International Journal on Advanced Science Engineering Information Technology

Vol.10 (2020) No. 4
ISSN: 2088-5334

Using the MLR and Neuro-Fuzzy Methods to Forecast Air Pollution Datasets

Osamah Basheer Shukur[a,*]

[a] *Department of Statistics and Informatics, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.*
E-mail: [*] drosamahannon@uomosul.edu.iq

***Abstract*— The forecasting of time series data is essential by following statistical and intelligent techniques. Particular matter (PM10) is a time series dataset used to scale the air pollution as a dependent variable while there are many types of pollutants used as independent variables. MLR model has been used as a traditional linear approach to forecasting PM_{10} data. Combining NF as a nonlinear intelligent method with MLR in a hybrid MLR-NF method has been proposed for improving PM_{10} forecasts and handling the nonlinearity of datasets. The forecasting results reflected that the hybrid method outperformed the traditional method. Although a multiple linear regression (MLR) model has been used for air quality forecasting depending on several meteorological variables in many recent studies, the MLR model is unable to identify the nonlinear pattern of these types of data. Malaysian datasets of PM10 and several climate pollutants will be studied in this paper. The objective of this study is to forecast PM10 and obtain the best results and minimum forecasting error. In this paper, the dependent variable will be forecasted by using traditional and intelligent methods. MLR has been used as a traditional method to forecast PM10. Neuro-fuzzy (NF) in the adapted copy, which calls the adaptive neuro-fuzzy inference system (ANFIS) is combined with MLR and used as an intelligent method to forecast PM10. The results reflect that MLR-NF outperformed MLR for forecasting PM10 data. In conclusion, MLR-NF can be used to forecast PM10 for more accurate results compared to traditional methods.**

***Keywords*— multiple linear regression (MLR); neuro-fuzzy (NF); PM10 datasets; forecasting; ANFIS.**

I. INTRODUCTION

Air quality and climate data have nonlinear nature, so forecasting will be a complicated process. Particulate matter (PM_{10}) data is a feature that can be used to measure air pollution depending on other meteorological datasets. The fluctuation in the pattern of PM_{10} is often the main reason for nonlinear behavior. High-quality results of PM_{10} forecasts depending on several meteorological datasets are important to control human health and environmental phenomena.

In of sample and out of sample, PM_{10} forecasting based on several pollutants as independent variables can be performed using multiple linear regression (MLR) model as a traditional method. MLR model reflects the relationship between the dependent or response variable and several explanatory or predictor variables that impact on dependent variable [1], [2] suggested MLR model forecast daily PM_{10} data with meteorological datasets as the best models that express the relationship between variables.

Neuro-fuzzy is suggested as an intelligent method to handle the nonlinearity and improve forecasting results. The adaptive neuro-fuzzy inference system (ANFIS) is adapted and improved copy of neuro-fuzzy as a perfect system to forecast nonlinear datasets in high-quality results. ANFIS method is subject to forecast for short terms. ANFIS was used to forecast three datasets through all seasons for urban areas in Romania. ANFIS is developed for forecasting of daily air pollution concentrations for five air pollutants datasets. Wongsathan analyzed PM-10 values to formulate a hybrid method to forecast PM_{10} and found the influence of exogenous variables in Chiang Mai Province, Thailand. [3], [4] presented MLR and Adaptive NF methods separately to forecast many types of datasets.

In this study, PM_{10} for 34 months from 1 January 2013 till 31 October 2015 was forecasted dependent variable influenced by Carbon Monoxide (CO), Sulphur Dioxide (SO_2), Nitric Oxide (NO), Ozone (O3) by using MLR and ANFIS. The full period is divided into training and testing series. The time-stratified (TS) method is proposed in this study to analyze the short-term effects of risk factors. TS can be used to separate the seasonal pattern effects. [5] It can be used to satisfy homogeneity and to reach more accurate results [6], [7]. Data belongs to the same season through different years are stratified timely. Each year has four seasons, and each season consists of three months. Full and time stratified datasets were forecasted in training and testing stages by using MLR. The structure of MLR input is

used to construct the structure of the ANFIS method. ANFIS as an intelligent method was combined with MLR in one hybrid method (MLR-NF) to improve the accuracy of PM_{10} forecast based on several pollutant variables.

II. Materials and Method

A. Data and Framework of the Study

In this study, Malaysian daily datasets of PM_{10} and several climate pollutants were studied for 34 months (1 January 2013 – 31 October 2015), which includes 1034 daily observations. The full period was divided into two periods for training and testing. The framework of this study includes the following [8]:

- Dividing the full period into two groups for training and testing.
- Modeling training data by using the MLR model.
- Simulating testing data by using the same MLR model.
- Modeling training data by using the proposed MLR-NF method.
- Simulating testing data by using the same proposed methods.
- Comparing MLR model as traditional methods and the proposed MLR-NF as an intelligent method to determine the best accuracy of forecasting results.

B. MLR model for PM10 Forecasting.

MLR model is used for modeling the effects of multiple explanatory variables on one dependent variable. Many researchers used MLR, which can be formulated as follows [9].

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_i \qquad (1)$$

where y_i is the dependent variable, $x_1, x_2, \ldots, x_p$ are the explanatory variables, β_0 is the cross point with y axis, $\beta_1, \beta_2, \ldots, \beta_p$ are the regression parameters, and e_i is the random error. Equation (1) can be written in matrices form as follows [10], [11].

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \varepsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \qquad (2)$$

Using matrix symbols, the linear regression model can be written as follows [12].

C. A hybrid MLR-NF Method

In this study, a hybrid MLR-NF method has been proposed to improve the forecasting results of PM_{10} based on several meteorological datasets. The framework of this proposed approach is as follows [13], [14].

- Constructing the best MLR model to be suitable for dataset.
- Constructing ANFIS based on the structures of MLR model to obtain MLR-NF hybrid method.
- Achieving training and testing processes to obtain forecasted series of the hybrid MLR-NF method.
- Comparing the results of the hybrid MLR-NF method to MLR for training and testing forecast.

A hybrid MLR-NF method is proposed for PM_{10} forecasting. MLR model was used only for drawing the structure of ANFIS inputs for training and testing. Hybrid MLR-NF was also proposed to handle the nonlinear data. Determining the method for generating a fuzzy inference system (FIS), optimization method, error tolerance, membership function types and parameters, rules, and other requirements were necessary to create the more suitable ANFIS structure [15].

ANFIS can be defined as a multi-layer feedforward neural network with learning algorithms and FIS through the inputs to the output [16]. Jang had introduced ANFIS to apply the main stage to build suitable FIS. ANFIS structure is tuning based on Sugeno-type of FIS [17]. First-order Sugeno-type of FIS includes a set of following rules for two if-then rules [18].

Rule 1: If x is equal to $\boldsymbol{A}_1$ and y is equal to $\boldsymbol{B}_1$, then $p_1 x + q_1 y + r_1 = f_1$

Rule 2: If x is equal to $\boldsymbol{A}_2$ and y is equal to $\boldsymbol{B}_2$, then $p_2 x + q_2 y + r_2 = f_2$

where $\boldsymbol{A}_1$, $\boldsymbol{B}_1$, $\boldsymbol{A}_2$, and $\boldsymbol{B}_2$ represents the membership functions for the input data x and y, while r_1, r_2, p_1, p_2, q_1, q_2, f_1, and f_2 represents the parameters of the output function. Input data are inserted as columns vector when the last column represents the target or dependent variable.

For generating FIS, there are two widespread partition methods (grid partitioning and subtractive clustering) that is used for FIS structure generating and rule establishing. Choosing grid partitioning includes uniformly partitioning the variable input ranges to generate input membership functions with single-output Sugeno FIS [19]. It is appropriate for the few input variables. Under this choosing, there will be one rule only for each of the input membership functions. By choosing subtractive clustering, using subtractive clustering of input and output data depends on deriving data clusters from generating Sugeno FIS and establishing membership functions and rules [20].

To determine the number of membership functions, trial and error principle are varied. Based on same principle, the optimal membership function can be the Gaussian. The structure of ANFIS can be presented such as follows in Fig.1.

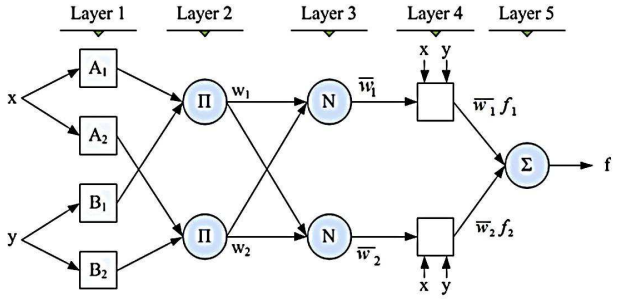

Fig.1 Architecture of five layers ANFIS method.

where Πrepresents layer 2 with fixed nodes, w_i represents the weight of the rule f_i, Nrepresents layer 3 with fixed nodes, $\bar{w}$represents the normalized weight, and $f$represents

the final output. The most widespread membership function types are presented in Fig.2 as follows [20], [21].

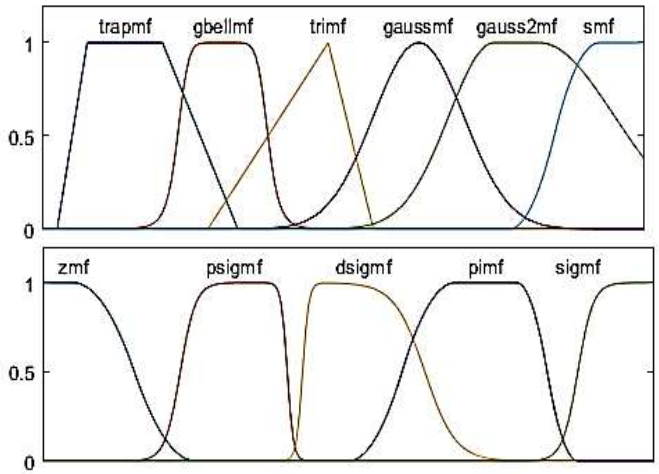

Fig.2 Membership function carves for ANFIS method [2]

In Fig.2 above, the presented curves were for triangular, trapezoidal, generalized bell, sigmoidal, *Z*, *S*, and *Pi* curves, and two different Gaussian curves in addition to the difference between two sigmoidal, and the product of two sigmoidal membership functions. Mean absolute percentage error (MAPE) is used in this study as forecasting error criteria, it can be written such as follows [22].

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{e_i}{y_i}\right| \times 100 \qquad (3)$$

where e_i is the random error of forecasting, n is the total number of observations, and y_i is the original series [13], [14].

III. Result and Discussion

Malaysian PM_{10} (less than or equal 10 micrometers) and several independent variables such as (Carbon Monoxide (CO), Sulphur Dioxide (SO_2), Nitric Oxide (NO), Ozone (O3)) in 1034 daily observation for 34 months from 1 January 2013 till 31 October 2015 will be studied in this paper. Datasets will be divided into training and testing data for full and TS datasets. TS method has been applied on a full dataset to result in four subgroups of data: S_1, S_2, S_3, S_4 time stratified datasets. Full-time series data were divided into 29 months (1 January 2013 – 31 May 2015) for training and five months of data (1 June 2015 – 31 October 2015) for testing. S_1 series (rain season) included six months (January 2013, February 2013, December 2013, January 2014 and February 2014, December 2014) for training and two months of data (January 2015 and February 2015) for testing. S_2 included six months (March 2013, April 2013, May 2013 and March 2014, April 2014, and May 2014) for training and 3 Months (March 2015' April 2015' and May 2015) for testing. S_3 (dry season) included six months (June 2013, July 2013, August 2013, June 2014, July 2014, and August 2014) for training and the other three months (June 2015, July 2015, and August 2015) for testing. S_4 included six months (September 2013, October 2013, November 2013, September 2014, October 2014, and November 2014) for training and two months (September 2015, October 2015) for testing.

A. *MLR model*

TABLE I
THE DETAILS OF MLR COEFFICIENT FOR FULL DATA.

Term	Coefficient	T-value	P-value
CO	79.01	25.82	0.000
SO2	3501.00	7.27	0.000
NO	-2029.00	-15.61	0.000
O3	197.60	2.97	0.003

From Table 1, all MLR coefficients in equation (4) are significant because of their p-values are less than the significant level of 5%. Therefore, the MLR model in equation (4) is fitted to datasets. After performing training and testing processes, MAPE values for MLR in equation (4) are 27.22 and 28.01 for training and testing, respectively. Fig.3 and Fig.4 explain the fitness between the original series and forecasted series for training and testing respectively for full data.

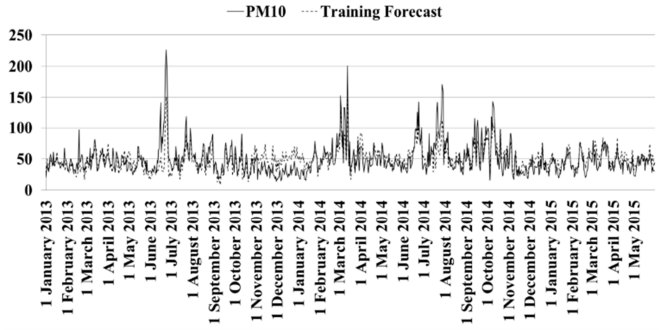

Fig.3: The fitness between the original series and training forecasted series for full data by using MLR model.

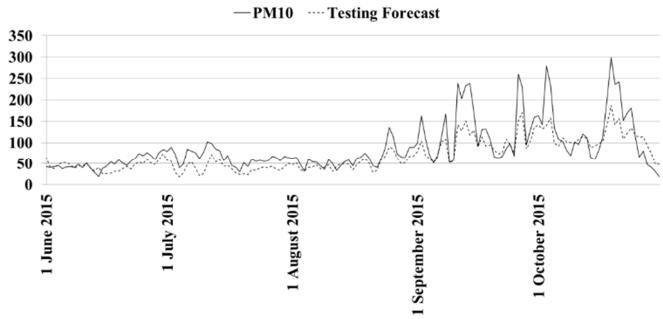

Fig.4: The fitness between the original series and testing forecasted series for full data by using MLR model.

The MLR model for S_1 dataset is as follows.

$$y = 16.46\,x_1 + 4357.00\,x_2 - 210.00\,x_3 + 954.00x_4 \qquad (4)$$

The details of MLR coefficients in equation (5) will be displayed in Table 2.

TABLE II
THE DETAILS OF MLR COEFFICIENT for S_1 DATA.

Term	Coefficient	T-value	P-value
CO	16.46	2.74	0.01
SO2	4357.00	4.05	0.00
NO	210.00	0.81	0.42
O3	954.00	7.62	0.00

From Table 2, Some of MLR coefficients in equation (5) are insignificant because of their p-values are greater than the significant level 5%. To compare the forecasting results of full and time stratified datasets, MLR model of time stratified data were taken with no changes. After performing training and testing processes, MAPE values for MLR in equation (5) are 25.32 and 25.03 for training and testing, respectively. Fig.5 and Fig.6 explain the fitness between the original series and forecasted series for training and testing respectively for S_1.

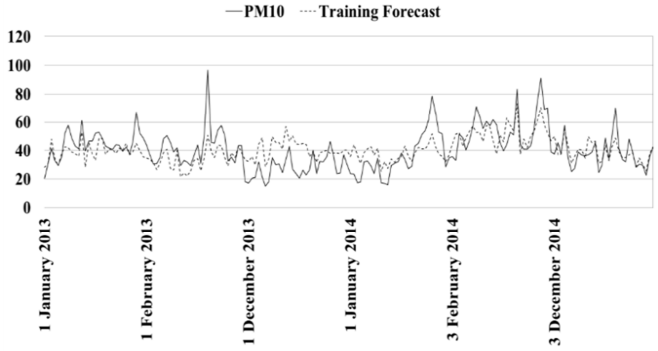

Fig.5: The fitness between the original series and training forecasted series for S_1 by using MLR model.

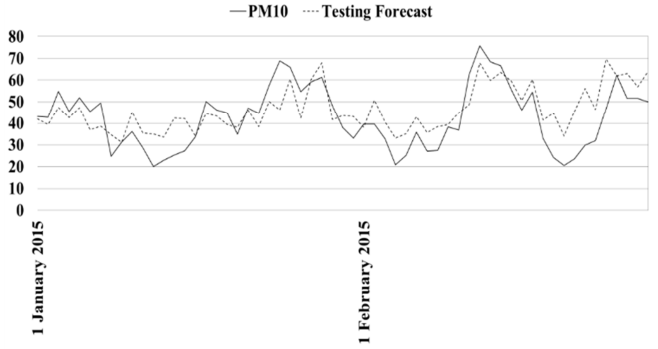

Fig.6: The fitness between the original series and testing forecasted series for S1 by using MLR model.

The MLR model for S_2 dataset is as follows.

$$y = 81.56x_1 + 1887.00\,x_2 - 2194.00x_3 + 308.30x_4 \qquad (5)$$

The details of MLR coefficients in equation (6) will be displayed in Table 3.

TABLE III
THE DETAILS OF MLR COEFFICIENT FOR S2 DATA.

Term	Coefficient	T-value	P-value
CO	81.56	16.51	0.00
SO2	1887.00	2.14	0.03
NO	-2194.00	-10.22	0.00
O3	308.30	3.46	0.00

From Table 3, all MLR coefficients in equation (6) are significant because of their p-values are less than the significant level of 5%. After performing training and testing processes, MAPE values for MLR in equation (6) are 20.74 and 22.06 for training and testing, respectively. Fig.7 and Fig.8 explain the fitness between the original series and forecasted series for training and testing respectively for S_2.

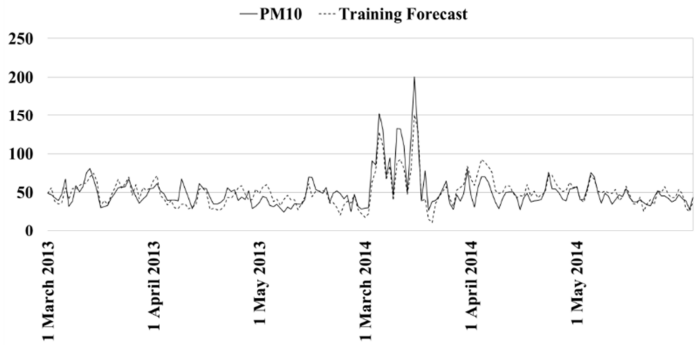

Fig.7: The fitness between original series and training forecasted series for S2 by using MLR model.

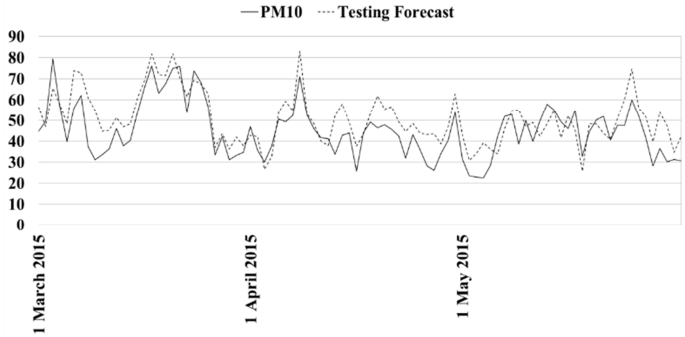

Fig.8: The fitness between original series and testing forecasted series for S2 by using MLR model.

The MLR model for S_3 dataset is as follows.
The details of MLR coefficients in equation (7) is displayed in Table 4.

TABLE IV
THE DETAILS OF MLR COEFFICIENT FOR S3 DATA.

Term	Coefficient	T-value	P-value
CO	122.76	18.99	0.00
SO2	-3117.00	-3.18	0.00
NO	-2179.00	-8.05	0.00
O3	358.00	1.95	0.05

From Table 4, All of MLR coefficients in equation (7) are significant because of their p-values are less than the significant level 5%. After performing training and testing processes, MAPE values for MLR in equation (7) are 24.82 and 19.03 for training and testing, respectively. Fig.9 and Fig.10 explain the fitness between the original series and forecasted series for training and testing respectively for S_3.

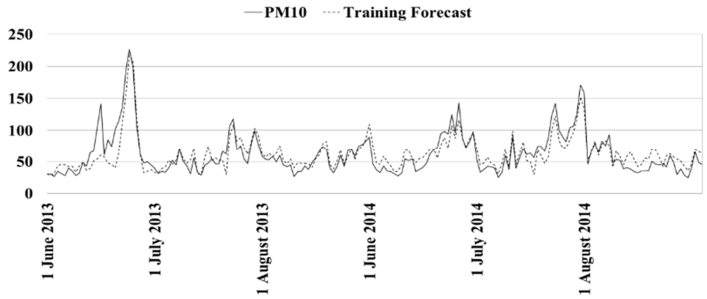

Fig.9: The fitness between the original series and training forecasted series for S_3 by using MLR model.

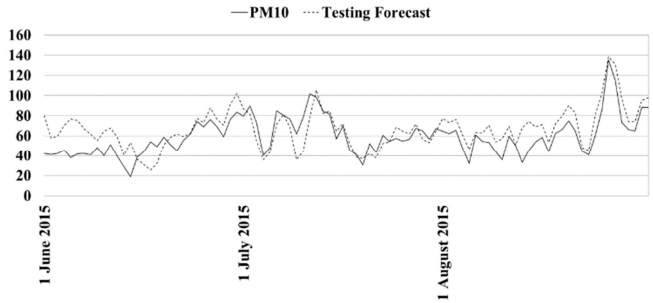

Fig.10: The fitness between the original series and testing forecasted series for S3 by using MLR model.

The MLR model for S_4 dataset in the form of MLR coefficients in equation (8) is displayed in Table 5.

TABLE V
THE DETAILS OF MLR COEFFICIENT FOR S4 DATA.

Term	Coefficient	T-value	P-value
CO	82.02	12.43	0.00
SO2	6149.00	5.48	0.00
NO	-2387.00	-9.59	0.00
O3	-120.00	-0.63	0.53

From Table 5, some of MLR coefficients in equation (8) are insignificant because of their p-values are higher than the significant level of 5%. After performing training and testing processes, MAPE values for MLR in equation (8) are 31.63 and 29.68 for training and testing, respectively. Fig.11 and Fig.12 explain the fitness between the original series and forecasted series for training and testing respectively for S_4.

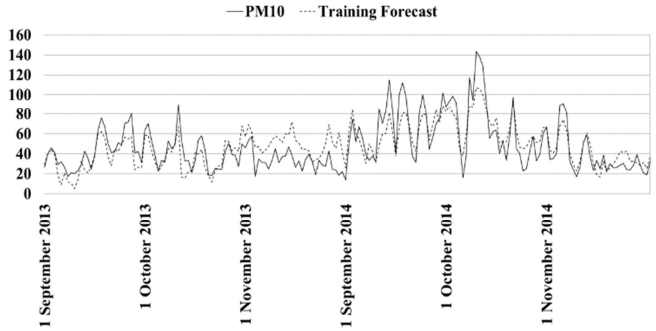

Fig.11: The fitness between the original series and training forecasted series for S_4 by using MLR model.

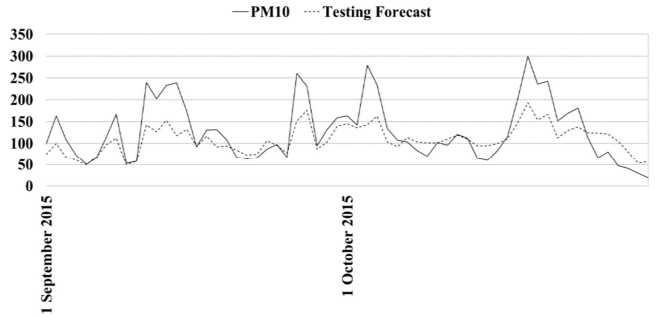

Fig.12 The fitness between original series and testing forecasted series for S4 by using MLR model.

Table 6 below presents the MAPE values of MLR model for full and time stratified datasets for the training process.

TABLE VI
MAPE OF FORECAST SERIES BY USING MLR

City	Training	Testing
Full	27.22	28.01
S_1	25.32	25.03
S_2	20.74	22.06
S_3	24.82	19.03
S_4	31.63	29.68

Table 6 and previous figures Fig.3 till Fig.12 refer that MLR model is resulted in better training and testing forecasting results for the time stratified S_1, S_2, S_3, and S_4 datasets than full datasets. Therefore, TS method can be suggested to improve forecasting results by satisfying the homogeneity for seasonal data.

B. Hybrid MLR-NF method.

Hybridizing linear and nonlinear approaches in one hybrid method may be proposed to improve forecasting results for any type of datasets. Combining MLR with ANFIS will result in a suitable method that can handle nonlinear data. The framework for forecasting by using MLR-NF hybrid method can be detailed practically as follows.

- Input variables are the same as the terms, including coefficients and variables on the right hand of MLR model.
- The target variable is the same as the dependent variable purely.
- The inputs variable and target variable should be entered into the workspace in MATLAB separately as columns vector; the last column should be specified for the target variable for training and testing separately.
- The training and testing data sets should be loaded from the workspace to the ANFIS toolbox in MATLAB.
- To generate FIS there are two options; a better option is choosing grid partition because there are just four input variables.
- In ANFIS training process, optimization method, error tolerance, and epochs are determined as hybrid, 0, and 100 respectively in this paper.
- The output variables of training and testing processes can be plotted and compared the fitness.

Fig.13 till Fig.22 present the fitness between original series and training forecasted series for full, S_1, S_2, S_3, and S_4 datasets by using hybrid MLR-NF method.

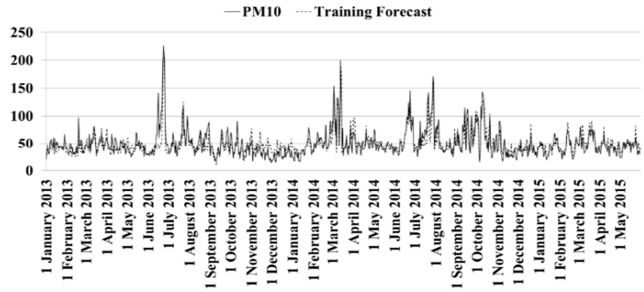

Fig.13 The fitness between original series and training forecasted series for full data by using hybrid MLR-NF method.

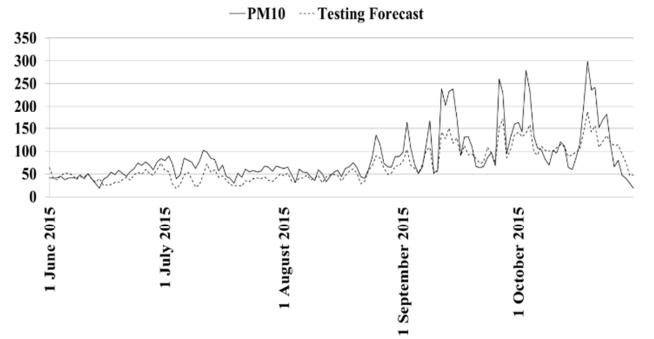

Fig.14 The fitness between original series and testing forecasted series for full data by using hybrid MLR-NF method.

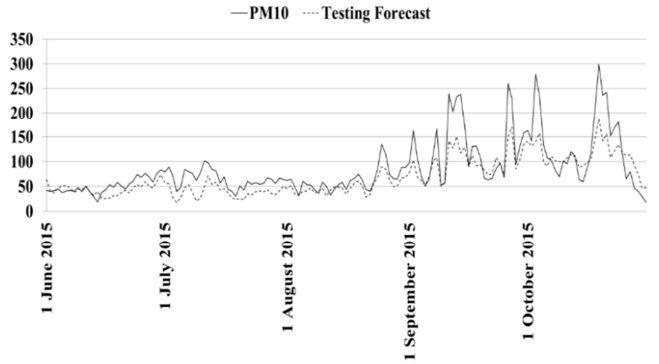

Fig.15 The fitness between original series and training forecasted series for S_1 data by using hybrid MLR-NF method.

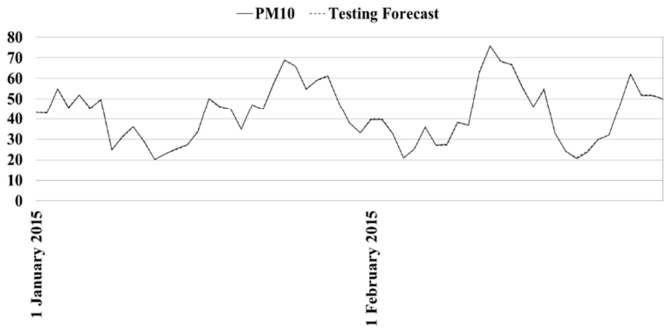

Fig.16 The fitness between original series and testing forecasted series for S_1 data by using hybrid MLR-NF method.

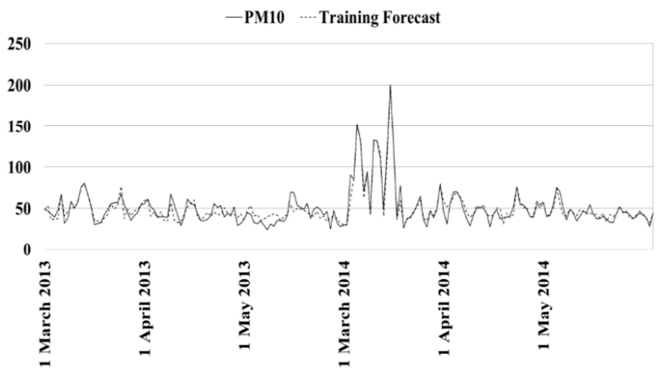

Fig.17 The fitness between original series and training forecasted series for S_2 data by using hybrid MLR-NF method.

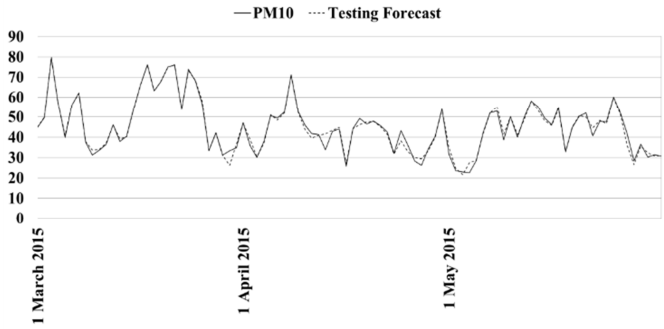

Fig.18 The fitness between original series and testing forecasted series for S_2 data by using hybrid MLR-NF method.

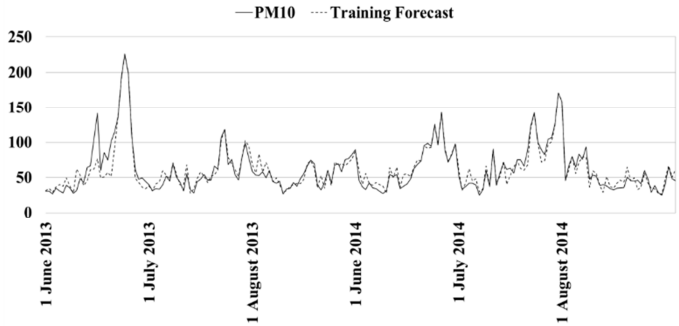

Fig.19 The fitness between original series and training forecasted series for S_3 data by using hybrid MLR-NF method.

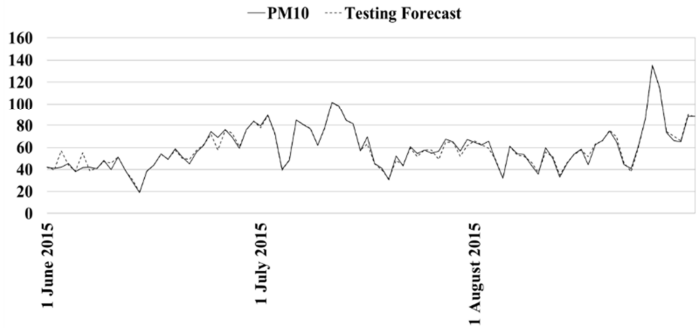

Fig.20 The fitness between original series and testing forecasted series for S_3 data by using hybrid MLR-NF method.

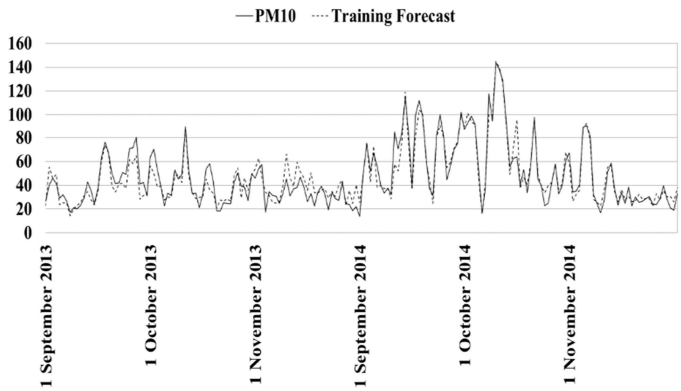

Fig.21 The fitness between original series and training forecasted series for S_4 data by using hybrid MLR-NF method.

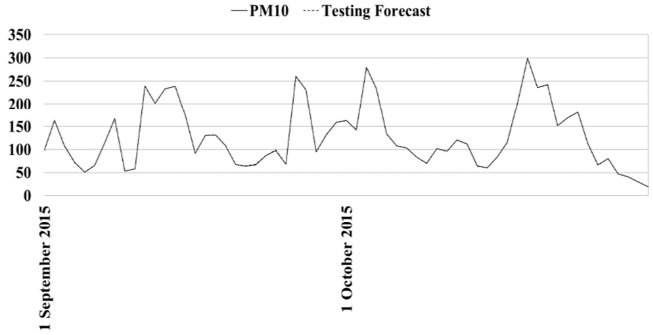

Fig.22 The fitness between original series and testing forecasted series for S_4 data by using hybrid MLR-NF method.

Comparing Fig.13 till Fig.22 to Fig.3 till Fig 12 one by one correspondingly, original series and forecasted series for MLR-NF for full, S_1, S_2, S_3, and S_4 datasets by using hybrid MLR-NF method are more fitted than those by using MLR model. Table 7 below presents the MAPE values of MLR-NF hybrid method for all studied datasets for training and testing forecasts.

TABLE VII
MAPE OF FORECAST SERIES BY USING MLR-NF.

City	Training	Testing
Full	20.70	10.37
S_1	14.79	0.04
S_2	11.68	3.03
S_3	13.89	3.60
S_4	16.07	0.20

Table 6, Table 7, and all of the previous figures confirm that all of the time, stratified datasets are resulted in better training and testing forecasts than full datasets by using MLR and $_{MLR}$-NF. The forecasting results by using the hybrid MLR-NF for all of the data styles outperformed the forecasting results by using the MLR model. Therefore, TS method can be suggested to improve forecasting results by satisfying the homogeneity for seasonal data. THE hybrid MLR-NF method can also be proposed for more improving the forecasting results.

IV. CONCLUSION

In this paper, MLR model has been used as linear traditional approach to forecast PM_{10} data. Combining NF as nonlinear intelligent method with MLR in hybrid MLR-NF method has been proposed for improving PM_{10} forecasts and handling the nonlinearity of datasets. The forecasting results reflected the hybrid method outperformed the traditional method. TS method was used in this paper for more homogeneity of studied datasets. Time stratified data outperformed full datasets in forecasting results for traditional and intelligent methods. In conclusion, the proposed hybrid method can be used to forecast PM_{10} in more accuracy of forecasting results. The hybrid method MLR-NF combines the MLR as a linear model with NF as a nonlinear method in one method can handle any type of data, especially the nonlinear type.

ACKNOWLEDGMENTS

The author is very grateful to the University of Mosul/ College of Computer Sciences and Mathematics for their provided facilities, which helped to improve the quality of this work.

REFERENCES

[1] Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. and Sliusarieva, A. (2012). Comparison of Multiple Linear and Nonlinear Regression, Autoregressive Integrated Moving Average, Artificial Neural Network, and Wavelet Artificial Neural Network Methods for Urban Water Demand Forecasting in Montreal, Canada. Water Resources Research, 48(1).

[2] Cavallaro, F. (2015). A Takagi-Sugeno Fuzzy Inference System for Developing a Sustainability Index of Biomass. Sustainability, 7(9), 12359-12371.

[3] Jung Wu, Nagi Gebraeel, Mark A Lawley, and Yuehwern Yih. A neural net- work integrated decision support system for condition-based optimal predictive maintenance policy. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(2):226–236, 2007.

[4] Zhiqiang Huo, Yu Zhang, Pierre Francq, Lei Shu, and Jianfeng Huang. Incipient fault diagnosis of roller bearing using optimized wavelet transform based multi- speed vibration signatures. IEEE Access, 5:19442–19456, 2017.

[5] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., Yidris, N., & Fattahi, A. (2020). Experimental and numerical investigation of the mechanical behavior of full-scale wooden cross arm in the transmission towers in terms of load-deflection test. Journal of Materials Research and Technology, 9(4), 7937-7946.

[6] Sharaf, H. K., Ishak, M. R., Sapuan, S. M., & Yidris, N. (2020). Conceptual design of the cross-arm for the application in the transmission towers by using TRIZ–morphological chart–ANP methods. Journal of Materials Research and Technology, 9(4), 9182-9188.

[7] Johari, A. N., Ishak, M. R., Leman, Z., Yusoff, M. Z. M., Asyraf, M. R. M., Ashraf, W., & Sharaf, H. K. (2019). Fabrication and cut-in speed enhancement of savonius vertical axis wind turbine (SVAWT) with hinged blade using fiberglass composites. In Proceedings of the Seminar Enau Kebangsaan (pp. 978-983).

[8] Asyraf, M. R. M., Ishak, M. R., Sapuan, S. M., Yidris, N., Johari, A. N., Ashraf, W., ... & Mazlan, R. (2019). Creep test rig for full-scale composite cr ossarm: simulation modelling and analysis. In Seminar Enau Kebangsaan (pp. 34-38).

[9] A. Environmentally, F. Tropical, and C. Hazard, "Journal of Advances in Modeling Earth Systems," pp. 223–241, 2018.

[10] L. Metz, N. Maheswaranathan, B. Cheung, and J. Sohl-Dickstein, "Learning Unsupervised Learning Rules," 2018.

[11] Indarto, "Penginderaan Jauh : Metode Analisis & Interpretasi Citra Satelit," no. June, 2017.

[12] A. Peytchev, A. Peytchev, and E. Peytcheva, "Reduction of Measurement Error due to Survey Length: Evaluation of the Split Questionnaire Design Approach," Surv. Res. Methods, vol. 11, no. 4, pp. 361–368, 2017.

[13] A. Zhang and Y. Xie, "Chaos theory-based data-mining technique for image endmember extraction: Laypunov index and correlation dimension (L and D)," IEEE Trans. Geosci. Remote Sens., vol. 52, no. 4, pp. 1935–1947, 2014.

[14] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., ... & Suveges, D. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays, and summary statistics 2019.

[15] Mohammad Reza Mahmoudi, Ali Abbasalizadeh, How statistics and text mining can be applied to literary studies?, Digital Scholarship in the Humanities, Volume 34, Issue 3, September 2019, Pages 536–541, https://doi.org/10.1093/llc/fqy069.

[16] Shelke, M., Deshpande, S. S., & Sharma, S. (2020). Quinquennial Review of Progress in Degradation Studies and Impurity Profiling: An Instrumental Perspective Statistics. Critical Reviews in Analytical Chemistry, 50(3), 226-253.

[17] Worster, A., & Haines, T. (2004). Advanced statistics: understanding medical record review (MRR) studies. Academic Emergency Medicine, 11(2), 187-192.

[18] Perer, A., & Shneiderman, B. (2008, April). Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In Proceedings of the SIGCHI conference on Human Factors in computing systems (pp. 265-274).
[19] Brazzale, A. R., Davison, A. C., & Reid, N. (2007). Applied asymptotics: case studies in small-sample statistics (Vol. 23). Cambridge University Press.
[20] Schaid, D. J., & Sommer, S. S. (1994). Comparison of statistics for candidate-gene association studies using cases and parents. American journal of human genetics, 55(2), 402.
[21] Morellato, L. P. C., Alberti, L. F., & Hudson, I. L. (2010). Applications of circular statistics in plant phenology: a case studies approach. In Phenological research (pp. 339-359). Springer, Dordrecht.
[22] Holey, E. A., Feeley, J. L., Dixon, J., & Whittaker, V. J. (2007). An exploration of the use of simple statistics to measure consensus and stability in Delphi studies. BMC medical research methodology, 7(1), 52.