Modeling and Forecasting of Air Quality Impact on Mortality Rates - A Case Study

THEODOR D. POPESCU
National Institute for Research and Development in Informatics
8-10 Averescu Avenue, 011455 Bucharest
ROMANIA
Theodor.Popescu@ici.ro

Abstract: The paper has as subject the modeling and forecasting of air quality impact on mortality rates, and present a case study, making use of time series analysis approach. After a general view on the time series models, regression and intervention models, to be used in modeling and forecasting of mortality, function of air quality, some methodological aspects of time series modeling and forecasting, based on Box-Jenkins methodology, are discussed with the emphasis on practical aspects. Finally, a case study using a multiplicative transfer function model with three exogenous variable representing ozone, daily average computed for the region, particulate matter 10 micrometers or less in diameter, daily average, and temperature mean, daily average, considered as risk factors, with the effect on mortality rate, as endogenous variable, is presented.

Key–Words: Time series analysis, Modeling, Forecasting, Box-Jenkins approach, Transfer function model, Air quality, Mortality, Case study.

1 Introduction

Air quality, weather and climate, and human health are closely linked. These interdependencies are becoming ever more evident and health professionals ever more reliant on meteorological and climate services to help anticipate and manage the health risks of poor air quality, [1]. Over the last century, poor air quality has become a critical environmental, economic, and health problem around the world as industrial growth and economic development have caused massive increases in air pollutants. The World Health Organization (WHO) has released alarming data on the impact of ambient (surrounding outdoor) air quality and climatic changes on human health, [2], among other reports. For such analysis it is important to have accurate information on the concentration-response relationships for the effects investigated, for example on the relationship between changes in daily air pollution and its impact on health.

Air pollution is defined as a phenomenon harmful to the ecological system and the normal conditions of human existence and development when some substances in the atmosphere exceed a certain concentration, with great effect on morbidity and mortality. Nitrogen oxides, ozone, volatile organic compounds, sulphur dioxide and particulate matter (PM) are accruing in our atmosphere, especially due to inefficiencies in transportation, energy production, energy use and industry. Chemical components and pollutants emitted into the atmosphere undergo chemical transformations and get transported far and wide, depending on the climate and weather. As a result, air pollution is now the worlds largest single environmental health risk, [3].

Time series studies of particulate matter and mortality and morbidity have provided evidence that daily variation in air pollution levels is associated with daily variation in mortality counts. These findings served as key epidemiological evidence for the recent review of the ambient air quality standards for particulate matter. As a result, methodological issues concerning time series analysis of the relationship between air pollution and health have attracted the attention of the scientific community and critics have raised concerns about the adequacy of current model formulations. Time series data on pollution and mortality are generally analyzed by using log-linear, Poisson regression models for overdispersed counts with the daily number of deaths as outcome, the (possibly lagged) daily level of pollution as a linear predictor and smooth functions of weather variables and calendar time used to adjust for time-varying confounders, [4]

In the face of increasingly serious environmental pollution problems, scholars have conducted a significant quantity of related research, and in those studies, the modeling and forecasting of their effects on mor-

bidity and mortality have been of paramount importance. Extensive research indicates that the methods in this field can be broadly divided in statistical forecasting methods, artificial intelligence methods, and numerical forecasting methods. Also, recently, some hybrid models have been proposed, to improve the modeling quality and forecast accuracy. The time series modeling and forecasting have been widely used in this field. Many studies provide a clear perspective on air pollution, and climatic changes effects on morbidity and mortality. So, [5] gives an overview of time series ideas and methods used in public health and biomedical research, with examples in public health including daily ozone concentrations, weekly admissions, in a health department in the U.S. The time series models are most commonly used in regression analysis to describe the dependence of the response at each time on predictor variables including covariances and possibly previous values in the series. In [6] are reviewed the history, methods, and findings of the time-series studies estimating health risks associated with short-term exposure to particulate matter (PM), though much of the discussion is applicable to epidemiological studies of air pollution in general. Time series methods are necessary to make valid inferences from data by accounting for the correlation among repeated responses over time. Estimation of health effects (morbidity and mortality) attributed to PM10 and PM2.5 exposure using an Air Quality model in Bukan city, from 2015-2016, is given in [7]. A multi-model (HTAP2), to estimate the premature human mortality due to intercontinental transport of air pollution and emission sectors, taking into account six source regions, three global emission sectors, power and industry, ground transportation, and residential, and one global domain, using an ensemble of global chemical transport model simulations coordinated by the second phase of the Task Force on Hemispheric Transport of Air Pollutants (TF HTAP2), and epidemiologically derived concentration response functions is given in [8].

The present paper aims to provide a case study on modeling and forecasting of air quality impact on mortality rates using time series analysis, for a data set provided by Kaggle, [9].

The paper is organized as follows. In Section II is given a general view on the time series models, regression and intervention models, to be used in modeling and forecasting of mortality, function of air quality. Section III discusses some methodological aspects of time series modeling and forecasting, based on Box-Jenkins methodology, with the emphasis on practical aspects. Section IV presents a case study using a multiplicative transfer function model with three exogenous variable representing ozone, daily average computed, particulate matter 10 micrometers or less in diameter, and temperature mean, daily averaged, considered as risk factors, with the effect on mortality rate, as endogenous variable.

2 Time series models

The statistical approaches adopted in time series modeling and forecasting usually rely on multiplicative $SARIMA$ (Seasonal Auto Regressive Integrated Moving Average) model. A such model has the following form for the time series z_t, [10]:

$$\phi(B)\Phi(B^s) \bigtriangledown^d \bigtriangledown_s^D z_t = \theta(B)\Theta(B^s)a_t \qquad (1)$$

where a_t is a white noise and

$$\phi(B) = 1 + \phi_1 B + \phi_2 B^2 + \cdots + \phi_p B^p;$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q;$$

$$\Phi(B^s) = 1 + \Phi_s B^s + \Phi_{2s} B^{2s} + \ldots + \Phi_{Ps} B^{Ps};$$

$$\Theta(B^s) = 1 + \Theta_s B^s + \Theta_{2s} B^{2s} + \ldots + \Theta_{Qs} B^{Qs};$$

with B the time delay operator, $Bz_t = z_{t-1}$, $\bigtriangledown z_t = (1 - B) = z_t - z_{t-1}$, nonseasonal differentiating operator, and $\bigtriangledown_s z_t = (1 - B^s) = z_t - z_{t-s}$, seasonal differentiating operator: d is the nonseasonal differentiating order, D is the seasonal differentiating order and s is the seasonal period of the series.

The model is defined as $SARIMA(p, d, q)(P, D, Q)_s$ where (p, d, q) denotes nonseasonal orders, and (P, D, Q) seasonal order of the model. The model is presented in Fig. 1.

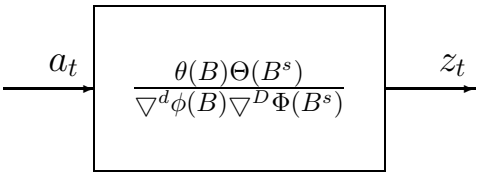

Figure 1: Multiplicative $SARIMA$ model

The multiplicative form of the model simplifies the stationarity and invertibility conditions checking; these conditions can be separately checked, for seasonal and nonseasonal coefficients of the model.

Starting from the general model form of the model $SARIMA$ it can be obtain related models: AR (Auto Regressive), MA (Moving Average), $ARMA$ (Auto Regressive Moving Average) and $ARIMA$

(Auto Regressive Integrated Moving Average), with or without seasonal components. These models are identified by the mean of the autocorrelation (ACF) and the partial autocorrelation functions ($PACF$).

In some situations, it is known that some external events can affect the variables for which the practitioner intends to forecast the future time series values. Dynamic models, used in this case, include several variables, as input variables, which are intended to take into account in the dynamics model, the mentioned exception events. A special kind of $SARIMA$ model with input series is called an intervention model or interrupted time series (ITS) model, [11]. In an intervention model, the input series is an indicator variable that contains discrete values that flag the occurrence of an event affecting the response series. This event is an intervention in or an interruption of the normal evolution of the response time series, which, in the absence of the intervention, is usually assumed to be a pure $SARIMA$ process. As examples of practical interventions can be mentioned: the effect of different promotions activities on the sales, the effect of strikes on the volume of the products and the price of the products, the effect of medication on the health of the patient, the effect of the exchange of the laws in the legislation on the mortalities resulting from car accidents, etc. In this case, some variables as step function, consisting of "zero" values and "unit" values, before and after application respectively change policy, medication, or exchange of laws are included in the model, as an external variable.

A such intervention model can be represented like a transfer function (TF) model (see Fig. 2), where z_t is the value of the endogenous variable at time t, $\mathbf{u}_t = [u_{1t}, \ldots, u_{rt}]^T$ is the vector of exogenous variables, and a_t is a white noise error.

$$\Omega_i(B) = \omega_{i0} + \omega_{i1}B + \omega_{i2}B^2 + \cdots + \omega_{in_i}B^{n_i}; \quad i = 1, 2, \ldots, r$$

$$\Delta_i(B) = 1 + \delta_{i1}B + \theta_{i2}B^2 + \cdots + \delta_{in\delta_i}B^{n\delta_i}; \quad i = 1, 2, \ldots, r$$

$\phi(B)$, $\theta(B)$,$\Phi(B^s)$ and $\Theta(B^s)$ have been described above.

3 Methodological Aspects

The time series model construction usually include the following stages, [10]:

- Identification (specification) of the time series model using some data analysis tools (different graphical representations, autocorrelation functions (ACF) and partial autocorrelation functions ($PACF$)) in order to determine the types of transformations to obtain stationarity and to estimate the degree of differentiation needed to induce stationarity in data, as well as the polynomial degrees of autoregressive and moving average operators in the model.

- Model parameter estimation of the time series implies the use of efficient methods (such as maximum likelihood, among others) for parameter estimation, standard errors and their correlations, dispersion of residuals, etc.

- Model evaluation (validation) aims to establish the model suitability, or to make some simplifications in structure and parameter estimates. Key elements for model validation refers to residuals which can not be justified, these being any residuals of abnormal value that can not be explained by the action of known external factors or other variables; also the correlations and partial correlations of the residuals prove useful tools in model evaluation.

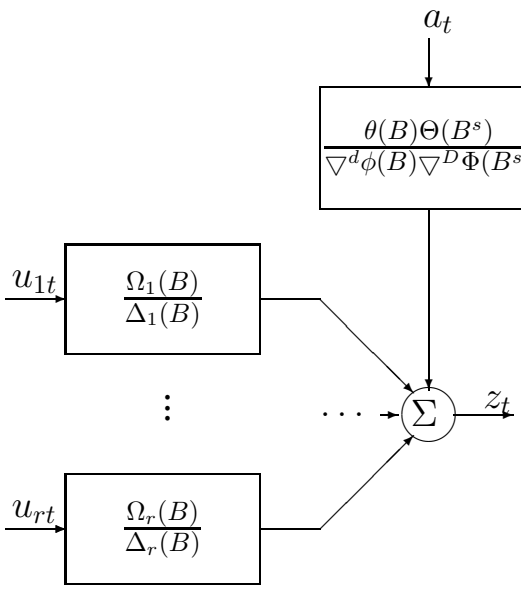

Figure 2: Transfer function (TF) model

More explanations of the process, e.g. [12], often add a preliminary stage of data preparation and a final stage of model application, or forecasting.

Visual analysis of series data allows a first image on the series' non-stationarity and on the presence of a seasonal pattern in the data. The final decision on the inclusion of seasonal elements in the time series

model will be taken after the autocorrelation function (ACF) and partial autocorrelation function ($PACF$) analysis, as well as after the estimation results analysis; the visual analysis of the data can provide useful additional information.

Significant changes in the mean value of the series data require non seasonal differentiation of the first order, while the varying of the rate for average value imposes the nonseasonal differentiation of the second order of the series. Strong seasonal variations usually require, not more than the seasonal differentiation of the first order of the series'data. Autocorrelation function of the series offers information on the nonseasonal and seasonal degrees to be used to obtain the stationarity of the data.

An $ARMA$ stationary process is characterized by theoretical autocorrelation and partial autocorrelation functions tending to zero. The autocorrelation function tends to zero after the first $q-p$ values of the delay, following the evolution of a exponential function or of a damped sinusoidal function, and the partial autocorrelation function is canceled after the first $p-q$ values of the delay, [13].

An AR or MA seasonal process is characterized by similar autocorrelation and partial autocorrelation functions, corresponding to nonseasonal processes, but the coefficients of autocorrelation and partial autocorrelation functions, significant for the seasonal process, appear at multiple seasonal delay values.

At the stage of model identification a special attention will be given to nonseasonal autocorrelation coefficients with absolute values of the associated t statistic test exceeding the value 1.6, [13]. Model parameters, associated to these coefficients prove to be significant from the statistical point of view, in the estimation stage.

In the identification and validation-diagnosis stages, the attention will be focused on the coefficients of seasonal autocorrelations with the absolute values of the t statistic test associated which overcome 1.25 value. The seasonal parameters estimates AR or MA , associated to these coefficients, will appear more significant in the estimation stage. If the residual autocorrelation function has zeros values, from statistical point of view, to seasonal delays: $s, 2s, \ldots$, and to the delays of the form $0.5s, 1.5s$, and in the vicinity of seasonal delays: $s+1, s-1, 2s+1, 2s-1, \ldots$, the same warning level will be used: 1.25. More information on the methodology used in this case can be find in [13] and [14].

In the estimation stage, the use of the initial estimates of the model parameters of the value of 0.1 leads to good results in most cases; better initial estimates for model parameters can be obtained based on the autocorrelation and partial autocorrelation functions, used to determine the structure of the model. In this stage as model parameters will be retain those for which $|t| \geq 2$, [13]. The criteria Akaike Information Criterion (AIC), Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC), [15], Adjusted Root Mean Square Error (ARMSE) and Absolute Mean Percent Error (AMPE), [13], offer information on the parameter estimation quality.

Forecasting is what the whole procedure is designed to accomplish. Once the model has been selected, estimated and checked, it is usually a straight forward task to compute forecasts. The forecasting problem can be solved, in the most direct way, using the multiplicative $ARIMA$ model of the form (1). The description of the model by an infinitely weighted sum of current values and the earlier noise is proving useful, in particular, to estimate the variance of forecasting values, as well as to determine their confidence intervals. Standards and practices for time series forecasting are given in [16].

4 Case Study - Forecasting Impact of Air Quality on Mortality Rates

The time series making the object of the case study represents the mortality rate (number of deaths per 10000 people), of a region in England, function of the ozone, daily average computed for the region, O3, particulate matter 10 micrometers or less in diameter, daily average, PM10, and temperature mean, daily average, in Kelvin degree, T2M, considered aa risk factors, and are provided in a Kaggle competition, [9]; the air quality data used in this competition is freely available from Copernicus Atmosphere Monitoring Service (CAMS) which is managed by European Centre for Medium Range Weather Forecasts (ECMWF).

The data containing each 364 values are graphically presented in Fig. 1, Fig. 2, Fig. 3 and Fig. 4, respectively.

After preliminary analysis of the data and different model structures resulted a transfer model function, with 3 exogenous variables: O3, PM10 and TempK, and Mort as output, of the form:

$$\begin{aligned} \text{Mort}_t &= \frac{\omega_{1,1}+\omega_{1,2}B}{1+\delta_{1,1}B+\delta_{1,2}B^2+\delta_{1,3}B^3}\text{O3}_t + \\ &+ \frac{\omega_{2,1}+\omega_{2,2}B}{1+\delta_{2,1}B+\delta_{2,2}B^2+\delta_{2,3}B^3}\text{PM10}_t + \\ &+ \frac{\omega_{3,1}+\omega_{3,2}B}{1+\delta_{3,1}B+\delta_{3,2}B^2+\delta_{3,3}B^3}\text{T2M}_t + \end{aligned}$$

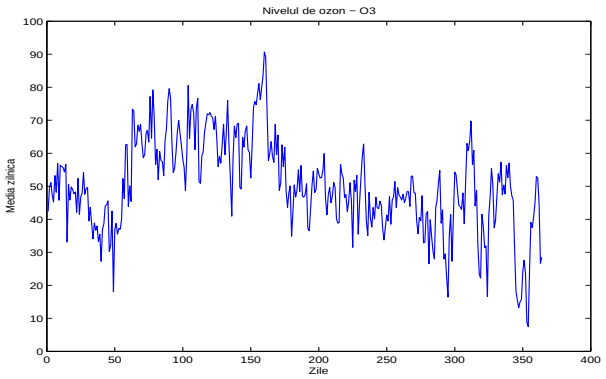

Figure 3: Ozone, daily average computed for the region, O3.

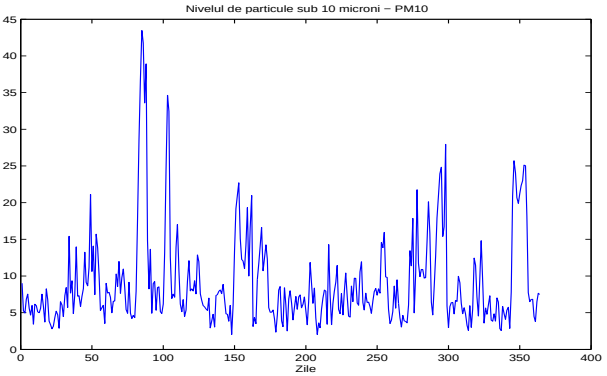

Figure 4: Particulate matter 10 micrometers or less in diameter, daily average, PM10.

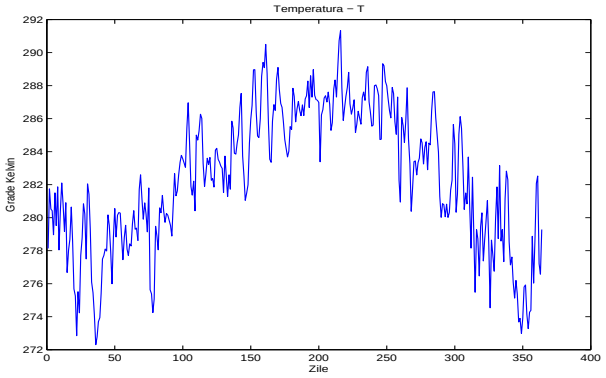

Figure 5: Temperature mean, daily average, in Kelvin degree, T2M.

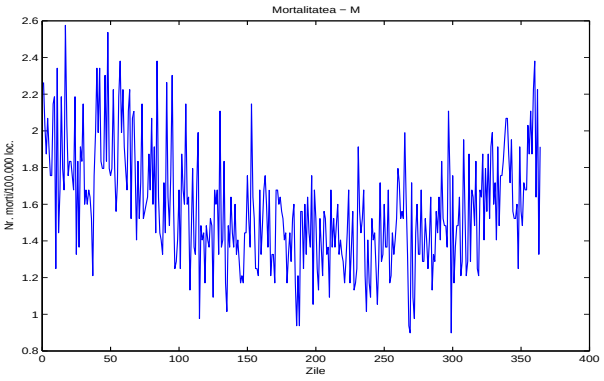

Figure 6: Mortality rate (number of deaths per 10000 people), Mort

$$+ \quad \frac{1+\theta_1 B}{1+\phi_1 B+\phi_2 B^2} a_t; \;\; v[a_t] = \sigma^2. \qquad (2)$$

with $s = 1$, due to the nostationarity of the data. For the model parameters and variance, σ^2, have been used as initial values 0.1. It was used implicit values for the optimization algorithm, excepting the maximum number of iteration, chosen 500. The following final values have been obtained for model parameters (see TABLE I):

Table 1: Final transfer function model parameters

Parameter	Estimate	Std. Dev.
ϕ_1	-1.0906	0.0002
ϕ_2	0.0906	0.0000
θ_1	-0.8683	0.0005
$\omega_{1,1}$	0.0055	0.0001
$w_{1,2}$	0.0001	0.0003
$\omega_{2,1}$	-0.0057	0.0001
$\omega_{2,2}$	0.0071	0.0001
$\omega_{3,1}$	0.0092	0.0001
$\omega_{3,2}$	-0.0097	0.0001
$\delta_{1,1}$	0.0311	0.0003
$\delta_{1,2}$	-0.1780	0.0005
$\delta_{1,3}$	-0.1007	0.0005
$d_{2,1}$	-1.5533	0.0004
$\delta_{2,2}$	1.5105	0.0004
$\delta_{2,3}$	-0.7825	0.0005
$\delta_{3,1}$	-0.8450	0.0005
$\delta_{3,2}$	0.3524	0.0005
$\delta_{3,3}$	0.1559	0.0005
$v_{1,1}$	0.1049	0.0004

with the objective function: 81.8254, nr. of iterations: 14, and information criteria: AIC = 0.5540 and SBC = 0.7574.

The model residuals are presented in Fig. 4, and the autocorrelation function fac partial autoccorelation function $pacf$ are given in Fig. 5. The confirm the model validation.

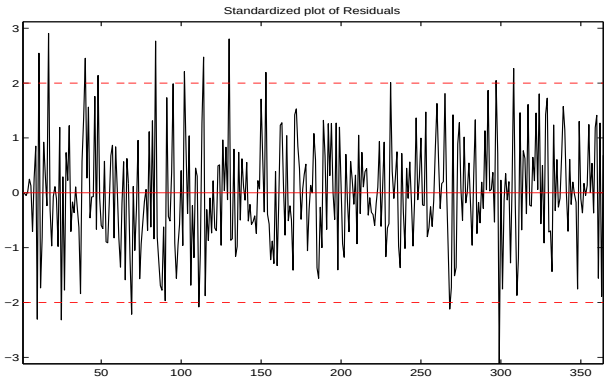

Figure 7: Model residuals

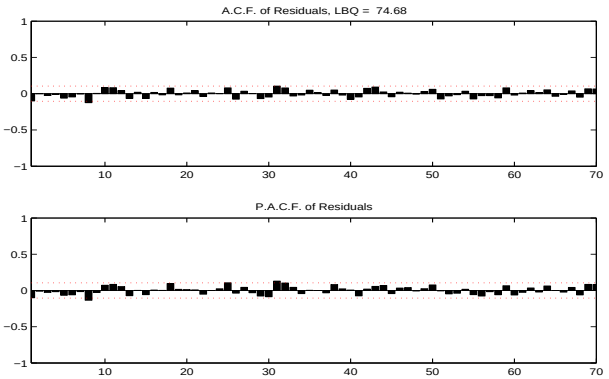

Figure 8: Residuals autocorrelation function, fac, and partial autoccorelation function, $pacf$

The forecasting values for mortality and confidence interval 95%, for a horizon of 14 days are graphical presented in Fig. 6; the numerical values are given in TABLE II.

5 Conclusions

The time series analysis of road traffic accidents using multiplicative $ARIMA$ models and the attractive features of the Box-Jenkins approach provide an adequate description to the data in this field. The $ARIMA$ processes are a very rich class of possible models and it is usually possible to find a process which provides an adequate description to the data. Monthly pattern was the best time process for forecasting. Also, the intervention analysis proved to be a useful approach to model interrupted time series, in this case, when such time series are generated as the

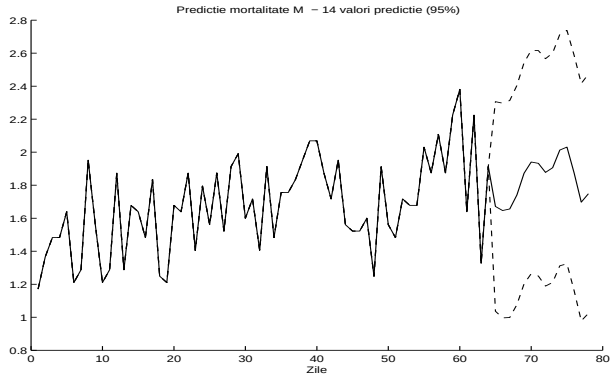

Figure 9: Forecasting values for mortality and confidence interval 95%, for a horizon of 14 days; only the the last 64 values of the original time series and forecasting results are presented.

Table 2: Forecasting values and confidence limits 95%.

Forecasting horizon	Inf. Lim. 95%	Forecasting Value	Upp. Lim. 95%
1	1.0363	1.6711	2.3059
2	0.9962	1.6465	2.2968
3	0.9985	1.6559	2.3133
4	1.0768	1.7407	2.4046
5	1.2021	1.8723	2.5425
6	1.2639	1.9404	2.6169
7	1.2501	1.9328	2.6155
8	1.1887	1.8776	2.5665
9	1.2109	1.9059	2.6009
10	1.3118	2.0128	2.7139
11	1.3238	2.0308	2.7379
12	1.1614	1.8743	2.5873
13	0.9784	1.6973	2.4162
14	1.0235	1.7482	2.4729

of training drivers to obey traffic laws such as using of the seat belt, some economical constraints, etc. The case studies presented in the paper proved the efficiency of the approach. Although originally designed for modeling time series with $ARIMA$ processes, the underlying strategy of Box and Jenkins is applicable to a wide variety of statistical modeling situations. It provides a convenient framework which allows an analyst to think about the data, and to find an appropriate statistical model which can be used to help answer relevant questions about the data.

Acknowledgments

The author thanks the Ministry of Research and Innovation for its support under the 2019-2022 Core Program, Cod PN 19-37, Project 03-01, RO-SmartAgeing.

References:

[1] J. S. Guillemot, L. Jalkanen, H. A. Rohani, *Air Quality and Human Health, a Priority for Joint Action*, Bulletin vol. 63, no. 2, 2014.

[2] H. R. Anderson et al., *Meta-analysis of time-series studies and panel studies of Particulate Matter (PM) and Ozone (O3)*, Report of a WHO task group, 2004.

[3] Qian Di,Yan Wang, Antonella Zanobetti, Yun Wang, Petros Koutrakis, Christine Choirat, Francesca Dominici, Joel D. Schwartz, Air Pollution and Mortality in the Medicare Population, *New England Journal of Medicine*, 376, 2017, pp. 2513-2522.

[4] R. D. Peng, F. Dominici, T. A. Louis, Model choice in time series studies of air pollution and mortality, *Journal of the Royal Statistical Society, Statistics in Society*, Series A, 169, 2006, pp. 179-390.

[5] S. L. Zeger, R. Irizarry, R. D. Peng, On time series analysis of public health and biomedical data, *Annu. Rev. Public Health*, 2006, pp. 5779.

[6] M. L. Bell, J. .M. Samet, F. Dominici, Time-series studies of particulate matter, *Annu. Rev. Public Health*, 25, 2004, pp. 247280.

[7] B. Kamarehie et al., Estimation of health effects (morbidity and mortality) attributed to PM10 and PM2.5 exposure using an Air Quality model in Bukan city, from 2015-2016 exposure using air quality model, *Environmental Health Engineering and Management Journal*, 4, 2017, pp. 137142.

[8] C. K. Liang, et al. HTAP2 multi-model estimates of premature human mortality due to intercontinental transport of air pollution and emission sectors, *Atmos. Chem. Phys.*, 18, 2018, pp. 10497-10520.

[9] * * * www.kaggle.com, *Predict impact of air quality on mortality rates*, UK Office for National Statistics, 2007, https://www.kaggle.com/c/predict-impact-of-air-quality-on-death-rates#

[10] G. E. P. Box, G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 2-nd Edition, Holden Day, San Francisco, 1976.

[11] G. E. P. Box, G. C. Tiao, Intervention Analysis with Applications to Economic and Environmental Problems, *Journal of the American Statistical Association*, 70, 1975, pp. 70-79.

[12] S. Makridakis, S. C. Wheelwright, R. J. Hyndman, *Forecasting: Methods and Applications*, 3-rd Edition, New York: John Wiley & Sons, 1998.

[13] A. Pankratz, *Forecasting with Univariate Box-Jenkins Models*, Wiley, New York, 1983.

[14] P. J. Brockwell, R. A. Davis, *Introduction to Time Series and Forecasting*, Springer-Verlag, New York, 1996.

[15] S. Konishi, G. Kitagawa, *Information Criteria and Statistical Modeling*, Springer, 2008.

[16] J. Scott Armstrong, Standards and practices for forecasting, *Principles of Forecasting: A Handbook for Researchers and Practitioners*, J. Scott Armstrong (ed.), MA: Kluver Academic Publishers, 2001, pp. 1-46.