IMPLEMENTATION OF GENETIC ALGORITHM IN MODEL IDENTIFICATION OF BOX-JENKINS METHODOLOGY

MOHD ZULARIFFIN MD MAAROF

UNIVERSITI TEKNOLOGI MALAYSIA

IMPLEMENTATION OF GENETIC ALGORITHM IN MODEL IDENTIFICATION OF BOX-JENKINS METHODOLOGY

MOHD ZULARIFFIN BIN MD MAAROF

A thesis submitted in fulfilment of the requirements for the award of the degree of Master of Science (Mathematics)

Faculty Of Science
Universiti Teknologi Malaysia

JUN 2013

Dedicated to:

My beloved parents,

Md Maarof Mardi, Ramlah Abdul Latif

My supportive siblings,

Zulfadli, Noradilah, Nor Rislah, Zulkhairi, Zul Amin

My dedicated lecturers,

My endless spirits

and all my friends.

This is for you.

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious and Most Merciful, all praise to Allah SWT, the Almighty, for His love has given me strength, perseverance, diligence and satisfaction in completing this project.

First and foremost, I would like to express my deepest gratitude especially to my supervisor, Prof. Dr. Zuhaimy Hj. Ismail who had taken a lot of effort to meticulously go through my work and come out with helpful suggestions. Not forgotten, million appreciations for my co-supervisor Dr. Norhisham Bakhary, and Dr. Noor Hazarina Hashim for their valuable critics and advices.

I would like to express my appreciation to Ministry of Science Technology and Innovation (MOSTI), MyBrain Programme and Research University Grant (RUG) for supporting the scholarship along my study.

Besides that, I also would like to acknowledge my special thanks to my supportive friends, Ezza Syuhada Sazali, Lili Ayu Wulandari and Siti Khadijah Mariam for their suggestions, comment and moral support. Their efforts are much appreciated. May Allah bless all of you.

Finally, I would like to express my greatest gratitude to my beloved family for their unstinting support and prayer. Without the family members support and prayer, this project would have been difficult at best. Thank You.

ABSTRACT

During the past several decades, a considerable amount of studies have been carried out on time series and in particular the Box-Jenkins (BJ) method. As with all techniques of statistical analysis, the conclusions of time series analysis are critically dependent on the assumptions underlying the analysis and BJ is a commonly used forecasting method that can yield highly accurate forecasts for certain types of data. Genetic Algorithm (GA) is a heuristic method of optimization. This study presents the study on developing an extrapolative BJ model with the use of GA method to produce forecasting models using time series data. BJ method has a cycle of four phases, the data transformation phase for model identification, parameter estimation, model diagnostic checking or validation, and finally producing the forecast. Although many researchers and practitioners have concentrated in the parameter estimation part of BJ model, the most crucial stage in building the model is in the data transformation and model identification where any false identification will lead to assuming a wrong model and will increase in the cost of re-identification. Hence, using GA a subset of artificial intelligence methods was introduced into the process of BJ to solve the problem in the model identification and parameter estimation phase. The data used in this study are the monthly data of international tourists arrival into Malaysia from 1990 to 2011. This is a case study in the implementation of GA-BJ model. The result from this study may be divided into two main parts, namely the result for the in-sample data (fitted model) and out-sample data (forecast model). The analysis shows that the out-sample values using GA-BJ model gives better forecast accuracy than the out-sample values for BJ model. This shows that the combination of BJ and GA methods gives a more accurate model than using a single method for forecasting. This study concludes that GA method can be an alternative way in identifying the right order of component in BJ model.

ABSTRAK

Dalam beberapa dekad yang lalu, sejumlah besar kajian telah dijalankan ke atas siri masa dan khususnya kaedah Box-Jenkins (BJ). Seperti semua teknik analisis statistik, kesimpulan analisis siri masa adalah amat bergantung kepada andaian analisis. BJ adalah satu kaedah yang lazim digunakan yang boleh menghasilkan ramalan yang sangat tepat untuk sesetengah jenis data. Kajian ini membentangkan hasil kajian kaedah ekstrapolatif model Box-Jenkins (BJ) bagi menghasilkan model Univariat dengan menggunakan data siri masa. Kaedah BJ mempunyai empat fasa utama iaitu model identifikasi, model penaksiran, model pengesahan, dan model peramalan. Walaupun banyak penyelidik dan pengamal telah tertumpu di bahagian anggaran parameter model BJ, peringkat yang paling penting dalam membina model adalah dalam transformasi data dan pengenalan model jika apa-apa pengenalan palsu akan membawa kepada andaian model yang salah dan akan meningkatkan kos semula membina model pengenalan. Oleh itu, dalam kajian ini, model algoritma genetik (GA) adalah subset bagi kaedah kepintaran tiruan yang diperkenalkan untuk menyelesaikan masalah yang dihadapi di fasa pertama dan kedua iaitu model identifikasi dan model penaksiran. Data yang digunakan dalam kajian ini adalah data bulanan pelancong antarabangsa melawat Malaysia mulai tahun 1990 sehingga 2011. Ini adalah kajian kes dalam implementasi model GA-BJ. Hasil analisis kajian ini dibahagikan kepada dua bahagian iaitu sampel dalam (model ujian) dan sampel luar (model ramalan). Di akhir kajian ini, model GA-BJ bagi sampel luar lebih tepat dan mempunyai ralat yang lebih kecil berbanding model asas iaitu model BJ bagi sampel luar. Ini menunjukkan bahawa model kombinasi kaedah BJ dan GA menghasilkan model ramalan yang lebih tepat berbanding menggunakan hanya satu model. Kesimpulannya, kajian ini menunjukkan bahawa kaedah GA boleh menjadi kaedah alternatif bagi mengenalpasti komponen model pengenalan BJ yang betul.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

e_t	-	The residual
p	-	The order of autoregressive model
q	-	The order of moving average model
P	-	The order of seasonal autoregressive model
Q	-	The order of seasonal moving average model
t	-	time
f_x	-	Value of fitness function
x	-	Number of chromosome
n	-	Number of observation in the time series
z_t	-	The number of monthly international tourist arrival
$\hat{z}_t$	-	The estimate numbers of international tourist arrival
H_0	-	Hypothesis one
H_1	-	Hypothesis two
S_1^2	-	Larger variance
S_2^2	-	Smaller variance
r	-	Correlation coefficient
$\bar{y}$	-	The mean of the time series
h	-	Maximum number of lag
k	-	the time lag
I	-	The difference of seasonal nor non-seasonal
ϕ_{kk}	-	Partial autocorrelation coefficient
ρ	-	population size
θ	-	Parameter for autoregressive model
ϕ	-	Parameter for moving average model

Φ	-	Parameter for seasonal moving average model
Θ	-	Parameter for seasonal moving average model
B	-	Backward shift operator
$\overline{w_t}$	-	The mean of difference time series data
w_t	-	The difference of time series data
$\frac{\partial y}{\partial \phi}$	-	Partial differential with respect to ϕ
$\frac{\partial y}{\partial \theta}$	-	Partial differential with respect to θ

LIST OF ABBREVIATIONS

ARIMA	-	Autoregressive integrated moving average
SARIMA	-	Seasonal autoregressive integrated moving average
AR	-	Autoregressive model
MA	-	Moving average model
ARMA	-	Autoregressive moving average model
GA	-	Genetic algorithm
BJ	-	Box Jenkins
GA-BJ	-	Genetic algorithm- Box Jenkins model
MSE	-	Mean square error
MAPE	-	Mean absolute percentage error
MAE	-	Mean absolute error
GA-SARIMA	-	Genetic algorithm-seasonal integrated moving average model
GA-ARIMA	-	Genetic algorithm- autoregressive integrated moving average model
SE	-	Standard error
ARFIMA	-	Autoregressive fractionally integrated moving average
ACF	-	Autocorrelation function
PACF	-	Partial autocorrelation function
FPE	-	Final prediction error
MEV	-	Minimum eigenvalue vector
MDL	-	Maximum distributed length
AIC	-	Akaike information criterion
AI	-	Artificial intelligence
ANN	-	Artificial neural network
VAR	-	Vector autoregressive model
ARDL	-	Autoregressive distributed lag

STSM	-	Structural time series model
TVP	-	Time varying parameter
GFS	-	Genetic fuzzy system

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Every definition of forecasting defined it as the process of predicting the future by organizing the past information. It is also described as a method to predict future events. Forecasting can be applied in areas as such, forecasting electrical demand, water demand, sales of product demand, tourisms demand and also for government policy making. For example, hotels management mostly use forecasting to determine operational requirements. Furthermore, it can helps in reducing the risk of decisions and the cost of expenditure for future planning.

The methods of forecasting can be categorized into two major groups which are qualitative methods and quantitative methods. Qualitative method requires no overt manipulation of data, and only the judgments of the forecaster were used. Meanwhile, quantitative method is a technique that can be applied when there is enough historical data. Most researchers and forecasters used quantitative method because this method involves a mathematical analysis of the historical data in developing a model for forecasting. Furthermore, quantitative methods can be categorized into two types namely, time series and causal methods. The common time series methods is Box-Jenkins (BJ) methods because it is one of the most powerful and accurate forecasting techniques for short term forecast-specially for univariate time series.

In this study, an heuristic method, genetic algorithm, (GA) is introduced and applied to BJ method in determining the right order of component BJ model (*p* ,*q*, *P*, *Q*), where *p*, *q* ,*P*, *Q* are the degree of autoregressive model, moving average model, seasonal autoregressive model and seasonal moving average model respectively.

The implementation of GA is to improve the weakness of BJ procedure in identifying the right order of BJ tentative model. Thus, to examine the effectiveness of this combination method, a comparison study was conducted between BJ model and the combination of GA-BJ model. Therefore, the following subsection describes the background of the problem.

1.2 Background of Problem

The stationary of univariate time series data are mostly been analyzed by using the application of Box and Jenkins [3] which introduced the autoregressive integrated moving average (ARIMA) model. However, when the time series data set contains seasonal effect, SARIMA(p,d,q)(P,D,Q) is applied. If there is no seasonal effect, SARIMA(p,d,q)(P,D,Q) will be reduced to pure ARIMA(p,d,q) model, and when the time series data set is stationary, a pure ARIMA(p,d,q) reduces to ARMA(p,q).

There are four stages to construct BJ models which are model identification, model estimation, model checking and forecasting. Although many researchers and practitioners have concentrated in estimation part of BJ model, the most crucial stage in building the model [4] is the identification part as the false identification will contribute to the increment of the cost of re-identification.

During identifying the order of BJ model, the intervention of a human expert is also required in order to identify the best model because it is also not fully automatic [3]. The current approaches are focusing on model estimation. The used of correloggram method is to identify the order of autoregressive model (*p*) or the order of moving average (*q*) of ARIMA or SARIMA model is complicated and not easily conducted. Hence, the identification model and the parameter estimation will be overlapped at diagnostic model checking stage until the best fitted model is found.

1.3 Statement of Problem

As stated in the background problem, the first step in building ARIMA or SARIMA models is by determining the fitting order for the model identification stage.

Since GA could works effectively by finding the approximate optimum solution in complex data set, this study seeks to use paradigms based on artificial intelligence GA for solving the best order for *p*, *q*, *P*, and *Q* of BJ model. Therefore, the correct parameter of BJ model can be properly estimated.

1.4 Research Question

Questions arise when developing GA model for ARIMA and SARIMA model identification procedure. It can be summarized as follow:

i. How to identify the order of ARIMA/SARIMA model using GA?
ii. How to estimate the parameter of ARIMA/SARIMA model using GA?
iii. How to model international tourist arrival time series data using GA-BJ model?

1.5 Objective of the Study

The main objective of this study can be categorized into three parts. The objectives are stated as follow:

i. To design and develop GA method for model identification in the BJ model.
ii. To implement GA in estimating parameters of BJ model.
iii. To develop a monthly international model for tourist arrival to Malaysia and forecast tourist arrival using GA-BJ model.

1.6 Scope of the Study

The scopes of this study are:

i. This study focuses only on Box-Jenkins methodology where the time series data used is based on historical past value.
ii. Forecast accuracy in this study will be defined by measuring the lowest error in term of mean square error (MSE)
iii. In genetic algorithm architecture, chromosomes are randomly generated using genetic operator.
iv. The data used as the secondary data of tourist arrival to Malaysia.
v. Time series data used in GA-BJ model is the process data that has been analysis in Chapter 4.
vi. GA-BJ model is used for modeling one type of case study which is k step ahead only.
vii. Out-sample data analysis is focuses on forecasting 12 months international tourist arrival to Malaysia.
viii. The analysis in Chapter 4 and Chapter 5 were based on off-line data only.

1.7 Research Contribution

The main contribution of this study is to develop an alternative way of using GA in BJ model identification method and estimation phases. The second contribution of this study is to the development of mathematical model combining GA-BJ forecast model and fitted model for monthly international tourist arrival to Malaysia. Last but not least, the third contribution of this study is to the forecast accuracy of this combining GA-BJ model is better than methods proposed by Box and Jenkins.

1.8 Research Data

The data used in this study are the secondary data provided by Malaysian Tourism Promotion Board. The data is an annual time series data that covered the period from 1990 to 2011.

1.9 Thesis Plan

This thesis was divided into six main chapters. Chapter 1 outlined the introduction parts which are including general background of study and some reviews from previous work and describes the problem statements which led to this research, objectives, scope and the contribution of study. Chapter 2 will describe the literature reviews, current knowledge and related theories regarding the identification method in BJ procedure. Next, Chapter 3 will explain in details the BJ procedures and GA-BJ procedures. The case study on forecasting monthly international tourist arrival to Malaysia using BJ model will be described in Chapter 4. In Chapter 5, the construction of forecasting model for monthly international tourists' arrival to Malaysia by using proposed GA-BJ model. Finally, Chapter 6 summarizes all the experimental findings and conclusions and presented some suggestions for future studies related to this research.

References:

[1] R. L. Haupt, S. E. Haupt, and A. J. Wiley, *Algorithms Practical Genetic Algorithms*.

[2] Tourism, *Malaysia tourism key performance indicators*. Malaysia Tourism Board, 2009, p. 43.

[3] G. E. . Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco CA, 1970.

[4] C. Chatfield, *Time Series Forecasting*. Chapman & Hall CRC FLorida, 2001.

[5] S. . Chuah, "Forecasting tourist arrivals to Malaysia," Universiti Putra Malaysia, 2001.

[6] M. Shitan, "Time Series Modelling of Tourist Arrivals to Malaysia," *interstat.statjournals.net*, pp. 1–12, 2008.

[7] H. Song, E. Smeral, G. Li, and J. L. Chen, "Tourism Forecasting :," 2008.

[8] C. Goh and R. Law, "The Methodological Progress of Tourism Demand Forecasting: A Review of Related Literature," *Journal of Travel & Tourism Marketing*, vol. 28, no. 3, pp. 296–317, Apr. 2011.

[9] F.-L. Chu, "Forecasting tourism: a combined approach," *Tourism Management*, vol. 19, no. 6, pp. 515–520, Dec. 1998.

[10] J. H. Kim and I. a. Moosa, "Forecasting international tourist flows to Australia: a comparison between the direct and indirect methods," *Tourism Management*, vol. 26, no. 1, pp. 69–78, Feb. 2005.

[11] C. Goh and R. Law, "Modeling and forecasting tourism demand for arrivals with stochastic nonstationary seasonality and intervention," *Tourism Management*, vol. 23, no. 5, pp. 499–510, Oct. 2002.

[12] C. Lim and M. McAleer, "Time series forecasts of international travel demand for Australia," *Tourism Management*, vol. 23, no. 4, pp. 389–396, Aug. 2002.

[13] V. Cho, "Tourism Forecasting and its Relationship with Leading Economic Indicators," *Journal of Hospitality & Tourism Research*, vol. 25, no. 4, pp. 399–420, Nov. 2001.

[14] H. Hom, "Effect of seasonality treatment on the forecasting performance of tourism," vol. 15, no. 4, pp. 693–708, 2009.

[15] P. Gustavsson and J. Nordström, “The impact of seasonal unit roots and vector ARMA modelling on forecasting monthly tourism flows,” *Tourism Economics*, vol. 7, no. 2, pp. 117–133, Jun. 2001.

[16] J. du Preez and S. F. Witt, “Univariate versus multivariate time series forecasting: an application to international tourism demand,” *International Journal of Forecasting*, vol. 19, no. 3, pp. 435–451, Jul. 2003.

[17] F.-L. Chu, “Forecasting tourism demand with ARMA-based methods,” *Tourism Management*, vol. 30, no. 5, pp. 740–751, Oct. 2009.

[18] C.-L. Chang, S. Sriboonchitta, and A. Wiboonpongse, “Modelling and forecasting tourism from East Asia to Thailand under temporal and spatial aggregation,” *Mathematics and Computers in Simulation*, vol. 79, no. 5, pp. 1730–1744, Jan. 2009.

[19] F. L. Chu, “A fractionally integrated autoregressive moving average approach to forecasting tourism demand,” *Tourism Management*, vol. 29, no. 1, pp. 79–88, Feb. 2008.

[20] E. Smeral, “Does Complexity Matter? Methods for Improving Forecasting Accuracy in Tourism: The Case of Austria,” *Journal of Travel Research*, vol. 44, no. 1, pp. 100–110, Aug. 2005.

[21] R. Y. C. Fan, S. T. Ng, and J. M. W. Wong, “Reliability of the Box–Jenkins model for forecasting construction demand covering times of economic austerity,” *Construction Management and Economics*, vol. 28, no. 3, pp. 241–254, Mar. 2010.

[22] G. E. . Box and G. Jenkins, *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco CA, 1970.

[23] H. Akaike, “X (n) X (n).,” pp. 243–247, 1969.

[24] H. Akaike, “A new Lool at the Statistical Model Identification,” *IEEE Transactions on Automotic Control*, vol. ac 19, no. 6, 1974.

[25] J. Hannan, E, “The Estimation of the order of an ARMA process,” *Ann. Stat*, no. 8, pp. 1071–1081, 1980.

[26] J. Rissanen, “Modeling by shortest data description,” *Automatica 14*, pp. 465–471, 1978.

[27] G. Liang, D. M. Wilkes, and J. a. Cadzow, “ARMA model order estimation based on the eigenvalues of the covariance matrix,” *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3003–3009, 1993.

[28] S. . Sivanandam and N. Deepa, S, *Introduction to Genetic Algorithm*. Springer, 2008.

[29] A. I. Journal, “Coefficient Estimation of IIR Filter by a Multiple Crossover Genetic Algorithm,” vol. 51, pp. 1437–1444, 2006.

[30] J.-C. Hung, “A genetic algorithm approach to the spectral estimation of time series with noise and missed observations,” *Information Sciences*, vol. 178, no. 24, pp. 4632–4643, Dec. 2008.

[31] S. Bisgaard and M. Kulahci, *Time series analysis and Forecasting by Example*. Wiley, 2011.

[32] D. C. Frechtling, *Forecasting Tourism Demand : methods and strategies*. Butterworth Heinemann, 2002, p. 112.

[33] Y. Ibrahim, “Forecasting International Tourism Demand in Malaysia Using Box Jenkins Sarima Application,” vol. 3, no. 2, 2010.

[34] S. Makridakis, S. C. Wheelwright, and R. J. Hyndman, *Forecasting Methods and Application*, Third Edit. John Wiley & Sons, Inc, 2004.

[35] P. Alan, *Forecasting With Univariate Box Jenkins Models : Concepts and Cases*. John Wiley & Sons, Inc, 1983.

[36] J. E. Hanke and D. W. Wichern, *Business Forecasting*, Eight Edit. Pearson Prentice Hall, 2005.

[37] M. Negnevitsky, *Artificial Intelligence, A Guide to Intelligent Systems*, Third Edit. Pearson, 2011.

[38] D. E. Goldberg, *Genetic Algorithms in Search, Optimisation and Machine Learning*. Addison-Wesley. Reading, MA, 1989.

[39] L. Davis, *Handbook on Genetic Algorithms*. Van Nostrand Reinhold, New York, 1991.

[40] M. . South, G. . Wetherill, and M. . Tham, “Hitch-Hiker’s guide to genetic algorithms,” *Journal of Application statistics*, vol. 20, pp. 153–175, 1993.

[41] C. Gathercole and P. Ross, “Small populations over many generations can beat large populations over few generations in genetic programming,” *Proceedings of the second Annual Conference, July 13-16 Standford University*, 1997.

[42] Z. S. Abo-Hammour, O. M. K. Alsmadi, A. M. Al-Smadi, M. I. Zaqout, and M. S. Saraireh, “ARMA model order and parameter estimation using genetic algorithms,” *Mathematical and Computer Modelling of Dynamical Systems*, vol. 18, no. 2, pp. 201–221, Apr. 2012.