COMPARISON OF SOME STATISTICAL FORECASTING TECHNIQUES WITH GMDH PREDICTOR: A CASE STUDY

Syed Misbah Uddin*, Aminur Rahman, Emtiaz Uddin Ansari
Industrial and Production Engineering Department,
ShahJalal University of Science and Technology, Sylhet, Bangladesh.
**Corresponding e-mail:misbah-ipe@sust.edu.*

***Abstract**: Demand forecasts are extremely important for manufacturing industry and also needed for all type of business and business suppliers for distribution of finish products to the consumer on time. This study is concerned with the determination of accurate models for forecasting cement demand. In this connection this paper presents results obtained by using a self-organizing model and compares them with those obtained by usual statistical techniques. For this purpose, Monthly sales data of a typical cement ranging from January, 2007 to February, 2016 were collected. A nonlinear modelling technique based on Group Method of Data Handling (GMDH) is considered here to derive forecasts. Forecast were also made by using various time series smoothing techniques such as exponential smoothing, double exponential smoothing, moving average, weightage moving average and regression method. The actual data were compared to the forecast generated by the time series model and GMDH model. The mean absolute deviation (MAD, mean absolute percentage error (MAPE) and mean square error (MSE) were also calculated for comparing the forecasting accuracy. The comparison of modelling results shows that the GMDH model perform better than other statistical models based on terms of mean absolute deviation (MAD), mean absolute percentage error (MAPE) and mean square error (MSE).*

Keywords: *Forecast, GMDH algorithm, Time series, MAPE, MSE.*

INTRODUCTION

In the past few decades, the cement industry has emerged as the fastest growing sector in Bangladesh due to massive construction work in private and public sector. Real estate has become an important source of economic activity, employment, tax revenue and income. Therefore, every company needs to understand its market demand to help formulate responsive policies on cement production and marketing properly. More accurately forecasting demand would facilitate for assisting managerial, operational and tactical decision making. Therefore the selection of forecasting model is the important criteria that will influence to the forecasting accuracy[1]. The GMDH algorithm has been successfully used to deal with uncertainty, linear or nonlinearity of systems in a wide range of disciplines such as economy, ecology, medical diagnostics, signal processing, fossil power plant process, electric power industry and control systems[2-6]. The revised GMDH algorithms[7,8] have been introduced to model dynamic systems in flood forecast and petroleum resource prediction with some success.

Group Method of Data Handling (GMDH) algorithm is a multivariate analysis method for modeling and identifying uncertainty on linear or nonlinearity systems. This algorithm was first introduced in 1967 by A.G. Ivakhnenko[9]. This approach from the very beginning was a computer-based method so, a set of computer programs and algorithms were the primary practical results achieved at the base of the new theoretical principles. The method was quickly settled in the large number of scientific laboratories worldwide due to open code sharing. At that time code sharing was quite a physical action since the internet is at least 5 years younger than GMDH. Despite this fact the first investigation of GMDH outside the Soviet Union had been made soon by R. Shankar in 1972. Later on different GMDH variants were published by Japanese and Polish scientists.

The main idea of GMDH is the use of feed-forward networks based on short-term polynomial transfer functions whose coefficients are obtained using regression combined with emulation of the self-organizing activity behind NN structural learning[10]. To improve the performance of the GMDH algorithm, Barron (1988) gave a comprehensive overview of some early developments of network, and introduced the polynomial network training algorithm (PNETTR). Elder (1996) proposed Synthesis of Polynomial Network (ASPN) algorithm to improve the GMDH algorithm.

J.A.Muller and Frank Lemke developed and improved self-organizing data mining algorithms on the basis of the above results in 1990s[11]. Further enhancements of the GMDH algorithm have been realized in the "Knowledge Miner" software. The GMDH algorithm has gradually become an effective tool for modeling, forecasting, and decision support

and pattern recognition of complex systems. There are processes for which it is needed to know their future or to analyze inter-relations.

The purpose of the study is to determine accurate model for demand forecasting of cement. For this secondary sales data of cement have been collected and forecasts have been made by applying different time series techniques. GMDH method has also been used to derive the forecast.The mean absolute percentage error (MAPE) and mean square error (MSE) have been calculated for comparing the forecasting accuracy among different techniques.

METHODOLOGY

This is a case study research based on time series data of cement industry. The data used in this case study are monthly sales data of cement. The data span the period from January 2007 to February 2016. The dataset consists of 110 months' time series data. Data were analyzed by using various time series model such as moving average, weighted moving average, single exponential smoothing, double exponential smoothing and least square method of simple linear regression.

In this study, we use the value of α 0.3 and 0.5 for single exponential smoothing method. Simple exponential smoothing does not do well when there is a trend in the data. In such situations, several methods were devised under the name "double exponential smoothing" or "second-order exponential smoothing. The basic idea behind double exponential smoothing is to introduce a term to take into account the possibility of a series exhibiting some form of trend. This slope component is itself updated via exponential smoothing. One method sometimes referred to as "Holt-Winters double exponential smoothing are followed here. One of two *smoothing factor is* α which is called data smoothing factor and it's value, $0 < \alpha < 1$, and the other one β is the *trend smoothing factor*, $0 < \beta < 1$. We also used the GMDH predictor version GMDH Data Science 3. 5. 9 to derive the forecast. Out of 110 data 58 months data are used for the training set and rest of the data are used for evaluation in checking set.

The GMDH method was originally formulated to solve for higher order regression polynomials especially for solving modelling and classification problem. General connection between inputs and output variables can be expressed by a complicated polynomial series in the form of the Volterra series, known as the Kolmogorov-Gabor polynomial:

$$Y(x_1, \dots\dots, x_n) = a_o + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} X_i X_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_{ij\,k} X_j X_j X_k + \dots\dots\dots$$

In this case, x represents the input to the system, *n* is the number of inputs and *a* are coefficients or weights. However, for most application the quadratic form are called as partial descriptions (PD) for only two variables is used in the form to predict the output.

$$y^{GMDH} = a_o + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i x_i + \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \quad (1)$$

To obtain the value of the coefficients *a* for each *m* models, a system of Gauss normal equations is solved. The coefficient of nodes in each layer are expressed in the form

$$A = (X^T X)^{-1} X^T Y$$

Where

$$Y = (Y_1\ Y_2 \dots\dots\dots\dots\dots\dots\dots\dots Y_M)^T$$

$$A = [a_o, a_2, a_3, a_4, a_5]$$

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1p} & x_{1q} & x_{1p}\,x_{1p} & x_{1p}^2 & x_{1q}^2 \\ 1 & x_{2p} & x_{2q} & x_{2p}\,x_{2p} & x_{2p}^2 & x_{2q}^2 \\ & & & \vdots & & \\ 1 & x_{Mp} & x_{Mq} & x_{Mp}\,x_{Mp} & x_{Mp}^2 & x_{Mq}^2 \end{bmatrix}$$

M is the number of observations in the training set.

2.1 Measurement of Forecasting Error

In order to evaluate the forecasting accuracy of different techniques various central tendency measures as the loss function were also calculated.

Mean Absolute Deviation

A common method for measuring overall forecast error is the mean absolute deviation. Heifer and Render (2001) noted that this value is computed by dividing the sum of the absolute values of the individual forecast error by the sample size (the number of forecast periods). The equation is:

$$\text{MAD} = \frac{1}{n} \sum_{n=1}^{n} |(\text{Actual} - \text{Forecast})|$$

n = the number of periods[12].

Mean Squared Error (MSE)

In statistics, the mean squared error (MSE) of an estimator measures the average of the squares of the "errors", that is, the difference between the estimator and what is estimated. MSE is a risk function, corresponding to the expected value of the squared error loss or quadratic loss. The difference occurs because of randomness or because the estimator doesn't account for information that could produce a more accurate estimate.

The MSE is the second moment (about the origin) of the error, and thus incorporates both the variance of the estimator and its bias. For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated. Jarrett (1991) stated that the mean square error (MSE) is a generally accepted technique for evaluating exponential smoothing and other methods. The equation is:

$$\text{MSE} = \frac{\sum_{k=0}^{n}\{\text{Actual}-\text{Forecast}\}^2}{\text{n}}$$

Where:

n = the number of periods[11].

Mean Absolute Percentage Error (MAPE)

Mean Absolute Percent Error (MAPE) is the most common measure of forecast error. MAPE functions best when there are no extremes to the data (including zero). With zero or near-zero, MAPE can give a distorted picture of error. The error on a near-zero item can be infinitely high, causing a distortion to the overall error rate when it is averaged in. For forecasts of items that are near or at zero volume, Symmetric Mean Absolute Percent Error (SMAPE) is a better measure[9].

MAPE is the average absolute percent error for each time period or forecast minus actuals divided by actual:

$$\text{MAPE} = \frac{1}{\text{n}}\sum_{n=1}^{n}\frac{|\text{Actual}-\text{Forecast}|}{\text{Actual}} * 100\%$$

DATA COLLECTION AND ANALYSIS

Data Collection is a significant aspect of any type of research study. The data used in this case study are monthly sales data of cement. The data span the period from January 2007 to February 2016. The time series plot is given Fig. 1.

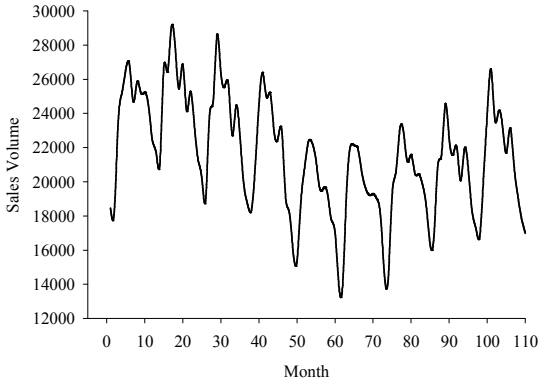

Figure 1: Monthly sales data (Jan 2007 to Feb 2016)

After collecting sales data GMDH algorithm and various statistical forecasting techniques were used to forecast. The mean absolute deviation (MAD), mean absolute percentage error (MAPE) and mean square error (MSE) were also calculated to assess forecasting performance of different models.

3.1 Analysis by GMDH algorithm

GMDH algorithm consists of set of steps that are described below:

Step 1: First *N* observations of regression-type data are taken. The collected load data are first normalized with respect to their individual base value in order to restrict the variation of data within the same level. Those normalized data are denoted by $(x_1, x_2, x_3, x_4, \ldots\ldots\ldots\ldots x_M)$ where M is the total number of input. The original data is separated into the training and test sets[14]. In this study total 110 data were separated into training (58) and test (52) sets. The 58 data is used for the estimation of the partial descriptions which describe the partial characteristics of the nonlinear system. The 52 data is used for organizing the complete description which describes the complete characteristic of the nonlinear system.

Step 2: Select $\binom{m}{2} = \text{m}(\text{m}-1)/2$ new input variables according to all possibilities of connection by each pair of inputs in the layer. Construct the regression polynomial for this layer by forming the quadratic expression which approximates the output *y* in equation (1).

Step 3: Identify the single best input variable out of these $\binom{m}{2}$ input variables, according to the value of mean square error (MSE). The input of variables that give the best results in the first layer, are allowed to form second layer candidate model of the equation (1). Set the new input $(x_1\, x_2 x_3\, x_4\, \ldots\ldots\ldots\ldots x_M)$ and $(M = M + 1)$ Models of the second layer are evaluated for compliance by using MSE, and again the input variables that give best results will proceed to form third layer candidate models. This procedure is

carried out as long as the MSE for the test data set decrease compared with the value obtained at the previous one as shown in Fig. 2. After the best models of each layer have been selected, the output model is selected by the MSE. The model with the minimum value of the MSE is selected as the output model[15].

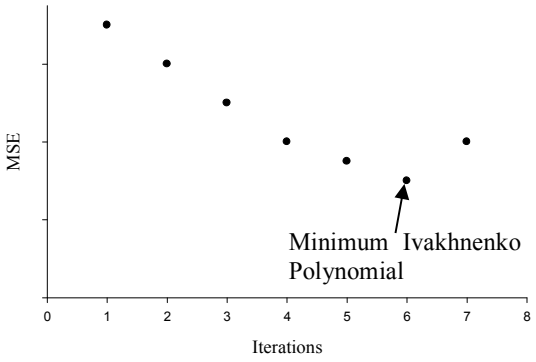

Figure 2. Stopping criteria of GMDH algorithm

3.2 Analysis by Statistical method

Various time series smoothing techniques such as exponential smoothing, double exponential smoothing, moving average and regression method were used for forecasting the load demand. Absolute deviations were also calculated. The mean absolute deviations (MADs) found from these calculations are listed in table 1.

Table 1. MAD of different forecasting methods

Method	MAD
3 month Moving Average	2306
6 month Moving Average	2791
12 month Moving Average	2230
Weightage Moving Average	2056
Regression	2459
GMDH Method	704
Exponential α=0.3	2286
Exponential α= 0.5	2053
Double Exponential α= 0.3, β= 0.5	2861

From Table 1 it is seen that the value of MAD due to forecasting by GMDH algorithm is 704. On the other hand all the statistical method gives four digits MAD. The mean absolute percentage error (MAPE) and mean square error (MSE) were also calculated and reported in Table 2 and Table 3 respectively. It is observed that the GMDH forecast with only 4% MAPE and nearest value is 9% which is done by exponential smoothing technique (α= 0.5). Form Table 3 it is clear that the model with the minimum value of the MSE is the GMDH model.

Table 2. MAPE of different forecasting methods

Method	MAPE
3 month Moving Average	11%
6 month Moving Average	14%
12 month Moving Average	11%
Weightage Moving Average	10%
Regression	12%
GMDH Method	4%
Exponential α=0.3	11%
Exponential α= 0.5	9%
Double Exponential α= 0.3, β= 0.5	13%

Table 3. MSE of different forecasting methods

Method	MSE
3 Month Moving Average	7994519
6 Month Moving Average	10355301
12 Month Moving Average	7710194
Weightage Moving Average	6291543
Regression	9177720
GMDH Method	824882
Exponential α= 0.3	7619269
Exponential α= 0.5	6220179
Double Exponential α= 0.3, β= 0.5	11913465

RESULTS AND DISCUSSION

After completing data analysis we have come out with some informative results. The calculated Mean absolute deviations (MADs) of forecasted data by different forecasting techniques are plotted in Fig. 3. It is seen that GMDH algorithm gives lowest value of MAD which is best suit.

The mean absolute percentage error (MAPE) and mean square error (MSE) are plotted in Fig.4 and Fig.5 respectively. The comparison of modelling results shows that the GMDH model perform better than other models based on terms of mean absolute percentage error (MAPE) and mean square error (MSE).

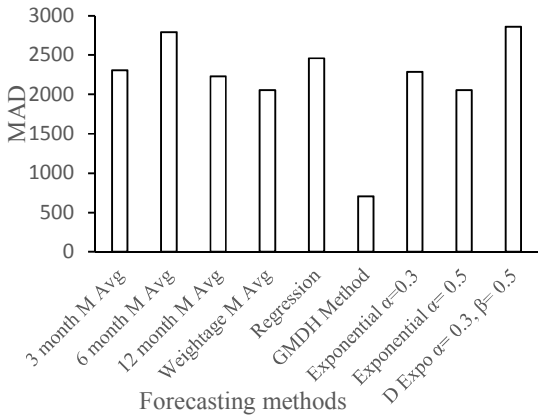

Figure 3.Comparison of MAD of different techniques

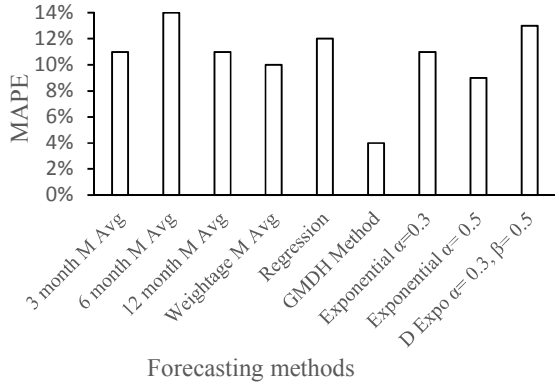

Figure 4.Comparison of MAPE of different techniques

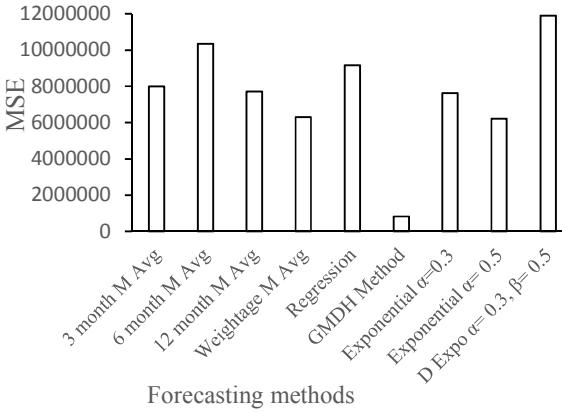

Figure 5.Comparison of MSE of different techniques

To assess the performance of GMDH modelling, last 52 months demand were forecasted and compared with the test set. The results of that model along with forecasting precision are shown in table 4. Normalize mean absolute error is found to be 4.65% whereas normalize RMS is 6%. The fitting accuracy of GMDH model algorithm is also very good as the value of R^2 is 0.90.

Table 4.Summary results of GMDH modelling

Metrics	Output / Value
Post processed result	Model fit
Number of observations	52
Normalize mean absolute error (NMAE)	4.65%
Normalize root mean square error (NRMSE)	6%
Standard deviation of residuals	5.8%
Coefficient of determination (R^2)	0.90
Correlation coefficient	0.95

Our findings have several important implications. Useless input variables are eliminated and useful input variables are selected automatically, the structure parameters and the optimum GMDH architecture can be organized automatically. The case study on the cement time series data testing demonstrated that the GMDH model is robust in the forecasting of nonlinear time series.

CONCLUSION

This paper examined the forecasting accuracy of different statistical techniques as well as GMDH predictor. For that purposes ten years secondary sales data of a cement were collected. There was low seasonal variation in their sales. Demand forecasting was performed using extrapolative time series methods, such as exponential smoothing with level, trend, and seasonal components. Besides that moving average, weighted moving average and regression method were also used for forecasting the demand. A nonlinear self-organizing model based on Group Method of Data Handling (GMDH) was also applied here to derive forecasts.We applied the GMDH predictor version GMDH Data Science 3. 5. 9.

In order to evaluate the accuracy of prediction, various performance measures such as MAD. MAPE and MSE were calculated. It is found that there is no result near to the GMDH predictor. GMDH algorithm forecast with only 0.0367 or 4% error which is substantially more accurate than statistical method.

References

1. Samsudin, R., Saad, P., and Shabri, A., 2010, "Hybridizing GMDH and Least squares SVM support vector machine for forecasting tourism demand,"IJRRAS, Vol. 3(3), pp. 274-279.

2. Ivakheneko, A.G., and Ivakheneko,G. A., 1995, "A Review of Problems Solved by Algorithms of the GMDH,"Pattern Recognition and Image Analysis,Vol. 5(4), pp. 527-535.

3. Onwubolu, G. C., Buryan, P. and Lemke, F, 2008 "Modeling Tool Wear in End-Miling Using Enhanced GMDH Learning Networks," International Journal of Advance Manufacture Technology, Vol.39(11) pp. 1080–1092.

4. Kondo, T., Pandya, A.S., and Nagashino, A.S., 2007, "GMDH-Type Neural Network Algorithm with a Feedback Loop for Structural Identification of RBF Neural Network," International Journal of Knowledge-Based and Intelligent Engineering Systems, Vol. 11, pp.157-168.

5. Puig, V., Witczak, M., Nejjari, F., Quevedo, J., and Korbicz, J., 2007, "A GMDH Neural Network-Based Approach to Passive Robust Fault Detection Using a Constraint Satisfaction Backward Test,"Engineering Applications of Artificial Intelligence,Vol.20, pp.886-897.

6. Li, F., Upadhyaya B. R., and Coffey, L.A., 2009, "Model-Based Monitoring and Fault Diagnosis of Fossil Power Plant Process Units Using Group Method of Data Handling,"ISA Transactions, Vol. 2,pp. 213-219.

7. Kondo, T., 2007, November. Nonlinear Pattern Identification by Multi-Layered GMDH-Type Neural Network Self-Selecting Optimum Neural Network Architecture. International Conference on Neural Information Processing (pp. 882-891). Springer Berlin Heidelberg.

8. Chang, F. J. and Hwang, Y. Y., 1999. "A Self-Organization Algorithm for Real-Time Flood Forecast," Hydrological processes, Vol. 13(2), pp.123-138.

9. Ivakhnenko A. G., 1970, "Heuristic self-organization on problems of engineering cybernetics", Automatic., Vol. 6(3), pp. 207-219.

10. Farlow, S. J., 1981, "The GMDH Algorithm of Ivakhnenko," The American Statistician, Vol. 35(4), pp. 210-215.

11. Muller, J. A., Lemke, F., 2000, "Self-Organizing Data Mining", Libri Books. Dresden, Berlin. pp. 67-110

12. Ezennaya, O. S., Isaac, O. E., Okolie U. O., and Ezeanyim O. I. C, 2014, "Analysis of Nigeria's National Electricity Demand Forecast (2013-2030)", International Journal Of Scientific & Technology Research, Vol.3, pp. 333-340.

13. Gooijer, J. G. D., and Hyndman,R. J., 2006, "25 Years of Time Series Forecasting, International Journal of Forecasting, Vol. 22, pp. 443– 473

14. Samsudin, R., Saad, P. and Skudai, 2009, "Combination of Forecasting Using Modified GMDH and Genetic Algorithm" International Journal of Computer Information Systems and Industrial Management Applications, Vol.1, pp.170-176.

15. Kondo, T. and Ueno, J., 2006, "Revised GMDH-type Neural Network Algorithm with a Feedback Loop Identifying Sigmoid Function Neural Network," International Journal of Innovative Computing, Information and Control, Vol. *2*(5), pp.985-996.

16. Sahu, P. K., Kumar,R.,2013 "Demand Forecasting for Sales of Milk Product (Paneer) in Chhattisgarh", International Journal of Inventive Engineering and Sciences, Vol. 1(9), pp.10-13.

17. Allen, P. G., 1994, "Economic Forecasting in Agriculture" International Journal of Forecasting, Vol. 10, pp. 81-135.