XIIth International Symposium «Intelligent Systems», INTELS'16, 5-7 October 2016, Moscow, Russia

Prediction of engine demand with a data-driven approach

Hudson Francis, Andrew Kusiak*

The University of Iowa, Iowa City, IA 52242-1527, USA

Abstract

Models predicting volume of engine demand from historical data are developed. To accommodate seasonal effects, neural networks and autoregressive integrated moving average (ARIMA) approaches are considered. Previous research on the effectiveness of neural networks to model phenomena with seasonality and trend using raw data has been inconclusive. In this paper, four predictive models for a linear time series with seasonality are developed and their accuracy is studied. Performance of a dummy variable linear regression model, a seasonal ARIMA model, a neural network model using raw historical data, and a hybrid linear model is compared. The seasonal ARIMA and linear regression models are found to perform better than the neural network model. The hybrid linear model is found to outperform the three individual models.

1. Introduction

Prediction of demand is an important part of business management, especially in manufacturing. Accurate forecasts reduce cost through better inventory management. A data set was obtained from a company servicing engines in the US and Canada. This company sells replacement parts for failed engines. These failures cause delays, becoming expensive to the customer. It is important that the turnaround time between the sale of a part and the delivery to the customer is minimized. The goal of this paper is to reduce customer downtime by developing a predictive tool estimating future sales of engines. Predictions are made at two levels of granularity: aggregate level (all parts) and group level (a subset of all parts). Several time horizons are considered for making predictions: one year, six months,

---
* Corresponding author. Tel.: 319-335-5934; fax: 319-335-5669.
  E-mail address: andrew-kusiak@uiowa.edu

one quarter, and one month in advance. This paper compares the aggregate level accuracy of four prediction methods: dummy-variable linear regression, neural network (NN), seasonal auto-regressive integrated moving average (ARIMA), and a linear hybrid predictive model.

The ARIMA method has become a standard time series forecasting tool since it was introduced by Box and Jenkins[1] in 1970. It effectively models linear data with seasonality and trend. Since then, advances in computer processing have allowed data-driven models such as neural networks to gain popularity. Neural networks can handle highly nonlinear data. Hornik[2] found that a multilayer NN can approximate any function given enough hidden nodes. This versatility has led to the application of NN based sales prediction models in a wide range of industries, including electronics[3,4], food[5,6,7], clothing[8], and footwear[9].

Despite their success, NNs do not universally outperform ARIMA models. As Chatfield[10] pointed out, the "best" forecasting method is situational. Ho et al.[11] found that recurrent neural networks and ARIMA models outperformed NN models when predicting compressor failures at a Norwegian power plant. Khashei and Bijari[12] noted that neural networks do not always model linear data well. In addition, using NNs for data with seasonality has yielded mixed results. One option is to use a preprocessing method to deseasonalize the data. Nelson et al.[13] analyzed 68 times series and found that deseasonalized NNs performed much better than NNs that were not deseasonalized. However, Sharda and Patil[14] analyzed 75 times series and concluded that NN models performed at least as well as ARIMA models and did not need to be deseasonalized. Alon et al.[15] found that neural networks generally outperformed ARIMA and linear regression models for US retail sales predictions, but that the ARIMA method was a strong competitor. Furthermore, they concluded that NN models did not require deseasonalization or detrending for that data set. Taskaya-Temizel and Casey[16] claimed that deseasonalization is not necessary if the NN is properly specified, but detrending through differencing will increase the accuracy of the model. Zhang and Qi[17] found that both detrending and deseasonalization preprocessing provided the best NN forecasting results in a case study of retail sales. Chu and Zhang[18] found that deseasonalization improved NN accuracy and that the deseasonalized NN outperformed ARIMA and dummy variable regression models. Finally, Zhang and Qi[19] concluded that NNs are not able to model seasonality directly in a case study of nine data sets. They compared seasonal ARIMA, NN, detrended NN, deseasonalized NN, and detrended and deseasonalized NN models. They found that the deseasonalized and detrended NN model outperformed all other models. Their research indicates that NN models without detrending and deseasonalization may be inferior to seasonal ARIMA models. Zhang and Qi[19] claimed that "a trend time series does not meet the conditions for universal approximation" (p. 513), therefore, preprocessing was necessary.

Individual forecasting methods are best suited for specific data characteristics. For example, ARIMA models can handle seasonality and trend, but cannot handle nonlinear data. Zhang[20] claimed that real data is rarely only linear or nonlinear. In this case, individual models may not be appropriate. Indeed, Khashei and Bijari[21] wrote "if a time series exhibits both linear and nonlinear patterns during the same time interval, neither linear models nor nonlinear models alone are able to model both components simultaneously" (p. 480). Research[22] has shown that combined forecasting methods often outperform individual methods. These combinations do not need to be complex. Clemen[23] observed that simply averaging the results of multiple forecasts can sometimes improve the prediction accuracy. Combined forecasts generally have less variability in accuracy than individual forecasts[24]. Hybrid models are one way to handle seasonality, as well. Tseng et al.[25] created a hybrid seasonal ARIMA and back-propagation NN model and compared it to individual seasonal ARIMA, differenced NN, and deseasonalized NN models. They found that the hybrid model performed best, especially with limited history. Combined forecasts do not always outperform individual forecasts[16,24,26], however. This could be due to the assumed relationship between the linear and nonlinear structures in the data. For example, if a linear-nonlinear relationship is assumed to be additive but is actually multiplicative, the individual model may outperform the hybrid model. Khashei and Bijari[21] presented a hybrid ARIMA-NN model that improves on previous hybrid models because it does not assume an additive relationship. They guaranteed that this method would not be worse than individual NN or ARIMA models and illustrated the results with three empirical examples.

The goal of this paper is to predict future engine sales at several time horizons using four prediction methods. This paper compares those results to determine if the neural network model without deseasonalization or detrending can outperform the seasonal ARIMA model. It also examines whether a hybrid model can outperform the individual models.

2. Data Description

A daily record of replacement part sales from November 1999 through October 2013 for the US and Canada was obtained from a company. All part numbers were combined to form an aggregate data set. Daily sales were combined by calendar month to create monthly sales spreadsheets. Figure 1 shows the aggregate level monthly sales from November 1999 through October 2013. May and June 2011 were identified as outliers. During these months, the company underwent a system change and was not manufacturing parts. The sales during these months were exclusively from inventory. Company project contacts indicated this was a planned event and should not be incorporated into the prediction model. Those two values were changed to the estimated linear regression values, as discussed by Tan[26], to avoid propagating errors due to the system change for subsequent predictions.

The graph in Figure 1 represents fourteen years of data. The data was divided into training and testing data sets. The size of the training set required for each prediction method varies. Therefore, some methods use smaller testing data sets than others. Since all methods have at least four years of testing data available, model comparison is done using fiscal years 2010-2013 (11/2009-10/2013). Figure 1 shows a gradual increasing trend in sales over time. In addition, a seasonal trend is present. This is due to the seasonal nature of industries served. Figure 2 depicts the mean aggregate sales by month for the training data set. The largest number of mean sales occur in May and October. The fewest number of mean sales occur in December, when weather and holidays decrease equipment use. This monthly seasonality is considered in the development of the linear regression, ARIMA, and hybrid models. The actual annual trends can vary from year to year, however. Figure 3 shows the monthly sales of the training data by fiscal year. The seasonal variability between fiscal years can be attributed to a variety of factors that affect the industries served, including weather, economic influences, and machine populations and age. For example, a long winter may delay the use of some equipment and affect the seasonal sales patterns of replacement parts

3. Methods

3.1. Dummy Variable Linear Regression

Although relatively simple, linear regression remains a popular and useful forecasting tool in many applications. In linear regression, a linear equation is used to represent the relationship between a dependent variable and one or more explanatory variables. The method of least squares is most commonly used to identify the linear equation by minimizing the sum of the squared residuals of the fitted value and actual value. An additive common-slope linear regression model with a dummy-variable regressor, as described by Fox[27], is shown in equation (1).

$$y(i) = \alpha + \beta X_i + \gamma D_i + \varepsilon_i \tag{1}$$

where α is the intercept for the baseline model, β is the coefficient for the quantitative explanatory variable X, γ is the coefficient for the dummy variable regressor D, and ε is the error. The dummy variable regressor D is coded such that $D = 1$ or $D = 0$. For the engine sales data set, the dummy variable regressor allows the model to capture seasonality by effectively changing the intercept for each month.

3.2. Neural Network

A neural network is a nonlinear machine learning construct consisting of a set of nodes and weighted links. Neural networks are useful forecasting tools with the ability to find nonlinear patterns within data. A general feedforward multilayer neural network, as described by Duda et al. [28], is shown in equation (2).

$$z_i = f \left(\sum_{j=1}^{n_H} w_{kj} f \left(\sum_{i=1}^{d} w_{ij} x_i + w_{j0} \right) + w_{k0} \right) \tag{2}$$

where z_k is the output or predicted value, n_H is the number of hidden nodes, d is the number of dimensions in the input layer, w_{ji} is the weight from input node i to hidden node j, w_{kj} is the weight from hidden node j to the output node k, x_i is the value of feature i, w_{j0} is the weight of the bias term from the input layer to the hidden node j, w_{k0} is the weight of the bias term from the hidden layer to the output node k, and f is an activation function. For this paper, the identity function and the logistic function were used as activation functions. The logistic function is shown in equation (3).

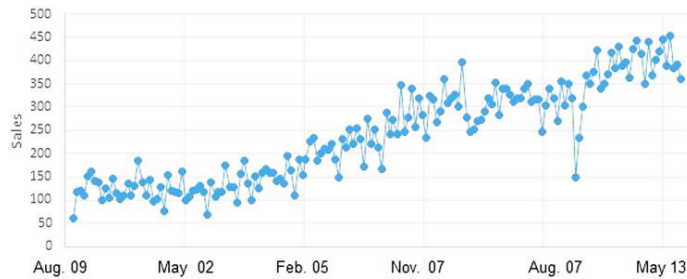$$f(x) = \frac{1}{1 + e^{(-x)}} \tag{3}$$

Fig. 1. Aggregate level monthly part sales (11/1999 - 10/2013)
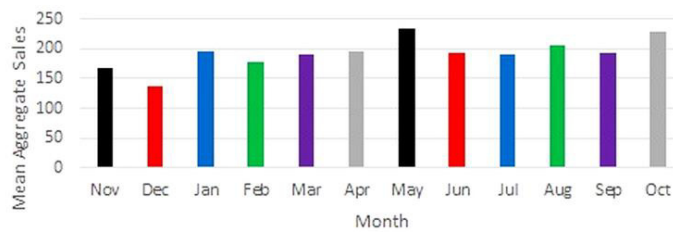

Fig. 2. Mean aggregate training data sales by month (11/1999 - 10/2009)
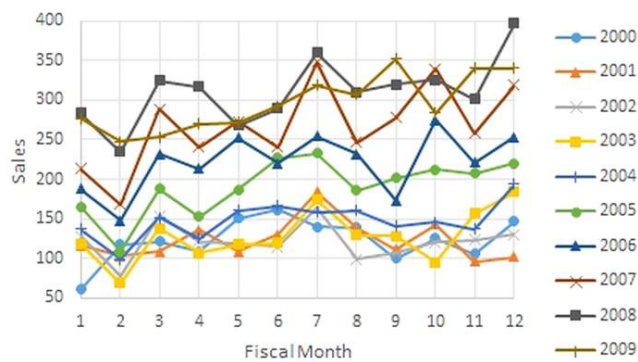

Fig. 3. Aggregate training data sales by fiscal month (11/1999 - 10/2009)

The weights of a neural network are optimized using a gradient descent algorithm. They are initialized with random values. An iterative loop uses the back-propagation technique to adjust weights to minimize the total sum of squared

errors in a set of training data. A momentum term can be used to avoid getting stuck at a local minimum. The accuracy of the model is estimated using a separate testing data set.

3.3. Seasonal Autoregressive Integrated Moving Average

An ARIMA model is a linear time series forecasting tool based on autocorrelations in the data. The order of the autoregressive polynomial is determined by examining the partial autocorrelation function. The order of the moving average polynomial is determined by examining the autocorrelation function. Stationarity is sometimes achieved through differencing. A general multiplicative seasonal ARIMA model, as described by Box et al. [29], is shown in equation (4).

$$\varphi_p\left(B\right)\Phi_P\left(B^s\right)\nabla^d\nabla_s^D z_t = \theta_q\left(B\right)\Theta_Q\left(B^S\right)a_t \quad (4)$$

where

$$\varphi_p\left(B\right)=1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^P \; , \; \varphi_p\left(B\right)=1-\varphi_1 B-\varphi_2 B^2-...-\varphi_p B^P \; ,$$

$$\Phi_P\left(B^s\right)=1-\Phi_s B^s-\Phi_{2s}B^{2s}-\cdots-\Phi_{Ps}B^{Ps} \; , \; \nabla^d=\left(1-B\right)^d \; , \; \nabla_s^D=\left(1-B^s\right)^D \; ,$$

$$\theta_q\left(B\right)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q \; , \Theta_Q\left(B^s\right)=1-\Theta_s B^s-\Theta_{2s}B^{2s}-\cdots-\Theta_{Qs}B^{Qs}$$

and B is the backward shift operator, z_t is predicted value at time t, and a_t is a white noise series from a fixed distribution with zero mean and variance σ_a^2. An ARIMA $\left(p,d,q\right)\times\left(P,D,Q\right)_s$ model with seasonal period s has a nonseasonal autoregressive order of p, a seasonal autoregressive order of P, a nonseasonal moving average order of q, a seasonal moving average order of Q, a nonseasonal difference of order d, and a seasonal difference of order D.

3.4. Metrics of Prediction Accuracy

Four different models are used to make monthly sales predictions. Each model is compared using mean absolute percentage error (MAPE) and total error (TE). Absolute error (AE), MAPE, and TE are shown in equations (5).

$$AE=\frac{|\hat{y}-y|}{y}\times 100\%, \; MAPE=\frac{\sum_{i=1}^{N}AE(i)}{N}, \; TE=\frac{\sum_{i=1}^{N}\hat{y}(i)-\sum_{i=1}^{N}y(i)}{\sum_{i=1}^{N}y(i)}\times 100\%, \quad (5)$$

where yˆis the predicted sales, y is the observed sales, and N is the number of data points in the time period of interest. The MAPE of the second equation is used to compute monthly MAPE and annual MAPE for the four year testing period. The TE of third equation is analogous to an inventory system. An over prediction one month is balanced by an under prediction for a subsequent month. The third equation is used to compute the TE for each year and the TE for the entire four year testing period.

4. Computational Results

4.1. Dummy Variable Linear Regression

An additive dummy variable linear regression model was fit to the sales data by the ordinary least squares method using R software. One quantitative explanatory variable (year) was used. Eleven qualitative dummy-variables (months) were used to account for seasonality. Each dummy variable was coded as 1 or 0. For example, a dummy

variable representing the month of November would equal 1 if the sales of a data point occurred in November and 0, otherwise. The interaction between month and year was found to be insignificant, and therefore not included in the model. Nine to ten years of the training data was used to specify the initial model parameters, depending on the prediction time horizon. Four years of testing data was used to assess the prediction accuracy. A series of loops was constructed to make each prediction at each time horizon. Predictions were rounded to the nearest integer. After each prediction, one month of "observed" sales was moved from the testing data set to the training data set. All model parameters were updated to reflect this new information before making subsequent predictions. The prediction error at each time horizon is shown in Table 1. TE is total error, MAPE is mean absolute percentage error.

Table 1. Error produced by linear regression model (11/2009-10/2013)

Prediction Time Horizon	Linear Regression		
	TE (%)	Annual MAPE (%)	Monthly MAPE (%)
1 month	-1.8	3.9	7.0
3 months	-1.8	4.0	7.1
6 months	-1.9	4.2	7.2
12 months	-2.0	4.5	7.2

4.2.   Neural Network

A neural network was developed using time series regression in Statistica software. The data was not preprocessed to remove seasonality or trend. A set of features was created using past sales at different lags. Feature selection from the time series was done with an exhaustive algorithm on the training data set using Weka software (Hall et al., 2009). Ten features were used to build a multilayer perceptron with 19 hidden nodes and one output node. A logistic activation function was used for the hidden nodes and an identity function was used for the output node. All predictions were rounded to the nearest integer. Four years of testing data were used to assess the prediction accuracy. Table 2 shows the error at each prediction time horizon rounded to the nearest tenth of a percent.

Table 2. Error produced by NN model (11/2009-10/2013)

Prediction Time Horizon	NN		
	TE (%)	Annual MAPE (%)	Monthly MAPE (%)
1 month	-0.3	5.0	8.8
3 months	-0.9	5.8	9.4
6 months	-2.0	6.3	8.9
12 months	-4.2	7.1	9.9

4.3.   Autoregressive Integrated Moving Average

A pool of four seasonal ARIMA models were developed using R software. Nine years of training data was analyzed to manually determine the order of each model using autocorrelation plots, statistical tests, and residual analysis. An automatic model specification function was tested but not used due to inferior results. The initial parameters for each model were specified using the training data. Predictions were rounded to the nearest integer. For each subsequent prediction, a loop was created to update the model parameters using the conditional sum of squares for each fixed order. Traditionally, ARIMA model comparison is done using the Akaike Information Criterion (AIC) (Cryer and Chan, 2008). However, since the pool of models used different data sets (transformed and untransformed), AIC could not be used. Instead, TE, MAPE, and residual analysis were used to determine the best ARIMA model. The final model chosen was of order $(0,1,1) \times (3,1,1)_{12}$ with the seasonal AR(1) and AR(2) parameters fixed to zero. The predicted error for this model at each time horizon is shown in Table 3.

Table 3. Error produced by ARIMA model (11/2009 - 10/2013)

Prediction Time Horizon	ARIMA (0, 1, 1) X (3, 1, 1)$_{12}$		
	TE (%)	Annual MAPE (%)	Monthly MAPE (%)
1 month	-0.7	1.6	7.0
3 months	-0.8	2.2	7.5
6 months	-1.3	2.0	8.2
12 months	-1.5	2.0	7.6

4.4. *Linear Hybrid Model*

Several hybrid models were developed in an attempt to improve upon the accuracy of individual models. From literature, it was determined that a hybrid ARIMA-NN model could capture seasonality and trend, as well as the linear and nonlinear aspects of the data. Hybrid models need multiple training data sets, however. The ARIMA model requires several years of training data to capture seasonal trends. This left an insufficient number of data points in the NN training set to accurately predict the four years of testing data. Due to this limited history, the hybrid ARIMA-NN model was less accurate than the individual NN and ARIMA model predictions.

A hybrid linear model was developed using dummy variable linear regression and nonseasonal ARIMA. Although neither method is able to capture nonlinearity in the data, this hybrid model improved upon the prediction accuracy of individual methods. A loop was created using R programming language to predict sales using dummy variable linear regression. The residuals from this model were fed into another loop and used to develop a nonseasonal ARIMA model. A built in function from the forecast package was used to specify the order and coefficients of the ARIMA model. The predicted residuals from the ARIMA model were added to the predicted sales from the linear regression model to create a hybrid sales prediction. Table 4 shows the prediction error of the hybrid model at each prediction time horizon. The smallest prediction error occurs at the six month prediction time horizon. At that horizon, the TE and monthly MAPE are lower than any other model tried.

Table 4. Hybrid linear error (11/2009 - 10/2013)

Prediction Time Horizon	Hybrid linear		
	TE (%)	Annual MAPE (%)	Monthly MAPE (%)
1 month	0.8	3.0	7.4
3 months	0.4	2.8	7.5
6 months	0.0	1.8	6.7
12 months	-0.3	1.9	6.8

5. Conclusion

In this paper, four models predicting engine demand were developed. Predictions were made at multiple time horizons. Model accuracy was compared using the following metrics: total error, annual mean absolute percentage error, and monthly mean absolute percentage error. These metrics were used to analyze the accuracy of linear, nonlinear, and hybrid algorithms on a seasonal time series with trend. The results presented in the literature were inconclusive whether neural networks could handle this type of data without preprocessing. We determined that the neural network model performed worse than the dummy variable linear regression, seasonal ARIMA, and hybrid linear models. This supports Khashei and Bijari's[12] claim that neural networks do not always model linear data well and Zhang and Qi's[19] claim that neural networks are not able to model seasonality directly. Many researchers have found that combining forecasting methods improves prediction accuracy. Ideally, the hybrid model would include both linear and nonlinear aspects. The limited available training data prevented us from developing an effective neural network ARIMA hybrid model. Instead, we considered a dummy variable linear regression-ARIMA hybrid model.

Despite the similarity of the models, this hybrid model outperformed all individual models. The best results were achieved at the six month time horizon. The algorithm is constructed to update model parameters as new information is received. This allows the model to adapt to changes in the data stream over time. Future research includes developing a neural network with detrended and deseasonalized data and examining more linear-nonlinear hybrid models.

References

1. Cryer, J. D., & Chan, K. S. Time Series Analysis With Applications in R (2nd ed.). New York: Springer; 2008.
2. Hornik, K., Stinchcombe, M., & White, H. Multilayer feedforward networks are universal approximators. Neural Networks; 1989; **2**: 359-66.
3. Chang, P. C., Liu, C. H., & Fan, C. Y. Data clustering and fuzzy neural network for sales forecasting: A case study in printed circuit board industry. Knowledge-Based Systems; 2009; **22**: 344-55.
4. Luxhøj, J. T., Riis, J. O., & Stensballe, B. A hybrid econometric-neural network modeling approach for sales forecasting. International Journal of Production Economics; 1996; **43**: 175-92.
5. Chen, C. Y., Lee, W. I., Kuo, H. M., Chen, C. W., & Chen, K. H., The study of a forecasting sales model for fresh food. Expert Systems with Applications; 2010; **37**:7696-702.
6. Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. Journal of Food Engineering; 2006; **75**:196-204.
7. Hamzaçebi, C. Improving artificial neural networks' performance in seasonal time series forecasting. Information Sciences; 2008; **178**:4550-59.
8. Thomassey, S. Sales forecasts in clothing industry: The key success factor of the supply chain management. International Journal of Production Economics; 2010; **128**:470-83.
9. Das, P., & Chaudhury, S. Prediction of retail sales of footwear using feedforward and recurrent neural networks. Neural Computing and Applications; 2007; **16**:491-502.
10. Chatfield, C. What is the 'best' method of forecasting?" Journal of Applied Statistics; 1988. **15**: 19-38.
11. Ho, S. L., Xie, M., & Goh, T.N. A comparative study of neural network and Box-Jenkins ARIMA modeling in time series prediction. Computers & Industrial Engineering; 2002.\; **42**:371-5.
12. Khashei, M., & Bijari, M. An artificial neural network (p, d, q) model for timeseries forecasting. Expert Systems with Applications; 2010. 37:479-89.
13. Nelson, M., Hill, T., Remus, W., & O'Connor, M. Time series forecasting using neuralnetworks: Should the data be deseasonalized first? Journal of Forecasting; 1999; **18**: 359-67.
14. Sharda, R., & Patil, R. B., Connectionist approach to time series prediction: An empirical test. Journal of Intelligent Manufacturing; 1992; **3**: 317-23.
15. Alon, I., Qi, M., & Sadowski R. J. Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. Journal of Retailing and Consumer Services; 2001; **8**: 147-56.
16. Taskaya-Temizel, T., & Casey, M.C. A comparative study of autoregressive neural network hybrids. Neural Networks; 2005. 18:781-789.
17. Zhang, G. P., & Qi, M. Predicting consumer retail sales using neural networks. In Neural Networks in Business: Techniques and Applications Hershey, PA: Idea Group Publishing; 2002. p. 26-40
18. Chu, C. W., & Zhang, G. W. A comparative study of linear and nonlinear models for aggregate retail sales forecasting. International Journal of Production Economics; 2003; **86**: 217-31.
19. Zhang, G. P., & Qi, M. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research; 2005; **160**:501-14.
20. Zhang, G. P. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing; 2003 **50**: 159-75.
21. Khashei, M., & Bijari, M. A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Applied Soft Computing; 2011; **11**:2664-76
22. Chatfield, C. What is the 'best' method of forecasting?" Journal of Applied Statistics; 1988; **15**: 19-38.
23. Clemen, R. T. Combining forecasts: A review and annotated bibliography. International Journal of Forecasting; 1989; **5**: 559-83.
24. Hibon, M., & Evgeniou, T. To combine or not to combine: Selecting among forecasts and their combinations. International Journal of Forecasting; 2005; **21**: 15-24.
25. Terui, N., & van Dijk, H. K. Combined forecasts from linear and nonlinear time series models. International Journal of Forecasting; 2002; **18**: 421-38.
26. Tan, P. N., Steinbach, M., & Kumar, V. Introduction to Data Mining. Boston: Pearson Education, Inc; 2006.
27. Fox, J. Applied Regression Analysis and Generalized Linear Models (2nd ed.). Los Angeles: Sage Publications, Inc; 2008.
28. Duda, R. O., Hart, P. E., & Stork, D. G. Pattern Classification (2nd ed.). New York: John Wiley & Sons, Inc; 2001.
29. Box, G. E. P, Jenkins, G. M., & Reinsel, G. C. Time Series Analysis Forecasting and Control (3rd ed.). Englewood Cliffs, New Jersey: Prentice Hall; 1994.