Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

Using Logistic Regression with Time-Stratified Method for Air Pollution Datasets Forecasting

Sura A. Mohammad and Osama B. Hannon

Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received October 21, 2019
Accepted November 30, 2019
Available online June 1, 2020

Keywords:
Logistic Regression (LR),
Time Stratified (TS),
Particular Matter (PM10),
Forecasting, Air Pollution.

Correspondence:
Sura A. Mohammad
Iraqmosul87@gmail.com

Abstract

Particular matter (PM10) studying and forecasting is necessary to control and reduce the damage of environment and human health. There are many pollutants as sources of air pollution may effect on PM10 variable. Studied datasets have been taken from the Kuala Lumpur meteorological station, Malaysia. Logistic regression (LR) is built by using generalized linear model as a special case of linear statistical methods, therefore it may reflect inaccurate results when used with nonlinear datasets. Time stratified (TS) method in different styles is proposed for satisfying more homogeneity of datasets. It includes ordering similar seasons in different years together to formulate new variable smoother than their original. The results of LR model in this study reflect outperforming for time stratified datasets comparing to full dataset. In conclusion, LR forecasting can be depended after datasets time stratifying to satisfy more accuracy with nonlinear multivariate datasets in which PM10 is to dependent variable.

DOI: 10.33899/IQJOSS.2020.165444 , ©Authors, 2020, College of Computer Science and Mathematics, University of Mosul.

1 INTRODUCTION

The particulate matter of size 10 micrometers can be symbolized as PM_{10}. It is the pollution matter that will be cause air pollution. In this case, it may be have negative effects on human health. Air pollutants has an important effects on human health. Particulate matter with an aerodynamic diameter smaller than or equal 10 µm (PM_{10}) is a measurement of air pollutant includes a mixture of organic and inorganic particles that are suspended in the air (Krzyzanowski *et al.*, 2005). There are many conventional contaminants as pollutants may be caused air pollution with high PM10 value such as the carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), sulphur oxide (SO2) and others. There are several studies about the conventional contaminants and their impacts on human health. Forecasting of air pollution is an important topic in recent years due to the health impact caused by air pollution. Previously, there are many researchers studied air pollution and air quality forecasts. Therefore, it is necessary to forecasting data on air pollution in order to control it and reduce its damage to the environment and human health.Usually, most air pollution datasets are non-linear and this may complicate the process of forecasting and reduce data homogeneity (Bai *et al.*, 2018; Vijayaraghavan & Mohan, 2016; Vong *et al.*, 2012). Logistic regression model (LR) is a special case of linear statistical methods employed for modeling and forecasting any type of multivariate datasets. Therefore, LR may reflect inaccurate results when it used with nonlinear and heterogeneous datasets. LR is used to explain the relationship between two or more independent variables and one dependent binary variable. To improve the results of modeling and forecasting, time stratified (TS) method in different styles is proposed for satisfying more homogeneity of datasets. (Vong *et al.*, 2012). TS method is widely used with time series generally and in meteorological datasets to analyze the short-term effects of risk factors, such as air pollution or temperature, on human health. TS is used to separate the seasonal pattern effects. (Tobias *et al.*, 2014). The data for the same seasons in different years will be stratified timely. Each year has four seasons, each season consists of three months.

2 MATERIAL AND METHOD

This section presents the methods used and suggested for forecasting air pollution based on several meteorological variables. This section begins with a detailed explanation of LR then the concepts of TS method.

2.1 LOGISTIC REGRESSION (LR) MODEL

The multiple linear regression (MLR) model is applicable when the dependent variable is continuous, and is not categorical, while LR is different from the MLR and used when the dependent variable is a binary variable and the independent variables are quantity, categorical variables, or both of them LR (Marill, 2004). Simple and multiple linear regression assumed linear relationship between the independent variables and the dependent variable, while logistic regression does not assume that. It is a useful tool for modeling and forecasting data that insists of binary dependent variable. Nowadays, the researchers state the fact that it is not appropriate to propose multiple linear regression for a binary dependent variable and the better choice institute of multiple linear regression is LR for accurate modeling and forecasting.

Categorical events will coded as binary variables with a value of one represents the positive outcome, or success as target outcome, and a value of zero represents the negative outcome, or failure (Hosmer *et al.*, 1997; Midi *et al.*, 2010; Pohlman & Leitner, 2003). LR is based on probabilities associated with the values of y . For simplicity, and because it is the case most commonly encountered in practice, we assume that y is dichotomous, taking on values of 1 and 0. In theory, the hypothetical, population proportion of cases for which $y=1$ is defined as $p=p(y=1)$. Then, the theoretical proportion of cases for which $y=0$ is $1-p=p(y=0)$. In the absence of other information, we would estimate p by the sample proportion of cases for which $y=1$. However, in the regression context, it is assumed that there is a set of predictor variables, $x_1,\ldots,x_p$, that are related to y and, therefore, provide additional information for predicting y . For theoretical, mathematical reasons, LR is based on a linear model for the natural logarithm of the odds (i.e., the log-odds) in favor of $y=1$: (Dayton, 1992)

$$\log_e\left[\frac{P\left(y=1\mid x_1,x_2,\ldots,x_p\right)}{1-P\left(y=1\mid x_1,x_2,\ldots,x_p\right)}\right]=\log_e\left[\frac{\pi}{1-\pi}\right] \qquad (1)$$

$$=\alpha+\beta_1 x_1+\ldots+\beta_p x_p=\alpha+\sum_{j=1}^{p}\beta_j x_j \qquad (2)$$

Where α is the constant term , β_i are the regression coefficients , x_i are independent variables and π represents the conditional probability of success which takes the form $P\left(y=1\mid x_1,\ldots,x_p\right)$ and depends on combinations of independent variables. $\log_e\left[\frac{\pi}{1-\pi}\right]$ term is the log-odds which defines as the logit transformation of p or the natural logarithmic of odd.

Using a simple exponential transformation that transfers the log-odds to probabilities such as in the following form.

$$\frac{P\left(y=1\mid x_1,x_2,\ldots,x_p\right)}{1-P\left(y=1\mid x_1,x_2,\ldots,x_p\right)}=\frac{\pi}{1-\pi}=e^{\alpha+\beta_1 x_1+\ldots+\beta_p x_p}=e^{\alpha+\sum_{i=1}^{p}\beta_i x_i} \qquad (3)$$

where the probability of success can be such as follows.

$$\pi=P\left(y=1|x_1,x_2,\ldots,x_p\right)=\frac{e^{\alpha+\sum_{i=1}^{p}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{p}\beta_i x_i}}=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}} \qquad (4)$$

and the probability of failure can be such as follows.

$$1-P\left(y=1|x_1,x_2,\ldots,x_p\right)=1-\frac{e^{\alpha+\sum_{i=1}^{p}\beta_i x_i}}{1+e^{\alpha+\sum_{i=1}^{p}}\beta_i x_i}=1-\frac{e^z}{1+e^z}=\frac{1}{1+e^z} \qquad (5)$$

Most rare events applications yield small estimates of $p\left(y_i=1\mid x_i\right)=\pi$ for all observations. However, if the logit model has some explanatory power, the estimate of π among observations for which rare events are observed (i.e., for which $y_i=1$) will usually be larger [and closer to 0.5, because probabilities in rare event studies are normally very small than among observations for which $y_i = 0$. The result is that $\pi(1-\pi)$ will usually be larger for ones than zeros, and so the variance (its inverse) will be smaller. Then the logit model should classify an observation as 1 if the probability is greater than 0.5, The observation is classified as 0 if the probability is less than 0.5. In order to estimate logistic regression coefficients, a higher probability method is used Maximum Likelihood Method is one of the most appropriate methods for all linear and nonlinear models, and the most likely method is a repetitive method based on repetition calculations multiple times, until the best estimate of the transactions is achieved view data can be interpreted. Also, since logistic regression predicts probabilities, rather than just classes, we can fit it using likelihood. For each training data-point, we have a vector of features, x_i ,and an observed class, y_i , the probability of that class was either $\pi(x)$, if $y_i = 1$ or $1-\pi(x)$, if $y_i = 0$.(Santner & Duffy, 1986).

2.2 TIME STRATIFIED METHOD

The TS method is an analytical tool for estimating the effects of the result catalysts from environmental exposure and ensures unbiased logistic regression estimates and avoids biased bias due to the time trend in the exposure chain. The specific layer can be adapted to control the changing time slots according to the design used. Extensive scope in environmental epidemiology to analyze the short-term effects of environmental risk factors, such as air pollution or temperature, and their impact on human health. The day when the health event occurs (the day of the case) is called the day of control chosen from the same month each year. It can be generally applied in other contexts and works to access more homogeneous data from data from the macro data to reach more accurate results (Malig *et al.*, 2015; Tobias *et al.*, 2014).

3 RESULTS AND DISCUSSION

In this section, the results of PM_{10} data forecasting using many forecasting methods discussed in previous section will be displayed with applicable details and discussed. The meteorological datasets in Malaysia for three years (2013-2015), will be studied as influential data on PM_{10} data. The seasonal patterns are detected in the studied variables. Therefore, time-stratified method will be used to handle the influences of these patterns on the accuracy of forecasting. In the seasonal pattern, four groups will be determined according to the nature of seasons.

3.1 PM_{10} data

In this study, A large part of the total data will be used for training and the remaining will be for testing. The training data for whole dataset is taken from 1 January 2013 to 30 April 2015 and the testing part of the data is taken from 1 May 2015 to 31 October 2015. Categorical events of the depended variable will coded as binary variables (positive outcome, and negative outcome). This step includes converting each value in the dependent variable PM_{10} to (-1,1), by taking the criterion of pollution equaled to 50. If the value is greater than or equal to 50, it will be denoted as 1, and if the value is smaller than 50, it will be denoted as -1. The daily data for PM_{10} values have been collected from Sek. Keb. Batu Muda, Kuala Lumpur, Malaysia meteorological station. Time- stratified method will be used for more accurate forecasting results using different seasonal patterns. The first time stratified series for first season S_1 (rain season) will be includes five months (January 2013, February 2013, December 2013, January 2014 and February 2014) for training and three months of data (December 2014, January 2015 and February 2015) for testing. The second time stratified series for second season S_2 will be includes six months (March 2013, April 2013, May 2013 and March 2014, April 2014, and May 2014) for training and three Months (March 2015' April 2015' and May 2015) for testing. The third time stratified series for third season S_3 (dry season) will be also includes six months (June 2013, July 2013, August 2013, June 2014, July 2014, and August 2014) for training and other three months (June 2015, July 2015, and August 2015) for testing. The last time stratified series for fourth season S_4 will be includes six months (September 2013, October 2013, November 2013, September 2014, October 2014, and November 2014) for training and three months (September 2015, October 2015, and November 2015) for testing Figure 1 and Figure 2 below explain the original training and testing series of full datasets respectively.

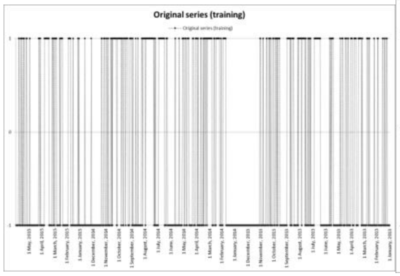


FIGURE 1: THE ORIGINAL TRAINING SERIES OF FULL DATA FROM 1/1/2013 TILL 31/5/2015.

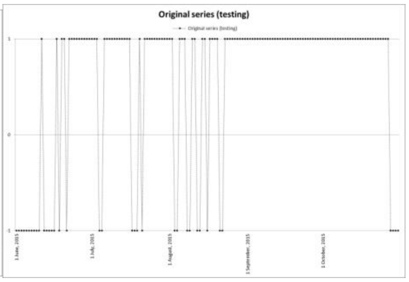


FIGURE 2: THE ORIGINAL TESTING SERIES OF FULL DATA FROM 1/6/2013 TILL 31/10/2015.

From Figure 1 and Figure 2 above, several seasonal patterns are detected PM_{10} variable for training and testing periods. Therefore, time-stratified method will be used to rearrange and divided the full dataset into four subgroups of data such as explained in details previously. Figure 3 and Figure 4 below explain the original time-stratified S_1 for training and testing periods respectively. From Figure 3and Figure 4 above, several seasonal patterns are detected time-stratified S_1 variable for training and testing periods. Figure 5 and Figure 6 below explain the original time-stratified S_2 for training and testing periods respectively.

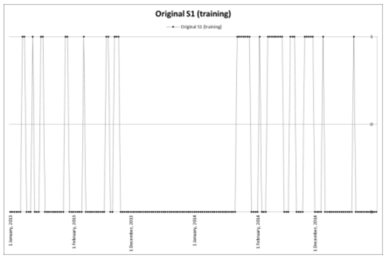


FIGURE 3: THE ORIGINAL TIME-STRATIFIED TRAINING SERIES S_1 FROM 1/1/2013 TILL 31/12/2014.

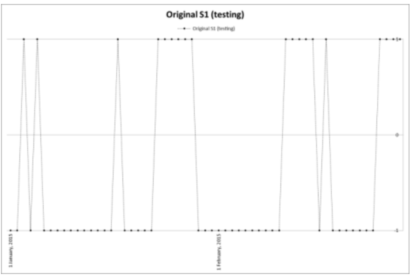


FIGURE 4: THE ORIGINAL TIME-STRATIFIED TESTING SERIES S_1 FROM 1/1/2015 TILL 31/2/2015.

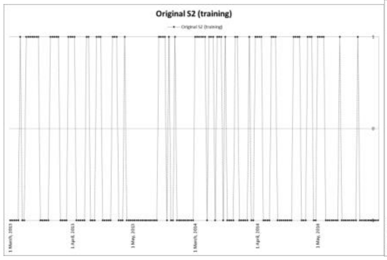


FIGURE 5: THE ORIGINAL TIME-STRATIFIED TRAINING SERIES S_2 FROM 1/3/2013 TILL 31/5/2014.

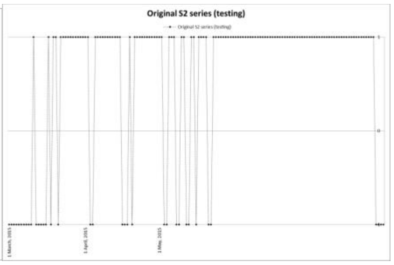


FIGURE 6: THE ORIGINAL TIME-STRATIFIED TESTING SERIES S_2 FROM 1/3/2015 TILL 31/5/2015.

From Figure 5 and Figure 6 above, several seasonal patterns are detected time-stratified S_2 variable for training and testing periods. Figure 7 and Figure 8 below explain the original time-stratified S_3 for training and testing periods respectively.
From Figure 7 and Figure 8 above, several seasonal patterns are detected time-stratified S_3 variable for training and testing periods. Figure 9 and Figure 10 below explain the original time-stratified S_4 for training and testing periods respectively.
From Figure 9 and Figure 10 above, several seasonal patterns are detected time-stratified S_4 variable for training and testing periods.

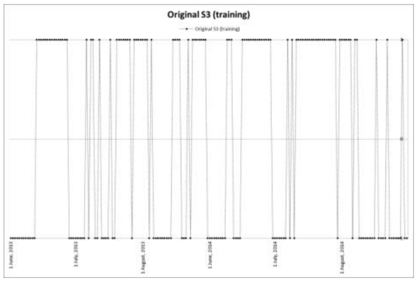


FIGURE 7: THE ORIGINAL TIME-STRATIFIED TRAINING SERIES S_3 FROM 1/6/2013 TILL 31/8/2014.

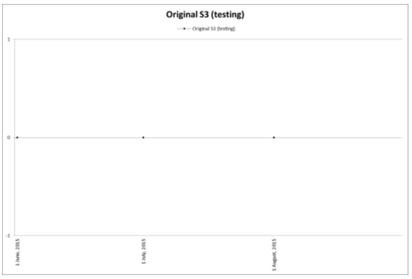


FIGURE 8: THE ORIGINAL TIME-STRATIFIED TESTING SERIES S_3 FROM 1/6/2015 TILL 31/8/2015.

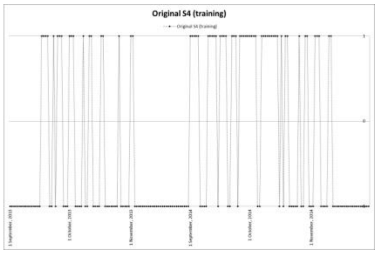


FIGURE 9: THE ORIGINAL TIME-STRATIFIED TRAINING SERIES S_4 FROM 1/9/2013 TILL 30/11/2014.

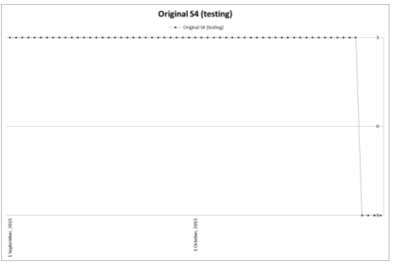


FIGURE 10: THE ORIGINAL TIME-STRATIFIED TESTING SERIES S_4 FROM 1/9/2015 TILL 31/10/2015.

3.1 LOGISTIC REGRESSION (LR) MODEL

LR is used when the dependent variable is a binary and the independent variables can be quantity, categorical, or both of them. To find LR models for PM_{10} and other meteorological variables, Minitab and Excel programs are used to perform that. Then the variable y (PM_{10}) can be evaluated according to the following model.

$$y = \beta_0 + \beta_1 CO + \beta_2 O_3 + \beta_3 SO_2 + \beta_4 NO_X + \beta_5 NO + \beta_6 AT + \beta_7 WS_{10m} \qquad (6)$$

Whereas: β_0 is the constant model, $\beta_1, \beta_2, \ldots, \beta_p$ are the parameters of regression model, CO: Carbon Monoxide, O_3: Ozone, SO_2: Sulphur Dioxide, NO_X: Nitrogen Dioxides, NO: Nitric Oxide, AT: Represent the temperature, WS_{10m}: Represents wind speed. The probability of success will be calculated according to the following equation.

$$\pi = \frac{1}{1+e^{-z}}$$

And the probability of failure $1-\pi$ will be calculated using the complement of π. Using a values of π and $1-\pi$, $\hat{z}$ will be obtained that will represent the logistic regression of the values of PM_{10} according to the following equation.

$$\hat{z} = Ln \frac{\pi}{1-\pi}$$

Then π variable will be converted to another categorical variable by using the value 0.5 which can be depended as fitted variable F that will compared with original z to get the classification or forecasting accuracy. If $\pi \geq 0.5$ then the value of π will be converted to 1, If $\pi < 0.5$ then the value of π will be converted to 0 as statistical term or -1 as in computer term as negative feature. The criterion of the classification accuracy will be calculated by using the following equation.

$$\text{Accuracy} = \frac{TP + TN}{N} \qquad (7)$$

Whereas:-

TN: The number of samples classified as negative (does not have the characteristic) is actually negative.

TP: The number of samples classified as positive (possessing the characteristic) is in fact positive.

N: Total number of samples. (Ferrer & Wang, 1999; Soderstrom & Leitner, 1997). By using Minitab and inserting the dependent and independent full datasets to find the best logistic regression model, the binary variable PM_{10} that was defined as dependent variable z and seven independent variables those were defined previously are inserted. The best logistic regression model of full dataset is as follows.

$$y = -23.8539 + 8.64487CO + 24.9608O_3 + 290.227SO_2 + 206.404NO_x - 346.32NO$$
$$+0.409440AT + 0.439284WS_{10m} \qquad (8)$$

The details of coefficients, are as in Table 1.

TABLE 1: THE DETAILS OF THE COEFFICIENTS OF LR MODEL FOR FULL DATA.

Predictor	Coef	Z-test	P
Constant	-23.854	-7.30	0.000
CO	8.645	8.61	0.000
O3	24.961	2.97	0.003
SO2	290.227	3.09	0.002
Nox	206.404	6.95	0.000
NO	-46.320	-8.78	0.000
AT	0.409	3.56	0.000
WS_{10m}	0.439	3.76	0.000

In Table 1 above, all coefficients are significant because of their P-values are less than 0.05. All of these variables will be used for all other models in order to compare among them.

Figure 11 and Figure 12 below explained the fitness of LR training and testing forecast series respectively with orig inal series for full dataset.

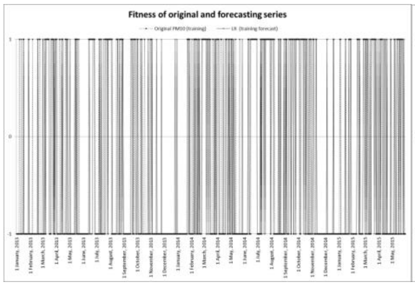


FIGURE 11: THE FITNESS OF ORIGINAL AND LR TRAINING FORECAST SERIES OF FULL DATA FROM 1/1/2013 TILL 31/5/2015.

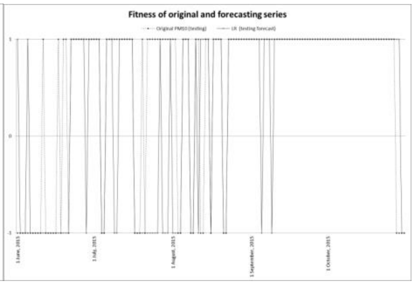


FIGURE 12: THE FITNESS OF ORIGINAL AND LR TESTING FORECAST SERIES OF FULL DATA FROM 1/6/2013 TILL 31/10/2015.

From Figure 11 and Figure 12 the fitness between the forecasting and original series are acceptable with some lags. The influences of different seasonal patterns may effect on the accuracy of LR forecasting results. Time-stratified model will be handle the problems of the seasonal patterns.

3.2 TIME-STRATIFIED METHOD.

In this study, the time-stratified method is used by following the seasonal pattern as mentioned above. Four random groups were identified. This method includes stratifying the similar seasons from different years one by one according to ascending order. To construct the first season according to time-stratified method as an example, the data of first season of 2014 will be taken and stratified directly after the data of first season of 2013 and so on. For time-stratified S_1 dataset, the binary dependent variable PM_{10} (y) and seven independent variables are inserted in Minitab. The LR model of time-stratified S_1 dataset will be as follows.

$$y = -19.7030 + 1.48464CO + 170.80003O_3 + 330.040SO_2 + 344.105NO_x - 261.084NO$$
$$+0.149419AT + 0.579733WS_{10m} x8 \qquad (9)$$

The details of coefficients are as in Table 2.

Table 2: the details of LR coefficients of model for time-stratified S_1 data.

Predictor	Coef	Z-test	P
Constant	-19.703	-2.22	0.026
CO	1.485	0.53	0.597
O3	170.800	3.46	0.001
SO2	330.040	1.03	0.305
Nox	344.105	4.08	0.000
NO	-61.084	-2.69	0.007
AT	0.149	0.47	0.639
WS_{10m}	0.580	2.32	0.020

In Table 2 above, the model of time-stratified S_1 reflect that there are some coefficients are insignificant because all items in companions should be have similar conations. Therefore, all models for full data and time-stratified data S_1,S_2 ,S_3 and S_4 should have similar structure of variables.

Figure 13 and Figure 14 below explained the fitness of LR training and testing forecast series respectively with original series for time-stratified S_1 dataset.

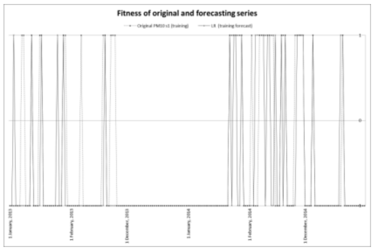


FIGURE 13: THE FITNESS OF ORIGINAL TIME-STRATIFIED S_1 AND LR TRAINING FORECAST SERIES FROM 1/1/2013 TILL 31/12/2014.

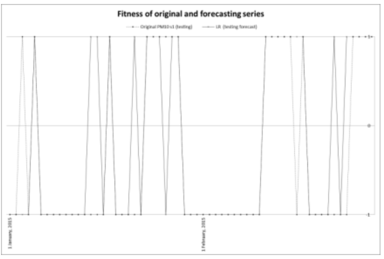


FIGURE 14: THE FITNESS OF ORIGINAL TIME-STRATIFIED S_1 AND LR TESTING FORECAST SERIES FROM 1/1/2015 TILL 31/2/2015.

From Figure 13 and Figure 14 for full data. The improvement in the accuracy of forecasting results belongs to using the time-stratified method and obtaining more homogeneous data.

The LR model of time-stratified S_2 dataset will be such as follows.

$$y = -15.6283 + 11.418CO + 16.3964O_3 - 80.5786SO_2 + 57.2452NO_x - 298.88NO$$
$$+0.261432AT + 0.116992WS_{10m} \qquad (10)$$

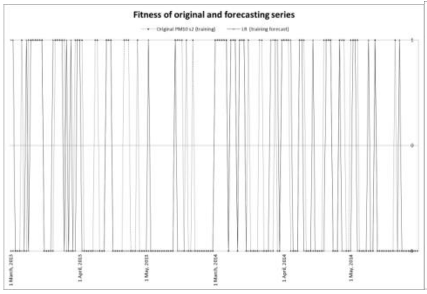
The details of coefficients are as in Table 3.

Table 3: The details of LR coefficients model for time-stratified S_2 data.

Predictor	Coef	Z-test	P
Constant	-15.628	-2.18	0.030
CO	11.418	5.03	0.000
O3	16.396	1.76	0.079
SO2	-80.579	-0.41	0.679
Nox	57.245	0.87	0.382
NO	-298.883	-3.44	0.001
AT	0.261	1.07	0.286
WS_{10m}	0.117	0.44	0.659

In Table 3 above, the model of time-stratified S_2 includes some insignificant parameters because of same reason mentioned above.

Figure 15 and Figure 16 below explained the fitness of LR training and testing forecast series respectively with original series for time-stratified S_2 dataset.



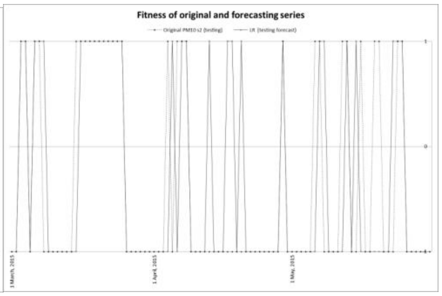


FIGURE 15: THE FITNESS OF ORIGINAL TIME-STRATIFIED S$_2$ AND LR TRAINING FORECAST SERIES FROM 1/3/2013 TILL 31/5/2014.

FIGURE 16: THE FITNESS OF ORIGINAL TIME-STRATIFIED S$_2$ AND LR TESTING FORECAST SERIES FROM 1/3/2015 TILL 31/5/2015.

From Figure 15 and Figure 16 for full data. The improvement in the accuracy of forecasting results belongs to using the time-stratified method and obtaining more homogeneous data.

The LR model of time-stratified S$_3$ dataset will be such as follows.

$$y = -52.3983 + 18.2970CO + 63.7141O_3 - 502.509SO_2 + 488.295NO_x - 618.144NO$$
$$+0.925373AT + 1.52072WS_{10m} \qquad (11)$$

The details of coefficients are as in Table 4.

Table 4: The details of the LR coefficients for time-stratified S$_3$ data.

Predictor	Coef	Z-test	P
Constant	-52.398	-4.75	0.000
CO	18.297	4.48	0.000
O3	63.714	0.97	0.330
SO2	-02.509	-1.85	0.064
Nox	488.295	3.64	0.000
NO	-18.144	-3.43	0.001
AT	0.925	2.61	0.009
WS$_{10m}$	1.521	2.81	0.005

In Table 4 above, all coefficients are significant except O3 and SO2 are insignificant because of their P-values are greater than 0.05. Figure 17 and Figure 18 below explained the fitness of LR training and testing forecast series respectively with original series for time-stratified S$_3$ dataset.

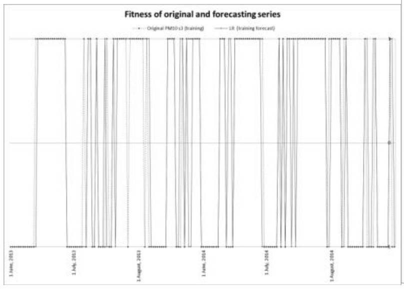


FIGURE 17: THE FITNESS OF ORIGINAL TIME-STRATIFIED S$_3$ AND LR TRAINING FORECAST SERIES FROM 1/6/2013 TILL 31/8/2014.

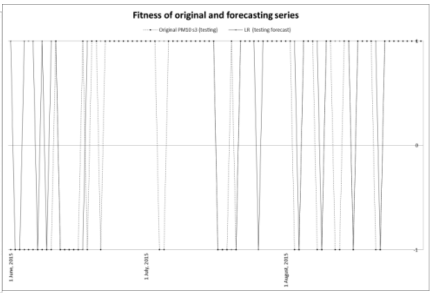


FIGURE 18: THE FITNESS OF ORIGINAL TIME-STRATIFIED S$_3$ AND LR TESTING FORECAST SERIES FROM 1/6/2015 TILL 31/8/2015.

From Figure 17 and Figure 18 for full data. The improvement in the accuracy of forecasting results belongs to using the time-stratified method and obtaining more homogeneous data.

The LR model of time-stratified S$_4$ dataset will be such as follows.

$$y = -23.2091 + 5.81558CO + 23.1442O_3 + 720.339SO_2 + 245.204NO_x - 433.476NO$$
$$+0.446819AT + 0.0078082WS_{10m} \qquad (12)$$

The details of coefficients, are as in Table 5.

Table 5: The details of LR coefficients for time-stratified S$_4$ data.

Predictor	Coef	Z-test	P
Constant	-23.209	-1.85	0.064
CO	5.816	2.24	0.025
O3	23.144	0.42	0.671
SO2	720.339	2.60	0.009
Nox	245.204	2.80	0.005
NO	-433.476	-3.69	0.000
AT	0.447	1.25	0.210
WS$_{10m}$	0.008	0.14	0.892

In Table 5 above, , the model of time-stratified S$_4$ includes some insignificant parameters because of same reason mentioned above. Figure 19 and Figure 20 below explained the fitness of LR training and testing forecast series respectively with original series for time-stratified S$_4$ dataset.

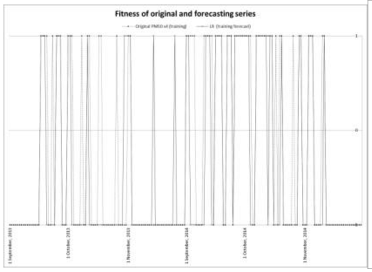


FIGURE 19: THE FITNESS OF ORIGINAL TIME-STRATIFIED S_4 AND LR TRAINING FORECAST SERIES FROM 1/9/2013 TILL 30/11/2014.

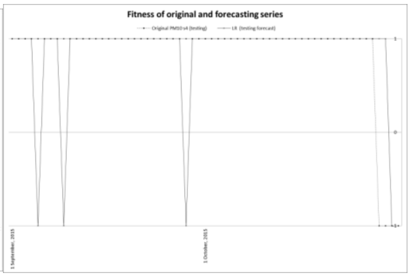


FIGURE 20: THE FITNESS OF ORIGINAL TIME-STRATIFIED S_4 AND LR TESTING FORECAST SERIES FROM 1/9/2015 TILL 31/10/2015.

From Figure 19 and Figure 20 the fitness between the forecasting and original series are acceptable and better. The improvement in the accuracy of forecasting results belongs to using the time-stratified method and obtaining more homogeneous data. The classification or forecasting was performed in Minitab program for all training datasets by supposing LR models similar to LR model of full dataset. For all testing datasets, the forecasting was performed by simulating LR models in training stages using Microsoft excel. The training and testing results are summarized as classification or forecasting accuracy measurements such as in

Table 6.

TABLE 6: CLASSIFICATION ACCURACY FORECAST BY USING LR METHOD.

Dataset	Training	Testing
Full	82.29%	82.35%
S1	86.67%	86.44%
S2	78.26%	81.52%
S3	88.59%	80.43%
S4	87.36%	91.80 %

4 CONCLUSIONS

PM_{10} forecasting is necessary to reduce the damage of environment and human health. LR forecasting can be depended after datasets time stratifying to satisfy more accuracy with nonlinear multivariate datasets in which PM_{10} is to dependent variable.

References

1. Bai, L., Wang, J., Ma, X., & Lu, H. (2018). Air pollution forecasts: An overview. *International journal of environmental research and public health, 15*(4), 780.
2. Dayton, C. M. (1992). Logistic regression analysis. *Stat*, 474-574.
3. Ferrer, A. J. A., & Wang, L. (1999). Comparing the Classification Accuracy among Nonparametric, Parametric Discriminant Analysis and Logistic Regression Methods.
4. Hosmer, D. W., Hosmer, T., Le Cessie, S., & Lemeshow, S. (1997). A comparison of goodness-of-fit tests for the logistic regression model. *Statistics in medicine, 16*(9), 965-980.
5. Krzyzanowski, M., Bundeshaus, G., Negru, M. L., & Salvi, M. C. (2005). Particulate matter air pollution: how it harms health. *World Health Organization, Fact sheet EURO/04/05, Berlin, Copenhagen, Rome, 4*, 14.
6. Malig, B. J., Pearson, D. L., Chang, Y. B., Broadwin, R., Basu, R., Green, R. S., & Ostro, B. (2015). A time-stratified case-crossover study of ambient ozone exposure and emergency department visits for specific respiratory diagnoses in California (2005–2008). *Environmental health perspectives, 124*(6), 745-753.
7. Marill, K. A. (2004). Advanced statistics: linear regression, part II: multiple linear regression. *Academic emergency medicine, 11*(1), 94-102.
8. Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics, 13*(3), 253-267.
9. Pohlman, J. T., & Leitner, D. W. (2003). A comparison of ordinary least squares and logistic regression.
10. Santner, T. J., & Duffy, D. E. (1986). A note on A. Albert and JA Anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 73*(3), 755-758.

11. Soderstrom, I. R., & Leitner, D. W. (1997). The Effects of Base Rate, Selection Ratio, Sample Size, and Reliability of Predictors on Predictive Efficiency Indices Associated with Logistic Regression Models.
12. Tobias, A., Armstrong, B., & Gasparrini, A. (2014). *Analysis of time-stratified case-crossover studies in environmental epidemiology using Stata.* Paper presented at the United Kingdom Stata Users' Group Meetings 2014.
13. Vijayaraghavan, N., & Mohan, G. (2016). Air pollution analysis for Kannur city using artificial neural network. *International Journal of Science and Research, 5*, 1399-1401.
14. 14.support vector machines. *Journal of Control Science and Engineering, 2012*, 4.

استخدام الانحدار اللوجستي مع طريقة الطبقية الزمنية للتنبؤ بمجموعات بيانات تلوث الهواء

سرى أمير محمد   و   أسامة بشير حنون

قسم الاحصاء والمعلوماتية، كلية علوم الحاسوب والرياضيات ،جامعة الموصل، الموصل، العراق

الخلاصة:ان دراسة الجسيمات المعلقة (PM_{10}) والتكهن بها ضروري للتقليل والسيطرة على الأضرار البيئية وصحة الانسان. هنالك العديد من مصادر التلوث او ما يسمى بالملوثات والتي ربما تؤثر على PM_{10}. بيانات الدراسة تم اخذها من محطة مناخية في كوالالمبور، ماليزيا. كل هذه المتغيرات تصنف بياناتها كغير خطية. نموذج الانحدار اللوجستي باستخدام النموذج الخطي المعمم كحالة خاصة من الطرق الاحصائية الخطية وبالتالي فقد يعكس نتائج غير دقيقة عند استخدامه مع مجموعات البيانات غير الخطية. طريقة التراصف الزمني في أنماط مختلفة تم اقتراحها لتحسين تلك النتائج وتحقيق التجانس ويتضمن مراصفة المواسم المتشابهة في السنوات المختلفة سوية لتكوين متغير جديد مختلف عن الاصلي. نتائج نموذج الانحدار اللوجستي هي افضل من النتائج للبيانات الاصلية الكلية لذلك نستنتج ان تكهنات الانحدار اللوجستي من الممكن اعتمادها بعد اخذ التراصف الزمني للبيانات بنظر الاعتبار مع البيانات غير الخطية متعددة المتغيرات عندما يكون PM_{10} كمتغير معتمد.

الكلمات المفتاحية: الانحدار اللوجستي (LR) ،الوقت الطبقي (TS) ،مسألة معينة (PM10) ،التنبؤ ، تلوث الهواء.