JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES
www.journalimcms.org

ISSN (Online) : 2454 -7190 Vol.-16, No.-9, September (2021) pp 74-84 ISSN (Print) 0973-8975

GASOLINE CONSUMPTION PREDICTION VIA DATA MINING TECHNIQUE

Soma Gholamveisy

Department of Industrial Engineering Islamic Azad University, South Tehran Branch, Tehran, Iran

Email: research.consultant12@gmail.com

https://doi.org/10.26782/jmcms.2021.09.00007

(Received: July 7, 2021; Accepted: August 30, 2021)

Abstract

Due to the increasing dependence of human life on energy, it plays a crucial role in the functioning of the various economic sectors of the countries, potentially and actually. Fuel products, especially gasoline, given their importance in the transportation sector, play major roles in the economic growth and development of countries. Hence, the authorities in each country have to control the fuel supply and demand parameters accurately with a more accurate prediction of fuel consumption and proper planning in the direction of consumption. The purpose of this study is to find appropriate methods and approaches for forecasting gasoline consumption in Tehran using data mining methods. For this purpose, daily consumption data of gasoline stations were collected in 5 different regions of Tehran during the period of 2008-2013. Then, these numbers were predicted on a daily, weekly, monthly, and seasonal basis for analyzing the consumption at different time intervals. The standardization method was also used to match the scales. After data pre-processing, gasoline consumption was predicted using the multi-layer perceptron (MLP) neural network method. The gasoline consumption forecast was evaluated based on the mean squared error (MSE), mean, and mean absolute error (MAE) criteria. The results indicate that the artificial neural network (ANN) can accurately predict gasoline consumption in five different regions of Tehran.

Keywords: data mining, gasoline consumption, ANN-MLP, prediction

I. Introduction

The refinery industry is one of the critical and infrastructural industries in every country. With the advancement of information technology (IT), it is now possible to extract signs and symptoms existing in time intervals. Because of the possibility of producing various products such as ethanol and methanol, this industry should be able to exploit all its capacities optimally.

Nowadays, the consumption of light petroleum products, especially gasoline is very important for many countries that face increasing gasoline consumption due to population growth. Gasoline is one of the most widely used consumer products in the world. This liquid is one of the energy carriers that is less productive than other carriers. This product has certain conditions in Iran such that billions of Iranian Rials and dollars annually are spent on its subsidies and imports

Iran has a wealth of energy resources, large oil, and natural gas reservoirs, huge underground mines, and energy potential. However, in Iran, efforts in the management sector focus on energy supply and less attention is paid to energy demand management. Meanwhile, energy demand management and efforts to optimize energy use in all developed countries of the world are among the most important drivers of sustainable industrial development.

The prediction of energy consumption demand can be of great assistance in determining the energy sector policies. Currently, the issue of limiting energy consumption, especially petroleum products such as gasoline, is a hot topic in the government's economic policies. Therefore, the issue of forecasting gasoline consumption has been paid special attention by researchers from various fields of civil engineering, traffic, computer science, chemistry, and organic materials.

Researchers in these fields using IT systems because of the nature of their information technology. Clearly, the use of these systems is not limited to the field of gasoline consumption and these systems are also applied in other areas such as the consumption of other fuels (Diesel fuel, Kerosene, etc.).

Various solutions have been offered for dealing with gasoline and other fuels consumption forecasts. One of the most important approaches for solving this problem is the intelligent approach. This approach is based on a black box that provides a limited interpretation of its operation. An advantage of these methods, such as data mining, is that they do not need many assumptions for solving the problem. Because of using advanced technologies, fuel organizations have a huge body of data. Analyzing these data using data mining techniques can help managers make the right decisions. Since the subject of this research is among the predictive issues in the field of data mining and the field of critical consumer goods, it is closely related to demand management and data mining. One of the methods of data mining in the prediction is artificial neural networks (ANNs). Because of the high ability of ANNs in data analysis, several studies have been done on fuels such as gasoline

Nasser and Badr [III] developed an ANN to forecast optimum gasoline consumption in Lebanon

Baba Zadeh [IX] investigated the hybrid approach of the ANN and time series to optimally predict gasoline consumption. This paper presents a hybrid approach based on an ANN and an autoregressive moving average model for assessing and forecasting gasoline consumption

Simie and Dindarlou [I] predicted fuel consumption of gasoline trucks using the ANN.

Assadi et al [II] predicted gasoline consumption in Fars province using ANN and time series models.

Nasla Toghan and Baizis [VI] forecasted torque and fuel consumption of the gasoline engine using ANNs. This research provides an ANN model for predicting the torque and fuel consumption of a gasoline engine. Rahimi Ajdadi and Abbaspour [III] applied ANN and stepwise regression to predict tractor fuel consumption. Also, ANNs were used to evaluate different gasoline characteristics using radial systems. A genetic algorithm (GA) was utilized as an optimization algorithm to optimize the maximum number of neurons and extend the model. The results show that the developed model is suitable for estimating effective and accurate empirical data. In addition, the comparison between this model and a previously reported model in the literature shows the superiority of this model

As can be seen, regarding the importance of the gasoline issue, several studies have been conducted on gasoline price and consumption forecast using smart methods. In the present study, we extend the ANN-based study of Baba Zadeh [IX] on gasoline consumption. The novelty of the present work is that none of the previous studies has forecasted gasoline consumption on daily, weekly, and monthly bases. Therefore, in this research, demand for gasoline is predicted based on the data from the time series data of the Statistical Center of Iran using the ANN method on daily, weekly, and monthly bases. These data include the amount of gasoline sold at the gasoline stations. The remainder of this multi-section research project is organized as follows. Section 2 shows the stages of the prediction model using data mining. Section 3 is an analysis of the prediction model. Finally, Section 4 provides some concluding remarks

II. Steps of the Gasoline Consumption Demand Forecast Model

The stages of each of these three phases can be summarized as follows:

- Data pre-processing
- Forecasting
- Data post-processing (method evaluation)

The most important feature of this study is introducing a comprehensive model for predicting gasoline consumption (daily, weekly, and monthly) and storing the obtained data in a database. To this end, different configurations of neural network methods were combined.

- The overall structure of the proposed approach is summarized in the following.

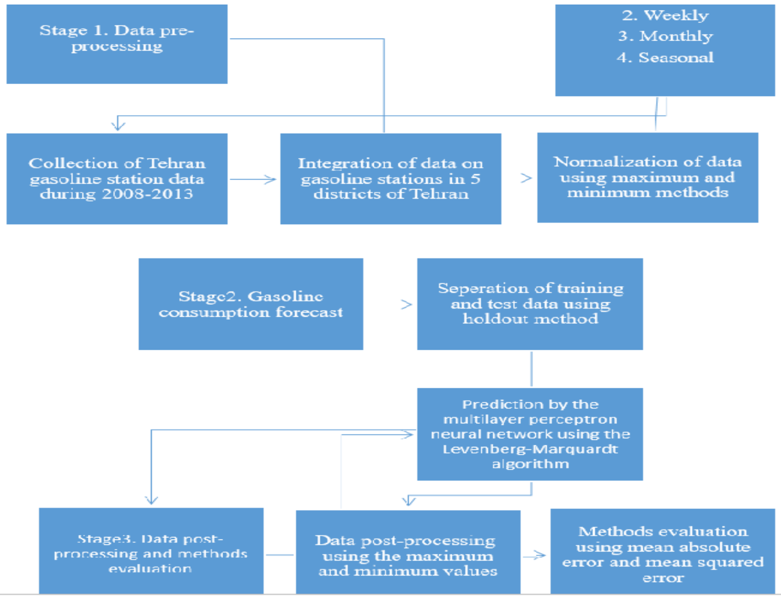
Stage 1. Data pre-processing
2. Weekly
3. Monthly
4. Seasonal
Collection of Tehran gasoline station data during 2008-2013
Integration of data on gasoline stations in 5 districts of Tehran
Normalization of data using maximum and minimum methods
Stage2. Gasoline consumption forecast
Seperation of training and test data using holdout method
Prediction by the multilayer perceptron neural network using the Levenberg-Marquardt algorithm
Stage3. Data post-processing and methods evaluation
Data post-processing using the maximum and minimum values
Methods evaluation using mean absolute error and mean squared error

Figure 1. Conceptual model of the research

The first stage in the conceptual model:

Data preparation

At this stage, which is the most elementary stage in the methodology, data in the fuel database are prepared for data mining and analysis. This section is divided into the following sub-sections:

- Data collection and aggregation
- Data normalization
- Missing data

Standardization of data

In this step, the normalization process is done on the data. Due to the difference in the unit of each of the indices, it is necessary to standardize the values of these indices based on a similar unit. Normalization has been used to normalize the data in this research. This method applies a linear transformation to the original data as follows:

The missing data

In this research, the identification technique was used to correct incomplete or missing data.

- Identification

Another way to deal with incomplete data is to use a common value to determine an incomplete attribute. In this way, it is not required to delete the entire record just due to the uncertainty of a particular value.

III. The Second Stage of Gasoline Consumption Forecast

At this stage, after pre-processing the data and preparing them for modeling, the training data are separated and the necessary tests are performed for using these data in the ANN. In this research, an ANN approach combined with the Levenberg-Marquardt (LM) algorithm was used in order to predict gasoline consumption.

Neural networks are based on simulating the behavior of biological nervous systems. Since the 1950s, ANNs have been used persistently to predict in the target attribute regression and classification fields. The neural network is a directional graph composed of a series of nodes called neurons. The nodes are connected with arches, called dendritic or synapses. Each arch has a specific weight, and there is an activation function on each node that is applied to the input values to the node. The inputs are imported to the node by arches, and each arch has its own weight. The learning phase is done by introducing the observations contained in the training data set and adjusting the weight of the arcs.

The reasons for using ANNs to predict fuel consumption are their high efficiency and no need for initial assumptions. In addition, this method is also used extensively in the literature.

The multi-layer perceptron (MLP) feedforward network is one of the most important ANN architectures. Typically, these layers consist of a set of sensory units (basic neurons) as an input layer, at least one hidden layer, and an output layer. The input signal is transmitted through the network and in the forward direction in a layer-to-layer manner. This type of network is commonly referred to as an MLP

Levenberg-Marquardt (LM) algorithm: The LM algorithm was selected among several training methods of error back-propagation because of the faster convergence in the training of medium-sized networks. The error back-propagation algorithm changes network weights and bias values in a given direction such that to reduce the performance function at a faster rate.

The ANN proposed in this study is comprised of an input layer (4 neurons), a hidden layer (3 neurons), and an output layer (1 neuron). These neurons are connected through their special weights, which are calculated after training the network.

Post-processing (estimating prediction model)

After estimating each prediction model, it is necessary to compare the ability and predictive power of the different models in order to determine the best prediction method. There are several criteria for assessing the performance of various predictive methods. Among these criteria, mean absolute error (MAE) and mean squared error (MSE) used to compare the predictive power and select the best method for prediction. These criteria can be expressed as relations (1) and (2).

IV. The analysis of the proposed model

In this section, gasoline consumption prediction is analyzed for different time intervals. For this purpose, the data are explained, the proposed approach is performed, and the results are expressed and analyzed.

Data collection

The data used in this study were collected from the Statistical Center of Iran. The data available in the database of this organization during the period of 2008-2013 (6 years) were used in order to forecast gasoline consumption. These data are time series and include the amount of gasoline sold at the gasoline stations. These values are also continuous.

Missing data

According to the existing database, we realized that on some days, some gasoline stations did not sell any gasoline. With further investigation, we found that these values were zero due to repairs at the gasoline stations. Therefore, assuming that the required quantities were supplied by other gasoline stations, the applied data of all gasoline stations in all regions were considered as the gasoline consumed on that specific day. The following tables summarize the data used for the various time intervals (daily, weekly, monthly,) for the five study areas.

Table 1: Basic information on raw data for each region on a daily basis

Regions	Region 2	Region 2	Region 3	Region 4	Region5
Average	1318709	1510390.83	1057699.19	2069779.57	1430992.71
Variance	44199758195	44204083470	26622826941	106309443184.49	20384641647
Standard Deviation	210237.38	210247.7	163165	326051.3	142774.8
maximum	2149699.37	2305497	1578662	3131645	1845252
minimum	655457.27	860698.6	494831.9	998313.2	791031.8

Table 2: Basic information on raw data for each region on a weekly basis

Regions	Region 2	Region 2	Region 3	Region 4	Region 5
Average	9226060.99	10566712.18	7400756.40	14481993	10012192
Variance	1.78819E+12	1.72379E+12	9.98363E+11	4.28E+12	6.86E+11
Standard Deviation	1337230.71	1312933.15	999181.24	2067875	828508.2
maximum	12932272.52	13707967.76	10312901.33	18652810	11575661
minimum	5947262.43	7542595.66	4313856.38	8770630	5998461

Table 3. Basic information on raw data for each region on a monthly basis

Regions	Region 2	Region 2	Region 3	Region 4	Region 5
Average	40147372.57	45983010	32201064	63013289	43565778
Variance	30776953878489.2	29860387571236.2	16170241796330.1	73037133530927.3	8390044896294.3
Standard Deviation	5547698.07	5464465.9	4021223.92	8546176.55	2896557
maximum	52086907.18	58955664.9	43551532.36	79486077.15	48825220
minimum	30912432.56	37172386.58	25258720.02	46174365.41	35468564

Aggregation of data

For comprehensive examination and accurate prediction of the amount of gasoline consumption, all the data of the regions in different periods of time were aggregated.

Normalization

According to the previous literature in the field of prediction, to increase the accuracy of prediction methods, an MLP with an LM algorithm and an MLP was designed and normalized based on the time interval. For this purpose, the maximum and minimum values are determined at each time interval, and then all input data are transmitted to an interval between 0 and 1.

V. Daily gasoline consumption forecast based on the proposed model

Regarding normalized daily data and considering 5 different delay modes, predictions were made and the MSE and MAE criteria were measured. These results are presented in Table 4.

Table 4. Prediction results of the designed networks based on the MSE and MAE MSE and MAE for the daily time interval

Time delay in days	**2**		**3**		**4**		**5**		**6**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MLP+LM	0.0052	0.0512	0.0050	0.0507	0.00502	0.05070	0.0050	0.04980	0.0050	0.490

As shown in Table 4, in the MLP-LM method, the error rate is almost equal for all time periods. Owing to the less complex computations, for a 2-day mode, this interval is recommended. In Fig. 2, the correlation coefficient (R) is presented based on model outputs and real values. As can be seen, the 3-day delay mode has obtained the best fit for this criterion (R = 0.89492).

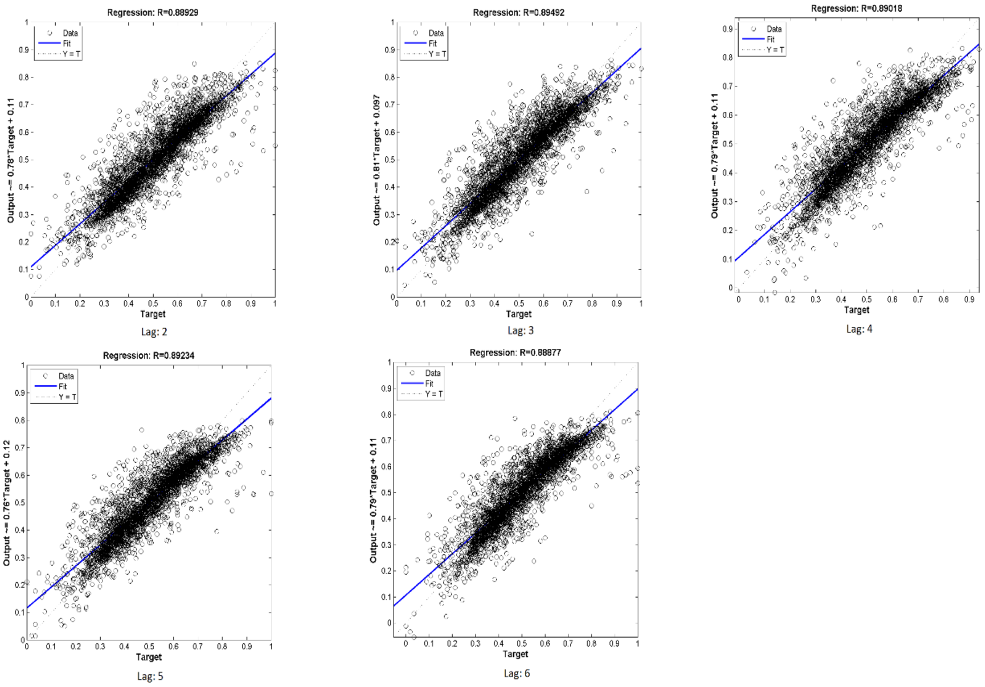

Fig. 2. The R-Chart using the proposed method for the daily time interval

The weekly gasoline consumption forecast based on the proposed model

Regarding normalized weekly data and taking into account 5 different delay modes, predictions were made, followed by measuring the MSE and MAE. These results are presented in Table 5.

Table 5. Prediction results of the designed networks based on the MSE and MAE for the weekly time interval

Time delay in days	**2**		**3**		**4**		**5**		**6**	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MLP+LM	0.477426	0.130714	0.011159	0.071915	0.01242	0.074824	0.011005	0.071731	0.01181	0.076387

In this method, the error rate is approximately equal for 3-week and 5-week intervals. Because of the less complex computations in the 3-week mode, this interval is recommended for this purpose.

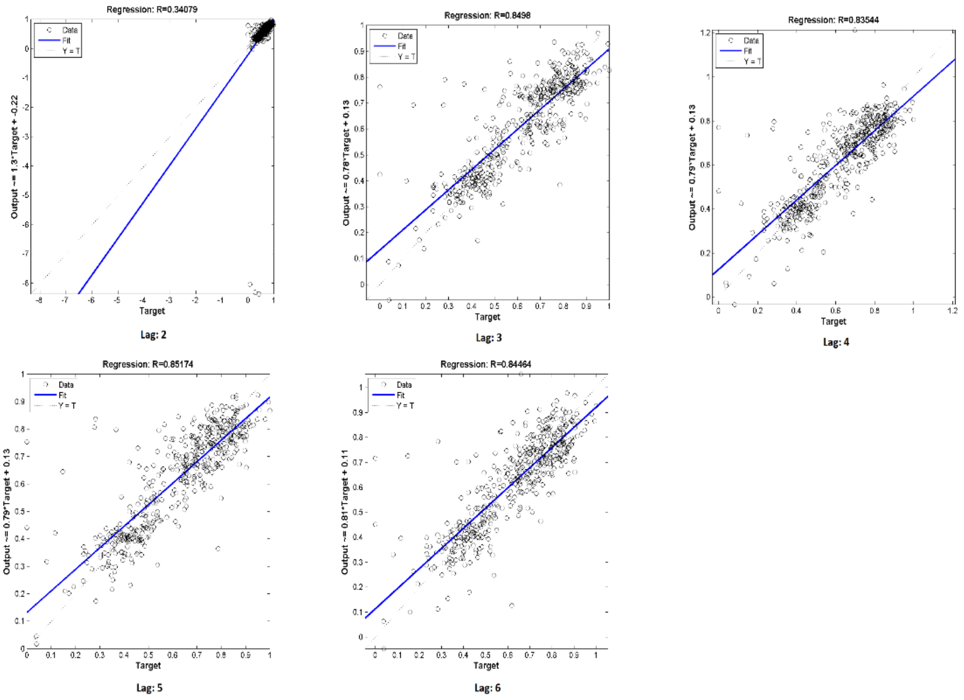

Fig. 3. The R-chart using the proposed method for the weekly time interval

The monthly gasoline consumption forecast based on the proposed model

Regarding normalized monthly data and taking into account 5 different delay modes, predictions were made and the MSE and MAE criteria were measured. These results are presented in Table 6.

Table 6: Prediction results of the designed networks based on the MSE and MAE for the monthly time interval

Time delay in days	2		3		4		5		6	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
MLP+ LM	0.750 608	0.245 696	0.025 512	0.118 071	5.237 976	0.484 325	2.433 155	0.286 458	0.108 577	0.198 854

As shown in Table 6, in this method, the error rate for the 4-month and 5-month time periods is very high and thus these intervals are not recommended for prediction.
In the following figures, the correlation coefficient is shown based on model outputs and actual values. The 3-month and 5-month delay modes have had the best and worst performance in this method, respectively.

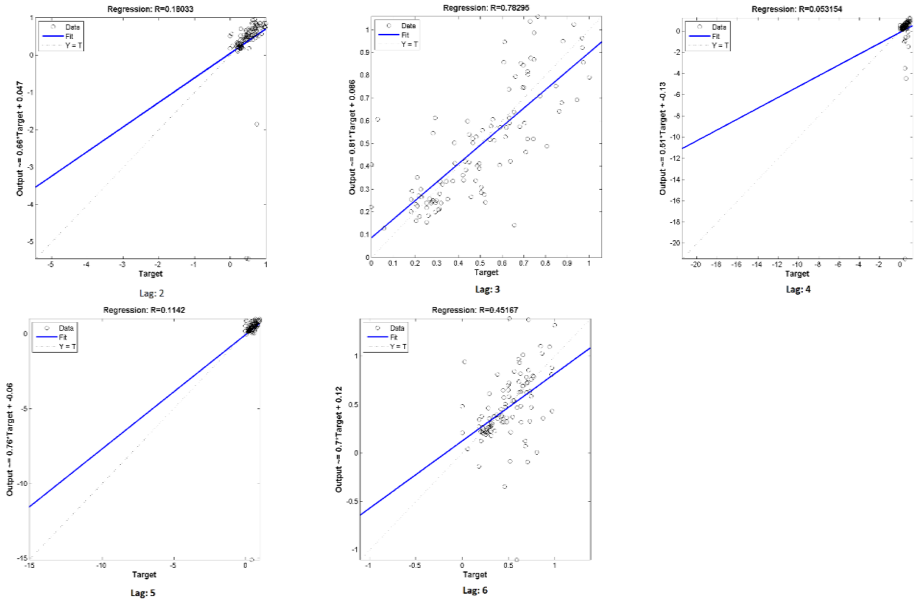

Fig.4. The R-chart using the proposed method for the monthly time interval

VI. Conclusion

The present study was conducted to propose appropriate methods and approaches for forecasting gasoline consumption in Tehran using data mining methods. The results demonstrate the high accuracy of the MLP ANN trained by the LM algorithm.

In addition, it was observed that when the time intervals become larger, the prediction power of the method is reduced and more errors are observed. According to the obtained results, it can be stated that at different time intervals, the appropriate period for prediction should be detected and then future forecasts should be made. The research could be updated for future gasoline consumption forecasts by adding newer data to the model. Also, the results obtained from this study can be used to plan and manage the production, distribution, and storage of gasoline in Iran.

In this study, a new approach was proposed to forecast gasoline consumption in Tehran. The following suggestion may be used to improve this approach.

- Using new and affective variables on gasoline consumption
- Using extensive data to increase the generalizability of the data and the accuracy of the models
- Using combined statistical and artificial intelligence techniques for more accurate diagnosis
- Considering gasoline consumption forecast on a seasonal basis

Conflict of Interest:

There is no relevant conflict of interest regarding this paper.

References

I. Elnaz Siami-Irdemoosa[a] Saeid R.Dindarloo, 2015 "Prediction of fuel consumption of mining dump trucks: A neural networks approach" Applied Energy.Volume 151, 1 August 2015, Pages 77-84.

II. Fatemeh Rahimi-Ajdadi Yousef Abbaspour-Gilandeh, 2011. Artificial Neural Network and stepwise multiple range regression methods for prediction of tractor fuel consumption, Measurement,Volume 44, Issue 10, December 2011, Pages 2104-2111.

III. G. E. Nasr E.A. Badr C.Joun, Backpropagation neural networks for modeling gasoline consumption, Energy Conversion and Management.Volume 44, Issue 6, April 2003, Pages 893-905.

IV. Karisa M. Pierce Janiece L. Hope Kevin J. Johnson Bob W.Wright Robert E.Synovec 2005" Classification of gasoline data obtained by gas chromatography using a piecewise alignment algorithm combined with feature selection and principal component analysis", Journal of Chromatography A Volume 1096, Issues 1–2, 25 November Pages 101-110.

V. Mohanad Aldhaidhawi, Muneer Naji, Abdel Nasser Ahmed. : 'EFFECT OF IGNITION TIMINGS ON THE SI ENGINE PERFORMANCE AND EMISSIONS FUELED WITH GASOLINE, ETHANOL AND LPG'. *J. Mech. Cont.& Math. Sci., Vol.-15, No.-6, June (2020) pp 390-401.* DOI : 10.26782/jmcms.2020.06.0003.

VI. Necla Kara .Togun Sedat Baysec, 2010" Prediction of torque and specific fuel consumption of a gasoline engine by using artificial neural networks" Applied Energy,Volume 87, Issue 1, January 2010, Pages 349-355.

VII. Pierhuigi Barbieri (2001) .Robust cluster analysis for detecting physico-chemical typologies of freshwater from wells of the plain of friuli .*Analytica Chimica Acta* ,,pp.161-170.

VIII. Răzvan Andonie. (2010) "Extreme Data Mining: Inference from Small Datasets"; International Journal of Computers Communications & Control, 5: 280-291.

IX. Reza Babazadeh ,2017"A Hybrid ARIMA-ANN approach for optimum estimation and forecasting of gasoline consumption", RAIRO-Oper. Res.Volume 51, Number 3, July-September 2017.