DEVELOPMENT OF DATA MODIFICATION METHOD FOR OPTIMIZATION OF FORECASTING PERFORMANCE

SEYEDNAVID SEYEDI

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Engineering (Industrial Engineering)

Faculty of Mechanical Engineering
Universiti Teknologi Malaysia

JUNE 2013

Thanks God to enable me performing this research, I would like to dedicate this dissertation to my beloved parents and my dear sister and brother for their continuous support and encouragement.

ACKNOWLEDGEMENTS

I praise Allah, continuously, though the praise of the fervent does not do justice to His glory.

I wish to express my sincere appreciation to my supervisor, *Dr. Syed Ahmad Helmi bin Syed Hassan* for his guidance, advices and encouragement. His support was really assisted me until the project have been presented completely.

I also would like to take this opportunity to express my gratitude and appreciation to other people who were involved directly or not for giving their criticism and suggestion.

ABSTRACT

Dynamic nature of influencing parameters on market variations prevents decision makers to have a broad vision about possible future changes as an important factor in an organization survival. A precise forecast of both price and demand is a vital issue to illustrate market changes, and prosperity of plans and investments. The main purpose of this study is to develop a quantitative method, which encompasses human user cognition in order to modify timeseries, before being used as an input for forecast models. Some studies conclude ARIMA-ANN hybrid model as the best forecasting model in comparison with its individual models. However, this claim is rejected in some cases. It is a reason to check the performance of individual models in addition to hybrid model in new cases. Historical data are collected from two case studies in manufacturing and service industries. These data are modified by the developed method. Both original and modified data are implemented as inputs for ARIMA, artificial neural network (ANN), and ARIMA-ANN forecast models. The developed method's performance is checked by comparing forecasted results' mean square errors (MSE) and mean absolute percentage error (MAPE). In both case studies, data modification method improves all forecast models' performance. In addition, there are significant changes among different models' performance. The study also finds that the hybrid model's forecasts are not as accurate as ANN's predictions.

ABSTRAK

Faktor-faktor pasaran yang dinamik mampu menghalang organasi daripada meramal masa hadapan dengan tepat agar ia mampu kekal di dalam pasaran. Ramalan yang tepat untuk harga dan permintaan adalah penting untuk melakar perubahan pasaran dan untuk menghasilkan pelan dan pelaburan organisasi yang berjaya. Tujuan utama kajian ini adalah untuk menghasilkan satu kaedah kuantitatif yang turut mengambil kira kemampuan pemikiran manusaia untuk mengubah suai data masa-bersiri sebelum digunakan sebagai input untuk model ramalan. Sesetengah kajian merumuskan bahawa model gabungan ARIMA dan rangkain neural (ARIMA-ANN) merupakan model yang lebih tepat berbanding model ARIMA dan model rangkaian neural (ANN). Kenyataan ini tidak dipersetujui oleh sesetengah kajian dan ini merupakan salah satu sebab untuk menganalisa kebolehan model gabungan dan model individu dalam situasi baru. Data diambil dari dua situasi iaitu industri pembuatan dan industri perkhidmatan dan data ini diubahsuai menggunakan kaedah yang dikaji. Data asal dan data yang diubahsuai digunakan sebagai input untuk ketiga-tiga model ramalan di atas: ARIMA-ANN, ANN dan ARIMA. Prestasi kaedah yang dikaji ditentukan dengan membandingkan purata ralat persegi (MSE) dan purata ralat peratusan mutlak (MAPE). Berdasarkan kepada kedua-dua ukuran prestasi ini, kajian menunjukkan bahawa data yang telah diubahsuai mampu meningkatkan prestasi ketiga-tiga model ramalan. Di samping itu, terdapat perubahan yay ketara didalam prestasi model-model ini. Kajian ini juga menunjukkan bahawa ramalan yang dihasilkan oleh model ARIMA-ANN tidak setepat model ANN.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AR	-	Autoregressive
MA	-	Moving Average
ARMA	-	Autoregressive Moving Average
ARIMA	-	Autoregressive Integrated Moving Average
ANN	-	Artificial Neural Network
MSE	-	Mean Squared Error
RMSE	-	Root Mean Squared
MAE	-	Mean Absolute Error
MAPE	-	Mean Absolute Percentage Error
AVR	-	Average Related Variance
SEP	-	Standard Error of Prediction
PI	-	Persistence Index
BIC	-	Bayesian Information Criterion
AIC	-	Akaike's Information Criterion
ACF	-	Autocorrelation Function
PACF	-	Partial Autocorrelation Function
DF	-	Dickey-Fuller test
ADF	-	Augmented Dickey-Fuller test
OLS	-	Ordinary Least Squared

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Artificial Neural Network (ANN) and Autoregressive Integrated Moving Average (ARIMA) forecasting models are in the spotlight in these days, because of their accuracy and ease of use. In addition to the large number of studies, more investigations are required in increasing models' accuracy as well as prediction performance. This chapter explains the objectives, scopes, methodology and literature review of this project.

1.2 Background of the Study

In this business dynamic environment and complicated economic situation, the role of an accurate planning to remain in the market and increase market share is critical. A specific planning for a certain period can reduce ambiguity and illustrates the way organization ought to pass to achieve companies' goals. In push processes, managers need to plan for their production, transportation and other planned. In a pull processes, the approximate level of available capacities and inventories are planned by managers.

Forecast utilization is a method to improve planning. A high performance, high accuracy forecast method can leads to providing direction, uncertainty reduction, waste and redundancies minimization, controlling standards, and goals establishment. This kind of forecast will alert managers if there is a change in business environment and there would be more time to evaluate and ameliorate directions and plans.

In today's economic, the growth in number of competitors in global markets forces organizations to improve continuously to fulfill their customers demand. A high accuracy forecast makes the supply chain more efficient and more responsive in serving customer needs. However, many factors such as political, economic and environmental in addition to competitors' plans and technology changes leads to instability in market demand, so high level of flexibility are necessary for adapting organizations plans and strategies to new situations.

Unfortunately, significant parameters, which influence the demand, vary from one circumstance to another, as a result there is no high accurate general forecasting method for all situations. Therefore, in spite of numerous research in this field, some gaps are not covered.

In this situation, forecasting the demand as an industrial engineering tool can play a critical role in increasing organizations flexibility, productivity and performance when, it provides managers with a high accurate future prediction and illuminates the way for making decisions.

Financial time-series prediction is encountered by financial investors and has gained researchers attractions as a major and significant task for financial decision-making, where small error in predicting following market movement may result to huge financial lost. Political and economic conditions, traders' expectations or even some rumors in addition to significant number of technical and fundamental parameters are able to influence stock market. Therefore, there is a need to a capable forecasting model to handle nonlinearities, discontinuities and high frequency of stock price timeseries (Hadavandi, 2010). Traditional forecast models cannot handle these levels of noise and complexity, so some new methods based on artificial intelligence and autoregressive moving average can lead to better results (Hsu, 2009). In this study, a proposed method to improve forecasted results is developed.

1.3 Problem Statement

Because of high concentration given on improving integrated quantitative forecasting methods, less attention is given to the role of visual and verbal data modification and forecasting methods in reducing forecast error. Most of the quantitative methods are too complicated or are developed for special circumstances; therefore, applying methods based on user knowledge and idea can results to easier and more common methods (Hong, 2011; Andrawis, 2011). Different groups of measurable, immeasurable, and unknown parameters influence historical data; so visual and verbal evaluation and modification of historical data before being used in forecasting could have a positive effect on forecasting.

Although some unexpected measured data are known as noise and are removed, there is no specific way to rectify the effects of temporary influencing factors. As a result, improving a method to use human user knowledge and idea in adjusting historical data before implementing in forecasting, could rise up forecasting reliability and performance without increasing model complicacy.

ARIMA, ANN and ARIMA-ANN hybrid model have been studied in different cases, while there are mismatched results based on diversity of influencing factors and studied circumstances. Consequently, the selection among these models could not be completely through literatures and it is needed to test them again for different cases.

1.4 Objectives of the Study

This study is based on the following objectives:

i. To develop a method based on human cognition and quantitative approaches in modifying historical data before being used as an input for forecasting models.

ii. Implement ARIMA, ANN, and the hybrid of ARIMA-ANN on manufacturing and service industries' historical data to check which method will produce the best forecast.

iii. To repeat the objective (ii) with the modified data as in objective (i) and to check the effect of data modification method on forecast improvement.

1.5 Scope of the Study

Three scopes of this study are as follow:

i. BINA Paint Integration and Amazon.com are selected as two case studies, representing manufacturing and service industries.

ii. MINITAB software is implemented for ARIMA model selection and forecasting.

iii. MATLAB software is used for ANN model training and forecasting.

1.6 Summary of Literature

Some factors like missing values and unusual data, in addition to seasonality and trend existence cause variation and instability in a time series. Data pre-processing is necessary to reduce this variation and instability, before using it in forecasting models, which leads to forecasts' results improvement (Zhang and Qi, 2005;Wichard, 2011).

ARIMA approach developed by Box and Jenkins (1970) has attracted a lot of attentions as a linear forecasting model. It contains three main steps, named: (i) checking stationary, (ii) parameter estimation and model identification, and (iii) diagnosis checking. Shi et al. (2012) suggests Akaike's information criterion (AIC) or Bayesian information criterion (BIC) methods instead of autocorrelation function (ACF) and partial autocorrelation function (PACF) in identifying appropriate model.

Multilayer ANN is a nonlinear forecasting model. It contains one input layer, one or more hidden layer, and one output layer. The learning ability is a significant advantage of ANN; weights are changed to make input-output behaviour in line with parameter real changes (Negnevitsky, 2005). ANN outperforms classical statistical methods and box Jenkins approach (Werbos, 1988), even though it is time consuming and it may not reach to global optimum answer (Hong et al., 2011).

Traditional models' limitations encourage researchers and decision makers to combine capable forecasting models (Andrawis et al., 2011). Zhang (2003) introduced a combination of ARIMA and ANN models as a general model for both linear and nonlinear cases. Improving forecast accuracy by applying ARIMA-ANN model is concluded by Gutierrez-Estrada et al. (2007). However, Shi (2012) and Taskaya-temizel (2005) report hybrid model inability in improving the result.

1.7 Expected Outcomes

Based on literatures, it is expected that the hybrid model results to a higher level of accuracy in comparison with individual models. Furthermore, adjusting data before implementing in models will have a positive effect on models' accuracy (Andrawis et al., 2011).

1.8 Conceptual Framework

Figure 1.1 represents the conceptual framework of this study. The steps are shown in input, process, and output levels. This study is based on two case studies from which timeseries are collected. These data will be modified; and both primary and modified data are used as inputs for ARIMA, ANN and hybrid of ARIMA-ANN model. Finally, models' performances are compared through Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) to find answers for the objectives.

1.9 Significance of the Study

Having a broad view about the future changes makes planning much easier and reliable. To reach this level of reliability, forecast method accuracy has an important effect. Different statistical methods have been developed to overcome variation and instability among timeseries which are important factors for prediction exactness reduction. Though, more concentration is needed to test and improve these methods in new circumstances.

The role of pre-processing in improving a model prediction, in addition to models ability to recognize correct trend or seasonality among data have been studied. Though, number of these studies is insufficient and there is a need to add human cognition to quantitative methods. In this study a new method for pre-processing data before being applied in forecast models is introduced. This method is based on shifting out bounded data point into most possible trend intervals which are selected by user.

ARIMA and ANN can be combined as ARIMA-ANN model which is suitable for linear and nonlinear data. The hybrid model fails to outperform individual models in some cases, so its performance needs to be tested in more cases.

1.10 Organization of the Study

The rest of the project report is organized as follow. The literature review about time series, forecast and seasonality, ARIMA, ANN, and ARIMA-ANN hybrid model is covered in Chapter 2; the methodology and necessary processes are represented in Chapter 3. Chapter 4 are about data collection and analysis, and the last chapter (Chapter 5) is about a summary of other chapters, results, further research works and conclusion of the study.

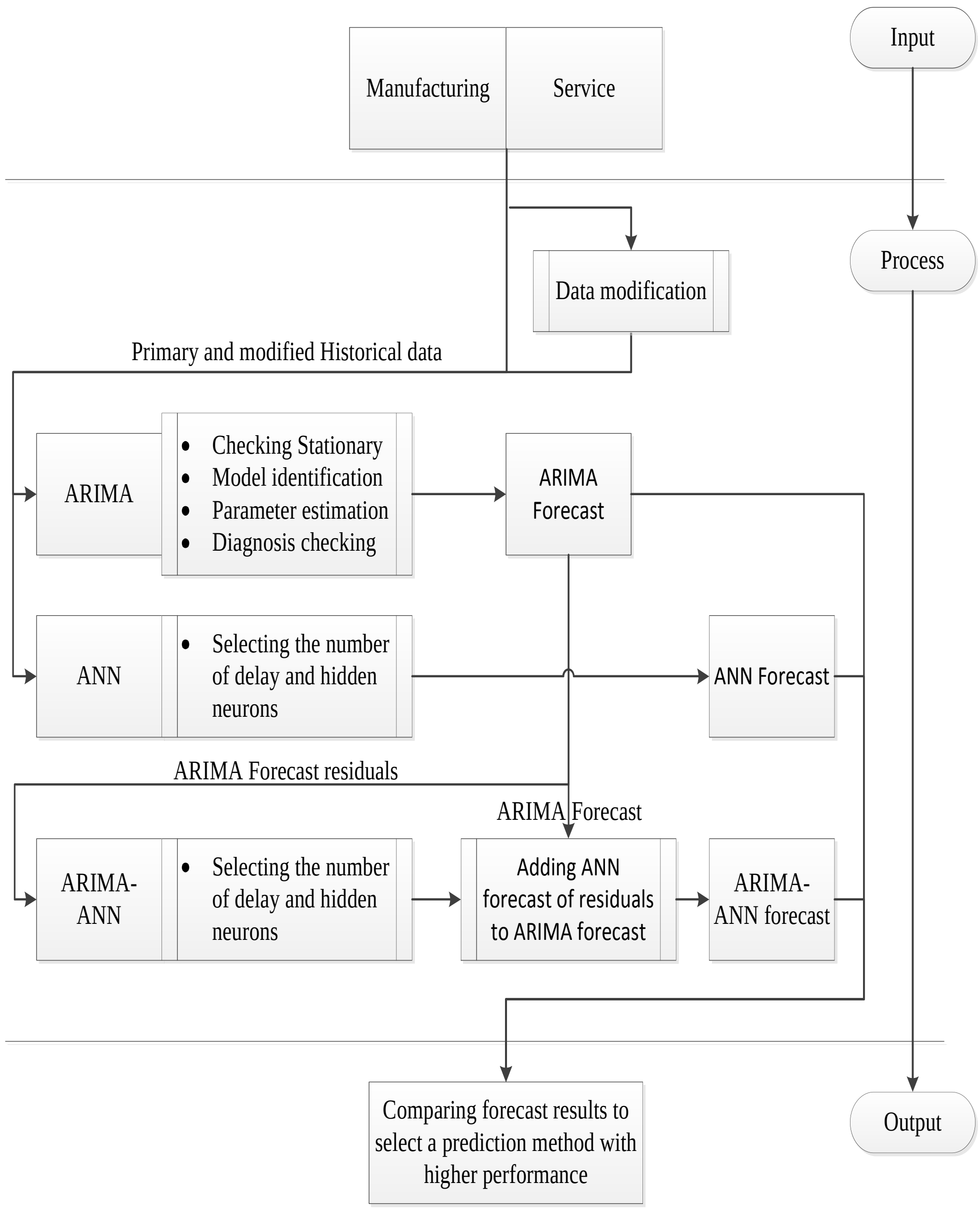

Figure 1.1 Conceptual Frameworks

1.11 Conclusion

Decision makers sometimes do not have the required broad vision regarding the way their market is going on; therefore, it is possible to loss their way to reach customers demand or loss in their financial investigations. By increasing the number of influencing factors, the traditional forecasting methods are not able to present accurate forecasts. In addition, historical forecasting methods present low-performance prediction for timeseries with fluctuated data and nonlinear trends. In these situations organizations have to implement new forecasting methods, which are important tools in manager's hands, to pave their ways to achieve their goals. Selecting the most appropriate model among ANN, ARIMA and ARIMA-ANN hybrid model for manufacturing and financial case studies, is aim of this study.

REFERENCES

Akaike, H. (1974). Markovian representation of stochastic processes and its application to the analysis of autoregressive moving average processes. *Annals of the Institute of Statistical Mathematics* 26(1): 363-387.

Andrawis, R. R., Atiya, A. F., and El-Shishiny, H. (2011). Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. *International Journal of Forecasting*, *27*(3), 672-688.

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*, Springer.

Asyraf, M. A. (2011). THE APPLICATION OF NEURAL NETWORK FOR PREDICTING HULL FORM PROPERTIES. Universiti Technologi Malaysia

Bates, J. M., and Granger, C. W. (1969). The combination of forecasts. *OR*, 451-468.

Ben Taieb, S., Bontempi, G., Atiya, A. F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert systems with applications*, 39(8), 7067-7083.

Bowerman, B. L., and O'Connell, R. T. (1987). *Time series forecasting*. Duxbury, Boston.

Box, G., and Jenkins, G. (1970). *Time series analysis; forecasting and control*. Holden-Day, San Francisco(CA).

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting, and Control.* Holden Day, San Francisco, CA.

Byers, A. (2006). Jeff Bezos: the founder of Amazon.com. *The Rosen Publishing Group*, 46-47

Chopra, S., and Meindl, P. (2010). *demand forecast in a supply chain. SUPPLY CHAIN MANAGEMENT*, PEARSON 198-226.

Crone, S. F., and Kourentzes, N. (2009). Input-variable specification for neural networks-an analysis of forecasting low and high time series frequency. *Neural Networks, 2009. IJCNN 2009. International Joint Conference on, IEEE*.

Degeratu, V., Degeratu, S., & Schiopu, P. (2005, August). Perceptron with one layer based on optical devices. In Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies II (pp. 59720V-59720V). International Society for Optics and Photonics.

Dekker, M., Van Donselaar, K., and Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, 90(2), 151-167.

De Leeuw, S., Van Donselaar, K., & De Kok, T. (1998). Forecasting techniques in logistics. *In Advances in distribution logistics*. Springer Berlin Heidelberg. 481-499.

Diaz-Robles, L. A., Ortega, J. C., Fu, J. S., Reed, G. D., Chow, J. C., Watson, J. G., and Moncada-Herrera, J. A. (2008). A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. *Atmospheric Environment*, 42(35), 8331-8340.

Durbin, J., and Koopman, S. J. (2001). *Time series analysis by state space methods*. Oxford University Press, Oxford.

Faraway, J., and Chatfield, C. (1998). Time series forecasting with neural networks: a comparative study using the air line data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(2): 231-250.

Fildes, R., and Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review/Revue Internationale de Statistique:* 289-308.

Franses, P. H., and Draisma, G. (1997). Recognizing changing seasonal patterns using artificial neural networks. *Journal of Econometrics* 81(1): 273-280.

Griñó, R. (1992). Neural networks for univariate time series forecasting and their application to water demand prediction. *Neural network world* 2(5): 437-450.

Gutiérrez-Estrada, J. C., Silva, C., Yáñez, E., Rodríguez, N., & Pulido-Calvo, I. (2007). Monthly catch forecasting of anchovy< i> Engraulis ringens</i> in the north area of Chile: Non-linear univariate approach. Fisheries Research, 86(2), 188-200.

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8), 800-808.

Harvey, A. C. , and GDA, P., (1979). Maximum likelihood estimation of regression models with autoregressive moving average disturbances. *Biometrika* 66: 49-58.

Hashem, S. (1993). *Optimal linear combinations of neural networks*, Purdue University. Ph.D.

Hashem, S., and Schmeiser, B. (1995). Improving model accuracy using optimal linear combinations of trained neural networks. *Neural Networks, IEEE Transactions on* 6(3): 792-794.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1), 1-12.

Hippert, H. S., Pedreira, C. E., & Souza, R. C. (2001). Neural networks for short-term load forecasting: A review and evaluation. *Power Systems, IEEE Transactions on*, 16(1), 44-55.

Hong, W. C., Dong, Y., Chen, L. Y., & Wei, S. Y. (2011). SVR with hybrid chaotic genetic algorithms for tourism demand forecasting. *Applied Soft Computing*, 11(2), 1881-1890.

Hsu, S. H., Hsieh, J. J., Chih, T. C., & Hsu, K. C. (2009). A two-stage architecture for stock price forecasting by integrating self-organizing map and support vector regression. *Expert Systems with Applications*, 36(4), 7947-7951.

Hylleberg, S. (1992). *Modelling seasonality*, Oxford University Press.

Hylleberg, S. (1994). Modelling seasonal variation. *Nonstationary Time Series Analysis and Cointegration, Hargreaves, CP (ed.)*: 153-178.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural networks* 1(4): 295-307.

Kitanidis, P. K., and Bras, R. L. (1980). Real-time forecasting with a conceptual hydrologic model 2. applications and results. *WATER RESOURCES RES.* 16(6): 1034-1044.

Kolarik, T., and G. Rudorfer (1994). *Time series forecasting using neural networks*. ACM SIGAPL APL Quote Quad, ACM.

Liu, G. Q. (2011). *Comparison of Regression and ARIMA Models with Neural Network Models to Forecast the Daily Streamflow of White Clay Creek*. University of Delaware

Makridakis, S., & Hibon, M. (2000). The M3-Competition: results, conclusions and implications. *International journal of forecasting*, 16(4), 451-476.

MINSKY, M. (1969). PAPER S., *Perceptron*, MIT Press, Cambridge, Mass.

Nash, J. E., and Sutcliffe, J. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology* 10(3): 282-290.

Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Addison-Wesley Longman.

Nelson, C. R., and Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: some evidence and implications. *Journal of monetary economics* 10(2): 139-162.

Nelson, M., Hill, T., Remus, W., and O'Connor, M. (1999). Time series forecasting using neural networks: Should the data be deseasonalized first?. *Journal of forecasting*, 18(5), 359-367.

Pankratz, A. (1983). *Forecasting with Univariate Box–Jenkins Models: Concepts and Cases*. John Wiley & Sons, New York.

Qi, M., and Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European journal of operational research* 132(3): 666-680.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* 65(6): 386.

Rumelhart, D. E., Hintont, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics* 6(2): 461-464.

Sharda, R., and Patil R. B. (1992). Connectionist approach to time series prediction: an empirical test. *Journal of Intelligent Manufacturing* 3(5): 317-323.

Shi, J., Guo, J., and Zheng, S. (2012). Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews*, 16(5), 3471-3480.

Shynk, J. J. (1990). Performance surfaces of a single-layer perceptron. *Neural Networks, IEEE Transactions on* 1(3): 268-274.

Shynk, J. J., and Bershad, N. J. (1992). Stationary points and performance surface surfaces of a perceptron learning algorithm for a nonstationary data model. *joint conference on Neural Networks*, Baltimore, MD.

Slutzky, E. (1937). The summation of random causes as the sources of cyclic processes. *Econometrica* 5: 105-146.

Taskaya-Temizel, T., & Casey, M. C. (2005). A comparative study of autoregressive neural network hybrids. Neural Networks, 18(5), 781-789.

Ventura, S., Silva, M., Perez-Bendito, D., and Hervas, C. (1995). Artificial neural networks for estimation of kinetic analytical parameters. *Analytical Chemistry*, 67(9), 1521-1525.

Wang, Y., Wang, J., Zhao, G., & Dong, Y. (2012). Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: A case study of China. *Energy Policy*.

Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Cambridge, MA., Harvard University. Ph.D.

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks* 1(4): 339-356.

Wichard, J. D. (2011). Forecasting the NN5 time series with hybrid models. *International Journal of Forecasting* 27(3): 700-707.

Yule, G. U. (1971(Published in 1927)). *On a method of investigating periodicities in distributed series with special reference to Wolfer's sunpspot numbers*. (Published in 1927.) Statistical papers of George Udny Yule, A. Stuart and M. Kendall, ed., Hafer Press, New York..

Zhang, G. (1998). *Linear and nonlinear time series forecasting with artificial neural networks*. Kent, OH., Kent State University, . Ph.D.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50: 159-175.

Zhang, G. P., and M. Qi (2005). Neural network forecasting for seasonal and trend time series. *European journal of operational research* 160(2): 501-514.