0000000000
W.18	coef99	0.000000000	0.000000e+00	0.000000000	0.1177669088	0.0000000000
W.24	coef100	0.016803558	0.000000e+00	0.000000000	0.0475838218	0.0000000000
Pasq.6	coef101	0.007324313	0.000000e+00	0.000000000	-0.0749103317	0.0047938365
Pasq.12	coef102	0.012509235	-2.109844e-02	0.000000000	-0.0674077858	0.0252960008
Pasq.18	coef103	0.028599392	0.000000e+00	0.000000000	0.0478092024	0.0236945906
Pasq.24	coef104	0.000000000	-1.890928e-01	-0.012583981	0.0687183488	0.0342083471
G20.6	coef105	0.012638397	0.000000e+00	-0.018971539	0.1383993551	0.0000000000
G20.12	coef106	-0.009606437	0.000000e+00	0.000000000	0.0000000000	0.0000000000
G20.18	coef107	0.000000000	-5.955500e-02	-0.063192746	0.0000000000	0.0000000000
G20.24	coef108	0.000000000	-5.214465e-02	-0.048745219	0.0000000000	0.0000000000
G50.6	coef109	0.000000000	-4.238070e-02	0.054972372	-0.0224475136	0.0307329320
G50.12	coef110	0.000000000	-6.334903e-02	0.019832814	0.0506459933	0.0000000000
G50.18	coef111	0.000000000	0.000000e+00	0.037433774	0.0000000000	-0.0254195333
G50.24	coef112	-0.025263469	-6.971288e-02	0.000000000	0.0000000000	-0.033898699
G120.6	coef113	0.000000000	0.000000e+00	0.000000000	0.0535780329	0.0000000000
G120.12	coef114	0.000000000	9.662234e-02	0.000000000	0.0428673330	0.0000000000
G120.18	coef115	0.000000000	1.748303e-02	0.000000000	0.0180826042	0.0167613933
G120.24	coef116	0.000000000	1.281062e-02	-0.007309878	-0.0899492122	0.0183185045
G200.6	coef117	0.030371709	0.000000e+00	-0.010292505	-0.0163103856	0.0526136892
G200.12	coef118	0.000000000	-1.493245e-02	-0.021874666	0.0271377014	0.0149430194
G200.18	coef119	0.000000000	0.000000e+00	0.000000000	0.0713682101	0.0038952050
G200.24	coef120	-0.008154209	-6.172984e-02	0.000000000	0.0952702554	0.0438944789
G300.6	coef121	-0.011064822	-9.520958e-02	0.000000000	0.0000000000	-0.0392621332
G300.12	coef122	0.000000000	0.000000e+00	-0.040942673	0.0157670973	-0.0177086237
G300.18	coef123	-0.047402507	0.000000e+00	-0.048709458	0.0000000000	0.0340923633
G300.24	coef124	0.000000000	-6.892620e-02	-0.036947382	0.0314772439	-0.0082909209
G500.6	coef125	0.000000000	-1.989024e-01	0.000000000	0.0511770501	0.0000000000
G500.12	coef126	0.000000000	2.968698e-02	0.000000000	0.0432590547	0.0000000000
G500.18	coef127	0.022805802	0.000000e+00	0.134242547	0.0106364272	0.0020708000
G500.24	coef128	0.032847718	0.000000e+00	0.010549103	-0.0429098728	-0.0331513015
G800.6	coef129	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0149512826
G800.12	coef130	0.000000000	0.000000e+00	0.000000000	0.0675772066	0.0000000000
G800.18	coef131	0.000000000	-6.774302e-02	0.000000000	0.0000000000	0.0000000000
G800.24	coef132	0.000000000	6.351414e-02	0.000000000	0.0998537508	0.0488359443
RAIN.6	coef133	-0.010603498	1.152388e-02	0.005024449	0.0000000000	0.0000000000
RAIN.12	coef134	0.000000000	0.000000e+00	0.001957564	0.0000000000	-0.0043567636
RAIN.18	coef135	0.000000000	0.000000e+00	-0.062482829	-0.0527414170	0.0000000000
RAIN.24	coef136	-0.007242985	0.000000e+00	0.000000000	0.0000000000	0.0000000000
RAINNC.6	coef137	-0.034072797	0.000000e+00	0.000000000	0.0000000000	0.0000000000
RAINNC.12	coef138	-0.019206332	0.000000e+00	0.000000000	0.0000000000	-0.0116647277
RAINNC.18	coef139	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
RAINNC.24	coef140	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
NO2.P.48	coef141	0.000000000	-7.911446e-02	-0.254272365	-0.1184760110	0.0000000000
NO2.P.52	coef142	0.000000000	6.422333e-02	0.000000000	0.0000000000	0.0416930341
NO2.P.56	coef143	0.000000000	0.000000e+00	0.018555380	0.0654765015	0.0698154844



NO2.P.60	coef144	0.109490417	0.000000e+00	0.012392805	-0.1477352458	0.0000000000
NO2.P.64	coef145	-0.004303657	1.418761e-01	-0.170389385	0.0000000000	-0.0118882410
NO2.P.68	coef146	0.000000000	3.620768e-02	0.000000000	0.0691378179	0.0000000000
NO2.P.72	coef147	-0.006528008	1.569382e-01	-0.031370367	-0.1978524104	-0.0247310423
NO2.H.48	coef148	0.000000000	0.000000e+00	0.276527895	0.1363387415	-0.0508349870
NO2.H.52	coef149	0.000000000	0.000000e+00	0.000000000	0.0000000000	0.0000000000
NO2.H.56	coef150	0.000000000	2.908832e-03	0.000000000	-0.0469595219	0.0000000000
NO2.H.60	coef151	-0.037097866	0.000000e+00	0.000000000	0.0000000000	0.0077396606
NO2.H.64	coef152	0.000000000	0.000000e+00	0.114863179	0.0082892334	0.0000000000
NO2.H.68	coef153	0.000000000	0.000000e+00	0.102362437	0.0000000000	0.0000000000
NO2.H.72	coef154	-0.015474939	1.712271e-02	0.000000000	0.1460981368	0.0000000000
O3.P.48	coef155	0.000000000	-5.626934e-02	-0.366012649	-0.1718591572	0.0000000000
O3.P.52	coef156	-0.023841011	0.000000e+00	-0.059528164	-0.0034195982	0.0000000000
O3.P.56	coef157	0.000000000	-1.357864e-01	0.000000000	0.0000000000	0.0993820952
O3.P.60	coef158	0.035951703	0.000000e+00	0.000000000	-0.0584834366	0.0000000000
O3.P.64	coef159	0.000000000	3.094197e-02	0.000000000	0.0036775570	-0.0241631974
O3.P.68	coef160	0.000000000	0.000000e+00	-0.183337712	0.0000000000	-0.1154098363
O3.P.72	coef161	-0.035899687	0.000000e+00	0.082932952	0.0393573336	0.0000000000
dispersion	sigma	0.190027788	6.168198e-01	0.396025509	0.3151685326	0.2176365432

13 References

- Abdul-Wahab, S.A. and Al-Alawi, S.M. (2002). Assessment and Prediction of Tropospheric Ozone Concentration Levels Using Artificial Neural Networks. *Environ. Modell. Softw.* vol. 17, 219-228.
- Antanasijevic, D.Z., Pocajt, V.V., Povrenovic, D.S., Ristic, M.D. and Peric-Grujic, A.A. (2013). PM10 emission Forecast Using Artificial Neural Networks and Genetic Algorithm Input Variable Optimization. *Sci. Total Environ.* vol. 443, 511-519.
- Bertaccini, P., Dukic, V. and Ignaccolo, R. (2012). Modeling the short-term effect of traffic and meteorology on air pollution in Turin with Generalized Additive Models. *Advances in Meteorology*, vol. 2012, Article ID 609328, doi:10.1155/2012/609328
- Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009), mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1-29.
- Brownlee, J. (2021). A Gentle Introduction to Mixture of Experts Ensembles. <https://machinelearningmastery.com/mixture-of-experts/>
- Brownlee, J. (2021a). Ensemble learning algorithms with python make better predictions with bagging, boosting, and stacking. *Machine Learning Mastery EBook Catalog*.
- Carvalho, A.X. and Tanner, M. A. (2005). Modeling nonlinear time series with local mixtures of generalized linear models. *The Canadian Journal of Statistics*, vol. 33(1), 97-113.
- Chen, D., Xu, T., Li, Y., Zhou, Y., Lang, J., Liu, X. and Shi, H. (2015). A Hybrid Approach to Forecast Air Quality during High-PM Concentration Pollution Period. *Aerosol and Air Quality Research*, vol. 15, 1325-1337.
- De Gooijer, J. G. (2017). *Elements of Nonlinear Time Series Analysis and Forecasting*. Springer.
- Gormley, I. C., and Murphy, T. B. (2018). MEclustnet: fitting the mixture of experts latent position cluster model. R package version 1.0.
- Grivas, G. and Chaloulakou, A. (2006). Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece. *Atmospheric Environment*, vol. 40(7), 1216-1229.
- Hastie, T., Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: prediction, inference and data mining*. Springer.
- Hyndman, R.J., and Athanasopoulos, G. (2021) *Forecasting: principles and practice*, 3rd edition, OTexts: Melbourne, Australia, OTexts.com/fpp3
- Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991). Adaptive mixtures of local experts. *Neural Computation*, vol. 3, 79-87.
- Luna, A. S., Paredes, M. L. L., de Oliveira, G. C. G. and Correa, S. M. (2014). Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil, *Atmospheric Environment* 98, 98 - 104
- Mazza, A., Punzo, A., and Ingrassia, S. (2017). flexCWM: Flexible Cluster-Weighted Modeling. R package version 1.7.
- McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, NewYork.
- Murphy, K. and Murphy, T. B. (2020). Gaussian parsimonious clustering models with covariates and a noise component. *Advances in Data Analysis & Classification*, 14(2): 293-325.

- Neykova, N. and Neytchev, P. (2019). Forecasting daily maximum ground-level ozone concentrations using stochastic models. In: AIP Conference Proceedings 2075, Mishonov, T. M. and Varonov, A. M. (eds.), 120008; <https://doi.org/10.1063/1.5091266>, AIP Publishing
- Neykova, N. and Neytchev, P. N. (2020). Missing data imputation for hourly PM₁₀ concentrations in Sofia stations. Bulg. J. Meteo. & Hydro., vol. 24 (1), 18-39.
- Ordieres, J.B., Vergara E.P., Capuz R.S., Salazar R.E. (2005). Neural network prediction model for fine particulate matter (PM 2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juarez (Chihuahua). Environmental Modelling and Software, 20 (5), pp. 547-559.
- Paschalidou, P. A., Kassomenos, P.A. and Bartzokas, A. (2009). A comparative study on various statistical predicting ozone concentrations implications to environmental management. Environ. Monit. Assess, vol. 148, 277-289.
- Pearce, J.L., Beringer, J., Nicholls, N., Hyndman, R.J., Uotila, P. and Tapper, N. J. (2011). Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modelling, Atmospheric Environment, vol. 45(1), Pages 128-136.
- Perez, P., Trier, A., Reyes, J. (2000). Prediction of PM_{2.5} concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, pp. 1189-1196.
- R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Russo, A., Lind, P.G., Raischel, F., Trigo, R. and Mendes, M. (2015). Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research, vol. 6, 540-549.
- Seni, G. and Elder, J. (2010). Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions (Synthesis Lectures on Data Mining and Knowledge Discovery)
- Song, Y.Z, Yang, H.L, Peng, J.H, Song, Y.R, Sun, Q. and Li, Y. (2015) Estimating PM_{2.5} Concentrations in Xi'an City Using a Generalized Additive Model with Multi-Source Monitoring Data. PLoS ONE 10(11): e0142149. doi:10.1371/journal.pone.0142149
- Städler, N., Bühlmann, P. and van de Geer, S. (2010). L1-penalization for mixture regression models. Test, vol. 19, 209-256.
- Turner, R. (2014). mixreg: functions to fit mixtures of regressions. R package version 0.0-5.
- Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S. and Kukkonen, J. (2011). Evaluation of a Multiple Regression Model for the Forecast of the Concentrations of NO_x and PM₁₀ in Athens and Helsinki. Sci. Total Environ. Vol. 409, 1559-1571.
- Wood, S. (2017). Generalized additive models. An introduction with R. 2nd ed. CRC press.
- Zhou, Z-H. (2012). Ensemble Methods: Foundations and Algorithms (Chapman & Hall/CRC Machine Learning & Pattern Recognition).