Forecasting Rail Transport Petroleum Consumption Using an Integrated Model of Autocorrelation Functions-Artificial Neural Network

Hanialhossein Abolfazli[1], Seyed Masoud Asadzadeh[2], Salman Nazari-Shirkouhi[3,*] Seyed Mohammad Asadzadeh[4], Kamran Rezaie[5]

[1]PMP®, RMP®, Business Development Division, Enermonde AB. Signe Tillischgatan 15, Stockholm, Sweden, 16973, Solna, Email: hani.abolfazli@enermonde.com

[2]Department of Software Engineering, University of Science & Culture, Tehran, Iran. P.O. Box: 13145-871; Email : masoud.asadzadeh@hamfekrangroup.org

[3,*]Young Researchers and Elite Club, Roudbar Branch, Islamic Azad University, Roudbar, Iran, P.O. Box 14579-44616, E-mail: snnazari@ut.ac.ir

[4] School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Tehran, Iran. P.O. Box 11155-4563. Email: smasadzadeh@ut.ac.ir

[5]School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Tehran, Iran. Email: krezaie@ut.ac.ir , P.O. Box 11155-4563

Abstract: This paper presents the application of time-series and artificial neural network for improvement of energy forecasting in rail transport section. An integrated artificial neural network (ANN) model is presented that uses autocorrelation and partial autocorrelation functions to determine the best input variables for ANN. The proposed ANN uses autocorrelation function (ACF) and partial autocorrelation function (PACF) extracted from time-series data to select appropriate inputs for ANN. For validation of the ANN results, they are compared with two auto regressive models using analysis of variance (ANOVA) technique. Weekly total gasoil consumption in Iran railway transportation from January 2009 to October 2011 is used for constructing and comparison time-series and ANN models. Result shows that ANN with inputs extracted from ACF and PACF analysis, performs well for estimation of railway gasoil consumption. The interaction between input variables as well as non-linearity in consumption data is properly handled by the proposed integrated ANN.

Keywords: Gasoil Consumption Forecasting; Railway; Artificial Neural Network; Time-series; Autocorrelation

1 Introduction

Economic development boosts different activities amon them transport activity. One of the pressing problems related to this increasingly important economy sector, is petroleum dependence. A great portion of energy (around 95%) used in the transportation sector is supplied by petroleum. Therefore, managing the use of petroleum in this sector is crucial for secure supply of energy.

During the last decade several new techniques have been used for energy demand management and specifically for accurate prediction of the future energy needs. Suganthi and Samuel [15] reviewed the existing models of energy demand forecasting. Among them, they discussed traditional methods as well as soft computing techniques such as neuro-fuzzy, GA, and ANNs. As new techniques of energy demand forecasting they discussed Support vector regression, ant colony and particle swarm [15]. These models can be categorized into three main approaches: time-series approach [3, 9], econometric approach [7, 12,] and artificial intelligence (AI) approach [14].

This variety of techniques also can be found in the literature related to the modeling and forecasting energy demand in transportation sector. Haldenbilen and Ceylan [10] used genetic algorithm (GA) and developed three forms of the energy demand equations (linear, exponential and quadratic) to improve transport energy estimation. Their model uses population, gross domestic product and vehicle-km as inputs. Forouzanfar et al. [4] proposed a multi-level genetic programming approach for demand forecasting of transport energy in Iran. It was shown that multi-level genetic programming approach outperforms neural network and fuzzy linear regression approaches for transport energy forecasting. Al-Ghandoor et al. [1] used ANFIS-double exponential smoothing to model and forecast the demand of transport energy in Jordan. Petrol and diesel consumption in the transpor sector of China is analyzed by Chai et al. [2]. They used the Bayesian linear regression integrated with Markov Chain Monte-Carlo method to establish a demand-forecast model.

AI techniques are increasingly diversifying today and among them Artificial Neural Networks have gained special attention in energy forecasting. Geem and Roper [6] developed an ANN model with four independent variables of *i*) gross domestic product (GDP), *ii*) population, *iii*) import, and *iv*) export to forecast transport energy demand. They showed that ANN performs better than linear regression model or an exponential model. Geem [5] developed ANN models to forecast South Korea's transport energy consumption and demonstrated that ANN produced more robust results. Murat and Ceylan [11] illustrates an ANN approach for the transport energy demand forecasting using socio-economic and transport related indicators.

Generally, econometric models which take into account the socio-economical variables are more powerful than time-series models. However, it should be noted

that econometric models need accurate data regarding its explanatory variables (e.g. production data, stock exchange market data, etc). In some countries, especially in developing countries accessing to accurate and reliable data is limited. Moreover, for forecasting purposes, the problem of forecasting the explanatory variables (e.g. production data, stock exchange market data, etc) is central and in fact an unsolved sub-problem for the main problem of forecasting the dependent variable (here gasoil consumption). Therefore, time-series models that directly model and forecast the dependent variable are of great interest because they don't need data of other variables. Time-series models may not capture the non-linearity and complex relationships in the modeling environment. Because of that, time-series modeling concept is integrated with ANN to improve forecasting in more complex cases.

To the best knowledge of the authors, the application of time-series integrated with ANN modeling is limited for railway transport energy forecasting and there is a research gap to study the advantages of combining time-series modeling and ANN modeling to reach improved models for railway energy forecasting.

The main objective of this paper is to use ACF and PACF of time-series data to construct ANN model to be used for gasoil consumption forecasting in rail transport sector. To this end, this paper presents a working algorithm to analyse autoregressive time series data and decide on input variables for ANN.

This paper is organized as follows. In section 2, the integrated ANN algorithm is described. Section 3 presents a case study which illustrates an experiment with the proposed algorithm. Section 4 presents the comparison, validation, and analysis of the results. Conclusions are presented in the last section.

2 Working Algorithm for ANN Modeling

In this section the working algorithm for ANN modeling is presented. At its early stages, this algorithm proposes to collect data and preprocess them using ACF and PACF. This preprocessing involves the calculation of confidence interval for ACF and PACF. With refer to Gujarati (2004), the algorithm decides on the most appropriate autoregressive moving average (ARMA) model. This decision is made based on the pattern in ACF and PACF. Table 1 provides the guidelines to choose ARMA models [8].

Moreover, significant ACFs and PACFs can be specified with reference to their confidence intervals. In the integrated algorithm of this paper, the inputs of ANN are selected as the lagged variables of the output. For modeling gasoil consumption, ANN output is the amount of gasoil consumed in time period t (Y_t). It is noted that the number of lagged variables that should be selected as ANN inputs is of question.

Table 1
How to choose ARMA model and its parameters

PACF pattern / ACF pattern	- Exponentially decreasing	- Very big for *p* lags and cut after *p* lags
- Exponentially decreasing	ARMA (p,q)	AR (p)
- Oscillating and damping sin curve	-	AR (p)
- Very big for *q* lags and cut after *q* lags	MA (q)	-

p is the number of autoreggresive lags and *q* is the number of moving averages

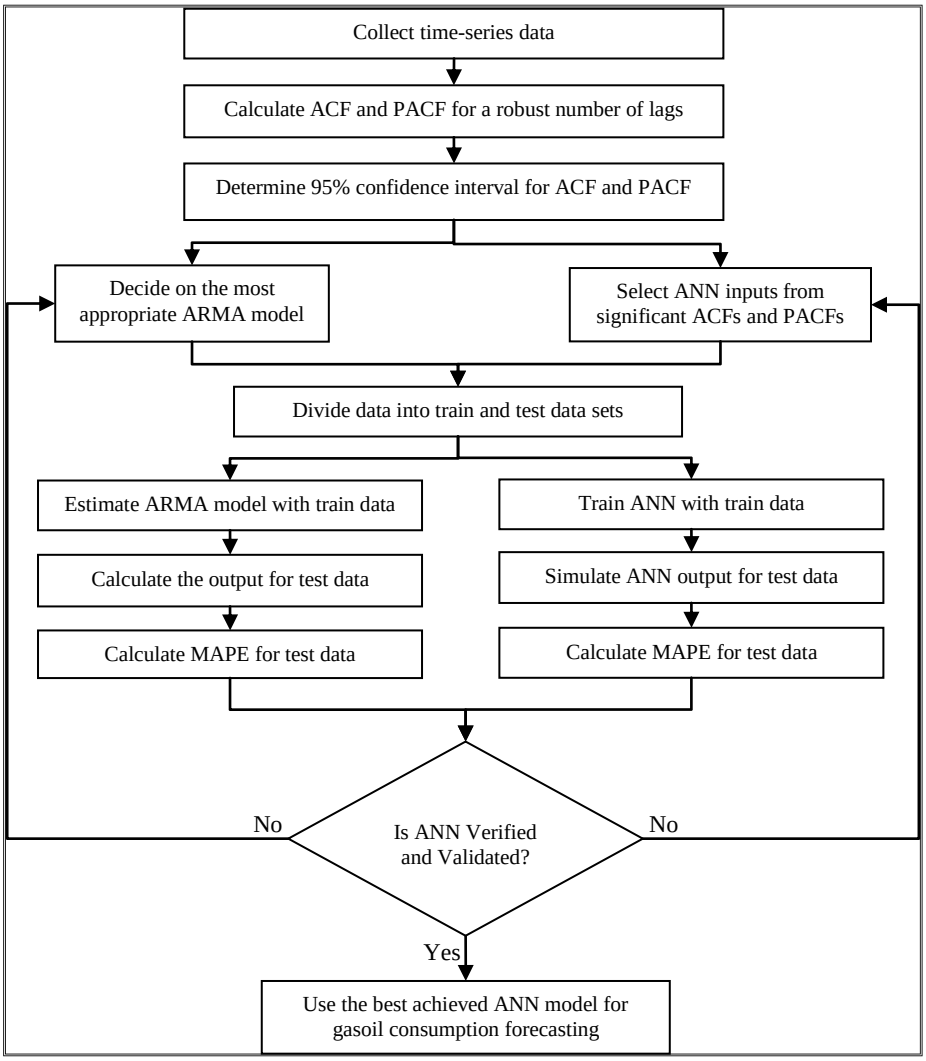

Figure 1
Working Algorithm for ANN Modeling

The inputs of ANN are specified as the lag variables which their ACF or PACF is significant. In some cases, the number of significant ACF or PACFs may be very high. In these cases, a robust number of input variables can be preset by the modeler. This procedure will determine the best inputs that could be used in ANN modeling. After that the data set is divided into two sets, train and test. With the train set, both ARMA and ANN models are trained and estimated and the performance of the models is examined with test data according to their Mean Absolute Percentage Error (MAPE). Finally, the result of ANN is verified and validated by comparison with the result of ARMA model. If the results be satisfactory, ANN could be used for gasoil consumption forecasting. Otherwise, more input variables should be used for both ANN and ARMA models. This integrated algorithm is presented in Figure 1.

2.1 Artificial Neural Network

Neural Networks can be configured in various arrangements to perform a range of analysis including function estimation and forecasting. ANNs are well suited for applications where we need to estimate a nonlinear function. ANNs consists of an inter-connection of a number of neurons. In Multi Layer Perceptron (MLP) network the data flows forward to the output continuously without any feedback. Figure 2 shows a two-layer feed forward model used for gasoil forecasting. The input nodes are the previous *m* time lagged consumptions while the output is the gasoil consumption for the current time period. Hidden nodes with linear or nonlinear transfer functions provide a basis for estimate the non-linear relation between inputs and output.

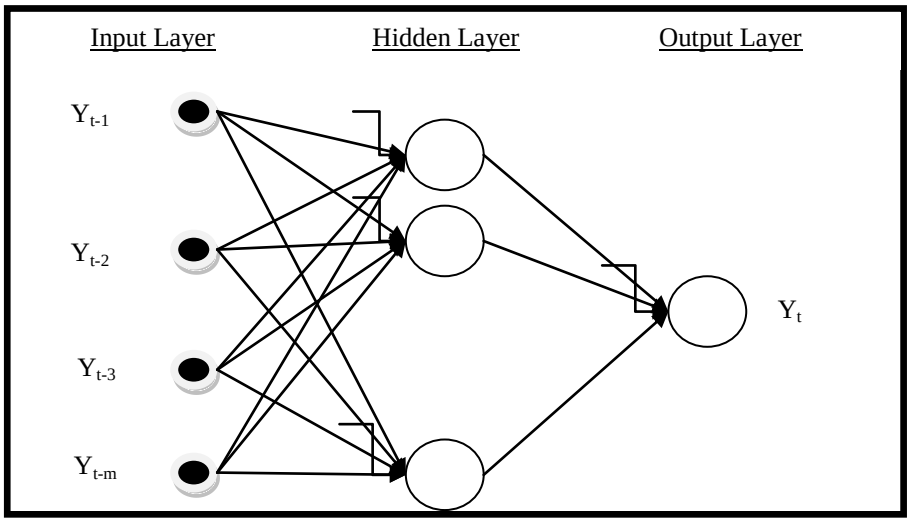

Figure 2
The structure of the ANN

The model can be written as:

$$y_t = \alpha_0 + \sum_{j=1}^{n} \alpha_j f\left(\sum_{i=1}^{m} \beta_{ij} y_{t-i} + \beta_{0j}\right) + \varepsilon_t \quad (1)$$

where m is the number of input nodes, n is the number of hidden nodes. f is a transfer function that can be tansig with $f(x) = 2/(1+\exp(-2x)) - 1$ or logsig with $f(x) = 1/(1+\exp(-x))$. Moreover, $\{\alpha_j,\ j = 1, \ldots, n\}$ is a vector of weights from the hidden to output nodes and $\{\beta_{ij},\ i = 1, 2, \ldots, m;\ j = 0, 1, \ldots, n\}$ are weights from the input i to hidden node j. α_0 and β_{oj} are bias weights. Note that in Equation (1) a linear transfer function is employed for the output layer and this is a good choice for forecasting problems. The MLP's most popular learning rule is the error back propagation algorithm. Back Propagation learning is a kind of supervised learning introduced by Werbos [16] and later developed by Rumelhart and McClelland [13].

2.2 Dividing Data into Train and Test Sets

There are two aspects in the procedure of data division that should be noted. First, the percentage of test data (hereafter call it Test Ratio) and second, which of data rows should be selected for test (hereafter call it Test Selection). Here, Test Ratios is selected randomly between 10%, 20%, 30%, 40%, and 50%. Block selection is an appropriate selection rule for forecasting purposes. In block selection, test data are selected from the most recent data rows.

2.3 Error Estimation Method

Error Estimation Method used in this study is Mean Absolute Percentage Error (MAPE). It can be calculated by the following equation:

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{x_t - x'}{x_t} \right| \quad (2)$$

In (2), x and x' are actual and estimated data, respectively. Scaling the output, MAPE method is the most suitable method to estimate the relative error because input data may have different scales.

3 Case Study

3.1 Gasoil Consumption in Iranian Railway Transport

To show the applicability and usefulness of the proposed ANN, an experiment of rail gasoil consumption estimation is conducted. Weekly total gasoil consumption in railway transportation of Iran from January 2009 to October 2011 is collected. The time-series is shown in Figure 3.

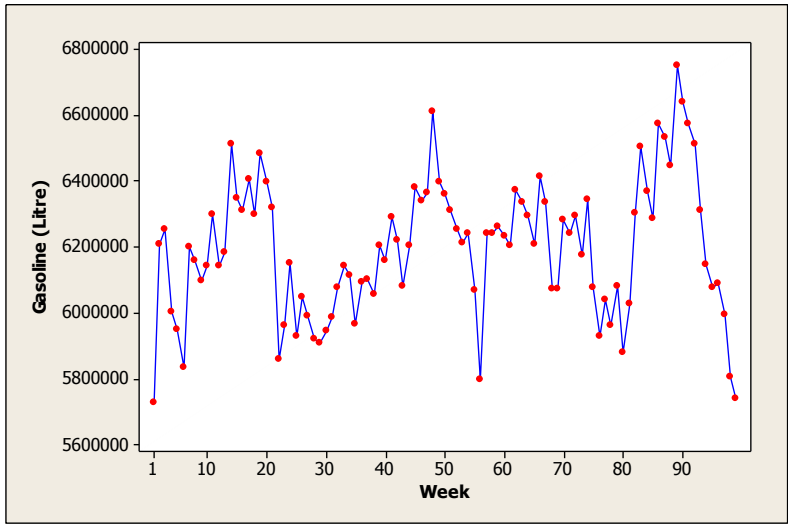

Figure 3
Weekly total gasoil consumption in railway transportation of Iran

3.2 ACF and PACF Analysis

Figure 4 shows the ACF and PACF of the gasoil consumption data. This figure also presents the lower and upper bound of 95% confidence intervals for the significance of ACF and PACF. In other words, those ACFs or PACFs which lie between these two bounds are not significant. For example the first four ACFs are significant but not the fifth ACF.

Analysis of the ACF graph shows that ACF has undergone a sin curve which is damped for lags after 15 (Y_{t-15}). For all lags, except for the first one, PACF has damped and the majority of partial autocorrelations are not significant or if significant, they are very close to the boundaries. Therefore, with respect to the guidelines in Table 1, this suggests that an autoregressive model AR(4) is the best time series for the data series under study.

3.3 Specifying ANN Input Variables

According to the significant ACFs, we can see that lagged variables 1, 2, 3, 4, 8, 9, 10, 11, 12, 13, and 14 are significant and can be used as ANN inputs. However, to have a robust number of input variables we select the first four variables as ANN inputs. Therefore, four lagged variables Y_{t-1}, Y_{t-2}, Y_{t-3}, and Y_{t-4} are selected as ANN inputs. In other words, gasoil consumption in 1 week before (Y_{t-1}), 2 weeks before (Y_{t-2}), 3 weeks before (Y_{t-3}) and 4 weeks before (Y_{t-4}) have been selected as ANN inputs where the gasoil consumption in the current week (Y_t) is the output.

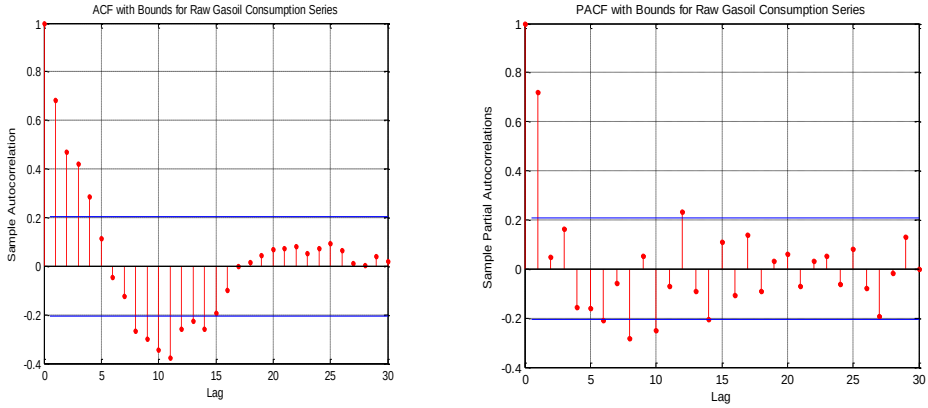

Figure 4
ACF and PACF for Gasoil consumption time-series

3.4 Auto-Regressive (AR) Modeling

Two autoregressive models are constructed for modeling gasoil data. The first model is a autoregressive model with 4 lags, AR(4), because the first four ACFs are the greatest of all. AR Model 1 is as equation (4)

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \varepsilon_t \tag{3}$$

The second autoregressive is an extension of AR Model 1 in which the interactions between every pair of lags are considered as explanatory variables. This formulation allows us to model the non-linear effects of interaction between variables.

$$\begin{aligned} Y_t = {} & \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + \beta_4 Y_{t-4} + \\ & \gamma_{12} Y_{t-1} Y_{t-2} + \gamma_{13} Y_{t-1} Y_{t-3} + \gamma_{14} Y_{t-1} Y_{t-4} + \gamma_{23} Y_{t-2} Y_{t-3} + \\ & \gamma_{24} Y_{t-2} Y_{t-4} + \gamma_{34} Y_{t-3} Y_{t-4} + \varepsilon_t \end{aligned} \tag{4}$$

4 Results and Analysis

4.1 ANN Results

The ANN model is used for estimation of weekly gasoil consumption and shows satisfactory results in terms of MAPE for test data. ANN model is trained with two different hidden transfer functions tansig and logsig (see section 2.1). Another parameter that is suspected to affect ANN performance is the number of neurons in the hidden layer. Different number of hidden neurons between 10 and 50 are considered to find the best ANN. Tables 2 and 3 show the results of MAPE for these ANNs and the best number of neurons in the hidden layer.

Table 2
Test MAPE of ANN with tansig hidden transfer function

Percent of test data	Hidden neurons					Best MAPE
	10	20	30	40	50	
10	2.2%	2.2%	2.5%	2.2%	1.8%	1.8%
20	2.3%	2.3%	2.3%	2.4%	2.5%	2.3%
30	2.0%	2.6%	2.1%	2.5%	2.5%	2.0%
40	2.1%	2.1%	2.2%	2.2%	2.2%	2.1%
50	2.3%	2.4%	2.4%	2.3%	2.3%	2.3%

Table 3
Test MAPE of ANN with logsig hidden transfer function

Percent of test data	Hidden neurons					Best MAPE
	10	20	30	40	50	
10	2.1%	1.4%	2.0%	1.6%	1.7%	1.4%
20	2.4%	2.3%	2.4%	2.6%	2.4%	2.3%
30	2.5%	2.6%	2.5%	2.2%	2.4%	2.2%
40	2.3%	2.2%	2.2%	2.3%	2.2%	2.2%
50	2.1%	2.3%	2.2%	2.3%	2.3%	2.1%

Figure 5 shows the actual and estimated gasoil consumption by ANN-tansig for train and test data when 30% of all data are used as test data. The MAPE error is calculated as 2.5%.

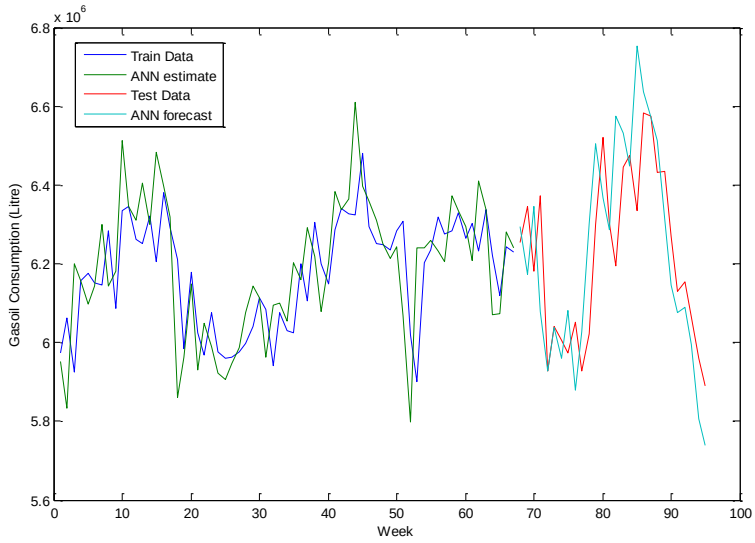

Figure 5
Actual and estimated gasoil consumption by ANN

4.2 AR Model Results

Two autoregressive models 1 and 2 are estimated with use of train data and then the estimated models are used to predict gasoil for test period. For these AR models the percentage of test changes from 10% to 50%. The results of MAPE are reported in Table 4.

Table 4
Test MAPE of ANN with logsig hidden transfer function

Percent of test data	Model 1 (MAPE)	Model 2 (MAPE)
10	2.3%	2.5%
20	2.3%	2.4%
30	2.3%	2.3%
40	2.0%	2.0%
50	1.9%	2.0%

4.3 ANN Verification and Validation

ANN results are verified with respect to the low value for average MAPE. The average MAPE of different ANN models is 2.1% that is a satisfactory result for rail gasoil consumption forecasting. To have a robust validation, we used a

analysis of variance (ANOVA) technique to test the difference between ANN and AR results. Since the percentage of test data could affect MAPE, we design a complete block design with four techniques: ANN-tansig, ANN-logsig, AR model 1, and AR model 2.

The results of ANOVA for this experiment are presented in Table 5. The null hypothesis for this ANOVA is:

H0: the MAPE for all techniques is the same

H1: at least two techniques have different MAPEs

The P-value is the risk of rejecting the null hypothesis *H0*. Because this risk is very high, 71%, we cannot reject *H0* and conclude that the MAPE for all techniques is the same. This conclusion serves as a validation for the ANN results.

Table 5
Two-way ANOVA: MAPE versus Technique, test ratio

Source	DF	SS	MS	F_0	P-Value
Technique	3	0.0000066	0.0000022	0.47	0.710
testratio	4	0.0000200	0.0000050		
Error	12	0.0000564	0.0000047		
Total	19	0.0000830			

Conclusion

This paper presented an integrated autoregressive-ANN algorithm to improve gasoil consumption estimation and forecasting in rail transportation sector. The proposed ANN uses autocorrelation function (ACF) and partial autocorrelation function (PACF) extracted from time-series data to select appropriate inputs for ANN. ANN results are assessed according to an error index namely mean absolute percentage error (MAPE). ANN results are validated by comparison with the results of two linear and non-linear auto regressive models. It was concluded that ANN provides satisfactory results hence it can be used for forecasting purposes. This is a unique study that introduces the application of ACF and PACF analysis for defining ANN inputs in time series modeling. Due to its mechanism, the integrated ANN of this study is capable of dealing data correlation, autocorrelation, complexity, and non-linearity.

References

[1] Al-Ghandoor, A., Samhouri, M., Al-Hinti, I., Jaber, J., Al-Rawashdeh, M., Projection of future transport energy demand of Jordan using adaptive neuro-fuzzy technique, Energy, 2012, 38 (1), pp. 128-135

[2] Chai, J., Wang, S., Wang, S., Guo, J., Demand forecast of petroleum product consumption in the Chinese transportation industry, Energies, 2012, 5 (3), pp. 577-598

[3] Forouzanfar M., Doustmohammadi A., Bagher Menhaj M., Hasanzadeh S., Modeling and estimation of the natural gas consumption for residential and commercial sectors in Iran, Applied Energy, 2010, 87, pp. 268–274

[4] Forouzanfar, M., Doustmohammadi, A., Hasanzadeh, S., Shakouri G, H., Transport energy demand forecast using multi-level genetic programming, Applied Energy, 2012, 91 (1) , pp. 496-503

[5] Geem, Z.W., Transport energy demand modeling of South Korea using artificial neural network. Energy Policy, 2011, 39, 4644–4650

[6] Geem, Z.W., Roper, W.E., Energy demand estimation of South Korea using artificial neural network. Energy Policy, 2009, 37, 4049–4054

[7] Gorucu, F. B. , Gumrah, F. 'Evaluation and Forecasting of Gas Consumption by Statistical Analysis', Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 2004, 26(3), pp. 267-276

[8] Gujarati D.N., Basic Econometrics, Fourth Edition, McGraw−Hill Companies, 2004

[9] Gutierrez, R. A. Nafidi, R. Gutierrez Sanchez, Forecasting total natural-gas consumption in Spain by using the stochastic Gompertz innovation diffusion model, Applied Energy, 2005, 80, pp. 115–124

[10] Haldenbilen, S., Ceylan, H., Genetic algorithm approach to estimate transport energy demand in Turkey. Energy Policy, 2005, 33, pp. 89–98

[11] Murat, Y.S., Ceylan, H., Use of artificial neural networks for transport energy demand modeling. Energy Policy, 2006, 34, pp. 3165–3172

[12] Primoz Potocnika, Marko Thaler, Edvard Govekar, Igor Grabeca, Alojz Poredos, Forecasting risks of natural gas consumption in Slovenia, Energy Policy, 2007, 35, pp. 4271–4282

[13] Rumelhart D.E., McClelland J.L., Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Foundations, MIT Press, Cambridge, MA. 1986

[14] Shakouri H., G., R. Nadimia and F. Ghaderi, A hybrid TSK-FR model to study short-term variations of the electricity demand versus the temperature changes, Expert Systems with Applications, 2008, 36, pp. 1765-1772

[15] Suganthi, L., Samuel, A.A., Energy models for demand forecasting - A review, Renewable and Sustainable Energy Reviews, 2012, 16 (2) , pp. 1223-1240

[16] Werbos P.I., Beyond Regression: new tools for prediction and analysis in the behavior sciences. Ph.D. Thesis, Harvard University, Cambridge, MA. 1974