Global Economy and Finance Journal
Volume 3. Number 2. September 2010. Pp. 13 - 26

Forecasting Financial Stocks using Data Mining

Luna C. Tjung, Ojoung Kwon, K. C. Tseng and Jill Bradley-Geist

This study presents a Business Intelligence (BI) approach to forecast daily changes in seven financial stocks' prices. The purpose of our paper is to compare the performance of ordinary least squares model and neural network model to see which model does a better job to predict the changes in the stock prices and identify critical predictors to forecast stock prices to increase forecasting accuracy. The BI approach uses a financial data mining technique to assess the feasibility of financial forecasting compared to a regression model using an ordinary least squares estimation method. We used eight indicators such as macroeconomic indicators, microeconomic indicators, market indicators, market sentiment, institutional investor, politics indicators, business cycles, and calendar anomalies to predict changes in financial stock prices. We found that NN provided superior performance with up to 96% forecasting accuracy compared with OLS model with only 68%.

Keywords: Financial modeling, market efficiency, investment strategies, data mining, forecasting techniques, and neural networks

1. Introduction

Burstein and Holsapple (2008, p. 175) stated that "business intelligence (BI) is a data-driven decision support system that combines data gathering, data storage, and knowledge management with analysis to provide input to the business decision process." According to Han and Kamber (2006, p. 5), "data mining is extracting knowledge from large amounts of data". Fama (1970, p. 383) defined efficient market hypothesis (EMH) where the idea is "a market in which prices provide accurate signals for resource allocation, that is, a market in which firms can make production-investment decisions, and investors can choose among the securities that represent ownership of firms' activities under the assumption that security prices at any time "fully reflect" all available information." Accordingly, it would be difficult if not impossible to consistently predict and outperform the market because the information that one would use to make such predictions is already reflected in the prices.

Although EMH has received some empirical support (Hess & Frost, 1982), it is not without flaws. The hypothesis does not, for example, take into account human cognitive biases and errors that can lead to imperfections in financial prices. In the current paper, we assume that predicting stock prices is difficult but doable to the extent that we can reduce the forecasting error by selecting some better model. The purpose of our paper is to compare the performance of ordinary least squares model and neural network model to see which model better predicts changes in stock prices.

Craig School of Business, California State University, Fresno, CA 93740-8001

In the remainder of the paper, we introduce business intelligence (BI) technique specifically Neural Network and discuss how NN can be used to predict stock prices. In the next section, we compare and contrast Ordinary Least Squares (OLS) and Neural Network (NN) models with regard to their accuracy and ease of use in stock forecasting. Next, we explain the methodology for testing our hypothesis including details on our predictors and data normalization. Finally, we present the results and discuss the implications of our findings.

2. Tools and Techniques of Stock Forecasting

The Ordinary Least Squares (OLS) model has many advantages. It is easy to use, to validate, and typically generates the best combination of predictors by using stepwise regression. However, OLS is a linear model that has relatively high forecasting error when forecasting a nonlinear environment which is common in the stock markets. Also, The OLS model can only predict one dependent variable at a time. On the other hand, a neural network model has high precision, is capable of prediction in nonlinear settings, and addresses problems with a great deal of complexity. Given these advantages, we expect the neural network will outperform OLS in predicting stock prices.

In addition to determining which factors best predict changes in stock prices, we will also be comparing the two analytical strategies of OLS and artificial neural networks. Hammad, Ali, and Hall (2009) showed that the artificial neural network (ANN) technique provides fast convergence, high precision and strong forecasting ability of real stock prices. Traditional methods for stock price forecasting are based on statistical methods, intuition, or on experts' judgment. Traditional methods' performance depends on the stability of the prices; as more political, economical and psychological impact-factors get into the picture, the problem becomes nonlinear, and traditional methods need a more nonlinear method like ANN, fuzzy logic, or genetic algorithms**.**

Along the same lines as Hammad et al., (2009), West, Brockett, and Golden (1997) concluded that the neural network offers superior predictive capabilities over traditional statistical methods in predicting consumer choice in nonlinear and linear settings. Neural networks can capture nonlinear relationships associated with the use of non-compensatory decision rules. The study revealed that neural networks have great potential for improving model predictions in nonlinear decision contexts without sacrificing performance in linear decision contexts.

However, neural networks are not a panacea. For example, Yoon and Swales (1991) concluded that despite neural network's capability of addressing problems with a great deal of complexity, as the increase in the number of hidden units in Neural Network resulted in higher performance up to a certain point, additional hidden units beyond the point impaired the model's performance.

Prior research and common wisdom have suggested several factors that might be used in OLS or a Neural Network model to predict stock prices. Grudnitski and Osburn (1993)

used general economic conditions and traders' expectations about what will happen in the market for their futures. Kahn (2006) stated that the sentiment indicator is the summation of all market expectation that is driven by volatility index, put/call ratio, short interest, commercial activity, surveys, magazines, emotions, and many more. Tokic (2005) showed that political events like the war on terror, fiscal policy to lower taxes, and monetary policy to lower short-term interest, and the increase in the budget deficit are related to stock prices. Nofsinger and Sias (1999) showed that there is a strong positive relation between annual changes in institutional ownership and returns over the herding interval across different capitalizations.

Moshiri and Cameron (2000) compared the most commonly used type of artificial neural network (the back-propagation networks (BPN) model) with six traditional econometric models (three structural models and three time series models) in forecasting inflation. BPN models are static or feed-forward-only (input vectors are fed through to output vectors, with no feedback to input vectors again); they are hetero-associative (the output vector may contain variables different from the input vector) and their learning is supervised (an input vector and a target output vector both are defined and the networks tend to learn the relationship between them through a specified learning rule). The three structural models include (1) the reduced-form inflation equation that follows from a fairly standard aggregate demand-aggregate supply model with adaptive expectations, (2) the inflation equation from Ray Fair's econometric forecasting model, and (3) a monetary model for forecasting inflation. The three time series models are (1) an ARIMA (Autoregressive integrated moving average) model is the single-variable model derived from Box-Jenkins methodology, (2) a Vector Autoregression or VAR model consider the joint behavior of several variables, and (3) a Bayesian Vector Autoregression or BVAR model is the combination of VAR model with prior information on the coefficients of the model and estimated using a mixed-estimation method. In one-period-ahead dynamic forecasting, the information contained in the econometric models is contained in the BPN, and the BPN contains further information; the BPN models are superior to all four comparisons. Over a three-period forecast horizon the BPN models are superior in two comparisons (VAR and Structural) and inferior in two (ARIMA and BVAR). Over a twelve-period forecast horizon, the BPN models are superior in two comparisons (VAR and Structural) and equally good or bad in two (ARIMA and BVAR). Moshiri (2001) concluded that the BPN model has been able to outperform econometric models over longer forecast horizons.

There are many examples of the successful applications of data mining. DuMouchel (1999) used Bayesian data mining to work with large frequency tables, millions of cells, for FDA Spontaneous application. Giudici (2001) used Bayesian data mining for benchmarking and credit scoring in highly dimensional complex datasets. Jeong, et.al. (2008) integrated data mining to a process design using the robust Bayesian approach.

3. Hypothesis

Because the neural network (NN) model can address problems with a great deal of complexity and improve its prediction in nonlinear settings, we expect that the neural network will outperform OLS in predicting stock prices.

H_1: NN Model better predicts stock prices than OLS Model

4. Data and Methodology

In order to forecast the changes in financial stock prices, we used the daily changes in stock prices of seven financial stocks from September 1, 1998 to April 30, 2008. This 10-year time period was selected because we wanted to include the dot com bubble and the early part of the recent global financial crisis into our forecasting horizon instead of including expansion periods only. We used financial stocks because they are relatively volatile and more sensitive to economic news. We provided seven financial stocks because we want to focus primarily on REIT industry as one of the indicators of the economic conditions.

4.1. Predictors

We used eight indicators such as macroeconomic leading indicators (global market indices), microeconomic indicators (competitors), political indicators (presidential election date and party), market indicators (USA index), institutional investor (BEN), and calendar anomaly as our independent variables to predict changes in daily financial stock prices. The calendar anomalies include daily, weekly, monthly, and pre holiday calendar. The daily calendar includes Monday, Wednesday, Thursday, and Friday. The weekly calendar includes week one, week three to week four. The monthly calendar includes January to August, November, and December. We also took into account business cycle factors such as the "dot-com bubble" in our forecasting horizon with dummy variables. We gathered our data through National Bureau of Economic Research (NBER), Yahoo Finance, Federal Reserve Bank, Market Vane (MV), NYSE, and FXStreet.

The macroeconomic indicators include the 18 major global stock indices. The microeconomic indicators include the competitors and companies from different industries. There are 213 of them. The daily market indicators include changes in price and volume of S&P500, Dow Jones Industrial, Dow Jones Utility, and Dow Jones Transportation. The sentiment indicators include the Volatility Index (VIX) and CBOE OEX Implied Volatility (VXO).

In OLS model, the political indicators include major election date and the political party in control. We used election data and political parties as dummy variables. The business cycle includes the recession from technological crash and current bear market

as dummy variables. On the other hand, we included all qualitative variables in the NN model.

4.2. OLS Model

We used SPSS to perform stepwise regression to create a unique regression model for each company.

4.3. NN Model

We ran the neural network with Alyuda NeuroIntelligence to create a NN model. We did data manipulation by using the changes or the first differences of our independent variables except the dummy variables. Also, we normalized the data because we have both negative and positive numbers.

We followed seven-step neural network design process to build up the network. We used the Alyuda NeuroIntelligence to perform data analysis, data preprocessing, network design, training, testing, and query.

We used hyperbolic tangent method to design the network. We used batch back propagation model with stopping training condition of MSE of 1% to find the best network during the network training. Hyperbolic tangent is a sigmoid curve and is calculated using the following formula: $F(x)= (e^{x}-e^{-x})/(e^{x}+e^{-x})$. "Back propagation algorithm is the most popular algorithm for training of multi-layer perceptrons and is often used by researchers and practitioners. The main drawbacks of back propagation are: slow convergence, need to tune up the learning rate and momentum parameters, and high probability of getting caught in local minima."(Alyuda NeuroIntelligence Manual, 2010)

Also, we used overtraining control such as retain and restore best network and add 10% jitter to inputs, weights randomization method such as Gaussian distribution of network inputs, and retrains network 2 times with the lowest training error to train the network. By retaining and restoring the best network, we can prevent over-training such as memorizing data instead of generalizing and encoding data relationships and thus reduce the network error. As a result, the validation errors rise while training errors may still decline in the training graph. By adding jitter, we not only can prevent over-training but also allow the network to escape local minima during training (a major drawback from the batch-back propagation) by adding 10% random noise to each input variable during training. By randomizing the weights, we avoid sigmoid saturation problems that cause slow training. We used Gaussian distribution because it is characterized by a continuous, symmetrical, bell-shaped curve.

4.4. Data Set for OLS Model

We used 2431 data points to build the OLS model by running stepwise regression. With stepwise regression, we can reduce our independent variables to only the statistically significant variables. By doing that, we reduced our independent variables range to between 31 and 61 variables. After having the OLS model, we tested the model by using randomly selected 152 data points by calculating the predicting error. Finally, we tested the forecasting accuracy of the OLS with NN methods by calculating the mean and standard deviation of the % error, that is error divided by the actual value of the stock price.

4.5. Data Set for NN Model

Unlike the OLS model, the NN model used all independent variables. There are three set of data used in the neural network model such as training set, validation set, and testing set. The training set is used to train the neural network and adjust network weights. The validation set is used to tune network parameters other than weights, to calculate generalization loss and retain the best network. The testing set is used to test how well the neural network performs on new data after the network is trained. We used training and validation data to train the network and come up with a model. Finally, we used testing data to test the forecasting errors between the actual and predicted values. Out of 2431 data, we have 152 testing data. The remaining is equally distributed among the training and validation data.

5. Data Normalization and Analysis

We measured our success by comparing the mean and standard deviation of the % error between NN and OLS model. After analyzing the results, the mean for NN is low (2.47% to 19.68%) but the standard deviation is high (218.73% to 584.26%). Similarly, the OLS model with mean of 7.29% to 167.43% also had a high standard deviation of 160.33% to 962.01%. Then, we realized that our % error has both positive and negative numbers because we are using the first difference for all our variables except dummy variables. So, we took the absolute value of the error percentage of all variables. Even after we took the absolute value of the error percentage, our mean (127% to 206%) and standard deviation (174% to 532.8%) for NN model are still high. For the OLS model, we got mean of 104%-381% and standard deviation of 127% to 849%.

So, we want to normalize the data to create better network training by adding 0.1 to the absolute value of the minimum value in each variable to avoid minus sign from the rounding down. For example, to normalized the data of company A, we added the absolute value of lowest negative numbers of company A, that is, change |-6.7| to 0.1. In this case, we have to add 6.8. Then we add 6.8 to all observations. We used the lowest numbers: 6.8 + (-6.7) = 0.1. The reason we used 0.1 is to avoid rounding error. To sum up, the formula we used to normalize the data is to make the lowest negative number to +0.1 and add that number to all observations in a given data set.

After we normalized the data, we have both lower mean (2.13% to 3.27%) and standard deviation (1.78% to 3.39%) for NN model. We have similar result for the OLS model with the mean ranging from 2.55% to 24.84% and standard deviation ranging from 1.88% to 5.37%. Finally, we conducted a paired-wise t-test to compare the performance between two models by using the % NN error and % OLS error.

6. Results

The Adj R^2 ranges from 0.523 to 0.766 with DW from 2.039 to 2.209.

FINANCIAL INDUSTRY	COMPANIES	OLS ADJ R^2	DW
1 Asset Management	T. ROWE PRICE GROUP INC. [TROW]	0.766	2.085
2 REIT - Diversified	PLUM CREEK TIMBER CO. INC. [PCL]	0.758	2.128
3 REIT - Healthcare Facilities	HCP INC. [HCP]	0.597	2.171
4 REIT - Hotel/Motel	HOST HOTELS & RESORTS INC. [HST]	0.695	2.153
5 REIT - Industrial	PUBLIC STORAGE [PSA]	0.818	2.085
6 REIT - Office	BOSTON PROPERTIES INC. [BXP]	0.811	2.039
7 REIT - Retail	SIMON PROPERTY GROUP INC. [SPG]	0.523	2.209

After normalized the data, we reduced our error significantly to < 3.28% for NN and < 33% for OLS model.

NORMALIZED DATA %ERROR

FINANCIAL INDUSTRY	COMPANIES	NN Mean (Stdev)	OLS |Mean| (Stdev)
Asset Management	T. ROWE PRICE GROUP INC. [TROW]	3.18% (2.94%)	32% (8%)
REIT - Diversified	PLUM CREEK TIMBER CO. INC. [PCL]	2.13% (1.78%)	4% (3%)
REIT - Healthcare Facilities	HCP INC. [HCP]	2.66% (3.25%)	25% (5%)
REIT - Hotel/Motel	HOST HOTELS & RESORTS INC. [HST]	3.22% (2.84%)	6% (3%)
REIT - Industrial	PUBLIC STORAGE [PSA]	2.82% (3.39%)	5% (3%)
REIT - Office	BOSTON PROPERTIES INC. [BXP]	2.54% (2.33%)	5.43% (2.46%)
REIT - Retail	SIMON PROPERTY GROUP INC. [SPG]	3.27% (3.3%)	6% (4%)

The Adj R^2 ranges from 0.523 to 0.766 with DW from 2.039 to 2.209.

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	T	df	Sig. 2-tailed
Pair 1	TROWnn – TROWols	-0.2929	0.08307	0.00674	-0.3063	-0.2796	-43.481	15	0
Pair 2	PCLnn - PCLols	-0.0181	0.03225	0.00262	-0.0233	-0.013	-6.933	15	0
Pair 3	HCPnn - HCPols	-0.2236	0.04953	0.00402	-0.2315	-0.2157	-55.661	15	0
Pair 4	HSTnn - HSTols	-0.0295	0.03996	0.00324	-0.0359	-0.0231	-9.111	15	0
Pair 5	PSAnn - PSAols	-0.0173	0.02853	0.00231	-0.0219	-0.0127	-7.47	15	0
Pair 6	BXPnn - BXPols	-0.029	0.034	0.003	-0.034	-0.024	-10.591	15	0
Pair 7	SPGnn - SPGols	-0.0273	0.04385	0.00356	-0.0344	-0.0203	-7.685	15	0

Based on the result from paired t-test, we reject the Ho that OLS better predicts stock prices. The negative sign in the t-statistics shows that NN has lower errors compared to the OLS model.

7. Conclusion and Future Research

The stock market is made of market participants with various risk and return characteristics, different perceptions and expectations about stocks and the economy, and how they interpret and react to the news. Each investor reacts to the market differently at a given point in time, focuses on different pieces of relevant information, and reaches different conclusions. It is unclear how important and how long are the impact of various pieces of information and economic data on the stock prices.

In conclusion, we found that the OLS model is easy to use and validate. It also works fast. However, it is a linear model with a relatively higher error to forecast non-linear environment in the stock market. Also, it only traces one dependent variable at a time.

In contrast, the NN model is complex and requires more efforts to train the network repeatedly to find the best model. Some critical factors may create the best model such as the network architecture (number of layers and neurons) and design (logistic/ hyperbolic tangent/ linear), training algorithms, and stop training conditions (number of iterations). Although we can choose low MSE, this does not guarantee that it is the best model because the network might be over-trained causing memorization rather than

learning. The Alyuda NeuroIntelligence used different sets of data each time we ran the network to avoid the memorization. The software can only reveal what is the best network architecture. Since it is an exhaustive and blind search, we cannot be certain if the model is the best or not when it comes to train the network. With these uncertainties, it is hard to measure the performance of the neural network. It takes more time to train and learn how to use the neural network.

Our results show that NN does a better job than OLS model. Furthermore, our paper shows a significant contribution to the financial forecasting where we can see how one industry affect the others. We also learned that data normalization can make a sizeable difference to the results.

One of our research limitations is that we are only comparing two methods while there are other possible models that may be considered and tested. Future researchers might include more techniques to find the best model for financial forecasting purpose especially for a learning algorithm that can handle market shocks, financial crisis, and business cycles. We provided seven financial stocks because we wanted to focus primarily on REIT industry as one of the indicators of the economic conditions. However, we plan to expand our dataset to see whether we can generalize our findings. Finally, there are many other learning algorithms in the NN to be explored.

References

Alyuda NeuroIntelligence. 2010. Alyuda NeuroIntelligence Manual. (http://www.alyuda.com/neural-networks-software.htm).

Brainmaker, 2010. California Scientific. (http://www.calsci.com/BrainMaker.html)

Burstein, F. and Holsapple, C. 2008. *Handbook on Decision Support System 2.* Springer Berlinn Heidelbert, pp. 175-193.

DuMouchel, W. 1999. “Bayesian Data Mining in Large Frequency Tables, With an Application to the FDA Spontaneous.” *American Statistician*. 53, p. 177.

Fama, E. 1970. “Efficient Capital Market: A Review of Theory and Empirical Work.” *Journal of Finance*, 25, pp. 383-417.

Giudici, P. 2001. “Bayesian Data Mining with Application to Benchmarking and Credit Scoring.” *Applied Stochastic Models in Business & Industry.* 17, pp. 69-81.

Grudnitski, G. and Osburn, L. 1993. “Forecasting S&P and Gold Futures Prices: An Application of Neural Network.” *Journal of Futures Markets*, 13, pp. 631-643.

Hammad, A.; Ali, S.; and Hall, E. 2009 “Forecasting the Jordanian Stock Price using Artificial Neural Network.” (http://www.min.uc.edu/robotics/papers/paper2007/Final%20ANNIE%2007%20Souma%20Alhaj%20Ali%206p.pdf)

Han, J. and Kamber, M. 2006. *Data Mining: Concepts and Techniques*. 2nd Edition. Morgan Kaufmann, p. 5.

Jeong, H.; Song, S.; Shin, S.; and Cho, B. 2008. “Integrating Data Mining to a Process Design Using the Robust Bayesian Approach.” *International Journal of Reliability, Quality & Safety Engineering*. 15, pp. 441-464.

Kahn, M. 2006. *Technical Analysis Plain and Simple: Charting the Markets in Your Language*. Financial Times, Prentice Hall Books.

Moshiri, S. and Cameron, N. 2000. "Neural network versus econometric models in forecasting inflation." *Journal of Forecasting*, 19, pp. 201-217.

Nofsinger, J.and Sias, R. 1999. "Herding and Feedback Trading by Institutional and Individual Investors." *Journal of Finance*, 54, pp. 2263-2295.

Tokic, D. 2005. "Explaining US stock market returns from 1980 to 2005." *Journal of Asset Management*, 6, pp. 418-432.

West, P.; Brockett, P.; and Golden, L. 1997. "A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice." *Marketing Science.* 16, pp. 370-391.

Yoon, Y. and Swales, G. 1991 "Predicting Stock Price Performance: A Neural Network Approach." *Proceedings of the IEEE 24th Annual International Conference of Systems Sciences*, pp.156-162.

Figures

FIGURE 1. TROW- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 751

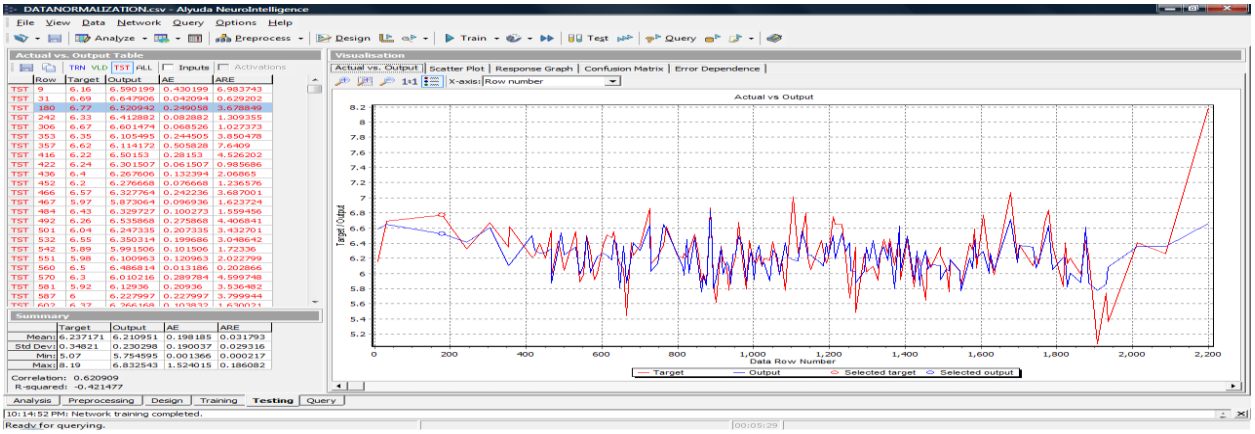

FIGURE 2. PCL- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001

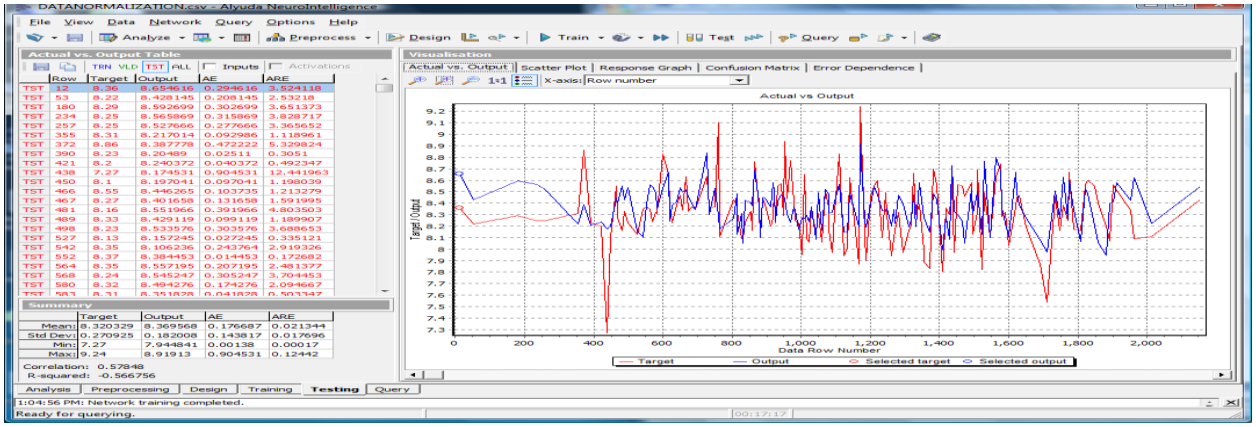

FIGURE 3. HCP- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001

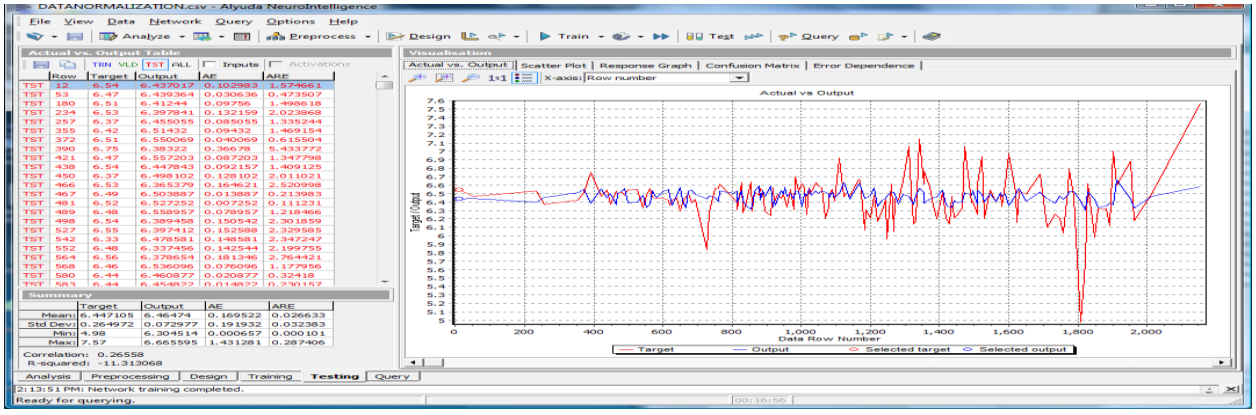

FIGURE 4. HST- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001

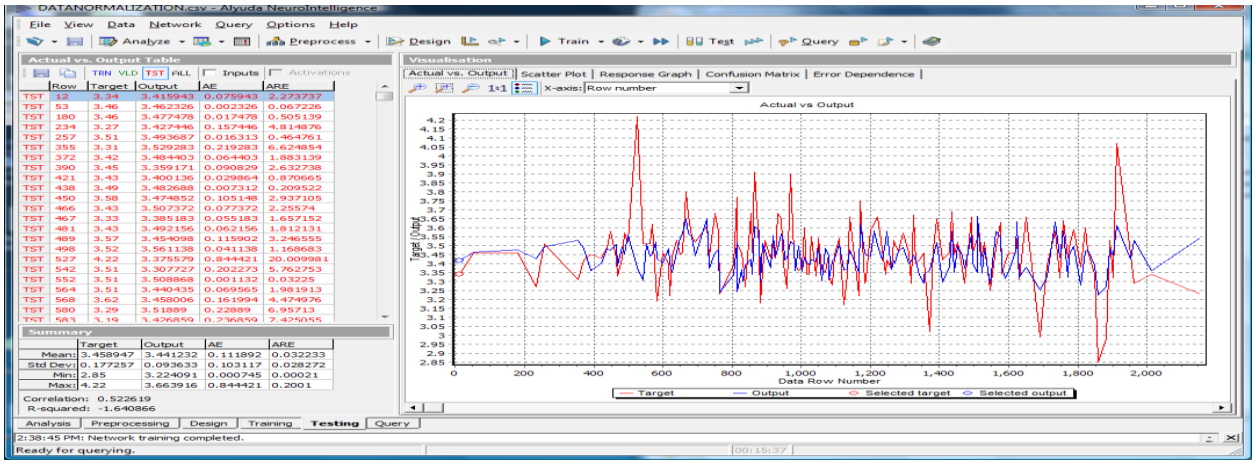

FIGURE 5. PSA- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001

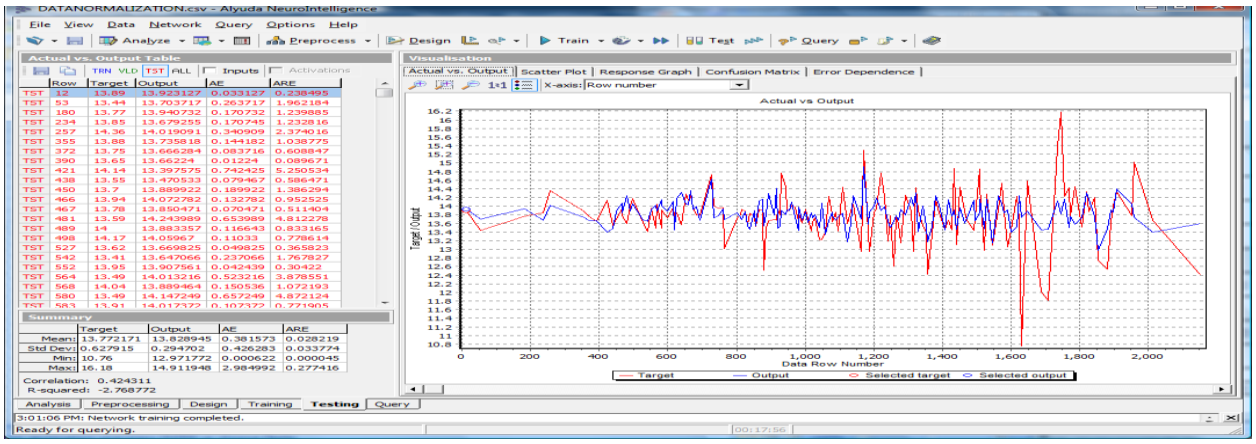

FIGURE 6. BXP- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001

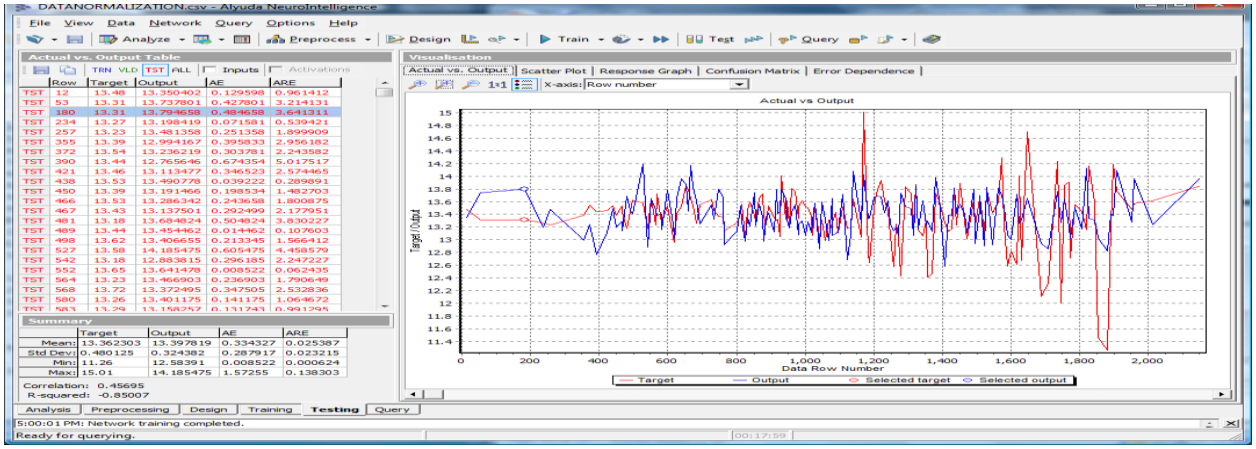

FIGURE 7. SPG- Batch Back Propagation: TESTING
Model Architecture: 267-40-1 No. of Iterations: 1001