BOX-JENKINS AND GENETIC ALGORITHM HYBRID MODEL FOR ELECTRICITY FORECASTING SYSTEM

KHAIRIL ASMANI B. MAHPOL

UNIVERSITI TEKNOLOGI MALAYSIA

BOX-JENKINS AND GENETIC ALGORITHM HYBRID MODEL FOR ELECTRICITY FORECASTING SYSTEM

KHAIRIL ASMANI B. MAHPOL

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

OCTOBER 2005

To my beloved mum, Asiah bt. Yin, my father, Mahpol Hj. Tahir
and to all my brothers, Khairil Izwan, Khairul Anuar, Khairil Azhar, Khairil Azrin, Khairil Azizi
and my sister, Juliana. Thanks for everythings.

ACKNOWLEDGEMENT

Alhamdulillah, by the grace of ALlah S.W.T and the Prophet Muhammad S.A.W, this research is finally completed. It is my greatest pleasure to take this opportunity to express my gratitude and thank you to all the people involved either direct or indirectly in my effort to successfully finish this project.

A deepest gratitude to my supervisor, Assoc. Prof. Dr. Hj. Zuhaimy Hj. Ismail who conducting and advising me on the excel of this research, my beloved parent, brothers and sister who support me and also to all my friends, thank you for your helps and advices that motivate me in completing this master degree. I also indepted to Universiti Teknologi Malaysia (UTM) and Ministry of Science, Technology and Environment for funding my study under UTM-PTP and IRPA Grant Vot. 74047.

To ensure this research follow the standard of academic, I have made thorough reference to books and papers that are relevant to the research title and depth fitering on any sources obtained through mass and electronic media.

Finally, I hope this work will benefited and useful for everybody especially who involved in this area. Any errors, omissions or obscurities which remain are entirely of my personal responsibility. I will be glad to hear from any reader who wishes to make constructive comments. Wassalam, thank you.

ABSTRACT

Energy is considered a prime agent in the generation of wealth and also a significant factor in economic development. There has been a strong relationship between the availability of energy to the economic activity, improvements in standards of living and the overall social well-being. In making a forecast for energy demand, accuracy is the primary criteria in selecting among forecasting techniques. Time Series method has always been used in a variety of forecasting applications. In this thesis, an approach that combines the Box-Jenkins methodology for SARIMA model and Genetic Algorithm (GA) will been introduced as a new approach in making a forecast. Data used in this study were collected from the year 1996 until year 2003 that has been classified into total monthly electricity generated in kWh unit. GA is widely known as a multi-purpose searching procedure commonly use in optimization and approximation field. The increasing popularity of GA is due to their adaptability and simplicity as a problem solution especially when they are applied into several complex problems. By adopting the GA blind search, the algorithm combines searching techniques and their capabilities to learn about the relationship of the pattern-recognition of the past data. This character helps GA to make a prediction of future values. This study proposed the possibility of using GA's approach as one of the unique forecasting method. It also represents a preliminary work in the current research and practices of GA. The investigation is simulated using Intelligent Electricity Forecasting System (IEFS) developed in this research which written in Borland Delphi 7.0 programming.

ABSTRAK

Tenaga merupakan sumber utama kekayaan, tidak dinafikan juga sebagai keperluan asas pembangunan ekonomi sesebuah negara. Kajian membuktikan penggunaan sumber tenaga berkadar langsung dengan aktiviti-aktiviti ekonomi seterusnya menjadi pemangkin pembangunan keperluan infrastruktur serta taraf hidup masyarakat. Ketepatan merupakan faktor utama bagi pemilihan sesebuah kaedah peramalan. Kaedah Peramalan Siri Masa sering kali digunakan bagi memenuhi penyelesaian di dalam pelbagai masalah peramalan. Selaras dengan itu, kajian akademik telah dijalankan ke atas data bulanan jumlah tenaga elektrik yang dihasilkan dalam negara bagi tujuan peramalan. Tesis ini mengetengahkan penyelesaian kaedah peramalan menggunakan satu pendekatan baru yang menggabungkan dua model iaitu SARIMA Box-Jenkins dan Algoritma Genetik. Algoritma Genetik (GA) dikenali sebagai kaedah pencarian kepintaran buatan yang diaplikasikan di dalam pelbagai penyelesaian pengoptimuman dan penghampiran. Peningkatan penggunaan algoritma ini adalah disebabkan ianya mudah diadaptasi serta dapat meringkaskan penyelesaian sesuatu masalah terutamanya apabila diaplikasikan terhadap masalah pengoptimuman yang rumit. Dengan menggunakan kaedah carian rawak GA, algoritma ini menggabungkan kaedah pencarian evolusi genetik berserta kebolehan kepintaran menelaah hubungan dan bentuk data dalam skop yang dispesifikasikan oleh pengguna. Ciri-ciri ini membantunya membuat telahan bagi masa hadapan. Di dalam penyelidikan ini, asas algoritma genetik telah diintegrasikan bagi membantu kaedah peramalan Box-Jenkins. Kajian literasi secara menyeluruh telah dibuat meliputi kajian-kajian terdahulu serta terkini berkaitan algoritma ini. Bagi merealisasikan hasil penyelidikan, sebuah sistem Intelligent Electricity Forecasting System (IEFS) yang berfungsi mensimulasikan peramalan telah dibangunkan dengan menggunakan perisian Borland Delphi 7.0 berorientasikan-objek Pascal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

GA	-	Genetic Algorithm
EP	-	Evolutionary Programming
ESs	-	Evolution Strategies
GUI	-	Graphical User Interface
MSE	-	Mean Square Error
MAPE	-	Mean Absolute Percentage Error
MAD	-	Mean Absolute Deviation
SSE	-	Sum Square Error
RMSE	-	Root Mean Square Error
SE	-	Standard Error
AR	-	Autoregressive
MA	-	Moving Average
I	-	Integrated
IMA	-	Integrated Moving Average
ARMA	-	Autoregressive Moving Average
ARIMA	-	Autoregressive Integrated Moving Average
SMA	-	Seasonal Moving Average
SAR	-	Seasonal Autoregressive
SARIMA	-	Seasonal Autoregressive Integrated Moving Average
MLE	-	Maximum Likelihood Estimation
LSE	-	Least Square Estimation
TNB	-	Tenaga Nasional Berhad
PTM	-	Pusat Tenaga Malaysia
OOP	-	Object-Oriented Programming

RW	-	Roulette Wheel
SC	-	Selected Chromosome
C/M	-	Crossover or Mutation Probability Rate
kWh	-	Kilowatt Hour Unit
fc	-	Final Electricity Consumption
ct	-	Consumption of Energy Transformed
dl	-	Losses of Electrical Energy
IEFS	-	Intelligent Electricity Forecasting System
STF	-	Short Term Forecast
IPP	-	Independent Power Producer
IRP	-	Integrated Resource Planning
ACF	-	Autocorrelation Function
PACF	-	Partial Autocorrelation Function
WAGs	-	Wild-Assed Guesses
$C.I$	-	Cyclical Irregular Component
$S.I$	-	Seasonal Index
$T.C$	-	Trend Component
TSP	-	Traveling Salesman Problem
SSR	-	Stochastic Sampling with Replacement
IDEs	-	Integrated Development Environments
APIs	-	Applications Programming Interfaces
MB	-	Mega Byte
GB	-	Giga Byte
PC	-	Personal Computer
GHz	-	Giga Hertz
RAM	-	Read Access Memory
NTFS	-	Network File System
SQL	-	Structured Query Language
O/S	-	Operating System
B	-	Backshift Operator
∇	-	Difference Operator

a_t , ε_t	-	A Series of Shocks Generated by a White Noise
Ln	-	Natural Logarithm
$p.d.f$	-	Probability Density Function
Cov	-	Covariance
ρ_k , r_k	-	Autocorrelation at lag-k
ϕ_{kk}	-	Partial Autocorrelation Function
A_t	-	Overall Smoothed Estimate
B_t	-	Trend Estimate
Sn_t	-	Seasonal Index Estimate
μ	-	Mean
t	-	Period of t
N	-	Recursive Iteration of ‘N’ Number

LIST OF APPENDICES

LIST OF PUBLICATIONS

NO.	TITLE
1	Zuhaimy Ismail & Khairil Asmani Mahpol (2005), *Emulation of Biological Reproduction System on Forecasting Model*, International Symposium on Bio-Inspired Computing, SCRG FSKSM, September 7-9, 2005, Puteri Pan-Pacific Johor Bahru.
2	Zuhaimy Ismail & Khairil Asmani Mahpol (2005), *An Intelligent Electricity Forecasting System (IEFS)*, Laporan Teknik FSUTM, LT/M Bil. 5/2005
3	Zuhaimy Ismail & Khairil Asmani Mahpol (2005), *$SARIMA_T$ Model for Forecasting Malaysian Electricity Generated*, Jurnal Matematika, UTM : UTM.26/18.11/1/21 Jld.8 (66).
4	Zuhaimy Ismail, Mohd Fuad Jamaluddin and Khairil Asmani Mahpol (2004), *Forecasting of Electricity Demand using ARIMA Models*, Proceeding of The Conference Focus Group, RMC UTM, Pulai Spring Golf Resort, Johor Bahru.
5	Zuhaimy Ismail, Khairil Asmani Mahpol & Mohd Fuad Jamaluddin (2004), *Fitting $SARIMA_T$ Model for Forecasting Malaysian Electricity Generated*, Laporan Teknik FSUTM, LT/M Bil. 5/2004.
6	Zuhaimy Ismail, Mohd Fuad Jamaluddin & Khairil Asmani Mahpol (2003), *Forecast of Energy Generated Using GA – A University-Industry Collaborative Research*, Proceeding of The Conference UNIPRENEUR2003, MARA, *11-12 June, 2003* Renaissance, Kuala Lumpur.
7	Zuhaimy Ismail & Khairil Asmani Mahpol (2003), *Forecasting of Malaysian Electricity Production using Winter's Method*, Proceeding of The Annual Fundamental Science Seminar 2003, IIS, May 20-21, 2003, Puteri Pan-Pacific Johor Bahru.

CHAPTER 1

INTRODUCTION

This chapter introduces the reader to the author's research project. It explains the problem background and description, research objective, research scope, research importance, research data, research contribution and the significance of this project. It then gives a brief description of the structure of the thesis organization.

1.1 Introduction to Forecasting

Forecasting is defined as an attempt to predict future events. It is also described as a concern of what the future will look like, that act as an aid for effective and efficient planning.

The layperson may question the validity and efficacy of a discipline aimed at predicting an uncertain future. However, it should be recognized that substantial progress has been made in forecasting over the past several centuries. There are a large number of phenomena whose outcomes can now be predicted easily such as the time of sunrise, the speed of a falling object, the trajectory of a satellite, rainy weather, and a myriad of other events (Makridakis et. al, 1998).

There are two categories of forecasting namely qualitative and quantitative. Qualitative forecasting is a very useful technique when there are no historical data. This method usually uses the opinions of experts to predict the future events subjectively via their experiences and knowledge. The most popular qualitative techniques are the Delphi method, subjective curve fittings and time independent technological comparison. Quantitative forecasting is a technique that can be applied when there are enough historical data to be modeled in making prediction. This technique involves a depth analysis of the historical data in developing a model. There are two types of quantitative techniques that are univariate and explanatory model. The common time series analysis is normally categorized into both of these types that will be described in details in our literature review.

Forecasting is a very important tool applied in a lot of sectors such as energy consumption, demand and supply, economy, business, sales and market analysis, weather forecast, engineering research and development, science and technology, management and development, industrial, politic and many more. It is very useful because domain experts often have knowledge of events whose effects have not been observed yet in a time series.

In the last decade, there has been increasing interest in imitating living beings to solve hard optimization problems. Simulating the natural evolutionary process of human beings result in stochastic optimization techniques called evolutionary algorithms, which can often outperform conventional optimization method when applied to difficult real life problem. Exploration into forecasting using heuristic method has increase for the past few years, namely the Genetic Algorithm (GA), Evolutionary Programming (EP), and Evolution Strategies (ESs). Among them, GA is one of the most widely known types of evolutionary algorithm today (Mitsuo and Runwei, 1997).

In the early 1970s, John Holland has introduced the concept of simulating the process of natural evolution known as GA in computer. Holland's GA was represented by a sequence of procedural steps for moving from one population of artificial chromosomes to a new population. It uses natural selection and genetic inspired techniques known as crossover and mutation. Each chromosome consists of a number of genes that each gene is represented by binary digit of 0 and 1.

GA has received considerable attention regarding its potential as an optimization technique for solving complex problems. It has been successfully applied in the area of industrial engineering that includes scheduling and sequencing, optimization, reliability design, vehicle routing, inventory control, transportation and many more. In this research, the researcher introduces GA as an approach for estimating the parameters in Box-Jenkins method of forecasting.

1.2 Problem Background

Box-Jenkins approach was developed in the 1960s for handling most time series of data analysis. This approach is about to match one Autoregressive Integrated Moving Average (ARIMA) model based on the historical data that currently available in making forecast. Box and Jenkins (1970) effectively put together in a comprehensive manner of the relevant informations required to understand and use univariate time series ARIMA models.

For stationary series, ARIMA models can be the Autoregressive (AR) or Moving Average (MA) or the combination of these two models known as Autoregressive Moving Average (ARMA). If the data is non-stationary, a simple modification of the ARMA model known as integrated (I) processes will be performed to produce an ARIMA model. If there is seasonality in the data, the models then become either Seasonal Autoregressive (SAR) or Seasonal Moving Average (SMA) or Seasonal Autoregressive Integrated Moving Average (SARIMA). Therefore the general Box-Jenkins model that allows seasonality is:

$$\phi_p(B)\Phi_P\left(B^T\right)\nabla^d\nabla_T{}^D\,\tilde{z}_i=\theta_q(B)\Theta_Q\left(B^T\right)a_i \qquad (1.1)$$

and is referred to as the multiplicative $(p,d,q)\times(P,D,Q)_T$ model where:

ϕ_p = unknown p^{th} AR parameter,

θ_q = unknown q^{th} MA parameter,

Φ_P = unknown P^{th} SAR parameter,

Θ_Q = unknown Q^{th} SMA parameter,

z_i = time series value at time i,

a_i = error term at time i,

B = backshift operator,

∇ = difference operator.

The value of the coefficient ϕ_p, θ_q, Φ_P and Θ_Q are restricted to lie between –1 and +1. Note the minus and plus sign on model is a convention for ARIMA models. The error, a_i, is normally distributed with a mean of 0 and variance of 1.

In conventional practices we are using a combination of a statistical maximum likelihood estimation (MLE) approach and the Newton-Raphson method to determine these parameters value. Finding MLE conceptually involves two steps. First, the likelihood function must be calculated. Secondly, the value of θ must be found that maximize this function. Then we use the Newton-Raphson numerical optimization to calculate the log likelihood function that is often known as the grid search method.

Many researchers have made comparative studies between GA and other searching techniques, such as random search, grid search, iterated search and simulated annealing. For example, grid search can be a very good method when there is a single unknown parameter to estimate. However, it quickly becomes intractable when the number of elements of θ becomes large.

Grid search method is generally referred to as hill - climbing. They can perform well on functions with only one peak. But on functions with many peaks, they suffered from the problem that the first peak found will be climbed, and this may not be the highest peak. Having reached the top of a local maximum, no further progress can be made (Beasley et. al, 1993). Usually global optimum can be found only if the problem possesses certain convexity properties that essentially guarantee that any local optimum is a global optimum.

1.3 Problem Statement

As describe in the previous section, after we have identified the appropriate ARIMA model, we proceed to estimate its parameters;

$$\left(\hat{\phi},\hat{\theta}\right)\equiv\left(\hat{\phi}_1,....,\hat{\phi}_p,\hat{\theta}_1,....,\hat{\theta}_q\right) \qquad (1.2)$$

that minimizes the shock sum of squares:

$$S(\phi,\theta)=\sum_{1}^{N}\alpha_i^2 \quad (1.3)$$

Where the $\alpha_i = \theta^{-1}(B)\phi(B)z_i$ are the estimated shocks given the model and the series. As described in problem background, it is normal to use a non-linear least squares procedure or MLE to obtain the vector of parameter estimates within the stationary invertible region.

According to Box and Jenkins (1970) the parameters value for ARIMA models are represented by the unknown ϕ_p, θ_q, Φ_P and Θ_Q where their values are restricted to lie between –1 and +1. In producing a forecast with the lowest possible error, we need to search the value for these unknown parameters which minimizes the Mean Square Error (MSE). Therefore the main aim of this research are;

1. To study the efficiency of using Genetic Algorithm (GA) to estimate the parameters ϕ_p, θ_q, Φ_P and Θ_Q in SARIMA models.
2. To develop an improved method based on GA for univariate forecasting model.

The ergodicity of evolution operators makes genetic algorithm very effective in performing the global search (Mitsuo and Cheng, 1997). This research proposes an alternative approach to the conventional searching method. One such method is the GA which have a potential in giving the optimize value for weighing ϕ_p, θ_q, Φ_P and Θ_Q thus improve the forecast result.

1.4 Research Objectives

The main aim of this study is centers around a few fundamental questions such as "What is the role of forecasting process been developed for predicting electricity generated? Would improve methods give any advantages over current practices? And how statistical method can be integrated to lead to substantial gains in accuracy?"

1.5 Research Importance

This research is highly significant for improving forecast accuracy. Implementation of GA heuristic search in estimating the parameters in SARIMA model is the major contribution in this study. This work also may improve the univariate Box-Jenkins forecasting model using GA, which may lead in producing a new forecast method using genetic approach.

1.6 Research Scope

1. This research focuses only on the univariate model where the time series based on the past value.
2. Forecast accuracy will be defined by measure the lowest error in term of MSE.
3. In Genetic Algorithm architecture, chromosomes are randomly generated using crossover and mutation operator.
4. Forecast simulation will be developed using Borland Delphi 7.0.

1.7 Research Data

The applications chosen for this study is forecasting the amount of Malaysian electricity generated at power plant. Data used here were a secondary data collected from 1996 to 2003, which has been classified into the total monthly electricity generated in kilowatt hour, kWh unit proposed by Tenaga Nasional Berhad (TNB) as the standard format unit recommended in the energy sectors. The flow of electricity represented by the following equations:

$$\textit{Electricity Generated} = fc + ct + dl$$
$$= \textit{Gross Inland Consumption}$$

Where (*fc*) is the final electricity consumption refers to the total quantity of electricity delivered to the end-user. The (*ct*) is the consumption of energy transformed while (*dl*) refers to losses of electrical energy, which occur outside the utilities and power plants, and the consumption of electricity by utilities and power plants for operating their installation.

1.8 Research Contribution

The main contribution in the research is the development of an algorithm to search for the best value of ϕ_p, θ_q, Φ_P and Θ_Q in SARIMA model to minimize the MSE in making a forecast based on the GA.

The second contribution is the development of a mathematical model for Malaysian electricity generated using Box-Jenkins methodology.

1.9 Thesis Organization

This thesis contains six chapters. The first chapter will introduce the problem background and description, research objective, research scope, research importance, research data, research contribution and the significance of this project.

The next chapter is a literature review. In this chapter it contains a discussion on the forecasting using time series method and GA. This chapter will describe the review of the literatures that are related to the past and present study on the electricity generated for Malaysia, time series forecasting and GA procedures, models and applications. It begins by reviewing literature related to time series forecasting model especially Box-Jenkins methodology for ARIMA models. Some comparison study on how selecting the appropriate time series model also discusses in this chapter. After that, researcher made some literature on the problem of searching the SARIMA parameters. Finally, a review of literatures related to GA procedures and application in solving complex searching problem is presented. In GA section, it contains an introduction to GA, a comparison study done by other researcher on GA with other searching technique, the GA operators, encoding and decoding of a chromosome, selection and the advantages of GA.

The research methodology is described in chapter three. This chapter involved reviewing the existing literature to develop a theoretical framework to form a structure for this research. The chapter begins with describing the theory of time series analysis and the Box-Jenkins methodology for ARIMA model. It then discusses the theory of GA and how it can be applied into Box-Jenkins forecasting.

Chapter four contains the implementation of forecast method at case study. Here, all the sample data of TNB for Malaysian electricity generated is analyzed and used to calculate the forecast step-by-step. All the calculations using hybrid of Box-Jenkins methodology and GA is described in detail in this chapter.

Chapter five discusses on the system develop by the researcher. Intelligence Electricity Forecasting System (IEFS) was developed to realize this research output as an effective decision tool program into computer.

Chapter six presented the results, discussions, conclusions and recommendations for further research. It gives a summary of this research and then discusses in further detail some of results and findings. Some conclusions are drawn and finally, some thoughts on possible directions in which future research in this area might be pursued are offered.