چهاردهمین کنفرانس بین المللی
مهندسی حمل و نقل و ترافیک
The 14th International
Conference on Traffic and
Transportation Engineering

A Predictive SARIMA Model for PM_{10} and $PM_{2.5}$ levels in Mashhad based on traffic flow and metrological data

Ramin Khavarzadeh[1], Navid Kalantari[2], Neda Alirezaei[3]
1- PhD Candidate, Department of Statistics, Mathematical Faculty,
Tarbiat Modares University.
2- Transportation Planning, PhD - Dept of Civil Engineering.

Abstract

Air pollution is known as a major cause of health and environment damages. Various factors are involved in increasing air pollution, among which road traffic is one of the main sources of these issues. Therefore, in this study, the effects of traffic flow and metrological parameters on the PM_{10} and $PM_{2.5}$ levels are investigated by a predictive model. Time series analysis is used to predict future daily levels of PM_{10} and $PM_{2.5}$ in Mashhad, based on predicted daily traffic flow on the major highways of Mashhad, temperature, wind speed and humidity. The major innovation of this paper is that the air pollution time series is modeled based on another time series (traffic volume). In other words, the time series model for air pollution contains some time series variables (as exogenous variables). These time series variables have some effects on each other, which are considered by Cross-Correlation Function (CCF). For each variable, Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are calculated. ACF defines the seasonal patterns of the observations, and PACF removes dependence of internal lags for each variable. A predictive SARIMA model, which estimates the future levels of PM_{10} and $PM_{2.5}$ is a result of this study. The R-Square of the proposed model is 0.714 and 0.676; and RSME of it is 8.667 and 9.374 for PM_{10} and $PM_{2.5;}$ respectively.

Keywords: ***SARIMA model, air pollution, traffic flow, PM_{10} and $PM_{2.5}$ levels.***

[1] Tarbiat Modares University, Tehran, Iran, Tel: +98(21) 82884424; Email: r.khavarzade@modares.ac.ir.
[2]Avand-e Tarh-o Andisheh Consulting Engineers, Tehran, Iran, Tel: +98(21) 77871287;
Email: kalantari@iust.ac.ir.
[3] Shahid Beheshti University, Tehran, Iran, Email:alirezaei@sharifdata.com.

1. INTRODUCTION

Urban air pollution causes different health damages. Various factors are involved in increasing air pollution, among which road traffic is one of the main sources of these issues, especially in the urban areas. Based on the technical paper of European Topic Center of Air Pollution (ETCAP), urban traffic caused 3 to 56% of the air pollution in Europe in 2012 (*13*). Another study revealed that 27% of the annual Particle Mater (PM) is emitted from vehicles (*3*). Additionally, Ministry of the Environment (MOE) of Ontario, Canada (*8*), has shown that the transportation sector has the highest share in NO_x, CO and PM emission (*7*). PM is one of the most dangerous types of air pollutant; a study suggested that the damage cost of PM inhalations is between $1000 and 9000 million US dollar per year (*7*). The same could be seen in Iran, where 1365 tons (daily average) of pollutants are emitted from mobile sources in Mashhad. Meanwhile, inversion has occurred on average in 270 days, annually. Consequently, studying the effects of the traffic flow on the air pollution has becomes an important topic in the recent years and many different acts has been proposed in different countries.
In order to better understanding the cause and effect relationship in air pollution modeling, time series analysis is used in many studies. Gouveia and Fletcher (2000) used time series analysis to investigate the short-term association between SO_2, NO_2 and O_3 and children morality. In this study, the correlation between changes in the temperature, humidity, air pollution parameters (SO_2, NO_2 and O_3) and children morality was been examined (*1*). In other studies, the short- term effects of air pollution and metrological factors are investigated on the preterm births in London (*6*) and allergic rhinitis in Beijing (16). In 2001, a time series model of NO, NO_2 and O_3 was presented for 19 stations in Paris by Romanowicz et al. (*9*). This time-space analysis demonstrated that on average, the maximum level of NO is measured every 11 days. The Structural Time Series (STS) analysis was also implemented in order to predict the average hourly concentration of NO_2 from a congested urban road (*5*). In this study an air quality model was developed which is parsimonious and computationally simple. The correlation between noise, air pollution and traffic was investigated in New York City based on a time series analysis (*10*). In this study, hourly noise level, average particle and NO, wind speed, wind direction and truck and bus traffic were investigated. The results have shown a strong correlation between noise and traffic, especially at nights. Kaushik and Melwani used a Seasonal Autoregressive Integrated Moving Average (SARIMA) in their study to predicting ambient air quality parameters (SO_2, NO_2, PM) (*4*). In order to investigate the influence of emission sources on the air pollutant concentrations, annual time series of CO and PM_{10} were presented for frequency analysis (*14*). In

this study, the spectrum and cross-spectrum analysis was used to reveal the differences between the influence of local traffic and long-range air pollution.
In the recent years the city of Mashhad has been shut down for many days due to air pollution. This has caused major economic losses in the city. Meanwhile, different traffic management strategies such as odd and even license plate restriction, not only did not reduce the air pollution level, but it was also debated that it might increase health risks. Therefore, it has been well known that predictive tools should be developed in order to forecast the future air pollution conditions and to devise some remedial measures before getting to the critical pollution levels. In this study, future daily levels of PM_{10} and $PM_{2.5}$ in Mashhad is predicted based on predicted daily traffic flow on the major highways, temperature, wind speed and humidity. The major innovation of this paper is that the air pollution time series is modeled based on another time series (traffic volume) using a Partial Autocorrelation Function. In other words, the time series model for air pollution contains some time series variables (as exogenous variables), which their effects on each other are considered, so the accuracy of estimation increases significantly. The paper is structured as follows: in the second section Materials and method are described. Section three is devoted to results, and section four to discussion. Finally section 4 concludes the paper.

2. MATERIAL AND METHOD

2.1. Study Area

Mashhad is the second largest city in Iran. The population of this city was estimated to be 3,069,941in 2011 (*12*). This city is well known in Iran, as the holy shrine of Emam Reza is located in it. More than a billion pilgrims visit the holy shrine, annually. This special feature of Mashhad has caused major traffic and environmental issues in the city. The traffic pattern in this city is also affected significantly by religious holidays, since during these holidays many pilgrims from all over the world travel to Mashhad. Consequently, this city has become one of the most polluted cities of Iran. This issue has attracted many studies to investigate the relationship between traffic and air pollution in this city and to come up with special predictive measure to resolve or manage the environmental issues.

2.2. Data collection

Mashhad has been recently equipped with many automatic data collection instruments. Traffic flow data on seven major highways of Mashhad are being collected automatically, and these data have been used as a variable in this study. The traffic data was provided by Mashhad Traffic and Transportation

Organization. The position of these sensors is shown in Figure 1. The level of PM_{10} and $PM_{2.5}$, humidity, wind speed and temperature data are gathered from Mashhad Metrological Organization. This information shows the traffic state and metrological conditions of Mashhad in each day, from March 2012 until March 2013. Additionally, in order to consider the effects of holidays on the level of PM_{10} and $PM_{2.5}$, another variable is added that indicates the special event in each year. The Time series pattern of PM_{10} and $PM_{2.5}$ are shown in Figure 2.

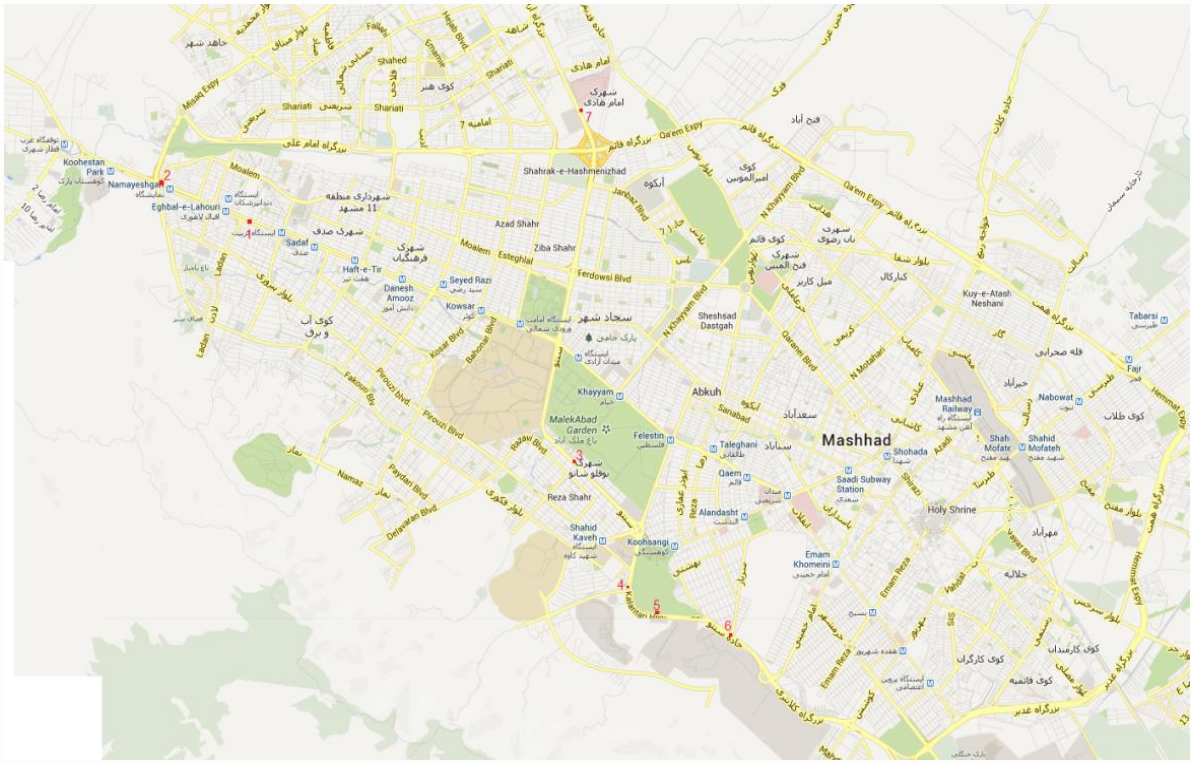
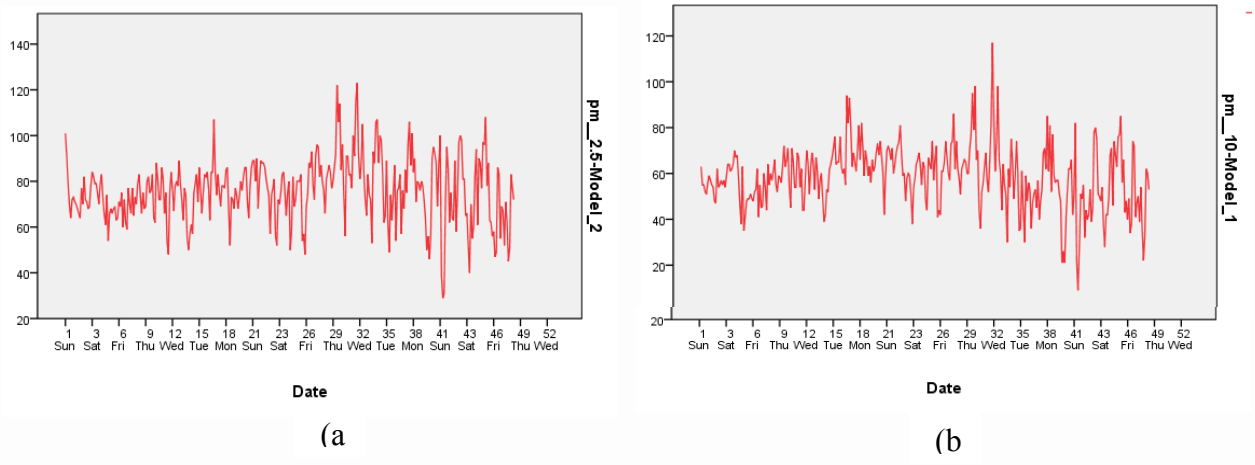


FIGURE 1: *Study area and location of traffic data collection*



(a (b

FIGURE 2: *Time series graphs for (a) $PM_{2.5}$, (b) PM_{10}*

2.3. Time Series analysis

Recently, the development of statistical methods to model and study traffic and environmental phenomenon has gained much interest. Time series are a set of observations, which each one is being recorded at a specific time (*11*). Time series is a powerful tool for statistical analysis when the observations are made at fixed time intervals. Time series function for a consecutive series of data is defined as Eq. 1.

$$y_t = \alpha_0 + \alpha_1 T^1 + \alpha_2 T^2 + \cdots + \alpha_n T^n \quad (1)$$

Where y_t is dependent the variable, α is autocorrelation factor and Ts are time independent variables. Cross Correlation Function (CCF) is used to investigate the effects of different time series on each other. In this case, the effective time interval is defined based on Eq. 2.

$$-(10+\sqrt{n}) \leq K \leq +(10+\sqrt{n}) \quad (2)$$

Where, K is the effective time delay in the downstream time series and n is the number of observations in the time series. To represent real world conditions, Auto Regressive Integrated Moving Average (ARIMA) models could be used, which is one of the most widely used time series patterns. ARIMA is defined by the autoregressive order or p, the degree of differencing or d and moving average parts or q. When seasonal behavior is presented in the time series model, ARIMA is converted to SARIMA (p, d, q)×(P,D,Q). In this model p, d and q are non-seasonal orders and P, D and Q are seasonal orders. In this paper, we have used to SARIMA model framework.

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) are used to investigate the value of seasonal and non-seasonal orders (*2*). In using

ACF and PACF, the values of p and q should be assumed, and then P and Q would be estimated, based on these assumptions. ACF calculates the correlation between the observations in each time period with the ones in previous time periods. This shows the correlation between the pollution in each day with the pollution in previous days. In the PACF the estimated correlation are independent from the days between the ones of interest. In other words, the correlations are estimated only based on the two days of interest and the effect of the other days (the days between the two) are isolated. This could give a more realistic estimate on the level of correlation between two days.

In order to estimate the goodness of fit, Bayesian Information Criterion (BIC) could be used. BIC factor is an independent criterion for choosing the best model. In this method, the goodness of fit is estimated based on the Sum of Square of Errors (SSE) and will be reduced as the number of variables (Degree of Freedom) increase. This is as a model with good prediction power and fewer variables is more favorable in statistical modeling; therefore BIC could be used to evaluate this property (*15*). In this study, in order to develop the time series model, first different models have been estimated then the best model is selected based on Bayesian information criterion (BIC).

3. RESULTS

3.1 Correlation

The first step for estimating the effects of each variable in the model is to calculate the correlation between the variables. In Table 1, the Pearson correlation coefficient and significance level of all the selected variables and PM_{10} and $PM_{2.5}$ are presented. Based on these results, the significance level of the selected variables is less than 0.05. Thus, the correlation between these variables and PM_{10} and $PM_{2.5}$ is significance. The correlations of traffic volume and temperature are positive with PM_{10} and $PM_{2.5}$, while wind speed and humidity have negative correlation with these variables, as expected.

TABLE 1: *Correlation Coefficient and Significance level of variables*

Variable		Pearson Correlation	Sig. level
PM_{10}	PM_{10}	1	-
	$PM_{2.5}$	0.734	0.000
$PM_{2.5}$	PM_{10}	0.734	0.000
	$PM_{2.5}$	1	-
Traffic Volume	PM_{10}	0.101	0.054
	$PM_{2.5}$	0.182	0.000
Temperature	PM_{10}	0.451	0.000
	$PM_{2.5}$	0.157	0.003
Wind Speed	PM_{10}	-0.076	0.146
	$PM_{2.5}$	-0.249	0.003
Humidity	PM_{10}	-0.495	0.000
	$PM_{2.5}$	-0.237	0.000

3.2 Autocorrelation and Partial Autocorrelation

After investigating the effects of each independent variable on the model dependent variables (PM_{10} and $PM_{2.5}$), the variables are tested for autocorrelation. Autocorrelation in time series modeling defines the relationship of observations in time. On other words, the seasonal patterns of the observations on each variable can be examined by the Autocorrelation Function (ACF). Partial Autocorrelation Function (PACF) is an extension of autocorrelation, when the dependence of internal lags is removed. Actually, a more distinct picture of serial dependencies for each lag is provided by partial autocorrelation (*17*). Therefore, in this study, after calculating autocorrelation of each variable, partial autocorrelation is also considered. These results are presented in Figure 3 and Table 2, respectively. As could be seen strong partial autocorrelation exist in the traffic count data. This property suggests the use of a time series model to predict the traffic condition in Mashhad. All other variable do not show any significant partial autocorrelation pattern other than the first lag.

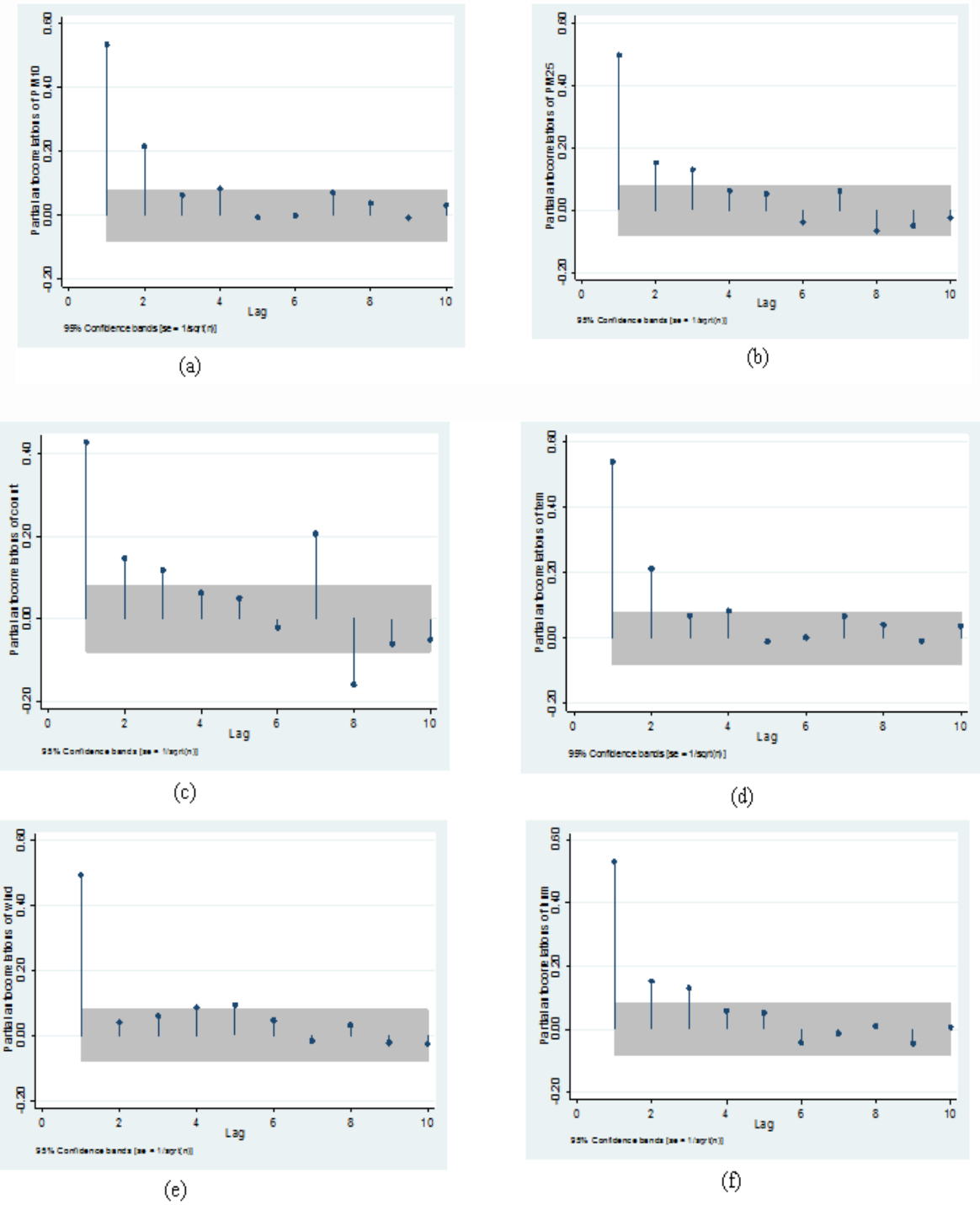


FIGURE 3: *Partial Autocorrelation Function for (a) PM_{10}, (b) $PM_{2.5}$, (c) Traffic volume, (d) Temperature, (e) Wind speed and (f*

TABLE 2: *Autocorrelation Coefficient for each variable*

Lag	Autocorrelation coefficient					
	PM_{10}	$PM_{2.5}$	Traffic Volume	Temperature	Wind Speed	Humidity
1	0.527	.492	.424	.541	.528	.489
2	0.286	.354	.296	.443	.387	.264
3	0.215	.306	.252	.350	.342	.166
4	0.216	.253	.205	.318	.286	.144
5	0.126	.222	.178	.241	.255	.162
6	0.068	.157	.125	.199	.176	.144
7	0.008	.180	.269	.204	.135	.093
8	0.04	.111	.085	.195	.121	.087
9	0.094	.051	.027	.160	.069	.059
10	0.13	.042	.017	.162	.059	.037
11	0.146	.059	.032	.144	.073	.004
12	0.089	.029	.006	.127	.042	-.012
13	0.134	.016	-.001	.101	.021	.005
14	0.183	.088	.186	.103	.034	.020
15	0.15	.033	.016	.080	.038	.008
16	0.135	.019	-.002	.061	.031	-.009

3.3 Cross Correlation

In this study, each variable is considered as a time series variable. In other words, the time series model that is estimated to predict the PMs' level is the summation of other time series. Therefore, considering the effects of these time series on each other is an important point that needs to be addressed in this study, as this is an innovation of this paper. The cross correlation coefficient of the time series variables, for seven days before and after each day (lag of -7 to lag of +7) with PM_{10} and $PM_{2.5}$ is presented in Figure 4 and 5, respectively.

The significance of the correlation between PM_{10} level and traffic volumes of one, two and three days before could be seen in Figure 4. Additionally, it could be seen that PM_{10} levels of 7 sequential days are correlated. Wind speed and PM_{10} have a negative correlation, as the PM_{10} level is reduced when the previous day is windy. Humidity can also reduce the level of PM_{10}, since a significant correlation between PM_{10} and humidity in previous seven days is seen. High levels of humidity could be an effect of precipitation, since rainfall reduces PM_{10} levels. It is worth mentioning that the data regarding the rainy days were not available in this study (was not gathered by the operators). Therefore, humidity was used as a proxy of rainfall in this paper. These patterns are also valid for cross correlation of $PM_{2.5}$ (Figure 5).

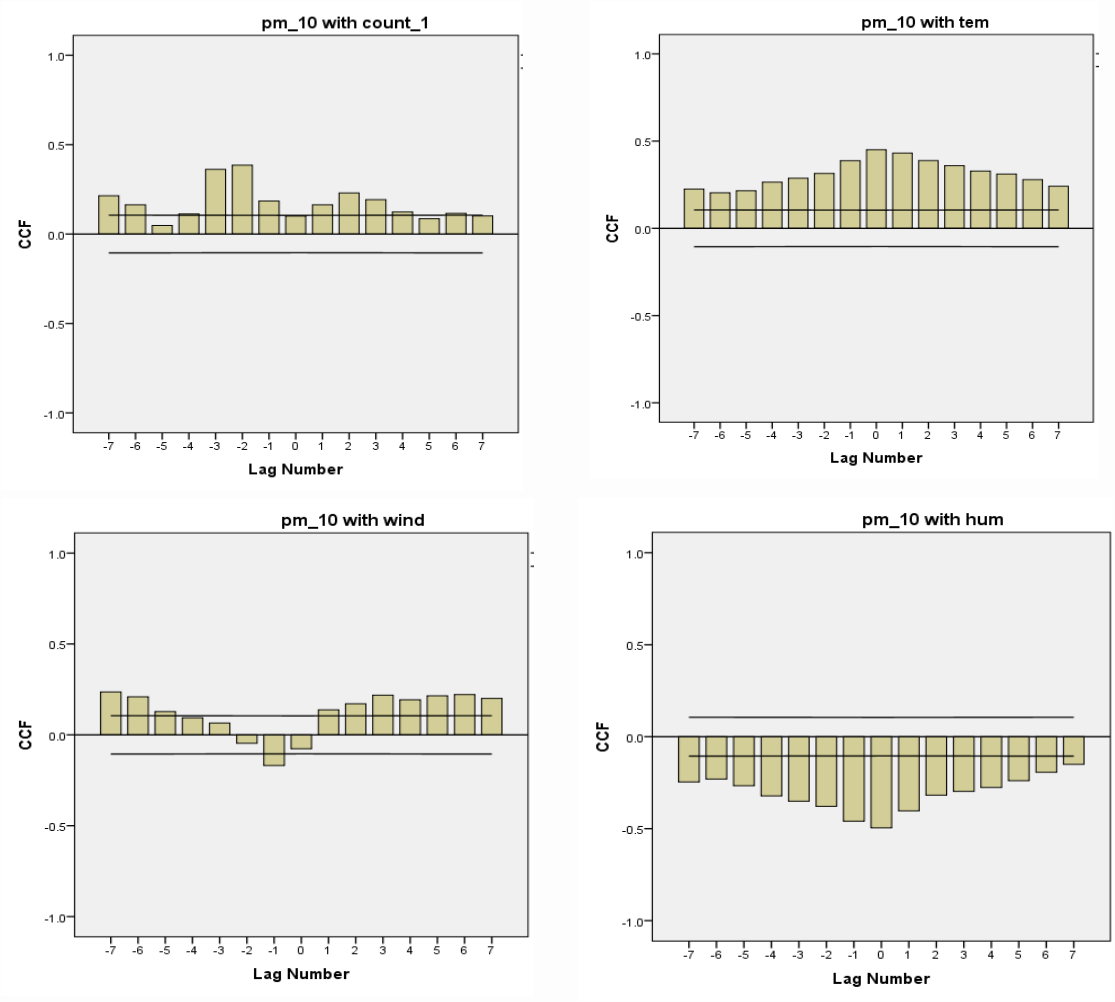


FIGURE 4: *Cross Correlation Function (CCF) for PM_{10} with(a) Traffic Volume, (b) Temperature, (c) Wind Speed and (d) Humidity.*

(a) *(b)*

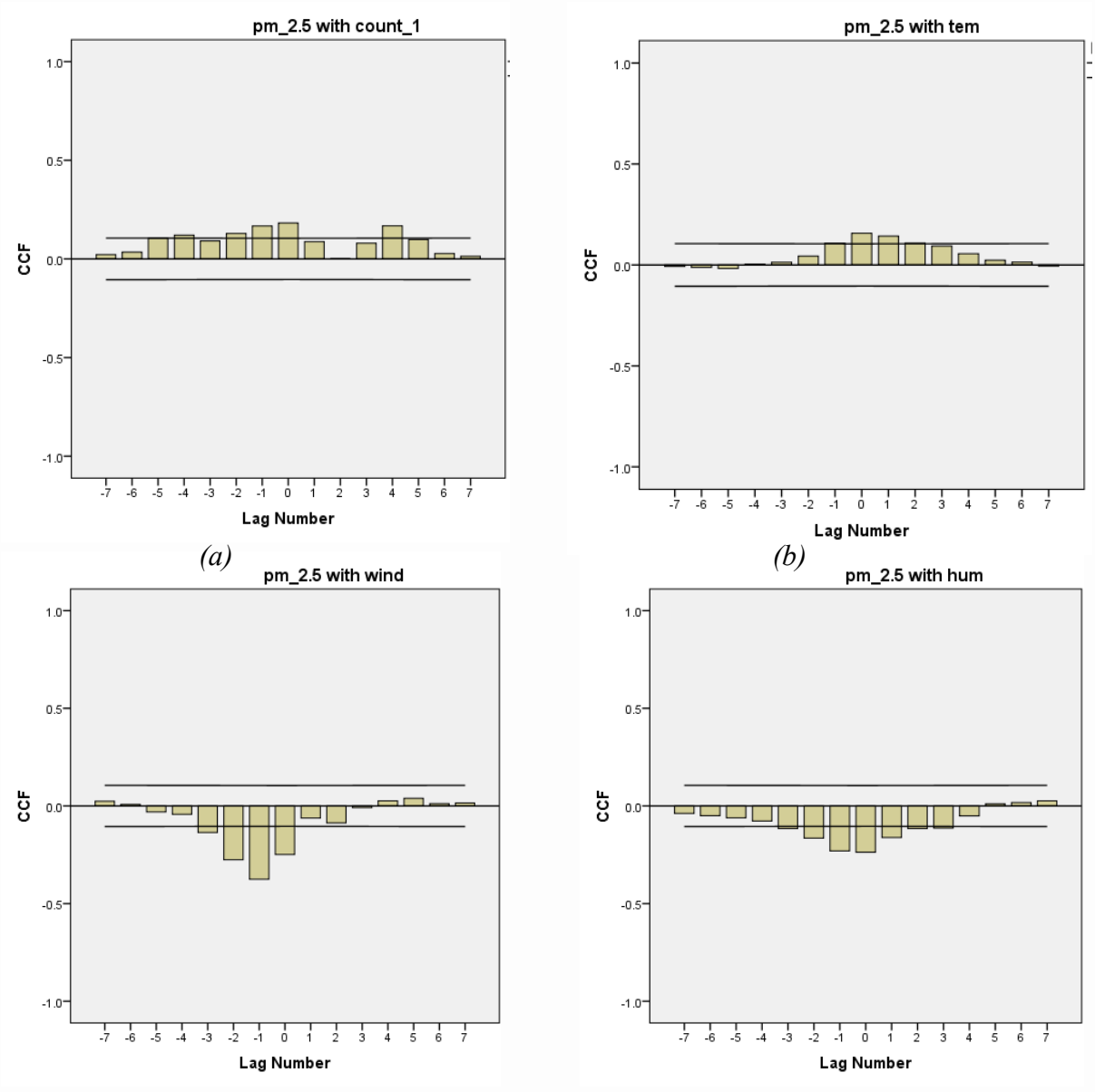


FIGURE 5: *Cross Correlation Function (CCF) for $PM_{2.5}$ with (a) Traffic Volume, (b) Temperature, (c) Wind Speed and (d) Humidity.*

3.4. SARIMA Estimation Results

Based on the previous results, in this section the final model for the prediction of PM_{10} and PM $_{2.5}$ levels would be presented. The SARIMA (1,0,2)(0,0,1)Model parameters for PM_{10} and PM $_{2.5}$ are presented in Table 3 and Table 4, respectively.

TABLE 3: *SARIMA model parameters for predicting PM_{10}*

Variable	SARIMA Parameters For PM_{10}				
		Lag	Estimate	T test value	Significance Level
PM_{10}	Constant	-	82.392	13.939	.000
	AR	lag 1	.897	10.870	.000
	MA	lag 1	.761	7.604	.000
		lag 2	.181	2.687	.008
	MA, Seasonal	Lag 1	.011	2.327	.021
Holidays	Numerator	Lag 0	-5.816	-4.581	.000
	Denominator	Lag 1	.704	7.476	.000
Traffic Volume	Numerator	Lag 0	6.800E-5	2.537	.012
	Denominator	Lag 1	9.140E-5	5.548	.000
Temperature	Numerator	Lag 0	1.223	6.713	.000
	Denominator	Lag 1	.398	2.824	.005
Wind Speed	Numerator	Lag 0	-.971	-7.233	.000
	Denominator	Lag 1	-1.015	-6.430	.000
Humidity	Numerator	Lag 0	-.044	-3.064	.002
	Denominator	Lag 1	-1.484	-17.484	.000

TABLE 4: *SARIMA model parameters for predicting $PM_{2.5}$*

Variable	SARIMA Parameters For $PM_{2.}$				
		Lag	Estimate	T test value	Significance Level
$PM_{2.5}$	Constant		115.643	12.789	.000
	AR	Lag 1	.793	4.813	.000
	MA	Lag 1	.704	8.466	.000
		Lag 2	.571	3.269	.001
	MA, Seasonal	Lag 1	.313	2.253	.025
Holidays	Numerator	Lag 0	-7.513	-5.008	.000
	Denominator	Lag 1	.702	8.031	.000
Traffic Volume	Numerator	Lag 0	6.020E-5	5.381	.000
	Denominator	Lag 1	7.970E-5	11.794	.000
Temperature	Numerator	Lag 0	1.234	5.568	.000
	Denominator	Lag 1	.171	2.220	.027
Wind Speed	Numerator	Lag 0	-1.048	-7.275	.000
	Denominator	Lag 1	-1.157	-8.106	.000
Humidity	Numerator	Lag 0	-.030	-2.181	.030
	Denominator	Lag 1	-1.544	-20.844	.000

The models presented in the previous section, indicates the effect of each variable on the PM_{10} and $PM_{2.5}$ levels. For both of these parameters, humidity in the first lag is the most effective variable. Holidays have shown to reduce the level of PMs. This is while, if the previous day is a holiday, the PMs' levels increase. The temperature of the zero lag had a higher effect compared with the previous day. This shows the instantaneous effect of temperature on PMs' level. On the contrary, the traffic volume of the previous day is more significant than the subject day.

The predictive ability of the proposed model has been tested using the R-square, RMSE, NRMSE and normalized BIC measures. Table 5 presents these values for each proposed model. Stationary R square values are 0.714 and 0.676 for the PM_{10} and $PM_{2.5}$in the proposed SARIMA models. These results confirm the accuracy of the proposed model. In order to validate the proposed models and to test for their predictive ability, a one month data (not used in model estimation) were used. Then these observations were compared with the model results. Figure 6 illustrates the observed versus forecast diagrams of PM10 and PM 2.5 levels for the validation data. As the proposed model uses the output of the traffic volume time series, the results of this time series model is also shown in Figure 7. It is worth mentioning that this study used one year air pollution data, this was as the data on the previous years were not available. This could be known as the major limitation of this paper, and is suggested for future research.

TABLE 5: *Validation results of the proposed model*

Fit Statistics	PM_{10} predicting model	$PM_{2.5}$ predicting model
R-Squared	0.714	0.676
RMSE	8.667	9.374
NRMSE	0.08025	0.09974
Normalized BIC	2.932	3.059

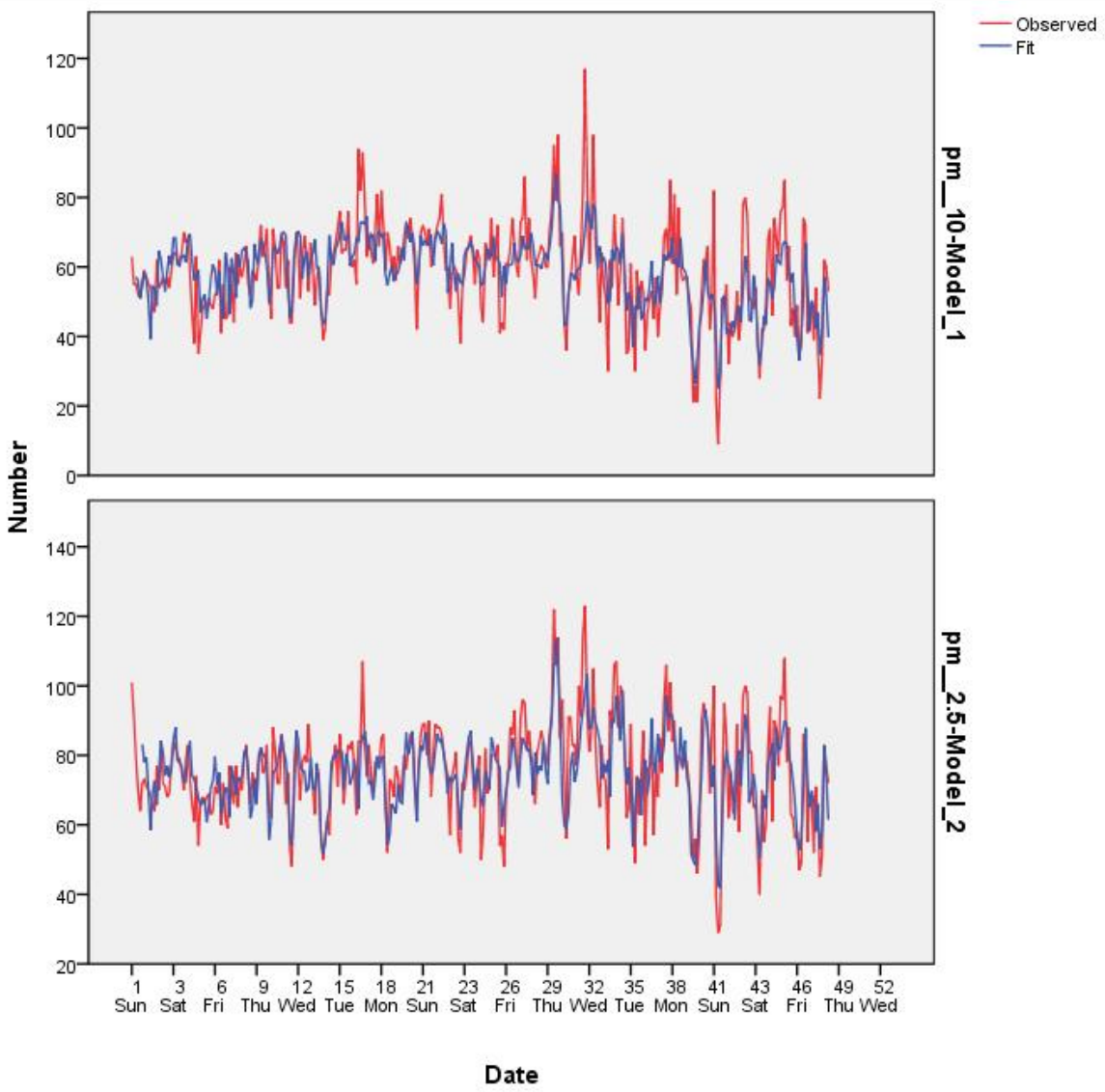


FIGURE 6: *Forecast and observed values of PM_{10} and $PM_{2.5}$ models*

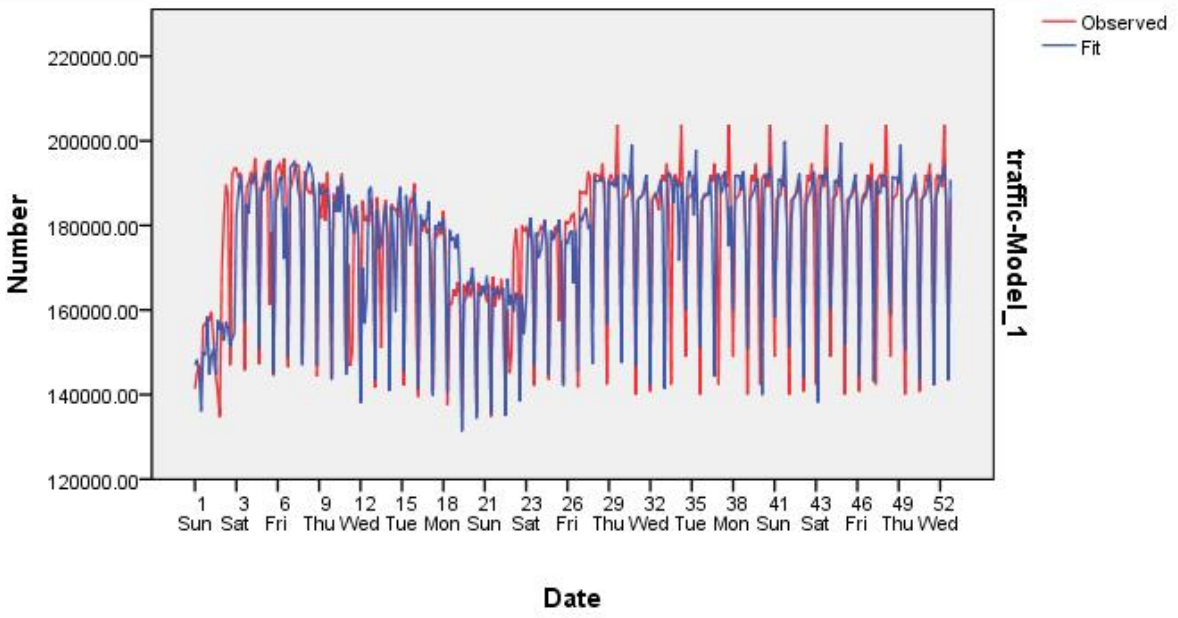


FIGURE 7: *Forecast versus observe traffic counts*

4. CONCLUSION

In this paper, a predictive SARIMA model for estimating future level of PM10 and PM 2.5 in Mashhad is presented. The model variables include: traffic volume, temperature, humidity and wind speed. The main innovation of this research is that some independent variables are considered as a time series variables (traffic volume) in the model.

The correlation between independent variables and PM_{10} and $PM_{2.5}$ is tested. The significance levels of the variables confirm the fact that the selected independent variables have a strong correlation with model parameters. For each variable the Autocorrelation Function (ACF) is calculated. ACF can define the seasonal patterns of the observations for each variable. This step is extended by calculating Partial Autocorrelation Function (PACF), which removes dependence of internal lags. Finally, the Cross-Correlation Function (CCF), which considers the effects of time series variables on each other, was computed. The results confirmed that the proposed model has a good predictive ability. The result of the proposed model has been implemented into a software package in order to predict the future air pollution levels, to put restrictive traffic policies in place.

5. REFERENCES

1. Gouveia N., and Fletcher, T. (2000). "Time series analysis of air pollution and morality: effects by cause, age and socioeconomic status," *Journal of Epidemiology Community Health*, 54, 750-755.
2. Hamilton, J.D. (1994). *Time Series Analysis*, Princeton: University Press.
3. Haohao, K., Ondov, J.M. and Rogge, W.F. (2013). "Detailed emission profiles for on-road vehicles derived from ambient measurements during a windless traffic episode in Baltimore using a multi-model approach," *Atmospheric Environment*, 81, 280-287.
4. Kaushik, I., and Melwani, R. (2007). "Time series analysis of ambient air quality at Ito intersection in Delhi," *Journal of Environmental Research and Development*, 2 (2).
5. Lawson, A.R., Ghosh, B. and Broderick, B. (2011). "Prediction of traffic – related nitrogen oxides concentrations using Structural Time-Series models," *Journal of Atmospheric Environment*, 45, 4719-4727.
6. Lee, S.J., Hajat, S., Steer, P.J. and Filippi, V. (2008). "A time-series analysis of any short-term effects of metrological and air pollution factors on preterm birth in London, UK," *Environmental Research*, 106 (2), 185-194.
7. Lee, Y.J., Lim, Y.W., Yang, J.Y., Kim, C.S., Shin, Y.C. and Shin, D.C. (2011). "Evaluating the PM damage cost due to urban air pollution and vehicle emissions in Seoul, Korea," *Journal of Environmental Management,* 92, 603-609.
8. Ministry of the Environment (Ontario) (2007). "Air Quality in Ontario, 2007 Report."
9. Romanowicz, R., Young, P., Brown, P. and Diggle, P. (2006). "A recursive estimation approach to the spatio-temporal analysis and modeling of air quality data," *Journal of Environmental Modeling and Software*, 21 (6), 759-769.
10. Ross, Z., Kheirbek, I., Clougherty, J.E., Ito, K., Matte, T., Markowitz , S. and Eisl, H. (2011). "Noise, air pollutants and traffic: Continuous measurement and correlation at a high-traffic location in New York City," *Environmental Research*, 111, 1054-1063.
11. Shumway R.H. and Stoffer, D.S. (2011). *Time series analysis and its applications (*3rd *Eition)*, Springer Texts in Statistics.
12. Statistical Center of Iran (2011). "Selected Findings of The 2011 National Population and Housing Census."
13. Sundvor, I., Balagure, N.C., Viana, M., Querol, X., Reche, C., Amato, F., Mellios, G. and Guerreiro, C. (2012). "Road traffic's contribution to air

quality in European cities," *European Topic Centre on Air Pollution and Climate Change Mitigation*.

14. Tchepel, O., and Borrego, C. (2011). "Frequency analysis of air quality time series for traffic related pollutants," *Journal of Environmental Monitoring*, 12, 544-550.
15. Wagenmakers E.J. and Farrel, S. (2004). "AIC model selection using Akaike weights," *Psychonomic bulletin & review*, 11 (1), 192-196.
16. Zhang, F., Wang, W., Lv, J., Krafft, T. and Xu, J. (2011). "Time-series studies on air pollution and daily outpatient visits for allergic rhinitis in Beijing, China," *Science of Total Environment*, 409, 2486-2492.