PARAMETER ESTIMATION OF BOX-JENKINS MODEL USING GENETIC ALGORITHM

NUR'AFIFAH BINTI RUSDI

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

JANUARY 2013

To my beloved parent and family,

thank you for everything.

ACKNOWLEDGEMENTS

In preparing this thesis, I am indebted to many people who have helped me and inspired me, in various ways to start, continue and complete this thesis. They have contributed towards my understanding and thought. In particular, I wish to express my sincere appreciation to my main thesis supervisor Prof Dr. Zuhaimy bin Hj. Ismail. His efforts for all necessary crucial information really much appreciate. In addition, I would like to thank Dr. Zaitul Marlizawati Zainuddin and Pn. Rashidah Ahmad for their beneficial comments on this work.

I also want to express my heartfelt thanks asspecially to my lovely and supportive parents, Rusdi bin Abu Bakar and Azarina binti Khanum for their endless support, love and pray.

My appreciation aso extends to Dr. Ali Sangawi who helped me a lot in doing this thesis. Lastly, I want to thank my lovely friends who give me undying support to complete this thesis.

Finally, I hope this work will benefit and useful for everybody who involved in this area. Thank you.

ABSTRACT

Malaysia is very fortunate to be free from natural disaster such as earth quake, volcano and typhoon. Unfortunately, the most severe natural disaster experiencing in Malaysia is flood. The probability of flood may occur had been increase due to the climate change and global warming that happened in Malaysia throughout the year. One of the major factor that contribute to flood is the heavy rainfall or maximum rainfall. Hence, in this study, mathematical analysis had been performed by studying the rainfall pattern of the past years and predict the future pattern. Ulu Sebol station situated in Johor was chosen as the rainfall data station since Johor is one of the state that experienced the worst flood in the year 2006. Accuracy plays an important role in choosing the forecasting techniques in order to make prediction of the future rainfall data. But, before forecasting can be made, estimation of the model parameter must be done. In this thesis, an approach that combines the Box-Jenkins methodology for ARIMA model and Genetic Algorithm (GA) had been introduced as a new approach in estimating the parameter and forecasting. A total of 127 series of data had been used in this study starting from January 2000 and these data were classified as monthly maximum rainfall data. MINITAB 16 computer package was used in analyzing the data and for the development of Box-Jenkins model. Meanwhile, JAVA was used in estimating the parameter of Box-Jenkins model by using Genetic Algorithm. The accuracy of the results were measured by concerning the minimum Mean Absolute Percentage Error (MAPE). By using MINITAB 16, ARIMA(0,1,1) was chosen as the best model that fits to the data. The best estimate of theta given by MINITAB is θ = 0.9857 with MAPE 0.6526. By adopting GA in searching the best parameter value, GA gives an outstanding performance with the best estimate of theta is 0.3427 and MAPE with 0.5416. Hence, Genetic Algorithm was proven to work well in estimating the parameter of Box-Jenkins model.

ABSTRAK

Malaysia bertuah kerana bebas daripada bencana alam seperti gempa bumi, gunung berapi dan taufan. Malangnya, bencana semulajadi yang paling teruk dialami di Malaysia ialah banjir. Kebarangkalian banjir berlaku meningkat disebabkan perubahan iklim dan pemanasan global yang berlaku di Malaysia. Antara faktor yang menyumbang kepada banjir ialah hujan lebat. Analisis secara matematik telah dilakukan dengan mengkaji corak hujan pada tahun-tahun lalu dan meramalkan corak hujan pada masa hadapan. Stesen Ulu Sebol, Johor dipilih sebagai stesen pengambilan data hujan kerana Johor merupakan salah satu negeri yang mengalami banjir terburuk pada tahun 2006. Ketepatan memainkan peranan penting dalam memilih teknik ramalan untuk membuat ramalan data hujan. Walaubagaimanapun, sebelum ramalan dibuat, penganggaran nilai parameter bagi model tersebut mesti dilakukan. Pendekatan yang menggabungkan kaedah Box-Jenkins untuk model ARIMA dan Algoritma Genetik telah diperkenalkan dalam menganggar parameter dan peramalan. 127 siri data telah digunakan dalam kajian ini bermula dari Januari 2000 dan data ini telah diklasifikasikan sebagai data maksimum hujan mengikut bulan. Pakej komputer iaitu MINITAB 16 telah digunakan untuk menganalisis data dan pembangunan model Box-Jenkins. JAVA telah digunakan dalam menganggar parameter bagi model Box-Jenkins dengan menggunakan Algoritma Genetik. Ketepatan keputusan bergantung kepada nilai minima purata peratusan kesilapan mutlak (MAPE). Dengan menggunakan MINITAB 16, ARIMA (0,1,1) telah dipilih sebagai model terbaik yang sesuai dengan data hujan ini . Anggaran terbaik theta yang diberikan oleh MINITAB adalah θ = 0.9857 dengan MAPE sebanyak 0.6526. Algoritma Genetik terbukti memberikan prestasi cemerlang dalam mencari nilai parameter dengan anggaran terbaik theta adalah 0.3427 dan nilai minima kesilapan sebanyak 0.5416. Algoritma Genetik terbukti berkesan dalam menganggar parameter untuk model Box-Jenkins.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GA	-	Genetic Algorithm
MAPE	-	Mean Absolute Percentage Error
MAE	-	Mean Absolute Error
MSE	-	Mean Square Error
AR	-	Autoregressive
MA	-	Moving Average
ARMA	-	Autoregressive Moving Average
I	-	Integrated
ARIMA	-	Autoregressive Integrated Moving Average
MLE	-	Maximum Likelihood Estimation
LSE	-	Least Square Estimation
ACF	-	Autocorrelation Function
PACF	-	Partial Autocorellation Function

LIST OF SYMBOLS

B	-	Backshift Operator
∇	-	Difference Operator
θ	-	Theta
ϕ	-	Phi
μ	-	Mean
δ	-	Mean
y_t	-	Observation at Time t
a_t	-	Random Shock
d	-	Order of Difference
t	-	Period of t
P_c	-	Probability of Crossover
P_m	-	Probability of Mutation
n	-	No of Population Size

CHAPTER 1

INTRODUCTION

1.1 Introduction

In a real world problem, forecasting is very significant in many type of organizations and the need of forecasting is increasing since predictions of future events must be incorporated into the decision making process. According to Bowerman and O'Connell (1993), predictions of future events and conditions are called forecasts and the act of making such predictions is called forecasting. In other word, forecasting plays an important role as it is needed to determine when an event will occur so that appropriate actions can be taken (O'Donovan, 1983). One such application are in business. In many events and situations, forecast is require such as in marketing departments, finance, production scheduling and others (Bowerman and O'Connell, 1993).

Forecasting are normally divided into two types namely Qualitative forecasting methods and Quantitative forecasting methods. The choice of forecasts method to be used depends on the availability of data and predictability of the quantity to be forecasted. Qualitative forecasting methods generally use the opinion of experts to predict future events (Makridakis *et al.*, 1998). This methods also used to predict changes in historical data patterns. As referred to Bowerman and O'Connell (1993), there are several common qualitative forecasting methods that have been used

such as subjective curve fitting, Delphi method and time independent technological comparisons.

On the other side, Quantitative forecasting methods involve the use of information of the historical data in an attempt to predict the future values (Makridakis *et al.*, 1998). Generally speaking, Quantitative forecasting methods can be categorized into two main parts namely causal models and time series models (O'Donovan, 1983). Causal models assume that the factor to be forecasted exhibits a cause. The aim of this model is to discover the form of the relationship and use it to forecast future values. Causal models give advantage especially in business world since they allow the management to evaluate the effect of various alternative policies (Bowerman and O'Connell, 1993).

For time series models, historical data are very valuable since the data will be used to identify the model. There are various type of time series models. One of them is Box-Jenkins model. This model is also known as Autoregressive Moving Average(ARMA). Suitable model is chosen from this class of model and it will be used for forecasting based on the study of the data. As compared to causal models, time series models has advantage in the sense when conditions are expected to remain constant (Bowerman and O'Connell, 1993).

In the last decades, imitating living things in solving optimization problems seems to be very popular among scientists and researchers. The father of the original Genetic Algorithm, John Holland invented the method in 1960 and it was then developed at University of Michigan in 1970 with his colleagues (Mitchell, 1996). The idea of Genetic Algorithm is to use the power of evolution to solve optimization problems.

According to Sivanandam and Deepa (2007), Holland proposed Genetic Algorithm as a heuristic method based on "survival of the fittest" and this method is very useful tool for search and optimization problems. Recently, Genetic Algorithm

received most attention and become popular because of their potential in solving complex problems. Apart from that, this method is widely used to solve many problems including scheduling and sequencing, reliability design, vehicle routing and others (Gen and Cheng, 1997). In past few years, exploration into forecasting and parameter estimation by using Genetic Algorithm has also increases.

1.2 Background Of Study

The method for forecasting, G.E.P. Box and G.M Jenkins introduced in 1960 and this method is widely known as Box-Jenkins method. According to Hoshmand (2009), there are five main factors why Box- Jenkins method is more suitable for the purpose of forecasting. Firstly, Box-Jenkins methodology is very suitable for any data patterns such as combination of a trend, random fluctuation and seasonal factors. Besides that, this method is able to identify the best model when there is given a set of data. Next, Box-Jenkins approach can handle with complex data patterns using relatively well specified rules. Fourth, statistical measurement can be used to test the reliability of forecasts. Lastly, Box-Jenkins is chosen instead of other forecasting methods because Box-Jenkins methodology does not make assumptions about the number of terms used in the models or the relative weight given to them.

In general, there are two main types of Box-Jenkins models namely seasonal and non-seasonal Box Jenkins models. Non-seasonal models consist of three main models, which are; Autoregressive models (AR), Moving Average model (MA) and Mixed Autoregressive Moving Average (ARMA). Box and Jenkins proposed a method known as Box-Jenkins methodology to ARIMA models. ARIMA model stands for an integration between AR and MA model.

Box-Jenkins is very popular for forecasting variety type of problems. The strength of this method is that it can be applied to any type of time series pattern. One of the early researches is in 1982 where this method is used to forecast U.S

Merchandise Exports. This study used both seasonal and non seasonal Box-Jenkins models (Dale and Bailey, 1982). Not only that, Lin *et al.* (1986) also trying to used Box-Jenkins approach to predict Louisiana's prison population. The result shows that the predictions are quite close to the actual values and sufficiently adequate to meet the needs of the correctional system for short-term planning.

Genetic Algorithm is an adaptive heuristic search based on evolutionary ideas. It simulates the survival of the fittest among individuals over consecutive generation for solving a problem. Basically, there are many other techniques that can be used for optimization problem. Genetic Algorithm is chosen instead of other methods because it gives bundle of benefits. Firstly, it is more robust. Even though there is changes in the inputs, Genetic Algorithm do not easily break.Besides that, GA are parallel as compared to other algorithms. Next, according to Sivanandam and Deepa (2007), the advantages of Genetic Algorithm are it is easy to discover global optimum, easily modified for different problems, the problem has multi objective functions and others. Other than that, it is noticed that most of the algorithms can only search the solution in one direction in a time. Meanwhile, GA can explore in various directions in a time. For example, if solution cannot be found or the path turn to be dead end, they can be eliminated and continue work using other paths that have possibilities to give optimal solution (Bajpai and Kumar, 2010).

According to Sivanandam and Deepa (2007), there are four major differences that exist between Genetic Algorithm and other conventional optimization techniques. One of them is Genetic Algorithm works with the coding of solution set and not with the solution itself. Not only that, this method differs from conventional optimization techniques in the way that it uses a population of solutions in each iteration (Deb, 2001). Besides that, GA uses fitness function for evaluation rather than derivatives. Lastly, Sivanandam and Deepa (2007) also noted that Genetic Algorithm use probabilistic transition operates compared to conventional methods that uses deterministic transition operates.

1.3 Problem Statemant

The main issue in Box-Jenkins modeling is the study of model parameter. The problem is, in estimating the parameter of Box-Jenkins model, it is a difficult task. The parameters in Box-Jenkins model are hard to estimate due to the large number of possible solutions. For non-seasonal model, the parameters that need to be estimated are ϕ and θ. Meanwhile, there are four parameters involve in seasonal model namely ϕ, θ, Φ and Θ. Nevertheless, this study will only focus on non-seasonal model which is ARIMA. The values of ϕ and θ are between 0 and 1.

There are many approaches that can be used to estimate the parameters such as Least Square Estimation, (LSE) and Maximum Likelihood Estimation (MLE). But, current predictions does not guarantee a good solution. Hence,this study explore the use of operational research tool namely Genetic Algorithm in estimating the parameters of Box-Jenkins model.So, Genetic Algorithm will be used in this research in searching the best value of ϕ and θ. Apart from that, this values will be used for finding the forecast accuracy.

1.4 Research Objectives

The objectives of this research are:

(i) To estimate the parameters in Box-Jenkins model by using Genetic Algorithm.

(ii) To define the best parameter estimate.

(iii) To implement the above model in forecasting the rainfall data.

1.5 Scope Of Study

The scope can be divided into three parts namely Box-Jenkins, forecast accuracy and Genetic Algorithm. For Box-Jenkins model, the research focus on ARIMA model since the data is non seasonal type data. The data will be collected from Department of Irrigation and Drainage Malaysia. There are 127 data and this data will be divided into two which are in-sample data and out-sample data. 115 data from in-sample data will be used to formulate the model and the balance data, which is out-sample data is used for calculating forecast accuracy.

Forecast accuracy plays an important role in order to ensure that the model that is chosen is the best model. There are number of ways to calculate forecast accuracy such as Mean Square Error (MSE), Mean Absolute Error (MAE), Mean Percentage Absolute Error (MAPE) and others. This study will only focus on MAPE method and the formula is as below:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|, \tag{1.1}$$

where

$$\begin{aligned} n &= \text{number of data}, \\ A_t &= \text{Actual values}, \\ F_t &= \text{Forecast values}. \end{aligned}$$

For Genetic Algorithm, this study will only use one point crossover as the operation and Roulette Wheel selection is chosen for the selection method. Four different population size will be used in this study which are 30, 50, 80 and 100. Besides that, empirical test will be used for the crossover probability and the values are 0.5, 0.7 and 0.9.

1.6 Significance Of Study

This research helps to improve the forecast accuracy. Other than that, the major contribution in this research is the implementation of Genetic Algorithm in estimating the parameters of Box-Jenkins model. The performance of Genetic Algorithm is important since it will be convenient to the government sector to predict floods.

1.7 Thesis Structure

This project consists of five chapters. This chapter includes the introduction, background of study, problem statement, research objectives, scope and the significance of the study. The next chapter will discuss the related works on Box-Jenkins demand, parameter estimation using Box-Jenkins, Genetic Algorithms demand and parameter estimation using Genetic Algorithms. Chapter 3 will briefly explain on how this study will be conducted. The implementation of Box-Jenkins and Genetic Algorithms in solving the problem will be discussed in detail in chapter 4. Chapter 5 will include the conclusion and suggestion for further research.

REFERENCES

Abo-Hammour, Z. S., Alsmadi, O. M.-K. and Al-Smadi, A. M. (2010). A Novel Technique for ARMA Modelling with Order and Parameter Estimation Using Genetic Algorithms.. In Zavoral, F., Yaghob, J., Pichappan, P. and El-Qawasmeh, E. (Eds.). *NDT (2)*. Vol. 88 of *Communications in Computer and Information Science*. Springer. 564–576.

Anderson, D., Sweeney, D., Williams, T., Camm, J. and Cochran, J. (2012). *Quantitative methods for business*. South-Western Pub.

Anderson, D., Sweeney, D., Williams, T. and Martin, R. (2007). *An introduction to management science: Quantitative approaches to decision making*. South-Western Pub.

Back, T., Fogel, D. and Michalewicz, Z. (1997). *Handbook of evolutionary computation*. IOP Publishing Ltd.

Bajpai, P. and Kumar, D. R. M. (2010). Genetic Algorithm–an Approach to Solve Global Optimization Problems. *Indian Journal of Computer Science and Engineering*. 1(3): 199–206.

Balakrishnan, N. (2010). *Methods and Applications of Statistics in Business, Finance, and Management Science*. Wiley.

Bandyopadhyay, S. and Pal, S. (2007). *Classification and learning using genetic algorithms: applications in Bioinformatics and Web Intelligence*. Springer Verlag.

Bowerman, B. and O'Connell, R. (1993). *Forecasting and time series: an applied approach.* Belmont.

BuHamra, S., Smaoui, N. and Gabr, M. (2003). The Box-Jenkins analysis and neural networks: prediction and time series modelling. *Applied Mathematical Modelling*. 27(10): 805–815.

Čepin, M. (2011). *Assessment of Power System Reliability: Methods and Applications*. Springer.

Chambers, L. D. (2001). *The Practical Handbook of Genetic Algorithms: Applications, Second Edition*. 2nd Edition. Boca Raton, FL, USA: CRC Press, Inc.

Chatfield, C. (2000). *Time-series forecasting*. CRC Press.

Chen, S. (2002). *Genetic algorithms and genetic programming in computational finance*. Springer.

Coley, D. (1999). *An introduction to genetic algorithms for scientists and engineers*. World Scientific Pub Co Inc.

Dale, C. and Bailey, V. (1982). A Box-Jenkins model for forecasting US merchandise exports. *Journal of International Business Studies*. 101–108.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms*. Vol. 16. Wiley.

ElMaraghy, H. (2011). *Enabling Manufacturing Competitiveness and Economic Sustainability: Proceedings of the 4th International Conference on Changeable, Agile, Reconfigurable and Virtual Production (Carv2011), Montreal, Canada, 3-5 October*. Springer Verlag.

Feigin, P., Kratz, M. and Resnick, S. (1996). Parameter estimation for moving averages with positive innovations. *The Annals of Applied Probability*. 6(4): 1157–1190.

Freitas, A. (2002). *Data mining and knowledge discovery with evolutionary algorithms*. Springer.

Gen, M. and Cheng, R. (1997). *Genetic Algorithms and Engineering Design*. Vol. 2. Wiley-Interscience.

Halim, S., Bisono, I., Sunyoto, D. and Gendo, I. (2009). Parameter estimation of space-time model using genetic algorithm. In *Industrial Engineering and Engineering Management, 2009. IEEM 2009. IEEE International Conference on*. IEEE. 1371–1375.

Hill, A. (2011). *The Encyclopedia of Operations Management: A Field Manual and Glossary of Operations Management Terms and Concepts*. Ft Press.

Hoshmand, A. (2009). *Business Forecasting: A Practical Approach*. Routledge.

Janacek, G. and Swift, L. (1993). *Time series: forecasting, simulation, applications*. Ellis Horwood New York.

Janeiro, F. and Ramos, P. (2009). Application of genetic algorithms for estimation of impedance parameters of two-terminal networks. In *Instrumentation and Measurement Technology Conference, 2009. I2MTC'09. IEEE*. IEEE. 602–606.

Kamrani, A. and Nasr, E. (2010). *Engineering design and rapid prototyping*. Springer.

Kureychik, V., Malyukov, S., Kureychik, V. and Malioukov, A. (2009). *Genetic Algorithms for Applied CAD Problems*. Vol. 212. Springer Verlag.

Lawrence, K., Klimberg, R. and Lawrence, S. (2009). *Fundamentals of forecasting using excel*. Industrial Press Inc.

Lee, Y., Park, S., Chang, D. et al. (2006). Parameter estimation using the genetic algorithm and its impact on quantitative precipitation forecast. In *Annales geophysicae*. Vol. 24. 3185–3189.

Liang, Y. (2009). Combining seasonal time series ARIMA method and neural networks with genetic algorithms for predicting the production value of the mechanical industry in Taiwan. *Neural computing & applications*. 18(7): 833–841.

Lin, B., MacKenzie, D. and Gulledge, T. (1986). Using ARIMA models to predict prison populations. *Journal of Quantitative Criminology*. 2(3): 251–264.

Liu, Y., Wang, D. and Ding, F. (2010). Least squares based iterative algorithms for identifying Box-Jenkins models with finite measurement data. *Digital Signal Processing*. 20(5): 1458–1467.

Madsen, H. (2008). *Time series analysis*. Vol. 77. CRC Press.

Makridakis, S., Wheelwright, S. and Hyndman, R. (1998). *Forecasting methods and applications*. Wiley-India.

Mandal, B. (2005). Forecasting sugarcane Production in India with ARIMA model. *Interstat. www. interstat. statjournals. net*. .

Martinez, E., Silva, E. and Fabbro, A. (2011). A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*. 44(4): 436–440.

Mashohor, S., Evans, J. and Arslan, T. (2004). Genetic Algorithms Based to Printed Circuit Board (PCB) Inspection System. In *IEEE International Symposium Consumer Eletronics*.

Mitchell, M. (1996). An introduction to genetic algorithms. *PHI Pvt. Ltd., New Delhi*. .

Morad, N. (2000). Genetic Algorithms Optimization for the Machine Layout Problem. *International Journal of the Computer, the Internet and Management*. 8(1).

Nochai, R. and Nochai, T. (2006). ARIMA model for forecasting oil palm price. In *2nd IMT-GT Reg. Conf, on Mathematics, Statistic and Apps., Univ. Sains Malaysia, Penang, June*.

Noraini, M. and Geraghty, J. (2011). Genetic Algorithm Performance with Different Selection Strategies in Solving TSP. . .

O'Donovan, T. (1983). Short term forecasting: An introduction to the Box-Jenkins approach.. *JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 256*. .

Promprou, S., Jaroensutasinee, M. and Jaroensutasinee, K. (2006). Forecasting dengue haemorrhagic fever cases in Southern Thailand using ARIMA Models. *Dengue Bulletin*. 30: 99.

Rajasekaran, S. and Pai, G. (2004). *Neural networks, fuzzy logic and genetic algorithms: synthesis and applications*. PHI Learning Pvt. Ltd.

Sharma, P., Chandra, A. and Kaushik, S. (2009). Forecasts using Box–Jenkins models for the ambient air quality data of Delhi City. *Environmental monitoring and assessment*. 157(1): 105–112.

Sivanandam, S. and Deepa, S. (2007). *Introduction to genetic algorithms*. Springer Publishing Company, Incorporated.

Sivaraj, R. and Ravichandran, T. (2011). A review of selection methods in genetic algorithm. *International Journal of Engineering Science and Technology*. 3(5): 3792–3797.

Tho, D. (2010). *Genetic Algorithms and Application in Examination Scheduling*. GRIN Verlag.

Wang, D., Dai, J. and Ding, F. (2009). Gradient-based iterative parameter estimation for Box-Jenkins systems with finite measurement data. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE. 239–243.

Winston, W. and Albright, S. (2009). *Practical management science*. South-Western Cengage learning.